ABSTRACT

ZHANG, YUAN. Every Moment Counts: Deep Variability Reasoning in EHR Data. (Under the direction of Min Chi).

Large-scale and systematic patient information documentation enables understating the medical history of patients, identifying interesting cohorts, predicting risks and evaluating interventions. Electronic Health Records (EHRs), one success of such effort, facilitate a systematic collection of temporal patient data and have provided unprecedented opportunities for exploratory analysis. However, such data collection also poses numerous challenges for traditional data-driven approaches due to the involved complex variabilities. In real-world EHR datasets, these variabilities mainly include unreliable ground-truth, irregular time series, heterogeneous patient groups, and highly frequent missing values.

Deep reasoning beyond these variabilities is the key to understand, study and improve the outcomes of a disease, and hence serves a better medical care delivery to public health. In this dissertation, we design a general framework to address the variabilities in multi-variate EHR data, which consequently assists a better clinical decision making. In particular, we apply the framework to predict septic shock, which is a life-threatening organ dysfunction and has an extremely high mortality rate.

Deep learning techniques have shown tremendous success for a wide scope of real-world tasks. In this work, we adopt a recurrent neural network (RNN) model, which is a widely used sequential deep learning model, to capture long-term temporal dependencies and underlying progressing patterns in septic shock. Our proposed framework is comprised of four thrusts described as follows:

Thrust 1: Multi-scale label learning. The learning performance of the conventional RNN model can be degraded given the difficulty in collecting reliable ground-truth labels in health-care domains. To this end, we design a learning process by jointly leveraging the imperfect labels from different resources (ICD-9 codes vs. clinical criteria) and at different scales (visit vs. event) to improve the prediction accuracy.

Thrust 2: Time-aware attention networks. In practice, clinicians refer to the records of several previous events to diagnose the medical condition of a patient. In some cases, the disease progression at a certain event may not rely on most recent events, but are more relevant to other events at earlier time steps. On the other hand, the time interval between two consecutive events is irregular, which impacts the dependency between consecutive events. To mimic the process of doctors reading records, we integrate attention mechanism with consideration of time irregularity to the conventional RNN model, to interpret the prediction results and identify the contributions of the historical events to the current event.

Thrust 3: Temporal domain adaptation. There exists heterogeneity among different patient

groups (domains), i.e. race, age and gender. This leads to discriminative distributions in the nature of EHR data. The learning model for one specific domain cannot directly apply to another domain. Hence, we utilize domain adaptation techniques to learn the correspondence among different patient groups. We propose a domain adaptation (DA) method, combining two temporal RNN-based recurrent components, to extract an invariant feature representation across different patient groups (domains) through an adversarial learning process. While learning the temporal developing patterns of the disease, the method also identifies the temporal changes of feature distributions across patient groups. Consequently, we aim to provide a more specific diagnosis.

Thrust 4: Variational RNN for missing data generation. Highly frequent missing values in EHR data can impact the decision-making process, which cannot be handled by most existing machine learning approaches. We propose to impute missing values using a temporal generative model (Variational Recurrent Neural Networks, VRNN). The generation of missing data enables the modeling of a complete disease progression, which consequently contributes to a better classification.

In this dissertation, we first validate each thrust separately and then evaluate the whole framework using a real-world EHR dataset on septic shock. Results demonstrate that the proposed framework achieved great success for identifying high-risk patients at an early stage. Through the deep reasoning in the variabilities of EHR data, the framework learns a more reasonable, realistic and informative representation of patients' health status, and hence sheds some lights on better understanding the disease progression and providing timely diagnosis and treatment for researchers and practitioners.

© Copyright 2019 by Yuan Zhang

All Rights Reserved

Every Moment Counts: Deep Variability Reasoning in EHR Data

by
Yuan Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2019

APPROVED BY:

Xipeng Shen	Ranga Raju Vatsavai
Julie Ivy	Min Chi Chair of Advisory Committee

DEDICATION

To my family, my teachers, the giants of human knowledge on whose shoulders I stand.

BIOGRAPHY

Yuan Zhang was born in Jinan, China. She obtained her bachelor's degree from Beijing University of Technology. She studied Law during her first college year and then transferred to the Department of Electronic Engineering. In the 4th year, she was awarded Outstanding Graduate and was recommended for graduate admission to Beijing University of Technology. In 2013, she obtained her master's degree in Electronic Engineering. During the period she was pursuing her master's degree, she worked as a research assistant in Digital Multimedia Information Processing and Communication Technology Laboratory (MIP&CT) and had been working on video-based automatic detection and prediction of the real-time traffic states. She developed her interests in challenging problems of algorithms, machine learning, data analysis, computer vision, etc.

After graduating from Beijing University of Technology, Yuan joined a start-up game company and contributed to developing the first 3D First Person Shoot web-page game in China. She made great progress in programming, system design and team communication. In 2014, she started pursuing her Ph.D. degree in Computer Science at North Carolina State University. She obtained a 4.0 cumulative GPA in her academic records and has been working as a research assistant under Dr. Min Chi at the Center for Educational Informatics (CEI). Most of Yuan's work was devoted to the fields of education and healthcare. The first project she was involved is Automatic Short Answer Grading (ASAG). Yuan applied natural language processing techniques to assess short student-authored answers and explored incorporating additional relevant information to improve the performance of conventional ASAG methods. Second, Yuan has also actively participated in an interdisciplinary research project, Sepsis Early Prediction Support Implementation System (S.E.P.S.I.S). She collaborated with students and faculties from Industrial Engineering, and clinicians and statisticians from Mayo Hospital and Christian Care Health System. By researching on specific challenges associated with EHR data and septic shock, she developed a deep neural network-based framework to provide realistic and reasonable prediction for the task.

Besides her academic life in school, Yuan has gained industrial experience from several internships, including eBay, Mayo Clinic and Apple, which gave her great opportunities to apply the knowledge into practice and contribute to the society.

ACKNOWLEDGEMENTS

I would like to sincerely thank my academic advisor, Dr. Min Chi, for her support, patience, and encouragement over the past four years. Her extensive knowledge and insightful advice have substantially broadened my views to the domains of education and healthcare. I am always inspired by her enthusiasm, perseverance and conscientiousness for research. Interactions with her were highly motivating and confidence boosting, especially during tough phases of my research. Besides, she is kind, caring and sweet, she is someone I will instantly love and forever be thankful to.

I would like to thank Dr. Xipeng Shen, Dr. Raju Vatsavai and Dr. Julie Ivy for kindly serving on my thesis committee. They provided instrumental suggestions regarding research organization, experimental design, statistical analysis and system implementation, which enable me to thoroughly investigate on my work and get further improvement.

I would like to thank all the faculties, Dr. Julie Ivy, Dr. Maria Mayorg, Dr. Min Chi and Dr. Osman Ozaltin, and the students from the Department of Industrial Engineering, Joseph Agor, Kendall McKenzie, and Nisha Nataraj, etc, involved in S.E.P.S.I.S project. Their hard work and excellent management keep our project proceeding smoothly and being productive. I would like to thank our collaborators, Dr. Muge Capan, Dr. Jeanne Huddleston, Dr. Santiago Romero Brufau and Dr. Ryan Arnold. They provided us with clinical insights and invaluable feedback whenever we encountered challenges and confusions from the data and from medical science.

I would like to thank my great lab mates as well as dearest friends Chen Lin, Xi Yang, Song Ju, Guojing Zhou, Yeojim Kim, and Farzaneh Khoshnevisan in CEI. It has been fortunate to work with them and I was always motivated from our discussions. I very much appreciate their kindnesses, support and encouragement for my personal life. I would like to thank my friends Jiyu Wang, Shuchi Liu, Yang Lei, Chi Zhang, Xinyu Liang, who have made my graduate life full of excitement and fun. We spent countless wonderful time together in Raleigh and I am always grateful for their friendship and support.

I would like to thank my parents and parents-in-law for their constant understanding and unconditional love. My husband, Xiaowei Jia, receive my deepest gratitude and love for his dedication to my work and life. He prepared me for major courses of computer science. He imparted rudimentary machine learning knowledge to me when I was a beginner. He set a good example for me by his passionate and serious attitude for research. He always enlightened and encouraged me whenever I was depressed or stressful. He is a good husband, as well as a good friend and teacher. Finally, I sincerely thank my dearest unborn son, Sinuo, who has brought enormous amount of motivation and happiness to my life.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **viii**

LIST OF FIGURES . . . . . **ix**

Chapter 1 INTRODUCTION **1**
1.1 Research Overview 1
1.2 Thesis Statement And Hypotheses 3
1.3 Contribution 6
1.4 Organization 6

Chapter 2 Background **7**
2.1 Machine Learning for EHR 7
2.2 Recurrent Neural Networks 9
2.3 Sepsis/Septic Shock Prediction 10

Chapter 3 DATASETS **13**
3.1 Data Description 13
3.2 Data Preprocessing 14
3.2.1 Tagging 15
3.2.2 Sampling 17
3.2.3 Data Selection 18
3.2.4 Feature Extraction 19
3.2.5 Data Aggregation 19
3.3 Data Split 20

Chapter 4 Thrust 1: Learning from Multi-scale Labels **22**
4.1 Overview 22
4.2 Related Work 24
4.3 Method 25
4.3.1 Problem Definition 25
4.3.2 LSTM for Sequential Classification 25
4.3.3 Learning from Multi-scale Labels 27
4.4 Experiments 28
4.4.1 Approaches 28
4.4.2 Experimental Setup 29
4.4.3 Evaluation and Setting 30
4.5 Results 30
4.5.1 Overall Validation of Multi-scale Label Learning 30
4.5.2 Validation on various machine learning classifiers 32
4.5.3 Validation Using Different Ground-truth Labels 32
4.5.4 Early Prediction 34
4.6 Summary 35

Chapter 5 Thrust 2: Time-aware Attention Networks . . . 36
5.1 Overview . . . 36
5.2 Related Work . . . 37
5.3 Method . . . 38
5.3.1 Three Attention Mechanisms. . . . 39
5.3.2 Time-aware Decay Function . . . 40
5.4 Experiments . . . 41
5.4.1 Approaches . . . 41
5.4.2 Experimental Setup . . . 42
5.5 Results . . . 42
5.5.1 Results of Overall Prediction . . . 42
5.5.2 Results of Early Prediction . . . 44
5.5.3 Case Study . . . 45
5.6 Summary . . . 46

Chapter 6 Thrust 3: Deep Domain Adaptation . . . 47
6.1 Overview . . . 47
6.2 Related Work . . . 48
6.3 Method . . . 49
6.3.1 Domain Adaptation with Adversarial Learning . . . 49
6.3.2 Temporal Adversarial Networks . . . 51
6.4 Experiments . . . 53
6.4.1 Data Grouping . . . 53
6.4.2 Approaches . . . 54
6.4.3 Experimental Setup . . . 55
6.5 Results . . . 56
6.5.1 Domain Adaptation Using Single-factor . . . 56
6.5.2 Domain Adaptation Using Two Factors . . . 61
6.6 Summary . . . 63

Chapter 7 Thrust 4: Variational Recurrent Neural Networks . . . 64
7.1 Overview . . . 64
7.2 Related Work . . . 65
7.3 Method . . . 65
7.3.1 Variational Autoencoders . . . 65
7.3.2 Variational Recurrent Neural Networks . . . 66
7.4 Experiments . . . 68
7.4.1 Missing Data Analysis . . . 68
7.4.2 Approaches . . . 69
7.4.3 Experimental Setup . . . 69
7.5 Results . . . 69
7.5.1 Results of Overall Prediction . . . 69
7.5.2 Results of Early Prediction . . . 71
7.6 Summary . . . 72

Chapter 8 Conclusion and Future Work **73**

BIBLIOGRAPHY **77**

APPENDIX **89**
Appendix A Results Integrating Thrust 1 to Thrust 4 90
A.1 Overview 90
A.2 Experiments 91
A.2.1 Experimental Setup 91
A.2.2 Evaluation and Setting 91
A.3 Results 91
A.3.1 Results of Overall Prediction 91
A.3.2 Results of Early Prediction 93
A.4 Summary 94

LIST OF TABLES

Table 3.1 Data distribution of CCHS according to the consistency between visit-level and event-level labels. . . . 18
Table 3.2 After sampling: data distribution of CCHS according to the consistency between visit-level and event-level labels. . . . 18
Table 3.3 After selection: data distribution of CCHS according to the consistency between visit-level and event-level labels. The number of events is in bold and the number of visits is in parenthesis. . . . 19
Table 3.4 After aggregation: data distribution of CCHS according to the consistency between visit-level and event-level labels. The number of events is in bold and the number of visits is in parenthesis. . . . 20
Table 3.5 Data distribution of training and testing sets from one sampling. The number of visits is fixed and the number of events might change from different samplings. 20

Table 4.1 Performance (±standard deviation) LSTM with or without the multi-scale label learning on different training sets. . . . 31
Table 4.2 Performance (±standard deviation) of multiple machine learning baselines with multi-scale label learning. . . . 33
Table 4.3 Performance (±standard deviation) of the framework on testing data with different ground-truth labels. . . . 33

Table 5.1 Performance (±standard deviation) of baselines and our approaches on septic shock overall prediction. . . . 43

Table 6.1 Distribution of consistent and inconsistent positive visits according to race and gender across the four age groups. . . . 54
Table 6.2 Performance (±std) of models with and without domain adaptation within race groups. . . . 57
Table 6.3 Performance (±std) of models with and without domain adaptation within age groups. . . . 57
Table 6.4 Performance (±std) of models with and without domain adaptation within gender groups. . . . 58
Table 6.5 Performance (±std) of models with and without domain adaptation within across different groupings. . . . 62

Table 7.1 Missing rates of 19 features from CCHS dataset. The first 9 features are vital signs, and the last 10 features are lab results. . . . 68
Table 7.2 Performance (±standard deviation) LSTM and VRNN with different imputation methods. . . . 70

Table A.1 Performance (±std) of LSTM and two integrated models with on all the test data or target domain data only. . . . 92

LIST OF FIGURES

Figure 1.1 The overall framework: originally data have a very high missing rate, we would like to fill up the missing values and then input them to the LSTM. In the learning process, we capture the informative previous events by time-aware attention networks and provide reliable ground-truth by jointly using two levels of labels. Then what have learned is used for prediction and other domains learning. The components in the graph are totally orthogonal. 3

Figure 3.1 The structure of EHRs: each block represents a visit and each bar represents an event. A visit consists of several events. At each event, the patient takes many measurements, such as vital signs of temperature and heart rate, and lab values of white cell count (WBC). 14

Figure 3.2 Pipeline of EHR data preprocessing. 15

Figure 4.1 (a) The structure of LSTM for sequential classification. (b) The internal structure of a LSTM cell. 26

Figure 4.2 The distribution of consecutive ones in non-shock and shock visits. 29

Figure 4.3 (a) The Original structure of LSTM. (b) The revised structure of LSTM for early prediction. 34

Figure 4.4 (a) AUC of the two models with shifting labels. (b) F1 of the two models with shifting labels. 35

Figure 5.1 Illustration of proposed ATTAIN model. ATTAIN accumulates the cell states from all the previous events and regularize the old memories through two sources of weights generated from attention mechanism ($\boldsymbol{\alpha}$) and time intervals ($\boldsymbol{\Delta}$). 39

Figure 5.2 AUC and F1 score of LSTM$_l$ when: (a) m is taken from $[1,199]$ with interval 10. (b) $m \in [20,40]$. 42

Figure 5.3 (a) AUC of the early prediction on five models. (b) F1-score of the early prediction on five models. 44

Figure 5.4 Attention weights for the 11th and 12th events achieved from (a) RETAIN; (b) LSTM$_l$; (c) ATTAIN$_l$. 45

Figure 6.1 Temporal Adversarial Networks. 51

Figure 6.2 Race and age distribution in positive septic shock visits. 54

Figure 6.3 Sensitivity of the domain adaptation process on the amount of unlabeled target domain data. 59

Figure 6.4 Unsupervised vs. supervised processes on male data. 60

Figure 6.5 AUC and F1-score of early prediction in: (a) Race: WA-AA; (b) Age: Sr-MID; (c) Gender: M-F. 61

Figure 7.1 Diagram of Variational Autoencoders. 65

Figure 7.2 Variational Recurrent Neural Networks: the blue line indicates the recurrent process, the yellow line stands for inference process and the green line stands for the generation process. 67
Figure 7.3 (a) AUC of early prediction on five models. (b) F1-score of early prediction on five models. 71

Figure A.1 (a) AUC of the early prediction on LSTM and two integrated models. (b) F1-score of the early prediction on LSTM and two integrated models. 93

CHAPTER

1

INTRODUCTION

1.1 Research Overview

Large scale and systematic patient information documentation enables understating medical history of patients, identifying interesting cohorts, predicting risks and evaluating interventions. Electronic Health Records (EHRs), one success of such effort, facilitates a systematic collection of temporal patient data and has provided unprecedented opportunities for exploratory analysis. However, EHRs also pose numerous challenges for traditional data-driven approaches due to the involved complex variabilities. In real-world EHR datasets, these variabilities mainly include *unreliable ground-truth*, *irregular time series*, *heterogeneous patient groups*, and *highly frequent missing values*.

Deep reasoning beyond these variabilities is the key to understand, study and improve the outcomes of a disease, and hence serves a better medical care delivery to public health. In this dissertation, we design a general framework to address the variabilities in multi-variate EHR data. In particular, we focus on the task of *Disease Progression Modeling* (DPM), which monitors the disease developing process and predicts future risks based on patients' historical information, and we focus on a specific disease septic shock, which is life-threatening organ dysfunction and has an extremely high mortality rate.

DPM is crucial for making clinical decisions and providing prompt medications. Given its importance, many recent works aim to develop automated solutions to DPM using machine learning approaches[Cho16a; Cho16b; Est16; Lip15; Zho11]. Among them, Recurrent Neural Network (RNN)

is one of the most widely used deep neural networks to handle the sequential EHR data. As an extension of RNN, the Long-Short Term Memory (LSTM) is specifically designed to capture long-term patterns that commonly exist over a long period of patients' records [Sun12]. LSTM-based approaches have found a huge success in a variety of tasks involving sequential data, such as video processing, climate changes detecting, traffic monitoring, and EHRs representation learning [Lip15; Jia17b; Jia17a; Est16; Zha17].

Despite its extensive use, significant barriers remain when applying the standard LSTM for modeling EHRs. The major research goal of this dissertation is to build a framework that further improves the performance of LSTM and hence better tackle the aforementioned variabilities in EHR data. Our proposed framework (Figure 1.1) is comprised of four trusts, which can be summarized in four research questions:

1. **Given the multi-scale imperfect labels (from visit levels and event levels), how to learn discriminative non-specific septic shock patterns in a long period?**

   LSTM is used to capture long-term temporal dependencies and underlying progressing patterns in septic shock. However, the performance of standard LSTM would degenerate due to the difficulty in collecting reliable ground-truth labels in health-care domains. To this end, we design a learning process by jointly leveraging the imperfect labels from different resources (ICD-9 codes vs. clinical criteria) and at different scales (visit vs. event).

   Solution: LSTM with Multi-scale Labels Learning.

2. **How to mimic the process of doctors reading records to capture informative historical moments?**

   In practice, clinicians refer to the records of several previous events to diagnose the medical condition of a patient. In some cases, the disease progression at a certain event may not rely on most recent events but are more relevant to another events at earlier time steps. On the other hand, the time interval between two consecutive events is irregular, which impacts the dependency between consecutive events. We integrate attention mechanism with consideration of time irregularity to the conventional RNN model, to interpret the prediction results and identify the contributions of the historical events to the current event.

   Solution: Time-aware Attention Networks.

3. **How to provide specific diagnosis considering heterogenous patient groups?**

   There exists heterogeneity among different patient groups (domains), i.e. race, age and gender. This leads to discriminative distributions in the nature of EHR data. The learning model for one specific domain cannot directly apply to other domain. Hence, we utilize domain adaptation techniques to learn the correspondence among different patient groups. We propose a domain

adaptation (DA) method, combining two temporal RNN-based recurrent components, to extract an invariant feature representation across different patient groups (domains) through an adversarial learning process. While learning the temporal developing patterns of the disease, the method also identifies the temporal changes of feature distributions across patient groups. Consequently, we aim to provide a more specific diagnosis.

Solution: Temporal domain adaptation.

4. **How to handle frequent missing values in EHR data?**

   We propose to impute missing values using a generative process (Variational Recurrent Neural Networks, VRNN). The generation of missing data enables the modeling of a complete disease progression, which consequently contributes to a better classification.

   Solution: Variational Recurrent Neural Networks (VRNN).

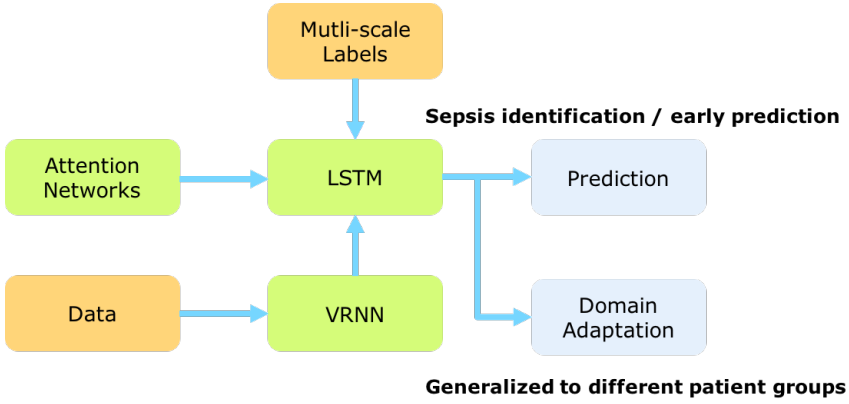


Figure 1.1 The overall framework: originally data have a very high missing rate, we would like to fill up the missing values and then input them to the LSTM. In the learning process, we capture the informative previous events by time-aware attention networks and provide reliable ground-truth by jointly using two levels of labels. Then what have learned is used for prediction and other domains learning. The components in the graph are totally orthogonal.

1.2 Thesis Statement And Hypotheses

In this work, we design a general framework to address the variabilities in multi-variate EHR data and specifically we will focus on prediction tasks for a specific type of disease, septic shock.

Our proposed framework is comprised of four thrusts, which correspond to the four research questions as stated in the research review. First of all, we design a learning process by jointly lever-

aging the imperfect labels from different resources (ICD-9 codes vs. clinical criteria) and at different scales (visit vs. event). Together with Long Short Term Memory (LSTM) networks, this method could reliably learn discriminative and non-specific septic shock patterns in a long period. Initially, the data are labeled at both visit level (ICD-9) and at event level (criteria defined by clinicians). In the training process, the framework first carries out a standard supervised learning cycle by training a LSTM model with event-level labels. Then we apply the learned LSTM model on the training data to calculate the probabilities that a patient is in the shock state at any given time point (event level). With the event-level output, we check the consistency between the event-level and the visit-level information (ICD-9), and revise the training objectives based on conditions. Then the trained LSTM model is updated using the training data with revised labels. The procedure is iteratively conducted until the model converges and no changes happen for the labels of training data. By utilizing this iterative training process, we expect the proposed framework to explore latent true labels of septic shock from combining two levels of noisy labels. In this way, the proposed framework can make accurate predictions to reflect real disease conditions.

Second, we propose time-aware attention LSTM networks to model the variabilities in EHR data over time dimension. In particular, the original LSTM pays attention to only one previous event when generating cell states. In time-aware attention networks, we accumulate all the cell states of previous events to get the historical information and scale each of them by assigning an attention weight. Attention weights are used to evaluate how important that event to the current event. Meanwhile, the time intervals are transformed to weights through a decay function so that the outdated events are more likely to play a less important role than recent events for predicting the outcome of the current event.

Third, there exists a shift between the distributions of different patient groups. We use domain adaptation techniques to learn an invariant feature representation across the patient groups (domains). Specifically, we investigate the effect of domain adaptation across different races, age levels and gender groups in our EHRs. In the feature learning process, we simultaneously optimize two discriminative classifiers, the label predictor to predicts class labels and the domain classifier to discriminate between different domains. In each iteration, the label predictor is optimized so as to extract a hidden representation which maximizes the classification performance while adversely degrading the domain classifier. In this way, the hidden representation is discriminative and also invariant across domains.

The last research question is regarding how to handle frequent missing values in EHR data. We propose to introduce the use of high-level latent random variables to model the variability observed in data. In the context of standard neural network models for non-sequential data, the variational autoencoder (VAE) offers an interesting combination of non-linear mapping between the latent random variables and the output and effective approximate inference. VRNN extends the VAE into a recurrent framework for multi-variate sequential data. There are three processes in VRNN, which

are recurrent process, inference process and generation process respectively. The inference process learns about what latent variables have most likely produced the input data, and then encodes the input to latent variables. The generation process will generate new data from the latent variables, using the learned distribution from the inference process. The recurrent process is responsible for modelling the transitions between time steps and is embedded into the other two processes.

The methods described in this work are applied to one real-world EHR dataset: the repository of Christiana Care Health System (CCHS). The raw data are under a complete preprocessing procedure and are transformed into a format suitable for machine learning systems and the analysis of current research, while minimizing the loss of information due to these transformations.

