UNIVERSITY OF CALIFORNIA

Los Angeles

# Advancing Temporal Modeling and Heterogeneous Data Analysis for Digital Health

A dissertation submitted in partial satisfaction of

the requirement for the degree

Doctor of Philosophy in Bioengineering

by

Yiwen Meng

2020

# ABSTRACT OF THE DISSERTATION

Advancing Temporal Modeling and Heterogeneous

Data Analysis for Digital Health

by

Yiwen Meng

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2020

Professor Corey Wells Arnold, Committee Co-Chair

Professor William F Speier, Committee Co-Chair

Recent development in electronic medical devices or systems has realized the effective collection and documentation of patients' health in real time. To date, the potential clinical impact of this healthcare data has not been fully realized. Specifically, patients' health data is heterogenous and sparse in nature, as it is composed of various modalities and is collected on different scales. In addition, processing this data efficiently in a temporal manner to take advantage of its sequential structure remains a barrier for medical records. This dissertation attempts to overcome these challenges by developing machine learning models to classify patient reported outcome (PRO) scores from activity tracker data and predict depression diagnoses based on data from patients' historical electronic health records (EHR). A temporal model based on hidden Markov models (HMM) is first proposed to classify PRO scores in various categories from human vital signs collected from Fitbit activity trackers. This approach is able to combine various vital signs on difference scales in a single model that tracks changes in PRO scores over time. Second, several end-to-end machine learning models were built to aggregate multimodal EHR data in a single model. A novel hierarchical embedding method achieved superior performance for predicting depression diagnosis, which lays a foundation for addressing the heterogeneity and sparsity of EHR data. Third, an innovative bidirectional sequence learning model with a transformer architecture was developed for representation learning on high dimensional EHR data, demonstrating significantly improved performance over the traditional forward-only method. Finally, methods to improve the interpretability of the aforementioned models have been developed, which is a critical step before clinical deployment. Relative feature importance factors are determined for each vital sign collected from the Fitbit and attention weights are found for each data modality in the sequential EHR data. Extensive experiments and results have demonstrated the effectiveness of these proposed methods. This dissertation provides methodologies that advance modeling and understanding of digital health datasets, which lays the foundation to construct clinical decision support systems in this domain which could potentially lead to early disease detection and intervention.

The dissertation of Yiwen Meng is approved.

Alex Anh-Tuan Bui

William Hsu

Michael K. Ong

Corey Wells Arnold, Committee Co-Chair

William F. Speier, Committee Co-Chair

University of California, Los Angeles

2020

Dedicated to my beloved family, friends, and the medical & engineering community!

And you shall know the truth, and the truth shall set you free.

--- John 8 : 32

For the commandment is a lamp, and the teaching a light, and the reproofs of instruction are the way of life.

--- Proverbs 6 : 23

# Table of Contents

# List of Figures

# List of Tables

# ACKNOWLEDGEMENTS

I would like to express my sincere and upmost appreciation and respect to my advisor, Dr. Corey Arnold, for his generous guidance and support for my PhD. I really appreciate his trust for me to join his lab and providing me with these interesting and challenging projects throughout my PhD. As my teacher and mentor, he has taught me so many technical and management skills—more than I could ever credit him for. I also want to express my gratitude to the co-chair of my committee, Dr. William Speier for his enormous patience and detailed instruction on my research, especially for invidual meetings of paper revision. I am thankful for all the time, effort, resources and support that he and Dr. Arnold have put in for my research as well as my personal growth. In addition, I would like to offer my genuine thanks and respect to the committee members, Dr Alex Bui, for introducing the opportunity in the Medical Imaging Informatics (MII) group and providing critical insights and comments for my research; Dr. William Hsu for his kindness and significant advice to improve this work; and Dr. Michael Ong for his support and help of providing clinical background information.

I am grateful to every current and past members in MII and the Computational Diagostics (CDx) group for holding a delightful working environment. In particular, I am thankful to Isabel Rippy, Denise Luna, Esther Lan and Patrick Langdon for their timely help on the administrive matters. Thank you to all the lovely and inspiring current and past labmates, Jiayun Li, Wenyuan Li, Karthik Sarma, Zichen Wang, Chao Chen, Jennifer Pelson, Haoyue Zhang, Leihao Wei, Tianran Zhang and Yannan Lin, Panayiotis Petousis, Johnny Ho, Shiwen Shen, Nova Smedley, Edgar Rios. Also, I'm thankful to Dr.Wentai Liu, Dr. Yikai Lo and Dr. Ying Li from for providing me with an opportunity to build a background in machine learning. I would like to express my sincere gratitude to Dr. Warren Grundfest who offered me the original funding and admission to UCLA. Last but not the least, I would like to thank UCLA for giving me an opportunity to pursue my PhD degree and make contributions to the machine learning and medical research community. I would also like to thank NIH for funding my research through grants R01 HL141773 and R01 HL135425.

Most importantly, I would like to express my gratitude to my parents Li Han and Yulin Meng for their greatest and devoted love, support and sacrifice throughout my education. Meanwhile, I am thankful to following people who have offered me necessary help and support throughout the time: Timothy Zhu, Yong Hu, Lumei Xu, Yuqian Chen, Xiaoyu Che, Hangbo Yang, Yanchao Yang, Haiwen Gao, Geer Chen, Kelly Yu and ZionYu's family, Yao Li and Tong Wu's family and Yingchun Wang and Hongchang Lin's family and Gibert Chang's family.

As this dissertation includes contents of the following articles, I would like to thank, and re-thank, all co-authors for their contributions:

Chapter 3 contains materials published in "Meng Y, Speier W, Shufelt C, et al. A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease Using Activity Tracker Data. IEEE Journal of Biomedical Health Informatics 2020, which is developed from one initial work published in "Meng Y, Speier W, Dzubur E, Spiegel B and Arnold C. W. Predicting Patient Health Status using Activity Tracker Data. AMIA Annual Symposium Proceeding, San Francisco, November 2-7, 2018".

Chapter 4 contains materials published in "Meng Y, Speier W, Ong M and Arnold C.W. Predicting Depression from Electronic Health records using Machine Learning Algorithms. Journal of Medical System. Submitted in May 2020"

Chapter 5 contains materials published in "Meng Y, Speier W, Ong M and Arnold C.W. HCET: Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Record for Predicting Depression. IEEE Journal of Biomedical Health Informatics. June 2020", which is developed from my initial work published in “Meng Y, Speier W, Ong M and Arnold C.W. Multi-Level Embedding with Topic Modeling on Electronic Health Records for Predicting Depression. AAAI Health Intelligence Workshop, New York, February 7-12, 2020”

Chapter 6 contains materials published in “Meng Y, Speier W, Ong M and Arnold C.W. Bidirectional Representation Learning with Transformer on Multimodal EHR for Chronic Disease Prediction. IEEE J Biomed Heal Informatics. Submitted in September 2020”.

Chapter 7 contains the specific components of improving interpretability of the models constructed through Chapter 3 to 6.

# VITA

| | |
|---|---|
| 2010-2012 | B.S. (Electronical Engineering), Department of Microelectronics and Solid State Electronics, University of Electronic Science and Technology of China, Chengdu Sichuan, China |
| 2012-2014 | B.S. (Electronical Engineering), Iowa State University |
| 2014-2015 | M.S. (Electronical Engineering), Iowa State University |
| 2015-2020 | Graduate Student Researcher, Department of Bioengineering, the Computational Diagnostics lab and Medical Imaging Informatics group, UCLA |
| 2019 | Deep Learning Engineer Intern (summer), Bosch |

# PUBLICATIONS

**Meng Y**, Speier W, Dzubur E, Spiegel B and Arnold C. W. Predicting Patient Health Status using Activity Tracker Data. AMIA Annual Symposium Proceeding, San Francisco, November 2-7, 2018

**Meng Y**, Speier W, Shufelt C, et al. A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease Using Activity Tracker Data. IEEE J Biomed Heal Informatics 2019;24:878–84. doi:10.1109/JBHI.2019.2922178

**Meng Y**, Speier W, Ong M and Arnold C.W. Multi-Level Embedding with Topic Modeling on Electronic Health Records for Predicting Depression. AAAI Health Intelligence Workshop, New York, February 7-12, 2020

**Meng Y**, Speier W, Ong M and Arnold C.W. Predicting Depression from Electronic Health records using Machine Learning Algorithms. Journal of Medical System. Submitted in May 2020

**Meng Y**, Speier W, Ong M and Arnold C.W. MLET: Multi-level embedding with topic modeling on electronic health records for predicting depression. Explainable AI in Healthcare and Medicine (pp. 241-246). Springer. Cham.

**Meng Y**, Speier W, Ong M and Arnold C.W. HCET: Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Record for Predicting Depression. IEEE J Biomed Heal Informatics. June 2020

**Meng Y**, Speier W, Ong M and Arnold C.W. Bidirectional Representation Learning with Transformer on Multimodal EHR for Chronic Disease Prediction. IEEE J Biomed Heal Informatics. Submitted in September 2020

# CHAPTER 1

# Introduction

## 1.1 Motivation

Rapid development in the field of digital technology has brought benefits to clinical domains, such as remote monitoring by ambulatory devices and clinical decision support systems (CDSS) based on electronic health records (EHR). Specifically, the advancement of machine learning algorithms could facilitate monitoring patients' health status and future diagnosis for several diseases, among which ischemic heart disease (IHD) and depression have high prevalence as well as high cost for treatment and monetary loss due to missed work. IHD is a major health problem worldwide and the top cause of death [1]. Approximately one third of adults in the United States (about 81 million) has some form of cardiovascular disease, including more than 17 million with coronary artery disease and nearly 10 million with angina pectoris [2,3]. IHD patient care cost $156 billion in the United States for both direct and indirect costs in 2008 [1]. Depression also has high prevalence as epidemiological studies have estimated that roughly 17% of Americans experience symptoms of depression during their lifetime [4]. The economic burden of depression was $210.5 billion in 2010, composed of direct medical costs and monetary loss due to disability [5]. Therefore, the large amount of medical data collected by activity trackers and EHR provides an opportunity for machine learning algorithms to explore and discover clinically relevant information. The main contribution and motivation of this work is to apply machine learning algorithms to constructure classification or predictions to certain clinical applications: 1) to build a remote monitoring system for IHD patients with wearable activity trackers; 2) to build a clinical decision support system for predicting future diagnosis of depression using patients' EHR.

### 1.1.1 Remote monitoring for patients with ischemic heart disease

IHD, also called coronary artery disease (CAD), is a type of cardiovascular disease caused by narrowed arteries that results in less blood and oxygen reaching heart muscle, which can ultimately lead to myocardial infarction (MI) [1]. IHD is a major public health problem in United States and worldwide. Approximately 25% of men and 16% of women have coronary artery disease among people in their 60s and 70s, and these figures rise to 37% and 23% among men and women in their 80s or older, respectively [2]. Although the survival rate of patients with IHD has been steadily improving, it was still responsible for nearly 380,000 deaths in the United States during 2010 [3].

Furthermore, IHD is the top cause of death in both men and women, among whom this condition accounts for 27% of deaths [1]. IHD accounts for the vast majority of the mortality and morbidity of cardiac disease, with more than 1.5 million patients having an MI each year. Many more are hospitalized for unstable angina and for evaluation and treatment of stable chest pain symptoms. Beyond the need for hospitalization, many patients with symptoms of chronic chest pain are temporarily unable to perform normal activities for hours or days and thus experiencing a reduced quality of life. IHD continues to be associated with considerable patient morbidity despite the decline in cardiovascular mortality rate. Patients who have had acute coronary syndrome, such as acute myocardial infarction, remain at risk for recurrent events even if they have no, or limited, symptoms and should be considered to have stable IHD (SIHD).

Despite the high prevalence, the costs of caring for patients with IHD are enormous, estimated at $156 billion in the United States for both direct and indirect costs in 2008. More than half of direct costs are related to hospitalization [1]. Another major expense is for invasive procedures and related costs as well as $13 million for outpatient visits for IHD that occur in the United States annually [4]. The estimated costs of outpatient and emergency department visits in 2000 by patients with chronic angina were $922 million and $286 million, respectively, and prescriptions accounted for $291 million. Long-term care costs, including skilled nursing, home health, and hospice care, were $2.6 billion, which represented 30% of the total cost of care for chronic angina [5]. Although the direct costs associated with SIHD are substantial, they do not account for the significant indirect costs of lost workdays, reduced productivity, long-term medication, and associated effects. The indirect costs have been estimated to be almost as great as the direct costs [2].

Therefore, much effort has been concentrated on increasing the diagnosis accuracy and risk assessment of SIHD both from clinical and technological perspectives. Recently, rapid developments in digital technology have led to biomedical applications for personal healthcare, where an increasing number of clinical trials have focused on the potential of using mobile devices and activity trackers in medicine [6]. This technology enables collecting human vital signs from people's daily lives with high accuracy, which was not feasible until recently [7,8]. For instance, studies have demonstrated the feasibility of the Fitbit Charge, a common commercial activity tracker, to record user's step counts and heart rate in real time [9]. However, few studies have sufficiently utilized activity tracker data to provide clinically relevant information on users' health [10,11]. Data collected from this minimally invasive method could potentially provide more reliable estimates of patient health status over time.

### 1.1.2 Early diagnosis of depression using electronic health records

Depression is a major cause of disability worldwide, often leading to a number of adverse outcomes, including increased risk of self-harm, premature mortality, and the development of comorbid general medical conditions, such as heart disease, stroke, and obesity [12]. Within a year of experiencing depressive symptoms, patients are 4.4 times more likely to develop major depressive disorder (MDD) [13], a heterogeneous spectrum disorder with a variety of onsets, treatment responses, and comorbidities. The economic burden of individuals with MDD was $210.5 billion in 2010, which increased from $173.2 billion in 2005. A large portion of this increase was attributed to higher direct medical costs and presenteeism, defined as being present but not fully functional in the workplace [14]. Depressive symptoms can be effectively improved by treatment with antidepressants [15] or psychotherapy to mitigate adverse outcomes. However, linking patients to care necessarily requires accurate and timely detection of depression by a qualified medical or mental health provider. Identifying depression in the primary care setting, particularly in patients with multiple comorbidities, can be inefficient. Thus, screening with self-reported questionnaires has emerged as an approach to aid primary care providers in identifying patients who may have depression but who do not have a diagnosis yet.

Despite high prevalence and high cost, the current diagnosis or screening method of high risk patients only generated a 50% true positive rate [16]. For instance, one prospective cohort studies found that only 17 patients (1% of the 1687 screened) started treatment for major depressive disorder. Screening for depression in high risk populations was thus deemed to be ineffective. The authors attributed this result mainly to low rates of treatment initiation [17]. The current evidence suggests that screening alone is not an effective strategy to improve the quality and outcomes of care. It is promising that screening in primary care carries important benefit when primary care practices can support accurate diagnosis, effective treatment and appropriate follow-up.

Recently, adoption of electronic health records (EHRs) has increased dramatically over the past several years [18]. As a consequence, there is an increasing number of healthcare providers equipped with computerized support tools that are explicitly designed to help health professionals comply with recommended clinical guidelines. These systems consist of heterogeneous data modalities, such as patient demographic information, diagnoses, procedures, medications or prescriptions, clinical notes, and medical images [19]. CDSS, a common feature of EHR systems, provide users with automated prompts when specified tests or screenings are indicated. A recent systematic review of research exploring the impact of CDSS on depression care yielded primarily positive results [20]. Notably, these results indicated that using CDSS can increase the number of individuals screened for depression. Thus, improving the diagnostic accuracy of depression prediction could allow patients to be screened at an early stage, enabling the possibility of careful monitoring or early intervention

## 1.2 Contributions

Both activity tracker data and EHR are in time series formats, which are similar to data structure of word sequences. Thus, recent advancement in sequential and temporal machine learning models developed in fields such as natural language processing (NLP) can be applied to medical datasets [21–23]. This dissertation aims to investigate the feasibility of constructing specific temporal models for each clinical application that can perform representation learning on time series activity tracker and EHR data. The main goal of the constructed models is to deliver clinical outcomes or predictive information for ischemic heart disease and depression while presenting novel approaches to alleviate the sparsity and heterogeneity nature of medical datasets. This proposed research work has three aims:

1. *To develop a temporal machine learning model to process human vital signs collected by activity trackers and provide information about a patient's health status.* This framework was

one of the earliest approaches to realize the clinical impact of vital signs collected from one activity tracker (Fitbit) by using them to classify PRO scores. A hidden Markov model (HMM) represented the change in PROs using transition probabilities, demonstrating superior performance over classical non-temporal machine learning algorithms. This framework also provided flexibility to be applied to multiple categories of PROs in both physical and mental health, such as physical function, anxiety and depression. This work demonstrates that data generated from activity trackers may be used in a machine learning framework to classify self-reported health status variables. These techniques could play a future role in larger frameworks for remotely monitoring a patient's health state in a clinically meaningful manner.

2. *To construct a single predictive model based on machine learning algorithms capable of aggregating multimodal EHR data to predict depression*. A novel Hierarchical Clinical Embedding with Topic modeling (HCET) architecture was proposed to address the heterogeneity and sparsity of EHR data while building a temporal model to predict depression onset at various prediction windows prior to diagnosis. The HCET model was able to aggregate various categories of EHR data and learn their inherent structure within hospital visits. The prediction performance was further improved after applying bidirectional learning with a transformer architecture, showing that bidirectional representation learning developed on word sequences can also be applied on longitudinal EHR sequence. The results demonstrated the model's ability to utilize heterogeneous EHR information to predict depression, which may have future implications for screening and early detection.
3. *To improve model's interpretability by revealing each feature's contribution to the classification task in Aim 1 and the importance factor of every modality of EHR data as well as the order of clinical visits in Aim 2*. Improving a model's interpretability is an important

area of machine learning research, especially in the biomedical domain. To address this problem, a random forest model being able to output feature importance factor were added to the model built on Aim 1, which revealed the relative feature importance for classifying PRO scores. Additionally, adding attention weights to HCET improved its interpretability by showing the relative importance of each EHR modality. Furthermore, adjusting the novel transformer architecture on EHR data and outputting self-attention for each code in sequences indicated the relationship between them in the task of predicting depression. Making these models more explainable allows physicians to better understand how decisions are made, which is an important step if they are going to be integrated as a clinical decision support system in the future.

Towards Aim 1, a machine learning framework was proposed to classify PRO scores from data collected from Fitbit activity trackers within a population of patients with stable ischemic heart disease (SIHD). The developed framework comprised two steps: 1) constructing an end-to-end machine learning system that uses data collected from activity trackers to predict or classify patient's health status assessed by clinical metrics; and 2) building a hidden Markov model (HMM) to process temporal data using learned transition probabilities of patient reported outcome (PRO) scores. The first model treated each week independently, whereas the second used an HMM to take advantage of correlations between successive weeks. After training, a retrospective analysis compared the classification accuracy of two models. The results demonstrated the ability of utilizing activity tracker data to classify patients' PRO scores over time and suggested that patients' health status can be monitored in real time by activity trackers.

For Aim 2, an EHR dataset was constructed that was composed of five data modalities: diagnoses codes, procedure codes, medications, demographic information, and clinical notes. The developed architecture is comprised of four steps: 1) investigating the effect of the length of a patient's medical history on depression prediction accuracy; 2) implementing a single model that aggregates multimodal EHR data and predicts future depression diagnosis; 3) constructing a temporal model capable of processing multimodal EHR data in a sequential manner; and 4) revising the representation learning of EHR sequences from a forward-only approach to a bidirectional forward-backward method to improve the model's prediction accuracy. Models constructed in the first two steps was able to aggregate multimodal EHR data into a single model for depression prediction, but they did not effectively resolve sparsity and heterogeneity nature of the data. Hence, HCET efficiently alleviated these challenges by constructing a hierarchical structure on multimodal EHR data with various embedding levels, while preserving the data's sequential nature. In this way, it learns the inherent interaction between EHR data from various sources within each visit and across multiple visits for an individual patient. Switching the architecture to a transformer for bidirectional representation learning significantly improved the prediction accuracy. These models could possibly be used as the basis for constructing a screening tool by utilizing the models' predictions to intervene with individuals who have a higher risk of developing depression.

Aim 3 was an extension work to models built on Aim 1 and Aim 2. The feature importance factor was enabled on a random forest model which discovered the relative contribution of each feature collected by activity tracker in the classifying of PRO scores. Secondly, attention weights were applied on the code level embedding of HCET model to reveal the contribution of each EHR data modality in predicting depression. In addition, applying the self-attention and multi-head mechanism within a transformer architecture further improved model's interpretability by exploring the relationship between various data modalities and clinical visits.

Collectively, the design and implementation of these Aims results in PRO classification and depression prediction models with high interpretability that presents a deeper understanding of vital signs collected by activity tracker and EHR representation learning for chronic disease prediction.

## 1.3 Organization of the Dissertation

The dissertation is organized as follows. Chapter 2 describes technical background information on activity trackers, EHR data, temporal and non-temporal machine learning models, text modeling techniques and metrics to evaluate them. Chapter 3 presents work on building models to classify patients' PRO scores from activity tracker data, including the advantage of using temporal models as described in Aim 1. Models constructed in Chapter 4 addresses first step of Aim 2 to predict depression onset and Chapter 5 describes the advanced deep learning models covering the second and third step of Aim 2. The work in Chapter 6 addresses the last step of Aim 2 by adjusting the BERT, originally developed in natural language processing to perform bidirectional learning on EHR data to predict depression. Chapter 7 is composed of the works on improving the interpretability of models built in the previous chapters, which addresses Aim 3. Chapter 8 concludes by summarizing the results, discussing the limitations of this work and suggesting future directions.

# CHAPTER 2

# Background

## 2.1 Wearable activity trackers

Activity trackers, also referred to fitness trackers, activity monitors or fitness bands are devices or applications for monitoring and tracking fitness-related metrics such as distance walked or run, calories consumed, and heart rate [24]. The word "wearable" refers to monitors that can be worn on the wrist or clipped to an individual's clothing, not including smartphones. Consumer activity monitors, such as Fitbit, Apple watch, Jawbone UP, Garmin Vivofit and others are now widely used in biomedical research to study therapeutic effects of self-monitoring, exercise therapy and behavioral interventions [25]. Fitness trackers provide an easy interface for adults to meet those guidelines and over 100 million units were sold in 2016 [24].

An intense area of research aims at estimating the association between physical activity and metabolic function as well as cognitive and neurological health using consumer activity trackers. Accurate and precise self-monitoring devices therefore provide potential benefits both to patients by providing real-time feedbacks on their specific physiological status and to healthcare providers, since they can collect and present a full set of information, including activity frequency, duration, intensity, heart rate (HR), and energy consumption. Previous works have validated the accuracy of heart rate monitoring specifically in the Fitbit Charge 2 [24]. The Fitbit hardware and its computational algorithms for calculating step counts and physical activity have been validated using other Fitbit devices [26,27]. The Fitbit Charge 2 estimates activity using metabolic equivalents (METs), which are calculated based on heart rate and distance traveled.

A key metric measured by fitness trackers is HR, namely the number of contractions of the heart per minute (bpm). Physical exercise, sleep, anxiety, stress, illness, and ingestion of drugs are all factors known to alter the normal HR, thus, HR has been used as an indicator of physiological adaptation and intensity of effort [28]. Methods used to detect changes in HR include: electrocardiogram (ECG), blood pressure, ballistocardiogram (BCG) and the pulse wave signal derived from a photoplethysmography (PPG). Recently, the need for affordable, simple and portable technology for both the primary care and community-based clinical settings, together with the wide availability of low cost and small semiconductor components, have raised attention around PPG [29]. PPG is an optical measurement technique that measures the amount of backscattered infrared light through a tissue to assess the variation of blood volume and thus the heart rate [29]. PPG requires a light source and a detector while their relative positions may vary. In quantitative PPG, the optical illumination in the measuring area is automatically adjusted for

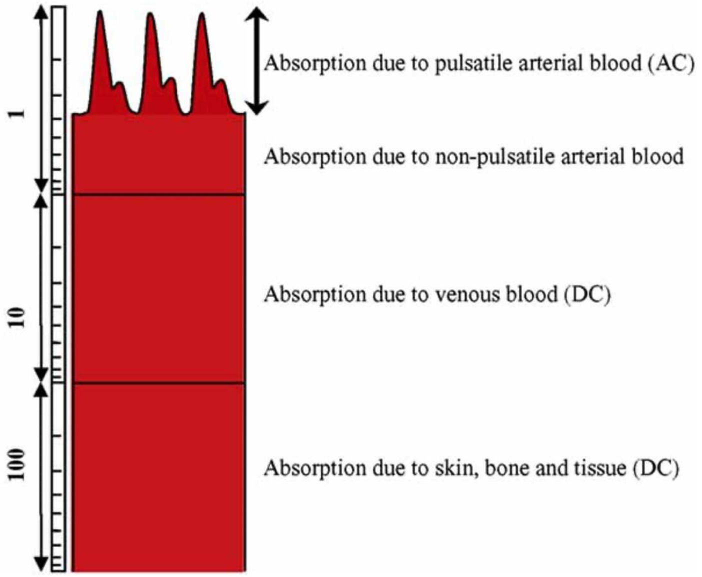

1
10
100
Absorption due to pulsatile arterial blood (AC)
Absorption due to non-pulsatile arterial blood
Absorption due to venous blood (DC)
Absorption due to skin, bone and tissue (DC)

Figure 2.1: The steady and pulsatile components of the PPG signal [30].

each type of skin until a predetermined level of reflected light is reached. With this technology, PPG measurements are independent of skin color, thickness and individual blood volume [30]. As shown in Figure 2.1, the PPG signal consists of AC and DC components. The AC component (for arterial pulse detection) is synchronous with the heart rate and depends on the pulsatile blood volume changes [31]. It has been suggested that the AC component is related to pulsatile blood volume changes because of varying lumen of the vessel and red cell orientation during each cardiac cycle [32]. Conversely, the DC component (commonly used for venous evaluation) of the signal varies slowly and reflects variations in the total blood volume of the examined tissue [33]. Recent studies suggest PPG is a simple, reliable and low-cost optical technique for measuring changes in blood volume in the microvascular bed of tissue with acceptable validity [34], although the accuracy often depends on the device used, the type and intensity of activity, and skin photosensitivity [35,36]. All wrist worn activity trackers rely on PPG to derive HR and several studies have investigated the accuracy of wearable devices for measuring HR recently.

Accelerometry is an objective method of quantifying physical activity and energy expenditure during the day and night [37]. Integrated chip sensors typically comprise a capacitive micro-electro-mechanical, piezoelectric, or piezo-resistive element that detects the change in acceleration of a small mass in the sensor. Raw data from the accelerometers (housed in both research-grade physical activity monitors and consumer physical activity monitors) are utilized in algorithms to count steps and sometimes to calculate parameters for other activities. Although the hardware and especially the algorithms vary, the newer research-grade physical activity monitors and consumer physical activity monitors generally employ a triaxial accelerometer, which records acceleration vectors in three planes, and represents a significant advance over devices using a uniaxial accelerometer in their ability to calculate energy expenditure [38]. Algorithms are continually improving to interpret this data to estimate the frequency, duration, and intensity of physical activity in steps, flights of stairs, or energy expenditure. One method involves using vertical and anterior-posterior (front and back) accelerations to calculate step frequency and determine walking patterns while using mediolateral (side-to-side) accelerations to predict stride frequency, which can also be used to assess abnormalities in gait [39].

## 2.2 Electronic health records (EHRs) systems

Like paper files, EHRs document the patient's health data, possibly including, but not limited to, patient encounter; demographic information; problem lists; active and past diagnoses; laboratory test orders and results; current prescriptions; radiological images and reports; hospitalization information; consultant reports; immunizations; pathology reports; social history; allergies; health screening study results; and physician, nurse, social worker, and physical therapy notes [40]. Therefore, EHRs are usually composed of heterogeneous data structures and modalities. In addition to having this information at the healthcare provider's fingertips in a searchable format on any securely connected computer, EHRs can include additional functionality such as access to clinical and public health guidelines, reminders about routine screenings or disease reporting responsibilities, and graphical display of trends in key parameters such as blood glucose for diabetic patients or blood pressure measurements in hypertensive patients. Some EHR systems can generate practice-level statistics.

Despite this promising array of possible features, there has been a lack of standardization and many EHRs have been developed with various designs and functionality. For example, a recent survey of office based physicians found that only 60.9% of EHRs could easily generate a list of patients by diagnosis, only 48.2% could easily track a patient referral to completion, and only

51.4% could easily generate a report on quality measures [40]. Therefore, it is significantly important to bring standardization and interoperability to EHRs. One purpose of EHRs is to enable providers to share patient information so that care can be delivered seamlessly across different settings and separate encounters. This practice helps avoid duplicate tests, prevents drug-drug interactions and enhances patient care. EHRs can also enable patients to access their records remotely and to use that information to better manage their health status and healthcare [40]. To ensure that EHRs reach their potential, networks are being developed to link EHRs so that healthcare providers can share information needed for care and patients can access their own records electronically. Such health information exchange systems, referred to in some states as regional health information organizations (RHIOs), also make it possible for public health workers to access EHRs to collect legally mandated disease reports.

While primarily designed for improving healthcare efficiency from an operational standpoint, many studies have found secondary use for clinical informatics applications [41]. In particular, the patient data contained in EHR systems has been used for tasks such as medical concept extraction [42], patient trajectory modeling [43], disease inference [44], clinical decision support systems [45] and more.

## 2.3 Non-temporal machine learning models

With the recent development of data collection, data storage and computational power machines, traditional machine learning algorithms have been widely adopted and advanced in many areas such as search ranking [46] and fraud detection [47]. In the meantime, they are also applied to the biomedical field, such as mortality prediction [48], disease classification [49] and drug discovery [50]. These algorithms include logistic regression, support vector machine, gradient boosting regression tree and random forest, which make decisions independently each time.

