UNIVERSITY OF CALIFORNIA

Los Angeles

Pediatric Asthma Management: System and Algorithms

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Anahita Hosseini

2019

ABSTRACT OF THE DISSERTATION

Pediatric Asthma Management: System and Algorithms

by

Anahita Hosseini
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2019
Professor Majid Sarrafzadeh, Chair

Pervasive health is an emerging field of research aimed at providing health monitoring to anyone at any time and anywhere. In contrast to existing healthcare systems, this paradigm is centered on prevention, wellness maintenance, and proactive care outside the hospital setting. In recent years, advances in sensing technologies and machine learning, as well as wide adoption of smart devices, have led to the development of successful pervasive health solutions such as daily activity monitoring.

Despite this growth, the full potential of pervasive health is yet to be unlocked. In this dissertation, we make several contributions to the application and methodological sides of pervasive health to advance it closer to its promises. We investigate design, implementation, and deployment of a highly context-aware and real-time solution for the complex application of pediatric asthma management, trying to predict the exacerbation of asthma ahead of time. On the methodological side, we investigate unmet data modeling challenges that arise in building a highly context-aware pervasive health solution and propose novel approaches to address them. In particular, we approach the challenges of integrative and interpretable modeling of heterogeneous health data, transferring and adapting models to new domains/individuals, and tackling sparsity of labels in real-world settings.

The dissertation of Anahita Hosseini is approved.

Yizhou Sun

Alex A. T. Bui

Kai-Wei Chang

Majid Sarrafzadeh, Committee Chair

University of California, Los Angeles

2019

To my beloved husband, loving parents, supportive brother, and joyful sister …

Table of Contents

List of Figures

List of Tables

Acknowledgments

My PhD journey has been a bright chapter of my life and for that, I have many people to thank. I would like to start by sincerely thanking my advisor Majid Sarrafzadeh. I am deeply grateful for the caring support, practical guidance, and freedom he offered me not only in my academic research but also in the ups and downs of my life in the past four years. I would also like to express my heartfelt gratitude to professor Yizhou Sun for being highly approachable in mentoring me to explore new directions in my research and for catalyzing my growth by sharing insightful comments and intuitive ideas. Further, I would like to thank Alex A. T. Bui who has been a compassionate and positive supervisor giving me space and opportunity to grow my capabilities. I would also like to acknowledge processor Kai-Wei Chang for serving in my committee and his valuable feedback.

I would like to thank my support system. First and foremost my beloved husband Shayan for being the sunshine of my life and giving me every reason to smile with no reason. I am grateful to him for helping me grow my curiosity and vision and for making learning a pleasurable part of our life. I would also like to thank my family especially my parents for their support that was felt very close despite the distance. They helped me to stay grounded at hard times and showed me the open windows when I saw all the doors closed. All I have achieved is for and because of them. I am also very grateful to my dearest friend Niloofar who has been a great source of joy and support in the ups and downs of this journey.

I would also like to acknowledge my fellow collaborators for their valuable contributions to this dissertation. A version of chapter 3 was published with the title ”Unsupervised Prediction of Negative Health Events Ahead of Time” [114] with the co-authorship of Majid Sarrafzadeh as the PI. Chapter 4 was published with the title ”Heteromed: Heterogeneous information network for medical diagnosis” [111] in which Wenjun Wu developed the model used for symptom extraction, Ting Chen provided feedback and editing, and Yizhou Sun and Majid Sarrafzadeh were the PIs. A version of chapter 5 is currently under submission [112] and Tyler Davis contributed to this work by fruitful discussions and assisting in the development of the baseline model and editing the manuscript. Chatper 6 borrows some text from a published work with the title ”Children Activity Recognition: Challenges and Strategies” [113] in which Shayan Fazeli contributed to the editing and final testing of the proposed model. Chapter 7 is a version of a work published under the title ”Domain Adaptation in Children Activity Recognition” [115] in which Davina Zamanzadeh contributed to the editing. The two aforementioned studies were conducted on a dataset collected by the efforts of Lisa Valencia, Eleanne van Vliet, and Rima Habre with Alex A. T. Bui serving as the PI and Majid Sarrafzadeh as the co-PI of both studies. Chapter 8 is an extended version of two studies, one published with the title ”HIPAA compliant wireless sensing smartwatch application for the self-management of pediatric asthma” [110] and one in preparation manuscript with the title ”Biomedical Real-time Health Evaluation (BREATHE): Towards an mHealth Informatics Platform” [32]. In these publications, I developed the server-side infrastructure for data collection and management and conducted all data analysis and Chris M Buonocore, Sepideh Hashemzadeh, Costas Sideris, and Arjun Arjun contributed to the development of mobile app and server security design. Furthermore, Rose Rochio, Nate Jacobs, Mindy Ross, Sande Okelo, Fred Lurmann, Sandrah Eckel, Genevieve Dunton, Frank Gilliland, and Dimitrios Stripelis provided clinical, and technical assistance as members of PRISMS consortium, a study with Alex A. T. Bui as PI and Majid Sarrafzadeh as co-PI. I would also like to acknowledge the grant funding majority of studies in this dissertation: NIH/NIBIB U54 EB022002.

Curriculum Vitae

2011 – 2015 B.S. in Software Engineering, Sharif University of Technology (SUT), Tehran, Iran.

2013 – 2015 Research Assistant, Human Computer Interaction Lab, Sharif University of Technology

2015– 2019 Teaching Fellow/Instructor, Computer Science Department University of California, Los Angeles.

2011 – 2019 Ph.D. student in Computer Science, eHealth Research Lab, University of California, Los Angeles (UCLA).

Publications

Hosseini, A., Davis, T., and Sarrafzadeh, M., 2019. Hierarchical Target-Attentive Diagnosis Prediction in Heterogeneous Information Networks. Submitted to 2019 19th IEEE International Conference on Data Mining (ICDM).

Hosseini, A. and Sarrafzadeh, M., 2019. Unsupervised Prediction of Negative Health Events Ahead of Time. In 2019 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).

Hosseini, A., Zamanzadeh, D., Valencia, L., Habre, R., Bui, A., and Sarrafzadeh, M., 2019. Domain Adaptation in Children Activity Recognition. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Hosseini, A., Chen, T., Wu, W., Sun, Y. and Sarrafzadeh, M., 2018, October. HeteroMed: Heterogeneous Information Network for Medical Diagnosis. In Proceedings of the 27th ACM

International Conference on Information and Knowledge Management (pp. 763-772). ACM.

Kachuee, M., Hosseini, A., Moatamed, B., Darabi, S. and Sarrafzadeh, M., 2017, November. Context-aware feature query to improve the prediction performance. In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 838-842). IEEE.

Hosseini, A., Buonocore, C., Hashemzadeh, S., Hojaiji, H., Kalantarian, H., Sideris, C., Bui, A., King, C. and Sarrafzadeh, M., 2017. Feasibility of a secure wireless sensing smartwatch application for the self-management of pediatric asthma. Sensors, 17(8), p.1780.

Shahmohammadi, F., Hosseini, A., King, C.E. and Sarrafzadeh, M., 2017, July. Smartwatch based activity recognition using active learning. In Proceedings of the Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (pp. 321-329). IEEE Press.

Kalantarian, H., Sideris, C., Le, T., Hosseini, A. and Sarrafzadeh, M., 2016, August. Computation offloading for real-time health-monitoring devices. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4971-4974). IEEE.

Hosseini, A., Kalantarian, H. and Sarrafzadeh, M., 2016, August. Adaptive data processing for real-time nutrition monitoring. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1882-1885). IEEE.

Hosseini, A., Buonocore, C.M., Hashemzadeh, S., Hojaiji, H., Kalantarian, H., Sideris, C., Bui, A.A., King, C.E. and Sarrafzadeh, M., 2016, June. HIPAA compliant wireless sensing smartwatch application for the self-management of pediatric asthma. In 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN).

CHAPTER 1

Introduction

1.1 Motivation

Over the past century, healthcare systems have experienced remarkable advancements, thanks to multi-disciplinary discoveries in science, from the introduction of evidence-based medicine to digital technologies such as medical imaging. Healthcare evolution has resulted in a 5.4 years increase of life expectancy, 2.4 billion growth in world population, and 44.2 of mortality rate reduction in infants between the years of 1985 and 2015 [69].

Despite these advancements, the healthcare system still faces serious challenges including access limitations, skyrocketing costs, and being occasional and reactive, meaning that clinical interventions take place only after the manifestation of a disease or a new condition. To tackle these, preventive and pervasive approaches to healthcare has gained remarkable attention from researchers. In this paradigm, there is a focus on wellness maintenance and proactive health monitoring outside the hospital setting, which not only improves the quality of care but can also result in a great reduction of healthcare costs by permitting early detection of health conditions and lowering unnecessary clinician and hospital visits. In recent years, advances in computer science and sensing technologies have paved the way for the rise of pervasive healthcare, a research domain working towards the ambitious goal of providing healthcare to anyone at any time and anywhere [171]. This goal can only be fully achieved when an integrated, accurate, and personalized view of a person's past medical history and real-time context can be obtained.

Existing pervasive health solutions mainly rely on environmental and physiological sensing to achieve context-awareness. With the recent prevalence of mobile devices and integration of sensing technologies, pervasive health has witnessed a fast growth in the data collection through mobile health (mHealth) applications [79]. Analysis of the collected data, which is a key element of pervasive health solutions, has also witnessed notable improvements. With the emergence of machine learning and more recently deep neural networks, they have successfully replaced the classical approaches formerly used in pervasive health and have brought state of the art results in various applications such as daily activity tracking [128], detection of mental health problems [6], and cardiovascular monitoring [15].

These advancements, although showing the promise of pervasive health, have not yet unlocked the full potential that it can offer. Aside from the need to explore applications of pervasive health in more diverse populations (e.g., children) and disease cohorts, there are still important unmet challenges in the design of analytical models and systems for achieving a fully context-aware healthcare solution, usable by diverse groups of people. The current challenges in training healthcare models include sparsity of labels, heterogeneity of data sources and types, the need for adapting to people differences, and providing interpretable explanations of the decision process. Furthermore, designed systems should meet domain-specific requirements such as data transfer security and flexibility in sensor integration.

In this dissertation, we investigate and address the aforementioned challenges in the pervasive health domain in the context of an impactful application, pediatric asthma management. We propose novel algorithms and techniques that can advance us closer towards the promises of pervasive health and employ them to study the feasibility of mitigating asthma symptoms and negative events in children.

1.1.1 Significance of Asthma Management

Asthma is a chronic inflammatory lung disease that can cause recurring periods of acute exacerbation with different severity degrees [161]. It is known as the most common chronic disease of childhood and affects nearly 7 million or 10% percent of pediatrics population (0 - 17 years) in the united states [167]. Asthma is also reported as the leading cause of hospitalization for those under the age of 15 [208] and today more than 57% of children with asthma experience an attack each year [120].

Asthma is associated with high direct and indirect costs and poor management of its symptoms can significantly impact one's quality of life. The direct healthcare costs for pediatric asthma in the USA is estimated to go up to $10 billion [120], with an estimated average yearly cost of care for a child with asthma being $1,450. Indirect costs of pediatric asthma are also considerable, including school absenteeism, which is experienced by 60 % of children with asthma, and negative impacts on a child's social and physical life [167].

It is clear that there is a significant need to alleviate asthma symptoms and reduce the costs and risks. Proper control of asthma has been shown to be extremely beneficial in reducing costs and the burden of this disease [58], with 4.6 times lower frequency of hospitalization and 1.8 times lower emergency visits compared to those with uncontrolled asthma [27].

1.2 Research Objectives

The purpose of this dissertation is to investigate the existing data modeling challenges in pervasive health, under the umbrella of pediatric asthma management application, in which we try to answer whether we can make an early, person-specific, and context-aware prediction of asthma exacerbation to mitigate and prevent such events in children?

Asthma is a complex disease with numerous impacting factors and considerable variability among different individuals. This complexity is particularly interesting as it can surface the special data modeling and system design needs that an advanced pervasive health solution should meet. Overall, the data modeling challenges approached in this dissertation are:

1. **Sparsity of labels**: The majority of the collected data in pervasive health applications remains unlabeled. This is largely due to uncontrolled and real-time data collection setting and the tediousness of offline labeling of data by domain experts. Therefore, traditional supervised models which heavily rely on the availability of labels, become disadvantageous in this domain and unsupervised approaches should be explored. To

this end, we propose an unsupervised model for the prediction of future negative events in pervasive health solutions in chapter 3.

2. **Heterogeneity of data types**: A successful pervasive health solution should be able to integrate real-time and historical data from various sources. However, the huge size, sparsity, and heterogeneity of health data pose technical challenges in building such an integrative model. To overcome these challenges, we propose a novel approach using heterogeneous information networks to represent healthcare data and model its relations which is presented in chapter 4. Chapter 5 extends this work by adding target-awareness to the representation learning process and expanding data integration.

3. **Poor model generalizabilty**: Due to person-specific variations in healthcare, the models trained for a specific group of individuals, are not always generalizable beyond that group. Furthermore, recalling the label sparsity challenge discussed above, separate modeling of each individual is not practically feasible in many applications. Therefore, there is a need for transferable models that can reuse the available knowledge, while adapting it to new individuals and groups. We present deep multi-task learning and domain adaptation frameworks for the pervasive health domain in chapters 6 and 7 respectively.

4. **Interpretablity of results**: Recognizing the sensitive nature of the healthcare domain, the decision process of the proposed solutions should be transparent to the clinicians and researchers, so that the validity of the decision can be examined. We propose integrative and unified modeling of health records and demonstrate the full interpretability of its prediction process in chapter 5.

Furthermore, we present a study on the design of a complex and highly context-aware pervasive health solution. Our proposed system for asthma management is designed with the goals of scalable and flexible sensor integration, data security, and collection of a comprehensive set of context information (e.g., physiological and environmental sensing and medical history) that can contribute to the detection of an asthma event.

1.3 Contributions

This dissertation offers several methodological and application contributions to the pervasive health domain. On the methodological side:

- we propose an unsupervised future event detection model that can capture complex and gradual changes in signal behavior prior to an event onset. We further show how the framework can be extended to semi-supervised settings (chapter 3).
- we introduce a novel approach for modeling electronic health records using heterogeneous information networks. Our model offers effective capturing of the rich structure, relation semantics, and heterogeneity of entities in healthcare (chapter 4). We further extend the model to be capable of integrating various data formats, especially time-series (chapter 5).
- we propose a hierarchical target-attentive health record embedding mechanism that can learn the importance of different records for a specific clinical prediction task. We show that our approach provides interpretability over the clinical decision process while improving the performance in clinical modeling tasks (chapter 5).
- we put forward the use of deep multi-task learning model for achieving improved personalized prediction performance (chapter 6).
- we present a deep domain adaptation model for unsupervised and semi-supervised knowledge transfer among groups with different data distributions with the aim of tackling the label missingness challenge in pervasive health applications (chapter 7).

On the application side:

- we present a secure and highly integrative pervasive health solution for pediatric asthma management (chapter 8).
- we present an effective model for early prediction of Bradycardia heart events in infants with no label supervision(chapter 3).

- we propose a fully integrative clinical diagnosis prediction model based on electronic health records, capable of achieving state of the art performance in recognizing both high-level disease cohorts and exact diagnosis codes (chapter 4 and chapter 5)

- we develop a smartwatch-based activity recognition model for children, previously missing in the field and shed light upon personalization needs in this setting (chapter 6).

- we show effective transfer of a benchmarked adult activity recognition model to a younger population with a different data collection setting (chapter 7).

1.4 Dissertation Outline

The organization of the rest of this dissertation is as follows: Chapter 2 provides background information on asthma, pervasive health applications, and algorithmic advances in modeling healthcare data. Chapter 3 introduces an approach for early detection of negative events in an unsupervised manner which is applied to an infant heart signal dataset. Chapter 4 proposes a heterogeneous information network modeling of electronic health records to improve the diagnosis prediction task. Chapter 6 introduces a children activity recognition model and provides a comparative analysis of personalization through fine-tuning and multi-task learning. Chapter 7 approaches the problem of domain transfer in pervasive health, taking the activity recognition task under the study. Chapter 5 extends chapter 4 by showing how time-series data can be integrated into the introduced model. It also Improves the prior model for improved interpretability and performance using attention mechanisms. Finally, chapter 8 provides an overview of our pediatric asthma management system design and our panel study specifications and results.

CHAPTER 2

Background

2.1 Pervasive Health

Various entities are involved in a pervasive health solution to achieve the goal of providing health monitoring to anyone at any time and anywhere. In this section, we provide a general overview of data sources, analysis goals, users, and primary application areas of such systems, all of which are summarized in Figure 2.1.

Diverse sources of data can be utilized to achieve a continuous and personalized view of an individual's health state. Ambient and wearable sensors that are placed in the individual's surrounding environment or on the body are among the most utilized sources of data. Inertial motion sensors (e.g., accelerometer and gyroscope), bio-electrical sensors collecting ECG, EMG, and EEG signals, sensors collecting vital signs (e.g., heart rate, glucose level) and those collecting environmental data (e.g., temperature and air quality) are among the most widely adopted types of sensors in pervasive health [16]. Cameras, RFIDs, and social media interactions are other data sources used in pervasive health systems [211]. Electronic health records (EHR) that contain detailed information regarding an individual's visit to the hospital and health clinics can also provide valuable healthcare information. Examples of records in EHR include those in a semi-structured format, such as symptoms and laboratory results, along with time-series data collected by bedside monitoring systems such as heart rate and respiratory rate. As researchers access to EHR is highly limited, partly due to privacy and security considerations, integration of EHR with pervasive health models has not yet been extensively studied. With the recent availability of anonymized EHR data and corresponding bedside monitoring time-series data, few studies such as [124] have approached

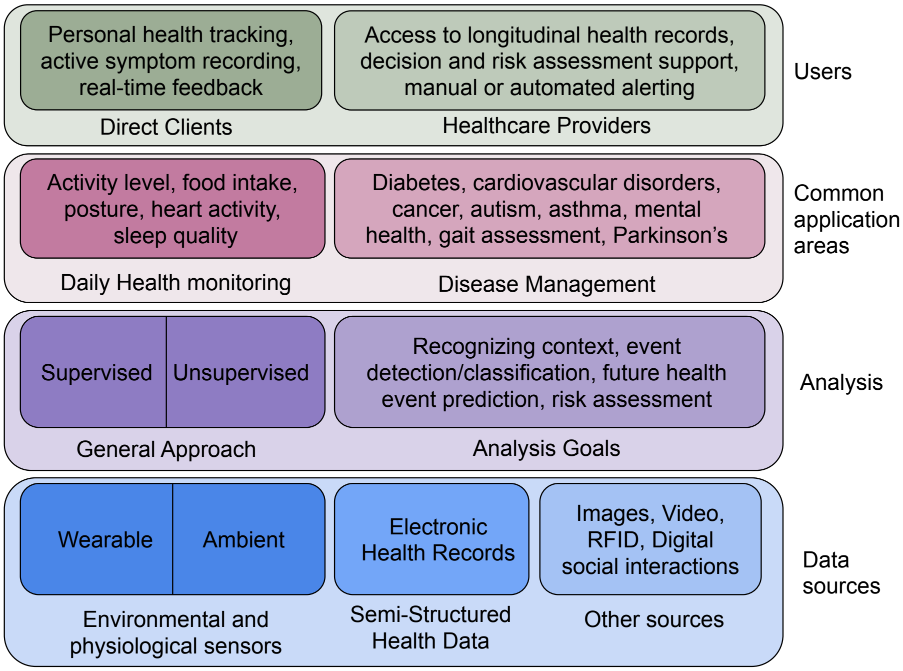

Personal health tracking, active symptom recording, real-time feedback
Access to longitudinal health records, decision and risk assessment support, manual or automated alerting
Users
Direct Clients
Healthcare Providers
Activity level, food intake, posture, heart activity, sleep quality
Diabetes, cardiovascular disorders, cancer, autism, asthma, mental health, gait assessment, Parkinson's
Common application areas
Daily Health monitoring
Disease Management
Supervised
Unsupervised
Recognizing context, event detection/classification, future health event prediction, risk assessment
Analysis
General Approach
Analysis Goals
Wearable
Ambient
Electronic Health Records
Images, Video, RFID, Digital social interactions
Data sources
Environmental and physiological sensors
Semi-Structured Health Data
Other sources

Figure 2.1: Overview of pervasive health data sources, analysis goals, users and prime applications areas.

the problem of combining this information.

Data analysis in pervasive health can provide various types of insights, depending on the application and data. Recognizing the real-time context (e.g., performed activity [220]), classifying negative events (e.g., type of anomalies in heart signal [95]), detection of negative events [151], prediction of future health events (e.g., hospitalization, clinic visit [48]), assessment of health risks (e.g., mortality [35]), and adherence improvement and monitoring (e.g., HIV medication adherence [187]) are among prime examples of analysis goals in pervasive health. These goals have been approached with supervised and more recently unsupervised methodologies depending on the availability of annotated data.

Applications of pervasive health, in general, can be categorized into daily health monitoring and chronic disease management. Examples of daily monitoring applications include tracking of activities using wearable devices such as Fitbit [64], food intake monitoring using throat movement and audio sensors [12], heart activity recording [123], sleep quality tracking [118], and posture improvement by slouch detection and intervention [87]. In chronic disease monitoring, diabetes management and non-invasive glucose assessment have long been a problem of interest [70]. Furthermore, monitoring of individuals with cardiovascular disorders with the aims of continuous heart activity tracking and re-admission prevention has been approached by prior studies [134]. Other application areas include obesity [145], cancer [155], Parkinson's [132], and asthma management [214] as well as gait assessment and fall detection for elderly individuals [207] and autism management in children [126]. Last but not least, stress and mental health monitoring along with suicide prevention through analysis of digital social interactions and other physiological signals such as activity and heart rate levels has gained considerable attention in recent years [6].

Pervasive health solutions serve both their clients as well as healthcare providers. Users and/or their caregivers can gain access to their real-time health tracking information and their medical records, leading to better self-monitoring and health management. They can also communicate instantly to their healthcare providers by actively recording their information, such as symptoms, and receive targeted alerts, reminders, or diagnostic results. On the other hand, pervasive health allows healthcare providers to obtain a more comprehensive longitudinal tracking of users' health status. Pervasive health systems can support clinical decision making and risk assessment while allowing transmission of automated or manual interventions and notifications to the users.

2.2 Sensor Data Modeling in Pervasive Health

Machine Learning has been an integral part of advancements in health informatics. In this section, we first provide a background on pervasive health data processing workflow and traditional processing techniques. We then review more recent data analysis trends, especially deep neural networks and how they are applied in the field.

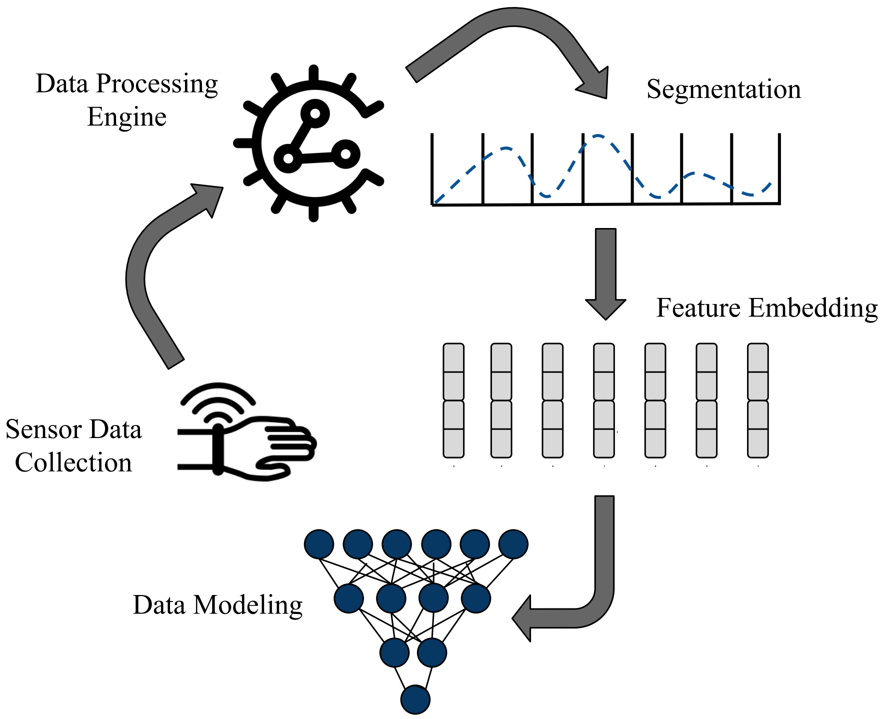


Figure 2.2: The general workflow of pervasive health data analysis

2.2.1 General Workflow

Pervasive health solutions follow a general workflow for analyzing the sensor data, which is illustrated in Figure 2.2. The process starts with acquiring data from sensors attached to users or placed in the environment. The data is then fed into a processing module which can be located in the cloud or as close as possible to the device, known as Edge Processing [195]. The time-series is then pre-processed, particularly with the aim of data cleaning and noise reduction, and is segmented into fixed or variable length windows, a process that is known as windowing. Prior studies have approached the problem of time-series windowing from two main directions. Time-based windowing, which is the most common approach, partitions the signal into fixed-size windows so that an event can be divided into two or more windows. Alternatively, event-based windowing tries to fit events into separate windows of variable length. The simplicity and lack of processing costs of time-based windowing have made this method well-suited to real-time processing applications. [127] provides a comprehensive survey of common windowing techniques used in healthcare informatics.

In the next step, extracted windows are embedded into representative feature vectors. Traditional processing approaches use feature engineering for this aim and there is a large body of research in the extraction of the most informative set of features from health signals, reviewed in [24]. These studies extract the features from both time and frequency domains of the signal, the later of which is obtained by applying the Fourier Transform decomposition [104].

Finally, the extracted features are fed into a prediction model in order to perform a supervised (e.g., classification of disease) or unsupervised task (e.g., clustering of patients based on similarity). Relying on a survey of machine learning techniques used in pervasive health before 2015, SVM, decision trees, and shallow neural networks were the most widely adopted models by researchers [67].

2.2.2 Background on Recent Sensor Data Modeling Advances

Deep Learning (DL) has recently drawn special attention of researchers in various domains following its massive success in achieving state of the art results in image and natural language processing [142]. Deep Learning has rapidly started to replace the traditional machine learning techniques in healthcare informatics as well. This trend can be specially attributed to the automatic representation learning that DL offers, allowing full or partial elimination of expert knowledge previously needed in feature engineering. In this section, we provide an overview of the design and applications of the most recent models used in healthcare as well as those appearing in further chapters of this thesis.

Deep Learning Introduction

Neural network models are designed on top of a simple feed-forward building block called Multilayer Perceptron (MLP), which consists of an input, an output and, a hidden layer. Each layer of a simple neural network is defined by a set of weights, an optional bias vector, and activation function that transforms the data before being fed into the following layer.

the MLP block can be formulated as follows:

$$\begin{aligned} h &= \sigma_1(W_1 X + b_1) \\ y &= \sigma_2(W_2 h + b_2) \end{aligned} \tag{2.1}$$

Where σ, W, and b indicate, the activation function, layer weights, and layer biases respectively. Selection of non-linear activation functions allows introducing non-linearity to the function learned by the model. The most commonly used activation functions are sigmoid, softmax, rectified linear unit (ReLU), and hyperbolic tangent (Tanh) that are all reviewed in [10]. To tackle more complex problems, a higher number of hidden layers can be employed in the network, known as a deep neural network.

Neural networks are trained by defining a customized loss function and minimizing the error observed in generated results for a single sample or a batch of samples. The error is back-propagated to the network layers using the derivative chain rule, allowing the model to update the weights and biases with an optimization algorithm such as Stochastic Gradient Decent [182]. More advanced optimization algorithms including Adagrad [68], RMSprop [91], and Adam [129] have been introduced to speed up and ease the process of finding an optimization point in the searching space.

Convolutional Neural Network (CNN)

CNN is a family of neural networks that initially showed considerable success in automatically extracting predictive features for classification of images [103]. Relying on that success, their application has been expanded to time series and sensor data analysis as well, which we focus on in this section.

CNNs allow extraction of time-invariant features of time series data which can convey important information around patterns and change processes. For this aim, a convolution filter is slid along the time series and is multiplied by the values in an element-wise manner, mapping the original signal into a new feature signal based on the following equation:

$$C_t = \sigma(W \cdot X_{t:t+l} + b) \tag{2.2}$$

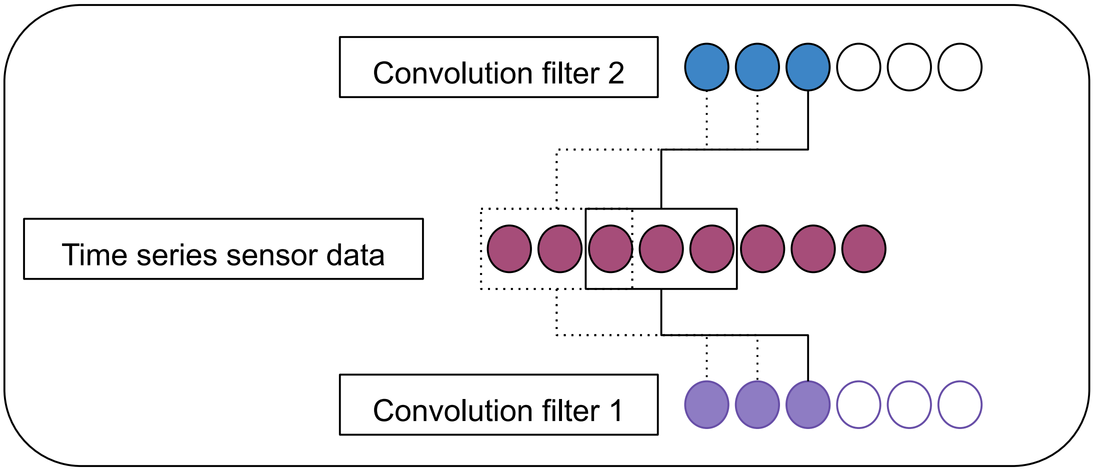


Figure 2.3: Feature extraction from time series sensor data using two convolution filters

where $X_{t:t+l}$ denotes a window of length l starting at time t, W and b are the shared weights and bias of the applied filter which are trained jointly by the model and σ denotes the non-linearity function. As each sliding filter extracts specific features of the signal, multiple filters can be employed to jointly represent features of time series data. Figure 2.3 depicts the process of feature extraction from time series data using two distinct filters, each generating a new feature signal as the filter slides along the data. These feature signals can be further used in the prediction layers of the network.

Applications of CNN in the analysis of healthcare signals has grown in recent years. In the analysis of ECG signals which record the electrical activity of heart, [7, 130, 239] are among studies that show the significance of convolution layers in classification of heartbeats and detection of arrhythmia, benchmarked on the MIT-BIH Arrhythmia publicly available database [88]. CNN has been also used for extraction of representative features for the gesture and hand movement classification based on Electromyography (EMG) signals in [172, 19, 83]. Detection and classification of seizures [8] along with advancing brain-computer interaction [135] are among other applications of CNN in health signal analysis. When working with everyday sensor-based monitoring, DeepSense [234] show that CNN can be used as a powerful tool for extraction of features from time and frequency domain signals in a unified sensor data processing framework. A more comprehensive list of CNN-based

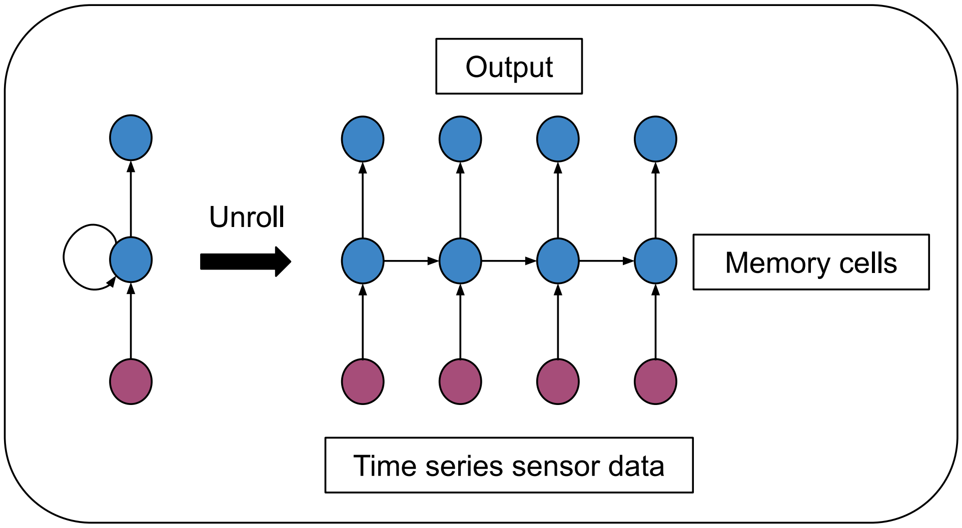


Figure 2.4: General structure of a recurrent neural network

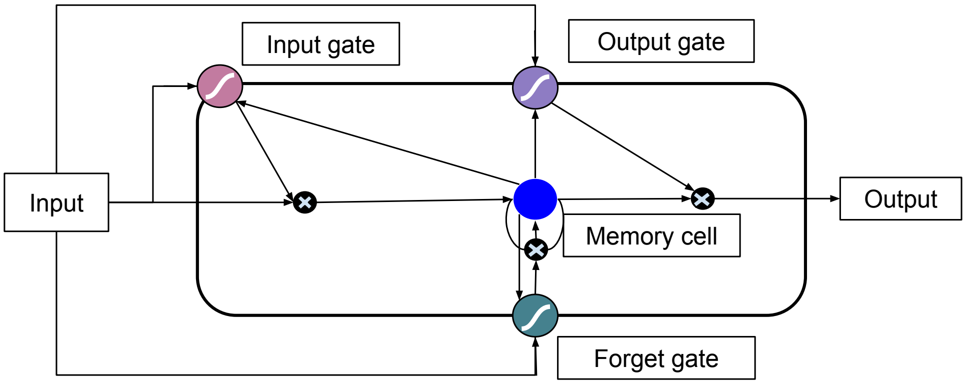


Figure 2.5: Internal design of an LSTM cell with three input, output, and forget gates.

studies in healthcare data analysis can be found in [75].

Recurrent Neural Network (RNN)

RNNs are another family of neural networks that are aimed at capturing the temporal relation of input data and have shown success in the analysis of sequential data especially natural language [185] and time-series [222]. RNNs can keep track of previously observed samples by storing the model output for them in an internal memory and passing it as an additional input for the next sample prediction. As illustrated in Figure 2.4, the network dynamically unrolls on the arrival of new data points to generate a new output using the new input and the last memory state.

Traditional RNN architectures are shown to suffer from the vanishing gradient problem during back-propagation when working with a long sequence of data [173]. Long Short Term Memory (LSTM) [106] network is a variation of RNN that was originally proposed to tackle this challenge by introducing a new data forgetting and remembering mechanism. Each LSTM cell contains three input, output, and forget gates that jointly work to control what information should be read, stored in the internal memory, or passed to the next cell. The Figure 2.5 shows the internal architecture of one LSTM cell.

Bidirectional LSTM (BiLSTM) [235] is an extension to LSTM cells that is shown to improve the performance in many applications. BiLSTM stacks two layers of LSTM cells to process the sequential data on in forward and one reverse order. Processing the data in reverse order is shown to help the model to better retain information near the end of the sequence. In this architecture, the concatenation of the hidden states of the forward and backward layers form the representation of the sequential data.

RNN has become one of the main models of interest in the analysis of health signals and sequential readings. LSTM-based models have produced state of the art results in daily health monitoring systems, especially human activity recognition, surveyed in [220]. In the study of sleep quality, BiLSTMs have been used on polysomnography signals to detect negative respiratory events during sleep [118, 31]. In cardiovascular disease monitoring, recurrent models have been used in the detection of congestive heart failure [221] as well as classification of negative heart events from ECG signals [15], especially when RNN is used in combination with CNNs studied in [95]. Moreover, state of the art performance in cuff-less monitoring of blood pressure has been obtained by LSTM-based models applied on ECG and PPG signals [190]. In the area of diabetes management, utilizing sequential models has led to a significant improvement in forecasting the glucose level and detection of glycemic events [163]. In mental health domain, LSTM has been applied on EEG signals to detect mental disorders [51], emotions [233], as well as detection of Parkinson's disease [132].

Furthermore, sequential analysis of speech signals has led to accurate detection of stress and nervousness in recent studies such as [99, 224]. In addition, [81] and [225] used two variants of LSTM-based models to analyze an individuals interactions in social media communications to detect suicide risk and stress respectively.