In the framework, all the thrusts are totally orthogonal. Each thrust is evaluated separately and compared to multiple strong baselines including the state of the arts. Finally, we combine them together and give an overall evaluation about the contribution of the whole framework.

The evaluation metrics include sensitivity/recall, specificity/true negative rate (TPR), precision/positive predictive value (PPV), F1 score, and area under the ROC (receiver operator characteristic) curve (AUC) [NH12]. Precision, recall, F1-score and AUC are widely used to measure the prediction performance for machine learning approaches. In medical science domain, researchers commonly refer to sensitivity (i.e. recall), specificity (i.e. TPR) and PPV for the annotation performance. Therefore in our tests, we include the metrics for both machine learning and medical science domains.

To summarize, this proposal proposes a general framework to solve the challenges from multiple aspects of variabilies in EHR data. The following hypotheses will be investigated in this proposal:

- Hypothesis 1: There exists vague and non-specific septic shock patterns at an early stage of the disease that can be captured by LSTM.
- Hypothesis 2: Either visit-level label and event-level labels are imperfect yet informative. The performance of the machine learning models will greatly be improved when combining them together to make use of good information from both of them.
- Hypothesis 3: The records of a patient in the current moment in not necessarily impacted by the most recent moment. Also, if the period between two successive moments is too long, the previous moment should not play an active role for the current moment.
- Hypothesis 4: There exists a shift in the distributions of the features collected from different groups (domains) of patients. The shift will degrade the performance of the traditional classification models.
- Hypothesis 5: The missing data would provide the incomplete information and would impact the conclusions from the data. Hence, we expect the generated missing data would be more realistic and increase the performance of the framework.

1.3 Contribution

The contributions of this dissertation are manifold. It investigates several aspects of variabilities, or non-deterministic factors, of EHR data and innovative integrate state of the art data mining and deep learning techniques to provide practical and efficient solutions.

First, the combination of multi-scale labels opens the way of learning from uncertainty. This is extremely useful since it is very difficult to collect ground-truth in the real world under tremendous applications. However, we may collect different sources of labels, learn informative information from each of them and make inaccurate information being corrected internally.

Second, the introduction of the attention mechanism with consideration to irregular time intervals and the domain adaptation among heterogeneous patient groups provide a more naturalistic hence more cogent way for clinicians to trust machine learning approaches. In practice, clinicians believe more in their experience and knowledge than in the conclusions draw from a black box. The attention networks are mimicking the process doctors look backwards at the records, and help clinicians target at the important moments quickly. Learning from different domains is helpful for cohort identification and give more specific diagnosis and treatment to different groups of patients.

Third, a lot of efforts have been spent on imputing missing data and hence get a better understanding of patient trajectories. However, the inherent disadvantage of deterministic methods would only generate the 'averaged' data. The averaged data may improve the performance of the models more or less, but it is meaningless in the real world. We set our foot on the solid foundations by applying generative models when generating data. Only striving for generating realistic data, we are approaching to deeply understand the disease progression with respect to each specific patient.

Last and equally important, we apply the whole framework on an extremely challenging disease, septic shock. Sepsis is a leading cause of all the world and takes a huge part of total hospital cost. The annualized incidence is still keeps rising. The proposed framework will help doctors identify the patients who are developing sepsis/septic shock and give immediate care for them, which would be a huge social good. On the other hand, the framework can be easily generalized to other types of disease and contribute to a better medical care delivery the public health.

1.4 Organization

The organization of this dissertation is as follows: Chapter 2 introduces the machine learning approaches that have been applied to EHRs, Recurrent Neural Networks and the background of sepsis/septic shock. Chapter 3 describes the dataset used in this dissertation and the data preprocessing procedure. Chapters 4-7 present a detailed description of the four research thrusts. For each thrust, the literature review, proposed framework, experimental design, and results are presented. Chapter 8 concludes this dissertation and discusses several possible directions for future research.

CHAPTER

2

BACKGROUND

In this section, we first briefly review the common machine learning approaches that have been applied to EHR data, and specially introduce the Recurrent Neural Networks, which is a basic model used in this dissertation. Next, we present an overview of the task of Disease Progression Modeling and sepsis/septic shock, the specific type of disease this dissertation focuses on.

2.1 Machine Learning for EHR

Electronic Health Records (EHRs) have reduced clinical errors [Sin12; Abr11; SG10], improved chronic illness care [Dor07; Sho09; Cro07] and improved the completeness, accuracy and timeliness of medical care to public health. On the other hand, EHRs also pose numerous challenges, such as they are sparse, noisy, heterogeneous, incomplete, high dimensional and systematic biased [Mio16]. Since EHR systems provide detailed care information of large patient populations with long follow-up times, these challenges have been addressed in machine learning community through data-driven healthcare research and advanced analytics.

Basically, machine learning approaches for healthcare can be broadly categorized as supervised or unsupervised: supervised methods take labels or outcomes as a learning guidance, while unsupervised methods simply learn the patterns from data. The temporal nature of EHR also suggests that these approaches can be further categorized into *temporal* methods and *non-temporal* methods, depending on if they take time into account.

Unsupervised methods have been widely used for identifying clusters of patients that have similar characteristics (e.g. demographics, medications, diagnosis codes, laboratory test results) and for finding associations between clinical concepts (e.g. medications, diagnosis codes and demographic attributes). For example, patient are stratified using hierarchical clustering, where the distance between patient records was computed by the cosine similarity of diagnosis codes [Roq11], or using medical concepts ((medication information, diagnosis codes, procedure codes), where the Jaccard index was used as the similarity measure [BM13]. Association rule mining is used to identify co-morbidities (e.g. non-insulin dependent diabetes mellitus (NIDDM) and cerebral infarction) which are strongly associated with hypertension [Shi10] and to detect transitive associations between laboratory test results and diagnosis codes [Wri10]. In addition, significant improvements have been achieved when temporal nature of EHRs is taken into account in these unsupervised methods. The temporal Hidden Markov Model (HMM) is used to cluster patient medical trajectories [Gha14]. The temporal abstraction framework has been frequently used to prepare patterns of from EHR data. Patterns can be fist abstracted using state representations (e.g. high, medium or low) or trend representations(e.g. increasing, decreasing, constant) [Sha97], and then be used to generate temporal association rules [Ver07; Hri15].

Generally, supervised methods aim to investigate the associations between risk factors and the outcome of interest. In non-temporal supervised methods, they usually transform time-dependent predictors to non-temporal predictors through temporal abstraction or summary statistics (e.g., mean, standard deviation), and then employ traditional predictive modeling techniques, such as decision trees [Man14] [Sar10], ensemble techniques (e.g. bagging, boosting, random forests) [Man14; Cho12; Kaw12; Kar12], logistic regression [Hua14; Cho12] and support vector machines [Man14] to predict a specific outcome. For example, these methods are widely used in predicting neonatal sepsis [Man14], potentially preventable events [SS13], risk of depression using diagnosis codes [Hua14], breast cancer survivability [Sar10], post-hospitalization VTE risk [Kaw12], atrial fibrillation [Kar12], mortality prediction in ICU [Her13] and so on. All of these non-temporal supervised methods rule out longitudinal analysis and do not adequately capture temporal relationships among medical events that are inherently existed in EHRs. On the other hand, these methods usually depend on feature engineering which is time-consuming and heavily based on clinical expertise.

Temporal supervised approaches have been extensively explored to learn the underlying temporal dynamics in EHRs. **Survival Modeling** have been used widely in the past to analyze the expected duration of time until one or more events happen (e.g. time until death in biological organisms). Cox regression [VR13] is one of the most commonly used survival regression models. Its semi-parametric nature with the mild assumption of the proportionality of hazards makes it fit to many fields including healthcare [Ike91] [Lum02]. For example, a stabilized sparse Cox model is proposed to estimate the feature graph derived from the temporal structure of the disease and intervention recurrences, and the hierarchical structure of medical knowledge and practices [Gop14]. Since cox models are

susceptible to overfitting as most other regression models, standard regularization techniques are applied, such as Lasso [Tib97] and elastic-net regularization [Sim11]. **Dynamic Bayes Networks** is another type of models that has been widely used to incorporate temporal relationships among EHRs. For example, DBN is applied to model temporal relationships between insulin and glucose homeostasis [Nac10]. DBN is used to assist physicians in monitoring the weight of patients suffering from chronic renal failure [Ros05], to model heart failure [Gat12] and model organ failure [Pee10]. These markovian models assumes that the patient's current medical state only depends on the his/her state on one previous visit/event, without considering the complexity of physiological processes in a long period and the nonlinear developing patterns among medical events. **Sequential Pattern Mining** has been recently implemented for monitoring and event detection problems in patients suffering from Type 2 diabetes mellitus (T2DM) through extending the temporal abstraction framework [Bat12]. Similar techniques are also used for detecting sequential rules associated with the early identification of brain tumors [SN13].

2.2 Recurrent Neural Networks

Deep Learning (DL) techniques have become exceptionally popular in healthcare domain since they can assist in mitigating the impact of noise and in learning complex relationships among medical events. Lots of prior works employ deep neural networks to map EHRs to clinical outputs in the tasks of phenotype learning [Las13; Mio16; Che15; Ham15] or representation learning [Cho16c; Mio16]. However, these deep neural networks do not fully address the temporal nature of EHRs and hence not capture the progressing patterns of patent's health status. To this end, Recurrent Neural Networks (RNN) [Elm90] that generates the outputs depending on not the current information but previous information is developed. Besides, the shared parameters over time enables the flexibility of dealing with sequences at varying length. Overviews of RNN for sequential learning and the broad applications are thoroughly described in multiple survey literature [Gra; DM15].

However, by the early 1990s, the vanishing gradient problem emerged as a major obstacle to RNN in long sequences [Pas12]. To tackle this problem, researchers have designed several effective variants of RNN, such as Long Short Term Memory (LSTM) networks [HS97]. LSTM is capable of memorizing temporal dependencies over a long period [Sun12] and have found a huge success in a variety of tasks involving sequential data, such as video processing, climate changes detecting, traffic monitoring, EHRs representation learning [Lip15; Jia17b; Jia17a; Est16; Zha17], etc. The ability of LSTM to recognize patterns in multivariate sequential EHRs has been validated on multi-label classification of diagnoses [Lip15].

Many efforts have been made, allowing RNN and LSTM to be better fit for EHRs applications. For example, the missing patterns led by irregularly spaced measurements in EHRs are investigated and the missing entries are further handled by forward or back-filling and missing indicators

[Lip16]. Some studies focus on addressing the episodic and irregular recording of EHRs. DeepCare [Pha16] introduces time parameterizations to enable irregular timing by moderating the forgetting dynamics in LSTM. GRU-D [Che16] achieves better prediction results on both synthetic and real-world healthcare datasets by incorporating informative missing patterns represented by masking and time intervals. The short-term memory is separated from the long-term memory at each time step of LSTM [Bay17]. Then the short-term memory is discounted according to the time spans between two consecutive moments while the global profile of the patient remains the same. Furthermore, it has shown to be effective that combining static information, such as demographics, and dynamic information, such as vital signs, of EHRs for clinical event prediction.

There have been active research in modifying the structure of RNN/LSTM and combining them with other efficient deep learning architectures, such as fully connected neural networks [Est], Convolutional Neural Networks (CNN) [Lai17], etc. RNN or LSTM with attention networks are widely developed to generate interpretations for EHRs [Cho16b; Ma17; SW17]. As frontier work, RETAIN [Cho16b] applied a two-level attention networks to identify meaningful visits and specific features that contribute to the prediction. Also, the performance of standard RNN/LSTM can be severely impacted due to the inherent heterogeneity among patient groups in EHRs [Nor17]. Domain adaptation, learning the common predictive patterns that help infer and transfer knowledge across different patient groups [Gan16], has become a natural solution to tackle such problems in healthcare field [HY09; Soc11; Pur16; Che18]. Generative models with recurrent structure, such as Variational Recurrent Neural Networks enables better modeling of variability in temporal progression, which has shown success in speech modeling, handwriting generation, and recommend systems [Chu15; Chr17]. VRADA, Variational Recurrent Adversarial Deep Domain Adaptation, combines the components of RNN, generative process and domain adaptation together and is the first work to empirically demonstrate that transferring the complex temporal latent dependencies between domains can help improve the model prediction for multivariate healthcare time-series data [Pur16].

2.3 Sepsis/Septic Shock Prediction

Disease Progression Modeling (DPM) monitors the disease developing process and predicts future risks based on patients' historical information. DPM is crucial for making clinical decisions and providing prompt medications. There have been many studies that focus on modeling the temporal progression of a specific disease, such as Alzheimer's disease [Ito10; Zho12; Suk12], glaucoma [Liu13], diabetes mellitus [DW06], and chronic kidney disease [Tan11], by taking advantage of intensive expert knowledge [DW06; Ito10; Tan11] or advanced statistics [Liu13; Suk12; Zho12]. On the other hand, progression models for a wider range of diseases have also been proposed. For example, Hawkes process is used to model multiple patients and multiple diseases [Cho15] and a graphical model based on Markov Jump Process is used to predict the stage progression of chronic obstructive

pulmonary disease (COPD) and its co-morbid diseases [Wan14].

In this dissertation, we validate the proposed framework on modeling the progression of a specific disease, septic shock. Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection [Sin16]. Sepsis is characterized with several negative factors including longer and larger number of hospitalizations, greater medical costs, and increased risk of death. As a leading cause of death in the United States, sepsis accounts for nearly $24 billion (6.2%) of total hospital costs in 2013 [TA16]. Each year 1.6 million people are diagnosed with sepsis, more than prostate cancer, breast cancer and AIDS combined [Eli11; HS17]. Septic shock is defined as a subset of sepsis in which particularly profound circulatory, cellular, and metabolic abnormalities are associated with a greater risk of mortality than with sepsis alone [Sin16]. As the most severe complication of sepsis, septic shock reaches a mortality rate as high as 50% [Mar03] and the annualized incidence keeps rising [DL08].

There exist plenty of general-purpose illness severity scoring systems to access health status of sepsis/septic shock patients [BL95; GZ11]. They include: Sequential Organ Failure Assessment (SOFA) Score [Vin96], Mortality in Emergency Department Sepsis (MEDS) [Sha03], Acute Physiology and Chronic Health Evaluation (APACHE II) [Boh88], Predisposition Insult Response and Organ failure (PIRO) [Rel09], Modified Early Warning Score (MEWS) [Sub01], Rapid Emergency Medicine Score (REMS) [Ols04], Simple Clinical Score (SCS) [Sub10], Simplified Acute Physiology Score (SAPS II) [LG93]. Even though these scores are useful for general deterioration and mortality, they typically cannot distinguish patients who are developing an acute condition (e.g. septic shock) with high sensitivity and specificity [Hen15]. For example, the AUC of SCS and REMS, which are concluded as the most appropriate clinical scores to predict the mortality of patients with sepsis [GZ11], are 0.76-0.79 and 0.74-0.79 respectively.

The broad adoption of EHRs in medical systems has promoted the development automating tools that identify patients with sepsis/septic shock and hence provide prompt treatment, e.g. early warning systems, track and trigger initiatives, sniffers and listening applications [Her11; NM14; Ums15; Nel11]. However, these tools can only detect patients currently experiencing sepsis/septic shock and cannot detect patients who are at great risk of developing sepsis/septic shock.

Prior studies have demonstrated that early diagnosis and treatment of septic shock can decrease patients' mortality and length of stay significantly [RM06; Kum06; CW11; Hen15]. As many as 80% of sepsis deaths could be prevented with rapid diagnosis and treatment[Kum06]. Conversely, every hour of delay in antibiotic treatment leads to an 8% increase in mortality for septic shock patients [Kum06]. Therefore, most recent studies have focused on early prediction of sepsis/septic shock [Giu07; Thi10; Hen14]. Given the rich clinical information (e.g., demographics, vitals, labs, treatment,etc) in EHRs, machine learning techniques are widely applied to construct predictive models by routine measurements [GG14]. For example, a septic shock early warning system (EWS) is developed using multivariate logistic regression to predict septic shock one hour prior to onset,

and achieves the AUC of 0.940, the sensitivity of 0.85 and the PPV of 0.70 [Sha07]. A targeted real-time early warning score (TREWScore) is fit in a Cox proportional hazards model and reports the AUC of 0.83, the specificity of 0.85 and the sensitivity of 0.74[Hen15]. A dynamic Bayesian network, modeling the progression of organ failure, reaches the AUC of 0.911 using 24-hour training data after admission and 0.944 using 3-hour training data after admission [Pee10]. SVM has also proven to be effective at sepsis-related classification tasks [Wan10; TM10]. Additionally, neural networks have been applied to medical problems for more than 20 years [Bax95; Bra02; Xun16]. In particular, temporal or sequential neural networks, e.g. RNN/LSTM, has been applied to prediction problems in genomics [Xu07], general diagnosis termed as phenotyping [Lip15; Est], etc, but not specifically implemented on the task of sepsis/septic shock prediction (to our best knowledge).

Nevertheless, several challenges remain to prevent early diagnosis and treatment for patients at risk of sepsis/septic shock. First, sepsis includes multiple subtypes and patients in different subtype might express drastically different symptoms [Rel09]. These symptoms at the early stage of sepsis/septic shock are often subtle and non-specific. Therefore, it is difficult for most sequential models to timely capture the discriminative patterns in sepsis/septic shock progression, especially when patient's stay spans over a long time. Second, there is continued disagreement between government agencies on the definition of sepsis and septic shock (e.g., Center for Medicare and Medicate, professional societies, and Third International Consensus Definitions for Sepsis and Septic Shock) [Sey16]. This leads to that different research institutes may adopt various biomarkers/indicators to identify sepsis-related population and use relatively small and carefully curated datasets with limited applicability to a broader population [Ho12; Ho14b]. To conclude, sepsis/septic shock prediction is a non-trivial task and robust models that can address these problems are in great demand.

CHAPTER

3

DATASETS

The framework described in this dissertation is applied to one real-world EHR dataset: the repository of Christiana Care Health System (CCHS). In this chapter, we briefly describe the details of the data. In addition, we perform a rigorous preprocessing procedure to transform and abstract the raw data into a format suitable for machine learning systems while keeping the loss of information minimized. Finally, we split the data into training and testing sets for future experiments of each thrust and the whole framework.

3.1 Data Description

We use retrospective EHRs collected from Christiana Care Health System (CCHS) as part of our collaborative National Science Foundation supported grant entitled S.E.P.S.I.S.: Sepsis Early Prediction Support Implementation System. CCHS is a none-profit healthcare system with over 53,000 hospital admissions per year and 1,100 hospital beds. The CCHS dataset consists of EHRs of hospitalized adult patients (i.e. age≥18) within CCHS from July 2013 to December 2015. In total there are 119,968 unique patients and 210,289 hospitalizations/visits. The length of hospital stays ranges from 30 minutes up to three years. Emergency department (ED) visits without subsequent hospital admission and visits with any surgical hospitalization are excluded and all the data are anonymized.

EHRs are multi-variate sequential data. In our data, each data point is a visit/episode/encounter of patients to hospital constituted by several events, and each event consists of a fixed number of clin-

ical measurements such as temperature, heart rate, etc. The measurements might leave blank if patients are not taken at the moment. The structure of our EHRs is shown in Fig. 3.1. Different visit may have various number (ranging from 1-3,464) of events, but the measurements/attributes/variables at each event are structured. In total, we have 10,412,729 medical events in 210,289 visits.

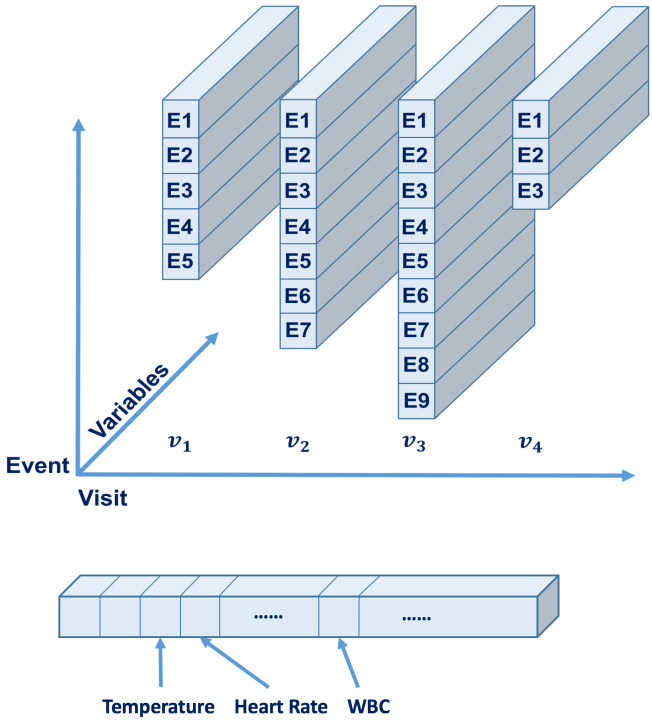


Figure 3.1 The structure of EHRs: each block represents a visit and each bar represents an event. A visit consists of several events. At each event, the patient takes many measurements, such as vital signs of temperature and heart rate, and lab values of white cell count (WBC).

3.2 Data Preprocessing

The raw EHRs are distributed in different tables in the system and have various formats. We perform a rigorous data preprocessing, determined by the leading clinicians with extensive experience on this subject from CCHS and Mayo Clinic, to prepare the data and narrow down the study population for our machine learning framework. The schematic of the overall process is shown in Fig. 3.2 and the details are described as follows.

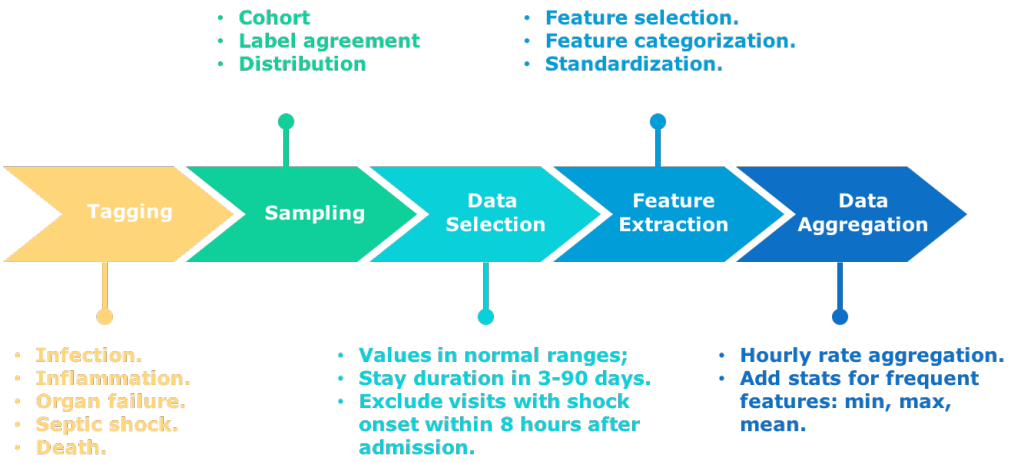


Figure 3.2 Pipeline of EHR data preprocessing.

3.2.1 Tagging

International Classification of Diseases, Ninth Revision (ICD-9) are widely used for clinical labeling (i.e., sepsis/septic shock or not). However, as visit-level labels, ICD-9 codes cannot tell when septic shock occurs at event level, and as administrative codes, they have been in doubt regarding the ability to establish gold standards for clinical conditions [Ho14a].. Therefore based on Third International Consensus Definitions [Sin16], *at each event*, our clinicians defined four stages of sepsis ordered by increasing severity: Infection, Inflammation, Organ Failure, and Septic Shock as follows:

- **Infection**: administration of a single dose of any anti-infective (antibiotic, antiviral, or anti-fungal) or any positive PCR test result.

  Rationale: our clinicians aim to include all the patients with *suspected infection*, or sepsis-related population, by setting a relatively low criteria. We define this population as our study population and narrow down to a dataset with 52,919 visits and 4,224,567 medical events from the original dataset with 10,412,729 medical events in 210,289 visits.

- **Inflammation**: the presence of an abnormality within any of the inflammatory criteria defines a positive inflammatory state, with each abnormality being summative representing the inflammatory burden:

  – Cellular Response:

  White blood cell > 12,000 cells;

  Bandemia of white blood cells > 10%;

  Erythrocyte Sedimentation Rate (ESR) > 20 mm/hr;

C-Reactive Protein (CRP) > 8 mg/L;

Procalcitonin > 0.15 ng/mL.

– Physiologic Response:

Heart rate >= 90 beats / minute;

Respiratory rate >= 20 breaths / minute;

Shock Index > 1.0;

Temperature >= $38^o C$ ($100.4^o F$);

Temperature < $36^o C$ ($96.8^o F$).

Rationale: inflammation exists on a broad spectrum that can be identified by different biomarkers. Expanding the inflammatory criteria can be useful for the risk stratification of sepsis. Its inclusion is mainly intended for investigation purpose.

- **Organ Failure**: the presence of an abnormality within any of the specified organ systems criteria defines a positive organ failure state. Each dysfunction is summative representing an organ failure burden:

  – Cardiovascular:

  Systolic blood pressure < 90 mmHg;

  Mean arterial pressure < 65 mmHg;

  Decrease in systolic blood pressure > 40 mmHg within an 8-hour window.

  – Metabolic:

  Lactate > 2.0 mmol/L.