### 2.3.1 Logistic regression (LR)

LR is a regression method for predicting a dichotomous dependent variable. In producing the LR equation, the maximum-likelihood ratio was used to determine the statistical significance of the variable. LR is useful for situations to predict the presence or absence of a characteristic or outcome based on values of sets of predictor variables [51]. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. LR model for $m$ independent variables can be written as

$$P(Y = 1) = f = \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}} \qquad (2.3.1)$$

, where $P(Y = 1)$ is the model's estimated probability for the positive label $Y$, and $\beta_0, \beta_1, \dots, \beta_m$ are regression coefficients, multiplied on each corresponding feature $X_i$ . Equation (2.3.1) is also named as a sigmoid function. There is a linear model hidden within the logistic regression model. The natural logarithm of the ratio of $P(Y = 1)$ to 1- $P(Y = 1)$ gives a linear model in $X_i$:

$$g(X) = ln \frac{P(Y = 1)}{1 - P(Y = 1)} = \beta_0 + \sum_{i=1}^{m} \beta_i X_i \qquad (2.3.2)$$

The function $g(X)$ has many of the desirable properties of a linear regression model. The independent variables can be a combination of continuous and categorical variables.  LR models can include the main effects and interaction terms. An important step in the process of modeling a set of data is determining whether there is evidence of interactions and confounder terms in the data. The term confounder is used to describe a covariate that is associated with both the dependent variable of interest and a primary independent variable. When both associations are present, the relationship between the independent variable and the dependent variable is said to be confounded. Any clinically important change in the estimated coefficient for the independent variable suggests that the covariate is a confounder and should be included in the model, regardless of the statistical significance of its estimated coefficient. One way to test for confounders and interactions in LR is to start with a main effects model and use a forward selection method to find interaction terms which significantly reduce the likelihood ratio test statistic [51]. Regularization can be applied on LR to reduce overfitting. The penalized log-likelihood function becomes:

$$\sum_{i=1}^{n}[Y_i log(f_i) + (1 - Y_i)\log(1 - f_i)] - \frac{1}{2\tau^2}||\beta||^{\alpha} \qquad (2.3.3)$$

, where $||\beta||$ is the norm of coefficients $\beta_0, \beta_1, \dots, \beta_m$ and $f_i$ is the model's logit for each label $Y_i$. $\tau \in (0, \infty)$ is the shrinkage parameter that controls the degree of shrinkage of $\beta_i$ toward 0 [52]. The term $\alpha$ determines the order of regularization. The whole model is called L1 norm or Lasso LR when $\alpha = 1$ while L2 norm or Ridge LR when $\alpha = 2$. One major difference between Lasso and Ridge is that, the former would shrink some coefficients to 0 and induce feature sparsity while the latter one generally provides non-zero coefficients.

### 2.3.2 Support vector machine (SVM)

SVM is a class of supervised learning algorithms that trades off accuracy for generalization error. SVMs build a hyperplane which divides examples such that examples of one class are all on one side of the hyperplane, and examples of the other class are all on the other side [53]. For one sample data point $(x_i\ ,y_i)$, where the vectors $x_i$ are in a dot space $H$ and $y_i$ are class labels. Formally, any hyperplane in $H$ is defined as

$$\{x \epsilon H| < w, x > + b = 0\}\ \ w \epsilon H, b \epsilon R \qquad (2.3.4)$$

, where $w$ is a vector orthogonal to the hyperplane and < > represents the dot product. The idea of SVM is to find the hyperplane that maximizes the minimum distance from any training data point as shown in Figure 2.2. The following constraint problem describes the optimal hyperplane.

$$\underset{w \epsilon H, b \epsilon R}{min} \ \tfrac{1}{2}\left||w|\right|^2,\ \ subject\ to\ y_i(< w, x > + b) \geq 1 \qquad (2.3.5)$$

The above problem can be solved by introducing the Lagrange multipliers $(\alpha_i \geq 0)$ which maximize the dual problem

$$\max\ W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\, \alpha_j y_i y_j < x_i, x_j > \qquad (2.3.6)$$

$$subject\ to\ \alpha_i \geq 0\ and \sum_{i=1}^{m} \alpha_i y_i = 0$$

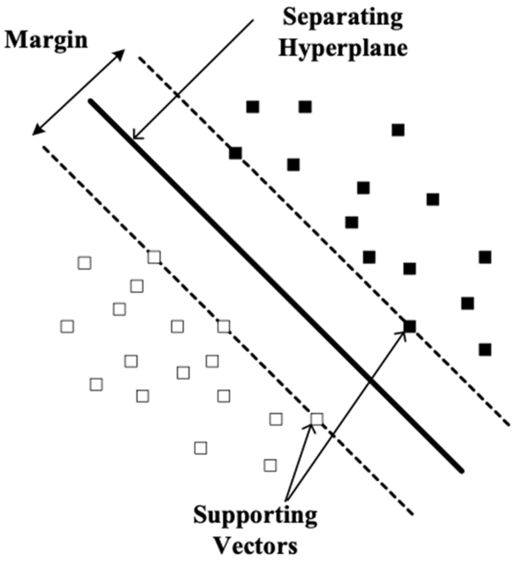


Figure 3.2: Maximum margin and optimal hyperplane.

The patterns $x_i$ that correspond to non-zero Lagrange coefficients are called support vectors. The resultant decision function has the following form

$$y(x) = sign\left(\sum_{i=1}^{m} \alpha_i y_i < x_i, x_j > + b\right) \tag{2.3.7}$$

Thus, the optimal margin hyperplane is represented as a linear combination of training points. Consequently, the decision function for classifying points with respect to the hyperplane only involves dot products between points. The algorithm that finds a separating hyperplane in the feature space can be stated entirely in terms of vectors in the input space and dot products in the feature space [53]. When the samples are not linearly separable, a kernel function is used to transform the data into a higher dimensional space where it is linearly separable. The kernel function gives the dot product of two examples in the higher dimensional space without actually transforming them into that space. This notion, dubbed the kernel trick, allows us to perform the transformation for the purpose of classification to large dimensional spaces. In the nonlinear case, the resultant decision function becomes

$$y(x) = sign\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x_j) + b\right) \tag{2.3.8}$$

, where the kernel function is a nonlinear mapping from the original space to the high dimensional space. Gaussian radial basis functions (RBF) kernel is one of the most commonly used kernel functions, formulated as

$$K(x, x') = e^{-\frac{||x-x'||^2}{2\sigma^2}} \tag{2.3.9}$$

, where $\sigma$ is the spread of the Gaussian function.

### 2.3.3 Gradient boosting regression tree (GBRT)

GBRT is an ensemble method on top of decision tree models, which constitute a highly interpretable machine learning technique, using a set of instances with known input and output variables to train a model that can both be applied for classification or regression tasks [54]. The key idea of GBRT consists of building a predictive model from an ensemble of simple models. The output of the ensemble is computed as the weighted sum of $M$ weak models. Formally, the output of GBRT for a test sample $x$ is computed as

$$H(x) = \sum_{i=1}^{M} \alpha_i h_i(x) \tag{2.3.10}$$

, where $h_i(x)$ is the output of the $i$-th weak learning model and $\alpha_i$ is a constant known as the stage length, which controls the contribution of the $i$-th model to the overall output. As opposed to other ensemble methods, these base models are not all trained to produce the same desired output. Instead, models are built sequentially, refining the output of the previous stage:

$$H_m(x) = H_{m-1}(x) + \alpha_m h_m(x) \tag{2.3.11}$$

Therefore, at each learning stage the current weak model $h_m(x)$ and the stage length $\alpha_m$ are trained to approximate the difference between the output of the previous stage and the desired overall output (i.e. the residuals). Of course, in the case of GBRT, the base models consist of regression trees with an advantage of allowing for the optimization of arbitrary differentiable loss functions. In each stage, the algorithm trains a set of binary regression trees on the negative gradient of the binomial or multinomial deviance loss function. GBRT is a generalization of boosting to arbitrary differentiable loss functions. They have good predictive power and robustness to outliers in the output space, but have increased complexity and phase scalability restrictions.

### 2.3.4 Random forest (RF)

RF is another ensemble method built upon decision tree models. It is based on the bagging technique instead of the boosting method used in GBRT, which enables parallel computing instead of sequential learning. The bagging method generates $K$ different training data subsets from the original dataset using a bootstrapping sampling approach, and then $K$ decision trees are built by training each subset. A random forest is finally constructed from these decision tree learners. Each sample of the testing dataset is predicted by all decision trees, and the final classification result depends on the votes from these trees [55].

In the step of bagging, $K$ training subsets are sampled from the original training dataset $S$ by bootstrapping. Namely, $N$ records are selected from $S$ by a random sampling and replacement method during each sampling. After this step, $K$ training subsets are constructed as a collection of training subsets $S_{Train} = \{S_1, S_2, \dots, S_K\}$. In an RF model, each meta decision tree is created by classification or regression tree algorithms from each training subset $S_i$. In the growth process of each tree, a subset of $m$ features out of the total set of $M$ are randomly selected to train each dataset $S_i$, whereas other models always all use the total $M$ features for training. This random selection of subsets of features explains why this algorithm is called random forest. In the node's splitting process of each tree, the gain ratio of each feature variable is calculated and the best one is chosen as the splitting node. This splitting process is repeated until a leaf node is generated [55]. Finally, $K$ decision trees are trained from $K$ training subsets in the same way. The $K$ trained trees are collected into an RF model, which is defined as

$$H(X, \Theta_j) = \sum_{i=1}^{K} h_i(x, \Theta_j), \qquad j \in [1, M] \qquad (2.3.11)$$

, where $h_i(x, \Theta_j)$ is a meta decision tree classifier, $x$ is the input feature vectors of the training dataset, and $\Theta_j$ is an independent and identically distributed random vector that determines the growth process of the tree. The random feature selection process leads to a feature decorrelating effect so that the most correlated features cannot dominate building the model for each subset, and when combined with the bagging technique, RF effectively reduces the model's variance.

Furthermore, RF has another advantage of providing better interpretation in the decision-making process than aforementioned models. It is able to output the relative feature importance factor by the information gain (IG). IG is usually computed by two methods: Gini index or entropy. The feature importance factor is computed by Gini importance. For a node $\tau$ within a binary tree $T$ of the random forest, the optimal split is sought using the Gini impurity $\varphi(\tau)$, which is a computationally efficient approximation of the entropy, measuring how well a potential split is separating samples of the two classes in this particular node [56]. Suppose there are $n$ samples in this node, where $n_0$ are negative and remaining $n_1$ are positive. The relative frequencies of the negative and positive samples, $f_0$ and $f_1$, respectively, are defined as

$$f_0 = \frac{n_0}{n}, f_1 = \frac{n_1}{n} \tag{2.3.12}$$

and the Gini impurity of this node $\varphi(\tau)$, can be calculated as

$$\varphi(\tau) = 1 - f_0^2 - f_1^2 \tag{2.3.13}$$

Then, for a split at this node that yields two sub-nodes $l$ and $r$, the decrease of the Gini impurity for this split is calculated as

$$\Delta\,\varphi(\tau) = \varphi(\tau) - f_l\varphi(\tau_l) - f_r\varphi(\tau_r) \tag{2.3.14}$$

, where $f_l$ and $f_r$ are the fractions of samples in that fall into $l$ and $r$, respectively [57]. Since the split happens on a certain feature $v$, this decrease in Gini impurity is also defined as the Gini decrease for $l$ and $r$ at the node $\tau$. Moreover, $v$ may be used as the splitting variable in more than one node. Let $I(\tau, v)$ be the indicator function that is equal to 1 when $v$ is the splitting variable of $\tau$ and 0 otherwise. The Gini decrease (GD) of $v$ in this tree is then defined as the summation of GDs for all nodes in which $v$ is the splitting variable, as

$$GD(T, V) = \sum_{\tau \epsilon N_T} \Delta\, \varphi(\tau) I(\tau, v) \tag{2.3.15}$$

, where $N_T$ is the collection of all nodes of the tree $T$. Finally, the summation of all GDs of $v$ over all trees in the forest is the Gini importance of $v$, as

$$GI(T, V) = \sum_{T} \sum_{\tau \epsilon N_T} \Delta\, \varphi(\tau) I(\tau, v) \tag{2.3.16}$$

, where $T$ is the collection of all decision trees in the random forest. Therefore, the random forest can output the Gini importance for each feature.

### 2.3.5 Training techniques

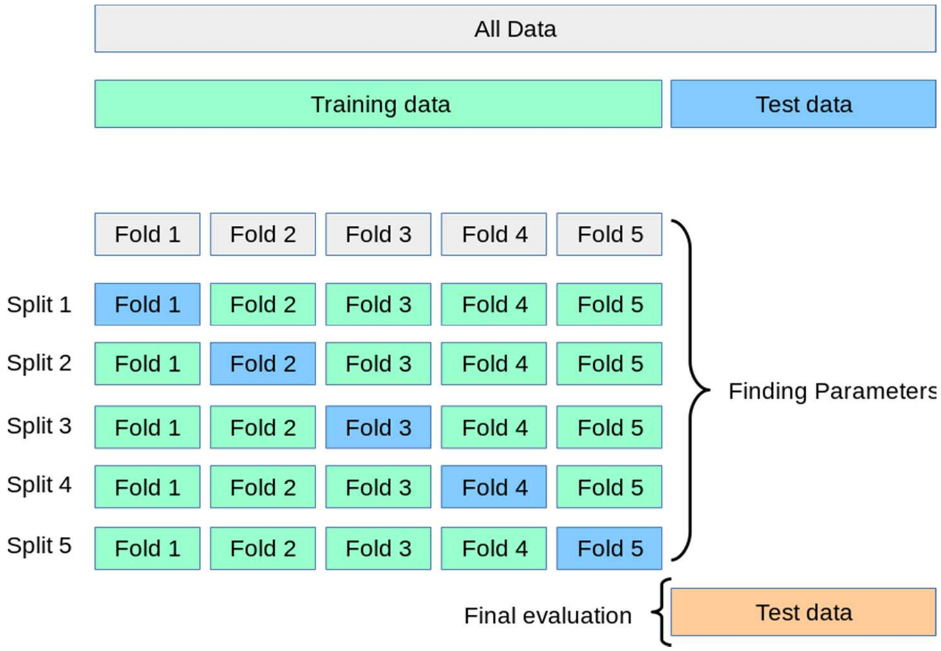


Figure 2.5: Illustration of data split and cross validation. The green block means the fold for the training set while the blue block means the fold for the test set.

Training strategy is one important step to find the best model as well as the optimal model parameters to train the dataset. Cross-validation (CV) is a computationally intensive technique, using all available examples as training and test examples. It mimics the use of training and test sets by repeatedly training the algorithm $K$ times with a fraction $\frac{1}{K}$ of training examples left out for testing purposes. This kind of hold-out estimate of the performance lacks computational efficiency due to the repeated training, but the latter is meant to lower the variance of the estimate [58].

Figure 2.3 illustrates the data split method for cross validation as one example of how five-fold cross-validation works. In practice, the dataset $D$ is first chunked into $K$ disjoint subsets or blocks of the same size: $m \triangleq \frac{n}{K}$ . $T_k$ stands for the $k$-th block and $D_k$ is the training set obtained by removing the elements in $T_k$ from $D$. The CV estimator is defined as the average of the errors on test block $T_k$ obtained when the training set is derived from $T_k$:

$$CV(D) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{\boldsymbol{Z}_i \epsilon T_K} L[(A(D_k), \boldsymbol{Z}_i) \qquad (2.3.17)$$

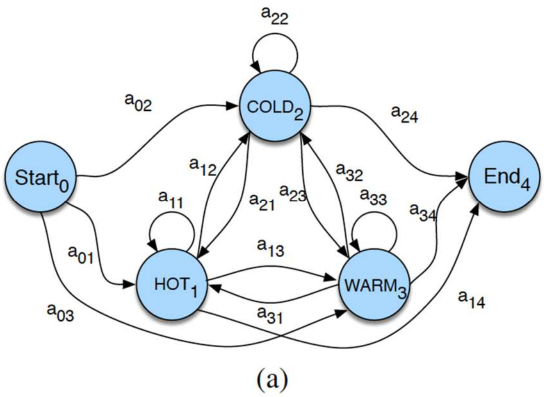
, which estimates the total error after training on the dataset $D$. After implementing on different models as wells as different sets of model parameters, the one with the minimum error would be selected.

## 2.4 Temporal machine learning and deep learning models

Classical or non-temporal machine learning methods are effective for making single decisions, but do not allow for adjustment to learn the sequential temporal information in the data. Temporal models are appropriate in the case of sequential or temporal observations where the value of the outcome may need to be updated over time. In particular, hidden Markov models (HMMs), recurrent neural network (RNN) and its variations have been widely applied to disease prediction [21,59], symptom progression [60] and classification of health questionnaires [61], demonstrating their outstanding performance over classical models in processing sequential or temporal data.

### 2.4.1 Hidden Markov models (HMM)

HMMs are sequence models. A sequence model or sequence classifier is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels [62]. An HMM is a probabilistic sequence model given a sequence of units (words, letters, morphemes, sentences, et al.) and it computes a probability distribution over the possible sequences of labels and chooses the best label sequence. Sequence labeling tasks come up throughout speech and language processing, a fact that is not too surprising if we consider that language consists of sequences at many representational levels.



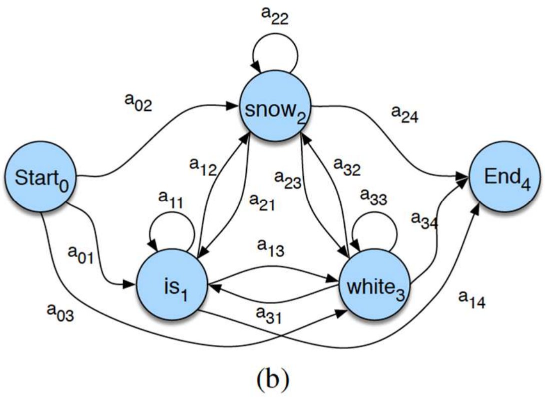


Figure 2.7: A Markov chain for weather (a) and one for words (b). A Markov chain is specified by the structure, the transition between states the start and end state. The probabilities on all arcs leaving a node must sum to 1.

HMM is one variation of the Markov chain, a special case of a weighted automaton in which weights are probabilities and the input sequence uniquely determines which states the automaton will go through. Because it cannot represent inherently ambiguous problems, a Markov chain is only useful for assigning probabilities to unambiguous sequences. Figure 2.4a shows a Markov chain for assigning probabilities to a sequence of weather events, for which the vocabulary consists of HOT, COLD, and WARM. Figure 2.4b shows another simple example of a Markov chain for assigning probabilities to a sequence of words $w_1 \dots w_n$. This Markov chain in fact represents a bigram language model. Given the two models in Figure 2.4, we can assign probabilities to any sequence from our vocabulary.

A Markov chain embodies an important assumption about these probabilities. In a first-order Markov chain, the probability of a particular state depends only on the previous state:

$$P(q_i|q_1 \dots q_{i-1}) = P(q_i|q_{i-1}) \tag{2.4.1}$$

, where $q_i$ is the state at $i$-th time point. In this way, the model is based on a strong assumption, but it efficiently reduces the computation for the input sequence.

When states are not directly observable, a Markov chain becomes an HMM. For instance, in the part-of-speech tags, we assign tags like Verb or Noun to words. However, we are only able to see the word sequence instead of part-of-speech tags. Thus, we call part-of-speech tags hidden because they are not observed. An HMM allows us to focus on both the observed events (like words we see in the input) and the hidden events (like part-of-speech tags) that we think of as causal factors in the probabilistic model. There are three algorithms embedded in HMMs: (1) Forward algorithm (likelihood): computes the likelihood for a given observation sequence; (2) Viterbi algorithm (decoding): finds the best state sequence for a given observation sequence; (3) Forward-backward algorithm (learning): learns the model parameter to compute transition and emission probability for a given observation sequence [62]. Forward algorithm will be explained in further details as it is most applicable to this work.

The Forward problem in HMMs computes the likelihood $P(O|\lambda)$ given an observation sequence $O$. An HMM model is defined as $\lambda = (A, B)$, where $A$ and $B$ stands for the transition and emission probability, respectively. Given this one-to-one mapping and the Markov assumptions expressed in Eq. (2.4.1), for a particular hidden state sequence $Q = q_0, q_1, \dots, q_n$; and an observation sequence $O = o_1, o_2, \dots, o_n$ , the likelihood of the observation sequence is

$$P(O|Q) = \prod_{i=1}^{n} P(o_i|q_i) \tag{2.4.2}$$

Since state sequences are hidden, we need to compute the probability of the observed sequence and sum over all possible state sequences, weighted by their probability. In general, the joint probability of a particular hidden state sequence $Q$, which generates a particular observed sequence $O$ is

$$P(O,Q) = P(O|Q) \times P(Q) = \prod_{i=1}^{n} P(o_i|q_i) \times \prod_{i=1}^{n} P(q_i|q_{i-1}) \qquad (2.4.3)$$

, where the first and second component are called the emission probability and transition probability, respectively. Then the total probability of observations can be computed simply by summing over all possible joint probabilities of each hidden state sequence:

$$P(O) = \sum_{Q} P(O,Q) = \sum_{Q} P(O|Q)P(Q) \qquad (2.4.4)$$

For an HMM with $N$ hidden states and an observation sequence of $T$, there are $N^T$ possible hidden sequences. Instead of using a brute force approach to compute every combination, generating time complexity of $O(N^T)$, we can use an efficient $O(N^2T)$ algorithm called the forward algorithm. The forward algorithm is one kind of dynamic programming algorithm, which uses a table to store intermediate values as it builds up the probability of the observation sequence [62]. The forward algorithm computes the observation probability by summing over probabilities of all possible hidden state paths that could generate the observation sequence, but it does so efficiently by implicitly folding each of these paths into a single forward trellis. Each path of the forward algorithm trellis $\alpha_t(j)$ represents the probability of being in state $j$ after seeing the first $t$ observations, given an HMM model $\lambda$. The value of each $\alpha_t(j)$ is computed by summing over probabilities of every path that could lead it to. Formally, each path expresses the following probability

$$\alpha_t(j) = \mathrm{P}(o_1, o_2, \dots, o_{t,} q_t = \mathrm{j}|\lambda) \qquad (2.4.5)$$

, where $q_t$ = j means "the $t$-th state in the sequence is state $j$". We compute this probability $\alpha_t(j)$ by summing over the extensions of all the paths that lead to the current path. For a given state $q_j$ at time t, the value $\alpha_t(j)$ is computed as

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i)\, a_{ij} b_j(o_t) \tag{2.4.6}$$

, where $\alpha_{t-1}(i)$ is the previous forward path probability from the previous time step. $a_{ij}$ is the transition probability from previous state $q_i$ to current state $q_j$ and $b_j(o_t)$ is the state observation likelihood or emission probability of the observation symbol $o_t$ given the current state j.

### 2.4.2 Long short-term memory (LSTM)

The Markov assumption ignores much information about the historical states, which cannot process long sequences. Thus, with development of deep learning and neural networks, RNN shows superior performance than HMM on processing sequential data. However, training RNNs suffers from the gradient vanishing and exploding problem due to the repeated multiplication of the recurrent weight matrix [63]. Several RNN variants such as the long short-term memory (LSTM) and the gated recurrent unit (GRU) have been proposed to address the vanishing gradient

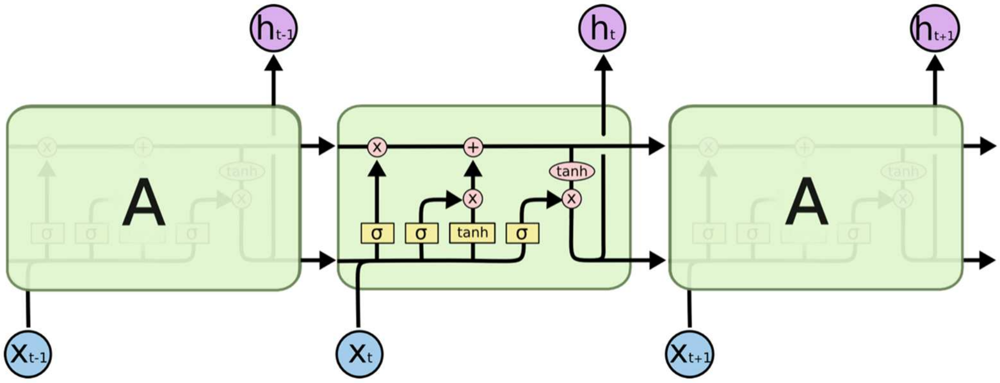


Figure 2.9: The repeating modules in an LSTM that contains four gate structures.

problems. LSTMs were originally introduced in [64], following a long line of research into RNNs for sequence learning. It can model varying-length sequential data, achieving state-of-the-art results for problems in natural language processing [65], image captioning [66] and genomic analysis [67]. LSTMs can capture long range dependencies and nonlinear dynamics. Some sequence models, such as Markov models, conditional random fields, and Kalman filters are ill-equipped to learn long-range dependencies. Other models require domain knowledge or feature engineering, offering less chance for serendipitous discovery. In contrast, neural networks learn representations effectively and can discover unforeseen structures.

Figure 2.5 shows the inner structure of LSTM and the repeating structure to process sequential inputs. LSTM is composed by four gate structures: (1) input gate; (2) forget gate; (3) cell state; (4) output gate, which are represented by Eqs. (2.4.7) to (2.4.11), respectively.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.4.7}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.4.8}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)\,, \qquad C_t = f_t * C_{t-1} + i_t \, * \tilde{C}_t \tag{2.4.9}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.4.10}$$

$$h_t = o_t * tanh(C_t) \tag{2.4.11}$$

, where $\tilde{C}_t$ is the cell candidate state and $\sigma$ is the sigmoid function. Notable earlier works [68,69] have realized backpropagation through time, where successfully trained RNNs were able to perform supervised machine learning tasks with sequential inputs and outputs. The design of modern LSTM memory cells has remained close to the original ones, with the commonly used addition of the forget gate.

### 2.4.3 Gated recurrent unit (GRU)

Another variation is using coupled forget and input gates. Instead of separately deciding what to forget and what new information to keep, it is more computationally efficient to make decisions together by only forgetting when there will be another input in its place. In the meantime, the model inputs new values to the state when forgetting something older. Therefore, Gated Recurrent Unit (GRU) was introduced as another variation of RNN [70]. It combines the forget and input gates into a single update gate $z_t$. It also merges the cell state and hidden state into the reset gate. The resulting model is simpler than the standard LSTM model and has become increasingly more popular. The formulation for GRU is as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.4.12}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_z) \tag{2.4.13}$$

$$\tilde{h}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_h) \tag{2.4.14}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{2.4.15}$$

## 2.5 Models for natural language processing

There has been fast development in NLP models recently. The typical models include topic modeling [71], word2vec embedding [72], transformer [73] and BERT [74], which have demonstrated outstanding performance in text generation, semantic feature extraction and entity recognition and language translation. Healthcare datasets such as activity tracker data or EHR are similar to sequential data structure of words and texts. Therefore, several studies have also applied NLP models on healthcare data for disease prediction [75] or representation learning of EHR [76].

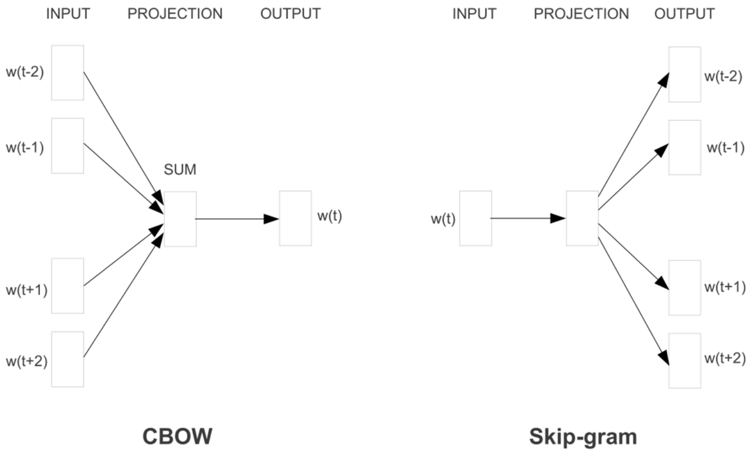


Figure 2.11: Model architectures for CBOW and skip-gram word2vec embedding.

### 2.5.1 Word2vec embedding

Before word2vec embedding was invented, words were treated as categorial variables and represented as one-hot vector with dimension of the vocabulary size. Each word has one column equaling one while the rest are zero. There are two main drawbacks in this method: (1) each vector is highly sparse, which is inefficient to compute; (2) it ignores the sematic meaning of words. Hence, the embedding technique resolves these two problems effectively. It is composed by two window based structures: continue-bag-of-words (CBOW) and skip-gram [77]. Figure 2.6 shows

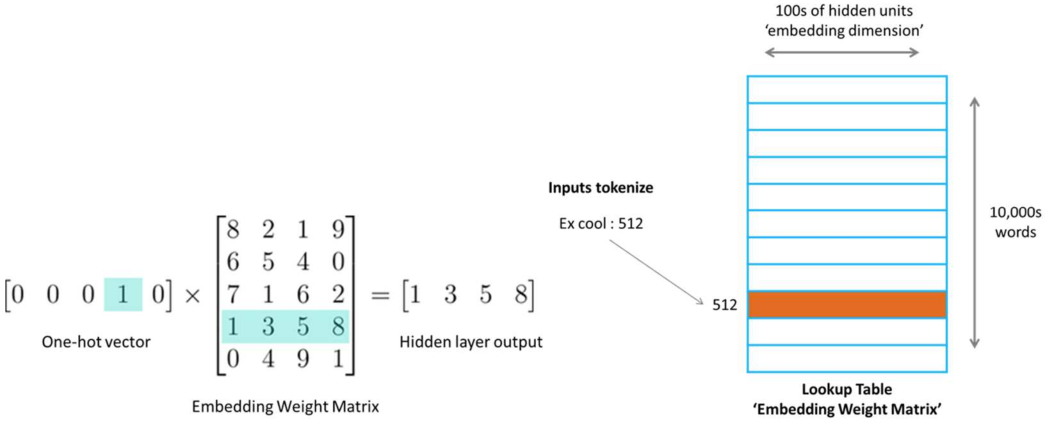


Figure 2.13: Illustration of word embedding method.

the structure for CBOW and skip-gram models. CBOW uses the context word to predict the center word whereas skip-gram uses the center word to predict context words. Figure 2.7 displays how the embedding works from one-hot encoding. The left of Figure 2.7 shows how the embedding weight matrix transforms the one-hot vector into a dense vector with the same dimension while the right figure extends the weight matrix as an embedding lookup table, where the one-hot encoding of each word with size of 10,000 is projected to a lower dimension as a dense vector with size of 100. Therefore, the embedding technique efficiently resolves the feature sparsity problem.