Attention

Attention is a technique that allows the model to better differentiate relevant and non-relevant information and focus on the former for making a prediction. Attention was initially applied in image recognition tasks to improve the model focus for supervised prediction [231]. Later, a great success was obtained in natural language processing domain, when attention was incorporated in sequence to sequence modeling in machine-translation tasks [147, 22]. Following this success, attention has been extended to other sequential prediction tasks such as time-series analysis. [50, 149] are among studies that used attention in the processing of health time-series data. As LSTM-based models show trouble in processing long sequential data, a recent study introduced Transformer architecture [212] that replaces sequential processing modules such as LSTMs with self-attention mechanism, so that long data history can be processed based on importance rather than temporal relation. Although this approach is proposed for the natural language processing domain, it has been recently used for processing other sequential data including historical health data [200].

Following the generic view of attention suggested in [212], attention is defined as a function between a query (q) and a set of keys (K), denoted as $f : q \times K \rightarrow [0, 1]$, in which each key is mapped to an importance score between zero and one, with respect to q. Moreover, it is commonly expected to have: $\sum_{k \in K} f(k, q) = 1$. The function f has been defined in various ways in past studies. [92] use a cosine similarity score among the feature vectors of k and q while [22] concatenates them and feed them into a MLP layer. Taking the dot product of q and k feature vectors is another approach used in [147], which is then modified in [212] to scaled dot product defined as $\frac{q\dot{k}}{\sqrt{n}}$ where n is the dimension of q and k feature vectors.

Other than performance improvements, attention offers the advantage of improved inter-

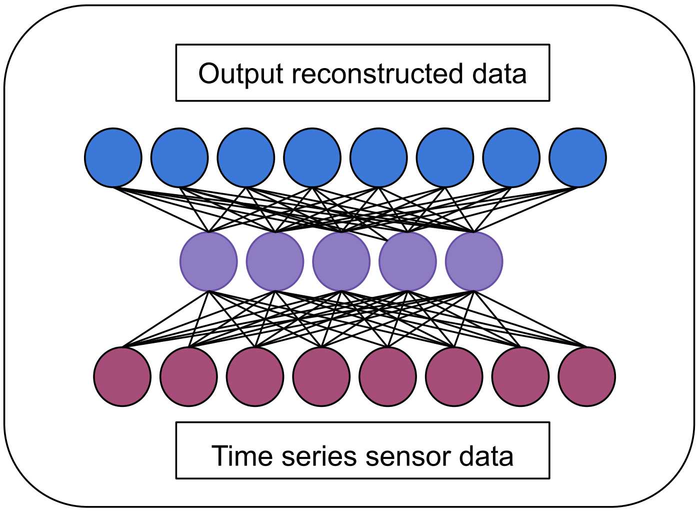


Figure 2.6: General structure of an autoencoder

pretability. The attention scores calculated for different keys can add transparency to the model decision process and explain what features have contributed most to the prediction.

Unsupervised Representation Learning

Although pervasive health systems can generate an enormous amount of sensor data, majority of it remains unlabeled, mainly because annotation process is costly, requires real-time expert knowledge, and high adherence of users to report changes in context and symptoms at the time of occurrence. Recognizing this challenge, there is an emerging trend of adopting unsupervised learning models in the analysis of healthcare data. Here we review two of the most commonly used models autoencoder and skip-gram based models.

Autoencoder networks are a family of unsupervised representation learning models. Autoencoders commonly consist of an encoder that embeds data into a compact feature vector and a decoder aimed at using the learned embedding for reconstructing the input. The model is then trained in an end to end manner with the objective of minimizing the reconstruction error. Figure 2.6 illustrates the general schema of an autoencoder. In the design of these networks, some training constraints are commonly introduced to drive extraction of representative features of the signal rather than learning an identity mapping [142]. The most common constraint used is adopting a smaller size for embedding layer compared to input and output to act as a bottleneck. The learned representations by the autoencoder can be used as an informative and compact encoding of data for further classification or clustering tasks [142].

Autoencoders have yielded new state of the art results in various domains including image-processing [141] and their application have been extended to other domains including speech recognition [76] and video processing [202]. Following their initial success, several variants of autoencoders were later introduced to improve the feature embedding quality and robustness. Sparse autoencoders [165] add increase training constraints by adding KL divergence penalty term to the cost function. This term ensures that most of the activation units in the hidden layer remain close to zero, leading to more representative extraction of feature in active units. Denoising autoencoders [216] are another group that attempts to reconstruct the data while removing the noise. For this aim, a certain amount of noisy data is intentionally added to the input. The model is then trained to reconstruct the original data from the noisy input. In addition to exploring new training constraints, various encoder and decoder architecture have been used as well. LSTM encoder-decoder is among successful architectures that were originally introduced in improving video representation learning [202]. In this model, an LSTM layer is used to encode an input sequence into a representation vector and another LSTM layer is used to decode it back to the original sequence. This approach was later used in a number of time-series analysis applications, mainly due to its capability of capturing temporal features of the signal as well [151, 218].

Applications of autoencoder have been recently studied in the pervasive health domain. Denoising autoencoders have been employed to approach the problem of missing value imputation [122, 229] which is a familiar challenge in sensor-data processing. Another application of unsupervised representation learning is the clustering of individuals or diseases based on their health data [160, 28]. Anomaly detection is another important application of autoencoders that have recently gained considerable attention in applications such as detection of anomalies in ECG signals [151] and detection depression [85]. Anomalies in these studies are generally identified by analyzing the reconstruction error in the time of prediction.

Skip-gram is another unsupervised learning technique originally introduced in natural language processing to learn word representations [158]. In this approach, feature vectors of input data are learned in a way that similar points reside in close proximity in the new feature space. The similarity is commonly defined as a neighboring relation. Skip-gram can be especially beneficial in the healthcare domain where similarity and co-occurrence of various data points can play a major role in various prediction tasks. Recently, a number of studies have combined autoencoders and skip-gram model to extract a dictionary of signal segments and learn their representations so that neighborhood relationship among segments is preserved [236, 232].

2.3 Interpreting Electronic Health Records

Electronic Health Records (EHR) provide detailed documented information on various clinical events that occur during a patient's stay in the hospital. Laboratory tests, medications, nurse notes, and diagnoses are examples of heterogeneous types of clinical records. EHR can play a major role in painting a more comprehensive picture of someone's health history and state and its availability in recent years has opened great opportunities for researchers to further explore computer-aided advancements in the healthcare domain. Variety of clinical prediction tasks have been studied on EHR including prediction of mortality, diagnosis, length of stay, and readmission [177, 102]. Regardless of the task at hand, there are a number of inherent challenges in EHR data modeling that prior studies have approached in different ways.

First, clinical records are heterogeneous and are sparsely distributed among patients. To tackle this, manual feature selection has been a method of choice in many studies [223, 201, 139, 46]. However, feature engineering for the clinical domain is a tedious task and requires expert knowledge. Moreover, missing values in EHR pose a great challenge to feature extraction [30]. Two recent studies have worked on addressing these by benchmarking the MIMIC III [125] EHR dataset and introducing sets of features to be manualy extracted for various prediction tasks [102, 177].

Recognizing the discussed challenges of manual feature extraction, recent studies have shifted towards unsupervised embedding of clinical records [90, 52, 74]. The main idea behind many unsupervised EHR embedding studies is inspired by the success of word2vec models [158, 159] in natural language processing. Earlier studies turned the clinical event records into an ordered sequence of words and employed skip-gram [159] to capture co-occurrence of clinical events and learn latent embeddings [53, 62]. To expand the idea, Med2vec [48] proposed a multilayer representation learning model for clinical code and visit embedding which uses skip-gram for initializing the representations but tailors them through the network training process. Similar EHR embedding approaches has been adopted by a number of later studies such as [149, 49, 50].

Another challenge in mining EHR data is modeling the complex structure and relations in EHRs. Recurrent Neural Networks (RNNs) have been one of the most widely adopted techniques, especially for sequential prediction tasks, such as prediction of future clinician visit based on admission history, in studies such as Dipole [149] and RETAIN [50]. However, for more complex tasks such as diagnosis prediction based on patients' clinical records, RNNs have some shortcomings. Firstly, RNNs lose efficiency and performance when working on long sequences, and clinical records may contain thousands of items. SAnD [200] is a recent study that tackles this loss of performance by using multi-head self-attention rather than an RNN. Secondly, relations among health records have various types and semantics and are not always sequential. GRAM [49] is a parallel study to above ones showing the promise of modeling EHRs as graphs, however, its design is very limited and only applicable on hierarchical medical codes. Observing these shortcomings, there is still a strong need for an integrative and structure-aware modeling of EHR data.

2.4 Asthma Monitoring

In this section, we provide a background on asthma control factors and exacerbation triggers and survey the prior mHealth studies on the improvement of asthma management.

2.4.1 Asthma Control and Triggers

Asthma is characterized by recurring attacks in the form of wheezing and breathlessness that can vary in frequency and intensity. The severity of an individual's asthma is measured by a combination of features, including the frequency of symptoms, medication requirements, reduced flow rates, and functional limitations. The individual is then assigned to one of 4 severity groups of intermittent, mild, moderate, and severe [169]. However, asthma and its symptoms can be controlled so that a patient with severe but well-controlled asthma can show symptoms of an intrinsically mild asthma [58]. Therefore, asthma is often described based on the control level that is evaluated by a number of impairment and risk components. Table 2.1 lists the main impairment metrics associated with three asthma control classes for children of age 5 or older. The full information on control factors and severity classifications of asthma can be found in guidelines for the diagnosis and management of asthma by National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health [169].

Although levels of asthma control are dependent on the underlying severity and responsiveness to treatment, proper management with high adherence to medication and care towards environmental and physiological triggers can play an important role in improving asthma control [237]. Variety of factors has been shown to trigger episodes of asthma exacerbation and reduce control. These factors include exposure to allergens (e.g., pollen, dust), cold weather, intense physical activity, extreme emotional arousal, and ambient air pollutants [215]. Contributing pollutants to asthma exacerbation, especially in children, are known to be ozone [140], particle matters with a diameter less than 2.5 (PM2.5) [84] and 10 (PM10) [65], smoke (carbon) [203] and No2 [43]. Prior studies have shown that these pollutants can increase significantly through basic activities such as cooking and vacuum-

Table 2.1: Asthma impairment assessment for different control levels

Control Impairment Components	**Age**	**Levels of Asthma Control**		
		Well Controlled	Not Well controlled	Very poorly controlled
Symptoms	5-11	2 days/week but 1x/day	$>$ 2 days/week or multiple times on 2 days/week	Throughout the day
	$>$ 12	$\leq$ 2 days/week	$>$ 2 days/week	
Interference with activity	All	None	Some limitation	Extremely limited
Nighttime awakenings	5-11	$<$ 1x/month	$>$ 2x/month	$>$ 2x/week
	$>$ 12	$\leq$ 2x/month	13x/week	$\geq$ 4x/week
Lung function (PEF,FEV$_1$)	$\geq$5	$>$ 80%	60-80%	$<$ 60%
Rescue medication use	All	$\leq$ 2 days/week	$>$ 2 days/week	Several times per day

ing [107, 119]. Furthermore, asthma triggers vary considerably from person to person, adding to the complexity of proper asthma management.

2.4.2 Pervasive Health Solutions

Recognizing the significance of proper asthma management on patients and their families quality of life, there has been a long interest in asthma control improvement through remote education and monitoring. [133, 162] are among early telehealth studies showing the positive impact of digital education and remote intervention in asthma control. Following this trend, early mHealth solutions were designed with the aim of improving medication adherence and symptom recording , however, they mostly relied on self-reports and diaries [108, 166, 217, 78] as well as reminders and alerts [148, 80]. Results of a number of randomized controlled trials reported improved control and a decrease in hospital visits through these approaches [157, 179, 37]. However, others showed that adherence to these solutions drops dramatically in a course of one year, attributing that to the demanding design of self-reporting systems [181].

Later, with the prevalence of ubiquitous computing, a number of wearable sensors were introduced for improved monitoring of asthma. These sensors include Respiri [38] that can measure and detect wheeze as a sign of airway obstruction, and Propeller Health sensors [226] that can be attached to asthma inhalers to record geographical and temporal information of medication use. Moreover, personal spirometers provide the opportunity for more comprehensive monitoring of lung function at home. Spirometry is the most common pulmonary function test used in asthma that allows assessment of airways and lung functionality [56] and is reported by two measure of: Forced expiratory volume in one second (FEV1) and Peak expiratory flow (PEF).

Recently, a number of pervasive health solutions have adopted the aforementioned sensors and other data sources for achieving enhanced asthma management and monitoring. AshtmaHealthApp (AHA) [39, 40] was introduced after Apple launched the health ResearchKit [123] and combines regional air quality tracking using phone GPS with self-reports. The current system design of AHA is very limited in terms of Asthma trigger and context monitoring. Khealth [13, 214] is another solution that combines public information sources (e.g., regional air quality) with local sensor-based data. Khealth shows limitation in the range and quality of variables it can monitor as in many scenarios it relies on the public information sources that do not offer desirable accuracy and reliability. Furthermore, it does not support the real-time assessment of an individual's asthma state in order to learn personalized triggers and reactions. Overall, systems presented above mostly deliver a report of collected information rather than building predictive models for detection of asthma exacerbation.

CHAPTER 3

An Unsupervised Approach to Ahead of Time Detection of Negative Health Events

Recognizing the sparsity of annotations in pervasive health and with the emergence of successful unsupervised prediction techniques in machine learning, special attention has been recently directed toward unsupervised and real-time detection of health anomalies. While anomaly detection happens at the time or after an event onset, in many applications, such as asthma control, there is a significant advantage in the early prediction of a forthcoming negative health event. Moreover, despite the commonly sharp changes of health signals on or after an onset, early changes before an upcoming negative event can be very subtle and gradual. In this chapter, we propose an unsupervised model for uncovering the early and complex changes prior to a negative health event. We first demonstrate our observations on the shortcomings of the widely adopted anomaly detection methods in this setting. Next, we propose a sequential autoencoder for the suitable embedding of the time and frequency domain features of health signals, followed by a noise-tolerant online clustering module for detection of abnormal signal behavior. We show how the framework can be extended to semi-supervised setting and benchmark our results on a publicly available heart ECG signal dataset which is collected from 10 infants experiencing Bradycardia events.

3.1 Introduction

Many health events such as asthma attacks, seizures, or cardiac arrests can be characterized by the changes they introduce to the physiological signals such as ECG and PPG and provide an opportunity for real-time analysis of such events. However, lack of annotations and interpersonal differences in health signals can hinder the use of established supervised training models, hence, there has been a recent growth of research interest in the use of unsupervised approaches.

Unsupervised models, especially autoencoders [23], have recently brought great success in learning an informative representation of data for improving further supervised or unsupervised tasks. When it comes to time series analysis, one important application of these representations has been anomaly detection [186] which has been successfully used in applications such as cyber-attack and financial loss surveyed in [96]. This success has been recently expanded to health care domain as well, especially in the analysis of ECG signals [151].

In spite of the recent application of unsupervised anomaly detection models on health signals, there has been a missing focus on the early prediction of forthcoming negative health events. Analysis of physiological and environmental sensor data can reveal early signs of a negative health event in the near future and a number of studies have already validated this hypothesis in case studies such as Bradycardia events (sudden heart rate drop) in infants [82]. Moreover, the commonly used anomaly detection models, cannot fully address the special challenges inherent in healthcare. Prior studies generally rely on the sudden increase of reconstruction error of an autoencoder trained on the normal signal to detect anomalies. The hypothesis in this approach is that such a model cannot reconstruct anomalous data points accurately. However, this approach works best when changes are abrupt and may lead to poor performance in addressing many health problems in which changes are gradual, complex, and subtle.

In this study, we draw the attention to the problem of early event detection and propose an unsupervised framework based on deep sequential autoencoders and online clustering of the internal representations to overcome shortcomings of prior methods. For this aim, we employ an LSTM encoder-decoder based architecture [151] to extract temporal and local features of the signal. We use unit-ball regularization of the learned representations [21] to optimize the results of further clustering phase. Furthermore, we use the wavelet transform [176] as a pre-processing step of the signal to better capture the time and frequency domain information when training the autoencoder model. We then analyze the clusters formed by the representation of signal segments using a variant of Denstream [34], an online and noise-tolerant clustering method. At last, we show how the model can be employed in a semi-supervised manner in scenarios when a sparse set of labeled samples are present. We evaluate our model on the publicly available Preterm Infant Cardio-Respiratory Signals Database (PICS) [175] which provides ECG signal prior to Bradycardia heart events in infants. As heart rate signals inherently show more complicated behavior compared to many other sensor data, this dataset is desirable to surface the advantages of our approach.

3.1.1 Contributions

The contributions of this study are: 1) we highlight the problem of in-advance detection of negative health events and propose an unsupervised and robust framework relying on real-time clustering of time-series representations. 2) We reveal the shortcomings of existing models relying on signal reconstruction error measurement. 3) We demonstrate the effectiveness of our approach in achieving a decent performance compared to prior supervised models by predicting Bradycardia events in infants 1.3 minutes ahead of time with 68% AUC score on average, using no supervision. 4) Finally, we demonstrate that the prediction performance can be further improved by incorporating a small set of labeled samples using the semi-supervised variant of our proposed framework.

3.2 Background

3.2.1 Autoencoders and Anomaly Detection

One important application of Autoencoders in time-series analysis has been anomaly detection [186]. The general workflow of anomaly detection using autoencoders is depicted in Figure 3.1. The process starts with training the model on normal segments of the time-series data. The model is expected to generate a compact representation of data and learn to reconstruct each segment from that. Having a model trained in this way, a sudden abnormal change in the input data is shown to increase the reconstruction error of the model at the time of prediction. This increase in discrepancy between the input signal and its reconstruction by the model is commonly thresholded to detect anomalies [186].

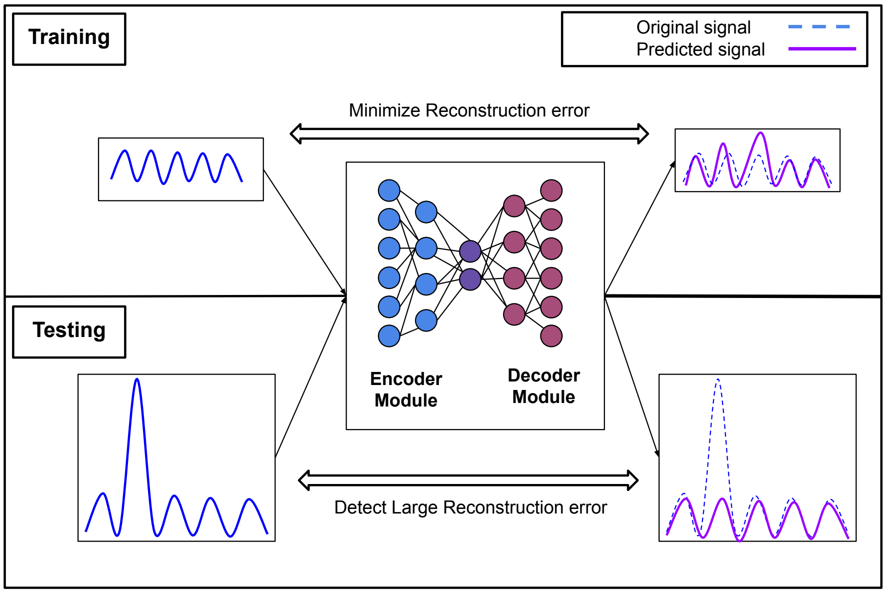


Figure 3.1: The general schema of anomaly detection using autoencoders

Two recent studies have shown improved performance of anomaly detection in more complex settings by utilizing the encoded representations from autoencoders rather than relying on the reconstruction error. [21] showed that the clustering of these representations can improve the results when the original normal data comes from multiple classes. This study, however, is not designed for time-series analysis and does not consider the real-time prediction setting of wireless health data. Moreover, [136] is another study on time series data but targeting the different task of data segmentation. They show the changes in internal representations in neighbor segments are an important sign for point of change detection in time series data.

3.2.2 Anomaly Detection in Healthcare

Many of the initially proposed anomaly detection models in healthcare relied on supervised prediction models [41]. However, as these models require large enough samples from abnormal data points, later studies shifted to more unsupervised models. Clustering of data points based on hand-engineered features for detection of outliers [178] and building future value prediction models for discovering unexpected readings [101] are among common approaches reviewed in [230]. Shortcomings of these methods, especially lack of generalizability, were later addressed with autoencoder based anomaly detection models. [151, 44, 152] are among studies that employed autoencoders on ECG signal to distinguish anomalous parts from the healthy ones. They all used reconstruction error to distinguish anomalous segments from normal ones.

3.2.3 Prior studies on PICS dataset

Prediction of Bradycardia in infants using the PICS dataset was approached before by publishers of the dataset with statistical methods [82]. They specifically used a point process analysis and tried to capture the differences in variance and mean of signal segments before a Bradycardia event. Although this study proves the feasibility of the task and achieves reasonable accuracy, their approach is supervised, hand-engineered, and heavily relies on the observance of multiple onsets of Bradycardia events in each infant, which is not always possible in the real-world setting. This is while our approach focuses on the straightforward collection of normal signals from individuals and the detection of changes in an unsupervised and automatic manner.

3.3 Methodology

In this section, we provide details of the two main components of our proposed model: sequential representation learning and real-time clustering.

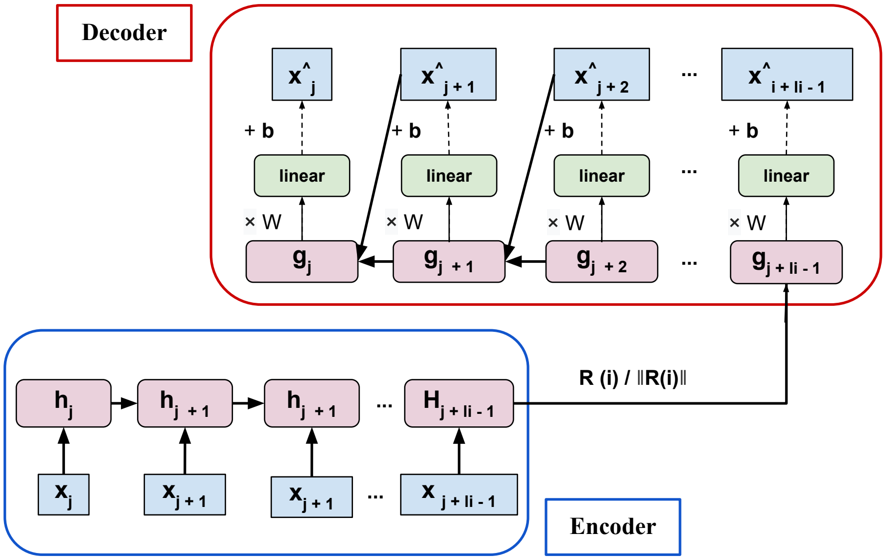


Figure 3.2: Design of the LSTM autoencoder with unit-ball regularization.

3.3.1 Sequential Representation Learning

Extracting representative features of a signal using unsupervised models such as autoencoders can provide beneficial insights into the detection of abnormal signal behavior. Moreover, temporal dependencies in health signals often carry important information regarding gradual changes. Therefore, we employ a deep sequential autoencoder to extract temporal and local features of health signal segments. In particular, we use a variant of LSTM encoder-decoder architecture similar to [151], design of which is discussed below.

Given a time series T, it is segmented into fixed or variable length windows denoted as $W = \{w_1, w_2, w_3, \cdots, w_n\}$ where each w_i is itself a list of readings of length l_i, hence w_i can be represented as: $w_i = \{x_j, x_{j+1}, \cdots, x_{j+l_i-1}\}$. The model first embeds w_i into a fixed length representation by feeding it into an encoder module, an LSTM based recurrent neural network (RNN) with l_i cells. The hidden state of the last (l_ith) cell can be considered as a compact and informative representation of w_i, which we call it $R(i)$. To make these representations more suitable for comparison using distance-based metrics and remove the impact of representation length, we apply $l2$ normalization on $R(i)$ in the training phase and before feeding to the decoder as suggested by [21].

$$R^*(i) = \frac{R(i)}{\|R(i)\|} \tag{3.1}$$

The decoder module that is aimed to reconstructs window w_i from $R^*(i)$ is also an LSTM RNN with a linear transformation layer on the output gates. The decoder uses $R^*(i)$ as the initial state to its first LSTM cell. Afterward, the output of each cell in the decoder, which represents the reconstructed value of the corresponding input by the model, is used as an input to the next decoding cell. Following the findings of prior studies on improved optimization of encoder-decoder architectures [202], our model is designed to predict each window w_i in reverse order. Figure 3.2 depicts the design of the encoder and decoder modules.

Considering decoder module as a function D, $D(R^*(i))$ denotes the output of the model for window w_i. When reconstruction error between $D(R^*(i))$ and input w_i is used as the objective function and both modules are jointly trained on normal intervals of data, the model learns to embed representative features of the input window w_i into $R^*(i)$. Therefore, the objective function can be written as:

$$L = \frac{1}{\|W\|} \sum_{w_i \in W} \frac{1}{l_i} \sum_{j=1}^{l_i} \|x_j - \hat{x_j}\|^2 \tag{3.2}$$

where x_j denotes the original reading and $\hat{x_j}$ denotes the corresponding prediction. We train the proposed autoencoder on event-free intervals to achieve representative features of a normal signal behavior.

3.3.2 Online Clustering

The sequential autoencoder is hypothesized to map similar windows into close feature vectors. Therefore, those related to the normal signal behavior may form one or more clusters in the feature space. This is while, any deviation from the norm in a signal window, would reflect in the features and increase the distance from those clusters. Therefore, a clustering approach is desirable to keep track of the clusters formed by normal windows and detect any abnormal outlier or cluster formation over time. Furthermore, outlier detection should be robust to noise and concept drift over time. To meets these needs, we employ an approach similar to Denstream [34], an online and noise-tolerant clustering model.

Training Phase

After training a subject-specific autoencoder, we feed features extracted from a set of normal training windows to find the main normal clusters (C). Similar to Denstream, our clustering model relies on finding and connecting the dense gathering of data points called micro-clusters. Each micro-cluster is represented by the two features of radius and weight. At each point in time, a micro-cluster is denoted as ”core” if its weight is more than the weight threshold (μ) and an outlier otherwise. On arrival of each data point, it is added to closest micro-cluster, with a priority on core ones, while keeping its radius below the threshold (ϵ). All micro-clusters evolve as the new data points arrive and outlier and core micro-clusters get updated. Once all windows are processed, we apply the DBSCAN [72] algorithm to connect overlapping core micro-clusters and form the final main clusters in C. Algorithm 1 provides detailed description of our clustering algorithm in the training phase.

Testing Phase

In the test time, as the incoming signal windows are encoded and fed into the clustering module in real-time, micro-clusters are updated and DBSCAN is used to generate the main clusters. An increased appearance or growth of clusters other than C in a short amount of time can be considered as an abnormality and a sign for a possible event onset. It is worth mentioning that as our clustering algorithm ignores outlier micro-clusters, any newly formed cluster is dense enough to show a real change in the signal behavior.

To robustly use the abnormal cluster formation and its growth trend in the prediction of an event we introduce ”confidence window”. In particular, We consider the last K received windows and keep track of the ratio that they join a cluster other than C. A threshold on this ratio(λ) can tune the sensitivity of the model against abnormal cluster formation. Algorithm 2 shows the process of our cluster-based real-time anomaly detection.

Algorithm 1 Training Phase Clustering

```
                                        ▷ core weight threshold $\mu$, radius threshold $\epsilon$
for a new normal signal window do
    compute the representation $r$
    try to merge $r$ to closest core micro-cluster
    if new radius $< \epsilon$ then complete the merge and record the new weight
    else try to merge $r$ to closest outlier micro-cluster
        if new radius $< \epsilon$ then complete the merge and and record the new weight
        else create new outlier micro-cluster with $r$ and record the new weight
        end if
        if new weight $> \mu$ then change the micro-cluster from outlier to core
        end if
    end if
end for
Use DBSCAN algorithm to find the main clusters of a normal signal ($C$).
```

Algorithm 2 Testing Phase Clustering

```
                                        ▷ Confidence Window length $K$, ratio threshold $\lambda$
cnt = 0
for a new test window do
    compute the representation $r$
    merge $r$ into the suitable core or outlier micro-cluster
    Use DBSCAN algorithm to find the clusters of data.
    if $r$ joined a cluster other than $C$ then $cnt++$
        if $\frac{cnt}{K} > \lambda$ then Trigger the alarm for a negative event
        end if
    else update cnt to include appearance of new clusters in the last $K$ windows
    end if
end for
```

3.3.3 Semi-supervised Variation of the Proposed Model

Our proposed algorithm is designed with the aim of prediction in the absence of labeled data, which is a known challenge in the healthcare domain. However, when a sparse set of labels are available, it is desirable to incorporate them into the prediction model. To this end, we employ a cluster tagging approach. Recalling the clustering phase in the training stage, on arrival of each window, if the window is labeled we use that to tag the assigned cluster as normal or abnormal. The cluster tag is determined by the majority vote of labeled samples inside that cluster. Different from the original approach which density of a cluster is considered as the main feature of a normal cluster, in this approach can detect sparse but normal clusters of signal and avoid false-positive triggering of alarms.

3.4 Experiments

In this section, we report the results of our experiments for evaluating our proposed approach in the detection of future negative events. The section starts with a general overview of the infant heart signal dataset used, followed by implementation details, experiment setup, and results.

3.4.1 Dataset and Data Processing

We benchmark our model on the Preterm Infant Cardio-Respiratory Signals Database (PICS), publicly available on [175]. This dataset contains 20 to 70 hours of ECG recordings of 10 infants with multiple onsets of Bradycardia events, in which an infant's heart rate stays below 100 beats per minute for at least two beats.

In this study, we rely on heart rate variability signal rather than the raw ECG, following the large body of research showing its power in indicating negative heart events [73, 131]. Heart rate variability is generated by taking the time difference between R-peaks in the original ECG signal, which is an indicator of how fast or slow the heart beats.

A number of studies have already shown the effectiveness of frequency domain analysis (along with the time domain) on improving the predictive power of heart rate variability signal [176]. Having this in mind, we employ Continuous Wavelet Transform (CWT) that permits simultaneous access to the time and frequency domain information at each point of time. As suggested by prior research [9], we use Morlet function as the mother wavelet in our CWT analysis and study the obtained signals on the low-frequency band (0.01-0.15 HZ), relying on initial results reported in [82].

3.4.2 Evaluation Measure

We evaluate our model by scoring the rate of true-positive alarms (recall), true negatives decisions (specificity), Area Under the Receiver Operating Characteristic curve (AUC), and earliest prediction time on true-positive alarms.

An alarm is considered as true positive if a negative event happens within a 3-minute time span. Moreover, 6 minutes after each onset is disregarded in evaluation to ensure the effects of the last event has passed. As a result, a true negative happens when no alarm is generated from 6 minutes after to 3 minutes before two consecutive events. The used time ranges are borrowed from an initial study on this database [82]. In all experiments we use 70% of data of each infant for training and the remaining 30% for testing.

3.4.3 Model Hyper-parameters

In the training of the autoencoder, we used window sizes of length 64 to processes heartbeat signal and hidden state size of 20 for the LSTM cells. In the clustering phase, radius threshold and minimum weight were set to 0.02 and 2 respectively. Moreover, we used the ratio threshold of 50% when using confidence windows in generation of event alarms.

3.4.4 Results

As tracking reconstruction error has been used as a common approach in detection of anomalies, we first share results of the performance of this approach in detection of early changes in the heart-rate signal before an event onset. Figure 3.3 illustrates an interval of heart-rate monitoring of infant 7 in which it shows multiple Bradycardia events. The figure also reports the corresponding reconstruction error computed by the autoencoder introduced in section 3.3.1. As it can be inferred, reconstruction error shows significant jumps at the time of events when the heart rate experience a sudden drop and other sharp changes. However, as discussed in [82] and observable here, the variability in heart rate, which translates into a similar variability in the reconstruction error, does not show a simple pattern before a Bradycardia onset. Therefore, although reconstruction error achieves good performance in unsupervised detection of sudden changes, it can perform poorly for prediction of forthcoming events, mainly due to the more complex nature of this task. This experiment validates our approach in employing deeper features of signal for uncovering the hidden changes before a negative event.

We next evaluate our proposed model qualitatively and quantitatively. Figure 3.4 visualizes a snapshot of the real-time trend of cluster formation by our method before three event onsets of infant 7. Each cluster is depicted by a unique color and level in the y-axis and forms as we move in time (on the x-axis) and process incoming signal windows. Moreover, clusters appearing in a 3-minute time span before an event onset (depicted by yellow triangles) are shown with a cross mark for easier readability. As the figure suggests, the two blue clusters appearing most often in the snapshot, especially in the times away from an event onset, represent the clusters corresponding to the normal behavior of data. The noisy samples that do not fit into any core micro-clusters are not represented in this figure. We can observe the sudden appearance of numerous new clusters in the 3 minute time spans before the onset of each event. This increase in the number of newly formed clusters is visibly a powerful sign for detection of an onset. Furthermore, it is noticeable that abnormal clusters appear far more sparsely at normal intervals of data. Confidence windows introduced in section 3.3.2

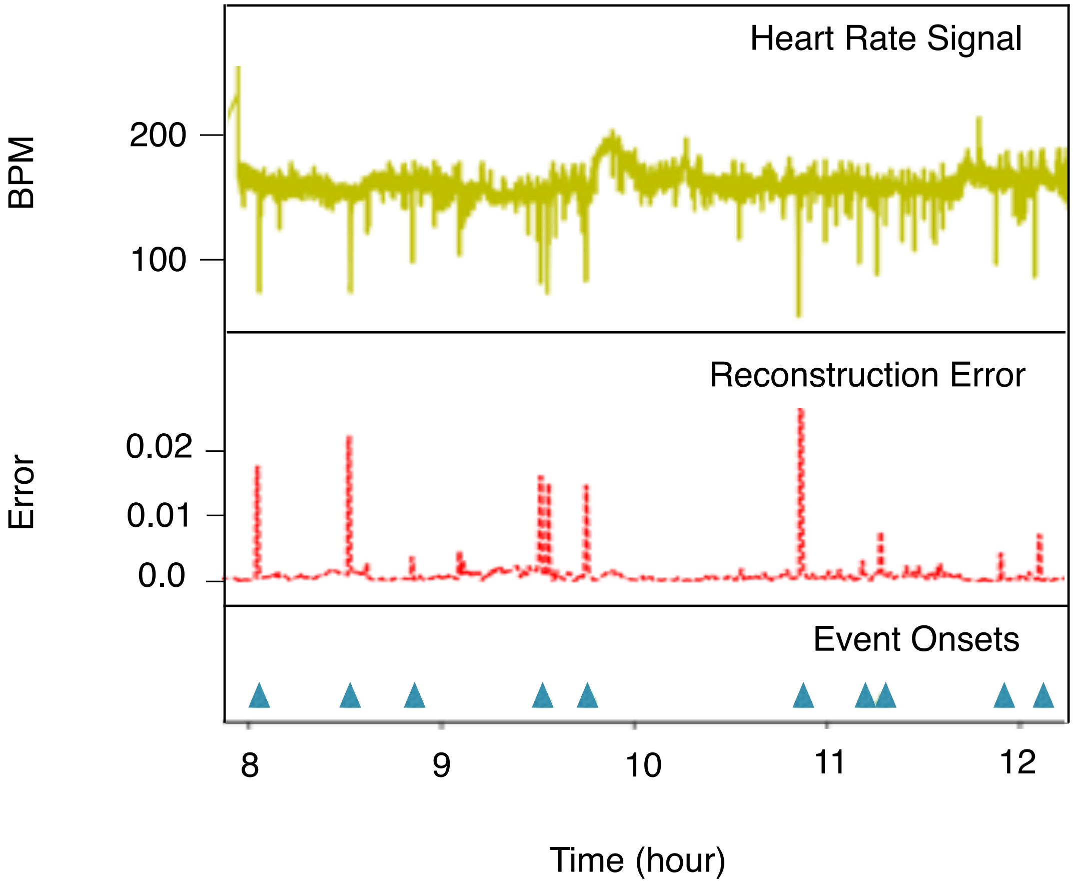


Figure 3.3: Correlation of heart rate variability signal and reconstruction error in a time span containing multiple onsets of Bradycardia related to infant 7.