  – Renal:

  Creatinine > 1.2 mg/dL;

  Creatinine increase > 50% from initial creatinine;

  Urine output < 500mL over 24 hours;

  BUN > 20 mg/dL.

  – Respiratory:

  New oxygen requirement;

  Mechanical ventilation requirement;

  Pulse Oxygen OR SpO2 < 90%;

  Fio2 > 21%;

  SpO2 / FiO2 ratio < 421.

- Nervous:

    Glasgow Comma Score (GCS) < 14;

    Verbal GCS <5.

- Gastrointestinal

    Bilirubin > 2 mg/dL.

Rationale: the criteria of the presence of organ failure with an active infection is widely used by previous works and scoring systems to identify patients with sepsis. However, the conditions defined in this criteria are not clearly explained and are practiced in different ways by different hospitals. Here our clinicians include such definition for future incorporation of novel predictive tools and re-design of EHRs.

- **Septic Shock**: the presence of an abnormality within any of the specified septic shock criteria defines a positive septic shock state:

    - Persistent hypotension as shown through two consecutive readings (>=30 minutes apart) for more than one hour;

        Systolic Blood Pressure < 90 mmHg;

        Mean arterial pressure < 65 mmHg.

    - Any vasopressor administration.

    Rationale: Traditional definition of septic shock usually presumes an Intensive Care Unit (ICU) admission. The definition here relies on patient-centered physiological criteria, i.e. having received vasopressor(s) or having had persistent hypotension, and is also able to identify non-ICU population. In this dissertation we use this definition to obtain event-level labels of septic shock.

3.2.2 Sampling

We cannot solely depend on either visit-level label, i.e. ICD9 codes, or event-level label, i.e. clinical criteria (described in Chapter 3.2.1), as ground truth of septic shock (described in Chapter 4.1). Here we define the consistency between visit-level label and event-level label to separate *consistent* data with more reliable labels and *inconsistent* data with less reliable labels as follows. The two-level labels are consistent (in blue) if: 1) for positive samples, ICD-9 indicates a septic shock visit and there exists at least one event in this visit meeting the shock criteria (i.e. with positive event-level label); and 2) for negative samples, ICD-9 shows no indication of septic shock and all events in this visit are not labeled as shock by the event-level criteria. Otherwise the two-level labels are inconsistent.

Table 3.1 shows the distribution of visits regarding the consistency between the two levels of labels in CCHS. When applying both ICD-9 codes and our clinical criteria, we identify 1,869 shock positive visits and 23,901 negative visits, referred to as *consistent* visits. Given the imbalanced ratio of positive and negative visits, we further conduct stratified random sampling on the 23,901 negative visits while keeping their underlying distribution of age, gender, ethnicity and length of stay the same as positive visits, and then get 1,869 consistent negative visits. Additionally, we keep 86 *inconsistent* visits (ICD-9(+) & Criteria(-)), and sample 6,147 *inconsistent* visits (ICD-9(-) & Criteria(+)). The data distribution after sampling is shown in Table 3.2.

Table 3.1 Data distribution of CCHS according to the consistency between visit-level and event-level labels.

All	Criteria(+)	Criteria(-)	Total
ICD-9(+)	1,869	86	1,955
ICD-9(-)	27,063	23,901	50,964
Total	28,932	23,987	52,919

Table 3.2 After sampling: data distribution of CCHS according to the consistency between visit-level and event-level labels.

All	Criteria(+)	Criteria(-)	Total
ICD-9(+)	1,869	86	1,955
ICD-9(-)	6,147	1,869	8,016
Total	8,016	1,955	9,971

3.2.3 Data Selection

EHRs contains errors and noise. We exclude the measurements in clinically unreasonable ranges. And we only include visits with stay duration from 3 days to 90 days since short visits do not hold sufficient information for analysis and long visits introduce data sparsity. More importantly, we exclude the visits that developed septic shock within 8 hours after admission to hospital, because our clinicians suggest that patients in such visits generally already show septic shock symptoms on admission and doctors can easily catch such cases. A more challenging yet more useful task is

to accurately predict septic shock after 8-hour admission. The data distribution, the number of events (in bold) and the number of visits (in parenthesis) after data selection is shown in Table 3.3. Note that we keep the number of consistent negative visits as the same as the number of consistent positive visits.

Table 3.3 After selection: data distribution of CCHS according to the consistency between visit-level and event-level labels. The number of events is in bold and the number of visits is in parenthesis.

All	Criteria(+)	Criteria(-)	Total
ICD-9(+)	**373,991** (1,791)	**7,980** (79)	**381,971** (1,870)
ICD-9(-)	**733,637** (6,147)	**114,658** (1,791)	**848,295** (7,938)
Total	**1,107,628** (7,938)	**122,638** (1,870)	**1,230,266** (9,808)

3.2.4 Feature Extraction

We include all the features shown in our tagging criteria as described in Chapter 3.2.1. We exclude features with a missing rate more than 90% to reduce the sparsity of the data. In our raw EHRs, categorical features are input from drop-down lists or in forms of free text. We first unify the categories with each categorical feature through natural language processing techniques and transform them into one-hot vectors. All the numeric features are standardized.

In brief, other than the patient information, such as gender, age, ethnicity, etc., each event at one visit consists of attributes/features of four categories: 1) vital signs: heart rate, temperature, etc.; 2) lab results: BUN, creatinine, white blood cell count (WBC), 18 culture tests, etc.; 3) interventions: FIO2, oxygen flow, etc.; 4) locations (e.g., emergency or nurse), descriptions, identifiers.

3.2.5 Data Aggregation

In the collecting process of EHRs, an event is created whenever a measurement/feature is taken. For example, one event is created when the patient's temperature and heart rate are measured, and a new event is created when the patient's blood pressure is measured a few minutes later. Since events are structured with a fixed number of features, the first event would have only two non-empty features and have the missing values for all other features. Similarly, the second event would be almost blank. To mitigate such data sparsity, we aggregate the events recorded within one-hour window to one event and take the mean value of each feature as values for the new generated event. For frequently measured features, such as vital signs, we also involve statistical features (e.g. min and max values in the same window). This process greatly reduces the number of events, as shown in Table 3.4,

and the missing rate of the whole dataset from ∼ 80% to ∼ 60%. We will investigate different data imputation methods in Chapter 7. In other experiments we adopt the baseline mean-imputation methods, which fills all the missing values with the mean value of the corresponding features.

Table 3.4 After aggregation: data distribution of CCHS according to the consistency between visit-level and event-level labels. The number of events is in bold and the number of visits is in parenthesis.

All	Criteria(+)	Criteria(-)	Total
ICD-9(+)	**151,037** (1,791)	**4,190** (79)	**155,227** (1,870)
ICD-9(-)	**368,913** (6,147)	**66,729** (1,791)	**435,642** (7,938)
Total	**519,950** (7,938)	**70,919** (1,870)	**590,869** (9,808)

3.3 Data Split

We always validate our framework on the data with more reliable ground-truth, i.e. *consistent* visits. To do so, we include ∼80% of the positive consistent visits (ICD-9(+) & Criteria(+): 1791) in the training set, and the remaining ∼20% of them in the testing set. Even in these positive consistent visits, the negative events are dominant and their ratio accounts for 87% of total number of events (131,056 out of 151,037). Since we perform evaluation at event-level and negative visits only contribute negative events, we only sample a few number of consistent negative visits (ICD-9(-) & Criteria(-): 1791) in training and testing set, and the ratio of positive events falls in the range of 10%∼15%.

Table 3.5 Data distribution of training and testing sets from one sampling. The number of visits is fixed and the number of events might change from different samplings.

(a) Training Set.

All	Criteria(+)	Criteria(-)
ICD-9(+)	**121,027** (1,425)	**4,190** (79)
ICD-9(-)	**90,525** (1,425)	**5,399** (150)

(b) Testing Set.

All	Criteria(+)	Criteria(-)
ICD-9(+)	**30,010** (366)	**0** (0)
ICD-9(-)	**0** (0)	**1,435** (40)

In some thrusts of the proposed framework, we also explore if the data with less reliable ground-truth, i.e. *inconsistent* visits, would help improve the model performance. Therefore, we sometimes

include inconsistent data in the training data. In specific, we keep all the 79 visits with labels ICD-9(+) & Criteria(-) and sample 1,465 visits (the same number of positive consistent visits) with labels ICD-9(-) & Criteria(+). The distribution of visits and events from one sampling is shown in Table 3.5.

CHAPTER

4

THRUST 1: LEARNING FROM MULTI-SCALE LABELS

4.1 Overview

This thrust investigates two major challenges associated with sepsis/septic shock prediction: the first one is the subtle progression of sepsis/septic shock in a long period and the second one is the lack of well-established definition for labeling sepsis/septic shock. Correspondingly, we provide a solution of learning from multi-scale labels.

As for the first challenge, the clinical signs and symptoms at the early stage of sepsis/septic shock are often subtle and non-specific. For example, only minor changes are reflected on white cell count and body temperature at the early stage of sepsis. Moreover, infection, a hallmark of sepsis, is highly likely to progress to other disease and hence not a specific symptom for sepsis. Therefore, it is critical to learn about the discriminative patterns of sepsis/septic shock and capture the informative progression during a patient's stay. It becomes especially challenging when the whole stay spans over a long time since it is difficult for most of the sequential models to memorize so many details and persistently connect previous information to the present without much loss.

To overcome this challenge, we model the septic shock progression using a variant of Long-Short Term Memory (LSTM) networks, which has shown extensive prospects in a variety of sequential labeling applications, such as climate changes, health-care records, traffic monitoring, etc [Lip15;

Jia17b; Jia17a; Est]. Compared with other sequential models like Recurrent Neural Networks (RNN), LSTM is capable of memorizing temporal dependencies over a long period [Sun12]. On the other hand, the electronic health records (EHRs) is a popular research platform with increasing availability to develop predictive models for acute medical conditions such as septic shock. However, EHR data also poses numerous challenges, such as they are noisy, fragmental and high dimensional. To this end the incorporation of temporal dependencies can assist in mitigating the impact of noise and in learning complex relationships among features.

The second challenge stems from no gold-standard definition or criteria of labeling sepsis/septic shock at any given time point. Due to varying purposes and expertise levels on the disease, different decision-making systems have certain *biases* on labelling sepsis/septic shock. During each hospital visit, a patient usually takes multiple tests and gets measurements at different time points. The assessment for the entire visit provides a *visit-level* label while the assessment for each time point provides an *event-level* label. We argue that though informative, both levels of labels are imperfect and they do not even agree for many cases.

At the visit level, septic shock patients are identified through the so-called International Classification of Diseases, Ninth Revision (ICD-9) code (“785.52") for billing purpose. As billing codes, ICD-9 is only coded for limited number of complications and dramatically different diseases can often share the same billing code as long as they have the same cost. Hence, ICD-9 may report false positives of septic shock and cannot fully represent the real medical states of patients. Indeed it has been widely argued that ICD-9 codes cannot be used for establishing reliable gold standards for various clinical conditions [Ho14b; Giu07].

At the event level, septic shock is usually determined by clinical criteria which is generally set based on current vital signs and lab tests, such as “hypotension" or “ongoing vasopressor therapy". The event-level labels are often noisy since the attributes/features involved in the criteria are not always observable and updated at each time point. For example, white cell count is only tested every 24 hours. Generally in this situation the unavailable values are filled with the ones from previous moments, which may lead to out-of-state evaluation for patients’ states. Moreover, the values of some attributes return to normal after patients receive a treatment, such as vasopressor administration, resulting in a negative event-level label even though the patient is still in septic shock. Finally and most importantly, there exists no well-established clinical criteria on whether a patient is in septic shock state at each time point.

To summarize, although visit-level labels and event-level labels allow us to grasp certain useful information of the disease, we cannot directly utilize them as ground truth of septic shock. Therefore, we propose a framework to learn the sequential patterns of septic shock by jointly leveraging both levels of imperfect yet informative labels. Initially, the data are labeled at both visit level (ICD-9 codes) and at event level (criteria defined by clinicians in Chapter 3.2.1). In the training process, the framework first carries out a standard supervised learning cycle by training a LSTM model

with event-level labels. Then we apply the learned LSTM model on the training data to calculate the probabilities that a patient is in the shock state at any given time point (event level). With the event-level output, we check the agreement between the event-level and the visit-level information (ICD-9) and revise the training objectives based on conditions. Then the trained LSTM model is updated using the training data with revised labels. The procedure is iteratively conducted until the model converges and no changes happen for the labels of training data. By utilizing this iterative training process, we expect the proposed framework to explore latent true labels of septic shock from combining two levels of noisy labels. In this way, the proposed framework can make accurate predictions to reflect real disease conditions.

We validate the proposed method on real-world EHRs (CCHS dataset described in Chapter 3) and compared it with multiple strong baselines. In practice there are various ways to manually adjust the data labels to approach ground truth, so we test the robustness of the framework by comparing its prediction results against three sets of ground-truth labels. Finally, we demonstrate that the proposed framework is also effective for early prediction of septic shock.

4.2 Related Work

Machine learning techniques have achieved considerable success in predicting sepsis/septic shock by using visit-level or event-level labels *but not both*. At visit level, most prior work used ICD-9 code as ground truth and considered the whole visit as one sample. For example, Brause *et al.*[Bra02] used neural networks to predict the critical states of sepsis; support vector machines (SVM) was proved to be effective at sepsis-related classification tasks [Wan10; TM10]. To truly grasp the temporal dependencies and capture the subtle progression of septic shock, it is important to predict at event level. Peelen *et al.*[Pee10] developed a dynamic Bayesian network (DBN) to model the progression of organ failure, a severe state developing into septic shock. Generally speaking, compared to works done at visit level, relatively less work was done at the event level for predicting septic shock, due to the fact that there is no well-established criteria on labelling septic shock at any time point. Some work relied on the labels manually annotated by clinicians [NH12], but such labeling process is often too expensive to be feasible for large volume of data.

There have been some studies take the intersection of ICD-9 code and event-level criteria to select positive and negative cases, e.g., septic shock patients are identified by ICD-9 code and the need for vasopressors within 24 hours of ICU transfer[Thi10]. A popular but non-specific indicator, Systemic Inflammatory Response Syndrome (SIRS), together with ICD-9 are used to screen sepsis cases[Cal16]. However, none of them combine two levels of labels to jointly learn the patterns of septic shock. Therefore, we propose to construct a general data-driven framework that leverages multiple sources of imperfect yet informative labels.

4.3 Method

4.3.1 Problem Definition

Our dataset is composed of multi-variate irregular time series data, and can be represented as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, where N is the total number of hospital visits. Each visit $\boldsymbol{x}_k$ consists of a sequence of events: $\boldsymbol{x}_k = \{\boldsymbol{x}_k^1, ..., \boldsymbol{x}_k^{T_k}\}$, where $\boldsymbol{x}_k^t$ represents the patient's event-level records at time step t in the visit $\boldsymbol{x}_k$. The value T_k denotes the number of events in the visit k and varies across different visits. For each event, we have $\boldsymbol{x}_k^t \in \mathbb{R}^D$, where D is the number of recorded measurements (i.e., features).

For each visit $\boldsymbol{x}_k$, we are provided with the event-level labels $\boldsymbol{y}_k = \{y_k^1, ..., y_k^{T_k}\}$ for every event within the visit. The label $y_k^t = 1$ indicates a patient is in the septic shock state at a given time t in the visit $\boldsymbol{x}_k$, otherwise $y_k^t = 0$. The goal of this work is to predict the $(t+1)$-th event-level label y_k^{t+1} given the clinical events from time 1 to t: $\boldsymbol{x}_k^1, \boldsymbol{x}_k^2, ..., \boldsymbol{x}_k^t$ in each visit k. For simplicity, we omit index k hereinafter when it does not cause ambiguity.

4.3.2 LSTM for Sequential Classification

Sequential data in real applications are often collected over a long span of time, such as health-care records, climate changes, traffic monitoring, etc. In our problem, the longest visit is up to three months. Therefore we utilize LSTM model in detecting septic shock for its capacity of capturing long-term sequential patterns.

We first briefly introduce the LSTM model, which is a variant of Recurrent Neural Networks (RNN). In essence, the LSTM model defines a transition relationship through an LSTM cell. The structures of LSTM and LSTM cell are shown in Fig. 4.1. Each LSTM cell contains a cell state $\boldsymbol{c}^t$, which serves as a memory and controls the information flow, added, removed or unchanged. Each LSTM cell at t also outputs a hidden representation $\boldsymbol{h}^t$, which is a high-level representation of information at current time step, and can be used for classification at t.

The cell state $\boldsymbol{c}^t$ is generated by combining $\{\boldsymbol{c}^{t-1}, \boldsymbol{h}^{t-1}\}$ and the information at t. Specifically, we first decide what information should be added to the current cell state by generating a new candidate cell state. We also generate an input gate to filter the new added information. The gating variables in LSTM cell, e.g input gate, are computed by a sigmoid function with combination of $\boldsymbol{x}^t$ and $\boldsymbol{h}^{t-1}$, while the candidate cell state is generated by a tanh(·) function, as follows:

$$\begin{aligned} \boldsymbol{i}^t &= \text{sigmoid}(\boldsymbol{W}_h^i \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^i \boldsymbol{x}^t), \\ \tilde{\boldsymbol{c}}^t &= \tanh(\boldsymbol{W}_h^c \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^c \boldsymbol{x}^t), \end{aligned} \tag{4.1}$$

where $\{\boldsymbol{W}_h^i \in \mathbb{R}^{H \times H}, \boldsymbol{W}_x^i \in \mathbb{R}^{H \times D}\}$ and $\{\boldsymbol{W}_h^c, \boldsymbol{W}_x^c\}$ denote two sets of weight parameters for gener-

ating the input gate and candidate cell state respectively. Hereinafter we omit the bias terms as they can be absorbed into weight matrices. Next, we create a forget gate $\boldsymbol{f}^t$ using a sigmoid function to remove information from the past:

$$\boldsymbol{f}^t = \text{sigmoid}(\boldsymbol{W}_h^f \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^f \boldsymbol{x}^t), \tag{4.2}$$

where $\{\boldsymbol{W}_h^f, \boldsymbol{W}_x^f\}$ denotes the weight parameters used to generate the forget gate layer $\boldsymbol{f}^t$.

In the way of forgetting old information from the old state $\boldsymbol{c}^{t-1}$ and filtering new information from the candidate cell state at time t, we obtain the new cell state $\boldsymbol{c}^t$ as follows:

$$\boldsymbol{c}^t = \boldsymbol{f}^t \otimes \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \otimes \tilde{\boldsymbol{c}}^t, \tag{4.3}$$

where $\otimes$ denotes entry-wise product.

Finally, we generate the hidden representation by filtering the the obtained new cell state by an output gate layer $\boldsymbol{o}^t$:

$$\begin{aligned} \boldsymbol{o}^t &= \text{sigmoid}(\boldsymbol{W}_h^o \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^o \boldsymbol{x}^t), \\ \boldsymbol{h}^t &= \boldsymbol{o}^t \otimes \tanh(\boldsymbol{c}^t), \end{aligned} \tag{4.4}$$

where $\{\boldsymbol{W}_h^o, \boldsymbol{W}_x^o\}$ are weight parameters that are used to generate the output gate layer. The output gate determines the information to output from $\boldsymbol{c}^t$ to $\boldsymbol{o}^t$.

With the hidden representation $\boldsymbol{o}^t$, we produce the probability of each event t at risk of septic shock using a sigmoid function with parameter U, as follows:

$$\boldsymbol{p}^t = \text{sigmoid}(\boldsymbol{U}\boldsymbol{h}^t). \tag{4.5}$$

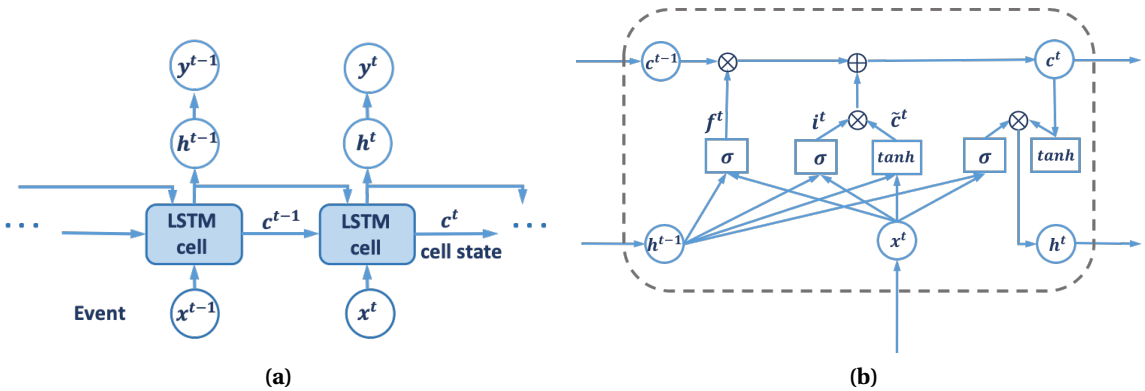


Figure 4.1 (a) The structure of LSTM for sequential classification. (b) The internal structure of a LSTM cell.

4.3.3 Learning from Multi-scale Labels

We have two levels of labels for septic shock, ICD-9 code as visit-level label and event-level labels generated by clinical criteria. Even though informative for septic shock progression, both of them are imperfect and contain false positives and false negatives. Therefore we propose to combine two sources of labels to the learning process of LSTM to jointly learn the sequential patterns of septic shock.

Basically, we reinforce the learning process by revising the provided labels with confidence. The LSTM model is firstly trained on the original event-level data. After training, the model grasps the event-level septic shock patterns and generates the probabilities of each event at the shock state, p_i^t. Then the predicted output labels, $\hat{y}_i^t$, are sampled according to p_i^t. Note that this output is different from original event-level labels in that it is generated by the model according to its learned shock patterns from training data.

With the obtained event-level output probabilities, we then check the agreement between event-level and visit-level information. Specifically, if the visit-level label y_i and the event-level output probabilities $\{p_i^t\}$ consistently indicate the same prediction results, i.e. shock or not, on some visits, event-level labels on those visits would be set accordingly to reinforce the learning objectives for the model. More formally, in a visit which is labelled as positive by ICD-9, if there exist m consecutive events at time steps $\{t_r, ..., t_r+m-1\}$, whose sum of the output probabilities is bigger than a predetermined threshold $\text{thr}_1 \times m$, this visit is identified as an agreed positive visit. Then we find the consecutive m events with largest summed probabilities, denoted by $\{t_r^*, ..., t_r^*+m-1\}$ and set the labels $\{y_i^{t_r^*}, ..., y_i^{t_r^*+m-1}\}$ as $\{1, ..., 1\}$. On the other hand, in a visit which is labelled as negative by ICD-9, if any m consecutive events, whose sum of the output probabilities is less than another predetermined threshold $\text{thr}_2 \times m$, this visit is identified as an agreed negative visit. . Then the labels of all the events in this visit are set to zero, i.e. $\{y_i^t\} = \{0\}$.

Then the LSTM model is updated using the data with revised labels by back-propagation algorithm. We repeat this process until no change is made to the labels and the model is convergent. The process of revising labels is only conducted every *epochs_update* iterations. The complete learning process is summarized in Algorithm 1.

In our implementation, we set the value of m in Algorithm 1 as two, which stands for the two consecutive events. This value is concluded from the data and confirmed by domain experts. Fig 4.2 shows the distribution of the longest sequence of consecutive ones in non-shock and shock visits from perspective of ICD-9 code respectively, where the horizontal axis is the percentile and the vertical axis indicates the length of the longest sequence of consecutive ones in a visit. It is observed that in the non-shock visits identified by ICD-9 code, there are less than ten percent of visits have one shock event. Conversely, in the shock visits, roughly 80% of visits have at least one shock event and 70% of visits have at least two consecutive shock events. In practice, clinicians usually regard

Algorithm 1: Learning process for proposed framework with LSTM.

Input:
$\{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, ..., \boldsymbol{x}_i^T\}_{i=1}^N$: A set of multi-variate sequences.
$\{y_i^1, y_i^2, ..., y_i^T\}_{i=1}^N$: Event-level labels of sequences.
$\{y_1, ..., y_N\}$: Visit-level labels of sequences.

Initialize parameters;
while *not converge & it++* $\leq$ *MaxIter* **do**
  for $i \leftarrow 1$ ***to*** N **do**
    // *In practice we use mini-batch update.*
    for $t \leftarrow 1$ ***to*** T **do**
      Compute latent output p_i^t by Eq. 4.4 and 4.5;
      Compute the predicted label $\hat{y}_i^t$;
    Update parameters for one iteration by back-propagation;
  if *mod(it,epochs_update)==0* **then**
    for $i \leftarrow 1$ ***to*** N **do**
      if $y_i = 1$ & $\exists t_r, \ni \sum_{t=t_r}^{t_r+m-1} p_i^t > thr_1 \times m$ **then**
        $t_r^* = \text{argmax}_t \sum p_i^{t:t+m-1}$
        if $\exists t \in \{t_r^*, ...t_r^* + m - 1\} \ni \hat{y}_i^t != y_i^t$ **then**
          $y_i^t = 1$
      if $y_i = 0$ & $\forall t_r, \ni \sum_{t=t_r}^{t_r+m-1} p_i^t < thr_2 \times m$ **then**
        if $\exists t \in \{1, ..., T\} \ni \hat{y}_i^t != y_i^t$ **then**
          $y_i^t = 0$

a visit as positive as long as there exists one positive event. Since the labels need to be carefully revised, it is more conservative to associate a positive visit with two consecutive positive events and no consecutive positive events would more safely indicate a negative visit.