### 2.5.2 Transformer and BERT

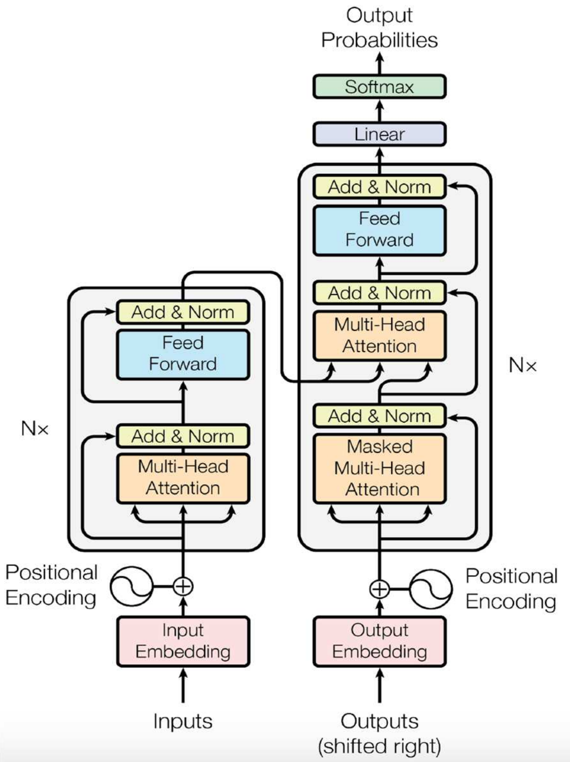


Figure 2.15: Architecture of the transformer. The left and right component are structure of encoder and decoder, respectively [73] .

Even though deep learning models have revealed the outstanding performance in classification and prediction tasks, model's interpretability is pretty low with the famous notation of "black box". Therefore, much effort has been put to improve models' interpretability and building explainable models [59,78], especially for deep neural networks. In addition, LSTM or GRU have two main drawbacks: (1) they take the input in a sequential order with one sample at a time, not allowing for parallel computing, so training every model takes much time; (2) they don't enable the method of pre-training on a large standard dataset and finetuning on customized tasks, which is successful and widely adopted in computer vision [66,79,80]. With the aid of the innovated encoder-decoder structure in neural machine translation [81], the transformer was then invented as a generalized language model [73]. Figure 2.8 describes the encoder-decoder structure, where the transformer is a multiple stacking of this structure.

The encoder takes all input words simultaneously, enabling parallel computing while the decoder takes every input sequentially for counting the next output. Positional embedding is adopted to annotate the position of each word in the sequence since the encoder ignores the position for parallel computing. Self-attention is one major technical breakthrough in the transformer architecture, which learns the inner relation between each word in the sequence. As shown in Figure 2.9, self-attention is computed from initialization of three matrices: Query; Key; Value with the following equation:

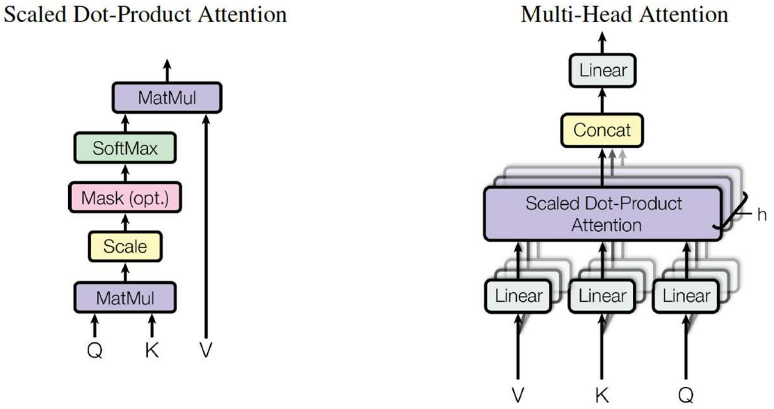


Figure 2.17: Structure of self-attention (left) and multi-head attention (right) [73].

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2.5.1)$$

, where $d_k$ is the dimension of the Key vector, serving as a scaling factor. In addition, transformer enables the flexibility of attention computation by adding more sets of Query, Key, Value vectors pairs with various dimension $d_k$, concatenated together to form the multi-head attention. Using multi-head attention allows the model to jointly tend to information from different representation subspaces at different positions [73]. Here is the mathematic formula:

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O \qquad (2.5.2)$$

$$where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Built up on that, the BERT model was implemented, which stands for Bidirectional Encoder Representations from Transformers [74]. As shown in Figure 2.10, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning one both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

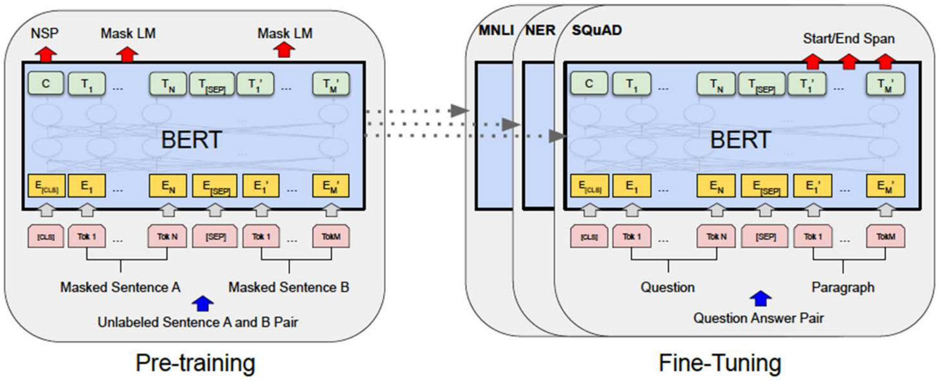


Figure 2.19: Pretraining (left) and finetuning (right) procedures for BERT. The model architecture is same other than the output layer [74].

The BERT model has demonstrated the power of bidirectional learning on word sequences rather than single or forward-only sequence learning. Furthermore, it first realized the pretraining and finetuning approach for NLP models, which started the enormous advancement in language or sequence modeling.

## 2.6 Metrics to evaluate machine learning models

### 2.6.1 Receiver operation characteristics area under the curve (ROCAUC)

A classifier operates at different thresholds or decision points, where the correct and incorrect classification are made after choosing one threshold. Therefore, it is often desirable to obtain one single metric to evaluate the classifier's performance [82]. The area under the curve for receiver operation characteristics (ROCAUC) is then derived and commonly used to evaluate the performance of a binary classifier [82]. The definition is followed by the computing the true positive rate (TPR) and false positive rate (FPR) from the confusion matrix computed at every operation point, iteratively. The four items in a confusion matrix are: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Then TPR and FPR are calculated using two equations below, respectively.

$$TPR = \frac{TP}{TP + FN} \tag{2.6.1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.6.2}$$

A curve is plotted based on the list of these two values and ROCAUC is computed using the curriculum sum of the curve which is also the area under the curve. The range of ROCAUC is between 0 and 1, inclusively, where 1 means a perfect classifier and 0.5 indicates a classifier behaves as a random classification. As all models developed in this dissertation are making binary classification, ROCAUC is chosen as the metric to evaluate and compare their performance.

### 2.6.2 Precision recall area under the curve (PRAUC)

ROCAUC is a common metric to evaluate the performance of binary classifiers as introduced in the previous section, but it fails to present the best result when the dataset is imbalanced [83,84]. Instead, precision-recall area under the curve (PRAUC) is developed to resolve this situation. Precision and recall are defined in the following equations:

$$Presicion = \frac{TP}{TP + FP} \tag{2.6.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.6.4}$$

As a similar approach to compute the AUC for ROC, PRAUC is computed by a curriculum sum of the precision recall curve at each decision threshold. According equation (2.6.1) and (2.6.2), each of them only consider the performance in classifying positive and negative classes, separately. On the contrary, precision calculated in equation (2.6.3) uses both positive and negative classes, which reveals the information of class distribution. Therefore, PRAUC is a better metric than ROCAUC when evaluating imbalanced datasets, which is suitable in our situation.

# CHAPTER 3

# Classifying Self-Reported Health Status by Machine Learning Algorithms and Activity Tracker

## 3.1 Overview

There has been significant effort in developing monitoring devices and protocols to diagnose patients remotely. However, device fatigue has been shown to be a barrier to adherence [85–87]. Commercially available devices, such as passive accelerometry, have been shown to overcome this barrier by reducing the burden of human intervention [88], and the accuracy of activity tracker data has been demonstrated to be sufficient for documenting health indicators in real-time [7,9,89]. With wireless connections to portable electronics, such as smartphones or tablets, monitoring by activity tracker is an easy-to-use, accessible means of providing personalized information to people's health and daily activities [90]. This approach creates a feedback loop that is capable of positively impacting health interventions with the goal of lifestyle change [10,91]. However, analysis of this data has largely been limited to simple correlations, and the ability to use this information to classify patient health status has not been explored [11,92].

In this chapter, I explored the use of machine learning methods to classify PRO scores over time [93]. The goal of this study was to investigate the feasibility of using machine learning models to classify PRO scores based on data collected using one type of activity tracker, the Fitbit Charge 2. In this study, I tested this goal within a population of patients with stable ischemic heart disease (SIHD). In the remainder of this chapter, Section 3.2 details the data and preprocessing steps used in this study. Section 3.3 describes the structure of two machine learning models: (1) a model that treats weeks independently; (2) an HMM that takes temporal information into account. Section 3.4 summaries the results of two constructed classification algorithms by evaluating through each PRO measure. Section 3.5 discusses the analysis of the results and suggests future directions for implementing such a classifier in a patient surveillance application. The content of this chapter have partly been published in [94].

## 3.2 Data Cohort and Data Preprocessing

A set of 200 patients with SIHD were recruited for a feasibility study conducted by Cedars-Sinai Medical Center from 2017 to 2018 to predict surrogate markers of major adverse cardiac events (MACE), including myocardial infarction, arrhythmia, and hospitalization due to heart failure, using biometrics, wearable sensors, patient-reported surveys, and other biochemical markers. The population size of this study is similar to several previous one that used activity trackers for patient monitoring [95,96]. The desired monitoring period was 12 weeks for each subject, during which time subjects wore personal activity trackers to record their physiological indices, including steps, heart rate, calories burned, and distance traveled. At the end of each week, they were asked to fill out eight PROMIS short forms as a self-report assessment of their health status [88].

### 3.2.1 Activity data

The Fitbit Charge 2 (Fitbit Inc., San Francisco, CA, USA) is a popular commercially available activity tracker that can record a person's daily activities and health indices like heart rate, steps, and sleep (Table 3.1). Previous works have validated the accuracy of heart rate monitoring specifically in the Fitbit Charge 2 [97]. The Fitbit hardware and its computational algorithms for calculating step counts and physical activity have been validated using other Fitbit devices [26,98].

Table 3.1: Summary of 17 types of feature collected from Fitbit per day.

| Type (units) | Mean ± Std |
|---|---|
| Steps (#) | 6138 ± 4031 |
| Total Distance (kilometers) | 4.18 ± 3.00 |
| Tracker Distance* (kilometers) | 4.18 ± 3.00 |
| Logged Activity Distance* (kilometers) | 0.02 ± 0.56 |
| Very Active Distance (kilometers) | 0.71 ± 1.49 |
| Moderate Active Distance (kilometers) | 0.36 ± 0.60 |
| Light Active Distance (kilometers) | 2.69 ± 1.90 |
| Sedentary Active Distance* (kilometers) | 0.01 ± 0.08 |
| Very Active Minutes | 12.21 ± 22.29 |
| Fairly Active Minutes | 12.78 ± 21.89 |
| Light Active Minutes | 176.81 ± 99.73 |
| Sedentary Minutes | 823.24 ± 323.90 |
| Calories | 2032 ± 610 |
| Floor (#) | 5.1 ± 11.8 |
| Calories BMR (basal metabolic rate) | 1428 ± 254 |
| Marginal Calories | 372 ± 317 |
| Resting Heart Rate (BPM) | 61.81 ± 7.45 |

*means that feature was eliminated for model input because it was highly sparse or redundant. Std represents the standard deviation.

The Fitbit Charge 2 estimates activity using metabolic equivalents (METs), which are calculated based on heart rate and distance traveled [99]. Heart rate during activity is also provided, however it has been shown to be inaccurate during activities [100]. Data quality was assured by verifying that there were no extreme outliers based on subject-specific inter-quartile range [101]. We aggregated the data for each day to compensate for noise and redundancy. After data preprocessing, tracker distance was eliminated because it was identical to total distance, and logged activity distance and sedentary active distance were also deleted because of high sparsity. As a result, there were 14 features per day for each patient in the model.

### 3.2.2 Patient reported outcome measures

Patient-Reported Outcomes Measurement Information Systems (PROMIS®) questionnaires are a library of instruments developed and validated to measure many domains of physical and mental health [102]. This analysis uses data from eight PROMIS instruments: Global Physical Health and  Global Mental Health, which are two composite scores from the Global-10 short form [103]; Fatigue-Short Form 4a; Physical Function-Short Form 10a; Emotional Distress-Anxiety-Short Form 6a; Depression-Short Form 4a; Social Isolation-Short Form 4a; and Sleep Disturbance-Short Form 4a. Each questionnaire either asks about current health or has a recall period of the previous seven days, so they are appropriate for weekly administration. The T metric method was used to standardize scores for each type to a mean of 50 and a standard deviation of 10, with a range between 0 and 100 [102,104]. Symptom (i.e., Fatigue, Anxiety, Depression, Social Isolation,

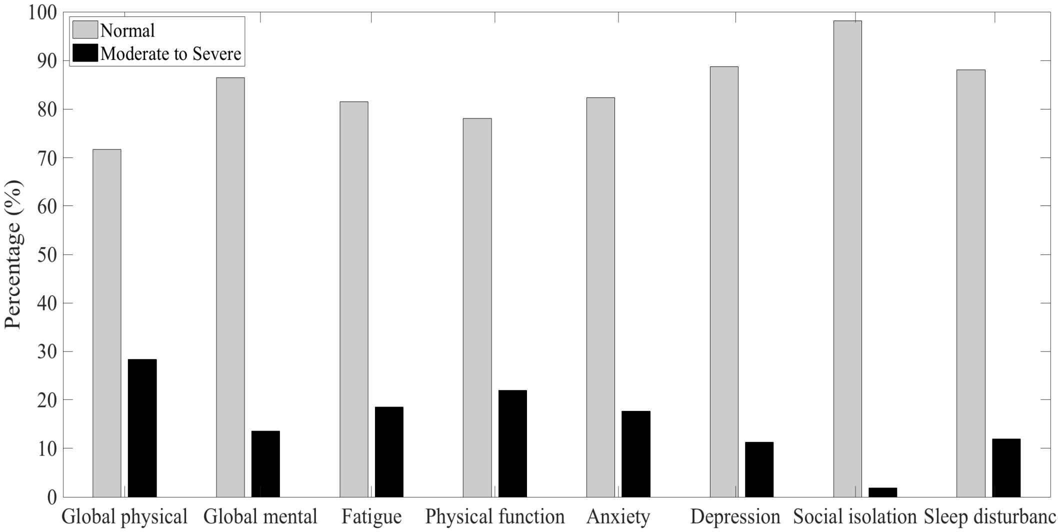


Figure 3.1: Distribution of normal and abnormal (moderate to severe) class for each PRO measure.

and Sleep Disturbance) scores of 60 or higher are one standard deviation above the average, which is defined as moderate to severe symptom severity. For function (i.e., Global Physical Health, Global Mental Health, and Physical Function), scores less than 40 are classified as moderate to severe, meaning less functional ability than normal. For this study, PRO scores were predicted in two ways: regression was used to predict PRO scores from patient activity tracker data, and classification was used to determine whether subjects' PRO scores were above the threshold for at least moderate severity. The distributions of PRO scores are shown in Figure 3.1. Because of a lack of moderate or severe cases for social isolation (<2%), this variable was eliminated for analysis in the model.

## 3.3 Methods

Missing data is a common concern when dealing with activity tracker data, which usually results from subjects either forgetting to wear their devices or removing them for charging. Patients were asked to fill out eight PROMIS questionnaires at the end of each week for a 12-week monitoring period. In total, 19.1 percent of weeks had missing PRO data and 16.6 percent of weeks had missing values from the activity tracker in four or more days. If data was available for at least four days in a week, missing values were permuted by using the average value of the rest of the week for steps or resting heart rate. Weeks with missing survey scores, as well as those without step and resting heart rate data for more than three days, were removed from the analysis.

A correlation analysis between subjects' missing Fitbit data and their average Global Physical Health and Global Mental Health scores showed a slight negative relationship (-0.11 and -0.09, respectively) that was not statistically significant (p=0.13 and p=0.23, respectively). The correlation coefficient between number of missing PROs and the average global health scores were

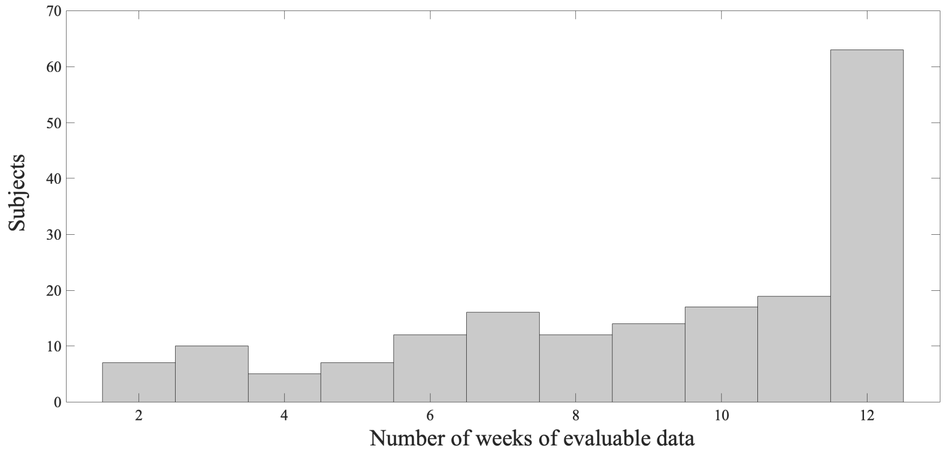


Figure 3.2: Histogram of number of weeks of evaluable data for the 182 subjects used in the dataset.

-0.17 (p=0.018) to -0.14 (p=0.048), respectively, indicating that the missing PROs were not significantly related to patient health. Another correlation analysis was performed between subject's age and number of missing values with $R^2 < 0.001$, which demonstrated no trend of more missing values for elder subjects. Finally, subjects with only one week of data were eliminated in order to ensure the continuity of transition of states from week to week when building the HMM model. After adopting this data preprocessing approach and using the classification criteria above, a total number of 182 subjects with a total of 1,640 weeks were collected, where the number of weeks of evaluable data for each patient ranged from two to 12 weeks as shown in Figure 3.2.

### 3.3.1 Independent per week model by machine learning algorithms

Since survey scores were generated per week, a naive approach is to treat each week independently. The left plot in Figure 3.3 illustrates the idea of the independent model as an example for one subject with 12 weeks of evaluable data. Features for each of the seven days were appended into a single feature vector, which was then used as the input for binary classification of each PRO score. Ensemble methods like AdaBoost, GBRT (gradient boosting regression tree) and Random Forest (RF) are relatively robust over unbalanced dataset and are capable of generating better classification accuracy than other types of machine learning algorithms [105]. Each of these methods was applied to the dataset using ten-fold cross-validation across subjects in conjunction with the grid search to find optimal parameters for every model. Paired t-test was applied to validate the statistical significance for each comparison of the result with different $\alpha$ values: 0.05, 0.01 as for different levels of significance. A sensitivity analysis was completed to investigate the model performance against missing values in the feature vector by randomly withholding values from one to six days within a week.

### 3.3.3 Hidden Markov model (HMM) with forward algorithm

In order to track changes in PRO responses over time, a model was built to incorporate temporal correlations of PRO scores across weeks. As shown in the right part of Figure 3.3, an HMM was used and formalized such that the state at each time point corresponded to the PRO score for that week, with features collected for that week treated as observations. The transition matrix was derived by counting the state transitions from week to week. The original number of states for each PRO was found by number of unique responses, ranging from 15 to 36. In order to make the transition matrix less sparse, we defined 10 states for all types of PROs based on the

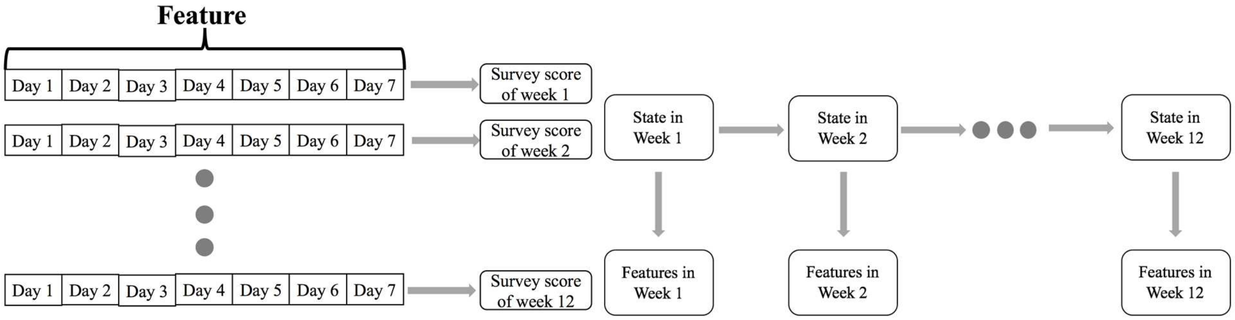


Figure 3.3: Illustration of independent week model (left) and Hidden Markov Model (right). For HMM, feature in each week was observed while the state of health status transits from week to week.

score distribution of each. The Forward algorithm computed the probability across states at time $t$, with the maximum probability representing the classified state,

$$S(y_t|y_{t-1},\dots,y_1,x_t,\dots,x_1) = P(x_t|y_t) * \sum \mathrm{P}(\mathrm{y_t}|\mathrm{y_{t-1}}) * \mathrm{S}(\mathrm{y_{t-1}}|\mathrm{y_{t-2}},\dots,\mathrm{x_{t-1}},\dots) \quad (3.3.1)$$

where the weekly PRO score was treated as the state $y_t$, with observation of features $x_t$. The emission probability, $P(x_t|y_t)$, computing the probability of the observed feature vector $x_t$ given state $y_t$, was derived from the random forest classifier and $P(y_t)$:

$$P(x_t|y_t) \propto \frac{P(y_t|x_t)}{P(y_t)} \quad (3.3.2)$$

At the first-time step, the transition probability distribution is undefined, so the state probability was:

$$S(y_1|x_1) \propto P(x_1|y_1)P(y_1) \quad (3.3.3)$$

For analysis, states were binarized according to the criteria defined above. Because dichotomizing PRO score values loses some information and precision, a regression analysis was conducted between the median value of HMM stages and actual scores for the HMM. This method of predicting PRO scores was compared against multinomial logistic regression to evaluate the accuracy of predicting PRO scores over time.

## 3.4 Evaluation and Results

Table 3.2: Mean and standard deviation ROCAUC of difference Algorithms.

| Type | AdaBoost | GBRT | Random Forest |
|---|---|---|---|
| Global physical health | 0.72 (0.03) | 0.69 (0.04) | **0.73 (0.01)*** |
| Global mental health | 0.53 (0.03) | 0.51 (0.03) | **0.55 (0.03) *** |
| Fatigue | 0.59 (0.04) | 0.60 (0.04) | **0.61 (0.03)** |
| Physical function | 0.74 (0.03) | 0.75 (0.03) | **0.75 (0.01)** |
| Anxiety | 0.48 (0.03) | 0.50 (0.03) | **0.54 (0.02)** † |
| Depression | 0.47 (0.04) | 0.50 (0.03) | **0.53 (0.02)** † |
| Sleep Disturbance | 0.55 (0.06) | 0.59 (0.05) | **0.61 (0.03)** |

* Significant improvement over GBRT.
† Significant improvement over both GBRT and AdaBoost. Bold values are the highest for a given PRO.

Table 3.2 shows the mean AUC for binary classification of PRO scores for the seven PROMIS measures using GBRT, AdaBoost and RF. The highest mean AUC was 0.75 using RF for classifying Physical Function, while the lowest was 0.47 using AdaBoost for Depression. The results indicated that RF significantly outperformed other models in classifying Anxiety and Depression ($p<0.05$), and it was also significantly better than GBRT for Global Physical Health and Global Mental Health ($p=0.01$ and $p=0.01$, respectively). The RF model was selected for the remaining analyses because its performance was equivalent to or better than other methods for classifying all PRO scores. Additionally, it was notable that the AUC related to self-reported physical health PROs such as Global Physical Health, Fatigue, and Physical Function were higher than those related to mental health such as Global Mental Health, Anxiety, and Depression.

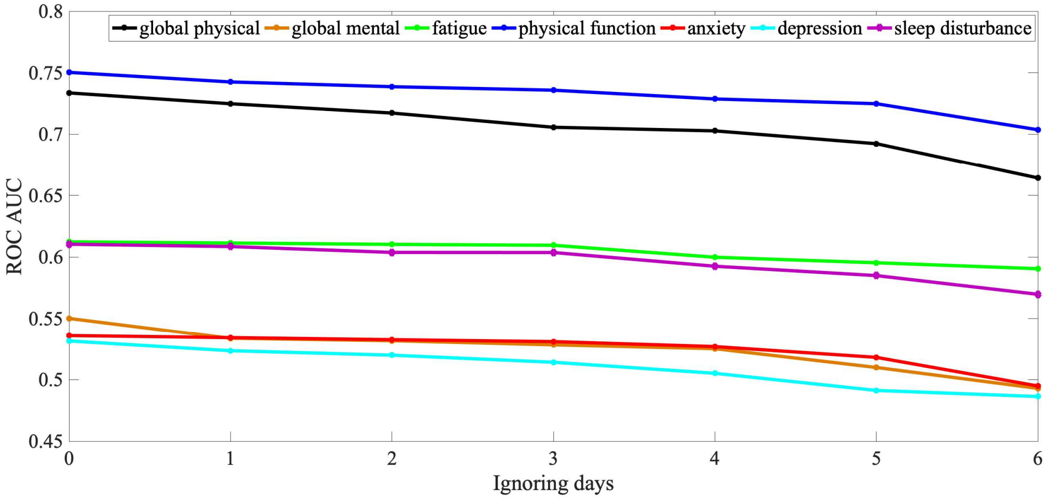


Figure 3.4: Plot of ROCAUC for each type of PRO after randomly withholding feature values from one day to six days within a week.

Figure 3.4 illustrates the results of sensitivity analysis on missing feature data on RF classification by randomly censoring data from one day to six days per a week. The results show that ROCAUC decreased monotonically as days were removed. For Global Physical Health, the value at missing four days dropped significantly compared to no missing data ($p=0.03$), while the difference at missing three days was not significant ($p=0.11$). This was why that cutoff was chosen for inclusion in the analysis.

Table 3.3 displays the comparison of means and standard deviations of the AUC for each PRO measure using the independent model and the HMM. AUCs derived using the HMM were significantly higher than those from the independent model in all domains other than Fatigue and Sleep Disturbance. Depression achieved the highest increase from 0.57 to 0.61. We also compared the $R^2$ value of the regression analysis between the HMM and a multinomial logistic regression. The values were 0.079 and 0.1526 from HMM in Global Physical Health and Physical Function. They were significantly better than the values achieved by the multinomial logit model (0.0016 and 0.0026, respectively; $p<0.001$ for both). This result suggested that HMM could also track the

Table 3.3: Mean and standard deviation of AUC values between the independent week model and the hidden Markov model.

| Type | Independent model | HMM |
|---|---|---|
| Global physical health | 0.73 (0.02) | **0.76 (0.02)*** |
| Global mental health | 0.58 (0.01) | **0.61 (0.02)*** |
| Fatigue | 0.64 (0.03) | **0.65 (0.03)** |
| Physical function | 0.76 (0.01) | **0.79 (0.02)*** |
| Anxiety | 0.57 (0.02) | **0.61 (0.04)*** |
| Depression | 0.56 (0.02) | **0.59 (0.02)*** |
| Sleep Disturbance | 0.64 (0.03) | **0.66 (0.05)** |

* Significant improvement over the independent model. Bold values are the highest AUC for a given PRO.

minor change of PRO scores with higher precision over time than baseline models like multinomial logistic regression.

## 3.5 Discussion

In this work, I proposed a temporal machine learning model can be used to classify self-reported health status in patients with SIHD using physiological indices measured by activity trackers. By constructing an HMM with an RF classifier, the resulting model can achieve an AUC of 0.79 for classifying Physical Function. The result indicates that data generated from activity trackers may be used in a machine learning framework to classify validated self-reported health status. These techniques could play a future role in larger frameworks for remotely monitoring a patient's health status in a clinically meaningful manner.

In general, the AUCs related to classifying physical health were relatively higher than mental health PROs, such as Global Mental Health, anxiety and depression. This result makes intuitive sense, as the collected data, such as steps, total distance, and calorie expenditure, are more directly related to physical health than mental health. In particular, the highest AUC was 0.79 from classification of Physical Function, which demonstrated the correlation between data collected from Fitbit and patients' physical health. However, AUC values also indicated that PROs cannot be completely determined by activity tracker data alone, suggesting that PROs, particularly those pertaining to mental health such as depression, contain additional information that was not captured in the tracking devices. While the current study demonstrates the use of activity trackers to capture information about patient's health status, in some cases PROs could be a preferable method. Internet access enables PRO data collection to be done outside of clinic through web or mobile apps, which provides convenience and reduces time commitment for patients.

In this study, I found inconsistency in sleep data and sleeping stages for subjects. It was likely that Fitbit was taken off for charging during nights. Therefore, future studies should notice user not always charge it during nights to collect sleep data. Moreover, the data elements that have been validated are generally only tested in specific devices, rather than across all activity trackers, so it is not clear how these validation results translate to other devices. Future studies should be conducted to validate these features. As indicated by the correlation between subject's average PRO scores and the number of missing PRO values, patients with moderate to severe health status were less likely to complete PRO questionnaires routinely, which may have introduced bias for data collection in this study. Future studies could try to provide incentives for continued participation, which may mitigate study attrition. Eight PROMIS instruments were used in this study, and some redundancy existed between the specific short forms such as fatigue or anxiety to the general Global-10 short form. Our current approach treated each score independently without considering this overlap. A possible future study could predict PRO scores simultaneously in a joint model such as Bayesian network, which considers the correlations between PRO scores.