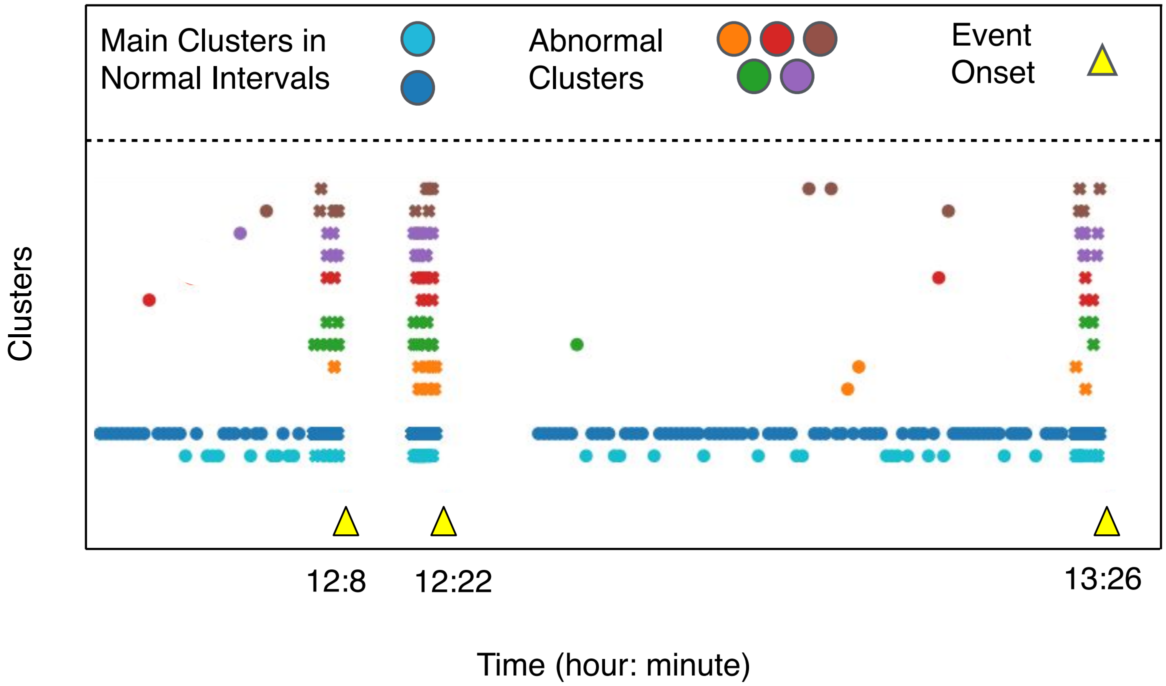


Figure 3.4: Example of online clustering in an interval of data from infant 7.

help in tuning the sensitivity of our model to these appearances.

The next experiment is used to analyze the impact of confidence window size (K) on the performance of our model and its results are outlined in Figure 3.5. In general, as we increase K, the earliest time to prediction decreases as the model has to wait for k windows to decide whether to generate an alarm or not. Moreover, with an increase of K the recall goes down and the specificity increases, meaning that the number of false alarms decreases, however, it comes with the cost of losing detection of some events. If we observe closer, for $K > 2$ to $K < 7$ we can see a stable performance. This is because having $K > 2$ ensures that a single appearance of an abnormal window does not generate an alarm. Furthermore, $K < 7$ corresponds to around 2 minutes before an event onset (recognizing the size of windows and their overlap used), which coincide with the time the main changes in signal is expected to happen [82]. We can also observe that AUC score is pretty stable as this metric is not dependent on our cut-off threshold (50% abnormal observations in a confidence window) and mainly measures how well our model can assign distinguishable scores to positive and negative labels.

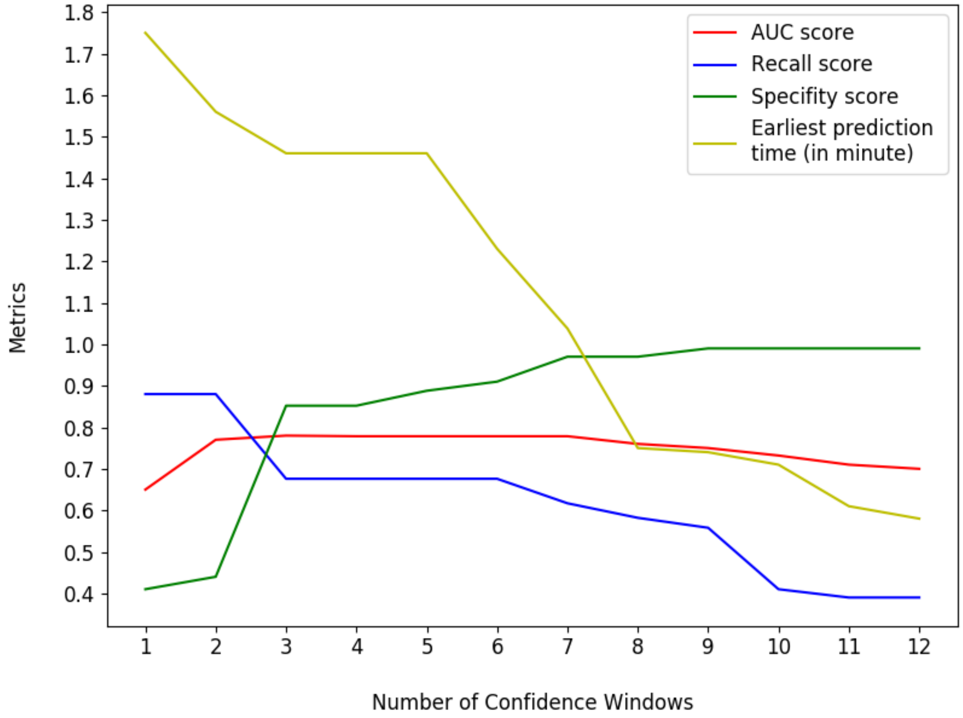


Figure 3.5: Comparison of evaluation metrics using different confidence windows

Table 3.1 reports AUC score and earliest time of prediction of our model using confidence window of size 5 for all 10 infants, achieving an average of 67.7% AUC and 1.29 minutes time difference between model alarm trigger and actual event onset. As expected, the results from our unsupervised model are lower than the supervised model introduced in an initial study of this dataset [82], achieving an average AUC score of 79%. However, our model shows the possibility of achieving competitive results with an unsupervised approach which is advantageous in the healthcare setting that labels are generally missing. This is while the supervised model relies on the observance of multiple negative events per infant in the training phase.

In the last experiment, we study the semi-supervised variant of our model and its impact on anomaly detection performance. To this end, we consider the addition of 10, 40, and 70 percent of labeled data to our model in the training stage and we perform the testing on the last 30% chunk of data as before. Mean AUC scores achieved from these models are reported

Table 3.1: Results of our unsupervised model for all the subjects

Infant Id	1	2	3	4	5
AUC score	0.71	0.65	0.61	0.68	0.63
Forecast time to onset (min)	0.87	1.31	1.24	1.42	1.28
Infant Id	6	7	8	9	10
AUC score	0.69	0.77	0.67	0.63	0.72
Forecast time to onset (min)	1.34	1.46	1.42	1.33	1.23

Table 3.2: Comparison of results from unsupervised and semi-supervised models with various labeled samples percentages.

Percentage of labeled samples	0%	10 %	40%	70%
Mean AUC score	0.68	0.73	0.75	0.76

in Table 3.2. As it can be inferred, the model is able to increase the prediction performance by 5% after observing only 10% of available labeled data. We can also observe that the performance gain between the scenarios when 40% and 70% of labeled data is available decreases significantly. These results suggest that few labeled samples can effectively guide the model toward the correct tagging of clusters.

3.5 Conclusion

In this study, we presented a framework for early detection of negative events from health signals. There are a number of improvements and extensions that can be considered when applying this framework to larger scale problems and for longer periods of time. Firstly, as the normal behavior of signal can change and drift over time, it is necessary to update clusters related to the normal signal behavior in a timely manner. For this aim, clusters can be regularly updated in an offline manner by recording the latest event-free intervals of signal or be updated online in intervals when no negative event happens. Moreover, in this study, we have not used the time decay on micro-clusters. This feature can be beneficial in capturing the time-related changes in the signal as well. Finally, the real-time clustering approach can be further optimized as calling the DBSCAN algorithm at each time a new point is added to the model can deliver sub-optimal performance.

CHAPTER 4

Heterogeneous Information Networks for Integrative Modeling of Health Records

Electronic Health Records(EHR) provide key information regarding an individual's health history which is essential in the design of an integrative pervasive health solution. However, EHR data is very high-dimensional, sparse, and structurally complex which pose major challenges to its modeling and integration into other systems. In this chapter, we introduce HeteroMed that represents entities and rich relationships in an EHR as a heterogeneous information network. HeteroMed automatically assigns each record to a type-aware feature embedding vector which allows for straightforward handling of missing values and heterogeneity of data. It also exploits metapaths to capture higher level and semantically important relations contributing to a specific prediction goal. Furthermore, it employs a joint embedding framework to tailor clinical event representations to the disease diagnosis goal. To evaluate the effectiveness of HeteroMed, we study it in the context of the diagnosis prediction task and show its superior performance compared to existing EHR modeling and diagnosis prediction approaches.

4.1 Introduction

Building predictive models for Electronic Health Records (EHR) requires suitable extraction of representative information available. Recognizing the high dimensionality of EHR data, commonness of missing values, and the need for extensive clinical knowledge to apply manual feature engineering, a recent group of studies have employed unsupervised representation learning of clinical events, patients, and visits [48, 52, 53, 62]. As discussed before in section 2.3, these methods that are mostly inspired by Med2vec [48], consider co-occurrence of clinical events in different patient records to extract latent embeddings of these entities. However, representations learned are general and not tailored to the prediction goal. More importantly, none of the above-mentioned approaches can capture the rich structure of EHR data and semantics of various relations it contains. This information makes a great contribution to the prediction task and a desirable model should be able to capture and reflect it into the learned representations. Therefore, any adopted EHR modeling approach should achieve two main goals:

- properly capture structure of EHR data and semantic of relations; and
- learn representations suitable for the prediction goal.

To address these requirements and shortcomings of prior models, we propose an EHR model based on Heterogeneous Information Network (HIN) [100]. These information networks that offer definition of various types of nodes and relations, have gained lots of attention in recent years due to their ability in distinguishing and learning the different semantics of relations among entities [193] and can be profoundly beneficial to better express the rich network of patients and clinical events in the EHR data.

We demonstrate how EHR can be translated into an HIN and introduce our node extraction strategies from different formats of data (e.g., raw text, numerical, categorical) present in EHR. We further show benefits of employing metapaths [100] to introduce composite relation semantics into our network and capture the informative ones for the prediction task. This allows our model to learn similarity of clinical events and patients with respect to semantically important paths rather than random sampling strategy used in prior skip-gram based diagnosis studies. To train the model, we employ an unsupervised heterogeneous embedding model [66] to learn representations of clinical events of the network and further tailor latent embeddings to diagnosis prediction goal by jointly training a supervised embedding model similar to [47]. While our diagnosis prediction model only utilizes diagnostic information for reasoning the disease, we propose exploiting the treatment information at the time of unsupervised representation learning to improve learned embeddings and capture similarity of clinical events in terms of outcome. Apart from relation-aware modeling and tailored representation learning, HINs also offer the advantage of straightforward handling of missing values, which is a common challenge in clinical data modeling.

In this study, we conduct model experiments on the task of diagnosis prediction and demonstrate that employing HIN for modeling EHR outperforms state of the art models in two levels of general disease cohort and specific diagnosis prediction. We also conduct two case studies to qualitatively reveal the strength of HeteroMed in capturing relations in clinical data which are validated by a clinician.

4.1.1 Contributions

In short, contributions of this study are: 1) We propose HeteroMed, an HIN-based medical model which captures the semantics of clinical entities and relations and learns record embeddings tailored for an aimed prediction task. To the best of our knowledge, this is the first study to use Heterogeneous Information Network for modeling clinical data. 2) We demonstrate how EHR data can be translated into an HIN to achieve optimal learning power. 3) We empirically show superior performance of HeteroMed compared to prior methods in the diagnosis prediction task at two levels of exact diagnosis code and general disease cohort. 4) We qualitatively demonstrate that HeteroMed outperforms the baseline models in capturing relations and similarities of clinical events.

4.2 Background

Heterogeneous Information Networks [100] are different from homogeneous ones in their ability of representing multiple types of nodes and relations. This capability has attracted lots of attention in different applications such as personalized recommendation [194] and malware detection [116]. Due to the large size of real-world networks and sparsity of data, representing nodes as a low-dimensional vectors is a widely adopted approach in network mining. Network representation learning techniques in general are inspired by word2vec [158], among which DeepWalk [174], LINE [206], and Node2vec [94] have been utilized in many network mining researches. Recent studies have adopted similar techniques for heterogeneous network representation learning [42, 66, 47]. They include the heterogeneity of nodes in the definition of relations and neighbors. Furthermore, it is demonstrated in [47] that a joint embedding approach in heterogeneous node representation learning can lead to improved supervised task performance. In this study, we employ heterogeneous network embedding alongside with the joint learning framework to learn clinical event representations.

4.3 Methodology

In this section, we first put forward the problem definition and terminology used in the study. Then, we introduce how EHR can be viewed as a heterogeneous network and discuss network construction techniques. Lastly, we discuss the training and prediction models adopted for our disease diagnosis task.

4.3.1 Problem Definition and Clinical Terminology

Each record in EHR data is conventionally called a clinical event. A clinical event e can be viewed as a triple: $e = (t, n, v)$ where t, n, and v respectively denote *type*, *name*, and *value* of it. *Glucose level of 60* is an example of a clinical event that has type = laboratory test, name = Glucose, and value = 60. Clinical events type may or may not have a value based on their type.

Furthermore, the set of clinical event types in EHR, denoted as $t_1, t_2, \ldots, t_{|T|} \in T$, can be broadly viewed as union of two main categories: $T = DIAGNOSTIC \cup TREATMENT$ where $DIAGNOSTIC = \{$*laboratory test*, *symptom*, *age*, *gender*, *ethnicity*, *microbiology test*$\}$ and $TREATMENT = \{$*prescription*, *procedure*, *diagnosis*$\}$. Diagnostic clinical events are the source of information for disease diagnosis. This is while clinical events in treatment

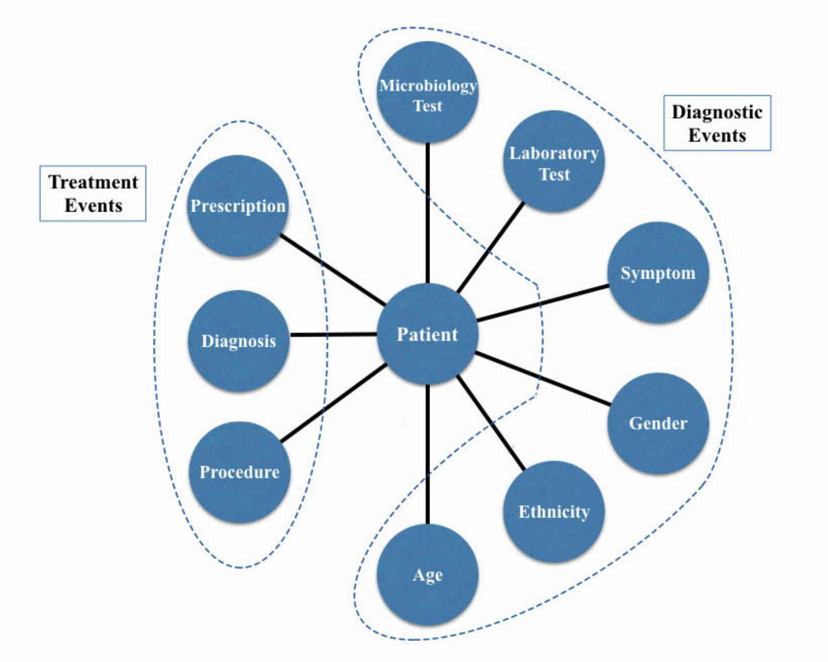


Figure 4.1: EHR heterogeneous network schema.

category happen after the diagnostic process and should not be directly used for diagnosis prediction.

We represent clinical events for a patient p as $E(p) = \{E_1(p), \ldots, E_T(p)\}$ where $E_t(p)$ denotes all clinical events of type t recorded for p. Consequently, the problem of disease diagnosis is formulated as prediction of p's diagnosis clinical events (D_p), denoted as $D_p = [E_t(p)|\ t = diagnosis]$, given the diagnostic clinical events of p: $\{E_t(p)|\ t \in DIAGNOSTIC\}$. Due to the large size of all possible diagnoses in EHR data, we define diagnosis prediction as a ranking problem such that top results of prediction model should ideally match the real diagnosis set (D_p) for patient p.

4.3.2 EHR from a Heterogeneous Network Point-of-View

Multiple types of clinical events and their various types of relations can be intuitively viewed as a heterogeneous network.

Definition 4.3.1. Heterogeneous Information Network is defined as a graph $G = (V, E)$ in which nodes and links between them can have various types. Nodes are mapped to their type by a node mapping function $g_v : V \to A$ where A is the set of all node types and similarly a link mapping function $g_e : E \to R$ maps links to their type where R is the set of all possible link types. By definition we have $|R| > 1$ or $|A| > 1$. Furthermore, $S_G = (A, R)$ denotes the network schema.

Different patients and clinical events form the nodes of our clinical heterogeneous network. The type of a node in this network is defined by the type of the clinical event mapped to it. Moreover, links of the network are designed based on the basic EHR relations which are mainly between a patient and a clinical event (e.g., patient's relation to his laboratory tests or symptoms). Figure 4.1 shows the abstract schema of the network illustrating node types and basic links. The figure also specifies if nodes belong to the treatment or diagnostic type category.

To further enrich the network with semantics of relation in EHR, new compositional relations can be defined using Metapaths [42].

Definition 4.3.2. Metapaths in HIN define higher order relations between two node types. Having the network schema $S_G = (A, R)$, a metapath schema is represented by its node and edge types connections: $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_m} A_{m+1}$.

A metapath is considered as a new link in the network and is added by creating a new connectivity between start and end nodes of any path matching the metapath schema. Metapaths allow our network to better learn the semantics of similarity among nodes. For instance, *patient*$\rightarrow$ *symptom* $\leftarrow$ *patient* can capture relation of patients in terms of experiencing similar symptoms.

4.3.3 Construction of HIN from Clinical Events

In this section, we introduce the proper modeling approach for construction of HIN from EHR data and our technique in extraction of some clinical events from raw text.

In general, having a clinical event $e = (t, n, v)$, it can be mapped into a node of type t

with identification of (n, v). For instance, a Glucose level of 60 can be mapped to a node of type *laboratory test* and identified by $(Glucose, 60)$. However, in many cases, different values of a unique clinical event convey identical semantic in terms of disease prediction. For instance, various numerical measurements in many laboratory tests are considered the same as long as they fall into one of the normal or abnormal ranges. Therefore, a proper modeling strategy should map clinical events with duplicate diagnostic semantic into the same node as failing to do so can negatively affect the power of the model in capturing similarity of nodes. Having this in mind, following steps are taken for mapping clinical events to nodes. Procedure and diagnosis clinical events are mapped based on the icd-9 coding system [18]. For each laboratory test, its name coupled with a reported flag which can be either normal or abnormal is considered as a unique node. The same strategy is employed when dealing with microbiology tests, where flags can be sensitive, resistant, or intermediate. Moreover, the age of patients is classified with threshold 15, 30, and 64 based on a statistical analysis of adverse events in different age groups studied by [146]. Finally, gender, ethnicity, and prescription, which are categorical events, are easily mapped by their unique category names.

Symptom Extraction

For extraction of symptoms that are commonly found inside raw-text clinical notes, we employed Autophrase [192] which is a novel phrase mining technique that learns high-quality phrases from a large corpus and allows for incorporating domain-specific knowledge bases for achieving highly domain-relevant results. We feed Autophrase with a pool of clinical phrases that are generated from two main sources: (1) Medical Subject Headings (MeSH) [1] vocabulary treasure which contains 90,000 medical entry terms, and (2) the ICD-10 [2] medical coding database. Quality phrases for symptoms are extracted from MeSH "signs and symptoms" category, code C23 and ICD-10 Chapter XVIII. We also run some final filtering steps on results of Autophrase to drop phrases that include measurements, adverbs,

[1] https://www.nlm.nih.gov/mesh/

[2] http://www.who.int/classifications/icd/en/

or symbols as they do not contribute to our diagnosis goal.

Having all nodes constructed, their connections are added based on schema in Figure 4.1 and selected metapaths are discussed in following sections. One of the advantages of HIN is that missing values in EHR only lead to the absence of some links and does not require further missing value management methods.

4.3.4 Heterogeneous Network Embedding for Clinical Events

Given the rich clinical information network, learning a latent and low-dimensional embedding of clinical events that can capture their internal relations is greatly beneficial for further analysis tasks. Inspired by the success of skip-gram [159] in learning latent word embeddings from the context of words in a corpus, most of homogeneous network embedding techniques [206, 94] rely on neighbor prediction paradigm. In this approach, given a network $G = (V, E)$, and an embedding function $f : V \rightarrow R^d$ that maps each node to a d dimensional vector, the objective is to maximize the probability of observing neighborhood of a node v, denoted as $N(v)$, conditioned on its representation $f(v)$ [94].

$$\underset{f}{\operatorname{argmax}} \prod_{v \in V} \prod_{c \in N(v)} Pr(c|f(v))$$

where the probability $Pr(c|f(v))$ is defined as a softmax function, normalized with respect to representation of all network nodes.

To exploit the rich structural information of EHR data and enrich semantics of similarity among different nodes, we employ an extension of above paradigm to heterogeneous networks that incorporates variety in node types and metapaths in the definition of node neighborhood and the objective function [47, 66]. In particular, with presence of multiple node types, neighborhood of a node v is defined as $N(v) = \{N_1(v), N_2(v), ..., N_T(v)\}$ where $N_t(v)$ denotes type t neighbors of v and T is the number of node types.

Moreover, having multiple types of paths leaving a node (simple or metapath), the neighbor prediction probability function $Pr(c|f(v))$ should be also conditioned on the type of path used. Specifically, the probability of visiting a neighbor c of a node v under path r with schema $V_1 \rightarrow ... \rightarrow V_l$, is defined as:

$$Pr(c|f(v), r) = \frac{exp(f(c) \cdot f(v))}{\sum_{u \in V_l} exp(f(u) \cdot f(v))}$$

As computation of above probability is very expensive in large networks, negative sampling [159] is employed to achieve following objective function:

$$Pr(c|v, r) = \log \sigma(f(c) \cdot f(v)) + \sum_{1}^{m} \mathbb{E}_{u_l \sim P_l(u_l)} \log \sigma(-f(u_l) \cdot f(v))$$

where m negative sample nodes are drawn based on their node degree and from nodes having the same type as r destination type (V_l). Therefore, a training step randomly samples a path schema r and two nodes v and c connected under r, along with m negative sampled nodes and employs Stochastic Gradient Descent (SGD) to update their embeddings. The tuple (v, c) is sampled based on the normalized number of links under the path r over each node tuples.

Although treatment clinical events should not be directly used in the diagnosis prediction, they can be profoundly beneficial in the unsupervised embedding model for capturing similarity of diagnostic clinical events in terms of consequent treatment. For instance, by including prescription and the metapath $symptom \leftarrow patient \rightarrow prescription$, into the embedding model, it can learn similarity among symptoms that lead to the same prescription. Therefore, for training the unsupervised embedding model, we first select the set of advantageous treatment nodes to be added to the embedding model by evaluating the performance-gain obtained from each or combination of them. Next, among many possible metapaths, we select candidate paths mainly from those that link a diagnostic event to a treatment one through a patient (such as the one above). We also compare candidate metapaths in terms of performance-gain when they are added to the network separately and incrementally and select the best configuration.

4.3.5 Diagnosis Prediction

When node embeddings are present, the process of diagnosis prediction for a new patient involves construction of the patient's representation based on his clinical events and ranking

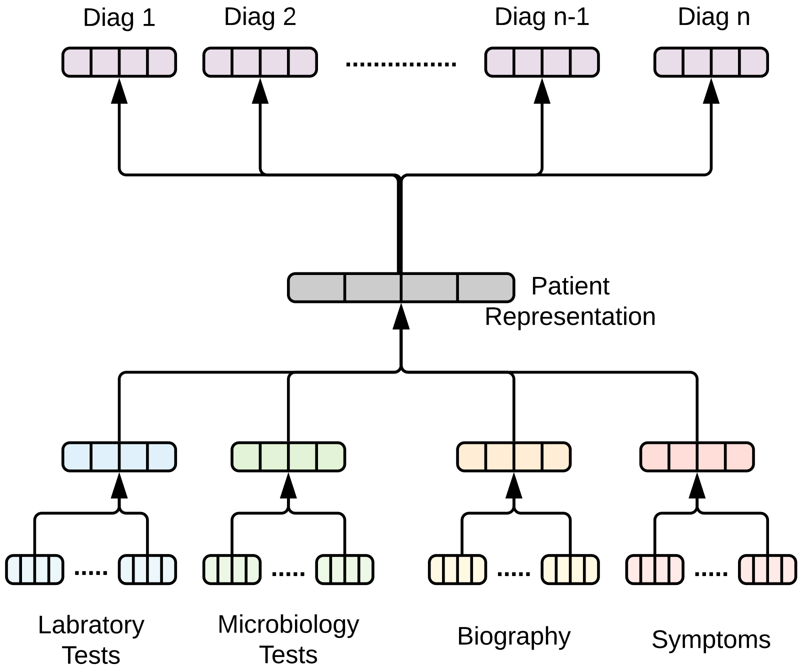


Figure 4.2: Diagnosis prediction flow.

diagnosis codes according to their dot product similarity to the patient's representation. Figure 4.2 shows the overview of the prediction flow. Given a patient p, his type t neighborhood ($N_t(p)$) can be summarized into a latent embedding ($f_t(p)$) by averaging its members:

$$f_t(p) = \sum_{n \in N_t(p)} \frac{f(n)}{|N_t(p)|}$$

Having clinical events of p grouped into latent type embeddings ($f_t(p)$), a representation for p can be intuitively achieved by aggregating them, but with different weights for each type (w_t) to capture importance of the type in diagnosis prediction.

$$f(p) = \sum_t w_t f_t(p)$$

Finally, a diagnosis d is scored and ranked by a dot product similarity between p and d embeddings: $s(d,p) = f(d) \cdot f(p)$.

4.3.6 Supervised Node Representation Learning for Diagnosis Prediction

The heterogeneous network embedding model discussed in section 4.3.4, does not have a direct guidance for learning representations that are specifically suitable for disease diagnosis aim and learns a general knowledge of the network. To add such guidance and provide diagnostic knowledge to the model, following [47] we employ the diagnosis prediction flow discussed in section 4.3.5 as a supervised embedding process and jointly use with the unsupervised model at the time of representation learning to tailor embeddings to disease diagnosis goal.

Recalling computation of prediction score from section 4.3.5, which is defined for a tuple of diagnosis d and patient p, we have:

$$s(d,p) = f(d) \cdot f(p) = f(d) \sum_t w_t f_t(p) = f(d) \sum_t w_t \left(\sum_{n \in N_t(p)} \frac{f(n)}{|N_t(p)|} \right)$$

We can employ a hinge loss ranking objective for the triple (p, d, $\sim d$) to update node embeddings (f) and node type weights (w_t) defined as:

$$max(0, -s(d,p) + s(\sim d,p) + \sigma)$$

where d and $\sim d$ are positive and negative sampled diagnosis for p and scores $s(d,p)$ and $s(\sim d,p)$ are calculated for them respectively.

To jointly learn embeddings, objectives of the two supervised and unsupervised models ($\mathbb{Z}_{supervised}$, $\mathbb{Z}_{unsupervised}$) are combined to form the joint objective as:

$$\mathbb{Z}_{joint} = \omega \cdot \mathbb{Z}_{unsupervised} + (1-\omega) \cdot \mathbb{Z}_{supervised} + \lambda \sum_n \|f(n)\|_2^2$$

where $\omega \in [0,1]$ is a pre-defined parameter for tuning importance of either models and a regularization term is added to prevent over-fitting of learned representations.

Therefore, a training step in the joint representation learning model works as follows. We draw one of the embedding models based on $Bernoulli(\omega)$. If the unsupervised model is drawn, its objective function is used on a mini-batch of randomly drawn triples (r, v, c) and

Table 4.1: MIMIC tables used in this study.

Table name	Main Columns	Description
patients_icd	gender, DOB, ethnicity	Name and demographic information of patients
procedures_icd	icd9_code	Procedure events such as brain monitoring, tubing, injection
prescriptions	generic_drug_name	Drugs prescribed in each admission
microbiologyevents	spec_itemid, interpretation	Microbiology tests and their sensitivity level; eg. fungi, bacteria
labevents	itemid, flag	Laboratory results and their flag (normal, abnormal); eg. Blood Glucose
Diagnosis_icd	icd9_code	Prescribed diagnosis codes.
noteevents	Category = D̈ischarge Summary¨	Raw text notes recorded by nurses which includes symptoms and other clinical information collected on admission time.

m negative samples to update representations. Otherwise, the supervised objective is used for a mini-batch of drawn triples (p, d, $\sim d$) to update type weights (w_t) and representations (f). Negative samples are drawn in both cases from a unigram distribution based on node degree [159].

4.4 Experiments

In this section we evaluate HeteroMed through three sets of experiments. First, it is evaluated under different design configurations. Then its diagnosis prediction performance is compared to various baseline models and finally it is quantitatively evaluated through two case studies.

4.4.1 Dataset

Experiments of this study are conducted on the publicly available Medical Information Mart for Intensive Care III (MiMIC III) [125] dataset. It contains a comprehensive clinical data for forty thousand patients admitted to the ICU department of BIDMC hospital during 11 years. The MIMIC dataset is organized into 26 tables containing clinical event records for each admission to the ICU and other general information such as definitions of clinical terms. Table 4.1 lists utilized database tables alongside with main columns used and a short description for each table. In this study, each admission of an adult subject (aged 15 years

Table 4.2: Node statistics for the HIN network.

Node Type	abbreviation	Train	Test
Patient stay	pati	36641	10000
Procedures	proc	1673	746
Prescription	pres	6000	3523
Microbiology	micro	212	63
Laboratory	lab	1870	1045
Diagnosis	diag	5605	2745
Symptom	symp	1602	435
Gender	gen	2	2
Age group	age	3	3
ethnicity	eth	40	32

or older) to the hospital is considered as a sample and called a *patient stay*. Few subjects with multiple ICU stays in a single hospital admission were excluded due to the insufficiency of diagnosis information provided for them in MIMIC.

Following these steps, we obtained a sample set of 46,641 patient stays from which 10,000 were randomly sampled for the test set and 36,641 remaining for the training set. The heterogeneous network was then constructed with the strategy discussed in section 4.3.3 using our train set. Table 4.2 lists statistical details for nodes of this network.

Furthermore, in addition to 9 length one basic links of the network, 9 other candidate metapaths were selected. Table 4.3 lists both types of paths with their frequency in the constructed network. As patient node is a central hub in our metapaths, each path is denoted only by its start and end node types. (e.g., lab-symp denotes the *laboratory test*← *patient* → *symptom* metapath).

Table 4.3: Edge statistics for constructed network.

simple links	count	metapaths	count
pati-proc	94,452	lab-diag	1,155,278
pati-pres	757,195	symp-diag	261,861
pati-micro	19,768	lab-proc	341,907
pati-lab	1,948,360	lab-pres	770,297
pati-age	36,641	symp-pres	223,666
pati-diag	292,473	symp-proc	63,424
pati-symp	307,325	lab-symp	214,356
pati-gen	36,641	micro-lab	14,394
pati-eth	35,342	micro-symp	8,696

4.4.2 Evaluation Strategies and Implementation Details

In this study, disease diagnosis is conducted in two levels. First, exact diagnosis code prediction as a ranking problem and second general disease cohort prediction as a multi-label classification problem which are evaluated with *MAP@k* and *AUROC score* respectively. *MAP@k* is a metric widely employed in information retrieval and reports the mean of average precision at k (*AP@k*) over all test samples. In this study, having a ranked diagnoses list returned by the prediction model, *AP@k* shows the averaged precision over all the positions in the list that the diagnosis is correct and has index less than k.

AUROC is a goodness of binary prediction metric based on different cut-off thresholds on classifier prediction score. Here, AUROC is computed for each of disease cohorts based on the scores computed by our supervised prediction model and baselines for each cohort.

For training HeteroMed and learning node embeddings, a mini-batch of 500 patients has been used at each training step with embedding vector size of 128. The unsupervised embedding model is selected with 4 times higher probability than the supervised model when performing the joint representation learning. Furthermore, each step of unsupervised

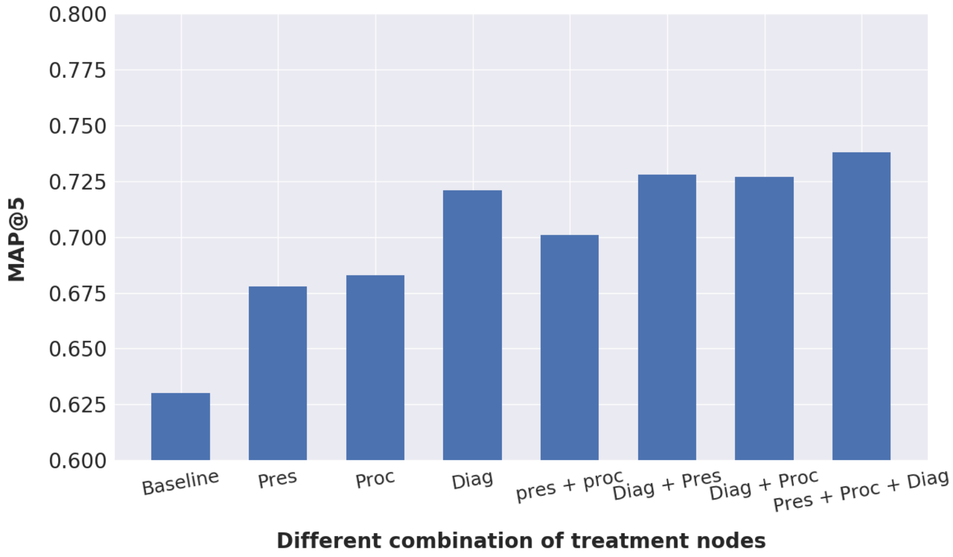


Figure 4.3: Treatment node selection evaluation.

approach draws 100 negative diagnosis samples for each patient based on the diagnosis node degree.

4.4.3 Evaluation of Proposed Method

In this section, we demonstrate experimental results of evaluating performance of HeteroMed under different metapaths and node selection configurations.

Treatment Node Selection

In this part, we evaluate performance-gain obtained, when each or a combination of treatment nodes (proc, pres, diag) are included in the network of unsupervised embedding model. Figure 4.3 illustrates the comparison to the baseline performance in which the network only contains diagnostic nodes.

We can observe that among treatment nodes, “diagnosis” show a great advantage to be added to the unsupervised embedding process. This is partly due to the fact that any improvement in diagnosis node embeddings directly impacts performance of diagnosis prediction. Procedure and prescription nodes also impact the performance in a positive way. This

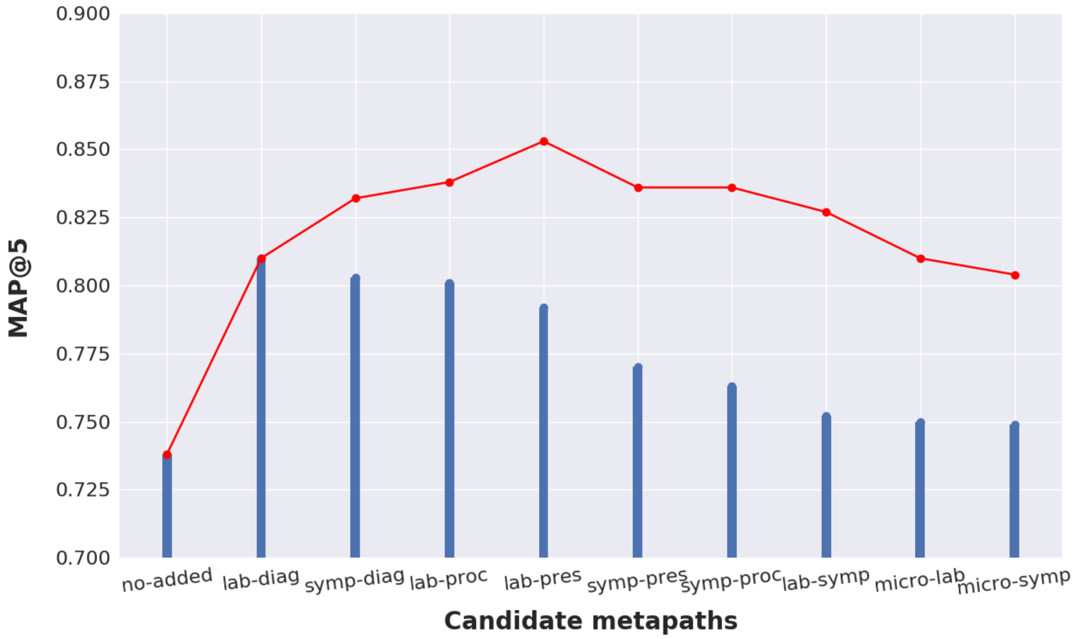


Figure 4.4: Metapath selection evaluation. Red line denotes additive performance and blue bars denote single path performance.

is while they are not included in the prediction step of diagnosis. Results of this experiment confirm the advantage of utilizing the whole set of available information in the unsupervised representations learning process.