4.4 Experiments

4.4.1 Approaches

Label Adjustment. In practice it is very hard to collect and confirm the ground-truth of septic shock. There exist multiple reasonable ways to adjust labels to approach the ground-truth, which usually generate different ground-truth sets. Following previous work [Sta14] and instructions from our clinicians, we use three ways to adjust the labels of the data:

- ***Prior***$_6$***:*** re-label all the events in the prior 6 hours of any shock event as positive; or

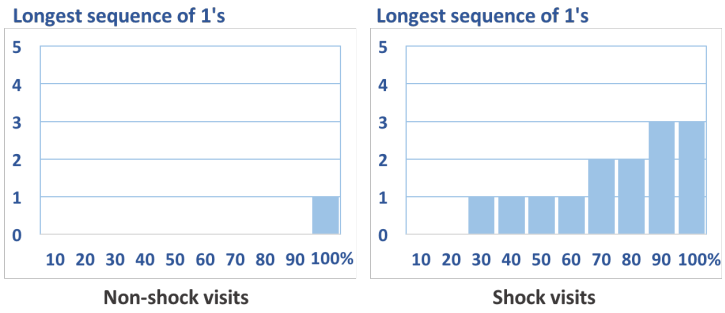


Figure 4.2 The distribution of consecutive ones in non-shock and shock visits.

- ***Post*$_8$*:*** re-label all the events in the following 8 hours of any shock event as positive; or
- ***Between*$_8$*:*** re-label all events between any two shock events as positive if their interval is less or equal to 8 hours.

The labels of the testing set are always adjusted according to one of the three ways and then used as ground-truth, since they are supposed to be highly reliable for evaluation purpose. The labels of *consistent* training set might be kept as original or adjusted with the same way with the testing data, denoted as 'train$_{adj}$'.

The effectiveness of the proposed multi-scale label learning is evaluated by comparing the performances of LSTM and LSTM$_{ms}$ (LSTM with the multi-scale label learning) on both train and train$_{adj}$, representing the consistent visits with original labels and adjusted labels respectively. In addition, we would like to explore if the inconsistent visits train$_{inc}$ can help improve the model performance since there are substantial such data in EHRs. Hence we compose two more training sets, {train, train$_{inc}$} and {train$_{adj}$, train$_{inc}$}, with involving the inconsistent visits train$_{inc}$.

4.4.2 Experimental Setup

We thoroughly explore the effect of the proposed multi-scale label learning framework in the following four subtasks. First, we compare the performance of the traditional LSTM to the LSTM$_{ms}$ that can leverage multi-scale labels. Second, to explore the generality of the proposed framework, we combine it with other popular machine learning approaches for the task of septic shock prediction and show that the framework reaches the best performance when combining with LSTM. Third, to demonstrate the robustness of the proposed framework, we compare the prediction results of the framework with LSTM using three different sets of ground-truth labels - *Prior*$_6$, *Post*$_8$, and *Between*$_8$. Finally, we conduct experiments to test the effectiveness of the proposed framework for early prediction of septic shock.

4.4.3 Evaluation and Setting

Our evaluation metrics include sensitivity/recall, specificity, precision/positive predictive value (PPV), F1-score, and area under the ROC (receiver operator characteristic) curve (AUC) [NH12]. Precision, recall, F1 and AUC are widely used for evaluating machine learning approaches. In healthcare domain, researchers commonly refer to sensitivity, specificity and PPV for the annotation performance. Therefore we include the metrics for both machine learning and healthcare domains.

All the models are implemented in Tensorflow using mini-batch stochastic optimizer and the batch size is 50. The same network settings and parameters are decided in a data-driven manner and are applied to all models, such as the number of epochs is 50 with early stopping, the learning rate is 0.01, and the number of hidden units for LSTM is 72. All the sequences are padded to have the same length with zero. We set the parameters in Algorithm 1 as follows: $epochs_update = 10$, $\text{thr}_1 = 0.7$ and $\text{thr}_2 = 0.5$.

In training process, we randomly split the *consistent* data into the training, validation and testing set in the ratio of [70%, 15%, 15%]. In different experiments, we might sample the *consistent* according to the distribution described in Chapter 3.3. Each experiment is repeated 10 times with random model initialization, and we report the average values with standard deviation for each evaluation metric.

4.5 Results

4.5.1 Overall Validation of Multi-scale Label Learning

In this experiment we use the adjusting way $\boldsymbol{Between}_8$, which is labelling all the events between two shock events which happened within 8 hours as positive events (as described in Chapter 4.4.1). The similar results are reflected when using the other two types of clinical adjustments.

Table 4.1 shows the results of comparing eight training models on our test dataset. The first four models are learned by applying either LSTM or LSTM_{ms} on the original training set or train_{adj} respectively without using inconsistent data. The last four models are learned by applying either LSTM or LSTM_{ms} on the original training set or train_{adj} using inconsistent data. Generally, the use of inconsistent data achieves a better prediction. And there exist similar patterns in the first four models and the last four models. Here we focus on the explanation for the last four models involving inconsistent data.

Overall, the combination of train_{adj} data, train_{inc} data and LSTM with the framework ({train_{adj}, train_{inc}, LSTM_{ms}}) achieves the best performance, which has the highest values of precision, AUC and F1 score. The other two evaluation metrics of this model, recall and specificity, with the highest values under the combination of {train_{adj}, LSTM_{ms}} and {train, LSTM}. However, the specificity of {train_{adj}, LSTM_{ms}} and the recall of {train, LSTM} are not competitive enough to be adopted.

Table 4.1 shows that {train$_{adj}$, LSTM} greatly outperforms {train, LSTM} (with or without using train$_{inc}$) on several important metrics including recall, AUC and F1 score. Such result is not surprising in that the LSTM learns better from the training data that is more similar to test data since both train$_{adj}$ and test data are processed with the same clinical adjustment. On the other hand, this result also shows that LSTM indeed captures the temporal dependencies that exist in the adjusted labels. Similarly, the comparison between {train, LSTM$_{ms}$}} and {train$_{adj}$, LSTM$_{ms}$} shows the latter outperforms the former (with or without using train$_{inc}$). Moreover, compared with the improvement from {train, LSTM} to {train$_{adj}$, LSTM}, relatively larger improvement was found from {train, LSTM$_{ms}$}} to {train$_{adj}$, LSTM$_{ms}$}}.

Additionally, the comparison of {train, LSTM} against {train, LSTM$_{ms}$}}}, and {train$_{adj}$, LSTM} against {train$_{adj}$, LSTM$_{ms}$}}} (with or without using train$_{inc}$) further demonstrates the effectiveness of our proposed framework. For example, when using train$_{inc}$, learning with the original training data, recall increases from 0.624 to 0.745, AUC increases from 0.707 to 0.741, and F1 increases from 0.644 to 0.687. When learning with the adjusted training data, the improvements are even more considerable.

As a brief summary, in the training process, our proposed framework can effectively update the learning objectives to approach the latent true labels of the data and then achieve a better performance. By jointly extracting useful information from two levels of labels, the model learns the discriminative patterns of septic shock and utilizes these patterns to automatically identify noisy labels.

Table 4.1 Performance (±standard deviation) LSTM with or without the multi-scale label learning on different training sets.

Data	Method	Sensitivity /Recall	Specificity	Precision	F1	AUC
train	LSTM	0.522(±0.031)	**0.778**(±0.024)	0.690(±0.018)	0.595(±0.025)	0.693(±0.022)
train$_{adj}$	LSTM	0.697(±0.023)	0.622(±0.021)	0.635(±0.020)	0.665(±0.021)	0.726(±0.020)
train	LSTM$_{ms}$	0.677(±0.019)	0.647(±0.020)	0.648(±0.018)	0.662(±0.018)	0.734(±0.016)
train$_{adj}$	LSTM$_{ms}$	**0.812**(±0.017)	0.478(±0.023)	0.599(±0.020)	0.689(±0.018)	0.753(±0.019)
train, train$_{inc}$	LSTM	0.624(±0.022)	0.700(±0.024)	0.666(±0.016)	0.644(±0.020)	0.707(±0.023)
train$_{adj}$, train$_{inc}$	LSTM	0.776(±0.013)	0.551(±0.012)	0.620(±0.010)	0.690(±0.011)	0.737(±0.016)
train, train$_{inc}$	LSTM$_{ms}$	0.745(±0.012)	0.570(±0.011)	0.638(±0.016)	0.687(±0.014)	0.741(±0.019)
train$_{adj}$, train$_{inc}$	LSTM$_{ms}$	0.781(±0.016)	0.680(±0.018)	**0.697**(±0.015)	**0.737**(±0.015)	**0.823**(±0.012)

⋆ The best number within each metric across the four models is in bold.

4.5.2 Validation on various machine learning classifiers

To explore the generality of the proposed framework, we incorporate it with other popular machine learning approaches for the task of septic shock prediction. Such popular approaches include three classic approaches: Logistic Regression (LR), Support Vector Machine (SVM), and Multi Layer Perceptron (MLP) which are non-sequential models and are ubiquitously implemented due to their efficiency and robustness, and two temporal sequential models: Recurrent Neural Networks (RNN) and Gated Recurrent Units (GRU). GRU, another variaent of RNN, has recently shown to have similar performance as LSTM, while employing a simpler architecture [Chu14], and has gained popularity recently.

For SVM baseline, we adopt RBF kernel, which is justified by the optimal performance in validation set. For MLP, we use two layers with hidden modes 40 and 20 respectively. For a fair comparison, we set RNN and GRU to share the same model settings with LSTM, e.g. the number of epochs and hidden units. The performance of these methods using the proposed framework is shown in Table 4.2. Here we use the combination of $\{\text{train}_{adj}, \text{train}_{inc}\}$ as the training data.

Table 4.2 shows that both GRU_{ms} and LSTM_{ms} obtain competitively best evaluation metrics. Generally, the ordering from low performance to high performance among these approaches is the same as the listed order in Table 4.2 except LR_{ms} outperforms SVM_{ms}. In particular, the specificity and precision of SVM are very low. As a result, their F1 scores are low for the detection task (as it is not allowed to miss too many true alarms of septic shock). This is largely because LR, SVM and MLP cannot capture the septic shock progression because both models do not take account of the temporal dependencies among the sequential events. On the other hand, the improvements made by GRU and LSTM over RNN demonstrate that incorporating the long-memory structure in the model greatly helps capture the underlying patterns of septic shock. Finally, the results of GRU and LSTM are very close with each other.

To summarize, sequential models are more appropriate for septic shock prediction since they are considering the sequential dependencies among the events, which is extremely useful for reasoning the development of septic shock. Compared with traditional RNN, GRU and LSTM are more preferable since they are capable of handling long-term dependencies and persistently keep the important long-term underlying patterns for future reference, e.g., the septic shock usually persists for a period.

4.5.3 Validation Using Different Ground-truth Labels

In practice it is very hard to collect and confirm the ground truth of septic shock. There exist multiple reasonable ways to adjust labels to approach the ground-truth, which usually generate different ground-truth sets. Here we implement our proposed framework on three different sets of ground-truth labels to testify its robustness. The three ground-truth labels of the test data are produced by

Table 4.2 Performance (±standard deviation) of multiple machine learning baselines with multi-scale label learning.

	Sensitivity/Recall	**Specificity**	**Precision**	**F1**	**AUC**
LR_{ms}	0.731(±0.026)	0.595(±0.019)	0.631(±0.023)	0.677(±0.024)	0.726 (±0.018)
SVM_{ms}	0.724(±0.019)	0.428(±0.020)	0.548(±0.021)	0.624(±0.020)	0.642(±0.015)
MLP_{ms}	0.695(±0.014)	0.675(±0.011)	0.673(±0.017)	0.684(±0.015)	0.752(±0.19)
RNN_{ms}	0.763(±0.021)	0.587(±0.016)	0.639(±0.014)	0.696(±0.017)	0.793(±0.013)
GRU_{ms}	**0.777**(±0.015)	**0.690**(±0.012)	**0.703**(±0.014)	**0.739**(±0.014)	**0.826**(±0.009)
$LSTM_{ms}$	**0.781**(±0.016)	**0.680**(±0.018)	**0.697**(±0.015)	**0.737**(±0.015)	**0.823**(±0.012)

⋆ The best number within each metric across the five approaches is in bold.

following the consultation of clinicians to adjust the labels as described in section 4.4.1, which are $Prior_6$, $Post_8$ and $Between_8$. Here we use the combination of $\{\text{train}_{\text{adj}}, \text{train}_{\text{inc}}\}$ as the training data.

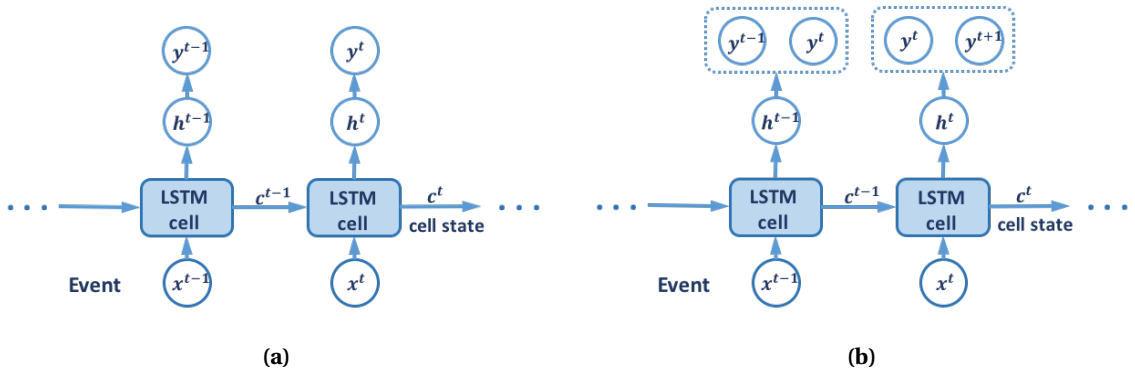
Since the settings $\{\text{train}_{adj}, \text{train}_{inc}, LSTM_{ms}\}$ lead to the best performance according to Table 4.1, in this experiment we focus on using $LSTM_{ms}$ and measure its performance with three different types of adjusted labels for both training labels and test labels. The results of the proposed framework on three groups of ground-truth labels are listed in Table 4.3. It can be observed that our proposed $LSTM_{ms}$ works well on all three types of clinical adjustment. More specifically, the values of recall and AUC of the $Prior_6$ and all of the metrics of $Post_8$ are as good as those of $Between_8$. On the other hand, $Prior_6$ achieves even higher AUC than $Between_8$. The specificity of $Prior_6$ is a little bit lower, but it dose have a comparable F1 score. Among three groups, the adjustment on $Between_8$ is most conservative since it labels positive events only when two positive events happen within eight hours. The results confirm that on all three groups of ground-truth labels, our proposed framework can timely and effectively learn the progression of septic shock and also discover latent true labels even from imperfect labels.

Table 4.3 Performance (±standard deviation) of the framework on testing data with different ground-truth labels.

Method	**Sensitivity/Recall**	**Specificity**	**Precision**	**AUC**	**F1**
$Prior_6$	0.832(±0.013)	0.528(±0.014)	0.663(±0.017)	0.738(±0.015)	0.831(±0.010)
$Post_8$	0.773(±0.011)	0.653(±0.015)	0.683(±0.010)	0.725(±0.010)	0.814(±0.009)
$Between_8$	0.781(±0.016)	0.680(±0.018)	0.697(±0.015)	0.737(±0.015)	0.823(±0.012)

4.5.4 Early Prediction

We explore the effectiveness of the proposed framework for early prediction, i.e., to detect the patients who are developing septic shock in advance. To do so, we utilize the future diagnostic condition as labels for each current event in training process. In specific, we shift the labels of prospective events to current events, as shown in Fig. 4.3. For example, in previous setting, we use $\{\boldsymbol{x}_k^i\}_{i=1}^t$ to predict the $(t+1)$-th event-level label y_k^{t+1}, one event in advance. In early prediction setting, we use $\{\boldsymbol{x}_k^i\}_{i=1}^t$ to predict $y^{t+\eta}$, η events in advance. Training data and testing data are processed with the same shifting procedure.



(a) **(b)**

Figure 4.3 (a) The Original structure of LSTM. (b) The revised structure of LSTM for early prediction.

For early prediction, it is extremely important to capture the positive class, such as the label pair of $\{0,1\}$ and $\{1,1\}$. Our data is heavily skewed with large volume of $\{0,0\}$, which accounts for more than 60% of the data. To fairly reflect the performance of our framework on septic shock detection, we keep all positive events and randomly sample similar number of negative events. Here we show the performance using F1 and AUC, as F1 score computes a balanced measure for positive class, and AUC reflects the relationship between false positives and true positives.

We conduct experiments on the baseline model LSTM and the proposed LSTM$_{ms}$. The training set of $\{\text{train}_{adj}, \text{train}_{inc}\}$ and the adjusting method of ***Between***$_8$ are applied. Since our data is aggregated on an hourly basis, η-event early guarantees at least η-hour early. Most of prior works report the performance up to 24-hour early for the task of septic shock prediction [Lin18]. The variation of F1-score and AUC with shifting labels are shown in Fig. 4.4. Note that the starting point '1' stands for one-event early prediction, which is equivalent to the results in overall prediction experiment.

Overall, LSTM$_{ms}$ is always better than the traditional LSTM, and both models have their performance gradually degraded when predicting future event labels. However, the integration of the

framework can mitigate the declining trends. It is obvious LSTM$_{ms}$ is more stable and decreases much slower in terms of both AUC and F1 metrics. While the mapping relation from features to prospective labels is difficult to learn, the temporal progression patterns can still be helpful for a reasonable prediction. By inferring true labels using our proposed framework, the LSTM model is more likely to capture the temporal septic shock progression using the revised event-level labels. The results indicate that the proposed multi-scale label learning can effectively and stably detect septic shock at an early stage with high performance.

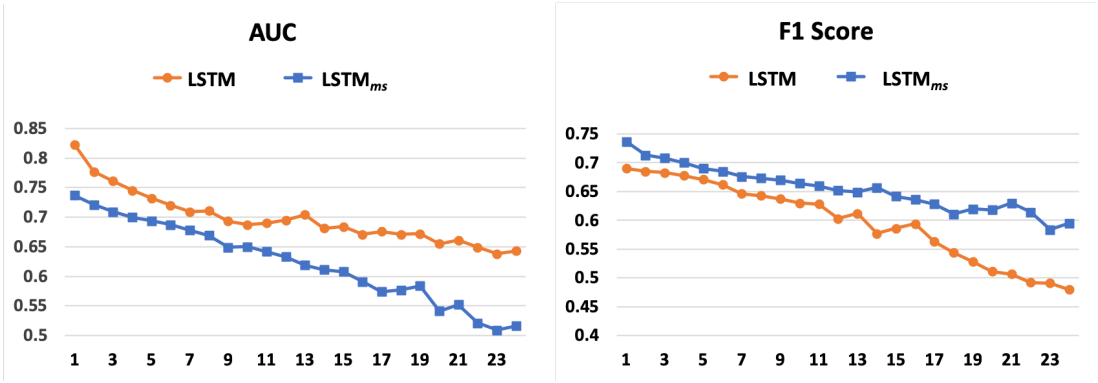


Figure 4.4 (a) AUC of the two models with shifting labels. (b) F1 of the two models with shifting labels.

4.6 Summary

In thrust 1, we focus on two major challenges associated with septic shock prediction: 1) how to capture the informative progression of septic shock in a long visit; and 2) how to obtain reliable predictions without well-established moment-by-moment ground-truth labels for septic shock.

To solve these two problems, we proposed a generic framework, multi-scale label learning, to predict septic shock based on Long-Short Term Memory (LSTM) model, which is capable of memorizing temporal dependencies over a long period. The framework integrates two unreliable sources of labels to reliably learn the discriminative patterns of septic shock: ICD-9 code as the visit-level label and the clinical criteria designed by domain experts as the moment-by-moment event-level label. The experimental results demonstrate the superiority of the proposed framework and the effectiveness of LSTM compared with multiple baselines. Its robustness is also validated on the test data with three different ground-truth labels. Besides, the proposed method can not only detect patients in the shock state, but also stably predict patients who are trending septic shock.

CHAPTER

5

THRUST 2: TIME-AWARE ATTENTION NETWORKS

5.1 Overview

The standard LSTM cannot provide an interpretation for prediction results. Also, it does not take into account irregular time intervals between consecutive events.

Interpretability of any computational models is extremely critical in healthcare-related domains. It is generally more important to learn about the discriminative *interpretable* patterns which capture informative progression of a disease than to induce an accurate predictive computational model. As a result, interpretability is a notable factor when we employ computational models to real hospital settings. Various attention-based neural networks are widely developed to generate interpretations for EHRs. [Cho16b; Cho17; Ma17; SW17]. As frontier work, RETAIN [Cho16b] applied a two-level attention mechanism to identify meaningful visits and specific features that contribute to the prediction.

Measurements in EHRs are commonly acquired with *irregular intervals*. For example, during a patient's stay at hospital, body temperature is often measured several times a day while white blood cells are only measured every other day. Hence the time intervals between events are also irregular as collections of multiple measurements. More importantly, the varying time intervals can reveal patient's health status on certain impending conditions. Intuitively, when a patient is

in a severe condition, events are likely to be recorded more frequently than when a patient is in a relatively “healthier" condition. Furthermore, it is important to consider the time intervals between temporal events to capture latent progressive patterns of a disease. For example, one criteria for defining the disease of septic shock is that two consecutive readings of systolic blood pressure are below 90mmHg and here consecutive readings have to happen within *8 hours*. There have been several previous works on handling the time irregularity [Bay17; Pha16; Cho16a; Che17]. Among them, Time-aware LSTM (T-LSTM) [Bay17] transforms time intervals into weights and uses them to adjust the memory content passed from previous moments.

In this thrust, we propose **ATTAIN**, an **at**tention-based **t**ime-aware dise**a**se progress**i**o**n** model, that incorporates the attention mechanism and models the time irregularity between events. Specifically, we adjust the memory content of LSTM when accumulating previous information. Instead of adding memory from one previous event, we retrospect memories of all/several previous events and discount them by weights generated from attention mechanism and the time intervals between those events and current event. The overall weights represent how important each previous event is for the current event to identify the progressing condition. Three attention mechanisms are explored: *global* ($\boldsymbol{g}$), *local* ($\boldsymbol{l}$), and *flexible* ($\boldsymbol{f}$), to calculate the attention weights. On the other hand, the time intervals are transformed to decay weights through a decay function so that the outdated events are more likely to play a less important role than recent events for predicting the outcome of the current event. We believe the obtained attention weights together with the decay weights would not only lead to an effective predictive model but also result in an interpretable and clinically reasonable model.

We validate the proposed ATTAIN on the task of early prediction of septic shock. Due to delicate progressions of septic shock (described in Chapter 4.1), variables in the before-shock stage may either be measured infrequently or not measured at all. Therefore, it is quite critical and challenging to identify previous indicative moments and give accurate predictions of patients' health status.

Our experimental results on real-world EHRs (CCHS dataset described in Chapter 3) show that ATTAIN outperforms the state-of-the-art models such as RETAIN and T-LSTM. Also, the generated interpretative time-aware attention weights shed some lights on the progression behaviors of septic shock.

5.2 Related Work

In recent years, attention mechanisms are extensively explored to interpret the model output. RETAIN [Cho16b] applied a reverse time attention mechanism in an RNN to identify meaningful visits and qualified features. GRAM [Cho17] employed a graph-based attention model for healthcare representation learning. These attention-based predictive models can also improve the prediction performance. Another challenge associated with EHR data, time irregularity, has also been tackled.

T-LSTM [Bay17] divides the previous cell memory as short-term and long-term, then introduces a time-aware mechanism to adjust the short-term memory while keeping the original long-term memory. In [Pha16], the time intervals are used to modify the forget gate to adjust the information passed from previous moments. Another way to handle the time irregularity is to impute the data by sampling new records to make regular time gaps between two consecutive events [Che18]. Finally, Health-ATM [Ma18] firstly extracts patient information patterns with attentive and time-aware models through a network composed of RNN and Convolutional Neural Networks (CNN). It integrated time stamps to calculate the weights of input in the convolution layer.

Compared with the previous works, our proposed approaches explore different attention mechanisms to generate weights for the past events while dealing with irregular time intervals in EHRs. For acute medical conditions such as septic shock, it is extremely significant to identify critical and timely meaningful events.

5.3 Method

In this thrust we adopt the same problem definition as stated in Chapter 4.3.1. The structure of the standard LSTM has been described in Chapter 4.3.2. Here we briefly describe the proposed time-aware attention networks, ATTAIN.

Fig. 5.1 shows the architecture of proposed ATTAIN framework. In the original LSTM, the cell state of a current event $\boldsymbol{c}^t$ obtains memory from its most recent event i.e., $\boldsymbol{c}^{t-1}$ as shown in Eq. 4.3. Due to the well-known vanishing gradient problem [Bay17], the learning of each cell state still heavily depends on most recent events. However, in disease progression, relationships of events are non-trivial in that the current condition is often not impacted only by recent moments. Single memory-reading also brings that it cannot interpret how critical of each past event. Conversely in practice, doctors usually review the patient's records to identify the critical previous events.