In this dataset of patients with SIHD based on adjudicated clinical data, HMMs achieved significantly higher classification accuracy than treating weeks independently because they took advantage of correlations in subjects' survey scores from week to week. I followed an data-driven approach that the model states were determined based on the distribution of PRO scores in the clinical study [106]. Score bins for the states were defined to limit the sparsity during training and make the number or states consistent across all PROMIS PROs tested. However, this may not be the optimal way to define the number of states for clinical representation of health status. Future studies could conduct some analysis to find out the optimum number of states, which may further increase the classification accuracy.

While activity trackers are able to produce patient information within seconds or minutes, the sampling periods for PROs like PROMIS [107] are on the order of weeks, requiring down-sampling of the Fitbit data for comparison. Given that the PROs measured in this study are unlikely to vary significantly from day to day, this temporal resolution is appropriate for the application of PRO prediction. However, predicting more acute events might require more temporal resolution, which could be addressed by using the activity tracker data at a finer time scale. Long term follow-up with patients including recordings of clinical events such as rehospitalizations could also allow us to evaluate the effect of mHealth monitoring on clinical outcome, an important step in determining the efficacy of such intervention. More details about the future work of improving the precision for this classification system will be presented in Chapter 8.

# CHAPTER 4

# Predicting Depression from EHR using Machine Learning Algorithms

## 4.1 Overview

With the development of machine learning algorithms and wide accessibility of electronic health record (EHR) data, an increasing amount of effort has been dedicated to improving prediction of depression diagnosis. The EHR is composed of myriad data sources: diagnosis codes, procedure codes, medications, patient demographics, and clinical notes. Previous works have applied natural language processing (NLP) models and latent semantic analysis (LSA) to structure clinical notes into predictive features. When representing a patient's EHR data as a concatenated feature vector of these datasets, issues of heterogeneity and sparsity can make it challenging to model trends across patients. For instance, several works utilized diagnosis codes with limited demographic information (like age and gender) as feature vectors fitted into predictive machine learning models [108–110]. In addition, Zhang et al. [111] and Bian et al. [112] added procedure codes in their analysis. However, clinical notes, which characterize disease progression in a temporal manner, were not included. Therefore,  Usama et al. [113] processed clinical notes along with demographics while LePendu et al. [114] used unsupervised learning methods to process clinical texts and diagnosis codes to investigate drug-drug interactions. Huang et al. identified medication drug terms from clinical notes combined with diagnosis codes and demographics for predicting future diagnoses of depression [115]. Miotto et al. applied topic modeling to extract semantic features from clinical notes along with other EHR data sources, but the ultimate model did not perform temporal prediction of depression [116].

This chapter investigates the performance of machine learning models on multimodal EHR data for depression prediction. The developed models are able to aggregate the aforementioned five types of EHR data sources to predict future diagnosis of depression. The remainder of this chapter is organized as follows. Section 4.2 describes the data and preprocessing steps used in this study. Section 4.3 the details methods to process various EHR data modalities by machine learning algorithms as well as the experimental setup of depression prediction. Section 4.4 summaries the results while Section 4.5 discusses several observations, limitations and future directions.

## 4.2 Data Cohort and Data Preprocessing

Patients with diagnoses of myocardial infarction (MI), breast cancer and liver cirrhosis were selected to capture a spectrum of clinical complexity. Generally, MI has the least temporal dynamics, with acute onset time, resolution, and management. Breast cancer is increasingly complicated in terms of diagnoses and treatment options. Finally, a patient with liver cirrhosis may have many sequelae, generating a complex EHR representation. Patients for this project were

Table 4.1: Statistics of patient cohort.

| | Depressed patients | Non-depressed patients |
|---|---|---|
| Number | 2,545 | 3,575 |
| Male | 705 (27.70%) | 1,094 (30.60%) |
| Female | 1,840 (72.30%) | 2,481 (69.40%) |
| Mean age (std) | 70.04 (16.10) | 69.82 (15.51) |
| Mean Length of record | 6.81 (2.80) | 5.93 (2.95) |

identified from our EHR in accordance with an IRB (#14-000204) approved protocol. Each patient visit had EHR data types consisting of diagnosis codes in ICD-9 (International Classification of Disease, ninth revision) format, procedure codes in CPT (Current Procedural Terminology) format, medication lists, demographic information, and clinical notes. Any patient record coded with ICD-9 values for MI, breast cancer, or liver cirrhosis from 2006-2013 was included. In this dataset, demographics were limited to the patient's gender and age at the time of each visit. Relevant cohort statistics are shown Table 4.1.

ICD-9 codes, CPT codes, medication lists, and demographic information can all be considered as categorical variables. Therefore, an intuitive approach is to encode these features in a multi-hot binary vector, where each column corresponds to a specific code or data element. ICD-9 codes are up to five digits long with three digits before a decimal point and two digits after, resulting in more than 12,000 unique codes in our data set. In order to reduce the dimensionality of the feature vector, ICD-9 codes were grouped by the three numbers before the decimal point. This approach has been used in previous work [117].

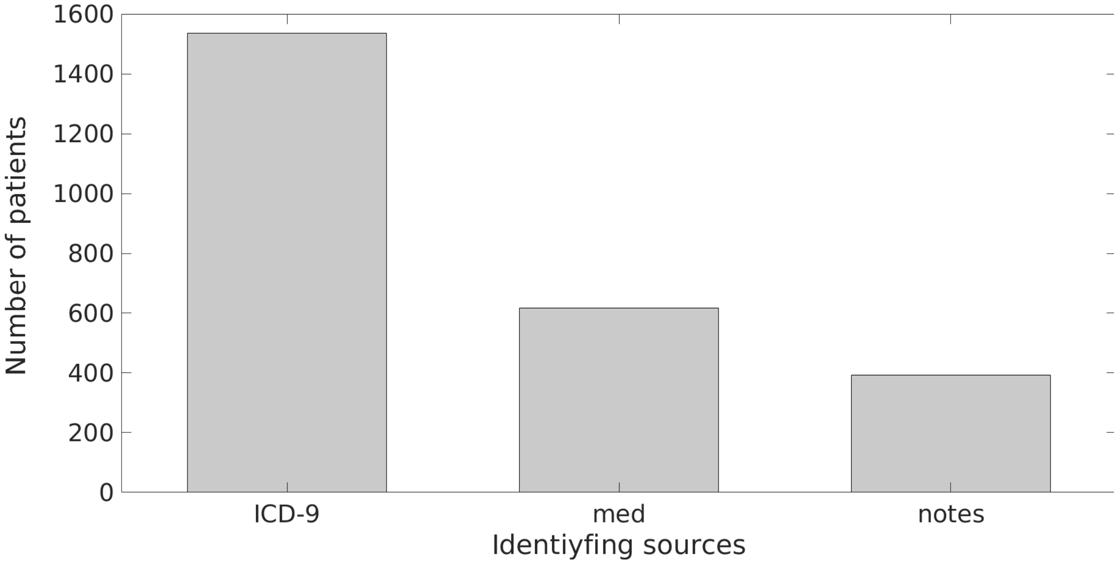


Figure 4.1: Distribution of depressed patients by three identifying methods.

### 4.2.1 Identification of patients with depression

Similar to the method used in [115], depression onset was identified by three methods: 1) a depression related ICD-9 code, 2) an antidepressant drug in a medication list, or 3) an mention of an antidepressant drug in a clinical note. The World Health Organization (WHO) drug index was used to generate a search list of drugs that could be used to treat depression.[1] For each patient, the earliest time stamp of an occurrence of any of these events was defined as the time of diagnosis with depression. The distribution of identified depressed patients from the cohort is shown in Figure 4.1. The total number of patients with depression was 2,545, with 1,536 identified by depression-related ICD-9 codes in their EHR, 617 identified by antidepressants in their medication lists, and 392 identified by antidepressant mentions in clinical notes.

## 4.3 Methods

### 4.3.1 Topic modeling of clinical reports

Statistical topic modeling is an approach that seeks to identify and quantify semantic themes from unstructured free text [118,119]. Previous works have applied topic modeling to understand and represent clinical notes [120–123]. Latent Dirichlet allocation (LDA) [118,119] is a common topic modeling technique that uses unsupervised learning approaches to learn underlying topics (semantic themes) based on the contextual co-occurrence of words in a collection. A topic is represented as a multinomial distribution over the unique words in a corpus, and a document is represented as a multinomial distribution over all topics. We used LDA to model clinical notes

[1] https://www.whocc.no/atc_ddd_index/?code=N06A

with 100 topics, thus generating a 100-feature vector representation of each note in the semantic topic space.

### 4.3.2 Predicting diagnosis of depression by machine learning algorithms

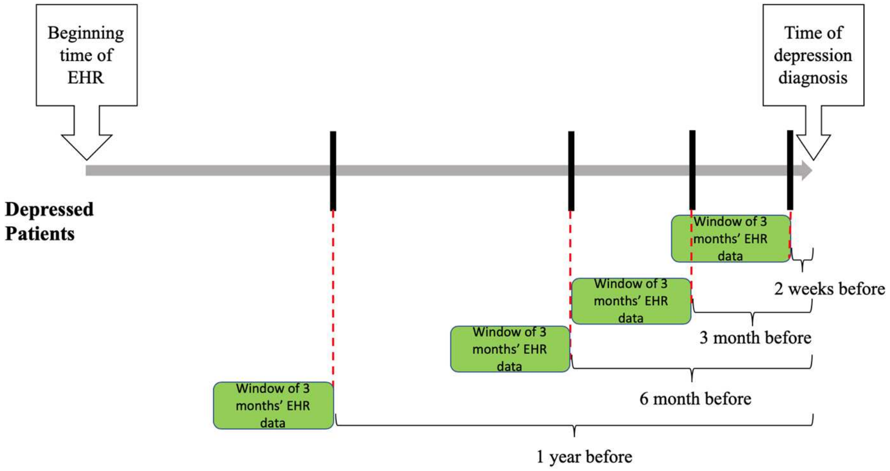

Beginning time of EHR
Time of depression diagnosis
Depressed Patients
Window of 3 months' EHR data
2 weeks before
Window of 3 months' EHR data
3 month before
Window of 3 months' EHR data
6 month before
Window of 3 months' EHR data
1 year before

Figure 4.2: Experiment design of 3 month time window of EHR data with four prediction time prior to diagnosis for patient with depression: 2 weeks, three months, six months, one year.

Predicting depression prospectively can be considered a continuous task that may be based on various temporal windows of the EHR. We therefore explored a spectrum of time windows and predictions horizons. Our first experiment was similar to previous works [23,75,124], in which the entire length of EHR data was utilized. As shown in Figure 4.2, we further defined four additional prediction windows: two weeks, three months, six months and one year prior to time of diagnosis, whereas Huang et al. utilized three windows, immediately preceding the diagnosis, six months prior to the diagnosis, and one year prior to the diagnosis [115]. Patients who had at least one of ICD-9, CPT, medication and topic feature in all four time windows were included in experiments. This inclusion criteria was enforced to ensure that the same set of patients was analyzed in each experiment. Different feature sets from the three EHR data modalities were analyzed, including all features, all features except topics, and all features except topics and CPT codes, which is the same set of the features used by Huang et al. [115]. I also tested each model after shrinking each patient's length of record to three months prior to the prediction horizon to evaluate model's performance on a limited set of EHR data. For patients labeled as not depressed, the last recorded time point in their EHR was substituted for the diagnosis time. Binary logistic regression (LR), support vector machine (SVM), and random forest (RF) classifiers were trained using ten-fold cross validation. Similar to [61], one advantage of using RF is its ability to output the importance factor of each feature, which facilitates the interpretation of features in the prediction task.

## 4.4 Evaluation and Results

Performance was measured using the receiver operator characteristic area under the curve (ROCAUC) and the precision recall aread under the curve (PRAUC) for each model. Figure 4.3 shows the bar plot of the mean ROCAUC and PRAUC for the entire EHR length with four prediction windows using all EHR data modalities. Based on the ROCAUC value, RF outperformed LR for prediction windows of two weeks and three months, with p-values of 5e-3

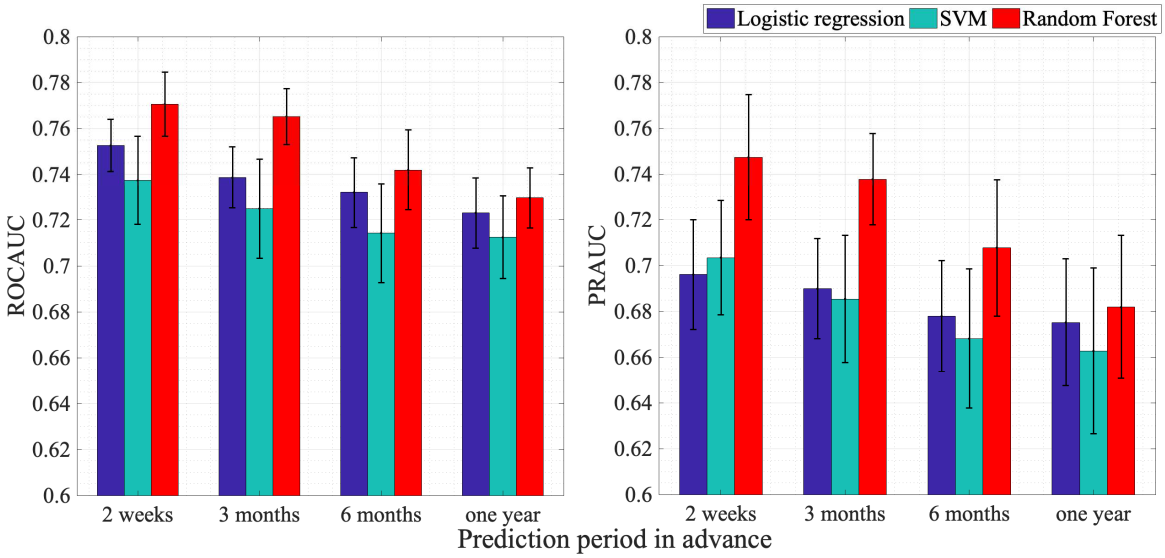


Figure 4.3: Prediction result of ROCAUC and PRAUC using all features for four predicting time periods in advance in left and right, respectively. The prediction accuracy was compared between logistic regression (blue), SVM (green) and random forest (red).

and 2e-2, respectively, compared to p-values of 0.19 and 0.32 for prediction windows of six months and one year, respectively. Additionally, RF outperformed SVM for all four time windows (p=3e-4, 2e-3, 2e-3, 4e-2), whereas the performance between SVM and LR was more variable. In total, for a prediction period of two weeks in advance, RF achieved its highest mean ROCAUC and PRAUC of 0.77±0.01 and 0.75±0.03, respectively. Whereas, LR reached 0.75±0.01 and 0.70±0.02 and SVM achieved 0.74±0.02 and 0.70±0.03. Additionally, there is a trend of decreasing ROCAUC and PRAUC as the prediction period progresses. Since RF outperformed LR and SVM, it was chosen for further experiments.

Figure 4.4 illustrates the results from RF using different combinations of EHR data modalities with the same prediction windows as described previously. The ROCAUC and PRAUC decreased by 0.01 and 0.02 on average after excluding topic features. After further excluding CPT codes, the ROCAUC decreased to 0.71±0.02 and 0.70±0.02 for the six-month and one-year prior to diagnosis horizons, respectively. Including CPT codes and topic features significantly improved the model's

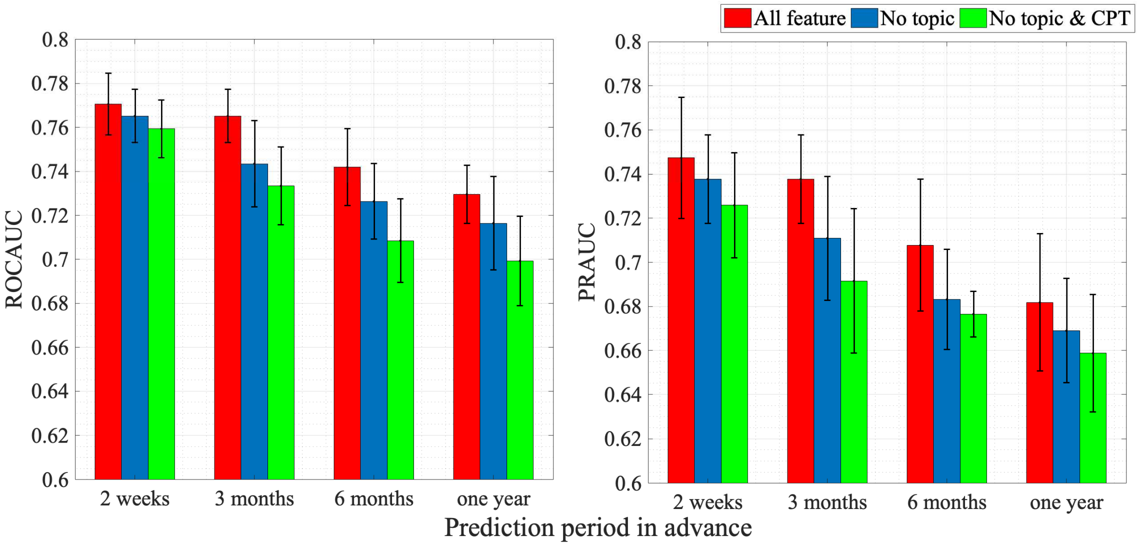


Figure 4.4: Plot of ROCAUC (left) and PRAUC (right) for the RF model with three EHR data modalities. The red bar used all features, which is the same as that in Figure 3, while the blue bar excluded topics, and the green bar excluded both topics and CPT.

prediction performance to 0.74±0.02 and 0.73±0.01, with p=7e-4 and p=9e-4, thus demonstrating the importance of these two data sources.

A comparison study was performed to test model's performance with limited views of EHR data (Figure 4.5). In general, ROCAUC decreased by 0.03 on average after narrowing the window size to three months, compared to using the entire EHR. Performance did not decline significantly at the two-week prior to diagnosis horizon (p=0.19, p=0.08, p=0.08 for three data modalities), but for the three-month horizon, results were significantly worse compared to using the entire EHR, with each p-value less than 0.05. This result indicated the important temporal relation of EHR data to the time of diagnosis compared to data temporally farther away. It is therefore feasible to use a limited amount data to predict depression with a similar level of accuracy to using all EHR data.

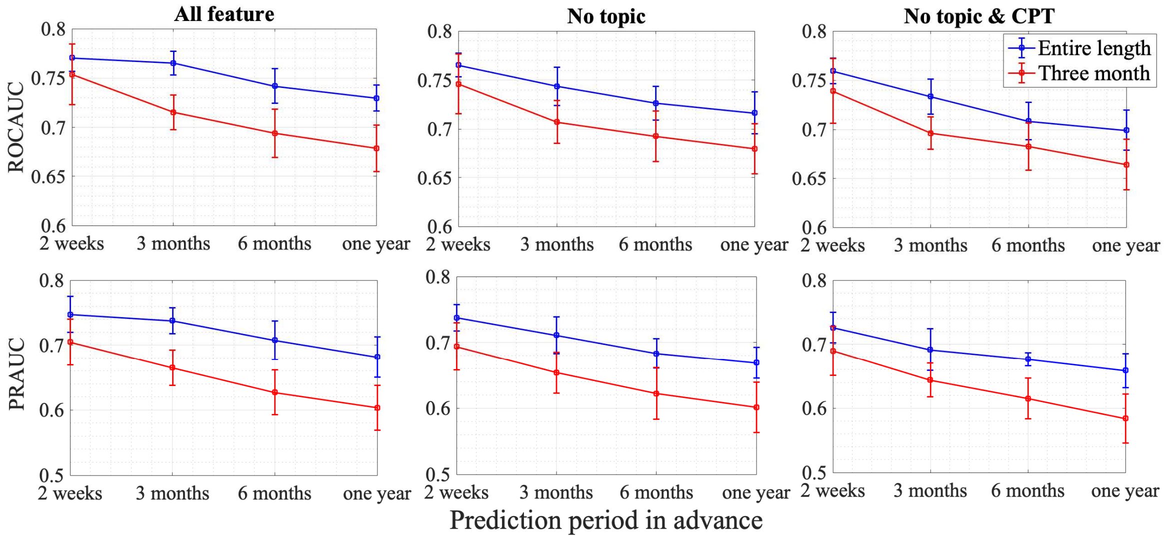


Figure 4.5: Comparison between full EHR length (blue) and three months (red) from RF using the same four prediction windows before. The two rows are shown ROCAUC and PRAUC, respectively while three columns are from three EHR data modalities as Figure 4.4.

## 4.5 Discussion

This chapter presents several machine learning models capable of predicting future diagnoses of depression based on heterogeneous EHR data sources and various time windows. The model was able to generate non-significantly declined accuracy in the prediction window of two weeks using limited length of EHR data. Our results demonstrate the feasibility of making early predictions of depression using five EHR data modalities in a single compact model and suggest the possibility of creating a surveillance system that can identify patients at risk for becoming depressed. Such a system could identify individuals for follow-up diagnostic testing and early intervention.

In general, results indicated that RF outperforms LR and SVM in the prediction task with various data sources and the majority of prediction windows, which demonstrated its robust ability to handle complex heterogeneous EHR data. Shown in Figure 4.4, the RF model's ROCAUC for six-month and one-year windows without including CPT and topic features were 0.71±0.02 and 0.70±0.02, respectively. These results are similar to previously published work [115], as they were 0.712 (95% CI 0.695 to 0.729) and 0.701 (95% CI 0.684 to 0.718), respectively. After including CPT and topic features, our results were significantly better. Specifically, the model performance was significantly improved after including CPT codes and topic features for all four prediction windows, indicating their important roles of diagnosing depression. The results also demonstrate the temporal nature of the task as we observed that performance decreased as prediction windows moved further away from diagnosis, which aligns with our experimental hypothesis (and likely depression etiology) that it is more difficult to predict diagnoses farther out in time. This was reinforced by our results showing an insignificant decrease of performance using only three preceding month of EHR data compared to the entire length when predicting two weeks before a diagnosis. This is noteworthy as it potentially makes our technique more generalizable, as it is easier to collect EHR data for three months rather than longer time spans. In our cohort, the average data length per patient was 6.3 years, thus, using three months of data was equivalent to using only 4% of patients' data on average.

All three machine learning algorithms used in the experiment are time independent, and thus lack the ability to take full advantage of the temporal nature of the data. Temporal models, such as recurrent neural network (RNN) and hidden Markov models (HMM), may be able to process input features in a sequential manner using their inner memory structure. Several studies have used temporal machine learning models for healthcare tasks [23,59,61,75,124]. On the other hand, there are alternative NLP models other than topic models, such as BERT [74] and XLNET [125]. Thus, future work could apply these techniques to process clinical texts, which may further increase the overall model's performance. More details about the future work of improving the precision for this classification system will be presented in Chapter 8.

# CHAPTER 5

# HCET: Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Records for Predicting Depression

## 5.1 Overview

With the rapid development of deep learning algorithms and widespread use of healthcare datasets, many models have presented state-of-the-art performance using patients' electronic health records (EHRs) for diagnostic tasks [21], disease detection [23], and risk prediction [126]. EHRs have been broadly adopted for documenting a patient's medical history [127]. They are composed of data from various sources, including diagnoses, procedures, medications, clinical notes, and laboratory results, which contribute to their high dimensionality and heterogeneity. Frequently, models built on EHR data have limited the number of data categories used [75,116]. Few studies have attempted to use data from a broad set of categories as data heterogeneity remains a technical barrier for utilizing all types of EHR data in one model. As a consequence, there is an ongoing effort to construct a single model that is able to aggregate data from different data modalities. An additional complication is that EHR data includes temporal information from different patient visits, with each visit producing data from various sources.

To construct a predictive model with high accuracy for prediction of depression and mitigate the heterogeneity and sparsity of EHR data, this chapter proposes Hierarchical Clinical Embedding with Topic modeling (HCET), which aggregates diagnoses, procedure codes, medications, and demographic information together with topic modeling of clinical notes. Inspired by [75], HCET builds a hierarchical structure on different categories of EHR data with various embedding levels, while preserving the data's sequential nature. In this way, it learns the inherent interaction between EHR data from various sources within each visit and across multiple visits for an individual patient. This chapter points to a potential method for targeting depression screening among individuals in a single health system who have conditions that are associated with high risk for depression**.** Depression is often not evaluated in primary care settings. This approach could help in clinical practice by identifying individuals potentially at risk for developing depression within a specific time interval who should be screened (and potentially treated) for depression.

In the remainder of this chapter, Section 5.2 details the data and preprocessing steps used in this study. Section 5.3 describes the architecture of HCET and several baseline models for comparison as well as the experimental setup for predicting future diagnosis of depression. Section 5.4 summaries the results while Section 5.5 discusses several observations and limitations. The content of this chapter have partly been published in [94].

## 5.2 Data Cohort and Data Preprocessing

To capture a spectrum of clinical complexity for our analyses, we selected patients based on three primary diagnoses: myocardial infarction (MI), breast cancer, and liver cirrhosis.  Generally, MI represents the least complexity, with acute onset, resolution, and straight-forward treatment. Breast cancer is increasingly complicated in terms of diagnoses and treatment options. Finally, a patient with liver cirrhosis may have many sequelae, generating a complex EHR representation. Patients for this project were identified from our EHR in accordance with an IRB (#14-000204) approved protocol. Each patient visit had EHR data types consisting of diagnosis codes in International Classification of Disease, ninth revision (ICD-9) format, procedure codes in Current Procedural Terminology (CPT) format, medication lists, demographic information, and clinical

Table 5.1: Statistics of EHR dataset.

| | |
|---|---|
| # of patients with MI | 2,943 (1,280 depressed) |
| # of patients with breast cancer | 5,568 (1,960 depressed) |
| # of patients with liver cirrhosis | 2,218 (772 depressed) |
| Gender | Male (27.46%), Female (72.54%) |
| Age | 68.78 ± 15.46, min: 18, max 98 |

notes. All patient records coded with ICD-9 values for MI, breast cancer, or liver cirrhosis from 2006-2013 were included. In this dataset, demographics were limited to the patient's gender and age at the time of each visit. Initially, there were 45,208 patients and after the preprocessing and patient including criteria in Section 5.2.4, 10,148 patients were included in the analysis. Table 5.1 shows statistics of the dataset. Note that some patients had more than one primary diagnosis.

### 5.2.1 Identifying diagnosis of depression

Because patients in this dataset were identified retrospectively and were not suspected for depression, common methods for identifying and assessing severity of depression such as Patient Health Questionnaire (PHQ-9) scores [128] were not available. Instead, depression onset was identified by three methods:

- depression related ICD-9 code [115]
- inclusion of an antidepressant drug in a patient's medication list
- appearance of an antidepressant drug in clinical notes (from https://www.whocc.no/atc_ddd_index/?code=N06A)

The earliest time stamp of an occurrence of any of these events was defined as the time of diagnosis with depression. In total, 3,747 out of the total 10,148 patients were identified as depressed, where the diagnosis time of depression occurred after the primary diagnosis for each patient.

## 5.3 Methods

ICD-9 codes, CPT codes, medication lists, and patient's gender can all be considered as categorical variables while ages are numerical. Therefore, an intuitive approach is to encode these features in a multi-hot vector, where each row corresponds to a specific code or data element. Each row has a binary value, where 1 indicates have this item and 0 for not during one visit. ICD-9 codes are up to five digits long with three digits before a decimal point and two digits after, resulting in 9,285 unique codes in our data set. In order to reduce the dimensionality of the feature vector, ICD-9 codes were grouped by the three numbers before the decimal point, as was previously done in [117]. Detailed descriptions of dimensionality reduction techniques for ICD-9, CPT, and medication lists are presented in Section 5.3.3, the definition of HCET. Embedding is a technique that has been widely adopted in NLP to project long and sparse feature vectors into a dense lower dimensional space [72]. This approach efficiently reduces the size of a model's parameters as well as decreases the training time. Recent models [75,117,129] have utilized embedding to process categorical data in EHRs, which we have adopted in the current model. The full definition is shown in Section 5.3.3.

### 5.3.1 Topic modeling of clinical notes

Latent Dirichlet allocation (LDA) is an unsupervised learning method to encode texts by assigning words to underlying topics (semantic themes). Briefly, a topic is represented as a multinomial distribution over unique words in a corpus, and a document is represented as a multinomial distribution over all topics. LDA is able to generate topics automatically from a corpus, providing generalized information. Recent works have applied topic modeling on clinical notes [120–123]. We chose to model clinical notes with 100 topics, each one contained five words with the top five probabilities to represent the semantic mean of the clinical notes, thus generating a 100-feature vector representation of the document in the semantic topic space. Topic vectors was dichotomize using each topic's average value as a threshold among our data. For patients with multiple clinical reports in the six-month time window, probabilities were averaged first to reach one feature vector and then dichotomized using the same method.

### 5.3.2 Baseline models

Predicting depression prospectively can be considered a continuous task that may be based on various temporal windows of the EHR. Traditional machine learning algorithms generally ignore temporal and sequential correlation among features by aggregating them over a time window for a patient. As mentioned in the first paragraph of Section 5.3, the feature vector for each patient is a multi-hot vector which concatenated all five EHR data modalities over multiple visits. In order to leave out the bias for more frequent codes, each row of vector is 1 when this code shows in any of the visits. As a compensation factor for temporal information, the number of records in ICD-9, CPT, medication lists, and clinical notes are added as addition factors to capture the of frequency of patients visits of records. 10-fold cross validation was adopted for each model. In addition, patients in the test set were separated by their primary diagnosis and the results were compared for three primary diagnosis individually.

**Lasso:** Previous work has applied Lasso for predicting depression [115], which was compared in the analysis. Lasso uses L1 regularization which brings sparsity to select the more correlated features for the task.

**SVM:** SVM was also compared in the experiment as it has been utilized to predict depression previously [111]. Here we used RBF kernel and five-fold cross validation with a grid search to finetune the regularization term.