Metapath Selection

In the second part of this experiment, we evaluate performance-gain obtained by using selected metapaths listed in Table 4.3. The results are elaborated in Figure 4.4. The blue bars in the figure show the performance for each metapath when added separately to the baseline and are sorted in descending order based on this measure. The red line, however, evaluates performance when these paths are accumulated incrementally to the model. Results of this experiment indicate that the combination of 4 first metapaths (lab-diag, symp-diag, lab-symp, lab-pres) provides us with the optimal performance for the disease diagnosis goal. This is while adding more paths leads to a gradual performance drop. This observation further clarifies the significant advantage of metapath-based neighbor sampling rather than the random neighbor sampling used in prior medical domain studies.

Based on these results, the model used in all succeeding experiments employs all treatment nodes and the four above-mentioned metapaths in its representation learning process.

4.4.4 HeteroMed Compared to Other Diagnostic Models

To further assess our model, we compare its diagnosis performance to selected state of the art models in two levels. First, when exact disease codes are to be predicted and second when disease cohorts are desired.

Exact Code Prediction

In this experiment, we try to rank exact disease codes for a patient stay. In this part, we compare HeteroMed only to embedding based models as the size of diagnosis codes is too large to be predicted by a supervised classifier. The baseline models in this task are:

Med2vec: Med2vec [48] is a multilayer medical embedding neural network which learns embeddings of medical events and visits using an approach inspired by word2vec. We modified the last softmax layer of this model to predict diagnosis codes for the current visit as the model originally predicted disease codes for last or future visits. Representation sizes in this model are chosen to be of size 100.

Skipgram-embedded: We use word2vec (skip-gram) to learn network node representations similar to the way it is employed in prior studies. In particular, all clinical events associated with an admission are considered as words and are concatenated to form sentences. The window size is set to the maximum length of sentences so that all clinical events (words) in an admission (sentence) can be sampled as neighbors. Node representations learned are fed into the supervised prediction model (section 4.3.5) to score diagnoses and rank them.

HeteroMed-embedded: We learn node embeddings by only employing the unsupervised

Table 4.4: Comparison of HeteroMed model to baselines for exact code prediction.

Model Name	MAP@3	MAP@5	MAP@10
Med2vec	0.75	0.78	0.79
Skipgram-embedded	0.73	0.76	0.77
HeteroMed-embedded	0.78	0.79	0.80
HeteroMed	**0.81**	**0.85**	**0.87**

Table 4.5: Similarity search results.

Diabetes		Cold		Anemia (lack of blood)	
HeteroMed	skipgram	HeteroMed	skipgram	HeteroMed	skipgram
peripheral neuropathy	**dietary change**	**general pain**	**fever**	**fatigue**	weight loss
sleep apnea	tightness	**fever**	**sick contact**	**malaise**	**allergy reaction to iron**
leg tingling	confusion	**chill**	constipation	**heart palpitation**	penile discharge
urinary frequency	speak difficulty	**sore throat**	**muscle pain**	**itchy skin**	sick contact
ulcers	nausea	swelling	**recent travel**	**bloody stool**	**shortness of breath**
dietary change	**rash**	**allergy**	limb pain	bruising	stuffy nose
burning	fever	**tightness**	urinary changes	**abdominal pain**	**leg tingling**
abdominal pain	mental status change	**sinus congestion**	**cough**	**nausea**	suicidal attempt
thirst	**numbness**	**cough**	stiff neck	**chills**	**abdominal pain**
itchy skin	sleepiness	blurred vision	**runny nose**	**cramps**	**jaundice**

representation learning approach introduced in section 4.3.4. We use the same set of metapaths as the main model for this aim. As the previous method, the supervised diagnosis model is then used to rank diagnoses.

Disease Cohort Prediction

Icd-9 diagnosis coding provides 20 code groups that correspond to 20 high level disease cohorts. In this part, we aim to predict all disease cohorts that a patient is diagnosed with. Having multiple diagnosis codes for a patient stay, different groups of diseases may be involved which turns the problem into a multi-label classification. When training our model for cohort prediction, only 20 disease nodes are constructed for the network and each disease code of patient is mapped to one of these nodes. Furthermore, in prediction time, scores for all 20 diagnosis nodes are computed to be evaluated. The baseline models are listed below:

Shallow supervised models: We use feature engineering along with common shallow models, from which Random Forest provided best results. We extracted the same features suggested by [177] but only from tables used to construct our network.

Deep Mimic Model: We finally compare our results to the ones from mimic learning model [46] which employs a deep neural network alongside with a Gradient Boosting Model for prediction of icd-9 diagnosis code groups.

Results

The exact code prediction evaluation is depicted in Table 4.4. As the results suggest, HeteroMed outperforms all the baseline models in exact diagnosis prediction. The out performance of HeteroMed-embedded model compared to skipgram-embedded model, reveals superiority of relation-aware embedding approach employed in this study to the skip-gram used in conventional clinical models. Furthermore, the Med2vec model outperforms the Skipgram-embedded model although they are both trained based on skip-gram embedding. This can be due to the fact that Med2vec incrementally updates the embeddings with back propagation in its model. However, it sill falls behind HeteroMed that employs a relation-aware embedding approach.

Results of the disease cohort prediction are illustrated in Figure 4.5. We can observe that HeteroMed performance exceeds baseline models in almost all code groups. In general, performance in some groups are lower than the others which generally corresponds to those diagnosis groups that are sparser in the MIMIC dataset.

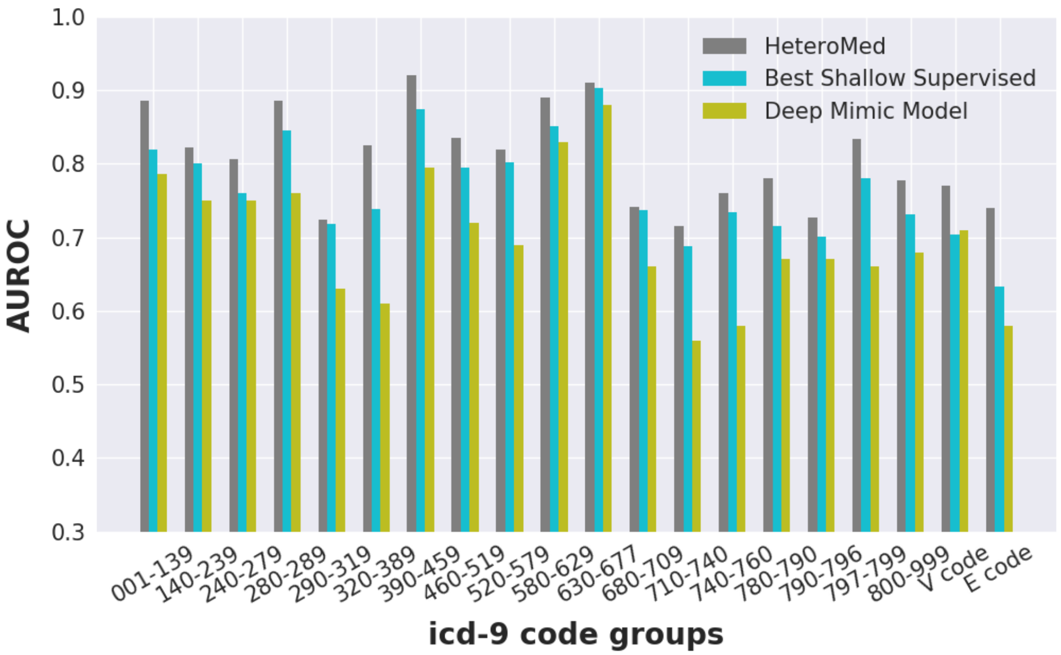


Figure 4.5: Disease cohort prediction evaluation.

4.4.5 Case Studies

In this section, we qualitatively evaluate modeling of EHR data using HeteroMed and validate sensibility of learned clinical event representations. First, we perform a similarity search to find relevant symptoms to three common diseases. We then review results of a sample prediction case. In both experiments, we compare the results to the Skipgram-embedded model introduced in the last section.

Table 4.5 lists top ten related symptoms and observations to three common clinical conditions: Diabetes, Cold, and Anemia. A dot product similarity has been employed to generate these results. To achieve better vision for comparison, results are validated by a clinical expert and relevant symptoms are shown in bold format. Recognizing the fact that symptoms can have hidden and complex relations to diseases, only directly related symptoms to each condition are considered as relevant.

Results of this experiment confirm the validity of learned representations by our model. Moreover, we can easily observe that HeteroMed ranks relevant symptoms higher than the

Table 4.6: Comparison of sample prediction results for a patient and real diagnosis codes. Star sign shows that the predicted code is not present in the ground truth codes.

Ground truth	Category	Skipgram	HeteroMed
4282	Circulatory system	2875	4273
4254	Circulatory system	3970	4282
2875	Blood organs	6841*	4583
4273	Circulatory system	281	2832*
3970	Circulatory system	7217*	2875
5303	Digestive system	427*	4254
4280	Circulatory system	4583	530*
281	Blood organs	4273	260*
4583	Circulatory system	2501*	281

Skipgram-embedded model and is vividly stronger in understanding relations of symptoms to diseases. One may notice that the intersection among results of two models is small. The large number of symptoms and the fact that a single complication can be described in multiple ways are the main reasons for this observation. For instance, leg tingling, numbness, and peripheral neuropathy can all refer to a similar complication caused by diabetes.

Table 4.6 shows a sample admission with 9 real diagnosis codes along with the 9 top-ranked predicted codes by each model. Wrong predictions are denoted by a star sign ($*$) on the top right corner. Furthermore, the main category of each disease code in ground truth is specified to provide better understanding of codes. The two methods rank a number of wrong codes in their first 9 predictions. However, the superior performance of HeteroMed is noticeable in two aspects. Firstly, we can observe that HeteroMed is able to detect all disease categories of the ground truth, although not predicting exact diagnosis codes. For instance, HeteroMed ranks the code 530, which is not present in ground truth, in its top 9 predicted codes. However, this code is a more general indication of the code 5303 in the ground truth and both show a digestive system disease with different specificity. This is while the top diagnosis codes ranked by Skipgram-embedded model do not cover this disease category. Secondly, all the top-ranked codes by HeteroMed are related to disease categories that are present in the ground truth. However, some of the diagnosis codes that are ranked high by Skipgram-embedded model are outside the ground truth disease categories. For instance, the code 7217 which relates to connective tissue diseases is ranked 5 by Skipgram-embedded model, while this disease category is not present in the ground truth.

4.5 Conclusion and Future Work

In this paper, we study the problem of disease diagnosis from a patient's diagnostic records available in EHR data. We propose modeling of clinical events as a heterogeneous information network, HeteroMed, to address shortcomings of previous methods pursuing same goals. Existing studies ignore the rich structure and relations in EHR data when learning representations of clinical events. HeteroMed is capable of capturing informative relations for the diagnosis goal and use the best relation sampling strategy when learning clinical event representations. It also allows for easy handling of missing values and learning embeddings tailored to the disease prediction goal using a joint embedding framework. Result of our study shows that HeteroMed can achieve significantly better results in diagnosis task and finding clinical similarities. This in turn confirms the benefits of employing heterogeneous information network in modeling clinical data.

CHAPTER 5

Hierarchical Target-Attentive Clinical Prediction using Heterogeneous Information Networks

In this chapter, we aim to extend our previously introduced approach in EHR modeling, HeteroMed (chapter 4), to achieve more integrative and target-aware modeling of patients' health records. In HeteroMed and existing studies relying on automatic EHR representation learning, record representations are often learned and aggregated without specificity for the target being predicted. Our model introduces a hierarchical attention mechanism that allows it to learn to attend to the most important clinical records for prediction of a specific target. Additionally, we propose a method for integrating non-categorical data such as real-time time-series data into our model which is essential in pervasive health solutions.

We evaluate our model using a publicly available benchmark EHR dataset on diagnosis prediction tasks and demonstrate that the use of target-aware attention significantly improves the performance compared to the current state of the art. We also demonstrate that the addition of time-series data leads to further performance improvements. Lastly, we show that the predictions made by our proposed model are easily interpretable.

5.1 Introduction

In this study, we revisit the problem of integrative health data modeling and take the diagnosis prediction as our target task. When mining EHR data, one recent approach for addressing the challenges of heterogeneity and sparsity of records has been unsupervised record embedding. As discussed in section 2.3, Med2Vec [48], as well as successive studies such as [49] and HeteroMed (chapter 4), use skip-gram [52] based techniques to learn latent representations for health records based on their co-occurrence relations. In these studies, predictions are commonly made by training supervised models on patient representations, which are obtained by aggregating the representation vectors of the items in a patient's health records.

One shortcoming in these past works stems from the relatively simple aggregation process they use, in which they treat records with equal importance regardless of what diagnosis is being predicted. Taking diabetes and kidney failure as an example, we can see how this is an issue: prior models generate a single patient representation by combining records with fixed weights, which is then used for the prediction of both diagnoses; however, the importance of tests should vary based on the diagnosis being predicted, with blood glucose levels being more important than blood albumin levels when predicting diabetes than when predicting kidney failure and vice versa.

Another shortcoming of these past approaches is that the predictions generated by these models are not easily interpretable, with no way for an end user to understand how the model arrived at its conclusion. Lastly, past approaches only make use of records whose values can be mapped to distinct categories, leaving out other important information such as time series data obtained from continuous monitoring of vital signs.

Inspired by the very recent success of attention mechanisms in network embedding [240, 117, 213], we propose HTAD, a novel approach for modeling EHR data that leverages hierarchical attention, to overcome these shortcomings. HTAD produces diagnosis-aware patient representations, as well as explainable predictions. We also suggest how non-categorical data, in particular, time series data, can be integrated into HTAD.

Considering EHR in the context of Heterogeneous Information Networks (HIN) with patients and records mapped to network nodes, our model's goal is to aggregate a patient's neighborhood such that the obtained representation is tailored to the prediction of a specific target diagnosis. Recognizing heterogeneity of nodes, we perform the neighborhood aggregation at two levels: first, at node-level and among nodes having similar type to obtain a set of type representations, and then at the type-level to achieve a comprehensive patient representation. In node-level aggregation, we propose employing a target-aware attention mechanism to learn the importance of various nodes with respect to the given diagnosis. We also show ways for the incorporation of time-series data at this level. We apply similar attention technique at the type-level to allow the model to learn preference towards various record types for the prediction of the specified disease. We then pass the resulting patient representation into our objective function for prediction. Importantly, attention weights generated in our model improve the interpretability by providing insight as to which nodes and types the model finds most important for the prediction.

We evaluate our proposed model's performance on two diagnosis prediction tasks: exact diagnosis code prediction and high-level diagnosis group prediction, using the publicly available MIMIC-III EHR dataset [125]. We compare HTAD to several existing models that represent the state of the art for diagnosis prediction using EHRs. Our experiments show that HTAD outperforms these benchmarked models on both tasks, in multiple cases beating them by a margin of over 10%. Additionally, we evaluate our model's interpretability, something that has not been explored in past models for diagnosis prediction that represented patients based on their aggregated EHR embeddings.

5.1.1 Contributions

In summary, we make the following contributions in this chapter: 1) we propose Hierarchical Target Attentive Diagnosis (HTAD), a novel approach in mining EHR data using HINs. 2) We show how attention mechanism can be exploited in EHR network mining for the calculation and incorporation of record importance with respect to various prediction targets. 3) We show how non-categorical data can be incorporated when mining EHR data represented as an HIN. 5) We demonstrate that HTAD significantly improves diagnosis prediction performance compared the state-of-the-art models. 6) We show that HTAD's use of target-aware hierarchical attention can provide full interpretability on its clinical decision process.

5.2 Methodology

In this section, we present our hierarchical target attentive representation learning model for multi-modal EHR mining using heterogeneous information networks.

5.2.1 Model Overview

Relying on the general design of HeteroMed introduced in chapter 4, we model EHR data as a heterogeneous information network, allowing record embedding vectors to contain information on the node types and their unique relations. Similar prior patient embedding studies, record representations are trained using an unsupervised skip-gram based approach and the patient representation is obtained through an aggregation of neighboring record representation vectors. However, different from prior studies where the relative importance of various context nodes in the prediction of the target is not considered, we propose a hierarchical target-attentive aggregation to include this information. In other words, instead of generating a single patient representation that is then used for scoring or classification of all target nodes, our model learns to generate a distinct patient representation for each target node, achieved by favoring the most predictive context nodes for that specific target. The overall architecture for our target-attentive patient aggregation is depicted in Figure 5.1.

As illustrated, we first aggregate context nodes based on their type using a node-level attention mechanism, generating type-specific embedding vectors. The attention weights are assigned based on the importance of the node in prediction of the diagnosis. We also present a type-level attention layer to learn the importance of each type in predicting the target, further helping to obtain a diagnosis-aware patient representation. Finally, to generate the aggregated type embedding for time-series nodes as well, we replace the node-level attention mechanism with a deep sequential model.

In the rest of this section, we provide the formal details of our proposed model. A summary of the notations used in the following sections along with their explanations are provided in Table 5.1.

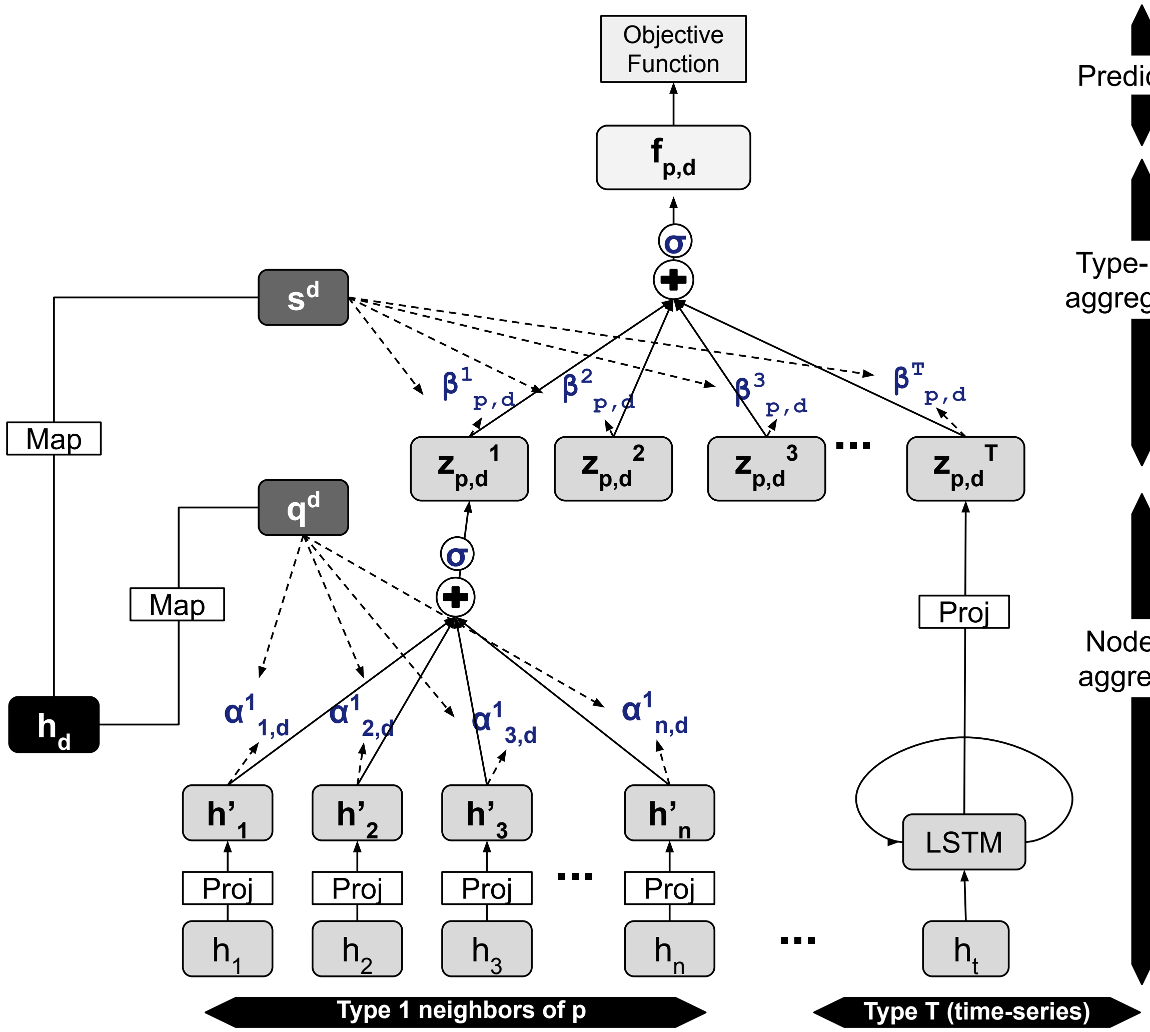


Figure 5.1: The architecture of the proposed hierarchical target-attentive HIN, illustrating the aggregation of patient p's context nodes with respect to diagnosis d.

Table 5.1: Notation and Explanations

Symbol	Explanation
h_n	Embedding of node n
$h_n^{'}$	Transformed embedding of node n
$N_t(p)$	Type t neighborhood of patient p
$z_{p,d}^t$	Aggregated embedding of nodes in $N_t(p)$ with respect to diagnosis d
q^d	Node-level attention vector for diagnosis d
s^d	Type-level attention vector for diagnosis d
$\alpha_{n,d}^t$	Node-level attention score assigned to node $n \in N_t(p)$ when predicting for diagnosis d
$\beta_{p,d}^t$	Type-level attention score assigned to type t representation of patient p, when predicting for diagnosis d
$f_{p,d}$	Aggregated patient p embedding with respect to diagnosis d
M	Node embedding lookup matrix
Q	Node-level attention lookup matrix
S	Type-level attention lookup matrix
W_c^t, b_c^t	Transformation parameters for context nodes with type t
W_d, b_d	Transformation parameters for target (diagnosis) nodes
W_q, b_q	Transformation parameters to obtain node-level attention
W_s, b_s	Transformation parameters to obtain type-level attention
W_t, b_t	Transformation parameters for time series type embedding

5.2.2 Problem Definition

Following the strategies introduced in section 4, each record in EHR is translated into a node in a heterogeneous information network. Having built a network using the EHR data, we can view the set of all nodes as a union of the target, patient, and context nodes.

Definition 5.2.1. Target/Context Nodes Target nodes are defined as the nodes for which we want to predict if a connection exists to a given patient p node. In this study, diagnosis nodes are considered to be target nodes. All nodes other than patient and target are considered context nodes.

Consequently, the prediction task in an HIN representing EHR data can be defined as:

Definition 5.2.2. Clinical Prediction in an HIN Setting Given a patient p with context nodes $N(p) = \{N_1(p), N_2(p), \ldots, N_T(p)\}$ where $N_t(p)$ denotes the type t neighborhood of p, predict p's target neighborhood: $N_d(p) = \{d_1, d_2, \ldots, d_{|N_d(p)|}\}$.

When working with diagnosis prediction task, it is important to note that many medical ontologies, such as the ICD-9 system [18], provide a hierarchical and multi-resolution view of diagnoses, with the highest level of the hierarchy identifying the general disease group (e.g., cardiovascular disorders) and lower levels providing more specificity as to the exact diagnosis. Importantly, clinicians may assign codes to a patient at any level. Therefore, the diagnosis prediction task can be defined at two levels:

- Low-level (exact) code prediction: Due to the large size of diagnosis codes, this task is approached as a ranking problem, with the aim of scoring positively labeled codes higher than others.
- High-level (grouped) code prediction: In this task, we aim to predict all diagnosis groups associated with a patient, formulated as a multi-label classification task.

5.2.3 Target-attentive Node-Level Aggregation

As EHRs are composed of data of heterogeneous types, each node type can carry specific semantic and diagnostic information. Therefore, we start the aggregation process of patient p's neighborhood by combining the context nodes based on their types, thus obtaining type representation vectors.

As motivated in section 5.2.1, ignoring the relative importance of nodes in prediction of a specific diagnosis and consequently making use of a single patient representation, may impede achieving optimal diagnosis prediction performance. To address this, we leverage an attention mechanism to generate diagnosis-aware type and patient representations. Given a patient p, its type t neighborhood denoted as $N_t(p)$, and a diagnosis node d with corresponding embedding vector h_d, the node level attention works as follows:

We first utilize a linear transformation layer, parameterized by a type-specific weight matrix $W_c^t \in \mathbb{R}^{F^{'} \times F}$ and bias vector $b_c^t \in \mathbb{F}^{'}$, to project p's context nodes into a new feature space that is more compact and expressive for attention-based node scoring:

$$h_n^{'} = W_c^t h_n + b_c^t \tag{5.1}$$

where h_n denotes embedding of context node $n \in N_t(P)$ with length F looked up from a trainable embedding matrix $M \in \mathbb{R}^{|N| \times F}$ and $h_n^{'}$ denotes the transformed node embedding with length $F^{'}$.

The importance of each node is then measured based on the similarity of its transformed embedding to a diagnosis-specific attention vector $q^d \in F^{'}$. In the most general design, this vector is obtained by applying a linear transformation, parameterized by weight $W_q \in \mathbb{R}^{F^{'} \times F}$ and bias vector $b_q \in \mathbb{F}^{'}$, to the diagnosis node embedding h_d, formulated as:

$$q^d = W_q h_d + b_q \tag{5.2}$$

where h_d is the original diagnosis node embedding.

However, when working with low-level diagnosis codes, there is a significant imbalance in their frequency in a real-world setting. Therefore, the prior approach may face trouble in learning attention vectors for sparser codes. As such, grouping together those with similar diagnostic processes and allowing them to share attention vectors can improve the expressive power of attention for sparser codes.

Following this idea and taking $D^{'}$ as the set of such a grouping with size $|D^{'}|$, q^d can be looked up from an attention matrix $Q \in \mathbb{R}^{|D^{'}| \times F^{'}}$, after mapping d to one of the $|D^{'}|$ diagnosis groups. Q is randomly initialized and jointly trained by the model. It is important to note that for high-level diagnosis classification task these groups can be defined the same as diagnosis groups we are predicting for. We refer to this approach in the rest of this paper as **group-based** attention.

Having the transformed node embedding $h_n^{'}$ and diagnosis attention vector q^d obtained, the importance score between them denoted as $e_{n,d}^t$, is calculated as:

$$e_{n,d}^t = \frac{q^d \cdot h_n^{'}}{\sqrt{F^{'}}} \tag{5.3}$$

where t shows the type of node n and division by $\sqrt{F^{'}}$ is used to scale the dot product for improved performance, following design of the scaled dot product attention [212].

We then normalize the node importance scores using a softmax function to obtain the attention coefficient $\alpha_{n,d}^t$.

$$\alpha_{n,d}^t = \frac{\exp(e_{n,d}^t)}{\sum_{n^{'} \in N_t(P)} \exp(e_{n^{'},d}^t)} \tag{5.4}$$

Lastly, the normalized attention coefficients are used as weights for linear aggregation of transformed node embeddings, which is then followed by a non-linearity function to form the type embedding:

$$z_{p,d}^t = \sigma\Big(\sum_{n \in N_t(p)} \alpha_{n,d}^t \cdot h_n^{'} \Big) \tag{5.5}$$

where $z_{p,d}^t$ denotes the aggregated embedding generated from patient p's neighbors of type t that is used to predict the presence of diagnosis d.

5.2.4 Node-Level Time Series Aggregation

As discussed in section 5.2.2, the node embeddings used in the node-level aggregation process are obtained using a shallow embedding lookup process. However, such a technique is not usable for records kept in a time series format, as these records cannot be easily mapped to a small fixed set of categorical values and as there would be too little sharing of nodes between patients if each unique time series were mapped to its own categorical node.

Therefore, to incorporate such records into our proposed information network, we employ a Long-Short Term Memory (LSTM) [106] sequential model similar to [102] to obtain the aggregated type representation.

Formally, given a patient p and its type t neighborhood $N_t(p)$ in which nodes have continuous values, the values are first sorted by time to form a sequence $S_t(p) = \{s_1, s_2, s_3, \ldots, s_T\}$ that is then fed to an LSTM-based recurrent neural network. Next, the hidden state of the last LSTM cell, denoted as v_t, is transformed to a vector with embedding size $F^{'}$, forming the type t representation formulated as:

$$z_{p,d}^{t} = W_t v_t + b_t \tag{5.6}$$

where W_t and b_t are the parameters of the transformation function. It is worth noting that the embedding obtained is not diagnosis specific, but we have included d to keep the type representation notation consistent throughout the paper.

5.2.5 Type-level Aggregation

After deriving the node-level aggregations on patient p's neighborhood with respect to diagnosis d to obtain type embeddings $Z_{p,d} = \{z_{p,d}^{1}, z_{p,d}^{2}, \ldots, z_{p,d}^{T}\}$, our next step is to combine them to generate the patient representation. Similar to nodes, the predictive power of the different types may vary across diagnoses. For example, the diagnosis of some diseases relies more upon the results of microbiology tests while others rely more on laboratory test results.

Therefore, we propose to use another layer of diagnosis-aware aggregation. To this end, the importance of each type is measured by taking the closeness of its aggregated representation to a type-level attention vector s^d. Similar to node-level attention, the type-level attention vector can either be obtained by applying a linear transformation parameterized by weight W^s and bias b^s to a disease node's embedding, or by the group-based attention introduced that maps d to one of the $|D^{'}|$ diagnosis groups and look up its attention vector from the attention-matrix $S \in \mathbb{R}^{|D^{'}| \times F^{'}}$.

The normalized attention coefficient between the aggregated type t representation ($z^t_{p,d}$) and attention vector s^d is defined as:

$$\beta^t_{p,d} = \frac{\exp \dfrac{s^d \cdot z^t_{p,d}}{\sqrt{F'}}}{\sum\limits_{z^{'}_{p,d} \in Z_{p,d}} \exp \dfrac{s^d \cdot z^{'}_{p,d}}{\sqrt{F'}}} \tag{5.7}$$

In the final step, the comprehensive patient representation, specific to prediction of diagnosis d, is denoted as $f_{p,d}$ and is obtained by combining the type representations as follows:

$$f_{p,d} = \sigma\Big(\sum_{t \in T} \beta^t_{p,d} z^t_{p,d}\Big) \tag{5.8}$$

5.2.6 Model Inference and Optimization

In this section, we describe how parameters and node embeddings are learned. We employ both supervised and unsupervised learning objectives, which we will provide their optimization details. Furthermore, recalling that our diagnosis prediction task is defined at two levels and as classification and ranking tasks, we formulate the supervised classification objectives separately for each of these tasks.

High-level Diagnosis Code Classification

In section 5.2.5, we explained how we obtain a set of patient representations $F_p = \{f_{p,d_1}, f_{p,d_2}, \ldots, f_{p,d_k}\}$ in order to predict each of the k diagnoses in $D = \{d_1, d_2, \ldots, d_k\}$ for a patient p.

To make a prediction using these representations, we first feed them into a Multi Layer Perceptron (MLP) that uses a sigmoid activation function on its last layer. The MLP maps $F_P \mapsto D$ and is implemented in two layers: the first one shared among all patient representations and the second one specific to each diagnosis group.

As the task is formulated as a multi-label classification problem, we use the following loss definition to optimize the model:

$$L = mean(l_1, l_2, \ldots, l_k) \tag{5.9}$$

where

$$l_i = y_i \log \sigma(x_i) - (1 - y_i) \log(1 - \sigma(x_i)) \tag{5.10}$$

where y_i denotes the ground-truth label for diagnosis d_i in patient p's records and x_i is the prediction made by the model.

Low-level Diagnosis Code Ranking

As this task is framed as a ranking problem, we rely on score calculation between a patient and diagnoses. In particular, given a patient representation $f_{p,d} \in \mathbb{R}^{F^{'}}$ learned with respect to diagnosis d, we first transform d's node embedding to the same space as $f_{p,d}$. In particular, we employ weight matrix $W_d \in \mathbb{R}^{F^{'} \times F}$ and bias vector $b_d \in F^{'}$ for this aim and have:

$$h_d^{'} = W_d h_d + b_d \tag{5.11}$$

where h_d and $h_d^{'}$ denote original and transformed diagnosis node embeddings, respectively.

The score of diagnosis d for patient p is then defined as the dot product between their representations:

$$score(p, d) = f_{p,d} \cdot h_d^{'} \tag{5.12}$$

Using this score definition, we optimize the model using a hinge loss formulated as:

$$\max(0, -score(d, p) + score({\sim}d, p) + \epsilon) \tag{5.13}$$

where ${\sim}d$ is a negative diagnosis sampled for this patient and ϵ is the hinge margin.

Unsupervised Node Embedding

Besides the guidance of the supervised task, similar to HeteroMed we employ an unsupervised network embedding to capture the network structure and relation of nodes. The unsupervised model is optimized by:

$$\log P(j|i;r) \approx \log \sigma(h_i \cdot h_j + b_r) + \sum_{l=1}^{k} \mathbb{E}_{j' \sim P_n^r(j')}[\log \sigma(-h_i \cdot h_{j'} - b_r)] \tag{5.14}$$

where i is a node in the network and j is its random neighbor under the path r. $Dest(r)$ is the set of all nodes that are possible destinations on a path of type r and h_i and h_j are the embedding vectors of nodes i and j, respectively. $j^{'}$ is a negative sample, a destination on path type r that is not connected to node i, and b_r is a learned bias term specific to path type r. We draw k negative samples, without replacement, using a probability distribution based on the weight of the edges connecting to each node.

Combining the Supervised and Unsupervised Models

The supervised objectives we introduced optimize the node embeddings for the diagnosis prediction task, while the unsupervised model embeds more general knowledge about the proximity of nodes and their relationships with one another. To combine these two types of models, we follow the joint optimization approach suggested in [47] and define the following objective:

$$\mathbb{L}_{joint} = \omega \mathbb{L}_{unsupervised} + (1-\omega)\mathbb{L}_{supervised} + \lambda \sum_i \|h_i\|_2^2 \tag{5.15}$$

where $\omega \in [0,1]$ sets the weight used when sampling a model to train at each training step. We incorporate L2 regularization to reduce overfitting.

5.3 Experiments

In this section, we provide qualitative and quantitative evaluations of HTAD, demonstrating its superior performance to existing models and its interpretability advantages.

5.3.1 Dataset

All evaluation experiments in this study are conducted using MIMIC-III, a freely accessible critical care database [125]. For data preparation and preprocessing, we follow the steps suggested by a recent work that has established a method for creating a standardized benchmarking dataset derived from MIMIC-III [102] and is adopted by recent studies. Accordingly, a total of 42,019 unique hospital admissions and their corresponding clinical records and diagnostic information are included for modeling, 35,725 of which are used for model training and 6,294 of which are used for testing. This work also proposes a set of manually selected dense features for model training, which we rely upon when training the LSTM-based time series node aggregation model.

As discussed earlier, the two main tasks we deal with in this study are exact and group-based diagnosis code prediction. To investigate the former, we utilize the 6016 exact ICD-9 codes recorded in MIMIC-III, with a mean of 11.33 diagnosis codes in each patient's records. For the task of high-level diagnosis group prediction, we rely on the 25 disease phenotype groups introduced in the aforementioned MIMIC-III benchmarking study [102]. Table 5.2 summarizes the statistics of the nodes we extract to build our proposed information network.

Table 5.2: Node statistics in our network

Abbreviation	Description	Used in supervised model	Unique Values (train)	Unique Values (test)
symp	Symptoms	Yes	802	486
micro	Microbiology tests	Yes	1,467	684
proc	ICD-9 procedure codes	Yes	1,788	1,088
labt	Lab tests	Yes	960	853
bio	Biographical data	Yes	9	9
pres	Medications Prescribed	No	2,594	1,547
pati	Patient stay identifier	No	35,725	6,294
diag	Low-level ICD-9 diagnosis codes	No	6,016	3,597

Evaluation Metrics

Prediction of high-level disease groups is considered a multi-label classification problem. Accordingly, we follow existing works and employ Micro, Macro, and Weighted AUC-ROC scores to evaluate this task.

On the other hand, the exact diagnosis code prediction task is considered a ranking problem. Following adopted approaches in the evaluation of large-scale ranking tasks [153], rather than scoring all diagnosis codes for each patient, the ranking is conducted on a list of 100 codes, consisting of the original positive codes and a number of negatively sampled diagnosis codes. We evaluate our performance on this task using the Mean Average Precision at K (MAP@K), where K is set to 4, 6, 8, and 10.

5.3.2 Baselines

We compare our proposed model, HTAD, to recent studies that have achieved state of the art results in diagnosis prediction, including those combining manual feature selection and supervised deep models as well as those relying on unsupervised EHR embedding. We also evaluate variants of HTAD in order to demonstrate the effectiveness of each of its components. A comprehensive list of models evaluated is as follows:

- Std-LSTM [102]: An LSTM-based model for predicting high-level diagnosis groups. This model was introduced as the standard baseline for benchmarking the diagnosis prediction task.