Here we propose to use attention mechanism to collect multiple previous memories and explore the relationships of events in patient trajectories. We expect it can help doctors make prompt decisions and achieve a better prediction.

The standard LSTM implicitly assumes that intervals between time steps, or events, in a sequence are uniformly distributed, whereas the frequency of collecting patients' records often vary greatly based on many factors. The interval between two consecutive events during a patient's visit can span from minutes to several weeks. In general, the events that occur long time ago tend to have less impact to the current event and thus we should properly reduce their contributions. Therefore, it is important to consider the elapsed time when predicting the current event's output. In addition, these irregular time intervals might be indicative for patients' health status on certain impending conditions. For example, when a patient is in a more severe condition, events are likely to be recorded more frequently than when a patient is in a relatively "healthier" condition.

To incorporate the elapsed time into the standard LSTM, we feed the intervals into a monotonic decreasing function. In specific, for each event, the memories collected from its previous events are discounted by a decay factor, which is estimated based on the time intervals between previous events and the current event. In this way, if the previous event occurs long time ago, the dependency from the previous moment would not play an active role for predicting the current event's output.

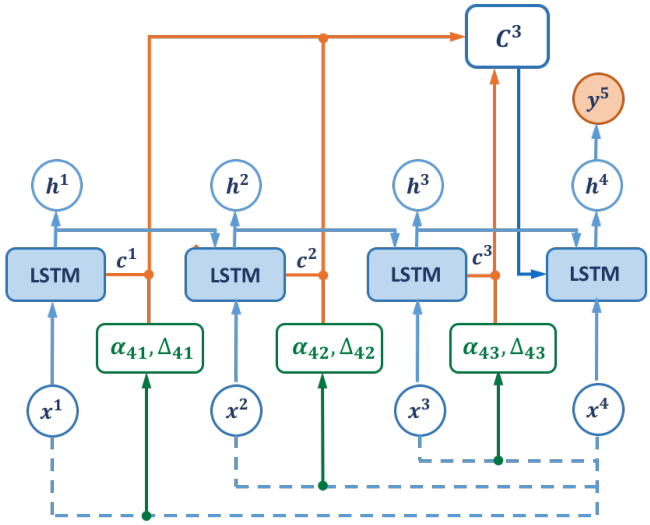


Figure 5.1 Illustration of proposed ATTAIN model. ATTAIN accumulates the cell states from all the previous events and regularize the old memories through two sources of weights generated from attention mechanism ($\boldsymbol{\alpha}$) and time intervals ($\boldsymbol{\Delta}$).

As shown in Fig. 5.1, in the proposed ATTAIN, instead of reading the information from just one previous cell state, we collect all/some previous cell states and regularize these old memories by weights generated by attention mechanism and time intervals as follows:

$$\boldsymbol{C}^{t-1} = \sum_{i=t-m}^{t-1} \alpha_{ti} \cdot \boldsymbol{c}^i \cdot g(\Delta t_{ti}), \tag{5.1}$$

where α_{ti} is the attention weight from i-th event to the current event t, Δt_{ti} is the time interval between i-th event to the current event, $g(\cdot)$ is a decay function, and m stands for the number of events to look backwards. Then we replace $\boldsymbol{c}^{t-1}$ in Eq. 4.3 with $\boldsymbol{C}^{t-1}$ for generating the new cell state.

5.3.1 Three Attention Mechanisms.

We explore three ways for determining m to generate attention weights:

- *Global Attention* generates attention weights α_{ti} from *all* the previous events for the current time step t, i.e, $m = t-1$ in Eq. 5.1.

$$\begin{aligned} e_{ti} &= (\boldsymbol{x}^i)^\top \boldsymbol{W}_\alpha \boldsymbol{x}^t, \text{for } i = 1, ..., t-1 \\ \boldsymbol{\alpha}_t &= \text{softmax}(e_{t1}, e_{t2}, ..., e_{t(t-1)}) \end{aligned} \tag{5.2}$$

  The advantage of the global attention is that it considers all past events and its disadvantage is that it is computationally expensive and attention weights are sparsely distributed. To address these problems, we also explore the following two options.

- *Local Attention* only considers a *fixed* number of previous events, m in Eq. 5.1. The hyperparameter m can be optimized by grid search or suggested by clinicians.

$$\begin{aligned} e_{ti} &= (\boldsymbol{x}^i)^\top \boldsymbol{W}_\alpha \boldsymbol{x}^t, \text{for } i = t-m, ..., t-1 \\ \boldsymbol{\alpha}_t &= \text{softmax}(e_{t(t-m)}, e_{t(t-m+1)}, ..., e_{t(t-1)}) \end{aligned} \tag{5.3}$$

- *Flexible Attention* learns the optimal number of previous events adaptively for different given events, i.e, m varies for different time step t. The intuition is that different number of previous events should be considered for different given events.

$$\begin{aligned} m(t) &= \lceil (t-1) \cdot \text{sigmoid}(\boldsymbol{v}_m^\top \tanh(\boldsymbol{W}_m \boldsymbol{x}^t)) \rceil \\ \boldsymbol{\alpha}_t &= \text{softmax}(e_{t(t-m(t))}, e_{t(t-m(t)+1)}, ..., e_{t(t-1)}), \end{aligned} \tag{5.4}$$

  where $\boldsymbol{W}_m$ and $\boldsymbol{v}_m$ are parameters to be learned, and $m(t) \in [1, t-1]$.

Previously RETAIN [Cho16b] introduced attention mechanism in learning health record representation. RETAIN generated attention weights from the hidden states of RNN: $e_{ti} = \boldsymbol{W}_\alpha^\top \boldsymbol{h}^i, i \in [1, t-1]$. However, it does not capture the relation between the targeting event t and past events is. Different with RETAIN, the aforementioned three attention mechanisms incorporate this relation when estimating attention weights. Additionally, when aggregating previous information (see Eq. 5.1), α_{ti} measures the weight of event i for the current event t and hence its computation should solely depend on the data at events i and t. Therefore we use the original inputs $\boldsymbol{x}$ (i.e. $\boldsymbol{x}^t$ and $\boldsymbol{x}^i$) instead of hidden states $\boldsymbol{h}$ (i.e. $\boldsymbol{h}^t$ and $\boldsymbol{h}^i$) since the pair $[\boldsymbol{x}_t, \boldsymbol{x}_i]$ extracts immediate relations among the medical variables of two events while $[\boldsymbol{h}_t, \boldsymbol{h}_i]$ encodes the information from multiple events.

5.3.2 Time-aware Decay Function

Time intervals might range from minutes, hours and even to days in healthcare domain. We integrate the intervals into the model and make LSTM predict in a time-aware way. As a guideline, the function

in Eq. 5.5 is validated and suggested by several prior works for long time periods in EHR data [Pha16; Bay17].

$$g(\Delta t) = 1/\log(e + \Delta t) \tag{5.5}$$

Previously T-LSTM [Bay17] also integrated the elapsed time to adjust the cell memory. It divided the previous cell memory as short-term ($\boldsymbol{c}_S^{t-1}$) and long-term ($\boldsymbol{c}_L^{t-1}$), then discounted only the short-term memory through the same decay function, i.e., $\boldsymbol{c}^t = \boldsymbol{f}^t \otimes (\boldsymbol{c}_L^{t-1} + \boldsymbol{c}_S^{t-1} \cdot g(\Delta t)) + \boldsymbol{i}^t \otimes \tilde{\boldsymbol{c}}^t$. T-LSTM depends on only one previous cell state to adjust the memory and is unable to interpret how important each event is for predicting the current outcome.

In summary, through the aforementioned memory adjustment, ATTAIN would identify the critical previous moments and take the effect of elapsed time into account. The two sources of weights, based on attention mechanism and time intervals, are simultaneously tuned together. The intuitions behind ATTAIN can be seen as mimicking doctors reading records for diagnosing in the real world in that they generally pay attention to both very important indicative previous events even though they happened long time ago but also the most recent events. As a result, we expect ATTAIN can lead to more reasonable predictions than the standard LSTM.

5.4 Experiments

5.4.1 Approaches

Baseline Approaches. We compare ATTAIN with three baselines: 1) LSTM without either attention or time-aware mechanisms; 2) RETAIN [Cho16b] with a two-level attention mechanism at both the event-level and the variable level; 3) T-LSTM [Bay17] with time-aware mechanism only.

Our Approaches. We integrate three attention mechanisms with the standard LSTM: LSTM$_g$, LSTM$_l$, LSTM$_f$. The attention weights are generated by Eq. 5.2 - 5.4 respectively. Old memory is obtained by accumulating all or selective previous cell states:

$$\boldsymbol{C}^{t-1} = \sum_{i=t-m}^{t-1} \alpha_{ti} \cdot \boldsymbol{c}^i. \tag{5.6}$$

Different from Eq. 5.1, it does not add the decay function.

Specifically, for local attention, we find the optimal setting of m by grid search on a validation set. Fig. 5.2a shows the model achieves competitive performance using tenth-interval $m \in \{20, 30, 40\}$. We hence further explore the performance for $m \in [20, 40]$ incremented by one and Fig. 5.2b shows that the optimal values are $m \in \{22, 26\}$. Here we use $m = 24$.

Furthermore, we add the time-aware mechanism to the three models above resulted in three ATTAIN models named ATTAIN$_g$, ATTAIN$_l$, ATTAIN$_f$ respectively. The $\boldsymbol{C}^{t-1}$ is calculated based on Eq. 5.1 in a way that the old memory is regularized by both attention weights and decay weights.

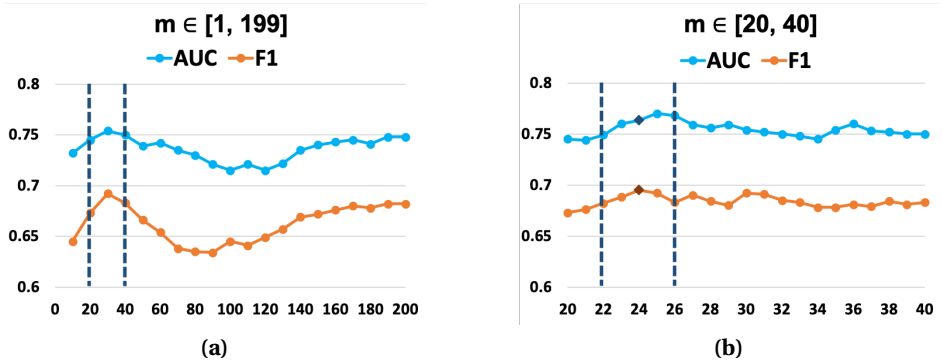


Figure 5.2 AUC and F1 score of LSTM$_l$ when: (a) m is taken from $[1,199]$ with interval 10. (b) $m \in [20,40]$.

5.4.2 Experimental Setup

Only *consistent* visits from CCHS dataset (described in Chapter 3.3) are used in the following experiments of this thrust. We adopt the same evaluation metrics and experimental settings as discussed in Chapter 4.4.3 for thrust 1.

We first compare the proposed ATTAIN with multiple baseline models without either or both attention-based mechanism and time-aware mechanism. Second, we conduct early-prediction experiments to test the capability of the proposed framework to detect patients who are progressing septic shock. Finally, we perform a case study to illustrate the interpretative time-aware attention weights generated from ATTAIN and how they are used for interpreting prediction results.

5.5 Results

5.5.1 Results of Overall Prediction

Table 5.1 shows the performance of all the models and from top to bottom, we have three baseline models, our proposed three attention mechanisms with LSTM, and the three corresponding ATTAIN models. Columns 'A' and 'T' indicate whether attention or time-aware mechanisms are used in the models. For each sub-session, the best results are marked in bold and the best model across all methods are labeled with $\star$.

Among the three baselines, both RETAIN and T-LSTM outperform the standard LSTM on almost all the evaluations, which shows that either the attention mechanism or time-aware mechanism can lead to a better prediction of the disease progression. As in the original work [Cho16b], RETAIN does not boost from LSTM much and this might come from two reasons. First, their attention mechanism did not incorporate the relation between the previous events and the current event. Second, due to high missing rate of features, the variable-level attention did not take into full effect.

All three attention-based models (in the middle sub-session) outperform three baselines across all evaluations with only one exception that LSTM_g has worse recall than T-LSTM. When comparing the three attention mechanisms, LSTM_l performs more stably than LSTM_g and LSTM_f, i.e., 0.695 vs. 0.682 and 0.695 (larger std) for F1 score and 0.763 vs. 0.748 and 0.755 for AUC. LSTM_g highlights on specificity and precision but the recall is much lower than the other two models. We found that the attention weights of LSTM_g are sparsely distributed in long sequences and cannot effectively detect critical moments. The result of LSTM_l validates that $m = 24$ is clinically reasonable. When only using attention mechanism, though close, LSTM_f, deciding how many previous events to pay attention in a data-driven manner, is not as competitive as LSTM_l.

Except that ATTAIN_g has a worse recall than T-LSTM, all three ATTAIN models (in the bottom sub-session) outperform three baselines with obvious improvement. Moreover, when comparing between ATTAIN and our three attention-based models, ATTAIN_g outperforms LSTM_g, and the same observation on ATTAIN_l over LSTM_l and ATTAIN_f over LSTM_f. Thus, when integrating both attention and time-aware mechanisms, the two sources of weights allow the model to comprehensively identify the meaningful previous events in terms of medical relation and time effect. For example, if attention weights indicate two previous events are equally important, we pay attention to the recent event first, exactly as doctors perform in real world.

Overall, ATTAIN_f achieves the best F1-score and AUC. ATTAIN_l is close to ATTAIN_f, i.e. 0.718 vs. 0.720 on F1 and 0.804 vs. 0.811 on AUC and gets the best recall. Since our data are aggregated at hourly rate, 24-event retrospect guarantees attending the last 24 hours of a patient status. In practice the clinicians prefer this local-based model because previous information can be sufficiently obtained with certainty.

Table 5.1 Performance (±standard deviation) of baselines and our approaches on septic shock overall prediction.

Method	A	T	Sensitivity /Recall	Specificity	Precision	F1	AUC
LSTM	-	-	0.627(±0.023)	0.632(±0.021)	0.635(±0.020)	0.631(±0.021)	0.716(±0.020)
RETAIN	✓	-	0.618(±0.015)	0.654(±0.016)	0.651(±0.016)	0.634(±0.016)	0.732(±0.010)
T-LSTM	-	✓	**0.643**(±0.009)	**0.680**(±0.012)	**0.702**(±0.013)	**0.671**(±0.010)	**0.745**(±0.013)
LSTM_g	✓	-	0.628(±0.013)	**0.798**(±0.015)	**0.747**(±0.018)	0.682(±0.016)	0.748(±0.014)
LSTM_l	✓	-	**0.684**(±0.007)	0.742(±0.013)	0.707(±0.012)	**0.695**(±0.011)	**0.763**(±0.008)
LSTM_f	✓	-	0.667(±0.022)	0.731(±0.017)	0.726(±0.016)	0.695(±0.019)	0.755(±0.016)
ATTAIN_g	✓	✓	0.636(±0.016)	⋆ **0.818**(±0.008)	⋆ **0.803**(±0.010)	0.710(±0.014)	0.782(±0.015)
ATTAIN_l	✓	✓	⋆ **0.695**(±0.014)	0.746(±0.012)	0.744(±0.015)	0.718(±0.014)	0.804(±0.010)
ATTAIN_f	✓	✓	0.686(±0.018)	0.767(±0.016)	0.758(±0.016)	⋆ **0.720**(±0.017)	⋆ **0.811**(±0.011)

· The best values of each metric within the session are in bold.
· The best values of each metric across all the models are labeled with ⋆.

5.5.2 Results of Early Prediction

Here we investigate the effectiveness of the proposed framework for early prediction, i.e. to detect the patients who are developing septic shock in advance. The method to perform early prediction using shifting labels is described in Chapter 4.5.4.

We conduct experiments on three baseline models: LSTM, RETAIN and T-LSTM respectively, the best attention-based model LSTM_l, and our best approach ATTAIN_f. We track the performance of each model on F1-score and AUC from one-event early up to 24-event early. Since our data is aggregated on an hourly basis, η-event early guarantees at least η-hour early. Most of prior works report the performance up to 24-hour early for the task of septic shock prediction [Lin18]. The variation of F1-score and AUC with shifting labels are shown in Fig. 5.3. Note that the starting point '1' stands for one-event early prediction, which is equivalent to the result of overall prediction in last experiment.

As expected, as η increases, the early prediction tasks become more and more challenging in that the performance of all the models has gradually degraded. Despite this, ATTAIN_f has been decreasing much slower and steadily staying on the top of all other models in terms of both AUC and F1.

Attention models RETAIN and LSTM_l seem more durable before eighth event, while time-aware model T-LSTM drops faster in the first few events and tends to be stable later. When taking advantages from both sides in different periods, ATTAIN captures the developing process of septic shock and gives a stable prediction up to 24-event early future.

To summarize, the results demonstrate that the proposed ATTAIN can effectively and stably detect septic shock at an early stage with high performance.

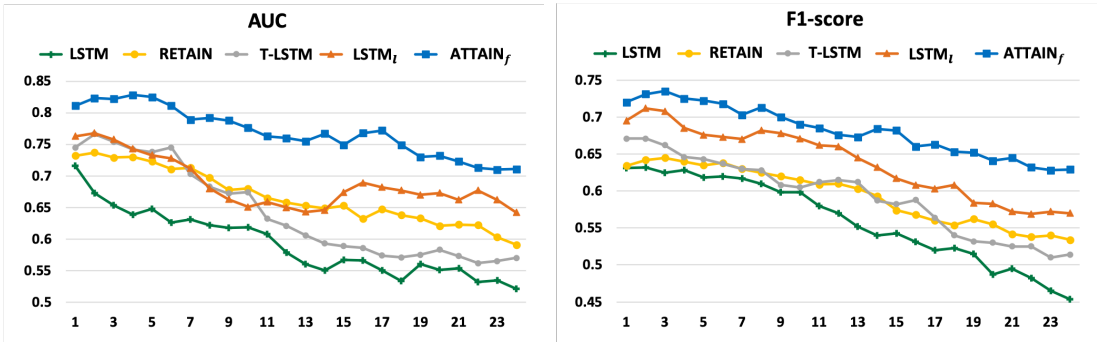


Figure 5.3 (a) AUC of the early prediction on five models. (b) F1-score of the early prediction on five models.

5.5.3 Case Study

To illustrate the interpretability of our proposed method, Fig. 5.4 shows a case study by comparing the weights learned from RETAIN, LSTM$_l$ and ATTAIN$_l$ for a sample patient's visit. To simplify, we fix $m = 10$ for all three models and the weights are normalized. The patient is in septic shock at events 11 and 12. For event 11, we expect the models can identify some critical previous event(s). To quantify it we consider *three* prior events with the highest weights to be *critical events*. At event 12, septic shock continues, we expect a reliable model should: a) identifies event 11 to be critical for event 12, because the former clearly explains why the patient was still in septic shock at event 12, and b) identifies the same critical events (other than event 11) as those for event 11 for event 12.

Fig. 5.4 shows that RETAIN indeed identifies that event 11 is critical for event 12, but it discovers different sets of critical events for the two events. LSTM$_l$ consistently detects the same set of critical events, $\{4,5,8\}$, for both events 11 and 12, but it fails to identify that event 11 is critical to event 12. Finally, our ATTAIN model successfully identify that event 11 is important and besides event 11, event 12 has the same critical set: $\{4,5,8\}$ as event 11.

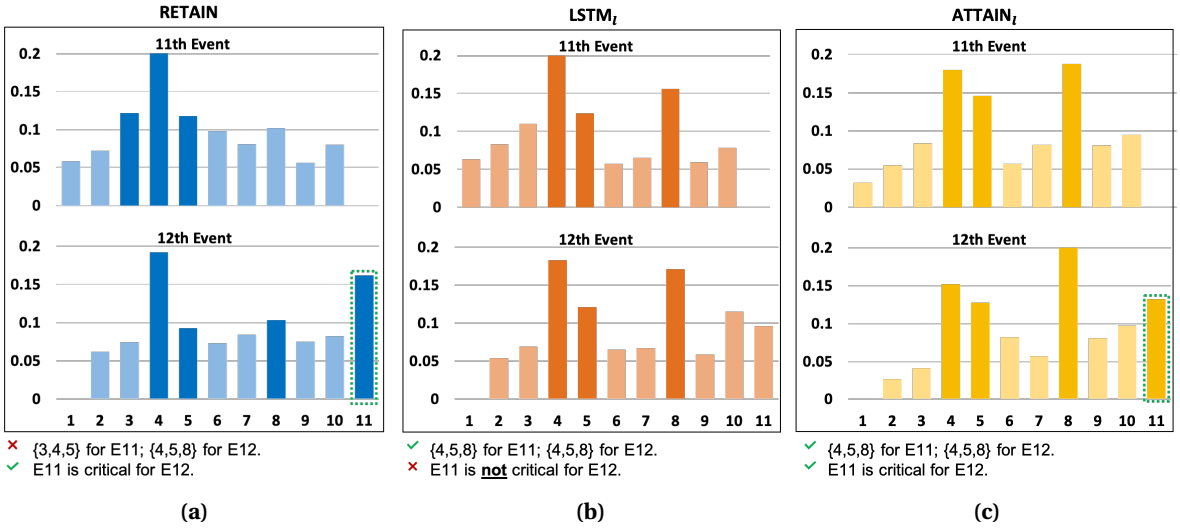


Figure 5.4 Attention weights for the 11th and 12th events achieved from (a) RETAIN; (b) LSTM$_l$; (c) ATTAIN$_l$.

5.6 Summary

In thrust 2, we propose ATTAIN, an attention-based time-aware LSTM networks to effectively improve the interpretability of LSTM while also modeling the irregular time intervals. In specific, ATTAIN employs attention weights combining with time decay function to identify the contributions of the historical events to the current event. The experimental results demonstrate the effectiveness of ATTAIN compared with state-of-the-art baselines. The case analysis shows that ATTAIN achieves an interpretable and reasonable prediction for modeling septic shock. One limitation of this work is it does not generate feature-level attentions at each event, since the feature-level attentions can vary drastically over multiple events in acute disease conditions, especially given the impact of high missing rate in real-world EHRs. We keep this challenging task as our future work.

CHAPTER

6

THRUST 3: DEEP DOMAIN ADAPTATION

6.1 Overview

LSTM-based approaches for disease progression modeling mainly focus on on discovering statistical relationships that match the distribution of training data without considering the intricate heterogeneity among different patient groups. Specifically, the heterogeneity of EHRs can be reflected in two aspects. First, the distributions of input measurements or features often differ among different patient groups [AW04; Lao04; SG04]. For example, the blood pressure of elderly people is generally much higher than that of young people [Blo]. The difference can be found on various medical measurements across different races, age levels and gender groups in our EHRs. Second, some patient groups are more dominant in the number of observations in EHRs than other groups (e.g., septic shock is more common among old adults than young adults). This leads to the bias of the learned model towards to dominant groups and thus can be inferior to minor groups.

We propose a general domain adaptation (DA) framework that takes both aspects of heterogeneity into consideration. The patient groups partitioned by specific factors, such as race, age, or gender, are referred to as *domains*. The framework aims to learn an invariant feature representation across different domains through an adversarial learning process. Specifically, our framework consists of two LSTM-based recurrent components: 1) a *label predictor* that learns the representation that is indicative of the discriminative temporal developing patterns of the disease; 2) a *domain classifier* that identifies the temporal changes of feature distributions across patient groups. Then the learned

representation becomes invariant across domains by adversely degrading the performance of the domain classifier. Compared with existing adversarial domain adaptation methods [Gan16; Pur16], our proposed framework assumes the shift between feature distributions of different domains is *temporal*. For example, it is quite likely that the young and elderly people share a quite similar feature distribution at a certain stage of the disease progression, but develop their exclusive distributions with time evolving. Modeling such temporal changes enables better revealing the relationships among different patient groups and hence improve the effect of domain adaptation process.

Given the temporal nature of both components, we refer to our proposed framework as Temporal Adversarial Networks in the following. We expect the duplex recurrent structure enables capturing complex temporal dependencies in healthcare data and generalizing well across different domains in spite of heterogeneity.

We validate the proposed ATTAIN on the task of early prediction of septic shock. There exists severe heterogeneity in sepsis-related cohort. For example, White American and elderly people are dominant among septic shock patients. It is very likely that the learned model performs poorly on other races and younger groups. Therefore, it requires development of new methods to overcome these challenges and deliver a specific diagnosis of patients' health status in delicate progressions of septic shock.

6.2 Related Work

Domain adaptation (DA) aims to leverage abundant labeled data from a source domain to learn a discriminative classifier and then generalize it to a target domain despite the data distribution discrepancy between the source and target domains. Existing DA techniques can be divided into two categories - semi-supervised DA [DI09] and unsupervised DA [GL14; Gop11]. Semi-supervised DA assumes the availability of certain amount of labeled data in the target domain, while there exists no labeled data in the target domain for the unsupervised DA. In this work, we focus on the unsupervised DA, although the proposed method can be generalized to the semi-supervised case.