**Multilayer perceptron (MLP):** Two layers of MLP with a tanh activation function and 256 nodes is also compared here, following the implementation from previous studies [59,75].

**RF:** Nevertheless, ensemble methods like random forests (RF) [130] and gradient boost regression trees (GBRT) [131] have produced competitive results in disease detection and outcome prediction for healthcare. These models also compute the significance factor for each feature, which provides valuable information on feature selection as well as dimension reduction. Therefore, RF was adopted as a baseline model in comparison with HCET. The hyper parameters were chosen using grid search with five-fold cross validation on the training set.

**VAE+RF:** [116] proposed pretraining autoencoder as the feature extractor for EHR and using RF for classification from the extracted features. This method was also compared.

**MiME*:** The MiME model demonstrated state-of-the-art performance in predicting heart failure onset [75]. It consists of a temporal model using GRUs that learn the temporal character of disease progression with external knowledge of linked relation between ICD-9 codes and associated CPT codes and medication lists during each visit. The MiME model required removal of visits that did not include diagnosis codes to make sure diagnosis codes were present to input the model. Since there was no direct linked relationship between ICD-9 codes, CPT codes, and medication lists in our EHR data, these three features were processed in the same level instead of the two-level structure proposed in MiME. In addition, there are many cases where procedure codes or medications are present in the EHR without associated diagnoses. Therefore, we revised the MiME model by removing this layer while keeping the remaining structure and some parameters, denoted MiME*. The performance of this modified model was compared to our HCET model.

$$L_{aux} = -\lambda_{aux} \sum_{t}^{T} \Big( \sum_{i}^{|v^{(t)}|} CE\left(d_i^{(t)}, \hat{d}_i^{(t)}\right) + \sum_{i}^{\left|M_i^{(t)}\right|} CE\left(m_{i,j}^{(t)}, \widehat{m}_{i,j}^{(t)}\right) \Big) \tag{5.3.1}$$

As shown above, MiME defined Equation (5.3.1) to compute the auxiliary loss, where $d_i^{(t)}$ denoted the diagnosis code in $t^{th}$ visit. Thus, calculating auxiliary loss required diagnosis codes present in each visit, which was not applicable to our dataset. On the other hand, we highly focused on the prediction accuracy of depression but not on other diseases or symptoms. Furthermore, the average performance after implementing this component was increased less than 0.01 from their reported results, so the auxiliary loss defined in MiME was not adopted in this study.

### 5.3.3 Definition of HCET

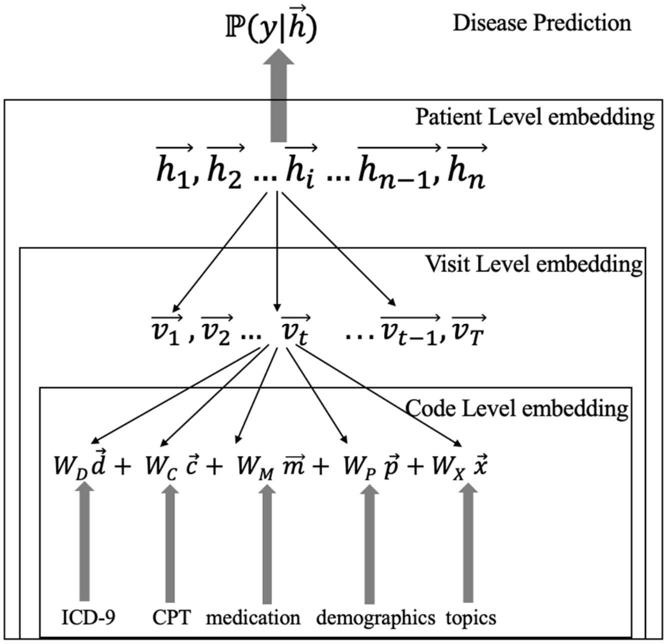

$\mathbb{P}(y|\vec{h})$
Disease Prediction
Patient Level embedding
$\overrightarrow{h_1}, \overrightarrow{h_2} \dots \overrightarrow{h_i} \dots \overrightarrow{h_{n-1}}, \overrightarrow{h_n}$
Visit Level embedding
$\overrightarrow{v_1}, \overrightarrow{v_2} \dots \overrightarrow{v_t} \dots \overrightarrow{v_{t-1}}, \overrightarrow{v_T}$
Code Level embedding
$W_D\vec{d} + W_C\vec{c} + W_M\vec{m} + W_P\vec{p} + W_X\vec{x}$
ICD-9
CPT
medication
demographics
topics

Figure 5.1: Illustration of HCET for EHR data. There are three levels of embedding: patient level, visit level and code level. The full explanation of symbols is described in Table 5.2.

Figure 5.1 illustrates the hierarchical structure of HCET. The ultimate goal of the model is to predict the probability of a chronic disease for patient $i$ given the feature embedding representing a sequence of visits, $\mathbb{P}(y_i|\overrightarrow{\mathrm{h_i}})$. While the model is designed to be generalizable, we focus here on the prediction of depression, $y_i$. $\overrightarrow{\mathrm{h_i}}$ stands for the patient level embedding of a patient's EHR, and each patient has multiple hospital visits from $\overrightarrow{v_1}$ to $\overrightarrow{v_t}$ , which compose the visit level embedding. During one visit $\overrightarrow{v_t}$ , the code level embedding $\overrightarrow{e_t}$ is the ensemble of multiple ICD-9 and CPT codes, medications, demographic information, and topic features extracted from associated clinical notes. Since there are five categories of EHR data, we built individual embedding for each first and aggregated them together.

Table 5.2 shows the full list of notation and corresponding definitions of symbols used in HCET. $\vec{d}$ is a multi-hot binary vector with dimension of $\mathbb{R}^{D\times 1}$, where each column corresponds to whether a specific ICD-9 code was assigned in the $t^{\text{th}}$ visit. A similar approach applies to $\vec{c} \in \mathbb{R}^{C\times 1}$, $\vec{m} \in \mathbb{R}^{M\times 1}$ , and $\vec{p} \in \mathbb{R}^{2\times 1}$ , which are the vector representations for CPT, medication, and demographic information, respectively. As described before, topic features are vector

Table 5.2: Notation used in the formulation of HCET.

| Notation | Definition |
|---|---|
| $D$ | Unique set of ICD-9 codes |
| $C$ | Unique set of CPT codes |
| $M$ | Unique set of medications |
| $X$ | Set of 100 topic features |
| P | Demographic information |
| $\lambda_j$ | Attention weight for one data modality, j ∈ (*D,C,M,X,P*) |
| $\overrightarrow{e_t} \in \mathbb{R}^z$ | Vector representation of summed EHR data at the $t$-th visit |
| $\overrightarrow{v_t} \in \mathbb{R}^z$ | Vector representation of $t \in [1 \dots T]$ visit EHR data for a patient |
| $\overrightarrow{\mathrm{h_i}} \in \mathbb{R}^z$ | Vector representation of EHR data for patient number $i$ |

The dimension of embedding $z$ is the same for associated vectors due to the residual connection used in HCET.

representations in the topic space, which represent the distribution of topic occurrences in the document. In order to match the embedding size of other data types, a threshold was defined to dichotomies each topic word, which was computed by the average probability of each topic value across all patients. Thus, $\vec{x} \epsilon \mathbb{R}^{100\times1}$ is a multi-hot binary vector representation of topic features, where each column denotes the unique 100 topics for the $t^{\text{th}}$ visit.

Equations (5.3.2), (5.3.3) and (5.3.4) describe mathematical formulation of HCET in the top-down view, denoting the *Patient level*, *Visit level*, and *Code level* embeddings, respectively.

$$\overrightarrow{\mathrm{h_i}} = f(\overrightarrow{v_1} \dots \overrightarrow{v_t} \dots \overrightarrow{v_T}) \tag{5.3.2}$$

Equation (5.3.2) shows the method to process temporal information for various visit level embeddings to compute a patient level embedding, where $f$ stands for the function to input visit information in a sequential order. As mentioned before, RNNs, LSTMs, and GRUs have been widely used to fulfill this task. Since RNNs often encounters the vanishing gradient problem and better performance has been shown for a GRU over an LSTM in previous work [75], we used a GRU in the current model.

$$\overrightarrow{v_t} = \alpha(W_e\overrightarrow{e_t}) + \overrightarrow{e_t} \tag{5.3.3}$$

In Eq. (5.3.3), visit level embedding is generated by first performing a matrix transformation with weight $\boldsymbol{W}_e \ \epsilon \ \mathbb{R}^{z\times z}$, followed by a non-linear ReLU transformation function $\alpha$, where $z$ is the embedding size. We omitted the bias term $\overrightarrow{b_t}$ here to formulate the residual connection [132].

$$\overrightarrow{e_t} = \beta(F) + F \tag{7.3.4}$$

$$F = W_D\vec{d} + W_C\,\vec{c} + W_M\,\vec{m} + W_P\,\vec{p} + W_X\,\vec{x} \tag{7.3.5}$$

Equations (5.3.4) and (5.3.5) define the code level embedding by summing individual embeddings from five EHR data sources with a non-linear transformation function $\beta$. As in equation (7.3.4), we use a ReLu for $\beta$. The $W_D \epsilon \mathbb{R}^{z \times D}$, $W_C \epsilon \mathbb{R}^{z \times C}$, $W_M \epsilon \mathbb{R}^{z \times M}$, $W_P \epsilon \mathbb{R}^{z \times P}$ and $W_X \epsilon \mathbb{R}^{z \times X}$ represent the weight matrices for transforming the feature vectors of ICD-9 codes, CPT codes, medication lists, demographics, and topic features with high and varied dimensionality into a latent space with the same lower dimension, respectively. For example, the diagnosis vector $\vec{d} \epsilon \mathbb{R}^{D \times 1}$, after multiplied with weight matrix $W_D \vec{d}$ ,results in a vector of dimension $\mathbb{R}^{z \times 1}$. Therefore, all vectors can sum up as in equation (7.3.5). In the same manner to Eq. (7.3.4), all of the corresponding biased terms were omitted to denote the residual connection. Finally, binary cross entropy was used as the loss function.

### 5.3.4 Predicting depression at different decision points

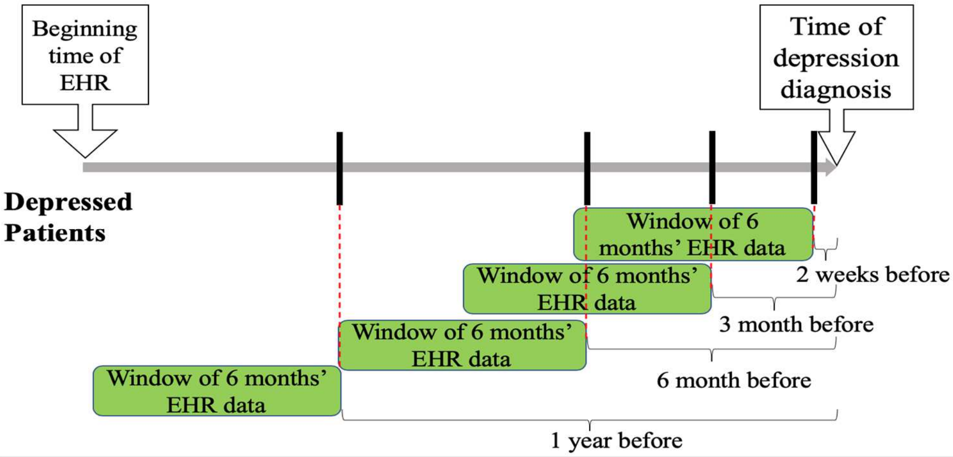

Beginning time of EHR
Time of depression diagnosis
Depressed Patients
Window of 6 months' EHR data
Window of 6 months' EHR data
Window of 6 months' EHR data
Window of 6 months' EHR data
2 weeks before
3 month before
6 month before
1 year before

Figure 5.2: Illustration of prediction at different time windows in advance of diagnosis of depression. The beginning time of EHR is defined by the timestamp of the primary diagnosis.

Previous studies [21,23,75,129] have used the data from the entire EHR for future disease prediction. This method could add bias for patients with longer medical histories. It also gives equal weight to old data that likely is not as useful as more recent data. As predicting the future risk of a disease in a prospective setting is an ongoing task, the time window of a patient's EHR is highly varied. Therefore, as a similar approach to [115], we defined four decision points in advance of the diagnosis of depression: two weeks, three months, six months, and one year. Figure 5.2 illustrates the four prediction windows for using EHR data to predict depression diagnosis. For non-depressed patients, the last time step of the EHR was substituted for the diagnosis time.

In order to test the effect of temporal information and data size on model's performance, previous work [75] used varying maximum lengths (visits) of the EHR. This resulted in a different number of patients in each of the four experiments as the number of visits was not consistent across patients. In our approach, we kept the number of patients consistent through the four predicting windows, which revealed the temporal nature of prediction as the time to diagnosis varies. In this case, patients who had at least one of ICD-9, CPT, medication and topic feature in all four time windows were included in experiments. After processing data based on this method, 10,148

Table 5.3: Statistics of data input for HCET.

| | |
|---|---|
| Total # of patients | 10,148 (Depressed:3,747; Non-depressed:6,401) |
| Total # of visits | 294,941 |
| Avg. # of visits | 29.06 |
| # of unique codes | $D$:1391, $C$:6927, $M$: 4181 |
| # of demographics per visit | 2 (Age, Gender) |
| # of topics per visit | 100 |
| Max / Avg. # of ICD-9 codes per visit | 69 / 1.74 |
| Max / Avg. # of CPT codes per visit | 106 / 3.23 |
| Max / Avg. # of medication per visit | 14 / 0.09 |
| Max / Avg. # of topics per visit | 30 / 1.87 |

patients were selected, where 3,747 were diagnosed with depression. Basic statistics of the data are shown in Table 5.3.

**Ablation study:** Three feature sets were generated to compare the contribution to predicting depression for demographics and topic features: all data types (ICD-9 codes, CPT codes, medication lists, demographics, and topic features); ICD-9 codes, CPT codes, medication lists and topic features; ICD-9 codes, CPT codes, and medication lists. This ablation study was only applied to HCET while all baseline models used all data types as input.

### 5.3.5 Training details

All models were implemented in TensorFlow 1.12 and trained on a workstation equipped with Intel Xeon E3-1245, 32 GB RAM and two NVIDIA Ti 1060 GPUs. Adam [133] was selected as the optimizer, with the same learning rate of $1e^{-3}$ as [75] for HCET. The number of parameters is 2.5M, which mainly depends on the size of embedding matrices. Reported results are averaged over 10 random data splits: training 70%, validation 10% and test 20%. Models were trained with the minibatch of 50 patients for a total of 2,000 iterations to guarantee convergence. The validation set was evaluated at every 100 iterations for early stopping. The vanishing gradient problem was avoided by using skip connections. To address over fitting, L2 regularization with coefficient $1e^{-4}$ was chosen for HCET models instead of using dropout. The embedding size $z$ was set as 200 and the number of nodes for the GRU was set at 256. The source code of HCET is available at https://github.com/lanyexiaosa/hcet.

Table 5.4: Comparison of prediction performance for different models.

| Metrics | Prediction window | Lasso | SVM | MLP | RF | VAE+RF | MiME* | HCET | **HCET** |
|---|---|---|---|---|---|---|---|---|---|
| ROCAUC | Two weeks | 0.66 (0.01) | 0.72 (0.02) | 0.72 (0.01) | 0.76 (0.02) | 0.76 (0.02) | 0.76 (0.02) | 0.76 (0.02) | **0.81† (0.02)** |
| | Three months | 0.65 (0.02) | 0.69 (0.01) | 0.70 (0.02) | 0.73 (0.02) | 0.74 (0.01) | 0.74 (0.01) | 0.75 (0.01) | **0.80† (0.01)** |
| | Six months | 0.63 (0.02) | 0.68 (0.02) | 0.69 (0.02) | 0.70 (0.02) | 0.71 (0.03) | 0.72 (0.03) | 0.73 (0.03) | **0.78† (0.03)** |
| | One year | 0.63 (0.02) | 0.68 (0.02) | 0.68 (0.02) | 0.69 (0.02) | 0.69 (0.02) | 0.70 (0.02) | 0.71 (0.02) | **0.75† (0.02)** |
| PRAUC | Two weeks | 0.55 (0.02) | 0.62 (0.03) | 0.64 (0.01) | 0.67 (0.03) | 0.67 (0.02) | 0.67 (0.02) | 0.68 (0.01) | **0.73† (0.02)** |
| | Three months | 0.52 (0.03) | 0.59 (0.02) | 0.60 (0.02) | 0.62 (0.03) | 0.64 (0.02) | 0.64 (0.02) | 0.65 (0.02) | **0.71† (0.02)** |
| | Six months | 0.51 (0.03) | 0.57 (0.02) | 0.58 (0.02) | 0.59 (0.02) | 0.60 (0.02) | 0.61 (0.01) | 0.62 (0.01) | **0.68† (0.02)** |
| | One year | 0.50 (0.03) | 0.57 (0.03) | 0.57 (0.02) | 0.58 (0.03) | 0.60 (0.02) | 0.61 (0.01) | 0.61 (0.01) | **0.66† (0.02)** |

Values in parenthesis refer to standard deviations across randomizations and bold values denotes the highest in each column. † indicates the value is significantly better than MiME* ($p<0.05$). All data modalities were input to Lasso, SVM, MLP, RF, VAE+RF, and **HCET**. MiME* took ICD-9, CPT, and medication lists as the input while HCET combined demographics with those.

## 5.4 Evaluation and Results

### 5.4.1 Comparison of performance in depression prediction

Table 5.4 displays the results from all baseline models and HCET with abalation analysis at four time points in advance of diagnosis in terms of receiver operating characteristic area under the curve (ROCAUC) and PRAUC. HCET using all EHR modalities outperformed other models for every prediction window. Lasso generated the worst accuracy. There was no significant difference between results from RF and VAE+RF. There was a consistent decrease of accuracy for each model as the prediction window moves further away from the time of diagnosis, where the number achieved the highest at the window of two weeks.

Adding demographic information and topic features improved the performance for HCET, which demonstrated their significant contribution in predicting depression as well as emphasized

Table 5.5 : Comparison of prediction performance for three primary diagnosis.

| Prediction window | Disease | Metrics | Lasso | SVM | MLP | RF | VAE+RF | MiME* | **HCET** |
|---|---|---|---|---|---|---|---|---|---|
| Two weeks | Breast cancer | ROCAUC | 0.67<br>(0.02) | 0.72<br>(0.02) | 0.74<br>(0.02) | 0.76<br>(0.02) | 0.76<br>(0.02) | 0.77<br>(0.01) | **0.81<br>(0.01)** |
| | | PRAUC | 0.54<br>(0.03) | 0.61<br>(0.03) | 0.63<br>(0.02) | 0.66<br>(0.03) | 0.67<br>(0.02) | 0.67<br>(0.02) | **0.73<br>(0.01)** |
| | MI | ROCAUC | 0.66<br>(0.02) | 0.71<br>(0.02) | 0.72<br>(0.02) | 0.74<br>(0.03) | 0.75<br>(0.02) | 0.75<br>(0.01) | **0.79<br>(0.02)** |
| | | PRAUC | 0.62<br>(0.03) | 0.68<br>(0.03) | 0.69<br>(0.02) | 0.71<br>(0.02) | 0.71<br>(0.01) | 0.70<br>(0.01) | **0.77<br>(0.01)** |
| | Liver cirrhosis | ROCAUC | 0.65<br>(0.02) | 0.71<br>(0.02) | 0.72<br>(0.02) | 0.75<br>(0.03) | 0.75<br>(0.02) | 0.76<br>(0.02) | **0.80<br>(0.01)** |
| | | PRAUC | 0.55<br>(0.02) | 0.60<br>(0.03) | 0.62<br>(0.02) | 0.65<br>(0.03) | 0.65<br>(0.02) | 0.67<br>(0.01) | **0.72<br>(0.01)** |
| One year | Breast cancer | ROCAUC | 0.64<br>(0.03) | 0.68<br>(0.03) | 0.69<br>(0.01) | 0.70<br>(0.03) | 0.70<br>(0.02) | 0.71<br>(0.02) | **0.78<br>(0.02)** |
| | | PRAUC | 0.49<br>(0.03) | 0.56<br>(0.03) | 0.56<br>(0.02) | 0.57<br>(0.03) | 0.58<br>(0.03) | 0.61<br>(0.01) | **0.67<br>(0.02)** |
| | MI | ROCAUC | 0.62<br>(0.02) | 0.67<br>(0.02) | 0.66<br>(0.01) | 0.67<br>(0.02) | 0.68<br>(0.01) | 0.69<br>(0.02) | **0.77<br>(0.01)** |
| | | PRAUC | 0.56<br>(0.04) | 0.62<br>(0.03) | 0.62<br>(0.02) | 0.63<br>(0.02) | 0.63<br>(0.01) | 0.64<br>(0.01) | **0.71<br>(0.01)** |
| | Liver cirrhosis | ROCAUC | 0.62<br>(0.04) | 0.66<br>(0.02) | 0.66<br>(0.02) | 0.67<br>(0.01) | 0.68<br>(0.02) | 0.70<br>(0.01) | **0.77<br>(0.02)** |
| | | PRAUC | 0.53<br>(0.02) | 0.55<br>(0.02) | 0.56<br>(0.02) | 0.57<br>(0.03) | 0.58<br>(0.02) | 0.61<br>(0.02) | **0.66<br>(0.02)** |

the advantage of building a model being able to aggregate EHR data from multiple sources. The values between MiME* and HCET were similar, while the difference between HCET and HCET shown in bold were relatively large. HCET with all types of EHR data achieved the highest accuracy at each prediction window than all baseline models. It generated the highest mean ROCAUC of 0.81 when predicting two weeks prior to the diagnosis, and the value dropped to 0.75 when predicting one year in advance.

### 5.4.2 Model's performance for each primary diagnosis

Table 5.5 shows the results for each of three primary diagnosis in predictions windows of two weeks and one year in ROCAUC and PRAUC. HCET also achieved the best performance for three primary diagnosis for two prediction windows. The low variance also indicated that it is more robust than other models. The ROCAUC for every model was quite similar even though the number of patients with breast cancer was substantially higher than the other diseases (Table 5.1), which indicated no bias toward any primary diagnosis in the prediction. On the other hand, it is noticeable that the PRAUC for patients with myocardial infarction was relatively higher than other two.

## 5.5 Discussion

In this chapter, I developed a temporal deep learning model, HCET, which was able to integrate five types of EHR data during multiple visits for depression prediction. HCET consistently outperformed the baseline models tested, achieving an increase in PRAUC of 0.07 over the best baseline model. The results demonstrated the ability of HCET as an approach to deal with data heterogeneity and sparsity in modeling the EHR.

Accoring to results in Table 5.4 and 5.5, Lasso generated the worst performance as it is a linear classifier which indicats that predicting depression from the EHR is a complicated task which requires more advanceted models. In addition, the Lasso method also provides sparsity of using more correlated features but the poor accuracy reveals that this task needs to include more features than only the most correlated ones. There was no significant different between results from RF and VAE+RF which indicated the power of classfication maily depends on RF. Models starting from MiME* are all temporal models and they all achieved higher performance than non-temporal ones, which further confirms the advantage of using temporal models over non-temporal ones in predicting chronic diseases. Furthermore, the performance consistently declined for each model as the prediction window moved further away from the diagnosis time point, which agrees with our expectation that records closer to the diagnosis are more likely to contain relevant information and provide better predictions.

The improvement of HCET over all baseline models demonstrated the advantage of utilizing temporal information and the hierarchical embedding to aggregate more heterogenous EHR data modalities to predict future diagnosis of depression. In the original implementation of the MiME model [75], interactions between diagnosis codes with associated procedures and medication were explicitly modeled, but this linked relation was not available in our EHR data, which  is a situation that commonly applies to other medical systems. Meanwhile, MiME also has another limitation of ignoring data when no diagnosis code is present for each visit. Our results indicate that treating all EHR data types in one level of code embedding during each visit is a viable solution in this scenario while being able to include all data modalities from each visit. Another adjustment in our model is the extension of embedding to process demographics and clinical notes, which further addresses the heterogeneity issue in EHR data.

In future work, HCET could possibly be used as the basis for constructing a screening tool by utilizing the models' predictions to intervene with individuals who have a higher risk of developing depression. More details about the future work of improving the precision of HCET will be presented in Chapter 8.

# CHAPTER 6

# Bidirectional Representation Learning with Transformer on Multimodal EHR to Predict Depression

## 6.1 Overview

Electronic health record (EHR) systems have become the main method of documenting patients' historical medical records over the last decade [40]. The latest report from the Office of the National Coordinator for Health Information Technology (ONC), stated that nearly 84% of hospitals have adopted at least a basic EHR system, which was a nine-fold increase from 2008 [134]. EHRs are composed of data from different modalities, documented in a sequence for each patient encounter, including demographic information, diagnoses, procedures, medications or prescriptions, clinical notes written by physicians, images, and laboratory results, which contribute to their high dimensionality and heterogeneity [18,94]. Deep learning algorithms enable the usage of EHR data not only as a documenting method for billing purposes, but also as a source of tremendous amount of data to construct classification or prediction models, which build the foundation for creating clinical decision support systems and personalized precision medicine. However, there is an unsolved challenge of achieving high accuracy while providing adequate explanation on model's decision-making process. Although several efforts have attempted to improve model interpretability [59,60,135], they did not address the problem of data heterogeneity that is pervasive in medical research as EHR is often composed of data from various modalities in a sequential structure.

The goal of this study is to create a model with high interpretability for predicting future diagnosis of depression while being able to accommodate the heterogeneity of EHR data and process it effectively in a temporal manner. I propose a Bidirectional Representation Learning model with a Transformer architecture on Multimodal EHR (BRLTM). This BRLTM is able to aggregate five EHR data modalities: diagnoses, procedure codes, medications, demographics and clinical notes. The remainder of this chapter is organized as follows. Section 6.2 details the data and preprocessing steps used in this study. Section 6.3 describes the architecture of BRLTM and several baseline models for comparison as well as the experimental setup for predicting future diagnosis of depression. Section 6.4 summaries the results while Section 6.5 discusses several observations and limitations. The content of this chapter has partly been under revision for a submitted manuscript to the IEEE journal of biomedical and health informatics.

## 6.2 Data Cohort and Data Preprocessing

Patients selected for this study were based on three primary diagnoses: myocardial infarction (MI), breast cancer, and liver cirrhosis, to capture a spectrum of clinical complexity. Generally, MI represents the least complexity, with acute onset, resolution, and straight-forward treatment. Breast cancer is increasingly complicated in terms of diagnoses and treatment options. Finally, a patient with liver cirrhosis may have many sequelae, generating a complex EHR representation. Patients for this work were identified from our EHR in accordance with an IRB (#14-000204) approved protocol. Each patient visit had EHR data types consisting of diagnosis codes in International Classification of Disease, ninth revision (ICD-9) format, procedure codes in Current Procedural Terminology (CPT) format, medication lists, demographic information, and clinical notes represented as 100 topics using LDA analysis [118]. All patient records coded with ICD-9 values for MI, breast cancer, or liver cirrhosis from 2006-2013 were included. Demographics were limited to the patient's gender and age. Initially, there were 45,208 patients and after the preprocessing to eliminate patients with fewer than two visits, 43,967 patients were included in the analysis. More importantly, data after the diagnosis time of depression was excluded for depressed patients to ensure no data leakage while all of them were included for non-depressed ones.

Patient Health Questionnaire (PHQ-9) scores [128], the most common way to identify the diagnosis of depression, were not available for this patient cohort during the data collection period. Hence, depression onset was identified by three methods: depression related ICD-9 codes, inclusion of an antidepressant drug in a patient's medication list, or appearance of an antidepressant drug in clinical notes, which has been used in [94]. More details are referred back to Chapter 5.2.1.