- MMDL [177]: A multi-modal deep model for diagnosis group prediction that relies on a comprehensive set of hand selected features extracted from categorical and time series records in EHR. We compare with MMDL to show the effectiveness of automatic health record embedding.

- SAnD [200]: SaND is a recent study on diagnosis group prediction that employs a self-attention mechanism when modeling the EHR data. This study relies on manual feature extraction as well.

- Med2Vec [48]: An influential skip-gram based model for embedding health records. As this model is used to learn node embeddings and not for prediction, we employ mean aggregation of the embeddings it learns to represent patients based on their records and rely on supervised prediction methods similar to those used in HTAD.

- HeteroMed (chapter 4): An HIN embedding method for modeling EHR data. Comparing to HeteroMed can directly reveal the benefits of learning record importance scores in diagnosis prediction as its basic architecture is similar to HTAD's.

- HeteroMed$_{\text{MLP}}$: A variant of HeteroMed that we introduce and use for the group-based diagnosis classification task, obtained by replacing the hinge loss ranking objective with our multi-label classification one. For fair comparison, the MLP classification layer is configured identically to HTAD's.

- HTAD$_{\text{noAttnGrp/noTS}}$: A variant of HTAD that does not employ the group-based attention introduced in section 5.2.3. This model also excludes time series data so that the performance comparison to HeteroMed is solely focused on the attention mechanism used.

- HTAD$_{\text{AttnGrp/noTS}}$: A variant of HTAD that employs the group-based attention. For fair comparison with HeteroMed, this model excludes the time series data as well.

- HTAD: Our proposed model, employing group-based attention along with time series node aggregation in both prediction tasks.

Implementation Details

We implemented HTAD in Python using TensorFlow [5]. The model parameters are randomly initialized, and the model is trained using the Adam optimizer [129]. The learning rate of the optimizer and the batch size are set to 0.001 and 32, respectively. Moreover, the node embedding size is set to 256 and the attention vector size is set to 128. When using grouped attention vectors, diagnosis groups are formed based on the Clinical Classification Software hierarchical coding system [60], which organizes diagnosis codes based on their overall disease cohorts. Furthermore, the LSTM model used in node-level time series aggregation is pre-trained using the model configuration proposed by the Std-LSTM model.

Our implementation of HeteroMed shares its code base with HTAD, particularly in network formation and unsupervised node embedding training. For a fair comparison, both models use the same set of hyperparameters and meta paths when training the unsupervised node embedding task. The metapaths used are: $labt \leftarrow pati \rightarrow diag$, $diag \leftarrow pati \rightarrow symp$, $labt \leftarrow pati \rightarrow symp$. Furthermore, we observed that running the unsupervised part as a pre-training step provided the best results for low-level prediction in HTAD, and as such for both models we do not employ joint training for this task. However, joint training is employed in all other tasks, with the model selection coefficient (ω) set to 0.5. Med2Vec is trained with an embedding size of 256, and the MMDL and SAnD models are run using the same parameters and setups suggested in their studies. Experiments were run on one NVIDIA GeForce RTX 2080 Ti GPU and two cores on an Intel Core i9-7920X CPU.

5.3.3 Evaluation of Disease Phenotype Classification

Table 5.3 lists the results obtained from evaluating our models on the diagnosis group classification task. Overall, we observe that HTAD outperforms all of the baselines we investigated. Inspection of results further demonstrates that:

- $\text{HTAD}_{\text{AttnGrp/noTS}}$ shows notably higher performance than $\text{HeteroMed}_{\text{MLP}}$. This comparison is important as it demonstrates the effectiveness of our target-attentive aggregation mechanism versus models that otherwise share the same structure.

- Compared to $\text{HTAD}_{\text{noAttnGrp/noTS}}$, $\text{HTAD}_{\text{AttnGrp/noTS}}$ shows slightly better performance. This indicates that defining independent attention vectors as in group-based attention can be easier to train even when we are working with limited set of diagnoses.

- HTAD shows better performance than $\text{HTAD}_{\text{AttnGrp/noTS}}$, which is expected as the

Table 5.3: Phenotype Classification Results

	AUC-ROC		
Model	Micro	Macro	Weighted
Std-LSTM	0.821	0.77	0.757
MMDL	0.819	0.754	0.738
SAnD	0.816	0.766	0.754
Med2Vec	0.815	0.748	0.741
HeteroMed	0.831	0.745	0.739
HeteroMed$_{\text{MLP}}$	0.864	0.788	0.786
HTAD$_{\text{noAttnGrp/noTS}}$	0.871	0.829	0.815
HTAD$_{\text{AttnGrp/noTS}}$	0.874	0.832	0.818
HTAD	**0.880**	**0.843**	**0.828**

latter does not utilize the time series information in our dataset.

- HeteroMed$_{\text{MLP}}$ outperforms HeteroMed by a considerable margin. This is in line with our expectations, as the original ranking objective used in HeteroMed may not be optimal for multi-label classification, and we expected that adjusting that could improve the performance.

- HeteroMed$_{\text{MLP}}$ shows performance distinctly superior to that of the methods that rely on deep neural networks (SAnD, Std-LSTM, MMDL). This can be attributed to the fact that information networks eliminate the need for manual feature selection and allow for the incorporation of all clinical records. HeteroMed$_{\text{MLP}}$ also outperforms Med2Vec, which is expected as it employs a more semantic-aware node representation learning approach.

Table 5.4: Exact Diagnosis Code Ranking

Model	MAP@4	MAP@6	MAP@8	MAP@10
Med2Vec	0.752	0.743	0.738	0.714
HeteroMed	0.866	0.843	0.814	0.805
$\text{HTAD}_{\text{noAttnGrp/noTS}}$	0.867	0.842	0.813	0.806
$\text{HTAD}_{\text{AttnGrp/noTS}}$	0.888	0.848	0.821	0.810
HTAD	**0.890**	**0.881**	**0.865**	**0.923**

5.3.4 Evaluation of Exact Disease Code Prediction

Few studies have approached the task of exact disease code prediction, partly due to the huge size of the prediction space, which consists of thousands of possible disease codes. Furthermore, many of the studies we evaluated for the phenotype classification task cannot be directly extended to the task of exact disease code prediction. In this study, we evaluate variants of our model against HeteroMed and a model based on Med2Vec. Results of these experiments are presented in Table 5.4, from which we can infer that:

- HTAD, which incorporates time series data as well as group-based attention, outperforms all other models.

- Similar to the high-level classification task, a comparison between $\text{HTAD}_{\text{AttnGrp/noTS}}$ and HeteroMed reveals the significance of employing hierarchical attention mechanism in node-aggregation.

- The performance gain of $\text{HTAD}_{\text{AttnGrp/noTS}}$ compared to $\text{HTAD}_{\text{noAttnGrp/noTS}}$ is significantly more greater in this task. This gain can better demonstrate the advantage of using the group-based attention mechanism. As discussed before, sharing attention

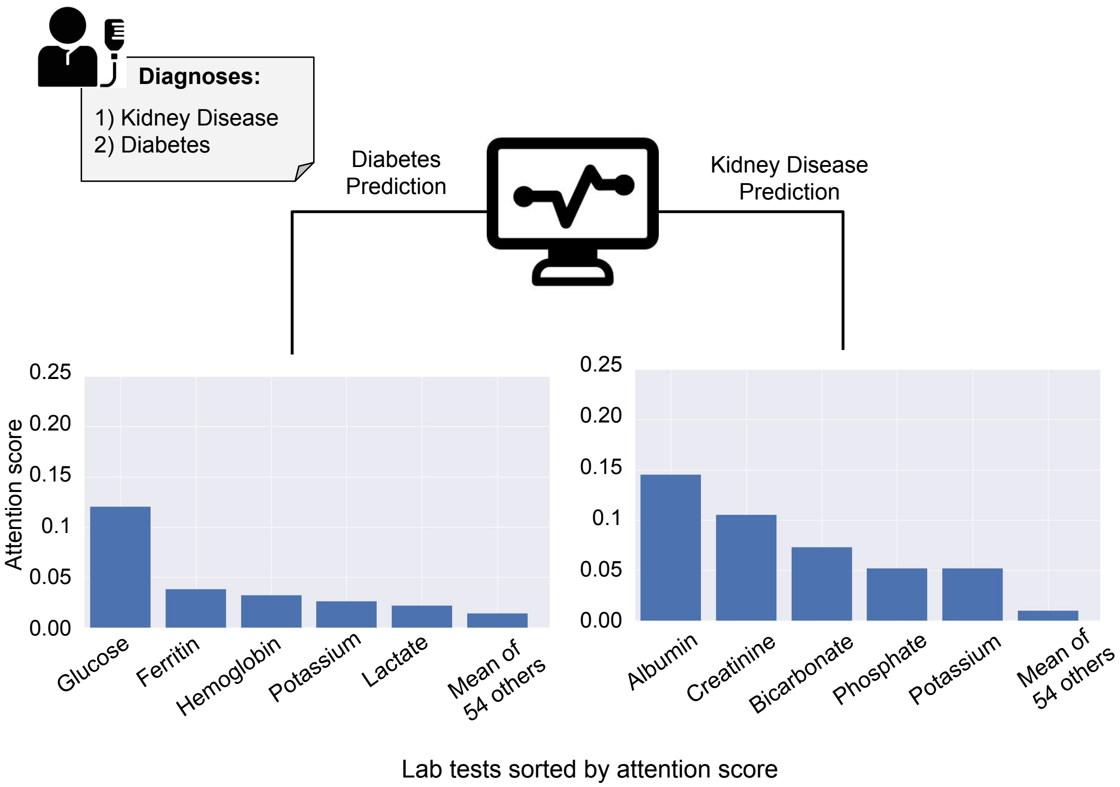


Figure 5.2: Distribution of attention scores for prediction of kidney disease and diabetes in a patient presenting with both conditions.

vectors among similar diagnoses can result in better performance for less common ones that otherwise remain under-trained.

5.3.5 Analysis of Attention Mechanism

Besides the performance improvement that our proposed hierarchical attentive architecture offers, one major benefit it provides is the interpretability of its results and the transparency for understanding which clinical events have led the model to a specific diagnostic decision.

To illustrate this interpretability in the node-level aggregation process, we consider a patient with stay id 20728. With diabetes and kidney failure both listed among this patient's high-level diagnosis groups, we study the importance score assigned to each of his 59 laboratory tests when predicting these two conditions. Recall that in our model, attention

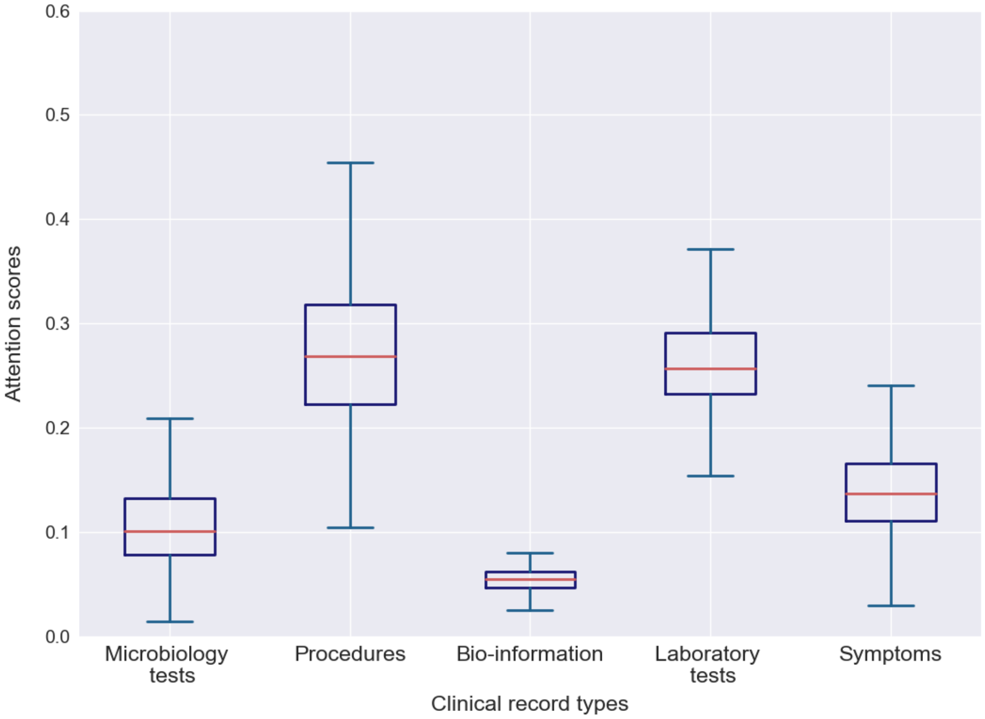


Figure 5.3: The distribution of attention weights among various types.

scores are calculated with respect to the diagnosis being predicted. The results of this case study are listed in Figure 5.2, which shows the five highest scored lab tests for each disease.

The first important observation we can make from this figure is that the set of laboratory tests the model attends to varies between the two diseases. As the figure shows, the highest attention score for the detection of diabetes is given to blood glucose level, which is a key predictor for diabetes. Similarly, the laboratory tests listed for kidney failure are highly indicative of this condition, as validated by a clinician.

Additionally, we observe a larger skewness in attention scores when predicting for diabetes, with glucose having a notably higher score than other labs, than we do when predicting for kidney disease, where attention scores are more evenly distributed. This can be attributed to the fact that kidney failure is indicated by multiple factors while blood glucose is a single key indicator of diabetes. Insights such as these can be highly beneficial in supporting the diagnosis decision process.

We next analyze the attention scores in the type-level aggregation process, where they represent the importance of various clinical record types in the prediction of different diseases. Figure 5.3 contains a box plot demonstrating the range of attention weights assigned to different type-level embeddings across the diseases in our test set. As we can see, the procedures and laboratory tests are overall our main predictors of diagnoses. However, there is more variance in procedure scores than in laboratory test scores, indicating that the predictive power of this category varies across diseases.

5.4 Conclusions

In this study, we introduced HTAD, an HIN based model incorporating a hierarchical attention mechanism for diagnosis prediction using EHRs. In HTAD, a patient representation is learned through the target-attentive aggregation of neighboring clinical records' embeddings, a process that allows distinguishing important record items for the prediction of a specific diagnosis. The novelty of this approach lies also in the interpretability it offers. Additionally, HTAD is capable of incorporating non-categorical records unused by past approaches. Experimental results demonstrate HTAD's superior performance compared to the previous state of the art methods and the interpretability of its predictions.

We suggest a few potential directions for future work. First, in this study, the target-aware selection process is done at a node-level. One possible direction for future work is the exploration of similar approaches for meta path-based relation selection. Another possible direction would be extending the model to other healthcare prediction tasks, and even incorporating it into multi-task learning methods. Finally, future work could investigate ways to incorporate information from multiple modalities in an end-to-end manner.

CHAPTER 6

Activity Recognition in Children: Personalization vs. Generalization

Physical activity is considered as an important health monitoring variable in diverse pediatric pervasive health applications including asthma management. Therefore, the development of an accurate activity recognition model can be of great value in achieving a better context of an individual's health. Despite the great advances in activity recognition in the past years, there has been a missing focus on building such models for children that are shown to have different motion patterns than adults. In this chapter, we present our smartwatch-based children activity tracker, trained on 20 children aged between 8 and 14 years old. We perform a comparative analysis of the prediction performance when personalized and generalized models are trained and propose a multi-task learning approach to combine the advantages of both approaches. Findings of this study can be beneficial in the development of other pediatric health monitoring applications as well.

6.1 Introduction

The long interest in Human Activity Recognition (HAR) using wearable motion sensors, has led to a large body of research and remarkable contributions to remote health monitoring [20]. As the motion in different body parts, including hip, waist, wrist, and foot can be indicative of activity, earlier HAR studies utilized multiple standalone sensor placements for activity recognition [57]. Although these studies concluded that combined sensor placement (e.g., on waist and wrist) can provide the best results, the special attention of researchers in recent years has been drawn to improving activity recognition models when using everyday devices such as smartphones, and more recently smartwatches, presented in studies such as [205, 128, 26, 188, 197]. This shift can be attributed to the rapid advancement of motion sensors on the smart devices and their widespread adoption by the public which allows achieveing continuous activity monitoring of users in their everyday life.

Among the many HAR models developed in the past years and surveyed in [196, 220, 168], few have worked with populations other than young adults, especially children. This is while continuous activity monitoring can be highly beneficial to various pediatrics health management applications, such as asthma, and [109] and is suggested as a possible way for prevention of health risks such as obesity and cardiovascular disorders in children [89, 14]. Prior studies on children activity monitoring are limited to few that mostly took place before the recent advancements of smart devices in motion sensing and their wide adoption in HAR research [71, 97, 105, 154]. Accordingly, although they demonstrate the poor transferability of established adult activity models to youngsters and the need for a children-focused model design, their approach in using standalone sensors placed on various body parts (e.g., ankle and waist) is limited for practical use in real-world settings. Therefore, there still remains a significant gap between adult and children activity models and there is a need for practical continuous activity monitoring solutions for children. As partly discussed in section 8.1.1, smartwatches are especially desirable in this setting as they offer undemanding data collection that can be used throughout the day and across different activities. Moreover, considering their widespread usage [55], they are deemed more cost-effective and usable.

Activity recognition models are generally trained with the assumption of similarity among activity patterns in the time of training and deployment. However, considering the observed variance among children and adults, a higher range of difference among children activity patterns can be expected as well. Hence, a comparative study across personalized and generalized models seems necessary. However, the limited size of labeled samples per person, in this study as well as real-world settings, pose some challenges on personalized model training. To tackle this, we propose employing multi-task learning to share the common knowledge among subjects while personalizing the model based on each subject's activity patterns.

In this chapter, we first present our study on building a smartwatch-based children activity recognition model and report the comparative performance results of several classical and state of the art pre-processing and data modeling techniques. Furthermore, we study the variance among children by comparing generalized and personalized prediction models. To tackle the challenge of label sparsity in person-specific modeling, we compare two transfer learning approaches, particularly simple sequential adaptation and multi-task learning and show that the later can achieve the best results for personalized activity modeling.

6.1.1 Contributions

Contributions of our work are as follows: 1) we introduce a smartwatch based activity recognition model for children, which to the best of our knowledge has been missing in prior pervasive health studies. 2) We perform a comparative analysis of generalized and personalized prediction models that can shed more light on the personalization requirements in this setting. 3) We propose a deep multi-task learning approach for addressing the personalization needs that arise in activity tracking, especially when working with children. We show that our approach obtains the best results compared to existing single-task training models.

6.2 Background

Inertial motion sensors are considered the most widely used devices for the study of activity recognition [156]. Triaxial accelerometer monitoring permits real-time measurement of a body part acceleration along three orthogonal axes and pairing it with a triaxial gyroscope for detection of rotation has brought the state of the art results of HAR modeling [196].

The general workflow of activity recognition models is similar to the one in pervasive health discussed in section 2.2.1. Therefore, in this section, we focus on providing an overview of prior studies in HAR, starting from strategies for building accurate prediction models using classical machine learning, followed by the recent success of deep learning models. We finally provide background on multi-task learning and its applications.

6.2.1 Classical Activity Recognition Models

Prior research in HAR has put the effort into improving various pre-processing and modeling steps in classical pervasive health data analysis. Signal segmentation and windowing of the motion signals can play a major role in the correct detection of activities. Therefore, earlier studies have focused on finding the best window length when using time-based windowing [25] or have worked on dynamic windowing approaches for addressing more complex scenarios such as overlapping activities [137, 191, 11, 219].

Recognizing the complexity of inertial signals, a great wealth of research has gone into finding and extracting the most advantageous set of features for HAR task. [77] is among early studies that outlined an all-inclusive pre-processing and feature extraction techniques for interpreting accelerometer data. This work is followed by later studies such as [33] that provide a thorough comparative analysis among the performance of HAR when different sets of features are used. This study also provides a publicly available smartphone-based activity data [17] that provides a full set of extracted features that can achieve over 90% accuracy in recognition of 6 activities with a simple prediction model.

6.2.2 Deep Learning Based Models

The recent success of Deep Neural Network models in other domains has drawn the researchers focus on their application on activity recognition. Convolutional and recurrent neural network models offer automatic embedding of signals into representative vectors that can replace the complex manual feature engineering techniques. Studies such as DeepSense [234] and [183] showed the superior performance of CNN in extracting representative features from time and frequency domain of the signal. Furthermore, RNN models, specifically BiLSTMs, have been shown by various studies reviewed in [220] to achieve the state of the art results in sensor-based activity recognition. These studies include [98] that gained 92.7%

F-score on Opportunity benchmark dataset [45] and [204] that achieved 88.87% on HAR public dataset [17]. The high performance of RNN models can be attributed to their ability in capturing temporal dependencies among signal readings that can carry key information regarding the activity performed. Studies using RNN, either used it on top of an initial feature extraction layer [234] or as a feature extractor and predictor on the raw signal itself [98], although the latter is shown to require more complex pre-processing steps to achieve state of the art results. The interested reader can find a more comprehensive list of studies on the role of deep learning in human activity recognition in two recent comprehensive surveys [220, 168]

6.2.3 Smartwatch-based models vs Smartphone-based models

Smartwatch based activity models have been recently approached by researchers in studies such as [188, 197] that provide a side by side comparison between the activity models relying on smartphones and smartwatches. Results of these studies suggest that smartphone-based models show higher range of prediction accuracy (with surprising 10% margin) in recognition of high-level activities such as running, walking, and sitting, when compared to smartwatch-based models, This is while smartwatches show superior performance on gesture-based activities such as detection of smoking and eating [198]. These observations can be explained by the fact that wrist-worn sensors may capture hand movements that are not directly correlated to the high-level activity performed. For instance, a person standing still may move his/her hands during this activity and introduce some complexity to prediction. In contrast, many smartphone-based activity recognition models rely on movement patterns of the phone when placed in a fixed position such as subject's waist [17], which can better capture the repetitive pattern of motion data compared to the wrist position. A combination of both devices is shown to bring the best of both worlds [197], however compromising comfortable wear. Having these in mind, one should select a suitable device for each application based on the trade-off between usability and performance.

6.2.4 Multi-task Learning

Multi-task learning (MTL) is an approach in machine learning in which a number of similar separate tasks are trained jointly so that the common knowledge required in learning them can be shared. In a deep neural network setting, this goal is often achieved by sharing the initial layers of the model among all tasks, while separating the top prediction layers for each of them. MTL has shown to bring great performance gain in many studies, part of which are reviewed in [238]. As discussed in [185], this observed gain in joint training of tasks compared to single-task training can be attributed to a number of features MTL offers. Firstly, MTL implicitly increases the data size for model training. This feature can be especially beneficial in the healthcare domain where labeled samples are commonly sparse. Furthermore, MTL improves the generalization of the model by extracting features of data that are informative for the prediction of various tasks. Finally, MTL acts as a regularizer for prevention of over-fitting.

In recent years, MTL has been applied in a number of pervasive health applications. [29, 121, 180, 102] are among studies that employed MTL for improving the prediction performance of similar health outcomes such as suicidal risk and mental health disorder. While the application of deep MTL for personalized prediction has been explored in other domains such as natural language processing [228] and speech recognition [164], it has not been well studied in improving the prediction based on health signals.

6.3 Data Collection

To collect children activity data, a smartwatch app was designed to record triaxial accelerometer and gyroscope signals in real-time and transfer the anonymized data to a web server for activity prediction. Figure 6.1 illustrates an overview of the data collection process. In this study, 25 children (10 girls and 15 boys) aged 8 to 14 years old participated to perform the following six activities: running, walking, standing, sitting, lying down, and stair climbing. Motion signals related to each activity were recorded with a frequency of 10 Hz for a duration

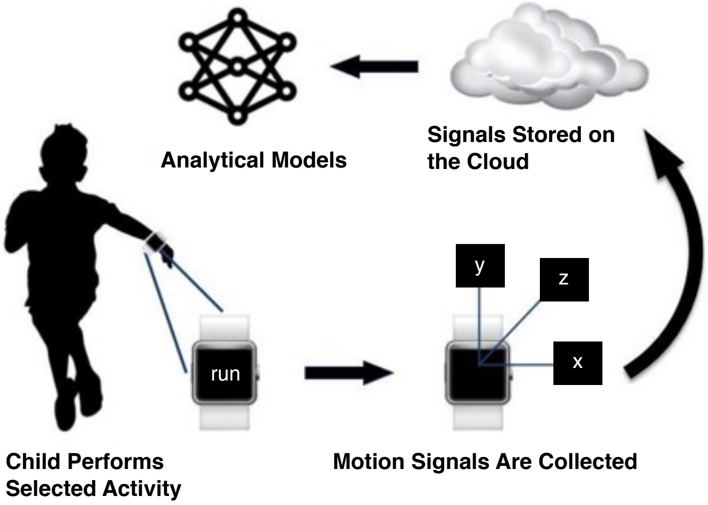

Analytical Models
Signals Stored on the Cloud
y
z
x
run
Child Performs Selected Activity
Motion Signals Are Collected

Figure 6.1: The data collection process for smartwatch-based activity monitoring.

of 10 minutes. Experiments took place in an open environment and children subjects were left with minimal interruptions to retain their natural movements and keep the dataset as close to real-world scenarios as possible. Data collection was manually stopped during time spans in which subjects required water break or needed to rest. Due to technical and experimental challenges involved when working with children, data from 5 participants contained a significant amount of noise and missing values which is not included in this study.

6.4 Methodology

In this section, we first review the pre-processing steps conducted on the collected dataset and then present the design of our main BiLSTM prediction model. We then explore the personalized model design and provide details of our proposed multi-task learning approach, along with fine-tuning that is employed as a baseline for evaluation of the MTL model.

6.4.1 Pre-processing

Figure 6.2 demonstrates the data processing and cleaning steps that we employed in this study. The pre-processing starts with removing noise from each one of 6 channels (x, y, z

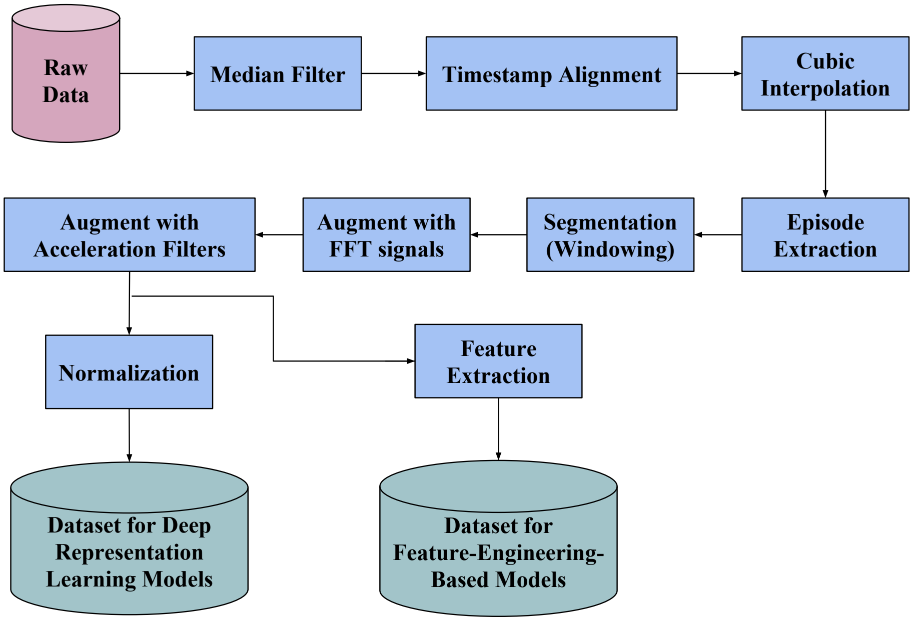


Figure 6.2: Pre-processing steps for data cleaning, quality improvement and augmentation.

per accelerometer and gyroscope sensors) by applying a median filter with kernel size of 3.

Next, there is a need to temporally fix and align the signal channels, mainly due to two challenges observed in smartwatch-based data collection. Namely, the sampling rate on the smartwatch can be loose and while the average number of readings per second remains close to the assigned frequency, the time difference between consecutive readings may vary. Besides, the sensor reading timestamps from the two sensors are not evenly matched. To address these, we first generate a 50 Hz equally spaced timestamp array, covering the whole data collection period. Afterward, each sensor reading is assigned into the closest timestamp and all missing values are injected using a cubic spline interpolation method [61], which is shown to improve the results of activity recognition models when the sampling rate is increased [137]. The 50 Hz frequency was also selected as it is considered as a confident range that can capture human movement [204].

Having all sensors aligned, we extract episodes of data that none of the signals were missing for more than 5 seconds. A fixed-length sliding window segmentation process was then performed on each of the 6 signals to generate temporally aligned windows. The next two steps are aimed at the integration of extra information to the processed dataset. In particular, Fast Fourier transform of all channels are added as new signals to capture the frequency domain information as well. Furthermore, 3 new signals are introduced by calculating the third order Butterworth filter [189] of each accelerometer channel with cutoff frequency of 0.3 Hz. This setting is shown to cancel the effects of gravity on acceleration and reflect the original acceleration happened due to hand movements [204]. Finally, the windows are normalized and scaled into $(0, 1)$ range.

To compare deep neural network models to those relying on feature engineering, each set of matching windows from 15 signals generated above, were mapped into a single representative vector by extracting the features employed in a benchmarked dataset [204].

6.4.2 Prediction Model

Following the proven success of Recurrent Neural Network in time-series modeling and relying on their ability in capturing temporal dependencies, we employ a Bi-directional LSTM (BiLSTM) model to capture representative local and temporal features of windows extracted from motion signals. The structure of the designed network is shown in Figure 6.3.

having the 15 aligned signals obtained by approaches discussed in section 6.4.1, 15 temporally matching windows can be extracted in each prediction step, which are denoted as: $W = W_1, W_2, ..., w_{15}$. Each window w_i wraps l sensor readings, represented as: $W_i = [x_{i,1}, x_{i,2}, x_{i,3}, ..., x_{i,l}]$. Given these, the input vector to the ith LSTM cell of our model is formed by combining the ith readings of windows in W forming the vector: $I_i = [x_{1,i}, x_{2,i}, x_{3,i}, ..., x_{15,i}]$.

The l input vectors constructed as above are fed into two layers of LSTM cells, one in forward and one in reverse order. The concatenation of the hidden states of the last LSTM cells in the forward and backward layers forms the representation of W. This representation is then fed into a fully connected layer followed by a softmax layer for activity prediction.

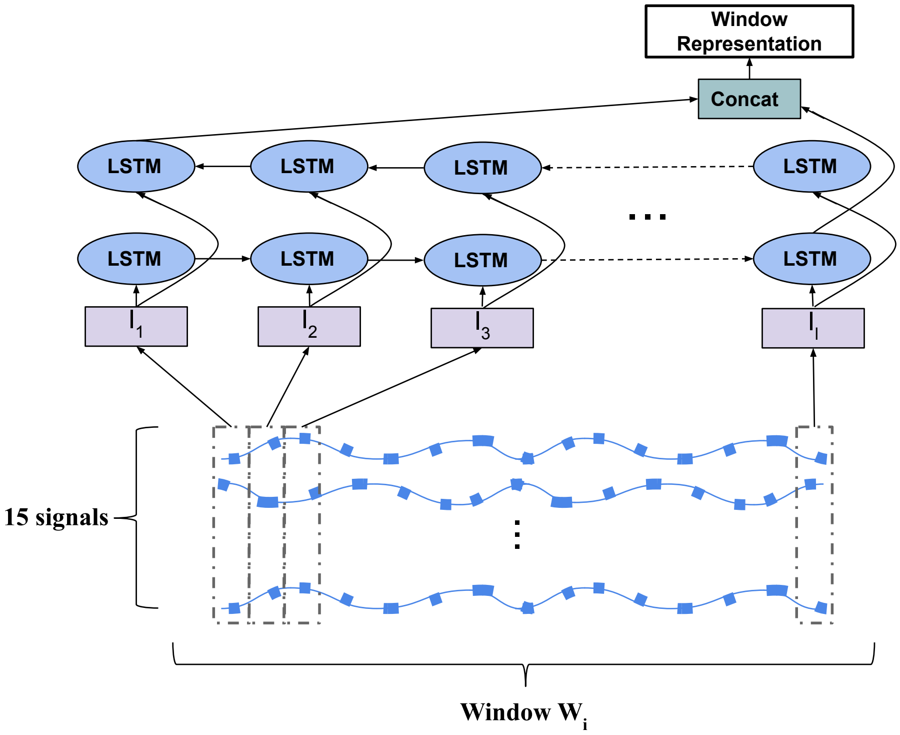


Figure 6.3: BiLSTM model for learning a single representation from 15 motion signals.

6.4.3 Person-Specific Model Design

To study and compare person-specific and generalized activity recognition model for children, the first straightforward approach is training a model for each subject. However, recognizing the limitation of obtaining labeled data samples for each subject in real-world settings (and in our study), we explore transfer learning approaches with the aim of leveraging the labeled data available from other subjects to develop general knowledge and tune the model to activity patterns of a new subject. In particular, we examine the two following methods:

- Fine Tuning: In this approach, the prediction model is first pre-trained on a set of labeled training samples. Afterward, the model is adapted to the subject by continuing the training on all or a set of layers. In this study, we freeze the BiLSTM layers and continue the training only on the last prediction layer.

- Multi-task Learning: In contrast to fine-tuning that performs the model transfer in a sequential manner, in this scenario, we allow the transfer of knowledge among subjects

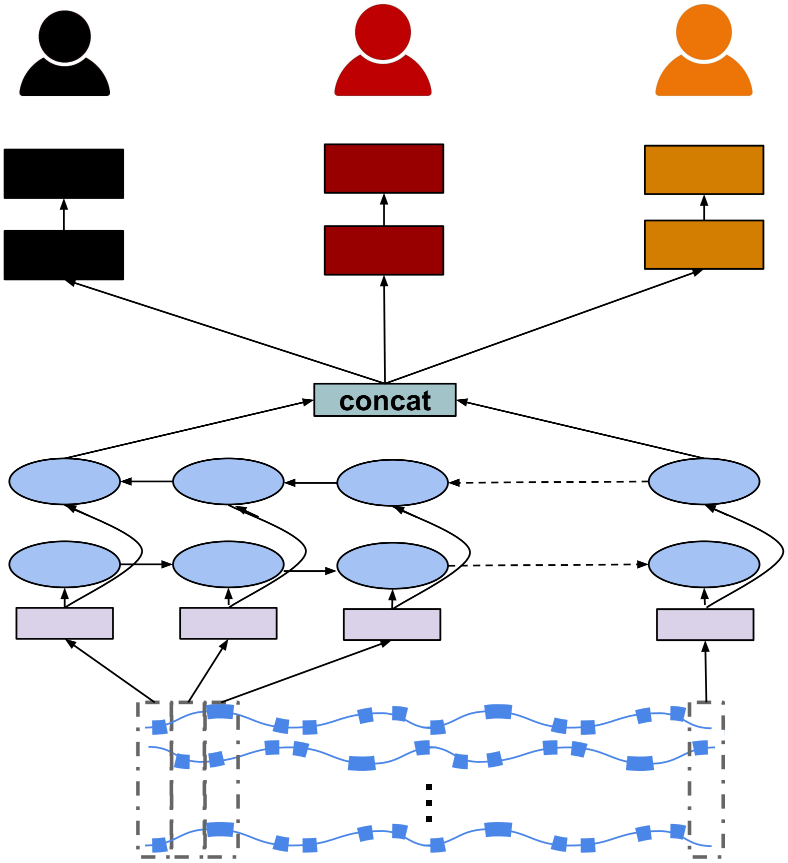


Figure 6.4: The general schema of our personalized multi-task learning model.

in a joint-learning setting. Recognizing the fact that various subjects share the basic activity patterns, each subject is regarded as a separate task that can be trained jointly to others. For this aim, the representation learning layers of the BiLSTM model is shared among all subjects, while keeping the prediction layers specific to each person. Two fully connected layers followed by a softmax activation are used in design of these layers. A general schema of this model is illustrated in Figure 6.4. As discussed before, this design pattern not only allows the model to use data from all subjects for learning an informative representation of motion signals, it tackles the over-fitting challenge present in fine-tuning approach by acting as a regularizer.

6.5 Experiments

In this section, we report the results of several experiments performed in this study. We first compare the performance of the proposed activity recognition model against various pre-processing and modeling methodologies. We then analyze the effect of personalized prediction on the performance improvement of our activity recognition model.

6.5.1 Implementation Details

To train our models we employed a window size of 8 seconds with 50% overlap. The hidden layer size was set to 32 in the BiLSTM model and to 16 in last prediction layer. The model is trained with Adam optimizer. The data are split into disjoint train and test sets with 70% to 30% ratio.