DA techniques have shown success in a variety of applications, including computer vision, text mining [Luo18; JZ07; Kan18; Gon12]. Many existing works [Lon15; Hou16; MHH15] learn transferable feature by minimizing Maximum Mean Discrepancies (MMD) between source domain and target domain, which measures the difference in both marginal and conditional distributions [Gre07]. With the recent advances in deep adversarial learning, researchers propose an alternative approach to learn consistent features across two domains by minimizing the classification accuracy of a separate domain classifier [Gan16]. Such adversarial training ensures that the extracted features from two domains fall into similar distributions which cannot be easily distinguished by the domain classifier. The adversarial training-based DA can be easily combined with deep learning models in that it integrates DA into the process of learning representation, so that the extracted features are both

discriminative and invariant to the change of domains [Gan16; Tze17].

DA has been popular to tackle challenging problems associated with EHRs learning [Alv18; Pur16], such as imbalanced data [Bha17], shift of distributions in different groups of patients [Nor17] and heterogeneous data sources [Gha15]. In health-care applications, the temporal patterns from long-term EHR data are essential for capturing the disease progression. Therefore, it requires that the DA technique can adapt the learned temporal patterns to sequential data for other patient groups. While most DA techniques focus on static data, recent works have demonstrated the effectiveness of DA for temporal data [HY09; Soc11; Pur16; Jae16]. For example, Che *et al.* [Pur16; Che18] uses adversarial regularization when training a recurrent sequential model. We are originally motivated by their approach of domain adaptation with adversarial learning.

Real-world EHR data commonly cover a long span of time. When predicting a specific target disease, it is critical to automatically discover the importance of different time periods in classification. To this end, researchers utilize attention structure on top of recurrent neural networks, which have shown superior performance on real-world sequential data (see Chapter 5). There also exist research works that combine attention mechanism with DA techniques [Has16; Kan18; Li17]. For example, in [Kan18], the authors claim that the performance of attention structure can be degraded when adapted to the target domain. To solve this issue, they add a constraint to regularize the shift of attention weights in adversarial training. In [Li17], an end-to-end model is proposed which automatically captures pivots by attention mechanism and also leverages domain adversarial training for cross-domain sentiment classification.

Different from prior works using non-temporal models as domain classifiers, our proposed Temporal Adversarial Networks is able to model the temporal changes of feature distributions within different patient groups. For acute clinical conditions such as septic shock, it is extremely significant to capture the timely changes within each patient group and give a more specific prediction.

6.3 Method

In this thrust we adopt the same problem definition as stated in Chapter 4.3.1. The structure of the standard LSTM has been described in Chapter 4.3.2. Here we briefly describe the process of domain adaptation with adversarial learning and our proposed Temporal Adversarial Networks.

6.3.1 Domain Adaptation with Adversarial Learning

Domain adaptation (DA) refers to the problem of leveraging labeled data in a source domain and unlabeled data in a target domain to jointly learn an accurate model for the target domain where labels are scarce or unavailable. Here each domain in our problem can be instantiated as a set of EHR samples that are taken from a certain group of patients. For example, we may have large amount

of labeled data for certain groups (e.g., elderly patients) while we have much less labeled data for the other groups (e.g., younger patients). Due to the difference in data sources or data collection process, there commonly exists a shift of data distributions between the source domain and the target domain. In EHR data, the distribution of many features, such as blood pressure, shifts across patients of different races, different age groups, and different genders [Blo].

Domain adversarial learning [Gan16] aims to extract domain-invariant features, which has shown to achieve the state-of-the-art performance for domain adaptation tasks. Specifically, the process involves one label predictor, G_y, and one domain classifier, G_d. The label predictor learns the discriminativeness among the classes of labeled data using standard supervised training process. The domain classifier is introduced to ensure that the learned representation is invariant across domains.

Specifically, the label predictor aims to map the input features to the corresponding labels. Our goal here is to train this label predictor such that its outputs, $\boldsymbol{p}$ (the same as described in Eq. 4.5), stay close to the ground-truth labels y. Therefore, we minimize the loss of the label predictor:

$$\min_{\theta} \mathscr{L}_y(\boldsymbol{\theta};\, \mathscr{D}_{\boldsymbol{S}}) = \sum_t E_{\boldsymbol{x},y\sim\mathscr{D}_{\boldsymbol{S}}}\left[y^t \cdot \ln \boldsymbol{p}^t\right] \tag{6.1}$$

where $\mathscr{D}_{\boldsymbol{S}}$ denotes the joint distribution over input $\boldsymbol{x}$ and label y. In this work we use LSTM as the label predictor.

At the same time, we need to ensure that the extracted feature representation by LSTM is invariant between source and target domains. Specifically, we minimize an adversarial loss such that the extracted features from the source domain and from the target domain cannot be easily distinguished even by a well-trained domain classifier G_d. The classifier G_d maps each $\boldsymbol{h}^t$ to [1,0], which correspond to two domains respectively. Here optimizing G_d aims to best separate between different domains while minimizing the adversarial loss produces $\boldsymbol{h}^t$ that confuses G_d. Hence, the adversarial loss can be defined as a min-max process, as follows:

$$\begin{aligned}\min_{\theta} \mathscr{L}_d(\boldsymbol{\theta};\mathscr{D}_{\boldsymbol{S}},\mathscr{D}_{\boldsymbol{T}}) = \sup_{G_d} \sum_t & E_{\boldsymbol{x}\sim X_{\boldsymbol{s}}}\left[\ln G_d(\boldsymbol{h}^t|\boldsymbol{x})\right] \\ & + E_{\boldsymbol{x}\sim X_{\boldsymbol{T}}}\left[\ln(1-G_d(\boldsymbol{h}^t|\boldsymbol{x}))\right]\end{aligned} \tag{6.2}$$

where $X_{\boldsymbol{S}}$ is the marginal input distribution for the source domain. Similarly, $(\mathscr{D}_{\boldsymbol{T}}, X_{\boldsymbol{T}})$ represent the joint distribution and marginal distribution for the target domain. The $\sup_{G_d}$ operation aims to find the optimal domain classifier to distinguish between the feature representation from the two domains. Combining the adversarial loss with the standard supervised loss, we get the final training loss as follows:

$$\min_{\theta} \mathscr{L}_y(\boldsymbol{\theta};\mathscr{D}_{\boldsymbol{S}}) + \lambda_d \mathscr{L}_d(\boldsymbol{\theta};\mathscr{D}_{\boldsymbol{S}},\mathscr{D}_{\boldsymbol{T}}) \tag{6.3}$$

where λ_d is a weighting factor.

6.3.2 Temporal Adversarial Networks

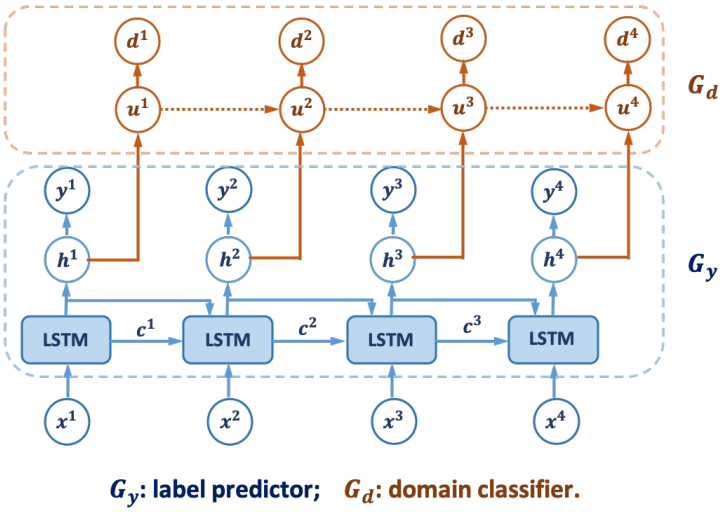


Figure 6.1 Temporal Adversarial Networks.

Domain adversarial learning methods find the discrepancy between two domains by the domain classifier and then update model parameters to decrease the discrepancy. In EHR data, the discrepancy across domains can be better identified by analyzing the temporal profiles. For example, the disease progression patterns for young people and old people might be quite different. However, in traditional adversarial networks, the feature representation $\boldsymbol{h}^t$ at each time step obtained from label predictor are directly fed into the domain classifier as independent input (see Eq. 6.2). Here the learned presentation $\boldsymbol{h}^t$ by label predictor focuses on extracting discriminative temporal knowledge for label prediction while containing less knowledge for the temporal changes of the discrepancy across domains. As a result, the temporal discrepancy between the distributions of the source domain and the target domain cannot be fully learned. Thus, our proposed Temporal Adversarial Networks uses temporal models as domain classifier to capture such temporal changes of the distributions in different domains and further improve the effectiveness of domain adaptation process.

6.3.2.1 Temporal domain classifier

In specific, the learned feature representation, $\boldsymbol{h}$, from label predictor G_y, are fed into a temporal domain classifier G_d. A set of hidden variables $\boldsymbol{u}$ are generated by any proper recurrent structure in the way that $p(\boldsymbol{u}|\boldsymbol{h}) = \prod_t p(\boldsymbol{u}^t|\boldsymbol{h}^1,...,\boldsymbol{h}^t)$. Intuitively, the hidden variable $\boldsymbol{u}^t$ at each each time step

t depends on the input representation $\boldsymbol{h}$ at both time step t and previous time steps. The domain classifier then produces the output d^t based on the generated hidden variables $\boldsymbol{u}^t$ at each time step by a fully connected layer. Compared with standard adversarial learning methods where d^t simply depends on $\boldsymbol{h}^t$ at each time step independently, the proposed domain classifier is able to learn the dynamic changes of the feature distribution and thus better identify the discrepancy across domains in the adversarial learning process.

6.3.2.2 Selective attention and time decay

Additionally, we adjust standard recurrent components, such as LSTM, to better fit into the disease progression modeling considering two facts: 1) the irregular time spans between different time steps in EHR sequences can also affect the patient status at the next moment [Bay17; Yan18]; 2) In the disease progression, the status at each moment are very likely to be relevant with several previous moments. Moreover, these indicative moments are not necessarily the most recent moments due to the complexity in disease progression.

We apply the time-aware attention networks, described in Chapter 5, to address the problem of time irregularity by introducing a decay function based on the length of each time interval. At the same time, we collect information from multiple previous moments by an attention mechanism. Here we maintain the structure of standard recurrent model where each $\boldsymbol{u}^t$ is computed using the input $\boldsymbol{h}^t$ and inherited temporal knowledge $\boldsymbol{u}^{t-1}$ from one previous moment, but we revise the $\boldsymbol{u}^{t-1}$ by incorporating the time irregularity and relevance with multiple previous time steps, as follows:

$$\boldsymbol{u}^{t-1}_{\text{new}} = \sum_{i=t-m}^{t-1} \alpha_{ti} \cdot \boldsymbol{u}^i \cdot g(\Delta t_{ti}), \tag{6.4}$$

where α_{ti} is the attention weight from the i-th moment to the current moment t calculated in Eq. 6.5, Δt_{ti} is the time interval between i-th moment to the current moment, and $g(\cdot)$ is a decay function as in Eq. 6.6, and m represents the number of events to look backwards, which can either be optimized by grid search or be decided by experienced clinicians (similar as Eq. 5.1).

Specifically, the attention weights are computed using local attention, similar with Eq. 5.3, as:

$$\begin{aligned} \boldsymbol{\alpha}_t &= \text{softmax}(e_{t(t-m)}, e_{t(t-m+1)}, ..., e_{t(t-1)}), \\ e_{ti} &= (\boldsymbol{h}^i)^\top \boldsymbol{W}_\alpha \boldsymbol{h}^t, \text{for } i = t-m, ..., t-1 \end{aligned} \tag{6.5}$$

The time decay function in Eq. 6.4 is the same as Eq. 5.5 and is defined as:

$$g(\Delta t) = 1/\log(e + \Delta t) \tag{6.6}$$

According to the above equations, we collect multiple previous memories in patient historical trajectories, as doctors usually review the patient's records to identify the critical previous events and the impacts of events that occur long time ago should be less than those of more recent ones.

6.4 Experiments

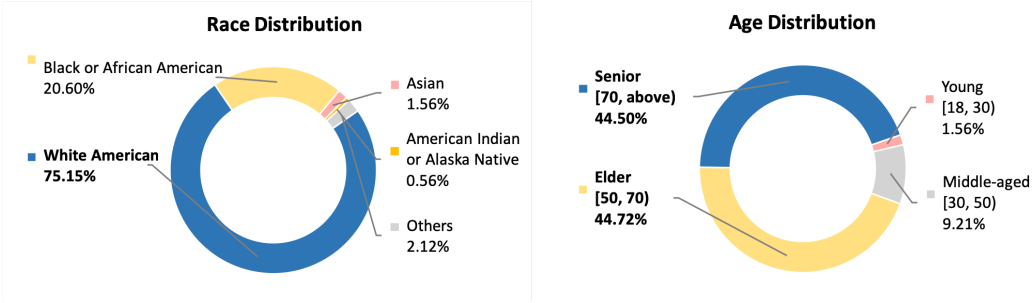
Both *consistent* and *inconsistent* visits from CCHS dataset (described in Chapter 3.3) are used in the following experiments of this thrust. We adopt the same evaluation metrics and settings as discussed in Chapter 4.4.3 for thrust 1. We conduct experiments to answer the following questions:

1. Does the inherent heterogeneity in EHR data impact the disease progression modeling?
2. How does the domain adaptation help tackle the heterogeneity? How different are the feature representations learned by various domain adaptation methods?
3. Whether there exist temporal changes of feature distributions among different patient groups? If so, how would that impact the domain adaptation process?
4. How close the unsupervised domain adaptation process is to the standard supervised learning if we have sufficient labeled data?

6.4.1 Data Grouping

We explore three grouping factors: race, age and gender, to partition patients into different subgroups and adapt learning models among them. Specifically for age, we split adult patients into four groups {young, middle-aged, elder, senior}, with the breakpoints of $\{30, 50, 70\}$ since it shows that the ratio of patients diagnosed as septic shock during their visit significantly differ among age groups according to $\tilde{\chi}^2$ test: $\tilde{\chi}^2(3, N = 52,919) = 307.38$, $p \approx 0 \ll 0.01$. Results from Kolmogorov-Smirnov test (KS test) show that the distributions of many features in our EHRs are indeed significantly different ($p < 0.05$) among different race, age or gender groups. Such findings confirm the existence of the first aspect of heterogeneity in our data. Fig. 6.2 shows the distribution of race and age groups in 1,791 positive shock visits. We can observe that some groups (domains), e.g., White American or senior patients, have sufficient labeled data while others, e.g., African American or young patients, only have few labeled data. This confirms the existence of the second aspect of heterogeneity in our data. Although female and male groups have equivalent number of visits, we still conduct domain adaptation across different genders because gender is a well-known important factor for septic shock progression [Sch98; Nas15].

Finally, Table 6.1 shows the distribution of both consistent (# con) and inconsistent (# inc) positive visits according to race and gender across the four age groups. Later we will describe how to use inconsistent data in domain adaption process.



(a) **(b)**

Figure 6.2 Race and age distribution in positive septic shock visits.

Table 6.1 Distribution of consistent and inconsistent positive visits according to race and gender across the four age groups.

Age		**Race**				**Gender**			
		White American (WA)		**African American (AA)**		**Female (F)**		**Male (M)**	
Group	Range	# con	# inc	# con	# inc	# con	# inc	# con	# inc
Young (YNG)	[18,30)	16	142	6	79	15	165	13	85
Middle-aged (MID)	[30,50)	108	460	47	236	89	412	76	347
Elder (EDR)	[50,70)	579	1,637	188	631	413	1,208	388	1,152
Senior (Sr)	$[70,\infty)$	643	2,263	128	431	366	1,574	431	1,204
Total		1,346	4,502	369	1,377	883	3,359	908	2,788

6.4.2 Approaches

In thrust 1, we have shown the superiority of LSTM for disease progression modeling over both non-sequential models, such as Logistic Regression (LR) and Support Vector Machines (SVM), and other sequential models, such as RNN [Zha17]. Therefore, to evaluate our proposed Time-aware Adversarial Networks, we fix the label predictor as LSTM for all the testing models and mainly focus on comparing different domain classifiers. More specifically, we have six testing models and they are described in three categories:

[–]. Our baseline model without domain classifier.

[DNN$_l$] & [DNN$_e$]. They both use a two-layer deep neural networks (DNN) as the domain classifier [Pur16]. The main difference is that DNN$_l$ performs domain adaptation only at the last time step while DNN$_e$ does it at every time step.

[LSTM] & [ALSTM] & [T-ALSTM]. As described in Section 6.3.2, LSTM, ALSTM (LSTM with attention mechanism), and T-ALSTM (LSTM with selective attention and time decay, or ATTAIN in thrust

2) are three types of domain classifiers used in our proposed Time-aware Adversarial Networks. We want to take advantage of various LSTM-based models to capture the temporal changes of feature distributions.

6.4.3 Experimental Setup

In domain adaptation tasks, the source domain is required to have large amounts of labeled (reliable ground-truth) data while target domain data can be partially labeled or completely unlabeled.

As shown in Table 6.1, the number of visits with reliable ground-truth for septic shock, i.e., *consistent* visits, is much smaller than the corresponding *inconsistent* visits. In experiments, we include inconsistent positive visits as part of target domain data and treat them as “unlabeled". We follow two different experimental settings:

In the first **single-factor** setting, we partition patients based on race, age or gender and adapt among different subgroups within each factor. Specifically, we thoroughly explore the impact of heterogeneity and the effect of different domain adaptation models in the following four subtasks:

1) Heterogeneity within race, age and gender. We investigate the effectiveness of the proposed framework using [LSTM] & [ALSTM] & [T-ALSTM] as domain classifiers in different source-target domain pairs within each factor:

- **Race.** Source domain: White American (WA) vs. Target domain: African American (AA).
- **Age.** Source domain: elder (EDR) or senior (Sr) vs. Target domain: middle-aged (MID). Note that we do not include the young group (YNG) because its sample size 28 is too small to provide a comprehensive evaluation.
- **Gender.** Since female (F) and male (M) groups have comparable number of visits, we explore both F-M and M-F domain-target pairs.

2) Sensitivity on the amount of target domain data. This subtask aims to explore how sensitive different domain adaptation models to the number of *inconsistent* visits included in target domain. We investigate whether the performance of our framework depends on the number of the unlabeled data and whether the large available volumes of unlabeled data can help to improve the domain adaptation performance.
3) Unsupervised domain adaptation vs. supervised learning. We investigate through adapting on target domain using unlabeled data, how close the performance of unsupervised domain adaptation process is to that of the supervised learning using labeled target domain data.
4) Early prediction. We investigate the effectiveness of domain adaptation for early prediction, i.e., to detect the patients who are developing septic shock at an early stage.

In the second **two-factor** setting, we explore heterogeneity across two factors by partitioning patients using race/age and gender/age. Here we aim to explore the shift between feature distributions of more specific patient groups.

6.5 Results

6.5.1 Domain Adaptation Using Single-factor

In this section, we first present the results of domain adaptation within each single factor and compare across the three factors in a combined manner. Second, we provide the sensitivity test of framework on the amount of target domain data. Third, we compare between the unsupervised domain adaptation and supervised learning. Finally, we demonstrate the effectiveness of the framework for early prediction.

6.5.1.1 Heterogeneity within Race, Age and Gender

Table 6.2 shows overall predictive performance of all the models on race groups. Similarly, Table 6.3 and Table 6.4 show the results for age and gender groups. Each sub-table stands for one source-target domain pair and contains three sub-sessions representing for three categories of domain adaptation methods. In each sub-table, we mark the best value for each evaluation metric with $\star$ and the second-best value in bold.

Domain adaptation models. In all the three tables, the five domain adaptation (DA) models consistently outperform the baseline, i.e., the LSTM without any domain classifier (the first row in each sub-table). Also, with one exception on the AUC of EDR-MID pair in Table 6.3, the overall evaluations, AUC and F1, of DNN_e obtain higher values than those of DNN_l. In other evaluation metrics, DNN_e is either higher or close to DNN_l. This demonstrates that the DA process at every time step is more effective than only at the last time step.

Furthermore, we can observe that the proposed Time-aware Adversarial Networks, i.e., the third session in each table/sub-table, generally outperform the traditional DA models without modeling the temporal changes of feature distributions in the second session. DNN-based DA models occasionally reach one or two best or the second-best specificity or precision, but the proposed framework takes over most of the best and the second-best metrics and always hits the best F1 and AUC. For example, in race factor, both F1 and AUC in WA-AA pair get increased by 4% using T-ALSTM compared to DNN_e. Similarly, the increments of F1 and AUC for two pairs in age grouping are 3% and 2%, and for (F-M, M-F) in gender grouping are (8%, 5%) and (5%, 5%). Note that recall values in the third sessions are greatly improved and this contributes to a more balanced model with a reasonable trade-off between precision and recall.

Next, we compare three individual models of Time-aware Adversarial Networks in the third session of each sub-table. The superiority of ALSTM and T-ALSTM can be easily observed. They interchangeably get the best and the second-best values of recall, and they also hit the best AUC and F1. With additionally modeling time irregularity of EHR data, the metrics of T-ALSTM stays above ALSTM or at least comparable, e.g., 0.813 vs. 0.816 of AUC in EDR-MID pair, or 0.668 vs. 0.671 of F1 in M-F pair. Hence, we conclude that the selective attention and time decay further improve the domain adpatation performance.

Table 6.2 Performance (±std) of models with and without domain adaptation within race groups.

S-T	Model	Sensitivity/Recall	Specificity	Precision	F1	AUC
WA-AA	-	0.565(±0.022)	0.769(±0.019)	0.621(±0.024)	0.592(±0.019)	0.698(±0.017)
	DNN_l	0.577(±0.017)	0.832(±0.015)	0.646(±0.020)	0.610(±0.018)	0.722(±0.024)
	DNN_e	0.551(±0.011)	⋆**0.863**(±0.012)	⋆**0.803**(±0.014)	0.653(±0.010)	0.751(±0.011)
	LSTM	0.615(±0.008)	0.811(±0.010)	0.694(±0.021)	0.652(±0.015)	0.762(±0.012)
	ALSTM	**0.638**(±0.012)	0.839(±0.020)	0.699(±0.018)	**0.667**(±0.016)	**0.774**(±0.011)
	T-ALSTM	⋆**0.645**(±0.014)	**0.851**(±0.016)	**0.742**(±0.011)	⋆**0.690**(±0.012)	⋆**0.791**(±0.008)

'S-T' is short for 'Source-Target';
The best and the second-best metrics across six models are labeled with ⋆ and are in bold respectively.

Table 6.3 Performance (±std) of models with and without domain adaptation within age groups.

S-T	Model	Sensitivity/Recall	Specificity	Precision	F1	AUC
EDR-MID	-	0.597(±0.023)	0.903(±0.009)	0.738(±0.020)	0.660(±0.022)	0.788(±0.018)
	DNN_l	0.601(±0.020)	⋆**0.922**(±0.007)	0.737(±0.022)	0.662(±0.021)	0.795(±0.016)
	DNN_e	0.603(±0.018)	**0.918**(±0.015)	0.741(±0.013)	0.665(±0.015)	0.793(±0.021)
	LSTM	0.612(±0.012)	0.887(±0.025)	⋆**0.762**(±0.015)	0.679(±0.014)	0.803(±0.009)
	ALSTM	**0.626**(±0.017)	0.904(±0.011)	0.753(±0.018)	**0.684**(±0.018)	⋆**0.816**(±0.014)
	T-ALSTM	⋆**0.637**(±0.013)	0.912(±0.016)	**0.756**(±0.020)	⋆**0.691**(±0.018)	**0.813**(±0.013)
Sr-MID	-	0.576(±0.029)	0.891(±0.019)	0.711(±0.023)	0.636(±0.024)	0.775(±0.017)
	DNN_l	0.589(±0.021)	0.890(±0.026)	0.710(±0.012)	0.644(±0.017)	0.778(±0.014)
	DNN_e	0.595(±0.017)	⋆**0.906**(±0.013)	0.715(±0.016)	0.650(±0.017)	0.784(±0.010)
	LSTM	0.618(±0.019)	0.892(±0.023)	0.723(±0.024)	0.666(±0.021)	**0.796**(±0.017)
	ALSTM	⋆**0.631**(±0.008)	0.871(±0.019)	**0.727**(±0.009)	**0.675**(±0.009)	0.794(±0.024)
	T-ALSTM	**0.624**(±0.011)	**0.894**(±0.014)	⋆**0.739**(±0.022)	⋆**0.677**(±0.020)	⋆**0.802**(±0.021)

Race. In Table 6.2, from the baseline model to the optimal DA model, the improvements of overall F1 and AUC are both around 10%. Race has the biggest gap regarding to the number of data in source domain and in target domain, due to the dominance of WA, and thus raises plenty of attention and concerns from clinicians. We found that even trained on a large number of WA data, the non-DA

Table 6.4 Performance (±std) of models with and without domain adaptation within gender groups.