## 6.3 Methods

In NLP implementations, BERT models process words sequentially. This method can be applied to EHR data by analyzing ICD-9 codes, CPT codes, medication lists, and topics as code sequences representing a patient's visits. Full ICD-9 codes are high dimensional that are sparsely represented with 9,285 distinct codes in our dataset. As in [117], dimensionality was reduced by grouping codes by the three numbers before the decimal point to reduce its feature dimension to 1,131. Each demographic is added as an individual feature and repeated for every sequence. Pretraining was conducted through masked language modeling (MLM) to predict the mask code based on EHR sequences [74]. After pretraining, the saved model was added a classification head to finetune for the downstream task of chronic disease prediction. The feature dimension as the unique number of codes for five modalities were listed as follows:

- Diagnoses: 1,131
- Procedures: 7,048
- Medications: 4,181
- Demographics: 2
- Topics: 100

### 6.3.1 BLRTM model for EHR representation learning

Eq. (6.3.1) shows a patient's EHR, composed by different visits ranging from $V_1$ to $V_L$, where L is the length of the EHR sequence. Two symbolic tokens $CLS$ and $SEP$ are adopted here: $CLS$ denotes the starting point of the EHR and $SEP$ denotes the separation between two consecutive visits.

$$EHR: (CLS, V_1, SEP, V_2, SEP, \dots, V_L) \tag{6.3.1}$$

Each of the visits $V_t$ is comprised of EHR codes $X$, as shown in Eq. (6.3.2), where the number of codes is $m_t$, which varies for each visit. Every code is from the vocabulary of the dataset: $D$: diagnosis, $C$: procedure, $M$: medication and $T$: topics.

$$V_t: \left(X_1, X_2, \dots, X_{m_t}\right), \qquad X \epsilon \{D, C, M, T\} \tag{6.3.2}$$

The original BERT model has three types of embeddings: token, position, and segment [74]. In our BRLTM model, we treated each token embedding as a code embedding and extended the model's ability to aggregate demographics by adding age and gender embedding, shown in Figure 6.1. There are five types of embeddings which are summed to generate the final output embedding

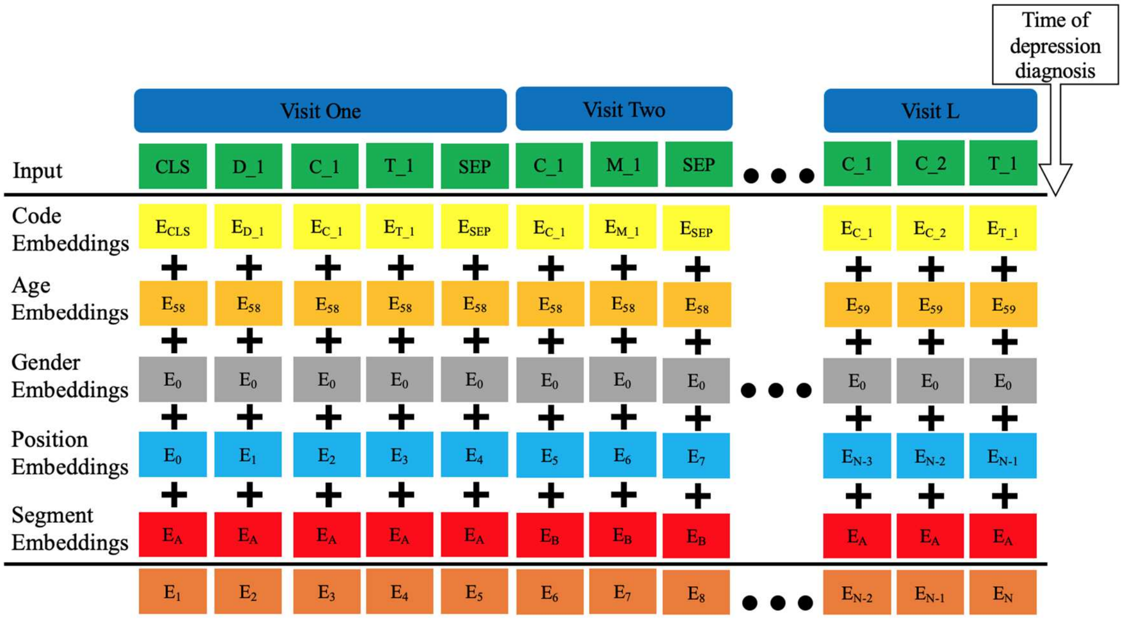

Time of depression diagnosis
Visit One
Visit Two
Visit L
Input
CLS D_1 C_1 T_1 SEP C_1 M_1 SEP C_1 C_2 T_1
Code Embeddings
$E_{CLS}$ $E_{D\_1}$ $E_{C\_1}$ $E_{T\_1}$ $E_{SEP}$ $E_{C\_1}$ $E_{M\_1}$ $E_{SEP}$ $E_{C\_1}$ $E_{C\_2}$ $E_{T\_1}$
Age Embeddings
$E_{58}$ $E_{58}$ $E_{58}$ $E_{58}$ $E_{58}$ $E_{58}$ $E_{58}$ $E_{58}$ $E_{59}$ $E_{59}$ $E_{59}$
Gender Embeddings
$E_0$ $E_0$ $E_0$ $E_0$ $E_0$ $E_0$ $E_0$ $E_0$ $E_0$ $E_0$ $E_0$
Position Embeddings
$E_0$ $E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_7$ $E_{N-3}$ $E_{N-2}$ $E_{N-1}$
Segment Embeddings
$E_A$ $E_A$ $E_A$ $E_A$ $E_A$ $E_B$ $E_B$ $E_B$ $E_A$ $E_A$ $E_A$
$E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_7$ $E_8$ $E_{N-2}$ $E_{N-1}$ $E_N$

Figure 6.1: Architecture of the BRLTM model for EHR representation learning. The subscripts show the original value for each embedding. CLS and SEP are symbolic tokens stands for the beginning of EHR and separation of two visits adjacent to each other, respectively. D, C, M, and T denote diagnoses, procedures, medications, and topics, respectively. The last row denotes the sum of the five embeddings as the final output embedding.

for training. Data after the depression diagnosis were excluded to avoid data leakage. For non-depressed patients, the last time step of the EHR was substituted for the diagnosis time. Code embeddings are from the four EHR data modalities mentioned in Eq. (6.3.2). Similar to the original BERT model, position and segment embeddings indicate the position of one code in the full sequence and distinguishes codes in adjacent visits, respectively. We adopted pre-determined instead of learned encodings for positional embeddings to avoid weak learning of the positional embedding due to high variety in a patient's sequence length. The position embeddings play an important role in sequence learning, equivalent to the recurrent structure in RNNs. Annotating the position of each code in the sequence enables the model to capture the positional interactions among EHR data modalities. However, position embeddings do not tell whether codes are from the same visit or not. Hence, segment embeddings are used to provide extra information to

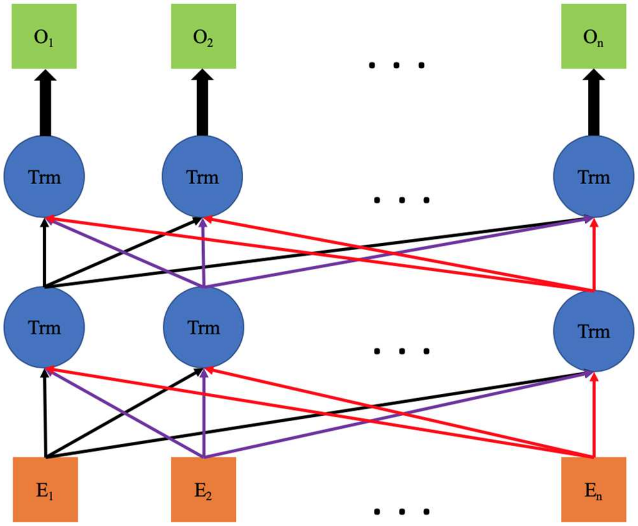


Figure 6.2: Illustration of bidirectional learning with the transformer architecture. The orange squares are the final out embeddings in Figure 6.1, which are the input sequences here. Trm stands for the transformer while the green squares denote the output sequence. O denotes the output for each code after learning.

differentiate codes in adjacent visits by alternating between two trainable vectors, depicted as A and B in Figure 6.1.

Age and gender embeddings are repeated in every position of the sequence. Combining code embeddings with the age embedding not only enables the model to use age information as a feature, but also provides temporal information in the sequence. As shown Figure 6.2, the final embeddings are the input for a bidirectional sequential learning step with the transformer architecture as in the BERT model [74]. The latent contextual representation of five data modalities in temporal EHR sequences can be efficiently learned from the aggregation of these five embeddings. This architecture is capable of aggregating multimodal EHR data into a single model and processing them in a temporal manner, as well as investigating the inner association contingency between them in various visits. In total, the model has the ability to perform representation learning on patient's EHR. More importantly, it realizes the common two-stage transfer learning approach on EHR modeling, which has been widely adopted and has achieved the outstanding performance in computer vision [66] as well as NLP [74,136].

### 6.3.2 Pretraining with mask language modeling (MLM)

An EHR is composed of multimodal code sequences, which is similar to the way that a language is composed of word sequences. Hence, we hypothesized that the advantage of deep bidirectional sequential learning in language modeling over either a left-to-right model or the shallow concatenation of a left-to-right and a right-to-left model can be transferred to EHR modeling. As a consequence, we adopted the same pretraining approach of MLM from the original BERT paper [74]. Namely, we randomly selected 15% of EHR codes and modified them according to following procedures:

- 80% of the time replace them with [MASK]
- 10% of the time replace them with a random disease word
- 10% of the time do nothing and keep them unchanged

This structure in MLM forces the model to learn the distributional contextual representation between EHR codes as the model does not know which codes are masked or which codes have replaced by a random code. EHR modeling is not affected significantly because only 1.5% (10% of 15%) of codes are randomly replaced. This random replacement brings a small perturbation that distracts the model from learning the true contextual sequences of the EHR and forces the model to identify the noise and continue learning the overall temporal progression. We used the precision score (true positives divided by predicted positives) at a threshold of 0.5 as the metric to evaluate pre-training MLM task. The average is calculated over every masked code over all patients. Similar to [137], we followed results from previous models [74,137] with random search to find the best set of hyperparameters during training. In addition, we investigated the contribution of each data modality by training the model with different combinations of CPT and topic features in an ablation study. The data used for MLM is shown in the second column of Table 6.1. Note that some patients had more than one primary diagnosis.

### 6.3.3 Finetuning to predict depression

After pretraining to learn the latent contextual representation of the EHR, one feed-forward classification layer was added to predict diagnosis of depression after finetuning on a specific dataset. We followed the same data selection criteria as in the HCET model [94] with four prediction windows prior to time of depression diagnosis: two weeks, three months, six months, and one year. The length of each data window was restricted to six months instead of patient's entire history to avoid bias towards patients with longer medical histories. Patients who had at least one ICD-9, CPT, medication or topic feature in all four time windows were included. After processing data based on this method, 10,148 patients were selected, where 3,747 were diagnosed with depression. Basic statistics of the data for this prediction task are shown in the third column

Table 6.1: Statistics of datasets for two training approaches.

| Datasets | pretraining | finetuning |
|---|---|---|
| Patients with MI | 10,616 (2,915 depressed) | 2,943 (1,280 depressed) |
| Patients with breast cancer | 23,3077 (4,483 depressed) | 5,568 (1,960 depressed) |
| Patients with liver cirrhosis | 11,757 (2,359 depressed) | 2,218 (772 depressed) |
| Gender | 70.18% female | 72.54% female |
| Age | 65.78 ± 14.99,<br>min: 18, max 100 | 68.78 ± 15.46,<br>min: 18, max 98 |
| Sequence length | 54.64 ± 45.37,<br>min:2, max: 1,186 | 54.64 ± 45.37,<br>min:2, max: 180 |

of Table 6.1. Finally, predicting performance of each model was evaluated in receiver characteristic area under curve (ROCAUC) and PRAUC.

### 6.3.4 Training details

The BRLTM model was implemented in Pytorch 1.4 and trained on a workstation equipped with an Intel Xeon E3-1245, 32 GB RAM and a 12G NVIDIA TitanX GPU. We followed the training scheduler with the Adam [133] optimizer used the original BERT model [74] and set the warmup proportion and weight to 0.01 and 0.1, respectively. The Gaussian error linear unit (GELU) rather than the standard ReLu was used as the non-linear activation function in the hidden layers. Pretraining of MLM used the first dataset with the minibatch of 256 patients for 100 epochs and evaluated at every 20 iterations. The dataset for finetuning of the prediction task underwent 10 random data splits: 70% training, 10% validation, and 20% test, and trained with minibatch of 64 patients for 50 epochs. Dropout of 0.1 was set to both hidden layers to address overfitting. The source code and more detailed description of the model is available at https://github.com/lanyexiaosa/bert_ehr.

### 6.3.5 Baseline models

Recent developments in natural language processing (NLP) provide a number of potential methods that can be applied to EHR data. Previous studies have applied temporal deep learning models on time-series medical data, particularly on EHR data to predict future diagnoses [21,117]. Retain first added a reverse time attention mechanism to RNN for heart failure prediction, which improved the model’s interpretability by showing the temporal effect of events [59]. Dipole exhibited the potential of bidirectional learning on EHR data using an RNN with concatenation based attention to predict diagnosis of diabetes [60] based on diagnosis codes and procedure codes.

Choi et al. enhanced their model's latent representation learning with a graph convolution transformer (GCT), but did not perform sequential learning, focusing only on a single patient encounter [140]. [138] also achieved improving their model's interpretability using self-attention, but only applied it on diagnosis and procedure codes. The BEHRT model [137] first realized a two-stage transfer learning approach with the BERT model [74], but only applied on diagnosis code with a low dimensional feature space (301). Meanwhile, it merely relied on diagnosis codes as the true label for every disease, which actually reduced the prediction sensitivity or specificity , due to inaccuracy and incompleteness in ICD codes and they are mostly for billing purposes [141]. MiME focused on learning the inner structure of an EHR by constructing a hierarchy of diagnosis level, visit level, and patient level embeddings [75]. The HCET model extended this hierarchical structure by removing the requirement of linked structure between diagnosis codes and procedure codes and medication while enabling attention on each EHR data modality to increase the model's interpretability [94].

Table 6.2: Results of pretraining with MLM.

| Data combination | All | No topic | No CPT | No topic+CPT |
|---|---|---|---|---|
| Vocabulary size | 12,460 | 12,360 | 5412 | 5312 |
| Precision | 0.4248 | 0.4324 | 0.4836 | 0.5086 |
| Learning rate | 1e-4 | 1e-4 | 1e-4 | 1e-4 |
| Embedding size | 216 | 240 | 252 | 264 |
| Attention layers | 9 | 9 | 6 | 6 |
| Attention heads | 12 | 12 | 12 | 12 |
| Intermediate layer | 512 | 512 | 256 | 256 |

The following models were used to compare the prediction performance to our BRLTM model: Dipole, MiME*, HCET, BERHT. I modified MiME to MiME* as the original MiME model requires external knowledge of linked relation between ICD-9 codes and associated CPT codes and medication lists during each visit, which was not applicable to use dataset.

## 6.4 Evaluation and Results

### 6.4.1 Pretraining on MLM

Table 6.2 presents the results for MLM including the optimal hyperparameter settings on various combination of EHR data modalities. According to the result, the precision score raises gradually from 0.4208 to 0.5086 as the vocabulary size decreases by excluding more data modalities. The optimal embedding size also follows this trend as 216 for all data and 264 for data without topics and CPT. The number of attention layers and the number of multi-head attention have the opposite trend, changing from 9 to 6 and 512 to 256, respectively.

### 6.4.2 Comparison of performance in depression prediction

Table 6.3 shows the ROCAUC and PRAUC from all baseline models and our BRLTM model at the four prediction time points. The BRLTM model achieved the highest performance in each prediction window with statistically significant improvements over the next best model (HCET). BEHRT generated slightly better results than MiME* in the two shortest time windows, but MiME* reached higher numbers in longer windows. This result follows those observed in HCET where ICD-9 possessed attention weights higher than the average in smaller prediction windows while it was lower than the average in larger windows. The prediction performance was slightly

Table 6.3: Comparison of prediction performance for different models.

| Metrics | Prediction window | MiME*<br>ICD-9+CPT+Med | BEHRT<br>ICD-9 | Dipole<br>ICD-9+CPT | HCET<br>All | **BRLTM**<br>All |
|---|---|---|---|---|---|---|
| ROCAUC | Two weeks | 0.76<br>(0.01) | 0.77<br>(0.02) | 0.78<br>(0.02) | 0.81<br>(0.02) | **0.85†**<br>**(0.02)** |
| | Three months | 0.74<br>(0.01) | 0.75<br>(0.02) | 0.76<br>(0.02) | 0.80<br>(0.01) | **0.84†**<br>**(0.01)** |
| | Six months | 0.72<br>(0.02) | 0.71<br>(0.01) | 0.75<br>(0.01) | 0.78<br>(0.03) | **0.83†**<br>**(0.01)** |
| | One year | 0.70<br>(0.01) | 0.69<br>(0.02) | 0.74<br>(0.01) | 0.75<br>(0.02) | **0.81†**<br>**(0.01)** |
| PRAUC | Two weeks | 0.67<br>(0.02) | 0.68<br>(0.01) | 0.70<br>(0.01) | 0.73<br>(0.02) | **0.78†**<br>**(0.01)** |
| | Three months | 0.64<br>(0.02) | 0.65<br>(0.01) | 0.67<br>(0.02) | 0.71<br>(0.02) | **0.76†**<br>**(0.02)** |
| | Six months | 0.61<br>(0.01) | 0.61<br>(0.02) | 0.65<br>(0.01) | 0.68<br>(0.02) | **0.74†**<br>**(0.02)** |
| | One year | 0.61<br>(0.01) | 0.60<br>(0.02) | 0.64<br>(0.01) | 0.66<br>(0.02) | **0.73†**<br>**(0.01)** |

Values in parenthesis refer to standard deviations across randomizations and bold values denotes the highest in each column. † indicates the value is significantly better than MiME* ($p<0.05$). The words after each model denotes the input data modalities where all means all five in our dataset.

improved with Dipole which used ICD-9 and CPT codes in a bidirectional learning method. Finally, there was a consistent decrease of accuracy for every model as the prediction window moved further away from the time of diagnosis.

### 6.4.3 Prediction performance for each primary diagnosis

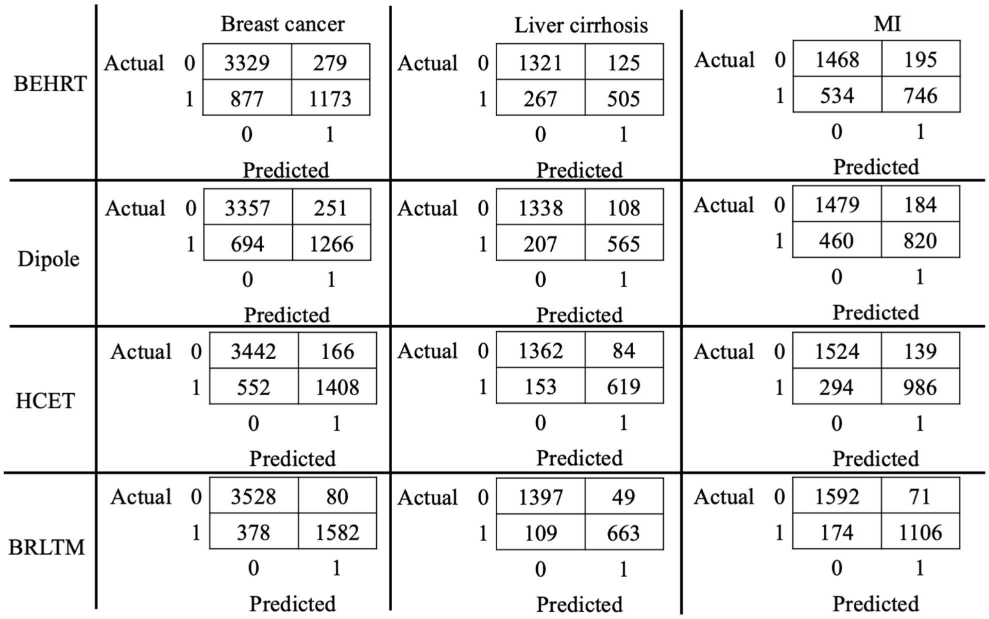
Table 6.4 displays the individual result for each of three primary diagnoses in prediction windows of two-weeks and one-year. Our BRLTM model also achieved the best performance for all three primary diagnoses in these two prediction windows. The ROCAUC from all three diseases

Table 6.4 : Comparison of prediction performance for three primary diagnoses.

| Prediction window | Disease | Metrics | MiME* | BEHRT | Dipole | HCET | **Ours** |
|---|---|---|---|---|---|---|---|
| Two weeks | Breast cancer | ROCAUC | 0.77<br>(0.01) | 0.78<br>(0.02) | 0.79<br>(0.02) | 0.81<br>(0.01) | **0.85**<br>**(0.01)** |
| | | PRAUC | 0.67<br>(0.02) | 0.68<br>(0.01) | 0.69<br>(0.02) | 0.73<br>(0.01) | **0.76**<br>**(0.01)** |
| | MI | ROCAUC | 0.75<br>(0.01) | 0.77<br>(0.01) | 0.78<br>(0.01) | 0.79<br>(0.02) | **0.85**<br>**(0.02)** |
| | | PRAUC | 0.70<br>(0.01) | 0.71<br>(0.02) | 0.71<br>(0.02) | 0.77<br>(0.01) | **0.78**<br>**(0.01)** |
| | Liver cirrhosis | ROCAUC | 0.76<br>(0.02) | 0.77<br>(0.01) | 0.78<br>(0.01) | 0.80<br>(0.01) | **0.84**<br>**(0.01)** |
| | | PRAUC | 0.67<br>(0.01) | 0.68<br>(0.02) | 0.69<br>(0.01) | 0.72<br>(0.01) | **0.75**<br>**(0.01)** |
| One year | Breast cancer | ROCAUC | 0.71<br>(0.02) | 0.70<br>(0.01) | 0.75<br>(0.01) | 0.78<br>(0.02) | **0.80**<br>**(0.01)** |
| | | PRAUC | 0.61<br>(0.01) | 0.59<br>(0.02) | 0.63<br>(0.02) | 0.67<br>(0.02) | **0.72**<br>**(0.01)** |
| | MI | ROCAUC | 0.69<br>(0.02) | 0.70<br>(0.01) | 0.74<br>(0.02) | 0.77<br>(0.01) | **0.81**<br>**(0.01)** |
| | | PRAUC | 0.64<br>(0.01) | 0.62<br>(0.01) | 0.66<br>(0.02) | 0.71<br>(0.01) | **0.74**<br>**(0.01)** |
| | Liver cirrhosis | ROCAUC | 0.70<br>(0.01) | 0.69<br>(0.02) | 0.74<br>(0.01) | 0.77<br>(0.02) | **0.80**<br>**(0.01)** |
| | | PRAUC | 0.61<br>(0.02) | 0.60<br>(0.02) | 0.63<br>(0.01) | 0.66<br>(0.02) | **0.71**<br>**(0.01)** |

within each model was similar even though the number of patients with breast cancer was substantially higher than other diseases (n=5,568), which indicated no bias toward any primary diagnosis for the task of predicting diagnosis of depression. It is notable that while the PRAUC for patients with myocardial infarction was relatively higher than other two, the difference in the BRLTM model was relatively small. Dipole generated a mean increase of around 0.02 both in ROCAUC and PRAUC across all diseases over BEHRT. In addition, HCET achieved better values than Dipole with higher improvement in PRAUC than ROCAUC. The BRLTM model further improved the performance from HCET with the highest increase of 0.06 in ROCAUC for MI in the window of two weeks and 0.05 in PRAUC for breast cancer and liver cirrhosis in the one-year window.

Figure 6.3 contains the confusion matrices individually for three primary diagnoses in the two-week prediction window from four models. The output probability was calibrated using the isotonic regression [142] with a threshold of 0.5, and numbers were aggregated from a 10-fold



| | | Breast cancer | | Liver cirrhosis | | MI | |
|---|---|---|---|---|---|---|---|
| | | Predicted | | Predicted | | Predicted | |
| | | 0 | 1 | 0 | 1 | 0 | 1 |
| BEHRT | Actual 0 | 3329 | 279 | 1321 | 125 | 1468 | 195 |
| | 1 | 877 | 1173 | 267 | 505 | 534 | 746 |
| Dipole | Actual 0 | 3357 | 251 | 1338 | 108 | 1479 | 184 |
| | 1 | 694 | 1266 | 207 | 565 | 460 | 820 |
| HCET | Actual 0 | 3442 | 166 | 1362 | 84 | 1524 | 139 |
| | 1 | 552 | 1408 | 153 | 619 | 294 | 986 |
| BRLTM | Actual 0 | 3528 | 80 | 1397 | 49 | 1592 | 71 |
| | 1 | 378 | 1582 | 109 | 663 | 174 | 1106 |

Figure 6.3: Confusion matrices for patients separated by three primary diagnosis at a window of two weeks for four models. The numbers are aggregated together with 10-fold cross validation. Label 0 means non-depressed while 1 means depressed.

cross validation. The class distribution was imbalanced with a smaller portion of depressed patients for each primary diagnosis. MiME* reached a higher portion of false positive than false negative and Dipole managed to reduce both numbers slightly. Our BRLTM model significantly decreased the false positives by almost 50% from HCET while reducing false negatives by roughly 40% for MI and 30% for breast cancer and liver cirrhosis. Hence, it achieved outstanding average precision and recall of 0.94 and 0.84, respectively, over the three primary diagnoses.

## 6.5 Discussion

This chapter presents a bidirectional deep learning model BRLTM to perform temporal representation learning on multimodal EHR data and successfully realized two-stage pretraining and finetuning. The results demonstrate the feasibility to apply the two-stage transfer learning approach on EHR modeling to overcome limitations in the amount of available data, which facilitates the development of clinical decision support systems for chronic disease prediction, such as a screening tool for patients at high risk depression, and thereby enabling early intervention.

According to results in Table 6.3 and 6.4, the BRLTM model sufficiently resolved the data heterogeneity issue by realizing bidirectional sequential learning and enabling the sturcture to aggregate multimoal EHR, which achieved the best performance in predicting future diagnoisis of depression in all four prediction windows. Additionally, the comparison to other models demonstrated the advantage of including more data modalities for the prediction task whereas BEHRT only took diagnosis codes in their study. In particular, MiME included two more data modalities than BEHRT. On the other hand, the better results from Diople over MiME* validates the advantage of bidirectional learning over single direction, as medication was less frequently present than ICD-9 and CPT, which Dipole did not take as the input. However, its lower performance than HCET, which adopted the forward-only sequnetial learning, highlights the importace to aggrete topics feature and demographics as Dipole only input ICD-9 and CPT codes, while HCET was capable of including all five modalities. Meanwhile, there is an observation in from these results that each model's performance consistently declines as the prediction window moves further away from the diagnosis time point, which agrees with our expectation that records closer to the diagnosis are more likely to contain relevant information and provide better predictions. Table 6.2 shows the observation that for a larger vocabulrory size or more data modalities a smaller embedding size should be used, but the number of attention layers and intermediate layer size should be increased, while no strong perference of the learning rate and number of attention heads. The results also approves model's flexible structure of several tunable hyperparameters, espeically in attetnion layers, enabling it to process various types of EHR data which may be collected from different insitutions.

More importantly, I sucessfully realized the common two-stage transfer learning apporach of pretraining and finetuing on modeling EHR data. Privacy issues related to EHR data restrict the ability of institutions to share data, which substantially hinders the development in this field. This two-stage transfer learning approach allows institutions with access to large amount of EHR data to be able to provide the pretrained model as a general EHR feature extractor so that others can take the advantage by only finetuning the pretrained model on the customized dataset for specific tasks [143,144]. This process benefits EHR representation learning and lays the foundation to facilitate adequate predictive power to models built on small EHR datasets. Furthermore, the BRLTM model also provides a generalized architecture that can be adopted with every EHR system by increasing the vocabulary of the code embedding or by stacking more embedding layers for additional data modalities not used in this work. More details about the future work of improving the precision of BRLTM will be presented in Chapter 8.

# CHAPTER 7

# Improving Interpretability of Machine Learning Models

## 7.1 Overview

Despite the unprecedented advancement and widespread of machine learning algorithms, they suffer from one main drawback of lacking adequate transparency or explanation, especially in the decision-making process, which highly impedes the development of their real-world applications [145]. In particular, the neural network based models act like a “black-box”, meaning no transparency to understand how they work in details [146].

In this chapter, I explore several techniques that are augmented to each model described in previous chapters individually. These efforts not only provide feasible approaches to increase transparency or interpretability of machine learning models for healthcare tasks, but also reveal insights on the tradeoff between improving model’s accuracy and interpretability [59]. The remainder of this chapter is organized as follows: Section 7.2 describes the dataset used in this study. Section 7.3 details every technique to improve models’ interpretability based on various dataset and tasks. Section 7.4 summaries the results while Section 7.5 discusses several observations and limitations. The content of this chapter have partly been published in [94,135].

## 7.2 Data Cohort

All of the dataset used in this chapter to test the effect of model’s interpretability are the same as previous chapters. Specifically, the cohort of 182 patients used in Chapter 3 for building an HMM is adopted here to study the importance factor of each of 14 features collected by Fitbit. As a similar approach, patient's EHR data in Chapter 4 is used here to investigate the importance factor of each of five modalities as well. In addition, the multimodal EHR dataset from Chapter 5 and Chapter 6 are adopted here to reveal the feature importance or contribution to the prediction task from each data modality or every EHR code.

## 7.3 Methods

### 7.3.1 Feature importance factor from RF

As described in Chapter 2.3.4, RF has another advantage of better interpretation in the decision process than other classical machine learning models. It can output the relative feature importance factor by information gain (IG). Therefore, I first utilized this function in RF to study the relative feature importance factor of 14 types of vital signs collected by Fitbit then normalized and compared for the classification task of each PRO in Chapter 3. Furthermore, this technique was also applied to reveal the feature importance factor of every element of various EHR data modalities to predict future diagnosis of depression in Chapter 4.

### 7.3.2 Attention for every EHR data modality in the code level embedding

In order to investigate the importance factor of every data modality in this prediction task as well as improve the interpretability of HCET, attention weights $\lambda_j$ are defined for each modality, where the sum of all weights equals to one, as shown in Eq. (7.3.1). A weighted sum of code level embedding $F'$ is input into HCET, indicates by Eq. (7.3.2), which substitutes $F$ in Equations (3.3.4) and (3.3.5). $d, c, m, p, x$ standard for diagnosis codes, procedure codes, medication, demographics and topics feature, respectively. After training, attention weights are able to reveal the importance factor of each feature type in the prediction task,

$$\sum \lambda_j = 1 \tag{7.3.1}$$

$$F' = \lambda_D W_D \vec{d} + \lambda_C W_C \vec{c} + \lambda_M W_M \vec{m} + \lambda_P W_P \vec{p} + \lambda_X W_X \vec{x} \tag{7.3.2}$$

### 7.3.3 Self-attention for every code in EHR sequences

Revealing the attention weight as the importance factor of each of five EHR data modalities still do not tell how each element in various patient encounter contributes the prediction task. Therefore, I followed the transformer architecture described first in Chapter 2.5.2, where the combination of self-attention and multi-head attention effectively presents a quantitative analysis of every element in a sequence. This technique was applied to the model developed in Chapter 6 to improve its interpretability.

## 7.4 Evaluation and Results

### 7.4.1 Relative importance factor for features collected by Fitbit

The interpretability of the model built in Chapter 3 was enhanced by outputting the importance factor of each feature contributing to the classification in the RF model. Table 7.1 displays the importance factor for the 14 feature types summed over seven days. Features that were significantly higher ($p<0.05$) than the average value for each classification were determined. Steps, total distance, calories, and calories BMR contributed to classify most of the PRO scores. The importance factor of light active distance was significantly better than other features for classifying global physical health and physical function, which were both related to a subject's physical health. On the other hand, resting heart rate contributed significantly more than other features for classification of mental health PROs such as anxiety and depression, while its importance factor was not significantly higher than other features in classifying PROs related to physical health.