6.5.2 Comparative Analysis of Generalized Activity Models

Focusing on building a general children activity recognition model, our first experiment reports the comparative results among baseline models with different pre-processing and representation learning approaches. In particular, the performance of the proposed BiLSTM model is compared against two shallow models using feature engineering, as well as a BiLSTM model but when the augmented signals introduced in section 6.4.1 are not used. The baseline models are as follows:

- Feature Engineering + Random Forest (FeatEngRF): In this approach, window representations are generated by manual feature extraction from the all 15 original and augmented signals. The representation of each window is then fed into a Random Forest model for activity prediction.

- Feature Engineering + FFN (FeatEngFNN): This model is different than above in using a three layer fully connected neural network model on window representations for activity prediction.

- Bi-LSTM Representation Learning/ No signal augmentation (BiLSTM/NoAug): This approach employs the proposed BiLSTM model, however, ignores the 9 additional augmented signals to evaluate their impact on the model performance.

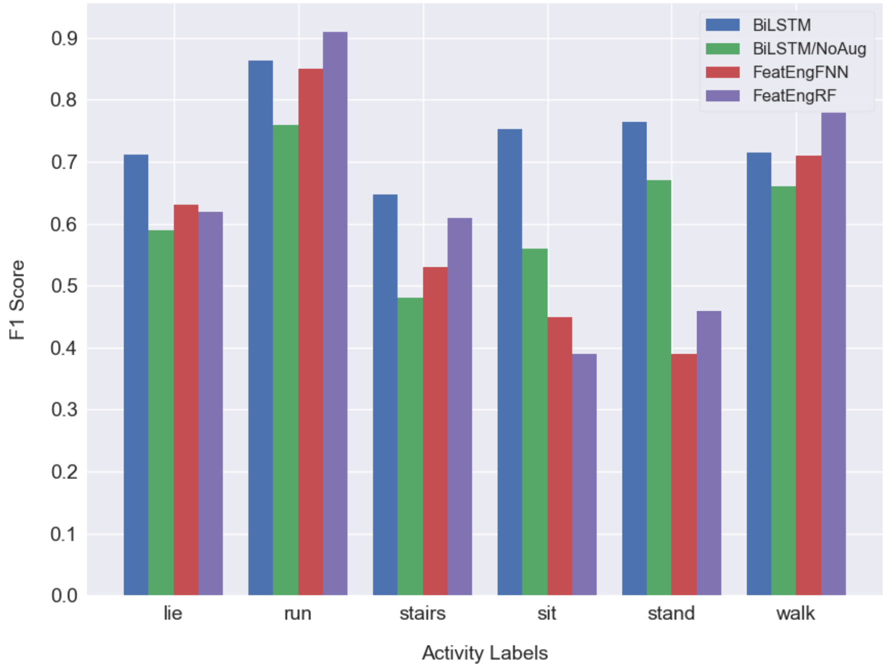
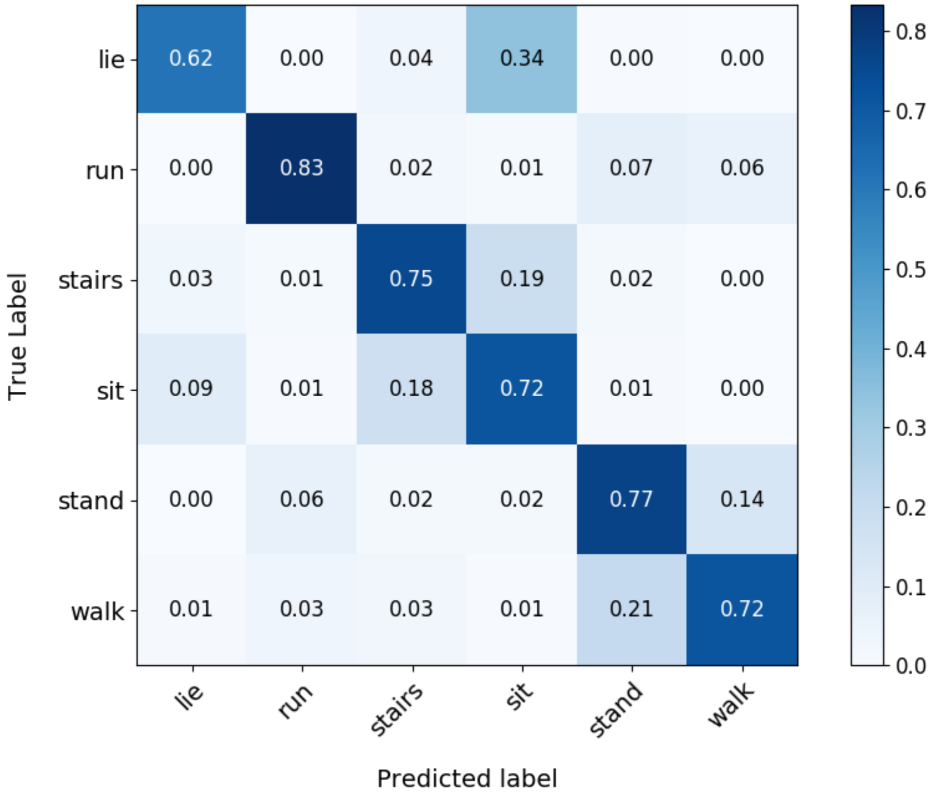


Figure 6.5: Comparative results of 4 activity prediction models in group training setting.

Figure 6.5 summarizes the results of this experiment. As the figure suggests, the BiLSTM model delivered considerably more consistent performance across all activities compared to other methods, specifically FeatEngRF and FeatEngRNN. These models although achieved a marginally higher performance in prediction of the run and walk activities, showed markedly inferior results on other activities especially standing, sitting and stair climbing. This can indicate that the established feature engineering approaches for HAR may not be easily applicable to new domains and proves the advantage of using automatic representation learning approaches such as the proposed BiLSTM model. Furthermore, comparison of BiLSTM and BiLSTM/NoAug results can clearly show the advantage of signal augmentation with frequency domain and pure hand acceleration information. The BiLSTM model is used in later experiments of this chapter.

The Confusion Matrix of the BiLSTM model is illustrated in Figure 6.6 and conveys important information regarding the prediction challenges. It is noticeable that the model faces trouble in distinguishing the lie and sit activities. This confusion can be partly attributed to the fact that the movement of the wrist in these two activities can be very similar to each other when the hand is left on a horizontal position when sitting. Other confusions are not as significant but observable among walking and standing, as well as stair climbing and sitting activities.



True Label	lie	run	stairs	sit	stand	walk
lie	0.62	0.00	0.04	0.34	0.00	0.00
run	0.00	0.83	0.02	0.01	0.07	0.06
stairs	0.03	0.01	0.75	0.19	0.02	0.00
sit	0.09	0.01	0.18	0.72	0.01	0.00
stand	0.00	0.06	0.02	0.02	0.77	0.14
walk	0.01	0.03	0.03	0.01	0.21	0.72

Figure 6.6: Confusion matrix of the BiLSTM model.

6.5.3 Person-Specific Analysis

This experiment is aimed at examining the hypothesis that variability in children activity pattern can bring sub-optimal performance in a general activity model. To test this hypothesis we designed and compared four scenarios related to person-specific activity recognition, as listed below:

- Person-specific training: In this scenario, the first 70% of each child's activity data is considered as train set and the model is tested on the remaining 30%.

- Leave-one-out: This model is trained to provide a baseline for the performance of the general training model on each child. The model is trained on all children except the child under the study and is tested on the last 30% of his/her data for a fair comparison to person-specific models.

- Fine-tuning: This approach relies on using the general trained model but tuning it to each subject. For this aim, the training on the Leave-one-out model corresponding to each child is restarted using the initial 30% of the child data. The model is then evaluated as above on the last 30%. To avoid over-fitting to the new data early stopping was employed.

- Multi-task learning: In this approach, the first 70% of all children data is used for training the shared and person-specific layers. The model is then evaluated on the last 30% of each child's data.

Table 6.1 lists the results of aforementioned models for all 20 subjects under the study. Results of this table indicate two distinguishable behavior of models. For a number of subjects (including 2, 4, 8 10, and 17) we can observe that the person-specific model performs notably better than the leave-one-out model. This can suggest that these participants show highly person-specific activity patterns that a general model cannot detect. Results of the fine-tuned model for these subjects also show higher performance than the leave-one-out model, showing that the fine-tuned model was able to capture person-specific activity patterns that were not learned by the general model. These observations can support the hypothesis of high variation in children activity pattern.

On the other hand, the leave-one-out model leads to higher performance for participants that arguably show more general activity pattern when compared to person-specific models,

Model Subject	Person-specific	Leave-one-out	Fine-tune	Multi-task learning
1	0.65	0.72	0.71	**0.72**
2	0.58	0.52	**0.61**	0.56
3	0.33	**0.72**	0.61	0.71
4	**0.82**	0.71	0.79	0.77
5	0.63	0.78	0.68	**0.79**
6	0.43	0.67	0.63	**0.69**
7	0.53	0.70	0.69	**0.71**
8	0.71	0.69	**0.81**	0.77
9	0.69	0.72	0.68	**0.73**
10	**0.83**	0.63	0.75	0.79
11	0.65	0.74	0.69	**0.75**
12	0.76	0.75	0.74	**0.76**
13	0.58	0.69	0.61	**0.71**
14	0.66	0.57	**0.72**	0.67
15	0.68	0.71	0.62	**0.71**
16	0.61	0.67	0.58	**0.68**
17	0.69	0.56	**0.75**	0.67
18	0.56	0.63	0.57	**0.64**
19	0.64	**0.73**	0.69	0.71
20	0.65	0.66	0.61	**0.67**

Table 6.1: Comparison of results among 4 person-specific activity prediction models.

partly due to the larger volume of training data and ability to capture a wider range of general activity patterns. Importantly, fine-tuned model show lower performance than the leave-one-out model for many participants which can be largely attributed to overfitting to the small new dataset.

The multi-task learning approach shows competitive results compared to the best result of all models as it allows tuning to the subject-specific patterns to some extent and can capture general patterns with more resistance against overfitting. We can especially observe a visible gain in participants that showed better results in person-specific models.

6.6 Discussion

A comparison between the results of this study and prior studies on adult activity recognition shows a considerable performance gap which can be attributed to a number of factors. First, the sampling rate adopted in this study stays in the lower range of suggested rates in prior research, mainly due to the current limitations of smartwatches and the existing trade-off of battery consumption and sensor data accuracy. The 10 Hz sampling rate allows BREATHE to run for a full day of data collection while maintaining a reasonable performance, especially for activity intensity prediction. Second, as this gap remains on person-specific models as well, we can conclude that smartwatch-based activity patterns of children can be noisy (due to uncorrelated hand movements) and more complex to analyze, when compared to adults. As discussed earlier, smartwatch-based adult activity recognition studies also report poorer results on the performance

Future directions include a more thorough study of the trade-off between sampling rate and battery consumption and how this can be optimized for children. This is while the constant improvement of smart devices batteries can help in reducing this trade-off as well.

CHAPTER 7

Deep Domain Adaptation in Pervasive Health

Among the major challenges in building supervised predictive models in pervasive health, is adapting them to new individuals or groups of people. This is not trivial largely due to possible differences in the distribution of data in training time and real-world deployments. For instance, an activity recognition model trained on adults may significantly lose accuracy when deployed to children or elderly people. The classical approach of addressing this issue is the collection of new annotated data from the representative groups under each study. However, this approach is very time-consuming and hinders the design of scalable pervasive health systems. In this chapter, we aim to tackle this problem by building upon the recent advancements in Domain Adaptation, which aims at transferring a model trained on a labeled dataset to a new unlabeled one, while taking the possible differences in the data distributions into account. To show the benefits of our approach, we transfer an activity recognition model trained on a popular adult dataset to children. We show that direct use of the adult model on children loses 25.2% F1-score compared to baseline on average, while our proposed transfer approach reduces this to 9%.

7.1 Introduction

Machine Learning methods used by many health studies rely heavily on annotated training datasets and can only perform as expected when the distribution of the data in the time of training and deployment is the same [209]. However, this assumption does not always hold in the real-world health applications, where the distribution of collected sensor data can vary based on various factors such as age, gender, or sensor placement [36].

A common approach in pervasive health applications is training models by collecting data from a group of representative individuals. However, this approach is not desirable as it introduces challenges when studying minorities and more importantly, it still fails to capture many person-specific differences. Annotation of each individual's data is not practically feasible as well, largely due to the real-time nature of data, its scale, and the significant effort required.

Therefore, there is a strong need for building transferable models that allow the deployment of pervasive health systems among individuals from various backgrounds and health conditions, denoted as new domains. The challenge of transferring a model to a new domain with no available labels has arisen in other applications, such as image recognition [184] as well and an appealing approach in addressing it has been Domain Adaptation (DA) [209]. In this approach, the model trained on a source domain (a group of individuals in our case) is transferred to solve the same problem in a target domain (a new individual), for which a considerable amount of data is available but the labels are missing. This setting is perfectly aligned with features of the pervasive health field since a huge amount of unlabeled sensor data is generally available from continuous monitoring of an individual while the size of the labeled data is limited.

Although applications of Domain Adaptation has expanded to diverse areas such as natural language processing [54], voice recognition [63], and brain-computer interaction [227], pervasive health has lagged behind in adopting this technique, especially after the rapid emergence of deep learning models. Prior studies have mainly relied on approaches such as fine-tuning the model to each subject [220] which is shown to lead to over-fitting and is only possible when labeled data for each subject is available.

In this study, we propose a deep sequential domain adaptation model for classification of health signals. Our model which is inspired by [209], tries to learn domain invariant features of data to allow label-free transfer among different datasets. In particular, we employ a Bidirectional LSTM Recurrent Neural Network for extracting representative features of the input signal along with Maximum Mean Discrepancy (MMD) [93] loss to minimize confusion between the source and target domains. We evaluate our model on activity recognition task and with the aim of transferring a model from a publicly available dataset for adult activity recognition to one we collected for children.

7.1.1 Contributions

Our contributions in this study are: 1) we propose a domain adaption framework to address the important and existing challenge of poor model generalizability in pervasive healthcare. Our proposed model relies on a deep sequential design for working with time-series data. 2) We demonstrate that our model is able to effectively transfer an activity recognition model between two groups having different data distributions due to the difference in age-group and data collection device type. 3) Our approach can be extended to many applications in pervasive health to extend the reach and inclusion of the designed solutions.

7.2 Background

In this section, we first provide background on Domain Adaptation and a general overview of its applications in other areas. We then review prior studies in wireless health that have approached the problem of model transfer.

7.2.1 Domain Adaptation

In many applications, a model trained on a dataset may not be generalizable to other datasets for solving a similar problem, due to the difference in data distributions. Domain Adaptation (DA) [209] tries to tackle this problem and transfer the knowledge from a labeled source domain to an unlabeled or sparsely labeled target domain.

A classic approach used by early studies was fine-tuning the source model with a small set of labeled data from the target domain [170]. However, this approach introduces over-fitting problems, especially in deep models when the labeled target dataset is small [184]. Later attempts were focused on different ways of building a shared representation space between the source and target domain. This was done by either learning a transformation from the source to the target domain or from both domains into space with a similar distribution [86].

With the emergence of deep learning and feasibility of automatic feature extraction, prior ideas were extended to extract features from the source and target domains so that they would map into a domain-invariant feature space. Models following this idea generally introduce an additional domain discrepancy loss term to the main objective function of the network to force the representations of the two domains to follow the same distribution. A widely adopted loss term is Maximum Mean Discrepancy (MMD) [93], which is utilized to measure the similarity of two distributions. [209] was among the first studies that used this approach to learn domain-invariant features from images of office appliances collected from three different domains: Amazon.com, a webcam, and an SLR camera. This study used shared network weights among domains and a single layer of discrepancy loss. On top of this study, other works proposed different architectures by adding extra loss terms in different layers [143], using sequential models [138], and using separate streams of data for source and target domain [184].

Following success of Domain Adaptation in image recognition, applications of it has been rapidly expanded in recent years to other areas such as voice recognition in determining individual emotions [63], natural language processing in improving machine translation [54], and brain computer interaction in calibration of the system [227]

In the domain of health signal processing using deep models, however, there has been a missing focus on this challenge. One study [36] has recently approached the problem of domain adaptation in ECG signal processing, however, with a focus on unsupervised signal reconstruction and anomaly detection. This is while in this study we aim to test feasibility of domain adaptation classification of health signal especially in application of activity recognition.

7.3 Methodology

In this section, we review our approach in building a domain invariant model using a labeled source and an unlabeled or partially labeled target dataset. This section starts with the problem definition and continues with providing design details of our proposed model.

7.3.1 Problem Definition

We are given a time series source dataset, D^s, which is segmented into fixed length windows denoted as $W^s = \{w_1^s, w_2^s, w_3^s, ..., w_n^s\}$ and is coupled with the corresponding label set $Y^s = \{y_1^s, y_2^s, ..., y_n^s\}$. The target time series dataset D^t, collected for a similar task but in a different domain, is also provided as a set of segmented windows $W^t = (w_1^t, w_2^t, w_3^t, ..., w_m^t)$, however, the label set Y^t is not be provided or we have $\|Y^t\| \ll \|Y^s\|$. Recognizing the motivations of our study, we also assume that the marginal probability distribution of the two domains do not match and we have $P(W^s) \neq P(W^t)$.

Our goal is to transfer the source prediction model to the target domain with minimal access to the later domain labels. For this aim, a transformation function F is trained to map W_s and W_t into a domain invariant feature space with similar probability distribution: $P(F(W^s)) = P(F(W^t))$. Consequently, the prediction model trained on the transformed source domain, $f(Y^s, F(W^s))$, can be used for predictions on the transformed target domain $F(D^t)$ as well. The overall design of our model is illustrated in Figure 7.1.

7.3.2 General Prediction Model

To learn the feature extraction function F, we rely on a Recurrent Neural Network (RNN) model, particularly a Bi-directional LSTM (BiLSTM) similar to the one introduced in section 6.4.2 with shared weights between the source and target domains. The motion signals from both domains are fed into the BiLSTM model to extract local and temporal information and encode them into a fixed length feature vector. These representations are used to train the prediction function f, by minimizing the prediction loss L_c on all labeled samples

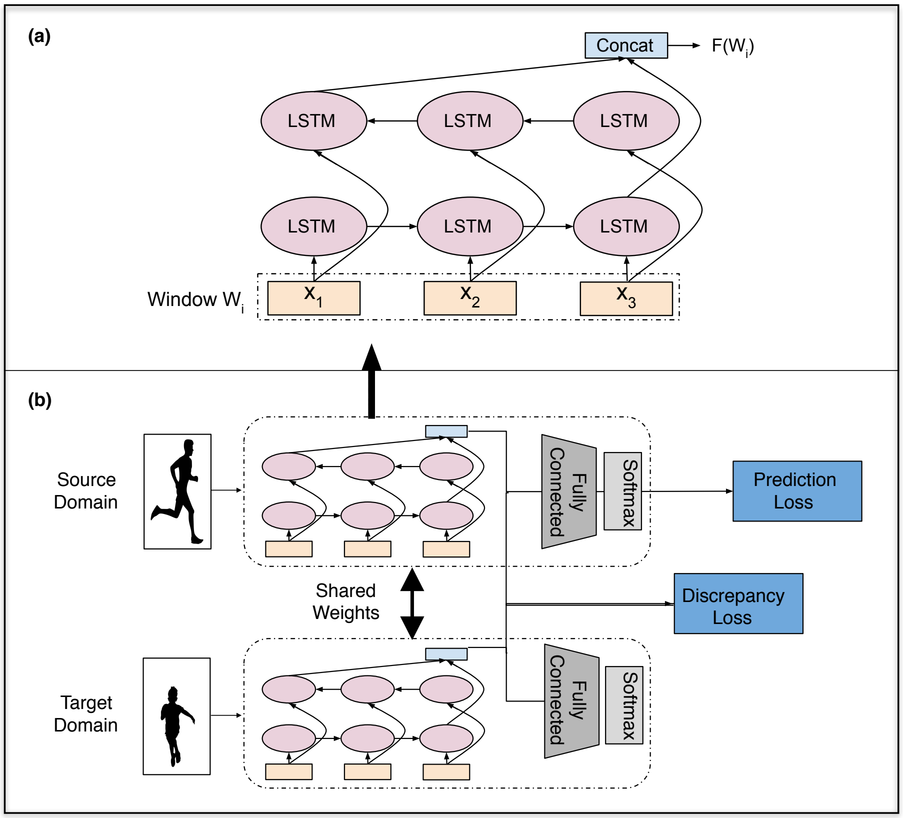


Figure 7.1: Structure of the BILSTM prediction model (a) and discrepancy loss layer between source and target domain (b)..

largely from W^s. For this aim, the transformed input samples by BiLSTM are fed into a fully connected layer followed by a softmax activation. Recognizing the dependence of learned representations to the objective of the model and optimization functions applied, there is a need to add an objective for minimizing the discrepancy between the representations learned from the target and source domains.

7.3.3 Domain Invariant Feature Training

The transformation function F obtained from the sequential model does not guarantee invariance between the source and target domains. In other words, if we feed W^s and W^t windows as input to the BiLSTM to achieve corresponding representation sets $F(W^s)$ and $F(W^t)$, the two may not share the same space. Consequently, the prediction function f trained on $F(W^s)$ cannot be safely transferred to $F(W^t)$. To address this, we add an additional domain discrepancy loss before the prediction layers to maximize the similarity between the distributions of $F(W^s)$ and $F(W^t)$ (Figure 7.1(b)).

In particular, we employ Maximum Mean Discrepancy (MMD) as a similarity measure between source and target domain representation distributions denoted as P_s and P_t. MMD is empirically estimated as the distance between the mean of sample representations in each domain, after mapping them to Reproducing Kernel Hilbert Space (RKHS) denoted as $\mathcal{H}$:

$$MMD^2(P_s, P_t) = \left\| \frac{1}{|W^s|} \sum_{w_s \in W^s} \phi(F(w_s)) - \frac{1}{|W^t|} \sum_{w_t \in W^t} \phi(F(w_t)) \right\|_{\mathcal{H}}^2 \tag{7.1}$$

Here F is our representation learning function and ϕ is a mapping function to RKHS which is commonly defined as a kernel function k. Accordingly, we can write the squared MMD by expanding the above formula and using the kernel trick as:

$$MMD^2(P_s, P_t) = \sum_{w_s, w_s^{'}} \frac{k(F(w_s), F(w_s^{'}))}{|W^s|^2} - 2 \sum_{w_s, w_t} \frac{k(F(w_s), F(w_t))}{|W^s|^2 |W^t|^2} + \sum_{w_t, w_t^{'}} \frac{k(F(w_t), F(w_t^{'}))}{|W^s|^2} \tag{7.2}$$

In this study, we employ RBF kernel function with a bandwidth of 1 as suggested by [185]. Based on this definition, we expect lower MMD scores to show higher similarity between $F(W^s)$ and $F(W^t)$. Therefore, the score can be directly used as a loss term (L_{MMD}) in model optimization.

7.3.4 Learning Objective

If we jointly optimize the BiLSTM model on the labeled data and minimize the discrepancy loss between D^s and D^t, the model is hypothesized to learn domain-invariant features from both domains, enabling the use of the prediction function f for D^t as well. Therefore, in training our model we optimize the following loss function:

$$L = L_c(W^s, Y^s) + L_{MMD}(W^s, W^t) + \lambda_r L_r \tag{7.3}$$

Where L_c is the classification loss, L_{MMD} is the MMD discrepancy loss and L_r is the regularization term on trainable weights, weighted by coefficient λ_r.

To train this network in an end to end manner, we feed it with mini-batches of data containing samples from both source and target domains. The labeled samples of each batch contribute to reducing the classification loss, while all samples from both domains are used to optimize the MMD score.

7.4 Experiments

In this section, we demonstrate the effectiveness of our approach in transferring a time-series health model between two domains and we report results on the case of transferring an adult activity recognition model to children. One level of adaptation complexity is inherent in the experiments of this study as the adult dataset is collected using a smartphone, while the child data comes from a wrist-worn smartwatch.

7.4.1 Datasets and Pre-processing

As our proposed model uses a shared design between the source and target data, we need to address a few differences between our two datasets. We use UCI-HAR [204] which is an open-sourced dataset for adult activity recognition as our source (labeled) domain. In this dataset, accelerometer and gyroscope signals are collected with a frequency of 50 Hz from 30 adult participants using a smartphone mounted on their waist. Our target domain dataset is similar to the one used in section 6.5.2 in which 20 child subjects used a wearable smartwatch to collect motion signals with a frequency of 10 Hz.

There are a number of steps required to match the two datasets to be fed into a shared model. Recognizing the lower data frequency of children dataset, we apply data alignment and interpolation discussed in section 6.4.1. Afterward, we follow all the discussed pre-processing steps on both datasets. Finally, to match the list of activities between the two domains, we use the following shared set of activities: walking, sitting, standing, stair climbing, and lying down. It is worth mentioning that stair climbing from UCI-HAR is obtained by combining upstairs and downstairs climbing sets and is down-sampled to be in balance with other activities.

7.4.2 Results

We first share the results of training the prediction model on the source dataset. Figure 7.2 shows the the performance of activity prediction using the introduced BiLSTM model (Section 6.4.2) on the UCI-HAR dataset.

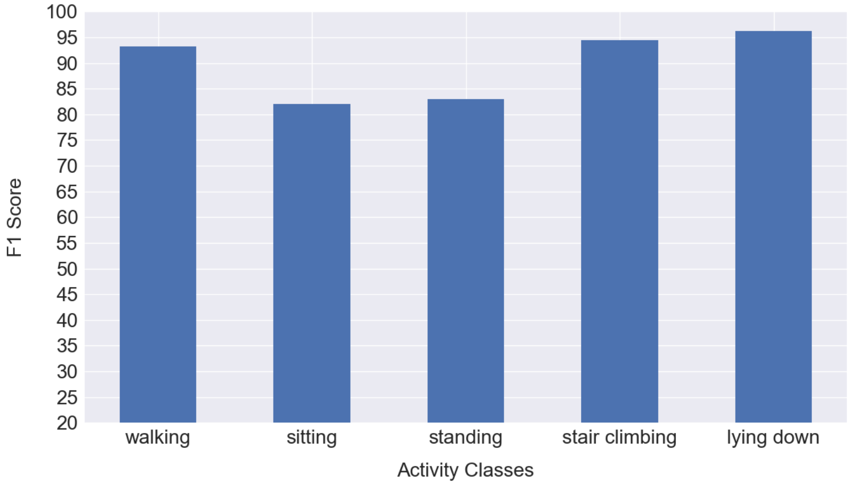


Figure 7.2: BiLSTM activity prediction results on the source domain.

To validate the benefits of our proposed approach in transferring the knowledge between domains, our first experiment compares the results of three child activity recognition models in the following scenarios:

- **Supervised Ground Truth**: To report a fair evaluation of the domain-adaption results, we consider the performance of the supervised baseline for children activity recognition. For this aim, the supervised BiLSTM model is trained on the 70% of children data and is evaluated on the remaining 30%, similar to section 6.5.2.

- **MMD Transfer**: Our proposed approach which learns a domain-invariant feature set

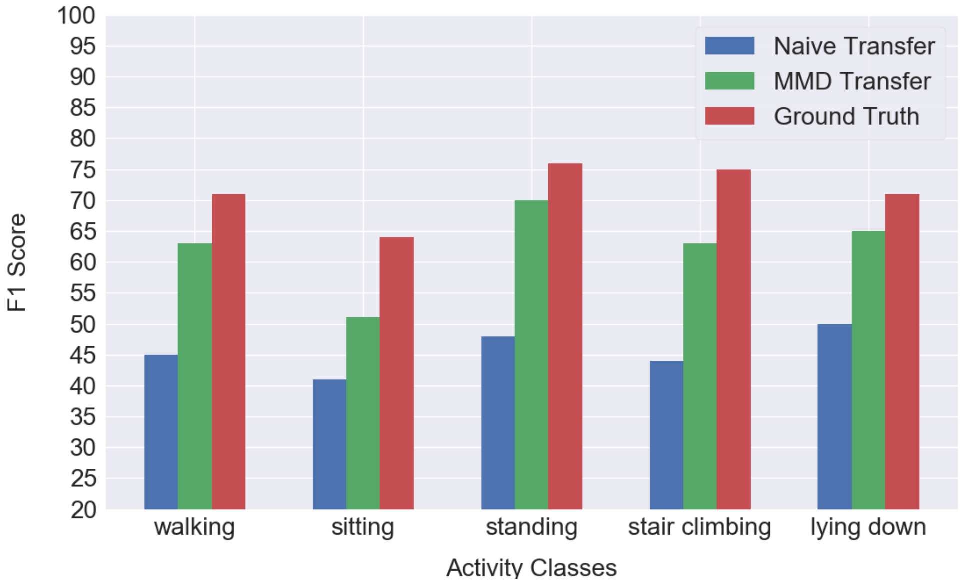


Figure 7.3: Comparative analysis of activity prediction.

between adults and children and makes predictions only by using labels of the adult dataset. This model utilizes 70% of unlabeled children data and is evaluated on the same test samples as above.

- **Naive Transfer**: In this scenario, the BiLSTM model is trained on the adult activity dataset and is directly used for prediction on the children with no domain adaptation considerations. Results of this setup will show how well or poorly distributions of the source and target domain match. This setup shares the same test samples as the first scenario as well.

Results of this experiment in terms of activity-scores are illustrated in Figure 7.3. As the figure suggests, the model trained on the adult dataset shows poor performance on the children data, confirming the difference in the distributions of these two domains and the need for domain adaptation. We can especially observe lower performance for walking and stair climbing activities that involve greater body part movements.

Furthermore, The MMD-Transfer model shows significant performance gain over the Naive-Transfer, reducing the performance gap compared to baseline from 25.2% to 9%. We

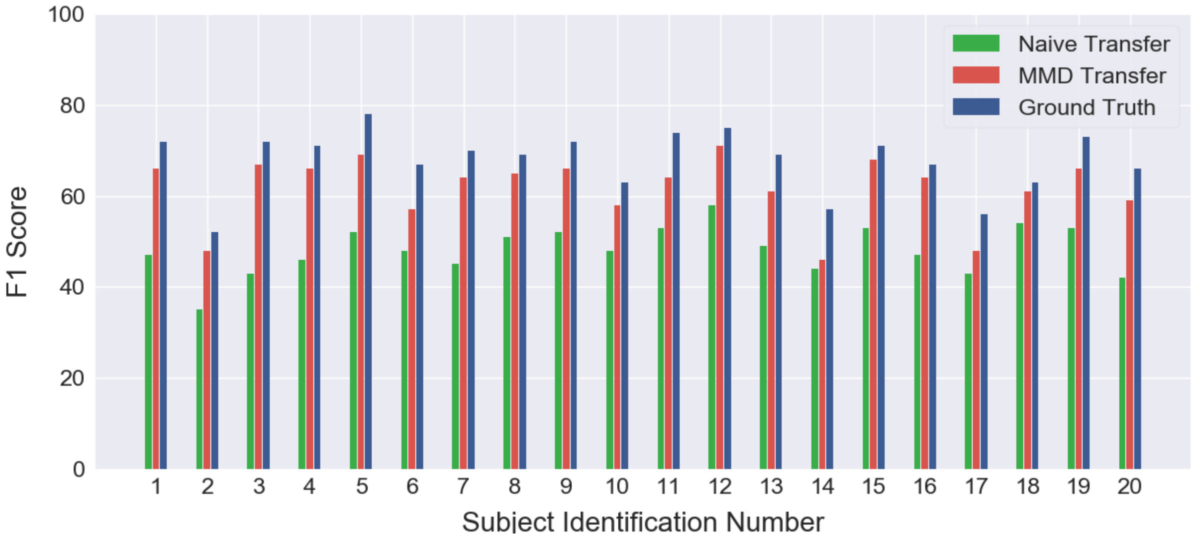


Figure 7.4: Comparative analysis of activity prediction.

can also observe that this gap becomes especially small for more steady activities. This can be partly due to the fact that movement patterns on a smartwatch are more diverse than a smartphone fixed on the waist.

Figure 7.4, reports the result of child-specific activity recognition using the MMD-Transfer model and provides comparative results of Naive-Transfer and supervised ground truth models as well. In an ideal real-world deployment scenario, the unlabeled target domain data can be easily collected from the child's activity records. However, as the duration of activity records for each subject is limited in this study, we included data from other subjects as well. It is worth mentioning that the supervised ground truth model for each subject in this experiment is trained in a leave-one-out setting. Similar to the last experiment, we can notice a noticeable performance gain using MMD-Transfer model for each subject.

Previous results were achieved in scenarios where the target labels were completely ignored. However, small supervision of target domain is hypothesized to improve the domain adaptation performance. To evaluate this, we gradually increase the ratio of labeled samples in the target domain from 10% to 100% in an exponential manner. Results, of this experiment, is listed in Figure 7.5. As it can be inferred, we can observe a performance jump after inclusion of 10% of target domain samples into the training model. We can also observe that

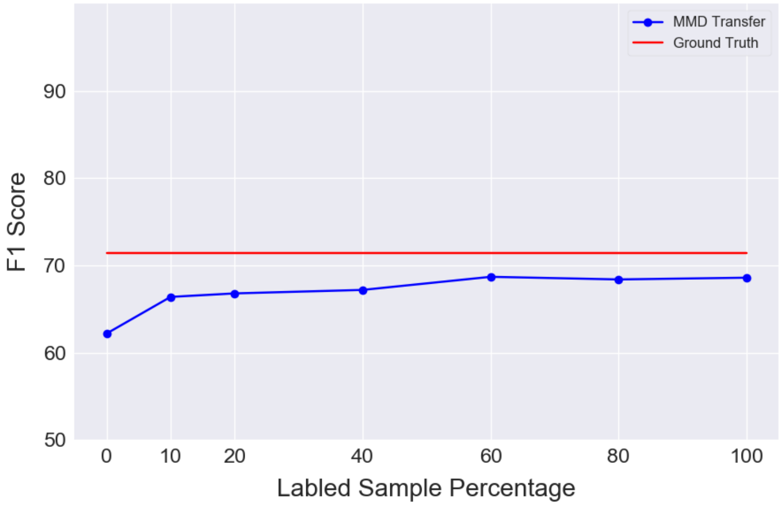


Figure 7.5: Domain Adaptation Performance in Presence of Labels.

the performance gain reduces afterward which can further indicate the significance of our approach in utilizing a small set of labels for achieving a competitive prediction performance to the supervised model trained on the target domain.

7.5 Conclusion

In this chapter, we introduced the significance of domain adaptation in the pervasive health field and proposed a sequential deep model for the transfer of activity recognition from adults to children. Results of our study show the promise of this approach which can be extended to other applications as well. As future directions for this study, we suggest exploration of this framework in broader applications and inclusion of deeper layers for reducing distribution discrepancy between the source and target domains

CHAPTER 8

BREATHE: A Pervasive Health Solution For Improvement of Pediatric Asthma Management

BREATHE (Biomedical REAl-Time Health Evaluation) is our pervasive health solution for improvement of pediatric asthma management. BREATHE provides a flexible framework for collection and integration of information from diverse sources including wearable sensors, online environmental data, and historical health records to obtain an expressive model of an individual's asthma exacerbation triggers and context. Our efforts in BREATHE are part of a larger initiative, PRISMS (Pediatric Research with Integrated Sensor Monitoring Systems), launched by National Institutes of Health in 2014. This chapter overviews BREATHEs architecture and components, followed by initial results from a study on 20 asthmatic children.

8.1 System Overview

BREATHE system is designed as a real-time and end-to-end framework with the main goals of flexible sensor integration, suitable data unification, and real-time analysis. BREATHE connects various data sources to analytical engines and keeps an on-demand interaction with children and their caregivers for providing health feedback or collecting self-assessment information. Different components and the complete life cycle of data in BREATHE are shown in Figure 8.1.