S-T	Model	Sensitivity/Recall	Specificity	Precision	F1	AUC
F-M	-	0.553(±0.011)	0.782(±0.023)	0.605(±0.014)	0.578(±0.013)	0.672(±0.025)
	DNN_l	0.552(±0.009)	0.826(±0.006)	0.622(±0.017)	0.585(±0.014)	0.711(±0.017)
	DNN_e	0.564(±0.017)	⋆**0.859**(±0.009)	0.639(±0.013)	0.599(±0.014)	0.728(±0.019)
	LSTM	0.625(±0.020)	0.830(±0.011)	0.667(±0.018)	0.645(±0.018)	0.761(±0.008)
	ALSTM	⋆**0.647**(±0.015)	0.836(±0.016)	**0.683**(±0.016)	**0.665**(±0.015)	**0.773**(±0.013)
	T-ALSTM	**0.641**(±0.005)	**0.843**(±0.013)	⋆**0.715**(±0.009)	⋆**0.676**(±0.007)	⋆**0.782**(±0.006)
M-F	-	0.571(±0.019)	0.767(±0.025)	0.614(±0.022)	0.592(±0.021)	0.703(±0.022)
	DNN_l	0.594(±0.018)	0.813(±0.014)	0.633(±0.015)	0.613(±0.016)	0.735(±0.015)
	DNN_e	0.583(±0.018)	0.842(±0.017)	0.648(±0.020)	0.614(±0.019)	0.749(±0.008)
	LSTM	0.617(±0.016)	⋆**0.866**(±0.009)	⋆**0.705**(±0.010)	0.658(±0.012)	**0.774**(±0.019)
	ALSTM	**0.643**(±0.012)	0.839(±0.010)	0.702(±0.014)	⋆**0.671**(±0.014)	0.772(±0.013)
	T-ALSTM	⋆**0.652**(±0.014)	**0.851**(±0.012)	**0.684**(±0.008)	**0.668**(±0.011)	⋆**0.797**(±0.017)

model performs poorly on AA domain. Compared with the baseline, the fundamental DA model DNN_e boosts ∼ 6% increase in the overall performance. We obtain an additional ∼ 4% increase by capturing the temporal changes of feature distributions among WA and AA. These findings shed light on adapting to other races with even fewer data and provide good reference for clinicians to give more delicate care and diagnosis for different race groups.

Age. Table 6.3 shows that it is also very beneficial to perform DA and apply Time-aware Adversarial Networks within age. Note that the improvement of overall F1 and AUC from the baseline to the optimal DA model are 3% and 3% for EDR-MID pair and are 4% and 5% for Sr-MID pair, which are relatively smaller than those in race grouping. This probably comes from two reasons. First, the discrepancy between the feature distributions of source-target domain pairs within age might not be significant or not be closely related to the septic shock. Table 6.3 shows that the performance of the baseline for both pairs of EDR-MID and Sr-MID is higher than the baseline in race grouping in Table 6.2, e.g., 0.660 (F1 of EDR-MID) vs. 0.592 (F1 of WA-AA), or 0.788 (AUC of EDR-MID) vs. 0.698 (AUC of WA-AA). Second, the target domain, MID, has less available inconsistent visits used for training and less consistent visits used for testing compared to the target domain, AA, in race grouping. The size of data may not support a comprehensive learning and evaluation. Specifically, although the best metrics of EDR-MID are completely higher than those of Sr-MID, Sr-MID achieves bigger improvement. Intuitively, the medical features of EDR and MID are more similar while the age gap is bigger between Sr and MID. Therefore, the domain-shared patterns between EDR and MID are relatively clear and serve a leading start in EDR-MID pair.

Gender. Similar as in race grouping, from the baseline to the optimal DA model, F1 and AUC get increased 10% and 11% for F-M pair and 8% and 10% for M-F pair in gender. Besides, as shown in Table 6.4, we obtain similar values and improvements using either F or M as the source domain and adapting to the other. This is because the target domains in both pairs of F-M and M-F have the

similar number of training and testing data. It indicates that female and male groups have their domain-specific feature distributions, and domain adaptation can be applied to the situations where the data has an imbalanced gender distribution.

6.5.1.2 Sensitivity on the amount of target domain data

We explore how the amount of target domain data would impact the effect of the DA process. In the training process, the number of target domain data can be used as many as the number of source domain data. Here we run the test on WA-AA pair since WA has the largest amount of source domain data in three groupings.

In Fig. 6.3, we track the AUC of the two DA models, i.e., DNN_e (optimal non-temporal DA framework) and T-ALSTM (optimal Time-aware Adversarial Networks), when increasing the amount of target domain data. At each point, we expand the target domain with 100 new random selected target domain data. At the start point, the target domain set is empty and hence it is the baseline without the DA process. It is immediately clear that more target domain data can improve the effect of DA process in both models. Furthermore, T-ALSTM manifests its effectiveness with more available target domain data, i.e., over 200. On the other hand, it requires less data ($\sim$ 1,000) to converge to its optimal performance than DNN_e ($>$ 1,400).

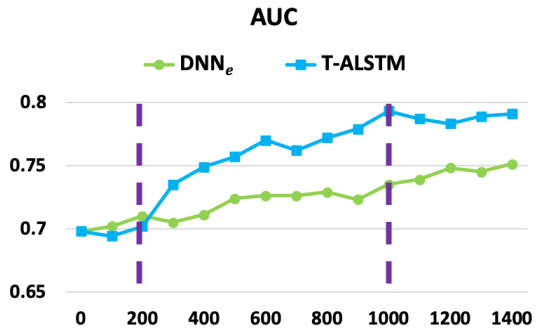


Figure 6.3 Sensitivity of the domain adaptation process on the amount of unlabeled target domain data.

6.5.1.3 Unsupervised domain adaptation vs. supervised learning

We investigate how effective of this unsupervised DA process compared to the supervised learning where the labeled target domain data are available in the training process. We pick up one of the pair under gender factor, F-M, to conduct this experiment since the target domain in gender factor has sufficient data to support a supervised learning.

We design four experiments using different training data while keeping the same testing set, i.e., ~30% of consistent male visits. The training data of four experiments include: 1) **F(con)**: labeled female data; 2) **M(con)**: the remaining 70% labeled male data; 3) **F(con)+M(con)**: labeled female data and labeled male data; and 4) **F(con)+M(inc)**: labeled female data and unlabeled male data. The first three experiments are supervised processes and the last one involves unsupervised DA process using T-ALSTM as the domain classifier.

Fig. 6.4 shows the F1 and AUC of four experiments and the ranking is clear in the order: **M(con)** > **F(con)+M(con)** > **F(con)+M(inc)** > **F(con)**. It is not surprising that **M(con)** obtains the best performance since the training and testing data come from the exact same distribution. On the other hand, **F(con)** obtains the worst performance since the model is unable to obtain information and transfer any knowledge that come from male data. **F(con)+M(con)** is close to **M(con)**, indicating that the model learns from available male information but it cannot fully distinguish male domain from female domain. The performance would further degrade if using fewer male data. **F(con)+M(inc)** is approaching to the optimal results of **M(con)** and **F(con)+M(con)** while is far exceeding **F(con)**. Hence through the unsupervised DA process, the learning model is able to achieve the performance as competitive as the supervised learning process, and not necessarily requires large amounts of labeled data.

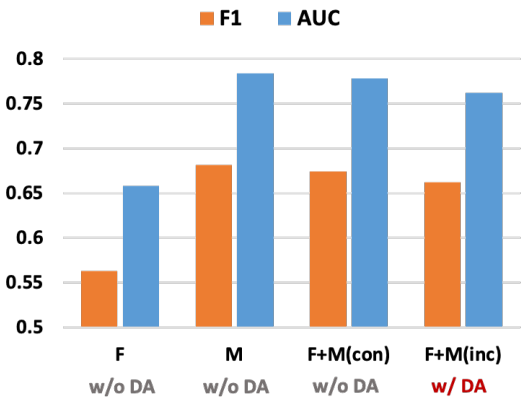


Figure 6.4 Unsupervised vs. supervised processes on male data.

6.5.1.4 Early Prediction

We explore the effectiveness of the proposed framework for early prediction, i.e., to detect the patients who are developing septic shock in advance. The method to perform early prediction using shifting labels is described in Chapter 4.5.4.

We conduct experiments on pairs of WA-AA in race, Sr-MID in age and M-F in gender. We compare between the baseline without DA process and the optimal model with T-ALSTM as the domain classifier. We track their performance by F1 and AUC from one-event early up to 24-event early. Since our data is aggregated on an hourly basis, η-event early guarantees at least η-hour early. Most of prior works report the performance up to 24-hour early for septic shock prediction [Lin18].

The results are shown in Fig. 6.5. Note that the starting point '1' stands for one-event early prediction, which is consistent with the results in previous experiments. We can observe that the performance of both models in three groupings has gradually degraded when predicting future event labels. However, T-ALSTM always outperforms the baseline. Also, the declining trends of F1 and AUC in T-ALSTM is mitigated especially in age pair Sr-MID and gender pair M-F. We can conclude that by learning domain-invariant features and modeling the temporal changes of feature distributions across domains, T-ALSTM effectively and stably captures the developing process of septic shock in different patient groups at an early stage with high performance. It is noteworthy that again the pair Sr-MID under age factor gains less improvement from the DA process compared to that in race and gender.

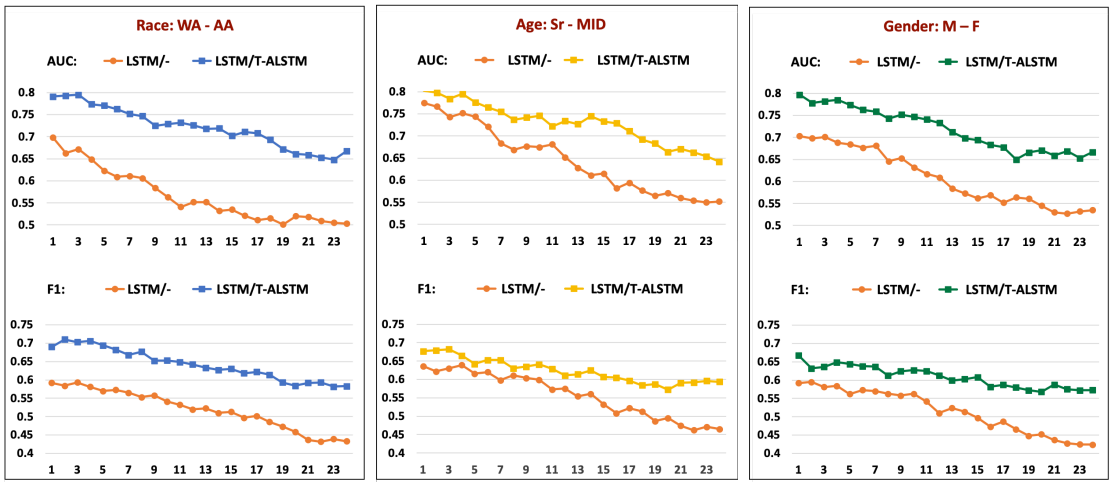


Figure 6.5 AUC and F1-score of early prediction in: (a) Race: WA-AA; (b) Age: Sr-MID; (c) Gender: M-F.

6.5.2 Domain Adaptation Using Two Factors

To further explore the heterogeneity among more specific groups, in Table 6.5, we show the performance of DA using race/age (top sub-table) and gender/age (middle and bottom sub-tables),

Table 6.5 Performance (±std) of models with and without domain adaptation within across different groupings.

Source-Target	DA	Sensitivity/Recall	Specificity	Precision	F1	AUC
(WA,MID)-(AA,MID)	-	0.617(±0.025)	0.798(±0.022)	0.618(±0.018)	0.617(±0.016)	0.742(±0.014)
	✓	0.649(±0.011)	0.812(±0.019)	0.628(±0.015)	0.638(±0.012)	0.769(±0.018)
(WA,EDR)-(AA,EDR)	-	0.591(±0.011)	0.842(±0.017)	0.745(±0.007)	0.654(±0.008)	0.778(±0.014)
	✓	0.631(±0.012)	0.848(±0.017)	0.769(±0.010)	0.693(±0.011)	0.815(±0.011)
(WA,Sr)-(AA,Sr)	-	0.613(±0.022)	0.874(±0.013)	0.765(±0.014)	0.681(±0.017)	0.781(±0.022)
	✓	0.639(±0.010)	0.896(±0.005)	0.790(±0.013)	0.707(±0.011)	0.826(±0.013)
(F,MID)-(M,MID)	-	0.586(±0.024)	0.803(±0.009)	0.601(±0.019)	0.600(±0.021)	0.714(±0.017)
	✓	0.614(±0.020)	0.822(±0.013)	0.622(±0.016)	0.618(±0.018)	0.739(±0.015)
(F,EDR)-(M,EDR)	-	0.625(±0.009)	0.825(±0.022)	0.673(±0.013)	0.648(±0.012)	0.767(±0.016)
	✓	0.661(±0.016)	0.844(±0.018)	0.709(±0.015)	0.684(±0.016)	0.805(±0.018)
(F,Sr)-(M,Sr)	-	0.627(±0.017)	0.830(±0.013)	0.658(±0.014)	0.642(±0.016)	0.756(±0.021)
	✓	0.636(±0.004)	0.857(±0.009)	0.712(±0.008)	0.672(±0.007)	0.787(±0.013)
(M,MID)-(F,MID)	-	0.577(±0.024)	0.752(±0.027)	0.597(±0.022)	0.587(±0.023)	0.697(±0.025)
	✓	0.621(±0.014)	0.833(±0.006)	0.632(±0.008)	0.626(±0.011)	0.736(±0.018)
(M,EDR)-(F,EDR)	-	0.619(±0.017)	0.825(±0.013)	0.658(±0.016)	0.638(±0.017)	0.766(±0.022)
	✓	0.657(±0.011)	0.861(±0.007)	0.679(±0.006)	0.668(±0.009)	0.792(±0.015)
(M,Sr)-(F,Sr)	-	0.631(±0.017)	0.819(±0.015)	0.670(±0.014)	0.650(±0.015)	0.771(±0.010)
	✓	0.648(±0.012)	0.843(±0.012)	0.735(±0.016)	0.689(±0.014)	0.814(±0.019)

respectively. For example, in the top sub-table, the pair (WA,MID)-(AA,MID) means that the source domain is middle-aged White American and the target domain is middle-aged African American. Since female group and male group have comparable number of visits, we interactively explore using female as the source domain (the middle sub-table) and using male as the source domain (the bottom sub-table) on specific age groups. In each sub-table, the first row shows the results of the baseline without the DA process and the second row reports the results with the DA process using T-ALSTM as the domain classifier.

Through all the cells, the model with the DA process consistently outperform the baseline without the DA process on all the evaluation metrics. In both race and gender, the ranking of the adaptation performance on specific age groups is MID<EDR<Sr. We note that the two factors DA is especially beneficial for EDR and Sr groups who are both more vulnerable and have higher risk of septic shock. In particular, we observe that the improvement of the two factors DA over no DA (baseline) is larger than the improvement of using one factor for these two groups: (WA/F/M, EDR)-(AA/M/F, EDR) > WA/F/M-AA/M/F and (WA/F/M, Sr)-(AA/M/F, Sr)> WA/F/M-AA/M/F. For example, the best F1 and AUC from WA-AA pair are 0.690 and 0.791, while for (WA,EDR)-(AA,EDR) are 0.693 and 0.815, and for (WA,Sr)-(AA,Sr) are 0.707 and 0.826. These results motivate us to employ the proposed framework to more specific patient groups given that DA can better capture the connections between these groups.

6.6 Summary

In this thrust, we propose Temporal Adversarial Networks to address the intricate heterogeneity among different patient groups and thus to effectively improve the generalizability of the conventional machine learning approaches such as LSTM. In specific, we combine two LSTM-based recurrent components to extract an invariant feature representation across different domains through an adversarial learning process. The incorporation of temporal models as domain classifier enables a complementary modeling of the temporal changes of the shift between different patient groups. Additionally, we address the time irregularity and collect information from multiple previous moments to better model the complex relationships among events in EHR sequences. As a result, the model learns an invariant feature representation that captures both the medical relation and the time interval effect.

We have several observations from the experimental results on real-world EHR data: 1) the heterogeneity widely exists in EHR data and can severely degrades the performance of conventional machine learning approaches; 2) the proposed framework can address such heterogeneity more effectively compared with state-of-the-art baselines; and 3) the framework is qualified to stably detect patients who are trending the disease at an early stage. Finally, due to complicated clinical reasons, the amount of data with reliable ground-truth for septic shock is greatly limited. The proposed Temporal Adversarial Networks is implemented in an unsupervised fashion. We show that it can significantly improve the prediction results using additional data even with unreliable labels.

CHAPTER

7

THRUST 4: VARIATIONAL RECURRENT NEURAL NETWORKS

7.1 Overview

In standard Recurrent Neural Networks (RNN), like LSTM, the internal transition structure is entirely deterministic. This can be an inappropriate way to model the kind of variability observed in highly structured data [Chu15], such as EHRs. In particular in the tasks of data generation, these models using deterministic methods to produce something between the possibilities since they are trained to minimize the expected distance among the data.

To generate more realistic missing data of EHRs, we propose to introduce the use of high-level latent random variables to model the variability observed in data. In the context of standard neural network models for non-sequential data, the variational autoencoder (VAE) offers an interesting combination of non-linear mapping between the latent random variables and the output and effective approximate inference. Variational Recurrent Neural Networks (VRNN) extends the VAE into a recurrent framework for multi-variate sequential data. In the following we will describe VAE first and then show how to extend it into VRNN, which we use for generating missing EHR data.

7.2 Related Work

Generative models refer to the set of machine learning methods which capture the data distribution. In other words, the generative models aim to learn a distribution which is close to the real data distribution. Most of these methods define the target distribution based on prior knowledge or strong assumptions about data distribution. Also, the construction of target distribution commonly involves latent variables, such as Latent Dirichlet Allocation (LDA) model [Ble03]. Training these models have been a serious challenge in machine learning community [Rob14; WJ08], which is further aggravated by the difficulties in inferring latent variables.

Recent advances in deep learning offer promise to use back-propagation algorithm to effectively build generative models. One of the most popular such methods is the Variational Autoencoder (VAE) [KW13]. Although VAE also involves approximation, the error introduced by this approximation is arguably small given high-capacity deep structured models [Doe16]. Additionally, VAE has been extended to temporal data [Chu15]. In this work, the authors design a generative recurrent network with similar generative process to VAE. This enables better modeling of variability in temporal progression, which has shown success in speech modeling, handwriting generation, and recommender systems [Chu15; Chr17].

7.3 Method

7.3.1 Variational Autoencoders

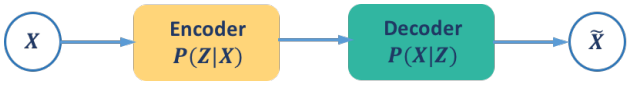


Figure 7.1 Diagram of Variational Autoencoders.

The diagram of VAE is shown in Figure 7.1. We encode the input to latent variables, and then generate new data from that latent variables. We want the generated $\tilde{X}$ is close to the distribution of the original input X. So we aim to maximize the probability of each X in the training set under the entire generative process, according to Equation 7.1.

$$P(X) = \int P(X|z) \cdot P(z) dz \tag{7.1}$$

But we cannot write down a closed form of $P(X)$, which is intractable. On the other hand, in practice, for most z, $P(X|z)$ will be nearly zero, and so they contribute almost nothing to our estimate

of $P(X)$. Then the key idea behind is to attempt to sample values of z that are most likely to have produced X. In this process, we need the function $P(z|X)$ which get us a distribution over z by taking values from X. However, $P(z|X)$ is also not solvable as $P(X)$. What we can do is to use an approximation, $Q(z|X)$, to approach $P(z|X)$ and then model their difference by Kullback-Leibler divergence (KL divergence or $\mathscr{D}$), as indicated by Equation 7.2.

$$\begin{gathered} Q(z|X) \longrightarrow P(z|X) \\ \mathscr{D}[Q(z|X)||P(z|X)] = E[\log Q(z|X)] - E[\log P(X,z)] + \log P(X) \end{gathered} \tag{7.2}$$

After several steps of transformations, we can get Equation 7.3. When we move $\log P(X|z)$ to the left we get 7.4. Notice that our goal is to maximize the likelihood of $\log P(X)$ and KL divergence is always positive, meaning that the right hand side of Equation 7.4 is a lower bound to $P(X)$, also reffered as Evidence Lower BOund (ELBO). Meanwhile, as we already know $E[\log P(X|z)]$ is a maximum likelihood estimation (MLE). In the other words, given an input z and an output X, we want to maximize the conditional distribution $P(X|z)$ under some model parameters. As for $\mathscr{D}[Q(z|X)||P(z)]$, in practice the choice for z is $P(z) = \mathscr{N}(0,1)$. Hence we want to make $Q(z|X)$ to be as close as possible to $\mathscr{N}(0,1)$ so that we could sample it easily. In [KW13], $Q(z|X)$ is a Gaussian $\mathscr{N}(\mu, \text{diag}(\sigma^2))$. Finally $\mathscr{D}[Q(z|X)||P(z)]$ can be written in a closed form.

Therefore, instead of maximizing $P(x)$, which is intractable, we can maximize the its lower bound, which is computationally tractable. Next, we can decompose ELBO to be in the terms of single data points and use standard back propagation for stochastic gradient descent with respect to all the parameters.

$$\begin{aligned} \mathscr{D}[Q(z|X)||P(z|X)] &= E[\log Q(z|X)] - E[\log P(X,z)] + \log P(X) \\ &= -E[\log P(X|z)] + \mathscr{D}[Q(z|X)||P(z)] + \log P(X) \end{aligned} \tag{7.3}$$

$$\log P(X) - \mathscr{D}[Q(z|X)||P(z|X)] = E[\log P(X|z)] - \mathscr{D}[Q(z|X)||P(z)] \tag{7.4}$$

7.3.2 Variational Recurrent Neural Networks

VRNN introduces a recurrent process to the VAR for modelling sequences. We describe the generation process and and inference process in each timestep as follows.

Generation: At each time step, VRNN can degrade to a VAE. However, these VAEs are conditioned on the hidden state h_{t-1} generated by an RNN, which helps VAE to model the temporal dependencies in sequential data. Istead of a standard Gaussian distribution, the prior probability on latent variables follows the distribution in Equation 7.5.

$$z_t \sim \mathscr{N}(\mu_{0,t}, \text{diag}(\sigma_{0,t}^2)), \text{ where } [\mu_{0,t}, \sigma_{0,t}] = \varphi_\tau^{\text{prior}}(h_{t-1}) \tag{7.5}$$

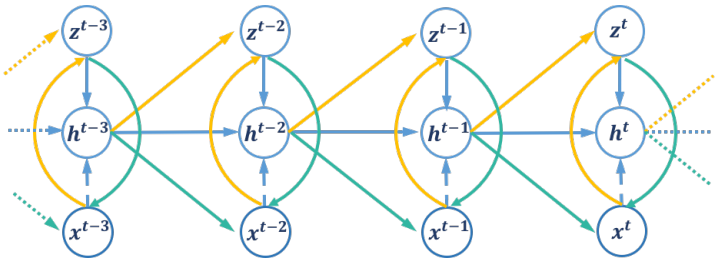


Figure 7.2 Variational Recurrent Neural Networks: the blue line indicates the recurrent process, the yellow line stands for inference process and the green line stands for the generation process.

Moreover, the generation distribution will condition on both z_t and h_{t-1}, as in Equation 7.6, where $\varphi_\tau^{\text{prior}}$ and $\varphi_\tau^{\text{dec}}$ can be any highly flexible functions such as neural networks. φ_τ^x and φ_τ^z can also be neural networks to extract features from x_t and z_t.

$$x_t|z_t \sim \mathcal{N}(\mu_{x,t}, \text{diag}(\sigma_{x,t}^2)), \text{ where } [\mu_{x,t}, \sigma_{x,t}] = \varphi_\tau^{\text{dec}}(\varphi_\tau^z(z_t),\ h_{t-1}) \tag{7.6}$$

Then the hidden states in RNN are updated using the recurrent process expressed in Equation 7.7, where f can be any other transition functions, like the one in LSTM.

$$h_t = f_\theta(\varphi_\tau^x(x_t),\ \varphi_\tau^z(z_t),\ h_{t-1}) \tag{7.7}$$

It is obvious that h_t is a function of $x_{\leq t}$ and $z_{\leq t}$, which means Equation 7.5 defines $p(z_t|x_{<t}, z_{<t})$ and Equation 7.6 defines $p(z_t|x_{<t}, z_{\leq t})$ respectively. Therefore the parameterization of the generative model results in the factorization:

$$p(x_{\leq T}, z_{\leq T}) = \prod_{t=1}^{T} p(x_t|z_{\leq t}, x_{<t}) p(z_t|z_{<t}, x_{<t}) \tag{7.8}$$

Inference: Similarly, the approximate posterior will be the function of both x_t and h_{t-1} as in Equation 7.9.

$$z_t|x_t \sim \mathcal{N}(\mu_{z,t}, \text{diag}(\sigma_{z,t}^2)), \text{ where } [\mu_{z,t}, \sigma_{z,t}] = \varphi_\tau^{\text{enc}}(\varphi_\tau^x(x_t),\ h_{t-1}) \tag{7.9}$$

It is noticeable that the encoding of the approximate posterior and the decoding for generation are both tied through the RNN hidden state h_{t-1}. Also, this conditioning on h_{t-1} results in the factorization as Equation 7.10.

$$q(z_{\leq T} \mid x_{\leq T}) = \prod_{t=1}^{T} q(z_t \mid z_{<t}, x_{\leq t}) \tag{7.10}$$

Learning: The objective function becomens a timestep-wise variational lower bound using Equation 7.8 and Equation 7.10.

$$E_{q(z_{\leq T}|x_{\leq T})}\left[\sum_{t-1}^{T}(-\mathscr{D}(q(z_t|x_{\leq t},z_{<t})||p(z_t|x_{<t},z_{<t}))+\log p(x_t|z_{\leq t},x_{<t}))\right] \tag{7.11}$$

$$E\left[\sum_{t-1}^{T}(\log P(x_t|z_{\leq t},x_{<t})-\mathscr{D}(Q(z_t|x_{\leq t},z_{<t})||P(z_t|x_{<t},z_{<t})))\right] \tag{7.12}$$

7.4 Experiments

Only *consistent* visits from CCHS dataset (described in Chapter 3.3) are used in the following experiments of this thrust. We adopt the same evaluation metrics and experimental settings as discussed in Chapter 4.4.3 for thrust 1.