Table 7.1: Importance factor of each feature for classifying various health status.

| Type | Global physical health | Global mental health | Fatigue | Physical Function | Anxiety | Depression | Sleep Disturbance |
|---|---|---|---|---|---|---|---|
| Step | **0.138 (0.010)** | **0.088 (0.006)** | **0.111 (0.009)** | **0.138 (0.007)** | 0.075 (0.005) | **0.080 (0.006)** | **0.104 (0.005)** |
| Total Distance | **0.122 (0.009)** | **0.081 (0.003)** | **0.100 (0.009)** | **0.123 (0.011)** | **0.079 (0.004)** | 0.075 (0.004) | **0.088 (0.005)** |
| Very Active Distance | 0.062 (0.005) | 0.066 (0.006) | 0.061 (0.004) | 0.068 (0.005) | 0.065 (0.003) | 0.056 (0.002) | 0.053 (0.004) |
| Moderately Active Distance | 0.059 (0.007) | 0.058 (0.002) | 0.052 (0.003) | 0.051 (0.003) | 0.052 (0.002) | 0.054 (0.002) | 0.0573 (0.003) |
| Light Active Distance | **0.088 (0.010)** | 0.071 (0.002) | 0.074 (0.001) | **0.086 (0.010)** | 0.070 (0.005) | 0.066 (0.003) | 0.069 (0.004) |
| Very Active Minutes | 0.054 (0.007) | 0.060 (0.006) | 0.052 (0.004) | 0.060 (0.006) | 0.057 (0.004) | 0.053 (0.002) | 0.055 (0.004) |
| Fairly Active Minutes | 0.055 (0.009) | 0.054 (0.002) | 0.050 (0.007) | 0.050 (0.006) | 0.0530 (0.003) | 0.051 (0.003) | 0.063 (0.004) |
| Light Active Minutes | 0.068 (0.006) | 0.072 (0.005) | 0.064 (0.003) | 0.061 (0.005) | 0.066 (0.002) | 0.073 (0.002) | 0.067 (0.004) |
| Sedentary Minutes | 0.046 (0.003) | 0.066 (0.005) | 0.055 (0.002) | 0.043 (0.001) | 0.065 (0.007) | 0.072 (0.007) | 0.060 (0.005) |
| Calories | 0.059 (0.006) | **0.088 (0.008)** | 0.074 (0.007) | 0.053 (0.005) | **0.098 (0.007)** | **0.094 (0.004)** | **0.082 (0.003)** |
| Floors | 0.055 (0.008) | 0.049 (0.001) | 0.056 (0.010) | 0.076 (0.019) | 0.048 (0.003) | 0.053 (0.003) | 0.051 (0.005) |
| Calories BMR | 0.073 (0.004) | **0.110 (0.016)** | **0.105 (0.007)** | 0.071 (0.009) | **0.117 (0.006)** | **0.128 (0.009)** | **0.105 (0.012)** |
| Marginal Calories | 0.066 (0.004) | 0.071 (0.003) | 0.060 (0.003) | 0.058 (0.005) | 0.070 (0.006) | 0.069 (0.003) | 0.077 (0.006) |
| Resting Heart Rate | 0.058 (0.012) | 0.067 (0.008) | 0.085 (0.012) | 0.063 (0.008) | **0.085 (0.003)** | **0.076 (0.002)** | 0.069 (0.016) |

Value in parentheses is the standard deviation. Bold values are significantly higher ($p<0.05$) than the average value for a feature (1/14 = 0.0714).

The analysis of classifying PRO was repeated using the RF classifier and only the significant features from Table 7.1 and were shown in Table 7.2. Because some studies such as [147] only used steps data to assess user's health status, we also compared the model's performance in the same manner. The results suggested that the RF model can generate significantly better classification accuracy with the selected features than all features from Fitbit for all PROMIS short form survey scores except for global mental health (p=0.37), with the highest AUC of 0.76 for classification of physical function.

Table 7.2 Mean and standard deviation ROCAUC of different Feature selection strategy.

| Type | Steps Only | All Feature | Selected Feature |
|---|---|---|---|
| Global physical health | 0.73 (0.03) | 0.73 (0.01) | **0.73 (0.02)** |
| Global mental health | 0.52 (0.02) | 0.55 (0.03)† | **0.58 (0.02)*** |
| Fatigue | 0.60 (0.05) | 0.61 (0.03) | **0.64 (0.03)*** |
| Physical function | 0.76 (0.03) | 0.75 (0.01) | **0.76 (0.01)*** |
| Anxiety | 0.50 (0.04) | 0.54 (0.02)† | **0.57 (0.02)*** |
| Depression | 0.51 (0.02) | 0.53 (0.02)† | **0.56 (0.02)*** |
| Sleep Disturbance | 0.59 (0.03) | 0.61 (0.03) | **0.64 (0.03)*** |

* Significant improvement from Selected Feature over All Feature. † Significant improvement from All Feature over Steps Only. Bold values are the highest for a given PRO.

### 7.4.2 Importance factor for individual EHR feature in predicting depression

Table 7.3 and 7.4 display the top 20 features ranked by their importance factors when using prediction windows of two weeks and one year, respectively. The "number of visits" feature was defined for each data modality that represents how many days in a patient's record that data type appears (e.g., if a patient has two reports on one day and one report on a different day, their

Table 7.3: Top 20 significant features in the prediction window of two weeks in advance.

| | |
|---|---|
| 1. Number of visits for ICD-9 | 11. CPT: Under Diagnostic/Screening Processes or Results |
| 2. Number of visits for CPT | 12. ICD-9: Symptoms involving digestive system |
| 3. Age | 13. ICD-9: Symptoms involving respiratory system and other chest symptoms |
| 4. ICD-9: Anxiety, dissociative and somatoform disorders | 14. **Topic: denies, family, use, alcohol, social** |
| 5. ICD-9 780: General symptoms | 15. CPT: Under Diagnostic/Screening Processes or Results |
| 6. Number of visits for medication | 16. CPT: Under Organ or Disease Oriented Panels |
| 7. Number of visits for reports | **17. Topic: normal, clear, bilaterally, extremities, soft** |
| **8. Topic: sleep, feels, week, visit, days** | **18. Topic: blood, year-old, rate, post, status** |
| **9. Topic: given, discussed, treatment, risk, prior** | 19. ICD-9: Other disorders of soft tissues |
| **10. Topic: continue, stable, daily, bid, prn** | 20. CPT: Under Established Patient Office or Other Outpatient Services |

Topic features are shown in bold and the feature of number of visits are underlined.

“number of visits for reports” feature would be two). For the two-week window, the “number of visits” features, were ranked 1, 2, 6, and 7. The high ranking of these features demonstrated the contribution of temporal frequency in predicting depression. When predicting two weeks before the diagnosis, there were six topic features in the top 20, and many of which contained words related to temporality, such as “week,” “continue,” and “stable.” When predicting with the one-year window, “number of visits” features for medication and reports were no longer in the list of top 20 features.

Table 7.4 Top 20 significant features in the prediction window of one year in advance.

1. Number of visits for ICD-9
2. Number of visits for CPT
3. Age
4. ICD-9: General symptoms
5. ICD-9: Anxiety, dissociative and somatoform disorders
6. CPT: Under Established Patient Office or Other Outpatient Services
7. ICD-9: Symptoms involving respiratory system and other chest symptoms
8. ICD-9: Symptoms involving digestive system
9. ICD-9: Other and unspecified disorders of back
10. ICD-9: Other and unspecified disorders of joint
11. ICD-9: Other disorders of soft tissues
12. CPT: Under Diagnostic/Screening Processes or Results
13. CPT: Under Diagnostic/Screening Processes or Results
14. CPT: Under Echocardiography Procedures
15. CPT: Under New Patient Office or Other Outpatient Services
16. ICD-9: Intervertebral disc disorders
17. ICD-9: Symptoms involving skin and other integumentary tissue
18. ICD-9: Nonspecific abnormal results of function studies
19. ICD-9: Nonspecific (abnormal) findings on radiological and other examination of body structure
20. CPT: Under Established Patient Office or Other Outpatient Services

Topic features are shown in bold and the feature of number of visits are underlined.

### 7.4.3 Attention weights for every EHR data modality in HCET

As mentioned before, one advantage of RF over the majority of deep learning models is the ability to provide information on the importance factor of each feature contributing to classification [135]. However, there is a consistent effort to improve the interpretability of deep learning models like HCET. Figure 7.1 shows the attention weights for each of EHR data modalities over four prediction windows. According the result, medication and demo both are below the average value

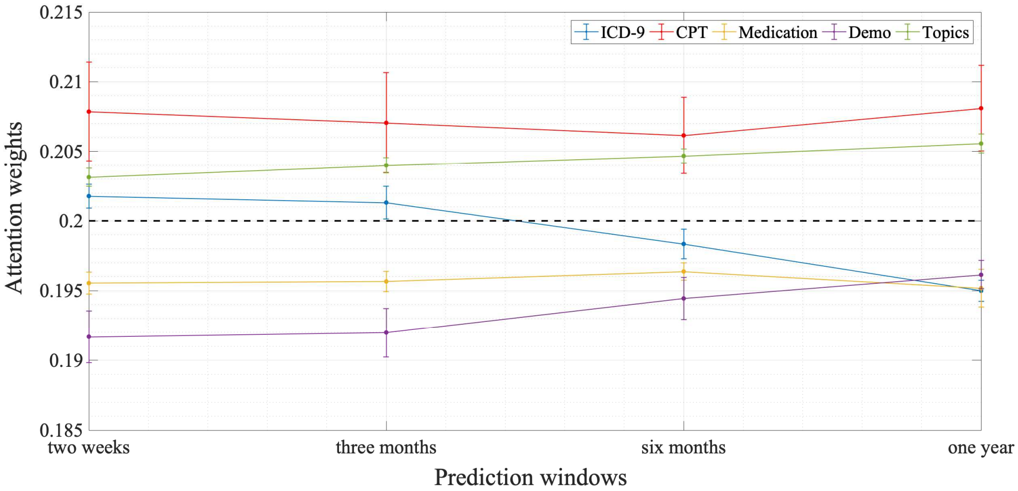


Figure 7.1: Attention weights from every EHR data modalities in four prediction windows. Error bars denotes the standard deviation. The black dash line is at threshold of 1/5, which indicates constant weights in HCET models before.

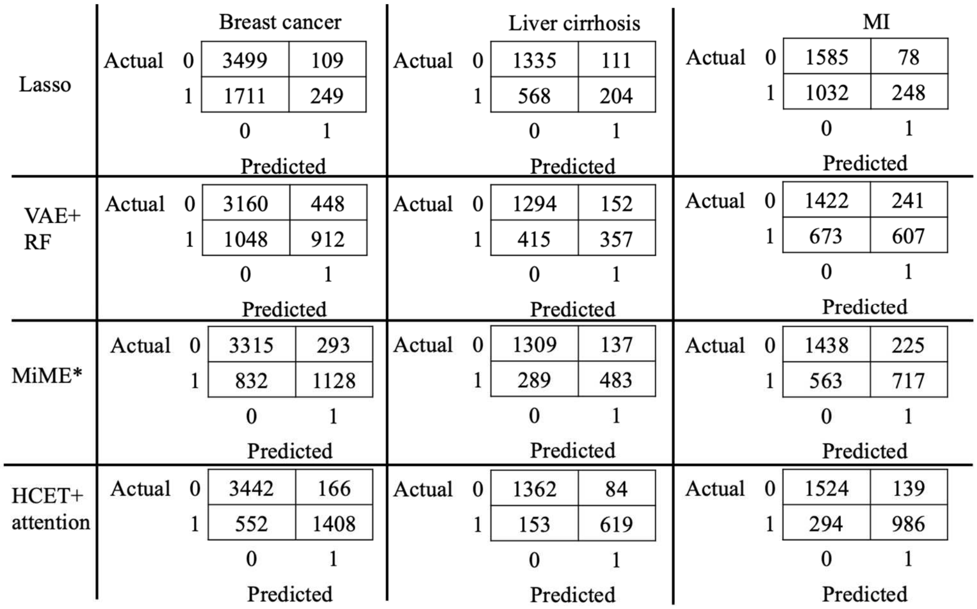
of 0.2 in four prediction windows. Attention for ICD-9 is above 0.2 in window of two weeks and three months, but it drops in six months and one year. There is a consistent increase of attention for topics while the attention from CPT always ranks top.

In the meantime, Table 7.5 reveals the effect of adding attention weights to each data modality at the code level embedding to the prediction task of HCET model. According to the result, the

Table 7.5: Comparison of prediction performance for HCET models.

| Prediction window | Two weeks | | Three months | | Six months | | One year | |
|---|---|---|---|---|---|---|---|---|
| Models | ROCAUC | PRAUC | ROCAUC | PRAUC | ROCAUC | PRAUC | ROC AUC | PRAUC |
| HCET | **0.81**<br>**(0.01)** | 0.73<br>(0.02) | 0.80<br>(0.02) | **0.71**<br>**(0.02)** | 0.78<br>(0.01) | 0.68<br>(0.02) | 0.75<br>(0.01) | 0.66<br>(0.02) |
| HCET + attention | 0.81<br>(0.01) | **0.73**<br>**(0.01)** | **0.80**<br>**(0.01)** | 0.70<br>(0.02) | **0.79***<br>**(0.01)** | **0.69**<br>**(0.01)** | **0.78***<br>**(0.01)** | **0.67**<br>**(0.01)** |

Values in parentheses refer to standard deviations across randomizations and bold values denotes the highest in each column. * denotes the value is significantly better than no attention ($p<0.05$).



| | | Breast cancer | | Liver cirrhosis | | MI | |
|---|---|---|---|---|---|---|---|
| | | Predicted 0 | Predicted 1 | Predicted 0 | Predicted 1 | Predicted 0 | Predicted 1 |
| Lasso | Actual 0 | 3499 | 109 | 1335 | 111 | 1585 | 78 |
| | Actual 1 | 1711 | 249 | 568 | 204 | 1032 | 248 |
| VAE+ RF | Actual 0 | 3160 | 448 | 1294 | 152 | 1422 | 241 |
| | Actual 1 | 1048 | 912 | 415 | 357 | 673 | 607 |
| MiME* | Actual 0 | 3315 | 293 | 1309 | 137 | 1438 | 225 |
| | Actual 1 | 832 | 1128 | 289 | 483 | 563 | 717 |
| HCET+ attention | Actual 0 | 3442 | 166 | 1362 | 84 | 1524 | 139 |
| | Actual 1 | 552 | 1408 | 153 | 619 | 294 | 986 |

Figure 7.2: Confusion matrix for patients separated by three primary diagnosis at a window of two weeks for four models. The numbers are aggregated together with 10-fold cross validation. Label 0 means non-depressed while 1 means depressed.

ROCAUC at six months and one year are significantly improved with p=0.04 and p=3e-5, respectively. Furthermore, Figure 7.2 contains the prediction results in confusion matrices with patients separated in three primary diagnoses in the prediction window of two weeks from four models at the same threshold of 0.5, after probability calibration using the isotonic regression [142]. The numbers were aggregated from 10-fold cross validation. For each primary diagnosis, the distribution was imbalanced due to a lower number of depressed patients. Lasso generated poor accuracy as it almost always predicted the negative class. VAE+RF slighted reduced false negative cases but the number of true negatives was worse than Lasso. MiME* both improved the numbers in true positives and true negatives while HCET with attention improved it further. The average precision and recall over three primary diagnoses from HCET with attention were 0.88 and 0.76, respectively.

### 7.4.4 Self-attention for every code in EHR sequence Fitbit

Figure 7.3 (a) and (b) exhibits the self-attention weights from two patients' EHR sequences, retrieved from the attention component of the last layer of the BRLTM model. It was illustrated in Figure 7.3 (a) that this patient was diagnosed with malignant neoplasm of the liver. The self-attention weight indicated its highest association with the topic feature associated with the words "transplant, tacrolimus, liver, renal, and daily" and second highest relation to the ICD-9 code for "Organ or tissue replaced by transplant." The topic feature with the words "liver, hepatitis, pain, hcc, and abdominal" described the fact that the patient was undergoing a liver transplant after the original diagnosis. Figure 7.3 (b) displays another patient was initially diagnosed with an unspecified joint disorder which led to a topic feature of "pain, knee, hip, fracture and shoulder" shown later in the EHR sequence. The darker color suggests the stronger association of this topic feature to the original diagnosis code (diagnostic radiology imaging) and a weaker latent relation

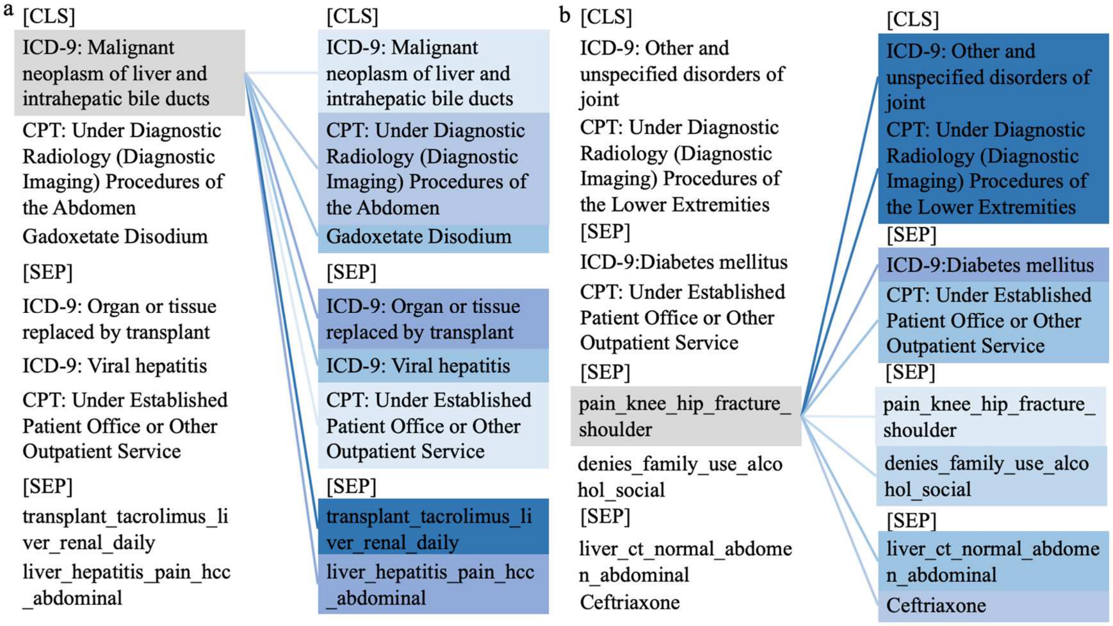

a
[CLS]
ICD-9: Malignant neoplasm of liver and intrahepatic bile ducts
CPT: Under Diagnostic Radiology (Diagnostic Imaging) Procedures of the Abdomen
Gadoxetate Disodium
[SEP]
ICD-9: Organ or tissue replaced by transplant
ICD-9: Viral hepatitis
CPT: Under Established Patient Office or Other Outpatient Service
[SEP]
transplant_tacrolimus_liver_renal_daily
liver_hepatitis_pain_hcc_abdominal
b
[CLS]
ICD-9: Other and unspecified disorders of joint
CPT: Under Diagnostic Radiology (Diagnostic Imaging) Procedures of the Lower Extremities
[SEP]
ICD-9:Diabetes mellitus
CPT: Under Established Patient Office or Other Outpatient Service
[SEP]
pain_knee_hip_fracture_shoulder
denies_family_use_alcohol_social
[SEP]
liver_ct_normal_abdomen_abdominal
Ceftriaxone

Figure 7.3: Quantitative analysis of self-attention from two patients' EHR sequences shown in color plots. CLS and SEP represent the beginning of the record and separators between visits, respectively. Topic features are represented as the five most commonly associated words. Each example is presented as two identical columns as the left one represents the code of interest colored in grey while the right one indicates the corresponding associations to the highlighted code on the left. The intensity of the blue color on the right column denotes the strength of the attention score; the deeper blue color suggests higher self-attention score and hence the stronger the latent association.

to the diagnosis of diabetes and the medication ceftriaxone. The attention scores demonstrated the association of the patient's health status with an original diagnosis of joint disorder and a comorbidity of diabetes developed later. This matches the meta-analysis that arthritic patients have 61% higher odds of having diabetes compared to the population without arthritis [148].

## 7.5 Discussion

In this Chapter, various techniques to improve the interpretability of previous constructed machine learning models by revealing quantitative analysis of features' latent association have been presented. According to Table 7.1, steps and total distance have significantly higher importance for classifying the majority of survey scores, while calories BMR significantly contributes to mental health scores, like anxiety and depression. Their importance factor may due to data quality, as  previous studies [7,9,89] have validated the data accuracy for step counts, distance traveled, and energy expenditure for activity trackers, while other features have not been validated in scientific works. Since Fitbits are not sold as medical devices, many of their features are not validated or regulated like other medical devices.

Feature importance factors listed in Table 7.3 and 7.4 revealed individual contributions to the prediction task and demonstrated the important contribution of topic features as they occupied six out of the top 20 features for predicting at a two-week time horizon. In contrast, no topic features were included in the one-year prediction window, which further indicated the temporal importance of topic features. In fact, the dimension of topics was set to 100, which is relatively small compared to the dimensionality of ICD-9 codes, CPT codes and medications. Nevertheless, including them boosted the prediction performance significantly. This result suggested that aggregating more EHR data sources especially like clinical notes could significantly increase model's prediction performance.

Adding attention weights effectively improved the interpretability of HCET model. In particular, attention weights were applied to each data modality, which revealed the relative importance of each data modality and the trajectory in four prediction windows. In addition, the attention weights of topics were consistently above the average value, demonstrating their important contribution in the prediction task. Finally, the self-attention from BRLTM models was able to provide quantitative analysis on latent association of every code in EHR sequences, which highly improves the transparency to trace how the model processes the dependency between them.

# CHAPTER 8

# Conclusion

## 8.1 Overview

This chapter summarizes the results and contributions of this dissertation. It also suggests the limitation and future work of potential research directions to continue the work conducted in this thesis. Finally, this chapter gives concluding remarks about building classifiers and the potential of prediction models for applications in digital health.

## 8.2 Summary and Results

This dissertation presents methods that improve modeling temporal datasets in two aspects of digital health: vital signs collected by activity trackers (Fitbit) and longitudinal EHR data. These models demonstrate the feasibility of creating classifiers to monitor a patient's physical and mental health status and predict diagnosis of depression using EHR, which could potentially be used in a clinical decision support system. Using these methods, researchers and clinicians can obtain more insight through modeling temporal patient digital health data and aggregating the various modalities available in the EHR. However, machine learning models are highly dependent on the quantity and quality of the training data, especially in the supervised learning paradigm. Therefore, addressing the need for large amounts of EHR data and the ability to integrate and standardize them between hospitals is critical to constructing accurate models. The availability of such datasets would be transformative, similar to how an image dataset such as ImageNet changed the landscape of computer vision. Since collecting and labeling patient's EHR takes a significant amount of time and effort and contains personally identifiable information, collaboration between several institutions is to enable data sharing (or alternatively, facilitate federated learning) could alleviate the bias on a small dataset by aggregating more patient cohorts from various locations, which eventually facilitate the development of modeling EHR and deployment of clinical decision support systems. The specific contributions of this dissertation are as follows:

- *A temporal machine learning model to process human vital signs collected by activity trackers and provide information about a patient's health status.* This model used an HMM to classify self-reported health status from activity tracker data.
- *A predictive model based on machine learning algorithms capable of aggregating EHR data across modalities to predict depression*. This study consisted of two approaches to aggregate multimodal EHR data to predict diagnosis of depression. The multi-hot encoding method was able to combine data into one vector. Furthermore, the novel hierarchical embedding model overcame the challenge of data heterogeneity and sparsity in the data structure of EHR.
- *A bidirectional and two-stage transfer learning approach to model multimodal EHR.* The BRLTM model is a novel adaptation of the BERT model, which realized the two-stage pretraining and finetuning approach on modeling EHR data. This framework facilitates the process of EHR modeling by efficiently leveraging the limited data available through the EHR.
- *Multiple techniques to improve the interpretability of machine learning models*. Several techniques were introduced to improve the interpretability of machine learning models built in the previous aims of this project, which enhanced the transparency and trust of models on the decision-making process by revealing latent feature associations.

There are relatively few previous studies that evaluate the clinical impact of activity tracker data, such as building a surveillance system for a user's health status. In Chapter 3 of this dissertation, I first developed a novel application of an HMM, capable of classifying various types of patients' self-reported health status from vital signs collected by Fitbit. The results demonstrate the superior performance of temporal machine learning models like HMM over non-temporal ones in processing temporal data such as activity tracker data and weekly PRO scores. Although this was a retrospective analysis, the results verified our assumption that human vital signs collected by activity trackers were accurate enough to reflect user's health status. We demonstrated that the sample size collected as part of the study was adequate to train machine learning models to classify patient's PRO. Hence, this approach supports the feasibility of constructing a surveillance system to monitor a user's health status in real-time. This model can be tested on data from other ongoing studies, which use activity trackers data to monitor or classify the surrogate of user's health status in real time.

Chapter 4 introduced a method to predict depression diagnosis with five heterogenous input data modalities. The study investigated the effect of length of a patient's EHR on the prediction task. Model performance reported from four prediction windows demonstrated that records closer to the diagnosis are more likely to contain relevant information.

Chapter 5 described HCET, an improved model that utilizes the hierarchical embedding architecture to alleviate the challenge of data heterogeneity and sparsity in EHR data. The results support the conclusion of Chapter 3 that incorporating temporal information into a model achieves better prediction accuracy than static models. More importantly, this model provides a more generalized architecture to aggregate EHR data modalities than the previous state-of-the-art, resulting in superior performance in predicting depression.

Chapter 6 describes the feasibility of applying a two-stage approach to model data from the EHR. The model first pretrained on a larger dataset and then finetuned for a specific task while performing bidirectional representation learning on multimodal EHR, which exhibited superior performance to previous models. These results demonstrated that the success of bidirectional sequence learning and the two-stage transfer learning can also facilitate modeling EHR sequences to build clinical decision support systems while alleviating the requirement of a large amount of training data for individual developers. The work described in Chapters 4-6 offers several approaches to address the heterogeneity and sparsity of EHR data for predicting depression diagnosis. These results tested our assumption that patient's historical EHR contains the relation to the future health status of depression, and the data we collected was sufficient to training a machine learning model to predict the future diagnosis. Meanwhile, the structure of EHR sequences is similar to word sequences so that NLP models are able to generalize the success of representation learning on languages to EHR.

Finally, Chapter 7 details several techniques to strengthen the interpretability and transparency of the machine learning models described in previous chapters, particularly for providing explanations of the decision-making process. The first approach enhances the transparency of the HMM model in Chapter 3 by providing the relative importance of each human signals, which provides guidance on the feature selection. The second work in Chapter 7 provides feature importance estimations for the five EHR modalities used in the prediction model built in Chapter 4. The third part in Chapter 7 improves the interpretability of the model built in Chapter 5 by showing the attention weight of each input data modality. The final work in Chapter 7 further enhances the transparency of the model in Chapter 6 by conducting a quantitative analysis of the latent association between every code in EHR sequences using self-attention and multi-head attention.

## 8.3 Future Work

There are several limitations of this work and ways to improve the existing methodologies. A sample of such improvements is briefly discussed in the following sections.

### 8.3.1 Improving precision for the health status classification system

The observation in Chapter 3 indicates the higher correlation between data collected by activity trackers with subjects' physical health than mental health. Thus, it might be useful to develop hardware to record data more related to the mental health for future studies. For instance, there has been effort to develop non-invasive and continuous blood pressure tracking [149] using wearable devices, which may improve classifying mental health [150]. Also, anxiety and depression were only measured by PROMIS instruments in this study, which lacks precision as mental health is a broad and complicated field. More thorough evaluations of subjects' mental states could provide more descriptive labels for training machine learning models, which could further improve the performance in predicting mental health status.

In addition, another future direction would be to approach this as a regression problem to predict actual PRO scores over time. Furthermore, sequential deep learning models, such as recurrent neural networks (RNNs) and long-short-term-memory (LSTM) networks have also demonstrated strong performance when dealing with sequential data [124,151]. Therefore, these techniques may hold potential for applications to sensor data to classify or predict health status. However, such methods generally require a large amount of training data, which was not available in the current study. In future studies, deep learning methods could be explored if a sufficiently large data set were collected.

### 8.3.2 Increasing the label accuracy for the diagnosis of diseases in the prediction task

As mentioned in Chapters 4-6, the clinical standard for depression diagnosis is the PHQ-9 questionnaire, which is not routinely collected clinically. Instead, the time of depression diagnosis was determined by either of three criteria: ICD-9 codes, prescription of antidepressant, or mention of antidepressants in a clinical report. This method likely added false positives as some antidepressants can be prescribed to treat other diseases. We viewed this method as a conservative baseline. Actually, in clinics, patients would go through mental state examination (MSE) to evaluate their mental health [152]. Thus, future studies could choose a more accurate way to evaluate depression diagnosis, such as collecting more data from the MSE or implementing the PHQ-9 questionnaire [128] at different time points to define the time of depression onset definitively. More robust predictive models could then be constructed to track the disease progression and enable early detection.

### 8.3.3 Enhancing the contextual representation of clinical notes

The results presented both in Chapters 4 and 5 demonstrate the contribution of topic features in temporal models for predicting depression. Future work could include a larger number of clinical notes in building models on EHR. LDA, a topic modeling method, was adopted to process the clinical notes in the study. It is based on the bag of words assumption, which may not be the ideal way to represent clinical text. Future studies could utilize more sophisticated NLP tools, such as BERT [74] or GPT-2 [136] to optimize the contextual representation of clinical notes, which could further improve the overall performance.

### 8.3.4 Extending the model to predict other diseases and aggregate more EHR modalities

The predictive power of models built in this dissertation was employed and limited to predict depression diagnosis, a binary classification task. Future studies could expand the models to perform multiclass prediction simultaneously for other highly prevalent chronic diseases such as hypertension, diabetes, or obesity. Finally, other EHR data modalities such as laboratory results [139] were not included in these models as they were unavailable for collection. Thus, future studies may extend the model to aggregate more data modalities to further utilize the heterogeneity of EHR data.