Successful modeling of asthma is dependent on how well data sources can cover different environmental and physiological variables that affect or indicate an asthma exacerbation

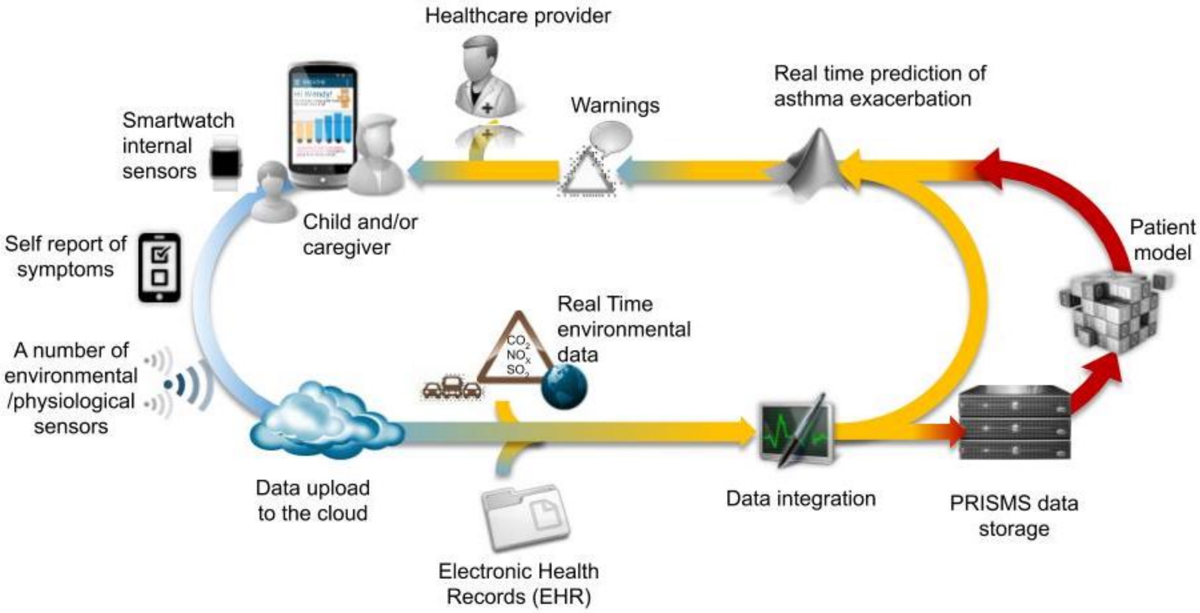

Healthcare provider
Real time prediction of asthma exacerbation
Warnings
Smartwatch internal sensors
Child and/or caregiver
Patient model
Self report of symptoms
Real Time environmental data
CO_2
NO_x
SO_x
A number of environmental /physiological sensors
Data upload to the cloud
Data integration
PRISMS data storage
Electronic Health Records (EHR)

Figure 8.1: An illustration of main entities and data flow in BREATHE.

(section 2.4.1 provides background on these variables). Therefore, BREATHE is motivated to gather as much relevant information as possible to achieve high context-awareness. For this aim, three main data sources are employed: real-time sensing, self-assessment surveys, and external online data.

8.1.1 Real-time Sensor Data Collection

BREATHE utilizes an Android-based smartwatch and smartphone pair as the central communication points between the system and clients. This design choice facilitates continuous and integrated sensing, allows flexible interaction with subjects, and takes the usability considerations for children into account.

- Smartwatch: The BREATHE's smartwatch acts as the primary connection hub among various environmental and physiological sensors and merges data from registered devices advertising via Bluetooth Low Energy (BLE) in real-time. Using the smartwatch for sensor connectivity is beneficial for a continuous data collection, buffering and transmission in a multi-sensor monitoring study—particularly in children—as it can be readily worn throughout the day and across different activities (e.g., school, exercise). Moreover, smartwatches offer integrated health monitoring signals including heart rate and motion tracking that can be seamlessly merged in. All data is then directed to the smartphone.

- Smartphone: As the power and processing resources on a smartwatch are limited, we use a smartphone to perform more computationally intensive tasks including data encryption and transmission to the cloud. Moreover, the larger screen of the smartphone allows more involved user interaction for collection of EMA surveys. Importantly, children do not have to always carry the smartphone given the BLE connection to the smartwatch that remains active when in the range of 20-100 meters (depending on the device and power [59, 144]). Upon receiving of real-time data from the smartwatch, it is normalized into a standard format, temporally and geographically tagged, appended with device status metadata (e.g., battery level), encrypted and reliably sent to the cloud.

The information regarding all sensor data collected in BREATHE along with the asthma variable they monitor is listed in Table 8.1.

8.1.2 Self-Assessment Surveys

A number of severe asthma exacerbation events can be recognized through rescue inhaler usage monitoring. However, developing an asthma model for recognizing events causing mild to moderate symptoms requires the inclusion of the subjects' self-assessment and validation as well. Ecological momentary assessments (EMA) permits the collection of asthma and environmental self-report from subjects, randomly or close in time to an event of interest (e.g., sudden increase of air pollution). In general, there are three types of EMA triggers used in our system: 1) context-sensitive EMAs that are triggered to assess the context around an event detected through real-time monitoring of data on the cloud, 2) scheduled EMAs that are collected based on a pre-set morning and night schedule, 3) random EMAs that

Table 8.1: List of integrated environmental and physiological sensors integrated.

Monitoring Variable	Device	Provider	Communication to BREATHE
Triaxial accelerometer	Moto360 [1]	Motorola Mobility	BLE
Triaxial gyroscope			
Heart rate			
PM 2.5 (air quality)	Airbeam v1.0 [3]	AirCasting	
Temperature			
Relative Humidity			
Black carbon (air quality)	MA200 [2]	AethLabs/PRISMS Columbia	
Ozone (air quality)	ASU sensor	PRISMS/Arizona State University	
VOCs (air quality)			
Temperature			
Humidity			
Forced expiratory volume (FEV1)	Asthm-1 BT Spirometer [4]	Vitalograph	
Peak expiratory flow rate (PEF)			
Medication usage	Inhaler sensor [226]	Propeller Health	BREATHE API

Table 8.2: Self-assessment surveys in BREATHE

EMA type	Trigger type	Trigger Event	Major Question Categories
Air pollution	Context-sensitive	Sudden increase in PM 2.5	Event validation, asthma symptoms, general context.
Rescue m'edication	Context-sensitive	Use of rescue inhaler	Event validation, asthma symptoms, general context.
Heart rate	Context-sensitive	Stay 2 minutes in moderate to vigorous heart activity zone	Event validation, asthma symptoms, general context.
Spirometry	Context-sensitive	Use of spirometer	Event validation, asthma symptoms, general context.
Morning survey	Scheduled	First EMA of the day	Asthma over night, asthma symptoms, general context.
Random survey	Random	Random	Asthma symptoms, general context.
End of day survey	Scheduled	Last EMA of the day	Allergy and sickness, asthma symptoms, general context.

are collected based on a pre-set or random schedule aimed at capturing the asthma state of subjects in their natural settings. Each EMA questionnaire contains several categories of questions relevant to the trigger type and event. These categories include: 1) asthma symptom questions that evaluate 4 major symptoms of wheezing, coughing, chest tightness, and troubled breathing, 2) event validation questions that are aimed at checking the accuracy of the detected event by the real-time analytical model, 3) general context questions that are designed to obtain a self-report of child's location, activity, and exposures, 4) the end or start of day context questions that are designed to obtain key information such as presence of a sickness, allergy, or night-time awakenings. Table 8.2 presents types of EMAs available in BREATHE along with their trigger type, trigger event, and categories of questions included in them.

8.1.3 External Clinical and Environmental Data

To expand the contextual information, BREATHE incorporates electronic health records of recruited subjects and real-time environmental information based on geographic location (e.g., traffic and regional air quality) from resources provided by Sonoma Technology Inc.

8.1.4 BREATHE Server

BREATHE cloud-server is the main component handling real-time data processing and storage. The server provides a RESTful API accepting various real-time requests including data push and EMA trigger generation. Through this API, data that can be received from any source that has registered secure access to the server, particularly BREATHE smartphone app and sensors not supporting BLE data transfer and requiring to transfer data directly to the server (see Table 8.1).

Once data is received and decrypted, it is parsed, reformatted into a standardized JSON, and relayed to PRISMS Data Software and Coordinating Integration Center (DSCIC) for long-term archive and offline data modeling. One day of data per subject is kept on the BREATHE server for real-time analysis purposes (e.g., event-based EMA trigger generation, subject-specific risk assessment). The separation of modeling and real-time data processing allows BREATHE server to retain high performance in scaled usage.

8.2 Data Security

In order to maintain a closed Health Insurance Portability and Accountability Act(HIPAA)-compliant system design, multiple levels of data security is employed. On data transfer between sensors to smartwatch and smartphone, we rely on the BLE stack which is compliant with HIPPA requirements of encryption [210]. To transfer the data from the smartphone to the server, we use a 128-bit hybrid Rivest Shamir Adleman (RSA)-Advanced Encryption Standard (AES) technique [199], which is illustrated in Figure 8.2. In this approach, AES generates a session key for encryption and decryption of data. However, as AES key is symmetric, meaning that the same key must be used to encrypt and decrypt the data, it should be transferred from the smartphone to the cloud in each session. To ensure secure key exchange, we employ the RSA algorithm to encrypt the AES symmetric key from the smartphone to the cloud. The RSA algorithm generates a pair of public and private keys on the cloud and the smartphone uses the public one to encrypt the AES symmetric key when sending it to the cloud. [199] showed that this hybrid encryption technique increases the privacy and confidentiality of data while showing a lower encryption time and cost compared to other techniques.

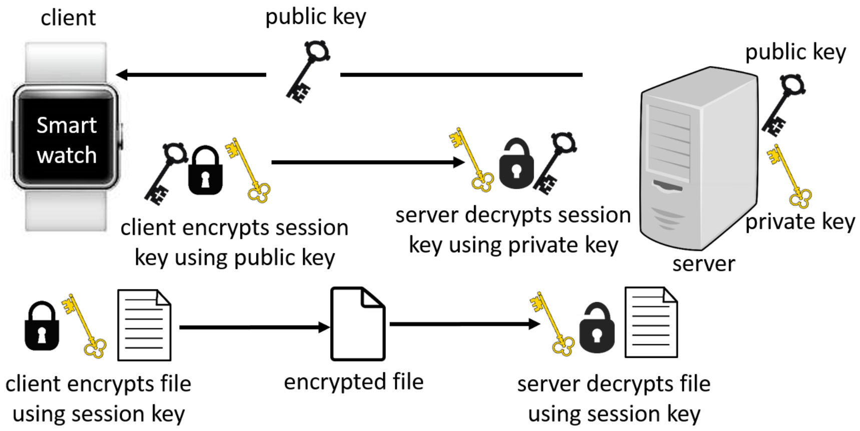

client
public key
public key
Smart watch
client encrypts session key using public key
server decrypts session key using private key
private key
server
client encrypts file using session key
encrypted file
server decrypts file using session key

Figure 8.2: Data security layers in BREATHE data transfer from a device to the server.

Table 8.3: Background information on recruited participants.

Age			Gender		Asthma level		
8-10	11-13	14-16	Male	Female	Mild	Moderate	Severe
3	11	6	11	9	5	9	6

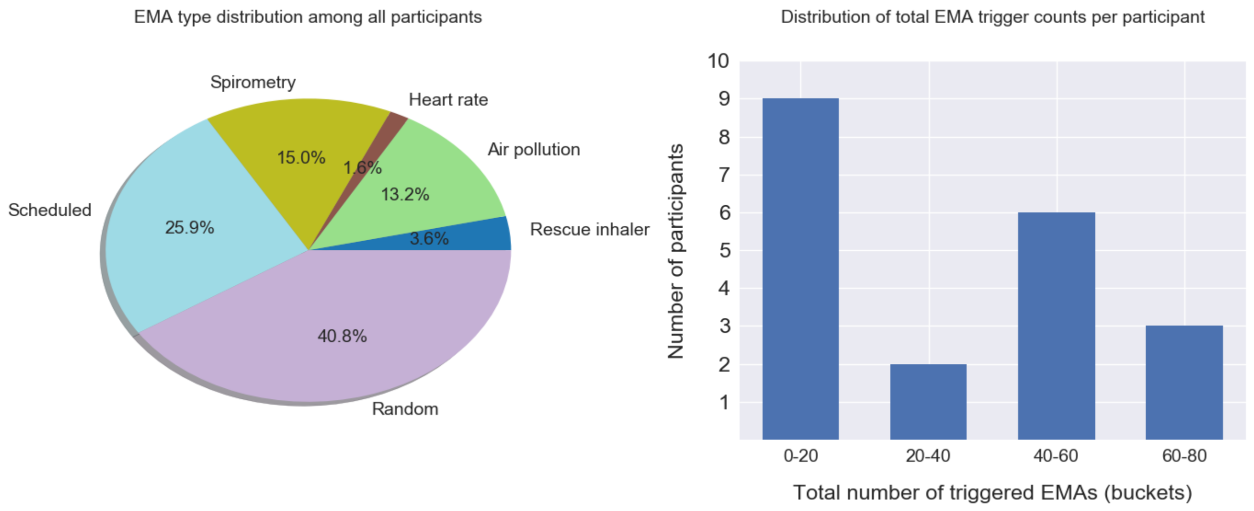


Figure 8.3: Distribution the type of triggered EMA questionnairs (left) and their total frequency among all participants (right).

8.3 Pilot Study

In this section, we provide results of our efforts in modeling asthma exacerbation using data collected from 20 participants recruited as part of the BREATHE pilot study. All individuals are consented, trained, and their asthma scores and sensor baselines are collected before entering the study. Following the recruitment session, they take the BREATHE kit to use during the child waking hours for a 2-week monitoring period. Table 8.3 provides demographic information on the recruited children. From 20 participants under the study, 19 completed and 1 left the study after 5 days. On average, each participant used the kit for 11 hours per day generating approximately 1.3 GBs of sensor data. It is important to mention that many participants reported trouble in using the kit during the school hours.

Participants were also asked to answer to EMA questionnaires within 10 minutes of their trigger event. Figure 8.3 demonstrates the distribution of triggered EMAs as well as their types among all participants. The observable difference in the number of total triggered EMAs can be attributed to the dependency of a subset of EMAs to the context events and its variation among the participants. Participants compliance with questionnaire completion has an average of 40%, ranging from a minimum of 11% to maximum of 85%.

8.3.1 Annotating Asthma Exacerbation Events

To form a model of asthma exacerbation for participants, we consider two sources of data: momentarily asthma symptom assessments and rescue inhaler usage. As discussed before, all EMA questionnaires contain a query on 4 asthma symptoms: chest tightness, wheezing, coughing, and having trouble breathing that are scored between 0 and 3 by participants. Asthma score which is used as a label in our study is defined as the summation of these 4 symptom scores. Furthermore, a rescue inhaler usage which is detected through Propeller sensors is considered as a label for a moderate to severe exacerbation of asthma. Figure 8.4 illustrates temporal changes in asthma scores and distribution of rescue inhaler usage among all participants. As both figures suggest, participants show a wide range of asthma exacerbation frequency and severity. However, the dataset distribution is skewed towards lower rates of rescue inhaler usage and minor asthma score changes.

In this study, we investigate the problem of predicting asthma exacerbation within the next two-hour time window. A positive label is assigned to any usage of rescue inhaler or a 2 point increase of asthma assessment score compared to the individual's baseline. These parameters are selected based on the observed distribution of collected data and consultation with domain experts in PRISMS. Table 8.4 summarizes the label distribution among subjects as well as the distribution of positive and negative labels in our dataset.

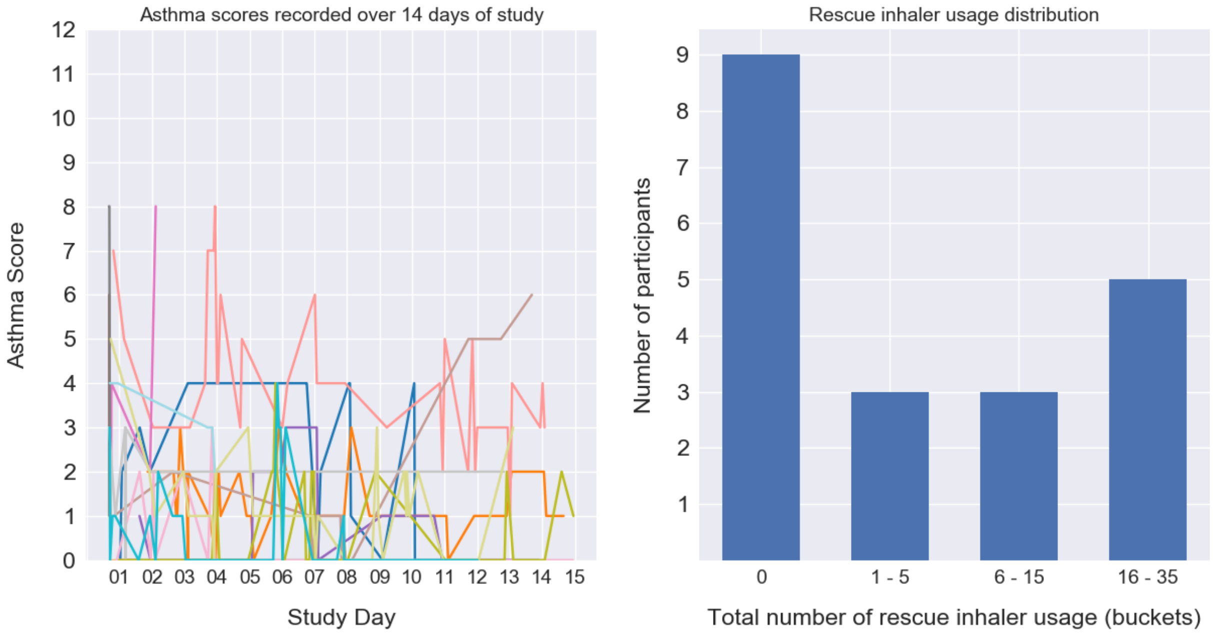


Figure 8.4: Temporal changes in asthma scores for all participants (left) and rescue inhaler usage distribution (right)

Table 8.4: The distribution of the number of collected data samples per participant and the average number of positive and negative labels in each bucket.

Number of samples per subject	0 – 10		10 – 20		20 – 30		30 – 40		40 – 50		> 50	
Number of subjects	6		2		1		5		3		3	
Sample label type	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Avrg number of label per subject	0.5	3.5	6	6	6	18	15	20	25.5	20.5	19.3	37.3

8.3.2 Asthma Model

Recognizing the sparsity of labels in this application, we employ a hierarchical prediction model. In the first step, we rely on supervised and unsupervised anomaly detection models to extract the environmental and physiological events in proximity to an annotated asthma assessment point. Next, an asthma model is developed for each individual to learn the relation of these events to possible exacerbation of symptoms.

Feature Event Extraction

A summarized view of how features/events are extracted from different components of BREATHE kit is illustrated in Figure 8.5. To detect changes in air quality and weather signals, we employ an online and unsupervised anomaly detection model, similar to the one discussed in section 3.2.1, which trains an autoencoder on intervals of data showing normal signal behavior for each individual and uses the reconstruction error for online detection of anomalies and possible asthma triggers. Furthermore, an increase in the intensity of activity, which is recognized as a possible trigger of asthma symptoms, is detected using the activity recognition model introduced in chapter 6. In order to detect heart rate events, we rely on [150] and examine the increase of heart rate reserve which is calculated as $HR_{max} - HR_{rest}$ with $HR_{max} = 208 - 0.7 \times age$. More than 50% increase of heart rate reserve for 2 minutes is considered as a heart event and possible asthma trigger. Experiencing an intense feeling such as anger or stress is also included as a feature through EMA questionnaires. Existence of these events is studied in a two-hour time-span before an annotated asthma assessment.

To include more long-term variables we also consider the presence of sickness and allergy in the past day, which is collected from the end of day surveys of participants. Furthermore, adherence to control medication is examined against each participant required dosage in the past two days and is integrated as a feature into the model. Poor lung function (detected by spirometer usage) and a close recorded exacerbation (through EMA or rescue inhaler usage) are two other features integrated into the model. It is worth mentioning that a few data

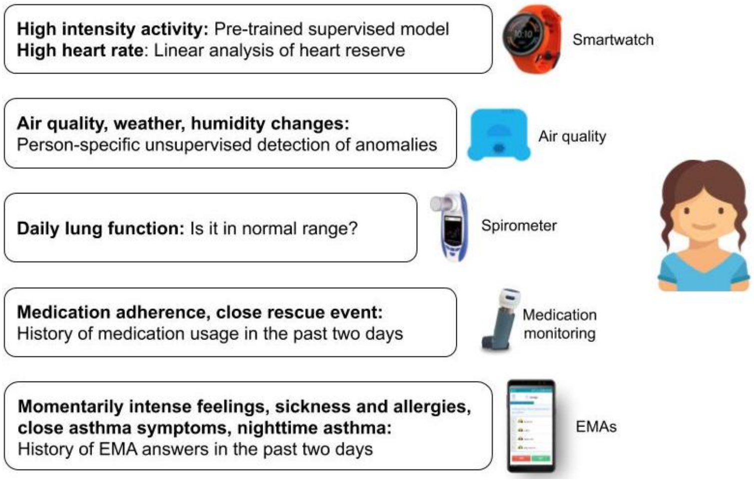

High intensity activity: Pre-trained supervised model
High heart rate: Linear analysis of heart reserve
Smartwatch
Air quality, weather, humidity changes:
Person-specific unsupervised detection of anomalies
Air quality
Daily lung function: Is it in normal range?
Spirometer
Medication adherence, close rescue event:
History of medication usage in the past two days
Medication monitoring
Momentarily intense feelings, sickness and allergies, close asthma symptoms, nighttime asthma:
History of EMA answers in the past two days
EMAs

Figure 8.5: Temporal changes of asthma scores for all participants (left) and rescue inhaler usage distribution (right)

sources introduced before have been excluded in our feature extraction process. In particular, ASU and MA200 sensors posed some deployment challenges such as loss of calibration on the time of deployment that left their records unreliable for some of the participants.

Figure 8.6 demonstrates the availability of each of the above-mentioned events prior to our labeled data samples. The numbers are reported separately for the two scenarios of rescue inhaler usage and self-assessment of asthma symptoms through EMAs. As it can be inferred from the left figure, the majority of detected events occur with higher frequency before positively labeled samples (presence of exacerbation) compared to negatively labeled ones. This difference in frequency is much lower for two event types, namely the increase of heart rate and experiencing an intense feeling, which can indicate a lower correlation either due to lack of accuracy or subjectivity of the event. On the other hand, we can observe a considerably lower frequency rate of activity and air quality events before usage of the rescue inhaler. A closer analysis shows no usage of BREATHE smartwatch before 35% of rescue inhaler usage events while this number is 16% for EMA questionnaires. To surface the roots of the observed usage behavior multiple factors such as battery issues and usability of the system when participants do not feel well should be further studied.

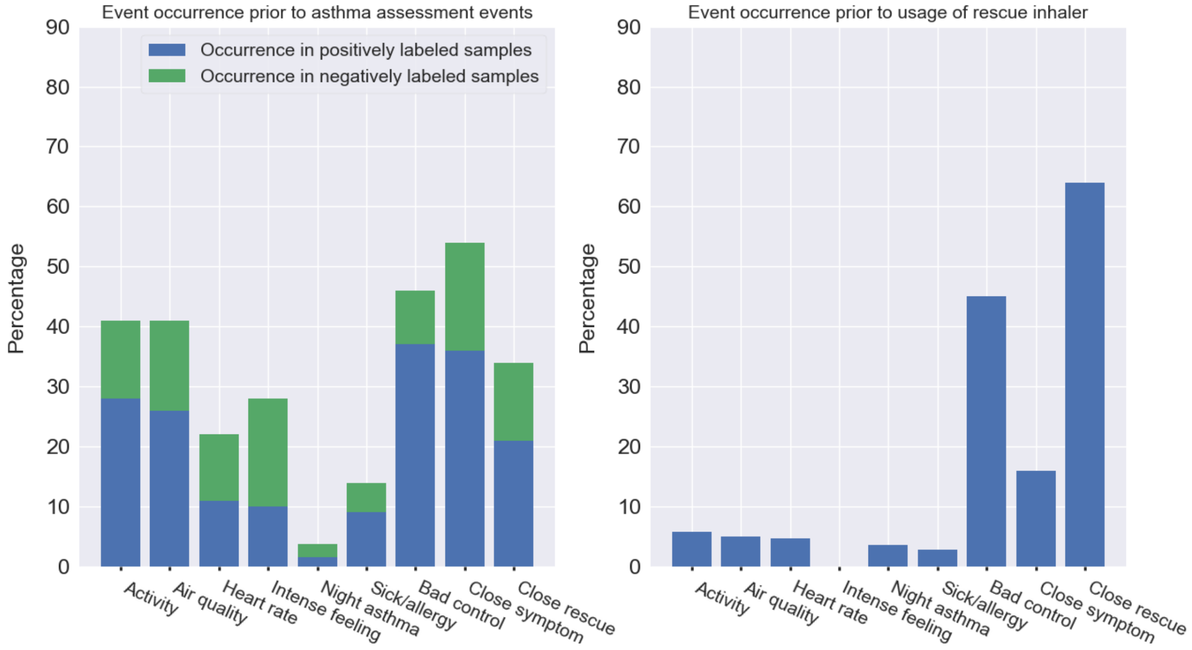


Figure 8.6: Frequency of various events prior to asthma assessments using EMA questionnairs (left) and rescue inhaler usage (right). Asthma assessment events are separated based on weather an exacerbation is observed (positive) or not (negative).

Prediction Results

Recognizing the limited availability of labeled samples and personal nature of asthma triggers, we rely on shallow models to build personalized predictors for each subject. For 11 out of 20 participants we were able to generate a large enough dataset (more than 20 samples) with a fairly balanced set of negative and positive labels. Figure 8.7 demonstrates the prediction results for these participants using a Random Forests classifier. The performance is reported based on accuracy, precision, recall, and average precision metrics. The average precision metric is especially important as it better shows how well the model scores the probability of an exacerbation onset. As the figure suggests, a mean average precision of 78.6% is achieved.

The importance of each feature in the prediction of the exacerbation is essential in the selection of optimal sensors and monitoring variables for asthma prediction. Figure 8.8

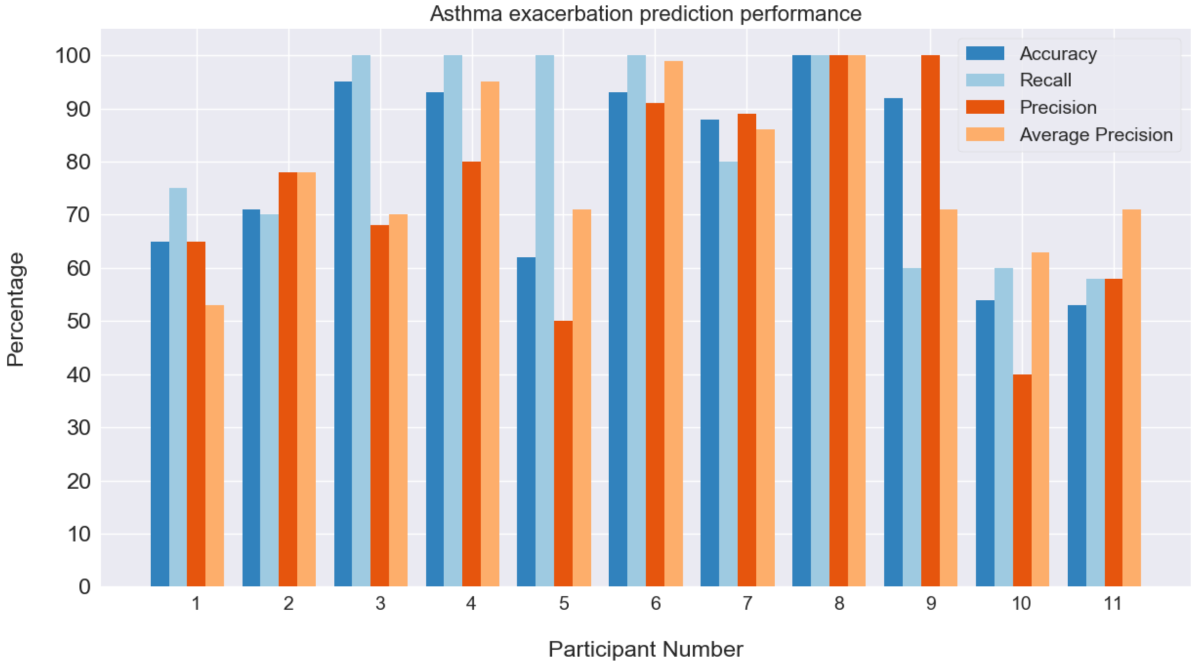


Figure 8.7: The prediction performance of personalized models trained on 10 participants.

reports the average importance score assigned by the predictor to each feature among all participants. The results interestingly show that a poor asthma control (lack of adherence to medications) and a close exacerbation are the main predictors of a future negative asthma event. High-intensity activity and sudden change in air quality follow these features in importance and show a significant importance margin against other features.

Results of this experiment show the considerable potential in personalized prediction of asthma exacerbation when enough positive labels (exacerbation events) are observed. However, for participants lacking enough labeled samples a general prediction model can be employed. To this end, three general models for mild, moderate and severe groups are trained. As the datasets collected from the remaining 9 participants are small and very skewed towards negative samples (with an average of 5 samples and less than 2 positive samples per participant) the prediction model for mild control group is employed which achieves an average of 92% accuracy, 57% recall, 71% precision and 61% average precision among these participants.

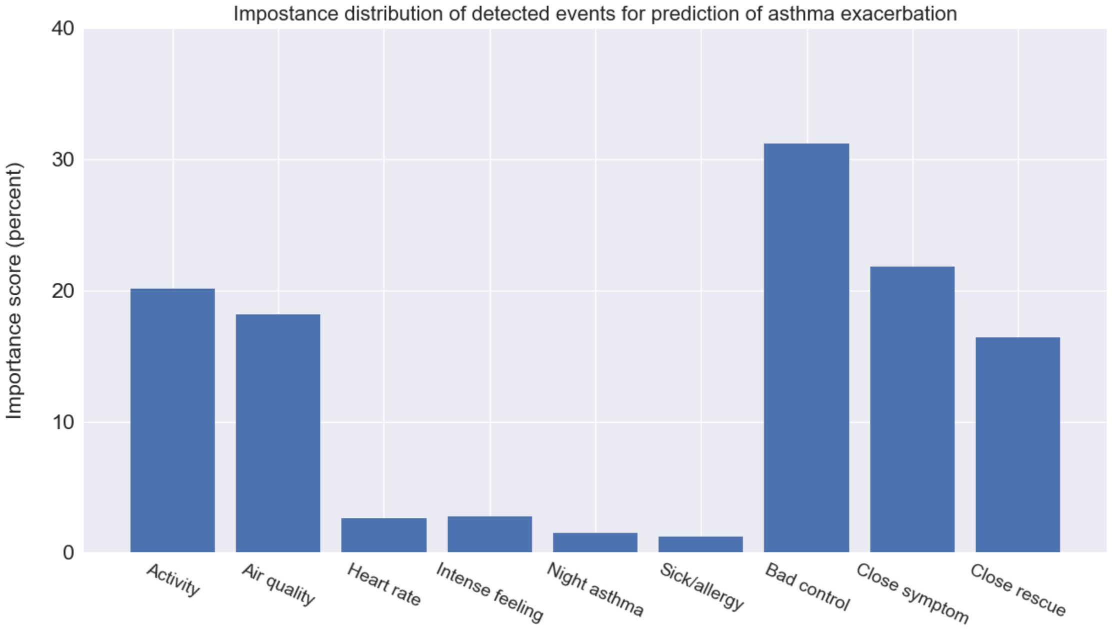


Figure 8.8: Average feature importance in prediction of asthma exacerbation.

8.4 Acknowledgment

BREATHE is a collaborative study between University of California Los Angeles (UCLA) and University of Southern California (USC) led by Alex A.T. Bui, director of UCLA Medical Imaging Informatics, as the principal investigator. BREATHE undertaking is divided between three closely interacting projects. Project 1 is our main effort in which we have been responsible for data analysis and building an integrated sensing solution and suitable infrastructure for secure data transfer from devices to analytical and data storage servers. Project 2 led by Rose Rocchio, director of the Mobile and Web Strategy Department of UCLA, is focused on visualizing the collected clinical and environmental data, as well as the design of applications and infrastructure for collection of EMA questionnaires. Project 3, led by Rima Habre from Keck school of medicine of USC, is aimed at developing testing protocols to assess the field performance of sensors and platform usability and provide feedback to projects 1 and 2 for refinements. Moreover, participant recruitment was conducted as a part of this project efforts. BREATHE has been also in close collaboration with other centers in PRISMS initiative especially PRISMS Data Software and Coordinating Integration Center

(DSCIC) led by Jose-Luis Ambite from USC Viterbi school of engineering and PRISMS center in Arizona State University and Aethlabs.

8.5 Conclusion

In this chapter, we introduced BREATHE, a pervasive health solution aimed at the prediction of asthma exacerbation ahead of time. Through this chapter, we visited the system design of BREATHE for secure and flexible data collection, transmission and analysis. We also investigated the modeling of asthma exacerbation on 20 participants. Results of our study show the potential of such a system in the prediction of future asthma events.

The future steps that are envisioned for this study include analytical and systematical improvements. On the modeling side, it is desirable to validate the obtained results on a larger set of participants with a diverse range of asthma symptoms. With the availability of larger datasets, further studies can be conducted on the design of reliable general models for participants with sparse labeled samples. On the system design side, we believe data quality can be improved by the adoption of more accurate heart measurement devices and increasing the frequency of motion signals collection. These improvements should be approached with the important challenge of battery consumption in mind. On the other hand, battery life improvements on smart devices can be of great value to such applications and removing the burden of recharging the kits from the user.

CHAPTER 9

Conclusion

In this final chapter, we summarize the objectives and contributions of this research and discuss the directions we envision for the future. This dissertation made several contributions to address unique data modeling challenges in the domain of pervasive health that were surfaced through our efforts in building a highly context-aware solution for the prevention of asthma exacerbation in children. These challenges and our approach in addressing them are summarized below:

- Sparsity of labels: Pervasive health solutions generate an enormous size of data, however, the majority of the collected samples remain unlabeled in a real-world setting. Therefore, diversifying pervasive health applications requires the adoption of unsupervised or semi-supervised approaches. To this end, we proposed a new unsupervised method for the detection of future negative events through physiological sensors in chapter 3. On the same direction, we also showed how a model can be transferred to a new domain with no or sparse set of labeled samples in chapter 7.

- Heterogeneity of data types: Developing a comprehensive model of health requires the integration of heterogeneous types of data in various formats. This is especially challenging due to the huge size of data in healthcare. In chapters 4 and 5, we proposed a unified integration using Heterogeneous Information Networks and showed that it can achieve the state of the art results in diagnosis prediction task.

- Poor model generalizability: One challenging aspect of healthcare is its high variability among individuals and diseases. Models trained for a specific group in the healthcare domain are hardly generalizable to individuals beyond that group. In chapters 6 and 7, we investigated multi-task learning to improve personalization and deep domain adaptation to address the important challenge of deploying a model in a new domain.

- Interpretability of results: Recognizing the sensitive nature of healthcare and the need for transparency of the decision process, interpretability is considered as an essential requirement in the design of models in this domain. To this end, we proposed an interpretable integration of healthcare data in chapter 5.

9.1 Future Directions

In this section, we discuss the lines of research that we find promising for further advancement of pervasive health and healthcare informatics in general. We envision these advancements in both methodological and application sides.

9.1.1 Methodological side

One important challenge in the current state of healthcare informatics is building longitudinal models of individuals' health state. This limitation is largely due to the lack of access to datasets containing long-term health records of subjects. In the future, we expect availability of such datasets that will in turn open new lines of research to address major unmet challenges, including adapting to an individual's health shift over time, balancing the impact of long-term and short-term events on prediction of an individual's health state, and learning the sequential relation of health events on different outcomes. On this direction, we also hope to see datasets and analytical models connecting electronic health records of individuals to the data collected from pervasive health solutions. Such integrative models will provide more detailed investigation of health trajectories and prevention opportunities.

As discussed in the objectives of our research, in order to expand the reach and scale of pervasive health solutions, we need to rely on learning transferable models for healthcare. In chapters 6 and 7, we approached this problem by showing how a time-series based model can be transferred among individuals and domains. However, there is still room for investigation of how models combining various types of data for an individual, such as the one we proposed in chapter 4, can be transferred. Furthermore, the transfer of models among various EHR datasets with a different distribution of records (e.g., symptoms and diseases) is a direction we expect to see in the future. This will allow tackling the existing high discrepancy among EHR datasets and eliminate the need for the development of dataset-specific models.

We similarly expect more work on personalization of healthcare models. Pervasive health solutions can currently collect a huge set of personalized data but there is still a great need for person-specific data modeling. In chapter 6, we showed the promise of this direction on improving the activity recognition task, which involved only a single source of data. In the future, we expect to see personalized models that integrate heterogeneous sources of data. Recalling the sparsity of labels in pervasive health, we also expect to see the investigation of active learning algorithms that query the individual only for highly informative labels that can improve the personalized prediction.

Last but not least, we expect to see a higher focus on unsupervised or semi-supervised models for real-time predictions in pervasive health. Following our work in chapter 3, we envision the investigation of how the model can be extended to incorporate various sources and types of information.

9.1.2 Application side

There have been many lessons learned through our efforts in building a highly context-aware pervasive health solution. We found battery life of smart devices one of the main existing barriers in the successful deployment of such solutions. Ideally, the battery should last for at least one day of constant use, however, our experience in the design of BREATHE (chapter 8) suggests much lower battery life expectancy, especially when the smart device is used as the main connection hub of other sensors. While we expect to see advancements in this direction, we envision a line of research on determining the minimum acceptable frequency of data collection for different types of sensors. This will ensure acceptable performance for models built on top of the collected data while minimizing energy consumption.

Furthermore, our experience in the integration of a wide range of sensors into BREATHE surfaces the strong need for the development of standardized APIs for flexible connection of sensors into pervasive health solutions. Although we established a normalized data view of available sensors, there is still a need to develop a specification of possible data elements and formats collected by environmental and physiological sensors so that previously unknown sensors can also be integrated into the system on-the-fly.

Finally, we expect to see advancements in the accuracy of built-in sensors in smart devices, especially smartwatches. Our experiments in BREATHE show high variation and low accuracy of heart rate signal collected by the smartwatch used in our study.

Bibliography

[1] 2019 motorola mobility llc. https://www.motorola.com.au/products/moto-360. Accessed: 2019-06-30.

[2] Aethlabs (2019) microaeth family. https://aethlabs.com/microaeth. Accessed: 2019-06-30.

[3] Airbeam (2019), wearable air monitor. http://www.takingspace.org/aircasting/airbeam/. Accessed: 2019-06-30.

[4] Vitalograph (2019) asma-1 bt. https://vitalograph.com/product/162431/asma-1-bt. Accessed: 2019-06-30.

[5] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[6] Saeed Abdullah and Tanzeem Choudhury. Sensing technologies for monitoring serious mental illnesses. *IEEE MultiMedia*, 25(1):61–75, 2018.

[7] U Rajendra Acharya, Hamido Fujita, Oh Shu Lih, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network. *Information sciences*, 405:81–90, 2017.

[8] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. *Computers in biology and medicine*, 100:270–278, 2018.

[9] Paul S Addison. Wavelet transforms and the ecg: a review. *Physiological measurement*, 26(5):R155, 2005.

[10] Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

[11] Fadi Al Machot, Heinrich C Mayr, and Suneth Ranasinghe. A windowing approach for activity recognition in sensor data streams. In *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 951–953. IEEE, 2016.

[12] Nabil Alshurafa, Haik Kalantarian, Mohammad Pourhomayoun, Jason J Liu, Shruti Sarin, Behnam Shahbazi, and Majid Sarrafzadeh. Recognition of nutrition intake using time-frequency decomposition in a wearable necklace using a piezoelectric sensor. *IEEE Sensors Journal*, 15(7):3909–3916, 2015.

[13] Pramod Anantharam, Tanvi Banerjee, Amit Sheth, Krishnaprasad Thirunarayan, Surendra Marupudi, Vaikunth Sridharan, and Shalini G Forbis. Knowledge-driven personalized contextual mhealth service for asthma management in children. In *2015 IEEE International Conference on Mobile Services*, pages 284–291. IEEE, 2015.

[14] Lars Bo Andersen, Maarike Harro, Luis B Sardinha, Karsten Froberg, Ulf Ekelund, Søren Brage, and Sigmund Alfred Anderssen. Physical activity and clustered cardiovascular risk in children: a cross-sectional study (the european youth heart study). *The Lancet*, 368(9532):299–304, 2006.

[15] Rasmus S Andersen, Abdolrahman Peimankar, and Sadasivan Puthusserypady. A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications*, 115:465–473, 2019.

[16] Javier Andreu-Perez, Daniel R Leff, Henry MD Ip, and Guang-Zhong Yang. From wearable sensors to smart implants—toward pervasive and personalized healthcare. *IEEE Transactions on Biomedical Engineering*, 62(12):2750–2762, 2015.

[17] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Esann*, 2013.

[18] American Medical Association. *International classification of diseases, 9th revision, clinical modification: physician ICD-9-CM, 2005: volumes 1 and 2, color-coded, illustrated*, volume 1. Amer Medical Assn, 2004.

[19] Manfredo Atzori, Matteo Cognolato, and Henning Müller. Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. *Frontiers in neurorobotics*, 10:9, 2016.

[20] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *23th International conference on architecture of computing systems 2010*, pages 1–10. VDE, 2010.

[21] Caglar Aytekin, Xingyang Ni, Francesco Cricri, and Emre Aksu. Clustering and unsupervised anomaly detection with l 2 normalized deep auto-encoder representations. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[22] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[23] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

[24] Hadi Banaee, Mobyen Ahmed, and Amy Loutfi. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors*, 13(12):17472–17500, 2013.

[25] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, 14(4):6474–6499, 2014.

[26] Akram Bayat, Marc Pomplun, and Duc A Tran. A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34:450–457, 2014.

[27] Richard Beasley, Alex Semprini, and Edwin A Mitchell. Risk factors for asthma: is prevention possible? *The Lancet*, 386(9998):1075–1085, 2015.

[28] Brett K Beaulieu-Jones, Casey S Greene, et al. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of biomedical informatics*, 64:168–178, 2016.

[29] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*, 2017.

[30] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

[31] A Brink-Kjaer, Alexander Neergaard Olesen, CA Jespersen, PE Peppard, PJ Jennum, HB Sørensen, and E Mignot. 0142 automatic detection of cortical arousals in sleep using bi-direction lstm networks. *Sleep*, 41(suppl_1):A55–A56, 2018.

[32] Alex AT Bui, Anahita Hosseini, Rose Rocchio, Mindy Jacobs, Nate Ross, Sande Okelo, Fred Lurmann, Sandrah Eckel, Genevieve Dunton, Gilliland, Majid Sarrafzadeh, and Rima Habre. Biomedical real-time health evaluation (breathe): Towards an mhealth informatics platform. *In prepration for publication in the Journal of the American Medical Informatics Association (JAMIA)*, 2019.

[33] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):33, 2014.

[34] Feng Cao, Martin Estert, Weining Qian, and Aoying Zhou. Density-based clustering over an evolving data stream with noise. In *Proceedings of the 2006 SIAM international conference on data mining*, pages 328–339. SIAM, 2006.

[35] Yu Cao, Peng Hou, Donald Brown, Jie Wang, and Songqing Chen. Distributed analytics and edge intelligence: Pervasive health monitoring at the era of fog computing. In *Proceedings of the 2015 Workshop on Mobile Big Data*, pages 43–48. ACM, 2015.

[36] Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi. Online anomaly detection for long-term ecg monitoring using wearable devices. *Pattern Recognition*, 88:482–492, 2019.

[37] Debora S Chan, Charles W Callahan, Virginia B Hatch-Pigott, Annette Lawless, H Lorraine Proffitt, Nola E Manning, Mary Schweikert, and Francis J Malone. Internet-based home monitoring and education of children with asthma is comparable to ideal office-based care: results of a 1-year asthma in-home monitoring trial. *Pediatrics*, 119(3):569, 2007.

[38] Johnny Yat Ming Chan, Stephen Anthony Tunnell, and Michael Joseph Thomas. Systems and methods for managing and dynamically predicting respiratory episodes, August 14 2014. US Patent App. 14/180,743.

[39] Yu-Feng Yvonne Chan, Brian M Bot, Micol Zweig, Nicole Tignor, Weiping Ma, Christine Suver, Rafhael Cedeno, Erick R Scott, Steven Gregory Hershman, Eric E Schadt, et al. The asthma mobile health study, smartphone data collected using researchkit. *Scientific data*, 5:180096, 2018.

[40] Yu-Feng Yvonne Chan, Pei Wang, Linda Rogers, Nicole Tignor, Micol Zweig, Steven G Hershman, Nicholas Genes, Erick R Scott, Eric Krock, Marcus Badgeley, et al. The asthma mobile health study, a large-scale clinical observational study using researchkit. *Nature biotechnology*, 35(4):354, 2017.

[41] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[42] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C Aggarwal, and Thomas S Huang. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 119–128. ACM, 2015.

[43] AJ Chauhan, Hazel M Inskip, Catherine H Linaker, Sandra Smith, Jacqueline Schreiber, Sebastian L Johnston, and Stephen T Holgate. Personal exposure to nitrogen dioxide (no2) and the severity of virus-induced asthma in children. *The Lancet*, 361(9373):1939–1944, 2003.

[44] Sucheta Chauhan and Lovekesh Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–7. IEEE, 2015.

[45] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[46] Zhengping Che and Yan Liu. Deep learning solutions to computational phenotyping in health care. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, pages 1100–1109. IEEE, 2017.

[47] Ting Chen and Yizhou Sun. Task-guided and path-augmented heterogeneous network embedding for author identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 295–304. ACM, 2017.

[48] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.

[49] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.

[50] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[51] Eun Jeong Choi and Dong Keun Kim. Arousal and valence classification model based on long short-term memory and deap data for mental healthcare management. *Healthcare informatics research*, 24(4):309–316, 2018.

[52] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.

[53] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.

[54] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*, 2018.

[55] Stephanie Hui-Wen Chuah, Philipp A Rauschnabel, Nina Krey, Bang Nguyen, Thurasamy Ramayah, and Shwetak Lade. Wearable technologies: The role of usefulness and visibility in smartwatch adoption. *Computers in Human Behavior*, 65:276–284, 2016.

[56] Giovanna Cilluffo, Salvatore Fasola, Velia Malizia, Giuliana Ferrante, Laura Montalbano, Marco Montalbano, Giovanni Viegi, and Stefania La Grutta. Measuring lung function in asthmatic children: A spirometry and forced oscillation technique (fot) comparison, 2016.

[57] Ian Cleland, Basel Kikhia, Chris Nugent, Andrey Boytsov, Josef Hallberg, Kåre Synnes, Sally McClean, and Dewar Finlay. Optimal placement of accelerometers for the detection of everyday activities. *Sensors*, 13(7):9183–9200, 2013.

[58] Donald W Cockcroft and Veronica A Swystun. Asthma control versus asthma severity. *Journal of allergy and clinical immunology*, 98(6):1016–1018, 1996.

[59] Mario Collotta, Giovanni Pau, Timothy Talty, and Ozan K Tonguz. Bluetooth 5: A concrete step forward toward the iot. *IEEE Communications Magazine*, 56(7):125–131, 2018.

[60] Healthcare Cost, Utilization Project (HCUP), et al. Beta clinical classifications software (ccs) for icd-10-cm/pcs.

[61] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

[62] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822. ACM, 2014.

[63] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 24(4):500–504, 2017.

[64] Keith M Diaz, David J Krupka, Melinda J Chang, James Peacock, Yao Ma, Jeff Goldsmith, Joseph E Schwartz, and Karina W Davidson. Fitbit®: An accurate and reliable device for wireless physical activity tracking. *International journal of cardiology*, 185:138–140, 2015.

[65] Kenneth Donaldson, M Ian Gilmour, and William MacNee. Asthma and pm 10. *Respiratory Research*, 1(1):12, 2000.

[66] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144. ACM, 2017.

[67] Sumeet Dua, U Rajendra Acharya, and Prerna Dua. *Machine learning in healthcare informatics*, volume 56. Springer, 2014.

[68] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[69] Hammad Durrani. Healthcare and healthcare systems: inspiring progress and future prospects. *Mhealth*, 2, 2016.

[70] Donna S Eng and Joyce M Lee. The promise and peril of mobile health applications for diabetes and endocrinology. *Pediatric diabetes*, 14(4):231–238, 2013.

[71] M Engels, FG Garre, et al. Identification of children's activity type with accelerometer-based neural networks. *Medicine and science in sports and exercise*, 43(10):1994–1999, 2011.

[72] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[73] Alessandra Fagundes, L.G. Fagundes, and E.A.A. Lo Schiavo. Wavelet concepts for electrocardiographic signal analysis. *Clinical and Experimental Medical Letters*, 54:169–178, 12 2013.

[74] Wael Farhan, Zhimu Wang, Yingxiang Huang, Shuang Wang, Fei Wang, and Xiaoqian Jiang. A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR medical informatics*, 4(4), 2016.

[75] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

[76] Xue Feng, Yaodong Zhang, and James Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1759–1763. IEEE, 2014.

[77] Davide Figo, Pedro C Diniz, Diogo R Ferreira, and João M Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, 14(7):645–662, 2010.

[78] Joseph Finkelstein, Manuel R Cabrera, and George Hripcsak. Internet-based home asthma telemonitoring: can patients handle the technology? *Chest*, 117(1):148–155, 2000.

[79] Susannah Fox and Maeve Duggan. *Mobile health 2010*. Pew Internet & American Life Project Washington, DC, 2010.

[80] Florian Gahleitner, Julian Legg, Emma Holland, Sarah Pearson, and Graham Roberts. The validity and acceptability of a text-based monitoring system for pediatric asthma studies. *Pediatric pulmonology*, 51(1):5–12, 2016.

[81] Jiahui Gao, Qijin Cheng, and LH Philip. Detecting comments showing risk for suicide in youtube. In *Proceedings of the Future Technologies Conference*, pages 385–400. Springer, 2018.

[82] Alan H Gee, Riccardo Barbieri, David Paydarfar, and Premananda Indic. Predicting bradycardia in preterm infants using point process analysis of heart rate. *IEEE Transactions on Biomedical Engineering*, 64(9), 2017.

[83] Weidong Geng, Yu Du, Wenguang Jin, Wentao Wei, Yu Hu, and Jiajun Li. Gesture recognition by instantaneous surface emg images. *Scientific reports*, 6:36571, 2016.

[84] Janneane F Gent, Elizabeth W Triche, Theodore R Holford, Kathleen Belanger, Michael B Bracken, William S Beckett, and Brian P Leaderer. Association of low-level ozone and fine particles with respiratory symptoms in children with asthma. *Jama*, 290(14):1859–1867, 2003.

[85] Walter Gerych, Emmanuel Agu, and Elke Rundensteiner. Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 124–127. IEEE, 2019.

[86] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[87] Gordon Glyck. Posture trainer, December 3 2013. US Patent 8,596,717.

[88] AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, P Ch Ivanov, RG Mark, JE Mietus, GB Moody, CK Peng, and HE Stanley. Components of a new research resource for complex physiologic signals. *PhysioBank, PhysioToolkit, and Physionet.*

[89] Michael I Goran and Margarita S Treuth. Energy expenditure, physical activity, and obesity in children. *Pediatric Clinics of North America*, 48(4):931–953, 2001.

[90] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, Russ B Altman, and Roded Sharan. A method for inferring medical diagnoses from patient similarities. *BMC medicine*, 11(1):194, 2013.

[91] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

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

[93] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

[94] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[95] Ma Guanglong, Wang Xiangqing, and Yu Junsheng. Ecg signal classification algorithm based on fusion features. In *Journal of Physics: Conference Series*, volume 1207, page 012003. IOP Publishing, 2019.

[96] Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, and Muhammad Imran. Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management*, 2018.

[97] Markus Hagenbuchner, Dylan P Cliff, Stewart G Trost, Nguyen Van Tuc, and Gregory E Peoples. Prediction of activity type in preschool children using machine learning techniques. *journal of Science and Medicine in Sport*, 18(4):426–431, 2015.

[98] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[99] Hyewon Han, Kyunggeun Byun, and Hong-Goo Kang. A deep learning-based stress detection algorithm with speech signal. In *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, pages 11–15. ACM, 2018.

[100] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S Yu. Mining knowledge from databases: an information network analysis approach. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1251–1252. ACM, 2010.

[101] Shah Ahsanul Haque, Mustafizur Rahman, and Syed Mahfuzul Aziz. Sensor anomaly detection in wireless sensor networks for healthcare. *Sensors*, 15(4):8764–8786, 2015.

[102] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[104] Michael T Heideman, Don H Johnson, and C Sidney Burrus. Gauss and the history of the fast fourier transform. *Archive for history of exact sciences*, 34(3):265–277, 1985.

[105] Yuki Hikihara, Chiaki Tanaka, Yoshitake Oshima, Kazunori Ohkawara, Kazuko Ishikawa-Takata, and Shigeho Tanaka. Prediction models discriminating between non-locomotive and locomotive activities in children using a triaxial accelerometer with a gravity-removal physical activity classification algorithm. *PloS one*, 9(4):e94940, 2014.

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

[107] Hannaneh Hojaiji, Haik Kalantarian, Alex AT Bui, Christine E King, and Majid Sarrafzadeh. Temperature and humidity calibration of a low-cost wireless dust sensor for real-time monitoring. In *2017 IEEE sensors applications symposium (SAS)*, pages 1–6. IEEE, 2017.

[108] Bree Holtz and Pamela Whitten. Managing asthma with mobile phones: a feasibility study. *Telemedicine and e-Health*, 15(9):907–909, 2009.

[109] Gaylor Hoskins, Colin McCowan, RG Neville, Giles E Thomas, Barbara Smith, and Sue Silverman. Risk factors and costs associated with an asthma attack. *Thorax*, 55(1):19–24, 2000.

[110] Anahita Hosseini, Chris M Buonocore, Sepideh Hashemzadeh, Hannaneh Hojaiji, Haik Kalantarian, Costas Sideris, Alex AT Bui, Christine E King, and Majid Sarrafzadeh. Hipaa compliant wireless sensing smartwatch application for the self-management of pediatric asthma. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 49–54. IEEE, 2016.

[111] Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, and Majid Sarrafzadeh. Heteromed: Heterogeneous information network for medical diagnosis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 763–772. ACM, 2018.

[112] Anahita Hosseini, Tyler Davis, and Majid Sarrafzadeh. Hierarchical target-attentive diagnosis prediction in heterogeneous information networks. *Submitted to 18th IEEE International Conference on Data Mining (ICDM 2019)*, 2019.

[113] Anahita Hosseini, Shayan Fazeli, Eleanne van Vliet, Lisa Valencia, Rima Habre, Majid Sarrafzadeh, and Alex Bui. Children activity recognition: Challenges and strategies. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4331–4334. IEEE, 2018.

[114] Anahita Hosseini and Majid Sarrafzadeh. Unsupervised prediction of negative health events ahead of time. *Accepted to 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019.

[115] Anahita Hosseini, Davina Zamanzadeh, Lisa Valencia, Rima Habre, Alex AT Bui, and Majid Sarrafzadeh. Domain adaptation in children activity recognition. *Accepted to 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

[116] Shifu Hou, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1507–1515. ACM, 2017.

[117] Binbin Hu, Chuan Shi, Wayne Xin Zhao, and Philip S Yu. Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1531–1540. ACM, 2018.

[118] KP Jacobsen, Alexander Neergaard Olesen, L Trap, PE Peppard, HB Sorensen, PJ Jennum, and E Mignot. 0328 automatic detection of respiratory events during sleep using bidirectional lstm networks. *Sleep*, 41(suppl_1):A125–A126, 2018.

[119] Utkarshani Jaimini, Tanvi Banerjee, William Romine, Krishnaprasad Thirunarayan, Amit Sheth, and Maninder Kalra. Investigation of an indoor air quality sensor for asthma management in children. *IEEE sensors letters*, 1(2):1–4, 2017.

[120] Junho Jang, Kwun Chuen Gary Chan, Hsiang Huang, and Sean D Sullivan. Trends in cost and outcomes among adult and pediatric patients with asthma: 2000–2009. *Annals of Allergy, Asthma & Immunology*, 111(6):516–522, 2013.

[121] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*, 2016.

[122] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 202–208. IEEE, 2017.

[123] Jennifer Jardine, Jonathan Fisher, and Benjamin Carrick. Apples researchkit: smart data collection for the smartphone era?, 2015.

[124] Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*, 2018.

[125] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[126] Stephen M Kanne, Laura Arnstein Carpenter, and Zachary Warren. Screening in toddlers and preschoolers at risk for autism spectrum disorder: Evaluating a novel mobile-health screening tool. *Autism Research*, 11(7):1038–1049, 2018.

[127] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. In *Data mining in time series databases*, pages 1–21. World Scientific, 2004.

[128] Adil Mehmood Khan, Y-K Lee, Seok-Yong Lee, and T-S Kim. Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis. In *2010 5th international conference on future information technology*, pages 1–6. IEEE, 2010.

[129] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[130] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, 2016.

[131] Robert E Kleiger, Phyllis K Stein, and J Thomas Bigger Jr. Heart rate variability: measurement and clinical utility. *Annals of Noninvasive Electrocardiology*, 10(1):88–101, 2005.

[132] Ana Francisca Rozin Kleiner, Ilaria Pacifici, Alessandro Vagnini, Filippo Camerota, Claudia Celletti, Fabrizio Stocchi, Maria Francesca De Pandis, and Manuela Galli. Timed up and go evaluation with wearable devices: Validation in parkinson's disease. *Journal of bodywork and movement therapies*, 22(2):390–395, 2018.

[133] Santosh Krishna, Benjamin D Francisco, E Andrew Balas, Peter Konig, Gavin R Graff, Richard W Madsen, et al. Internet-enabled interactive multimedia asthma education program: a randomized trial. *Pediatrics-English Edition*, 111(3):503–510, 2003.

[134] Mars Lan, Lauren Samy, Nabil Alshurafa, Myung-Kyung Suh, Hassan Ghasemzadeh, Aurelia Macabasco-O'Connell, and Majid Sarrafzadeh. Wanda: An end-to-end remote health monitoring and analytics system for heart failure patients. In *Proceedings of the conference on Wireless Health*, page 9. ACM, 2012.

[135] Hyeon Kyu Lee and Young-Seok Choi. A convolution neural networks scheme for classification of motor imagery eeg based on wavelet time-frequecy image. In *2018 International Conference on Information Networking (ICOIN)*, pages 906–909. IEEE, 2018.

[136] Wei-Han Lee, Jorge Ortiz, Bongjun Ko, and Ruby Lee. Time series segmentation through automatic feature learning. *arXiv preprint arXiv:1801.05394*, 2018.

[137] Kenan Li, Rima Habre, Huiyu Deng, Robert Urman, John Morrison, Frank D Gilliland, José Luis Ambite, Dimitris Stripelis, Yao-Yi Chiang, Yijun Lin, et al. Applying multivariate segmentation methods to human activity recognition from wearable sensors data. *JMIR mHealth and uHealth*, 7(2):e11201, 2019.

[138] Zheng Li, Ying Wei, Yu Zhang, Xiang Zhang, Xin Li, and Qiang Yang. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. *arXiv preprint arXiv:1811.10999*, 2018.

[139] Rong-Ho Lin. An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, 47(1):53–62, 2009.

[140] Shao Lin, Xiu Liu, Linh H Le, and Syni-An Hwang. Chronic exposure to ambient ozone and asthma hospital admissions among children. *Environmental Health Perspectives*, 116(12):1725–1730, 2008.

[141] Jin Liu, Yi Pan, Min Li, Ziyue Chen, Lu Tang, Chengqian Lu, and Jianxin Wang. Applications of deep learning to mri images: A survey. *Big Data Mining and Analytics*, 1(1):1–18, 2018.

[142] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[143] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Transferable representation learning with deep adaptation networks. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[144] Angela Lonzetta, Peter Cope, Joseph Campbell, Bassam Mohd, and Thaier Hayajneh. Security vulnerabilities in bluetooth technology as used in iot. *Journal of Sensor and Actuator Networks*, 7(3):28, 2018.

[145] Ivo M Lopes, Bruno M Silva, Joel JPC Rodrigues, Jaime Lloret, and Mario L Proença. A mobile health monitoring solution for weight control. In *2011 International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–5. IEEE, 2011.

[146] Jake Luo, Christina Eldredge, Chi C Cho, and Ron A Cisler. Population analysis of adverse events in different age groups using big clinical trials data. *JMIR medical informatics*, 4(4), 2016.

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

[148] Yanhua Lv, Haijin Zhao, Zhenyu Liang, Hangming Dong, Laiyu Liu, Dandan Zhang, and Shaoxi Cai. A mobile phone short message service improves perceived control of asthma: a randomized controlled trial. *Telemedicine and e-Health*, 18(6):420–426, 2012.

[149] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1903–1911. ACM, 2017.

[150] Anthony D Mahon, Andrea D Marjerrison, Jonah D Lee, Megan E Woodruff, and Lauren E Hanna. Evaluating the prediction of maximal heart rate in children and adolescents. *Research quarterly for exercise and sport*, 81(4):466–471, 2010.

[151] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

[152] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings*, page 89. Presses universitaires de Louvain, 2015.

[153] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[154] Andrea Mannini, Mary Rosenberger, William L Haskell, Angelo M Sabatini, and Stephen S Intille. Activity recognition in youth using single accelerometer placed at wrist or ankle. *Medicine and science in sports and exercise*, 49(4):801, 2017.

[155] Gabriela Marcu, Anjali Misra, Karina Caro, Meghan Plank, Amy Leader, and Andrea Barsevick. Bounce: Designing a physical activity intervention for breast cancer survivors. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 25–34. ACM, 2018.

[156] Ruth E Mayagoitia, Anand V Nene, and Peter H Veltink. Accelerometer and rate gyroscope measurement of kinematics: an inexpensive alternative to optical motion analysis systems. *Journal of biomechanics*, 35(4):537–542, 2002.

[157] Susannah McLean, David Chandler, Ulugbek Nurmatov, Joseph Liu, Claudia Pagliari, Josip Car, and Aziz Sheikh. Telehealthcare for asthma: a cochrane review. *Cmaj*, 183(11):E733–E742, 2011.

[158] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

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

[160] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

[161] Jeanne E Moorman, Lara J Akinbami, CM Bailey, HS Zahran, ME King, CA Johnson, and X Liu. National surveillance of asthma: United states, 2001-2010. *Vital & health statistics. Series 3, Analytical and epidemiological studies*, (35):1–58, 2012.

[162] Deborah Morrison, Sally Wyke, Karolina Agur, Euan J Cameron, Robert I Docking, Alison M MacKenzie, Alex McConnachie, Vandana Raghuvir, Neil C Thomson, and Frances S Mair. Digital asthma self-management interventions: a systematic review. *Journal of medical Internet research*, 16(2):e51, 2014.

[163] Clara Mosquera-Lopez, Robert Dodier, Nichole Tyler, Navid Resalat, and Peter Jacobs. Leveraging a big dataset to develop a recurrent neural network to predict adverse glycemic events in type 1 diabetes. *IEEE journal of biomedical and health informatics*, 2019.

[164] Markus Müller, Sebastian Stüker, and Alex Waibel. Language adaptive multilingual ctc speech recognition. In *International Conference on Speech and Computer*, pages 473–482. Springer, 2017.

[165] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[166] Flory L Nkoy, Bryan L Stone, Bernhard A Fassl, Karmella Koopmeiners, Sarah Halbern, Eun H Kim, Justin Poll, Joseph W Hales, Dillon Lee, and Christopher G Maloney. Development of a novel tool for engaging children and parents in asthma self-management. In *AMIA Annual Symposium Proceedings*, volume 2012, page 663. American Medical Informatics Association, 2012.

[167] Carlos Nunes, Ana Margarida Pereira, and Mário Morais-Almeida. Asthma costs and social impact. *Asthma research and practice*, 3(1):1, 2017.

[168] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-Garadi, and Uzoma Rita Alo. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, 105:233–261, 2018.

[169] US Department of Health, Human Services, et al. Expert panel report 3: Guidelines for the diagnosis and management of asthma. *Bethesda, MD: HHS, National Heart, Lung and Blood Institute, National Institutes of Health. Publication*, (07-4051), 2007.

[170] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[171] Carsten Orwat, Andreas Graefe, and Timm Faulwasser. Towards pervasive computing in health care–a literature review. *BMC medical informatics and decision making*, 8(1):26, 2008.

[172] Ki-Hee Park and Seong-Whan Lee. Movement intention decoding based on deep learning for multiuser myoelectric interfaces. In *2016 4th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–2. IEEE, 2016.

[173] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[174] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[175] PhysioBank PhysioToolkit. Physionet: components of a new research resource for complex physiologic signals. *Circulation. v101 i23. e215-e220.*

[176] Vincent Pichot, Jean-Michel Gaspoz, Serge Molliex, Anestis Antoniadis, Thierry Busso, Frédéric Roche, Frédéric Costes, Luc Quintin, Jean-René Lacour, and Jean-Claude Barthélémy. Wavelet transform to quantify heart rate variability and to assess its instantaneous changes. *Journal of Applied Physiology*, 86(3):1081–1091, 1999.

[177] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets. *arXiv preprint arXiv:1710.08531*, 2017.

[178] Sutharshan Rajasegarar, Christopher Leckie, and Marimuthu Palaniswami. Hyperspherical cluster based distributed anomaly detection in wireless sensor networks. *Journal of Parallel and Distributed Computing*, 74(1):1833–1847, 2014.

[179] Linda M Rasmussen, Klaus Phanareth, Hendrik Nolte, and Vibeke Backer. Internet-based monitoring of asthma: a long-term, randomized clinical study of 300 asthmatic subjects. *Journal of Allergy and Clinical Immunology*, 115(6):1137–1142, 2005.

[180] Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*, pages 73–100, 2016.

[181] ERVM Rikkers-Mutsaerts, AE Winters, MJ Bakker, HF van Stel, V van der Meer, JC de Jongste, JK Sont, and SMASHING Study Group. Internet-based self-management compared with usual care in adolescents with asthma: A randomized controlled trial. *Pediatric pulmonology*, 47(12):1170–1179, 2012.

[182] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[183] Charissa Ann Ronao and Sung-Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244, 2016.

[184] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[185] Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing.* PhD thesis, NATIONAL UNIVERSITY OF IRELAND, GALWAY, 2019.

[186] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, page 4. ACM, 2014.

[187] Gabrielle M Salib, Juan Fernando Maestre, Kenneth B Nimley, Nadia Dowshen, and Gabriela Marcu. The role of reflection and context in medication adherence tracking for people living with hiv. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page LBW042. ACM, 2018.

[188] Rubén San-Segundo, Henrik Blunck, José Moreno-Pimentel, Allan Stisen, and Manuel Gil-Martín. Robust human activity recognition using smartwatches and smartphones. *Engineering Applications of Artificial Intelligence*, 72:190–202, 2018.

[189] Ivan W Selesnick and C Sidney Burrus. Generalized digital butterworth filter design. *IEEE Transactions on signal processing*, 46(6):1688–1694, 1998.

[190] Ümit Şentürk, Ibrahim Yücedağ, and Kemal Polat. Repetitive neural network (rnn) based blood pressure estimation using ppg and ecg signals. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–4. IEEE, 2018.

[191] Ahmad Shahi, Brendon J Woodford, and Hanhe Lin. Dynamic real-time segmentation and recognition of activities using a multi-feature windowing approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 26–38. Springer, 2017.

[192] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *arXiv preprint arXiv:1702.04457*, 2017.

[193] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37, 2017.

[194] Chuan Shi, Zhiqiang Zhang, Ping Luo, Philip S Yu, Yading Yue, and Bin Wu. Semantic path based personalized recommendation on weighted heterogeneous information networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 453–462. ACM, 2015.

[195] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016.

[196] Muhammad Shoaib, Stephan Bosch, Ozlem Incel, Hans Scholten, and Paul Havinga. A survey of online activity recognition using mobile phones. *Sensors*, 15(1):2059–2085, 2015.

[197] Muhammad Shoaib, Stephan Bosch, Ozlem Incel, Hans Scholten, and Paul Havinga. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors*, 16(4):426, 2016.

[198] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul JM Havinga, and Ozlem Durmaz Incel. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 591–596. IEEE, 2015.

[199] Bruno M Silva, Joel JPC Rodrigues, Fábio Canelo, Ivo C Lopes, and Liang Zhou. A data encryption solution for mobile health apps in cooperation environments. *Journal of medical Internet research*, 15(4):e66, 2013.

[200] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[201] Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[202] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[203] Richard P Stankus, Prem K Menon, Roy J Rando, Henry Glindmeyer, John E Salvaggio, and Samuel B Lehrer. Cigarette smoke-sensitive asthma: challenge studies. *Journal of Allergy and Clinical Immunology*, 82(3):331–338, 1988.

[204] Odongo Steven Eyobu and Dong Han. Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. *Sensors*, 18(9):2892, 2018.

[205] Xing Su, Hanghang Tong, and Ping Ji. Activity recognition with smartphone sensors. *Tsinghua science and technology*, 19(3):235–249, 2014.

[206] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[207] Maojie Tang and Dongxiu Ou. Fall detection system for monitoring an elderly person based on six-axis gyroscopes. In *2018 3rd International Conference on Electrical, Automation and Mechanical Engineering (EAME 2018)*. Atlantis Press, 2018.

[208] William R Taylor and Paul W Newacheck. Impact of childhood asthma on health. *Pediatrics*, 90(5):657–662, 1992.

[209] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[210] Juha T Vainio. Bluetooth security. In *Proceedings of Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, Seminar on Internetworking: Ad Hoc Networking, Spring*, volume 5, 2000.

[211] Upkar Varshney. Pervasive healthcare and wireless health monitoring. *Mobile Networks and Applications*, 12(2-3):113–127, 2007.

[212] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[213] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[214] Revathy Venkataramanan, Dipesh Kadariya, Hong Yung Yip, Utkarshani Jamini, Krishnaprasad Thirunarayan, Maninder Kalra, and Amit Sheth. Determination of personalized asthma triggers from evidence based on multimodal sensing and mobile application. *arXiv preprint arXiv:1811.10073*, 2018.

[215] Margaret K Vernon, Ingela Wiklund, Jill A Bell, Peter Dale, and Kenneth R Chapman. What do we know about asthma triggers? a review of the literature. *Journal of Asthma*, 49(10):991–998, 2012.

[216] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[217] Sandra Voorend-van Bergen, Anja A Vaessen-Verberne, Anneke M Landstra, Hein J Brackel, Norbert J van den Berg, Daan Caudri, Johan C de Jongste, Peter J Merkus, and Mariëlle W Pijnenburg. Monitoring childhood asthma: web-based diaries and the asthma control test. *Journal of Allergy and Clinical Immunology*, 133(6):1599–1605, 2014.

[218] Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1041–1044. ACM, 2016.

[219] Jie Wan, Michael J O'grady, and Gregory M O'hare. Dynamic sensor event segmentation for real-time activity recognition in a smart home context. *Personal and Ubiquitous Computing*, 19(2):287–301, 2015.

[220] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[221] Ludi Wang and Xiaoguang Zhou. Detection of congestive heart failure based on lstm-based deep network via short-term rr intervals. *Sensors*, 19(7):1502, 2019.

[222] Yingxu Wang, Henry Leung, Marina Gavrilova, Omar Zatarain, Daniel Graves, Jianhua Lu, Newton Howard, Sam Kwong, Phillip Sheu, and Shushma Patel. A survey and formal analyses on sequence learning methodologies and deep neural networks. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 6–15. IEEE, 2018.

[223] Cheng-Hsiung Weng, Tony Cheng-Kui Huang, and Ruo-Ping Han. Disease prediction with different types of neural network classifiers. *Telematics and Informatics*, 33(2):277–292, 2016.

[224] Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6204–6208. IEEE, 2018.

[225] Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6204–6208. IEEE, 2018.

[226] How It Works. Propeller health. 2013. web.

[227] Dongrui Wu. Online and offline domain adaptation for reducing bci calibration effort. *IEEE Transactions on human-machine Systems*, 47(4):550–563, 2017.

[228] Fangzhao Wu and Yongfeng Huang. Personalized microblog sentiment classification via multi-task learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[229] Cheche Xie, Sheng Bi, Min Dong, and Yongfa Li. Recovery method for missing sensor data in multi-sensor based walking recognition system. In *2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 558–563. IEEE, 2018.

[230] Miao Xie, Song Han, Biming Tian, and Sazia Parvin. Anomaly detection in wireless sensor networks: A survey. *Journal of Network and Computer Applications*, 34(4):1302–1325, 2011.

[231] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[232] Guangxu Xun, Xiaowei Jia, and Aidong Zhang. Context-learning based electroencephalogram analysis for epileptic seizure detection. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 325–330. IEEE, 2015.

[233] Yilong Yang, Qingfeng Wu, Ming Qiu, Yingdong Wang, and Xiaowei Chen. Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[234] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 351–360. International World Wide Web Conferences Steering Committee, 2017.

[235] Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, Aliaksei Ivanou, and Yao Qian. Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 338–345. IEEE, 2015.

[236] Ye Yuan, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang. Wave2vec: Deep representation learning for clinical temporal data. *Neurocomputing*, 324:31–42, 2019.

[237] Hatice S Zahran, Cathy M Bailey, Xiaoting Qin, and Jeanne E Moorman. Assessing asthma control and associated risk factors among persons with current asthma–findings from the child and adult asthma call-back survey. *Journal of asthma*, 52(3):318–326, 2015.

[238] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

[239] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *International Conference on Web-Age Information Management*, pages 298–310. Springer, 2014.

[240] Sheng Zhou, Jiajun Bu, Xin Wang, Jiawei Chen, Bingbing Hu, Defang Chen, and Can Wang. Hahe: Hierarchical attentive heterogeneous information network embedding. *arXiv preprint arXiv:1902.01475*, 2019.