## 8.4 Concluding Remarks

In digital health, the feasibility of exhibiting clinical impact on users' health status using activity trackers has not been fully explored. In the meantime, current screening tools used in clinics for depression only produced a true positive rate of 50%. These challenges highlight the need for assisting clinicians with better clinical decision support systems to monitor a patient's health status or predict future diagnoses. The contributions of this dissertation provide the foundation towards modeling datasets for clinical decision support systems in digital health. These contributions include developing a temporal machine learning approach to classify various types of self-reported health status using activity tracker data, which validates the feasibility of building a real-time surveillance system to monitor users' health status. Furthermore, this dissertation presents a variety of methods to process multimodal EHR data to predict future diagnoses of depression. These models provide a potential application of building clinical decision support systems to assist clinicians for depression screening based on EHR data. This dissertation successfully applies the two-stage transfer learning and bidirectional sequence learning approaches to EHR modeling. The model pretrains on a large standard dataset and then finetunes for specific tasks, effectively resolving the limitation of data scarcity in modeling multimodal EHR. Finally, this work's contributions include numerous techniques to improve the interpretability and transparency of machine learning models by revealing the importance factor for each data modality and feature. This effort could provide decision-making guidance for clinicians in health status classification or disease prediction and ultimately shorten the time to deploy machine learning models in clinics.

# REFERENCES

1 Fihn SD, Gardin JM, Abrams J, *et al.* 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart diseass. *Circulation* 2012;**126**:354–471. doi:10.1161/CIR.0b013e318277d6a0

2 Lloyd-Jones D, Adams RJ, Brown TM, *et al.* Executive summary: Heart disease and stroke statistics-2010 update: A report from the american heart association. *Circulation* 2010;**121**:46–215. doi:10.1161/CIRCULATIONAHA.109.192667

3 Ford ES, Capewell S. Coronary Heart Disease Mortality Among Young Adults in the U.S. From 1980 Through 2002. Concealed Leveling of Mortality Rates. *J Am Coll Cardiol* 2007;**50**:2128–32. doi:10.1016/j.jacc.2007.05.056

4 Schappert SM, Burt CW. Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 2001-02. *Vital Health Stat 13* 2006;:1—66.http://europepmc.org/abstract/MED/16471269

5 Javitz HS, Ward MM, Watson JB, *et al.* Cost of illness of chronic angina. *Am J Manag Care* 2004;**10**:S358—69.http://europepmc.org/abstract/MED/15603245

6 Bakker JP, Goldsack JC, Clarke M, *et al.* OPEN A systematic review of feasibility studies promoting the use of mobile technologies in clinical research. *npj Digit Med* Published Online First: 2019. doi:10.1038/s41746-019-0125-x

7 Meyer J, Hein A. Live long and prosper: Potentials of low-cost consumer devices for the prevention of cardiovascular diseases. *J Med Internet Res* 2013;**15**:1–9. doi:10.2196/med20.2667

8 Alharbi M, Straiton N, Gallagher R. Harnessing the Potential of Wearable Activity Trackers for Heart Failure Self-Care. *Curr Heart Fail Rep* 2017;**14**:23–9. doi:10.1007/s11897-017-

0318-z

9 Ferguson T, Rowlands A V., Olds T, *et al.* The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study. *Int J Behav Nutr Phys Act* 2015;**12**:1–9. doi:10.1186/s12966-015-0201-9

10 Shuger SL, Barry VW, Sui X, *et al.* Electronic feedback in a diet- and physical activity-based lifestyle intervention for weight loss: a randomized controlled trial. *Int J Behav Nutr Phys Act* 2011;**8**:41. doi:10.1186/1479-5868-8-41

11 Zan S, Agboola S, Moore SA, *et al.* Patient engagement with a mobile web-based telemonitoring system for heart failure self-management: a pilot study. *JMIR mHealth uHealth* 2015;**3**:e33. doi:10.2196/mhealth.3789

12 Akincigil A, Matthews EB. National rates and patterns of depression screening in primary care: Results from 2012 and 2013. *Psychiatr Serv* 2017;**68**:660–6. doi:10.1176/appi.ps.201600096

13 Horwath E, Johnson J, Klerman GL, *et al.* Depressive Symptoms as Relative and Attributable Risk Factors for First-Onset Major Depression. *Arch Gen Psychiatry* 1992;**49**:817. doi:10.1001/archpsyc.1992.01820100061011

14 Greenberg PE, Fournier A-A, Sisitsky T, *et al.* The Economic Burden of Adults With Major Depressive Disorder in the United States (2005 and 2010). *J Clin Psychiatry* 2015;**76**:155–62. doi:10.4088/JCP.14m09298

15 John Mann J. The medical management of depression. *N Engl J Med* 2005;**353**:1819–34. doi:10.1056/NEJMra050730

16 Mitchell AJ, Vaze A, Rao S, *et al.* Clinical diagnosis of depression in primary care : a meta-analysis. *Lancet* 2009;**374**:609–19. doi:10.1016/S0140-6736(09)60879-5

17 Baas KD, Wittkampf KA, Van Weert HC, *et al.* Screening for depression in high-risk groups: Prospective cohort study in general practice. *Br J Psychiatry* 2009;**194**:399–403. doi:10.1192/bjp.bp.107.046052

18 Shickel B, Tighe PJ, Bihorac A, *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Heal Informatics* 2018;**22**:1589–604. doi:10.1109/JBHI.2017.2767063

19 Xiao C, Choi E, Sun J. Review Opportunities and challenges in developing deep learning models using electronic health records data : a systematic review. 2018;**25**:1419–28. doi:10.1093/jamia/ocy068

20 Triñanes Y, Atienza G, Louro-González A, *et al.* Development and impact of computerised decision support systems for clinical management of depression: A systematic review. *Rev Psiquiatr y Salud Ment (English Ed* 2015;**8**:157–66. doi:https://doi.org/10.1016/j.rpsmen.2015.05.004

21 Lipton ZC, Kale DC, Elkan C, *et al.* Learning to diagnose with LSTM recurrent neural networks. In: *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. 2016. 1–18.

22 Lang JM, Beck J, Zimmermann M, *et al.* Clinical evaluation of intraparenchymal Spiegelberg pressure sensor. *Neurosurgery* 2003;**52**:1455–9. doi:10.1227/01.NEU.0000065136.70455.6F

23 Choi E, Schuetz A, Stewart WF, *et al.* Using recurrent neural network models for early detection of heart failure onset. *J Am Med Informatics Assoc* 2017;**24**:361–70. doi:10.1093/jamia/ocw112

24 Benedetto S, Caldato C, Bazzan E, *et al.* Assessment of the fitbit charge 2 for monitoring

heart rate. *PLoS One* 2018;**13**:1–10. doi:10.1371/journal.pone.0192691

25 Wright SP, Hall Brown TS, Collier SR, *et al.* How consumer physical activity monitors could transform human physiology research. *Am J Physiol - Regul Integr Comp Physiol* 2017;**312**:R358–67. doi:10.1152/ajpregu.00349.2016

26 Tully MA, McBride C, Heron L, *et al.* The validation of Fibit ZipTM physical activity monitor as a measure of free-living physical activity. *BMC Res Notes* 2014;**7**:952. doi:10.1186/1756-0500-7-952

27 Diaz KM, Krupka DJ, Chang MJ, *et al.* Fitbit®: An accurate and reliable device for wireless physical activity tracking. *Int J Cardiol* 2015;**185**:138–40. doi:10.1016/j.ijcard.2015.03.038

28 Achten J, Jeukendrup AE. Heart rate monitoring applications and limitation. *Sport Med* 2003;**33**:517–38.

29 Alnaeb M, Alobaid N, Seifalian A, *et al.* Optical Techniques in the Assessment of Peripheral Arterial Disease. *Curr Vasc Pharmacol* 2006;**5**:53–9. doi:10.2174/157016107779317242

30 Schultz-Ehrenburg U, Blazek V. Value of quantitative photoplethysmography for functional vascular diagnostics: Current status and prospects. *Skin Pharmacol Appl Skin Physiol* 2001;**14**:316–23. doi:10.1159/000056362

31 Millasseau SC, Guigui FG, Kelly RP, *et al.* Noninvasive assessment of the digital volume pulse: Comparison with the peripheral pressure pulse. *Hypertension* 2000;**36**:952–6. doi:10.1161/01.HYP.36.6.952

32 Simpson CR, Kohl M, Essenpreis M, *et al.* Near-infrared optical properties of ex vivo human skin and subcutaneous tissues measured using the Monte Carlo inversion technique. *Phys Med Biol* 1998;**43**:2465–78. doi:10.1088/0031-9155/43/9/003

33 Kamal AAR, Harness JB, Irving G, *et al.* Skin photoplethysmography - a review. *Comput Methods Programs Biomed* 1989;**28**:257–69. doi:10.1016/0169-2607(89)90159-4

34 McCombie DB, Shaltis PA, Reisner AT, *et al.* Adaptive hydrostatic blood pressure calibration: Development of a wearable, autonomous pulse wave velocity blood pressure monitor. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*. 2007. 370–3. doi:10.1109/IEMBS.2007.4352301

35 Wallen MP, Gomersall SR, Keating SE, *et al.* Accuracy of heart rate watches: Implications for weight management. *PLoS One* 2016;**11**:1–9. doi:10.1371/journal.pone.0154420

36 Parak J, Korhonen I. Evaluation of wearable consumer heart rate monitors based on photopletysmography. *2014 36th Annu Int Conf IEEE Eng Med Biol Soc EMBC 2014* 2014;:3670–3. doi:10.1109/EMBC.2014.6944419

37 Plasqui G, Westerterp KR. Physical activity assessment with accelerometers: An evaluation against doubly labeled water. *Obesity* 2007;**15**:2371–9. doi:10.1038/oby.2007.281

38 Rowlands A V., Stone MR, Eston RG. Influence of speed and step frequency during walking and running on motion sensor output. *Med Sci Sports Exerc* 2007;**39**:716–27. doi:10.1249/mss.0b013e318031126c

39 Kavanagh JJ, Menz HB. Accelerometry: A technique for quantifying movement patterns during walking. *Gait Posture* 2008;**28**:1–15. doi:10.1016/j.gaitpost.2007.10.010

40 Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health* 2015;**36**:345–59. doi:10.1146/annurev-publhealth-031914-122747

41 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits*

*Transl Sci* 2010;**2010**:1–5.http://www.ncbi.nlm.nih.gov/pubmed/21347133%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3041534

42 Jiang M, Chen Y, Liu M, *et al.* A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Informatics Assoc* 2011;**18**:601–6. doi:10.1136/amiajnl-2011-000163

43 Ebadollahi S, Sun J, Gotz D, *et al.* Predicting Patient's Trajectory of Physiological Data using Temporal Trends in Similar Patients: A System for Near-Term Prognostics. *AMIA Annu Symp Proc* 2010;**2010**:192–6.

44 Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J Biomed Inform* 2011;**44**:859–68. doi:10.1016/j.jbi.2011.05.004

45 Very a NJ a, Andhi TEKG, Urns GEB, *et al.* Medication-related Clinical Decision Support in Computerized Provider Order Entry Systems : A Review. *J Am Med Informatics Assoc* 2007;**14**:29–40. doi:10.1197/jamia.M2170.Introduction

46 Agichtein E, Brill E, Dumais S. Improving Web Search Ranking by Incorporating .pdf. ;:19–26.http://delivery.acm.org/10.1145/1150000/1148177/p19-agichtein.pdf?ip=133.51.28.240&id=1148177&acc=ACTIVE SERVICE&key=D2341B890AD12BFE.5A1FE8BDE322CB75.B9D506412D243992.4D4702B0C3E38B35&__acm__=1530114788_b686d8ca7157db1f54e46486d3d134f0

47 Adewumi AO, Akinyelu AA. A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *Int J Syst Assur Eng Manag* 2017;**8**:937–53. doi:10.1007/s13198-016-0551-y

48 Karhade A V., Thio QCBS, Ogink PT, *et al.* Development of Machine Learning Algorithms for Prediction of 30-Day Mortality after Surgery for Spinal Metastasis. *Clin Neurosurg* 2019;**85**:E83–91. doi:10.1093/neuros/nyy469

49 Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. *J Intell Learn Syst Appl* 2017;**09**:1–16. doi:10.4236/jilsa.2017.91001

50 Vamathevan J, Clark D, Czodrowski P, *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019;**18**:463–77. doi:10.1038/s41573-019-0024-5

51 Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl* 2008;**34**:366–74. doi:10.1016/j.eswa.2006.09.004

52 Liao JG, Chin KV. Logistic regression for disease classification using microarray data: Model selection in a large p and small n case. *Bioinformatics* 2007;**23**:1945–51. doi:10.1093/bioinformatics/btm287

53 Bhatia S, Prakash P, Pillai GN. SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features. *Lect Notes Eng Comput Sci* 2008;**2173**:34–8.

54 Ramos-González J, López-Sánchez D, Castellanos-Garzón JA, *et al.* A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. *Comput Biol Med* 2017;**86**:98–106. doi:10.1016/j.compbiomed.2017.05.010

55 Chen J, Li K, Tang Z, *et al.* A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment. *IEEE Trans Parallel Distrib Syst* 2017;**28**:919–33. doi:10.1109/TPDS.2016.2603511

56 Menze BH, Kelm BM, Masuch R, *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 2009;**10**:1–16. doi:10.1186/1471-2105-10-213

57 Jiang R, Tang W, Wu X, *et al.* A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* 2009;**10**:1–12. doi:10.1186/1471-2105-10-S1-S65

58 Bengio Y, Grandvalet Y. No Unbiased Estimator of the Variance ofK-Fold Cross-Validation Yoshua. *J ofMachine Learn Res* 2004;**5**:1089–1105. doi:10.1016/S0006-291X(03)00224-9

59 Choi. E, Bahadori MT, Kulas JA, *et al.* RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Adv Neural Inf Process Syst 29 (NIPS 2016)* Published Online First: 2016. doi:10.1063/1.859355

60 Ma F, Chitta R, Zhou J, *et al.* Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017. doi:10.1145/3097983.3098088

61 Meng Y, Speier W, Shufelt C, *et al.* A Machine Learning Approach to Classifying Self-Reported Health Status in a cohort of Patients with Heart Disease using Activity Tracker Data. *IEEE J Biomed Heal Informatics* 2019;**PP**:1–1. doi:10.1109/JBHI.2019.2922178

62 Fine S, Singer Y, Tishby N. The hierarchical hidden Markov model: Analysis and applications. *Mach Learn* 1998;**32**:41–62. doi:10.1023/A:1007469218079

63 Li S, Li W, Cook C, *et al.* Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*

2018;:5457–66. doi:10.1109/CVPR.2018.00572

64 Hochreiter S, Urgen Schmidhuber J. Long Shortterm Memory. *Neural Comput* 1997;**9**:17351780.http://www7.informatik.tu-muenchen.de/~hochreit%0Ahttp://www.idsia.ch/~juergen

65 Zaremba W, Sutskever I, Vinyals O. Recurrent Neural Network Regularization. *arXiv* 2014;:1–8.http://arxiv.org/abs/1409.2329

66 Russakovsky O, Deng J, Su H, *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 2015;**115**:211–52. doi:10.1007/s11263-015-0816-y

67 Xu R, Wunsch DC, Frank RL. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Trans Comput Biol Bioinforma* 2007;**4**:681–92. doi:10.1109/TCBB.2007.1057

68 Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. *Calif Univ San Diego La Jolla Inst Cogn Sci* Published Online First: 1985.https://apps.dtic.mil/docs/citations/ADA164453

69 Elman JL. Finding structure in time. *Cogn Sci* 1990;**14**:179–211. doi:10.1016/0364-0213(90)90002-E

70 Cho K, Bahdanau D, Bougares F, *et al.* Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *arXiv* Published Online First: 2014. doi:10.1074/jbc.M608066200

71 Arnold CW, Oh A, Chen S, *et al.* {E}valuating topic model interpretability from a primary care physician perspective. *Comput Methods Programs Biomed* 2015;**In press**. doi:10.1016/j.cmpb.2015.10.014

72 Mikolov T, Chen K, Corrado G, *et al.* Distributed Representations of Words and Phrases

and Their Compositionality. *Adv Neural Inf Process Syst 2013)* 2013;:3111–9. doi:10.1162/jmlr.2003.3.4-5.951

73 Vaswani A, Brain G, Shazeer N, *et al.* Attention Is All You Need. *Adv Neural Inf Process Syst* 2017;:5998–6008.http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

74 Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1181004805* Published Online First: 2018.http://arxiv.org/abs/1810.04805

75 Choi E, Xiao C, Sun J, *et al.* Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In: *Advances in Neural Information Processing Systems*. 2018. 4547–57.

76 Choi E, Bahadori MT, Searles E, *et al.* Multi-layer Representation Learning for Medical Concepts. 2016;:1495–504.http://arxiv.org/abs/1602.05568

77 Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. In: *arXiv:1301.3781*. 2013. 1–12.

78 Holzinger A. From machine learning to explainable AI. *DISA 2018 - IEEE World Symp Digit Intell Syst Mach Proc* 2018;:55–66. doi:10.1109/DISA.2018.8490530

79 Ouyang W, Wang X, Zeng X, *et al.* Training deformable object models for human detection based on alignment and clustering. In: *IEEE conference on computer vision and pattern recognition*. 2015. 2403–12. doi:10.1007/978-3-319-10602-1_27

80 Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2014. doi:10.1109/CVPR.2014.81

81 Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align

and translate. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2015. 1–15.

82 Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997;**30**:1145–59. doi:10.1016/S0031-3203(96)00142-2

83 Davis J, Goadrich M. The relationship between precision-recall and ROC curves. *ACM Int Conf Proceeding Ser* 2006;**148**:233–40. doi:10.1145/1143844.1143874

84 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:1–21. doi:10.1371/journal.pone.0118432

85 Ong MK, Romano PS, Edgington S, *et al.* Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure the better effectiveness after transition-heart failure (BEAT-HF) randomized clinical trial. *JAMA Intern Med* 2016;**176**:310–8. doi:10.1001/jamainternmed.2015.7712

86 Shaw RJ, Steinberg DM, Bonnet J, *et al.* Mobile health devices: Will patients actually use them? *J Am Med Informatics Assoc* 2016;**23**:462–6. doi:10.1093/jamia/ocv186

87 Black JJT, Romano PSP, Sadeghi B, *et al.* A remote monitoring and telephone nurse coaching intervention to reduce readmissions among patients with heart failure: study protocol for the Better Effectiveness After Transition-Heart Failure (BEAT-HF) randomized controlled trial. *Trials* 2014;**15**:124–35.

88 Speier W, Dzubur E, Zide M, *et al.* Evaluating utility and compliance in a patient-based eHealth study using continuous-time heart rate and activity trackers. *J Am Med Informatics Assoc* 2018.

89 Alharbi M, Straiton N, Gallagher R. Harnessing the Potential of Wearable Activity Trackers

for Heart Failure Self-Care. *Curr Heart Fail Rep* 2017;**14**:23–9. doi:10.1007/s11897-017-0318-z

90 Franklin NC, Lavie CJ, Arena RA. Personal health technology: A new era in cardiovascular disease prevention. *Postgrad Med* 2015;**127**:150–8. doi:10.1080/00325481.2015.1015396

91 Smith-spangler C, Gienger AL, Lin N, *et al.* CLINICIAN ' S CORNER Using Pedometers to Increase Physical Activity A Systematic Review. *Clin Corner* 2014;**298**:2296. doi:10.1001/jama.298.19.2296

92 Hale TM, Jethwani K, Kandola MS, *et al.* A Remote Medication Monitoring System for Chronic Heart Failure Patients to Reduce Readmissions: A Two-Arm Randomized Pilot Study. *J Med Internet Res* 2016;**18**:e91. doi:10.2196/jmir.5256

93 Cella D, Riley W, Stone A, *et al.* Initial Adult Health Item Banks and First Wave Testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) Network: 2005-2008. *J Clin Epidemiol* 2011;**63**:1179–94. doi:10.1016/j.jclinepi.2010.04.011.Initial

94 Meng Y, Speier W, Ong M, *et al.* HCET : Hierarchical Clinical Embedding with Topic Modeling on Electronic Health Record for Predicting Depression. *IEEE J Biomed Heal Informatics* Published Online First: 2020. doi:10.1109/JBHI.2020.3004072

95 Cadmus-bertram LA, Marcus BH, Patterson RE, *et al.* Physical Activity Intervention for Women. *Am J Prev Med* 2015;**49**:414–8. doi:10.1016/j.amepre.2015.01.020

96 Wang JB, Cadmus-bertram LA, Natarajan L, *et al.* Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial. 2014;:18–23. doi:10.1089/tmj.2014.0176

97 Benedetto S, Caldato C, Bazzan E, *et al.* Assessment of the fitbit charge 2 for monitoring heart rate. *PLoS One* 2018;**13**:e0192691. doi:10.1371/journal.pone.0192691

98 Diaz KM, Krupka DJ, Chang MJ, *et al.* Fitbit®: An accurate and reliable device for wireless physical activity tracking. *Int J Cardiol* 2015;**185**:138–40. doi:10.1016/j.ijcard.2015.03.038

99 Reddy RK, Pooni R, Zaharieva DP, *et al.* Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise : Evaluation Study Corresponding Author : 2018;**6**. doi:10.2196/10338

100 Jo E, Lewis K, Directo D, *et al.* Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. 2016;:540–7.

101 Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinol Metab* 2012;**10**:486–9. doi:10.5812/ijem.3505

102 Liu H, Cella D, Gershon R, *et al.* Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *J Clin Epidemiol* 2010;**63**:1169–78. doi:10.1016/j.jclinepi.2009.11.021

103 Hays RD, Bjorner JB, Revicki DA, *et al.* Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Qual Life Res* 2009;**18**:873–80.

104 Spiegel BMR, Hays RD, Bolus R, *et al.* Development of the NIH patient-reported outcomes measurement information system (PROMIS) gastrointestinal symptom scales. *Am J Gastroenterol* 2014;**109**:1804–14. doi:10.1038/ajg.2014.237

105 Ogutu JO, Piepho HP, Schulz-Streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc* 2011;**5**:3–7. doi:10.1186/1753-6561-5-S3-S11

106 Stolcke A, Omohundro S. Hidden Markov Model Induction by Bayesian Model Merging. *Neural Inf Process Syst* 1993;**5**:11–8.http://papers.nips.cc/paper/669-hidden-markov-

model-induction-by-bayesian-model-merging.pdf

107 Pilkonis PA, Yu L, Dodds NE, *et al.* Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. *J Psychiatr Res* 2014;**56**:112–9. doi:10.1016/j.jpsychires.2014.05.010

108 Ryu E, Chamberlain AM, Pendegraft RS, *et al.* Quantifying the impact of chronic conditions on a diagnosis of major depressive disorder in adults: A cohort study using linked electronic medical records. *BMC Psychiatry* 2016;**16**:1–9. doi:10.1186/s12888-016-0821-x

109 Jin H, Wu S, Di Capua P. Development of a Clinical Forecasting Model to Predict Comorbid Depression Among Diabetes Patients and an Application in Depression Screening Policy Making. *Prev Chronic Dis* 2015;**12**:1–10. doi:10.5888/pcd12.150047

110 Lin Y, Huang S, Simon GE, *et al.* Data-based Decision Rules to Personalize Depression Follow-up. *Sci Rep* 2018;**8**:4–11. doi:10.1038/s41598-018-23326-1

111 Zhang J, Xiong H, Huang Y, *et al.* M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data. *Proc - 2015 IEEE Int Conf Big Data, IEEE Big Data 2015* 2015;:2569–77. doi:10.1109/BigData.2015.7364054

112 Bian J, Barnes LE, Chen G, *et al.* Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis. *2017 IEEE EMBS Int Conf Biomed Heal Informatics, BHI 2017* 2017;:457–60. doi:10.1109/BHI.2017.7897304

113 Usama M, Ahmad B, Wan J, *et al.* Deep feature learning for disease risk assessment based on convolutional neural network with intra-layer recurrent connection by using hospital big data. *IEEE Access* 2018;**6**:67927–39. doi:10.1109/ACCESS.2018.2879158

114 LePendu P, Iyer S V., Bauer-Mehren A, *et al.* Pharmacovigilance using clinical notes. *Clin*

*Pharmacol Ther* 2013;**93**:547–55. doi:10.1038/clpt.2013.47

115 Huang SH, LePendu P, Iyer S V, *et al.* Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Informatics Assoc* 2014;**21**:1069–75. doi:10.1136/amiajnl-2014-002733

116 Miotto R, Li L, Kidd BA, *et al.* Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016;**6**:1–10. doi:10.1038/srep26094

117 Choi E, Bahadori MT, Schuetz A, *et al.* Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *Proc Mach Learn Healthc 2016* 2016;**56**:301–18.http://www.ncbi.nlm.nih.gov/pubmed/28286600%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5341604

118 Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *J Mach Learn Res* 2003;**3**:993–1022.

119 Griffiths TL, Steyvers M. Finding scientific topics. *Proc Natl Acad Sci U S A* 2004;**101**:5228–35. doi:10.1073/pnas.0307752101

120 Arnold CW, El-Saden S, Bui AAT, *et al.* Clinical case-based retrieval using latent topic analysis. *AMIA Annu Symp Proc* 2010;:26–30.

121 Arnold CW, Speier W. A topic model of clinical reports. In: *35th international ACM SIGIR conference on Research and development in information retrieval*. 2012. 1031–2.

122 Arnold CW, Oh A, Chen S, *et al.* Evaluating topic model interpretability from a primary care physician perspective. *Comput Methods Programs Biomed* 2015;**124**:67–75. doi:10.1016/j.cmpb.2015.10.014

123 Speier W, Ong M, Arnold C. Using phrases and document metadata to improve topic

modeling of clinical reports. *J Biomed Inform* 2016;**61**:260–6.

124 Juba B, Musco C, Long F, *et al.* Principled Sampling for Anomaly Detection. 2015;:1–18. doi:10.14722/ndss.2015.23268

125 Yang Z, Dai Z, Yang Y, *et al.* XLNet : Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1190608237* 2019;:1.

126 Pham T, Tran T, Phung D, *et al.* Predicting healthcare trajectories from medical records : A deep learning approach. *J Biomed Inform* 2017;**69**:218–29. doi:10.1016/j.jbi.2017.04.001

127 Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. *Risk Manag Healthc Policy* 2011;**4**:47–55. doi:10.2147/RMHP.S12985

128 Kurt Kroenke, MD; Robert L. Spitzer M. The PHQ-9 : A New Depression Measure. *Psychiatr Ann* 2002;**32**:509–515.

129 Choi E, Bahadori MT, Song L, *et al.* GRAM: Graph-based attention model for healthcare representation learning. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017. 787–95. doi:10.1145/3097983.3098126

130 Hsich E, Gorodeski EZ, Blackstone EH, *et al.* Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ Cardiovasc Qual Outcomes* 2011;**4**:39–45. doi:10.1161/CIRCOUTCOMES.110.939371

131 Limsopatham N, Macdonald C, Ounis I. Learning to Combine Representations for Medical Records Search. *Proc 36th Int ACM SIGIR Conf Res Dev Inf Retr* 2013;:833–6. doi:10.1145/2484028.2484177

132 Kaiming He, Xiangyu Zhang, Shaoqing Ren JS. Deep Residual Learning for Image Recognition Kaiming. In: *IEEE conference on computer vision and pattern recognition*. 2016. 770–8. doi:10.1002/chin.200650130

133 Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015*. 2015. 1–15.

134 Henry, J., Pylypchuk, Y. ST& P V. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. *ONC Data Brief, no35* 2016;:2008–15.https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php

135 Meng Y, Speier W, Shufelt C, *et al.* A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease Using Activity Tracker Data. *IEEE J Biomed Heal Informatics* 2020;**24**:878–84. doi:10.1109/JBHI.2019.2922178

136 Radford A, Wu J, Child R, *et al.* Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 2019;**1**.http://arxiv.org/abs/2007.07582

137 Li Y, Rao S, Solares JRA, *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020;**10**:1–17. doi:10.1038/s41598-020-62922-y

138 Bai T, Egleston BL, Zhang S, *et al.* Interpretable representation learning for healthcare via capturing disease progression through time. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min* 2018;:43–51. doi:10.1145/3219819.3219904

139 Suresh H, Hunt N, Johnson A, *et al.* Clinical Intervention Prediction and Understanding using Deep Networks. *arXiv Prepr arXiv170508498v1* 2017;:1–16.http://arxiv.org/abs/1705.08498

140 Choi E, Xu Z, Li Y, *et al.* Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer. *Proc AAAI Conf Artif Intell* 2020;**34**:606–13. doi:10.1609/aaai.v34i01.5400

141 Wei WQ, Teixeira PL, Mo H, *et al.* Combining billing codes, clinical notes, and medications

from electronic health records provides superior phenotyping performance. *J Am Med Informatics Assoc* 2016;**23**:20–7. doi:10.1093/jamia/ocv130

142 Guo C, Pleiss G, Sun Y, *et al.* On Calibration of Modern Neural Networks. In: *34th International Conference on Machine Learning*. 2017. 1321–30.https://arxiv.org/pdf/1706.04599.pdf

143 Soni S, Roberts K. Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering. In: *12th Conference on Language Resources and Evaluation (LREC 2020)*. 2020. 5532–5538.https://rajpurkar.github.io/

144 Liu D, Miller T. Federated pretraining and fine tuning of BERT using clinical notes from multiple silos. *arXiv:200208562* Published Online First: 2020.http://arxiv.org/abs/2002.08562

145 Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 2018;**6**:52138–60. doi:10.1109/ACCESS.2018.2870052

146 Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv:170808296 (2017)* Published Online First: 2017.http://arxiv.org/abs/1708.08296

147 Petrella RJ, Koval JJ, Cunningham DA. A Self-Paced Step Test to Predict Aerobic Fitness in Older Adults in the Primary Care Clinic. 2001;:632–8.

148 Dong Q, Liu H, Yang D, *et al.* Diabetes mellitus and arthritis: Is it a risk factor or comorbidity? *Med (United States)* 2017;**96**:1–6. doi:10.1097/MD.0000000000006627

149 Kachuee M, Kiani MM, Mohammadzade H, *et al.* Cuff-Less Blood Pressure Estimation Algorithms for Continuous Health-Care Monitoring. *IEEE Trans Biomed Eng*

2016;**9294**:1–11. doi:10.1109/TBME.2016.2580904

150 Fallo F, Barzon L, Rabbia F, *et al.* Circadian blood pressure patterns and life stress. *Psychother Psychosom* 2002;**71**:350–6. doi:10.1159/000065996

151 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning for electronic health records. *npj Digit Med* 2018;:1–10. doi:10.1038/s41746-018-0029-1

152 Tombaugh TN, Mclntyre NJ. The Mini-Mental State Examination : 1992;:922–35.