7.4.1 Missing Data Analysis

Each visit consists of irregular multi-variate time series events with missing values and missing attributes. This is caused by different attributes are measured at different events. For example, vital signs are measured every 8 hours while lab values are measured only every 24 hours. Hence there doesn't always exist available readings for lab results when a new event is created for vital signs. In this thrust, we explore different data imputation methods to generate missing values for 19 important vital sign or lab test features. They are either numeric or categorical and their missing rates are listed in Table 7.1.

Table 7.1 Missing rates of 19 features from CCHS dataset. The first 9 features are vital signs, and the last 10 features are lab results.

Feature	Type	Missing Rate	Feature	Type	Missing Rate
HeartRate	Vital	28.83%	Bands	Lab	98.25%
PulseOx	Vital	35.44%	BUN	Lab	87.31%
RespiratoryRate	Vital	32.23%	Lactate	Lab	96.99%
MAP	Vital	47.40%	WBC	Lab	88.16%
SystolicBP	Vital	43.22%	Platelet	Lab	88.06%
DistolicBP	Vital	43.33%	Creatinine	Lab	87.29%
Temperature	Vital	52.21%	BiliRubin	Lab	97.13%
GlasgowCommaScore	Vital	98.84%	Procalcitonin	Lab	99.95%
GlasgowBestVerbalResponse	Vital	98.80%	CReactiveProtein	Lab	99.89%
			SedRate	Lab	99.87%
Total		53.37%			94.29%

7.4.2 Approaches

Essentially, VRNN can be regarded as a classifier and a data imputation method. From perspective of classifier, we compare it with LSTM.

As a data imputation method, we compare it with three types of popular imputation methods. The details of three baseline imputation methods are listed as follows:

- ***Mean***: fills all the missing values with the mean value of the corresponding features, which is the default imputation method in previous experiments of thrust 1-3.
- ***Forward***: fills the missing values with the last observation until the next value is observed. If the values are still missing, they are filled with feature-wise mean values.
- ***Expert***: is defined by our clinicians and fills the missing values with the last value within the fixed length of forward time window (8 hours for vital signs and 24 hours for lab tests), and fills the remaining ones with feature-wise mean values.

Also, we combine the three imputation methods with missing indicator (MI), which is shown to be an effective way to handle missing data in prior works [Lip16]. The above imputation are only applied to the 19 features listed in Table 7.1. We fill up feature-wise mean value for other features with missing values.

7.4.3 Experimental Setup

We thoroughly explore the effectiveness of VRNN as a classifier and a data imputation method. First, we use *Mean*/*Forward*/*Expert* with or without MIs to impute missing values in our EHRs and then input the imputed data without missing values to LSTM and VRNN to compare the model performance. Next, we use VRNN to impute missing values in EHRs and feed them into LSTM and we compare the imputation performance of VRNN against three baseline imputation methods with or without MIs. Finally, we conduct experiments to test the effectiveness of VRNN for early prediction.

7.5 Results

7.5.1 Results of Overall Prediction

Table 7.2 shows the performance of LSTM and VRNN under different imputation methods. The first sub-table shows the results of LSTM and the second sub-table shows the results of VRNN. Each sub-session in a sub-table represents one imputation method with or without MIs. In each sub-table,

Table 7.2 Performance (±standard deviation) LSTM and VRNN with different imputation methods.

Classifier	Imputation	Sensitivity/Recall	Specificity	Precision	F1	AUC
LSTM	Mean	0.618(±0.013)	0.622(±0.011)	0.634(±0.009)	0.626(±0.010)	0.709(±0.023)
	Mean-MI	0.620(±0.017)	0.619(±0.013)	0.630(±0.015)	0.625(±0.015)	0.717(±0.013)
	Forward	0.587(±0.028)	0.641(±0.015)	**0.654**(±0.018)	0.619(±0.021)	0.682(±0.019)
	Forward-MI	0.633(±0.012)	0.668(±0.018)	0.652(±0.017)	0.642(±0.014)	0.726(±0.008)
	Expert	0.642(±0.019)	0.651(±0.008)	0.641(±0.024)	0.641(±0.021)	0.727(±0.022)
	Expert-MI	**0.663**(±0.008)	**0.684**(±0.011)	0.652(±0.014)	**0.657**(±0.010)	**0.748**(±0.014)
	VRNN	⋆**0.691**(±0.009)	0.702(±0.019)	0.666(±0.015)	0.678(±0.013)	0.759(±0.016)
	VRNN-MI	0.684(±0.013)	⋆**0.730**(±0.007)	⋆**0.687**(±0.016)	⋆**0.685**(±0.014)	⋆**0.765**(±0.014)
VRNN	Mean	0.743(±0.012)	0.637(±0.015)	0.678(±0.008)	0.709(±0.011)	0.797(±0.017)
	Mean-MI	0.721(±0.017)	0.691(±0.014)	0.688(±0.014)	0.704(±0.015)	0.801(±0.017)
	Forward	0.719(±0.017)	0.670(±0.018)	0.686(±0.014)	0.702(±0.016)	0.788(±0.013)
	Forward-MI	0.703(±0.006)	0.720(±0.013)	0.716(±0.009)	0.709(±0.007)	0.800(±0.008)
	Expert	⋆**0.748**(±0.015)	0.690(±0.016)	0.694(±0.009)	0.720(±0.011)	0.815(±0.010)
	Expert-MI	0.745(±0.016)	⋆**0.736**(±0.013)	⋆**0.732**(±0.014)	⋆**0.738**(±0.014)	⋆**0.821**(±0.007)

The best metrics in each sub-table are labeled with ⋆.
For LSTM, the best metrics among three baseline imputation methods are in bold.

we mark the best value for each evaluation metric with ⋆. In specific, in the sub-table of LSTM, the best values among the three baseline imputation methods, i.e. *Mean*/*Forward*/*Expert*, are in bold.

We first illustrate the performance of VRNN as a classifier and compare the performance of LSTM and VRNN fed with the data from three baseline imputation methods. For LSTM, *Expert*, either with or without MI, shows an advantage over *Mean* and *Forward*. *Forward* outperforms *Mean* only when MIs are used. The effectiveness of MIs can reflect on the pairs of *Forward* vs. *Forward*-MI and *Expert* vs. *Expert*-MI while no clear improvement reflects on *Mean* vs. *Mean*-MI. The similar conclusions can be drawn for VRNN with the three baseline imputation methods, except that the improvement from MIs is not as big as that in LSTM. For example, AUC got increased ∼4% from *Forward* to *Forward*-MI of LSTM, while the jump is ∼1% from *Forward* to *Forward*-MI of VRNN. This might come from VRNN itself can extract the similar useful information as MIs. We can observe that the three imputation methods have already achieved competitive performance before using MIs. As a classifier, VRNN completely beats LSTM in that the best metrics obtained from VRNN (the numbers labeled with ⋆ in the second sub-table) are much higher than those obtained from LSTM (the numbers in bold in the first three sub-sessions of the first sub-table). The conclusion also holds when we check the results of each individual imputation method.

Next, we explore the performance of VRNN as an imputation method. When VRNN performs as a classifier, it directly generates the classification results from the latent random variables (i.e. z). When we use VRNN as an imputation method, we need to go through the generation process (i.e. data imputation) in the decoder, and then feed complete data without missing values into other classifiers, such as LSTM. The results of LSTM on VRNN imputed data are shown in the last

sub-session of the first sub-table in Table 7.2, i.e. the rows of *VRNN* and *VRNN*-MI. Clearly, *VRNN* is standing out when compared against the three baseline imputations.

Note that the metrics achieved from LSTM with *VRNN*, which are the best metrics in the first sub-table, cannot catch up the best metrics in the second sub-table where VRNN is used as a classifier. One possible reason is that data generation process (i.e. $z \rightarrow x$) would bring information loss. On the other hand, it also demenstrate that VRNN is a better classifier than LSTM since LSTM cannot win over VRNN even with the complete data without missing values.

7.5.2 Results of Early Prediction

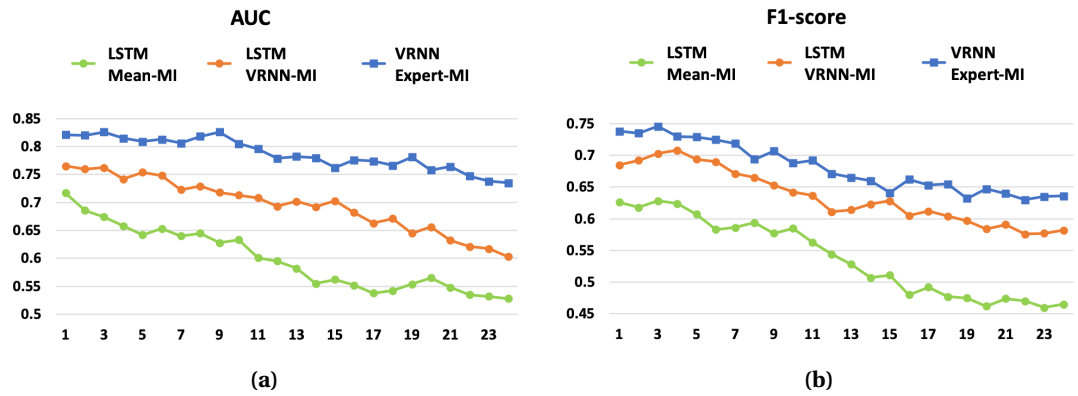


Figure 7.3 (a) AUC of early prediction on five models. (b) F1-score of early prediction on five models.

Here we investigate the effectiveness of VRNN for early prediction, i.e. to detect the patients who are developing septic shock in advance. The method to perform early prediction using shifting labels is described in Chapter 4.5.4.

We conduct experiments on the following three models: LSTM as the classifier and *Mean* with MI as the imputation (LSTM Mean-MI), LSTM as the classifier and *VRNN* with MI as the imputation (LSTM VRNN-MI), and VRNN as the classifier and *Expert* with MI as the imputation (VRNN Expert-MI). We track the performance of each model on F1-score and AUC from one-event early up to 24-event early. Since our data is aggregated on an hourly basis, η-event early guarantees at least η-hour early. Most of prior works report the performance up to 24-hour early for the task of septic shock prediction [Lin18]. The variation of F1-score and AUC with shifting labels are shown in Fig. 7.3. Note that the starting point '1' stands for one-event early prediction, which is equivalent to the result of overall prediction in last experiment.

As expected, as η increases, the early prediction tasks become more and more challenging in

that the performance of all the models has gradually degraded. Despite this, when using LSTM as the classifier, *VRNN* imputation has been steadily staying on the top of the baseline *Mean* imputation. On the other hand, equipped with its best imputation method in Table 7.2, VRNN decreases slower and remains its high performance through all the hours in terms of both AUC and F1. We can conclude that VRNN is both an effective imputation method and classifier for early prediction of septic shock.

7.6 Summary

Missing data pervading in multivariate time series EHRs pose significant challenges in deriving robust predictive models for health-care applications. Most of prior works use deterministic methods for data imputation to generate missing values in EHRs, which are unable to reflect the data variability.

In this thrust, we investigate different missing data imputation methods with LSTM. We adopt an effective generative model, VRNN, to introduce high-level latent random variables to model the variability observed in data and hence to generate more realistic missing data of EHRs. The experimental results show that the completeness of the data contributes to a more comprehensive learning of the disease progression. Especially, *VRNN* imputation shows an overwhelming advantage in generating missing data for EHRs. Additionally, VRNN is a more efficient classifier than LSTM in both tasks of septic shock identification and early prediction. Therefore, we consider the future work of adopting VRNN as the basic module of the whole framework, and implementing its integration with other three thrusts.

CHAPTER

8

CONCLUSION AND FUTURE WORK

In this dissertation, we investigate data-driven approaches for knowledge discovery from large-scale EHRs. The innovations of the proposed framework have significantly improved the detection and generalization performance for the task of disease progression modeling (e.g., septic shock prediction). In particular, we develop RNN-based models to discover discriminative temporal patterns from multi-scale EHRs. To overcome the limitation of the standard recurrent sequential models that memorize only a few previous events, we use an attention model and a time irregularity function to selectively learn from previous events. We further propose an adversarial learning-based method to enhance the generalizability of the extracted temporal patterns. Additionally, we show that the long-term temporal dependencies captured by a VRNN model can be used impute missing values in EHRs. The dissertation is organized into four thrusts, each of which has been deeply reasoning about a specific aspect of the variabilities of EHRs.

In thrust 1, we propose **Multi-scale Label Learning** and focus on two major challenges associated with septic shock prediction: 1) how to capture the informative progression of septic shock in a long visit to hospital of a patient; and 2) how to obtain reliable predictions without well-established moment-by-moment ground-truth labels for septic shock. To solve the first challenge, LSTM is used to capture long-term temporal dependencies and underlying progressing patterns in septic shock. Despite this, the performance of standard LSTM would degenerate due to the difficulty in collecting reliable ground-truth labels in health-care domains. To this end, we design a learning process by jointly leveraging the imperfect labels from different resources (ICD-9 codes vs. clinical criteria)

and at different scales (visit vs. event). The results demonstrate LSTM's capability that memorizing temporal dependencies over a long period compared with multiple non-sequential and sequential models. Also, the proposed multi-scale label learning framework is shown to effectively update the training objectives and obtain the latent genuine labels, and hence achieve a better model performance.

In thrust 2, we propose **Time-aware Attention Networks (ATTAIN)** and focus on the two limitations of the standard LSTM when it is applied to EHRs' learning: 1) it is unable to interpret the prediction results; 2) it ignores the irregular time intervals between consecutive events. The proposed ATTAIN incorporates the attention mechanism and models the time irregularity between events. Specifically, it adjusts the memory content of LSTM when accumulating previous information. Instead of adding memory from one previous event, it retrospects memories of all/several previous events and discount them by weights generated from attention mechanism and the time intervals between those events and current event. The overall weights represent how important each previous event is for the current event to identify the progressing condition. ATTAIN mimics the the process that doctors reading records in practice to capture informative historical moments, improves the interpretability of LSTM and hence provides a cogent way for clinicians to trust machine learning methods. The experimental results validate the design of combining attention and time-aware mechanisms together as ATTAIN achieves an interpretable and reasonable prediction for modeling septic shock.

In thrust 3, we propose **Temporal Domain Adaptation** to address the two aspects of inherent heterogeneity of EHRs: 1) the distributions of input measurements or features often differ among different patient groups (domains, i.e. race, age and gender); 2) some patient groups are more dominant in the number of observations in EHRs than other groups. We propose a domain adaptation (DA) method, combining two temporal RNN-based recurrent components, to extract an invariant feature representation across different patient groups (domains) through an adversarial learning process. While learning the temporal developing patterns of the disease, the method also identifies the temporal changes of feature distributions across patient groups. Consequently, we aim to provide a more specific diagnosis. We thoroughly conduct experiments to investigate the heterogeneity and the effectiveness of the proposed temporal domain adaptation framework. We find that the heterogeneity widely exists in EHR data and can severely degrades the performance of conventional machine learning approaches. The proposed framework can address such heterogeneity more effectively compared with state-of-the-art DA methods. Finally, the proposed framework is implemented in an unsupervised fashion using large amounts of data without reliable labels and hence enables improving the model performance given limited data with reliable ground-truth.

In thrust 4, we use **Variational Recurrent Neural Netrowks (VRNN)** to generate realistic missing values for EHRs. Highly frequent missing values in EHR data can impact the decision-making process, which cannot be directly handled by most machine learning approaches. Prior works

are using deterministic methods for data imputation and are unable to reflect the data variability. As a temporal generative model, VRNN introduces the use of high-level latent random variables to model the variability observed in data and hence generate more realistic missing values for EHRs. In VRNN, the inference process learns about what latent variables have most likely produced the input data, and then encodes the input to latent variables. The generation process generates new data from the latent variables, using the learned distribution from the inference process. The recurrent process is responsible for modelling the transitions between time steps, and is embedded into the other two processes. Therefore, VRNN can be used as a data imputation method or a classifier. From the extensive experiments, we get to know that the generation of missing data enables modeling a complete disease progression, which consequently contributes to a better classification performance. LSTM obtains all the best evaluation metrics when integrated with *VRNN* imputation. In addition, we surprisingly find VRNN is a more efficient model for septic shock prediction, which is a good direction for our future work to replace LSTM with VRNN as the basic module of the whole framework.

We demonstrate the utility of the proposed framework in modeling progression of septic shock and we have following observations from each of four thrusts: 1) The use of multi-scale label learning can lead to a more reliable and robust classification. The proposed method in thrust 1 also has the potential to fix wrong labels. 2) The proposed time-aware attention networks in thrust 2 enables modeling complex temporal dependencies among different events in the progression of septic shock. It also provides better interpretability to the classification. 3) Many traditional machine learning models greatly suffer from the heterogeneity across patients. In contrast, the proposed temporal domain adaptation with adversarial training model introduced in thrust 3 can extracts invariant temporal patterns for different patient groups. 4) Deterministic sequential models are prone to learn averaged temporal profiles. By capturing non-stationary temporal patterns using a probabilistic structure, the VRNN can impute missing data that are closer to real EHR data.

This dissertation has made great contributions to both medical domain and machine learning domains on the following aspects. First, the proposed framework has shown great success in early diagnosis of septic shock. Hence, it can provide timely assessment for patients at high risks so priority actions can be informed accordingly. Second, the proposed framework provides interpretation of results from multiple aspects by answering which labels out of multi-scale labels are more reliable, which previous events are more relevant, and what is the discrepancy between different patient groups. These results can help clinicians better analyze patients' EHR profiles and understand the predictions made by machine learning models. Third, this dissertation addresses multiple challenges (e.g., diverse data scales, variable temporal dependencies, data heterogeneity, and missing data) that ubiquitously exist in many sequential datasets. Hence, the proposed framework is a widespread applicability to many other disciplines.

In the future, we will apply the proposed framework to larger real-world EHR datasets (e.g.,

MIMIC-III and the dataset from our collaborator Mayo Clinic) and also to different types of diseases. Besides, the utility of proposed framework could be explored in other applications and scientific studies that deal with the sequential data of a complex system. For example, in climate research, many climate drivers are collected by sensors onboard earth observing satellites. These drivers are measured under a regular time interval and commonly contain missing values due to natural disturbing factors (e.g., clouds and aerosols). Our proposed framework can potentially help monitor climate conditions using such sequential censored data. It is also possible to apply the proposed framework to analyze sequential customer behaviors for personalized recommendation in e-Commerce applications. Most customers are very likely to click on relevant web pages consecutively, which is indicative of their shopping preference. These scenarios offer opportunities to discover temporal data patterns in applications where automated prediction or classification is urgently needed.

Another set of research direction could focus on further enriching the disease progression modeling with the health-care data collected in people's daily life. For example, many smart wearable devices can track users' heart rate, blood pressure and other movements in real time. The availability of such real-time data offers opportunity to develop efficient online algorithms to capture users' temporal patterns under a higher temporal frequency and consequently detect diseases at an earlier stage. The incorporation of these data sources to our EHR-based framework also makes it possible to monitor patients' status after they leave hospitals/clinics. The advancement of analyzing the real-time health-care data allows developing intelligent assistant to help track health conditions and provide advice for people who do not regularly visit hospitals or clinics.

BIBLIOGRAPHY

[Abr11] Abramson, E. L. et al. “Electronic prescribing within an electronic health record reduces ambulatory prescribing errors”. *The Joint Commission Journal on Quality and Patient Safety* **37**.10 (2011), pp. 470–478.

[AW04] Alemayehu, B. & Warner, K. E. “The lifetime distribution of health care costs”. *Health services research* **39**.3 (2004), pp. 627–642.

[Alv18] Alves, T. et al. “Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation”. *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 1328–1336.

[BL95] Barriere, S. L. & Lowry, S. F. “An overview of mortality risk prediction in sepsis”. *CCM* (1995).

[Bat12] Batal, I. et al. “Mining recent temporal patterns for event detection in multivariate time series data”. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, pp. 280–288.

[BM13] Bauer-Mehren, A. et al. “Network analysis of unstructured EHR data for clinical research”. *AMIA Summits on Translational Science Proceedings* **2013** (2013), p. 14.

[Bax95] Baxt, W. G. “Application of artificial neural networks to clinical medicine”. *The lancet* (1995).

[Bay17] Baytas, I. M. et al. “Patient subtyping via time-aware LSTM networks”. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 65–74.

[Bha17] Bhattacharya, S. et al. “ICU mortality prediction: A classification algorithm for imbalanced datasets”. *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[Ble03] Blei, D. M. et al. “Latent dirichlet allocation”. *Journal of machine Learning research* **3**.Jan (2003), pp. 993–1022.

[Blo] *Blood Pressure Chart: Low Normal High Reading by Age Table*. `https://www.disabled-world.com/calculators-charts/bloodpressurechart.php`. Accessed: 2018-11-28.

[Boh88] Bohnen, J. M. et al. “APACHE II score and abdominal sepsis: a prospective study”. *Archives of Surgery* **123**.2 (1988), pp. 225–229.

[Bra02] Brause, R et al. “Septic shock diagnosis by neural networks and rule based systems”. *Studies in Fuzziness and Soft Computing* (2002).

[Cal16] Calvert, J. S. et al. “A computational approach to early sepsis detection”. *Computers in biology and medicine* (2016).

[Che17] Che, C. et al. “An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson’s Disease”. *SDM*. SIAM. 2017, pp. 198–206.

[Che15] Che, Z. et al. “Deep computational phenotyping”. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 507–516.

[Che16] Che, Z. et al. “Recurrent neural networks for multivariate time series with missing values”. *arXiv preprint arXiv:1606.01865* (2016).

[Che18] Che, Z. et al. “Recurrent neural networks for multivariate time series with missing values”. *Scientific reports* **8**.1 (2018), p. 6085.

[Cho15] Choi, E. et al. “Constructing disease network and temporal progression model via context-sensitive hawkes process”. *2015 IEEE International Conference on Data Mining*. IEEE. 2015, pp. 721–726.

[Cho16a] Choi, E. et al. “Doctor ai: Predicting clinical events via recurrent neural networks”. *Machine Learning for Healthcare Conference*. 2016, pp. 301–318.

[Cho16b] Choi, E. et al. “Retain: An interpretable predictive model for healthcare using reverse time attention mechanism”. *Advances in Neural Information Processing Systems*. 2016, pp. 3504–3512.

[Cho17] Choi, E. et al. “GRAM: Graph-based attention model for healthcare representation learning”. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 787–795.

[Cho16c] Choi, Y. et al. “Learning low-dimensional representations of medical concepts”. *AMIA Summits on Translational Science Proceedings* **2016** (2016), p. 41.

[Cho12] Cholleti, S. et al. “Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions”. *AMIA Annual Symposium Proceedings*. Vol. 2012. American Medical Informatics Association. 2012, p. 103.

[Chr17] Christodoulou, P. et al. “A Variational Recurrent Neural Network for Session-Based Recommendations using Bayesian Personalized Ranking” (2017).

[Chu14] Chung, J. et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. *arXiv preprint arXiv:1412.3555* (2014).

[Chu15] Chung, J. et al. “A recurrent latent variable model for sequential data”. *Advances in neural information processing systems*. 2015, pp. 2980–2988.

[CW11] Coba, V., Whitmill, M., et al. “Resuscitation bundle compliance in severe sepsis and septic shock: improves survival, is better late than never”. *Journal of intensive care medicine* (2011).

[Cro07] Crosson, J. C. et al. “Electronic medical records and diabetes quality of care: results from a sample of family medicine practices”. *The Annals of Family Medicine* **5**.3 (2007), pp. 209–215.

[DI09] Daumé III, H. “Frustratingly easy domain adaptation”. *arXiv preprint arXiv:0907.1815* (2009).

[DM15] De Mulder, W. et al. “A survey on the application of recurrent neural networks to statistical language modeling”. *Computer Speech & Language* **30**.1 (2015), pp. 61–98.

[DW06] De Winter, W. et al. “A mechanism-based disease progression model for comparison of long-term effects of pioglitazone, metformin and gliclazide on disease processes underlying type 2 diabetes mellitus”. *Journal of pharmacokinetics and pharmacodynamics* **33**.3 (2006), pp. 313–343.

[DL08] Dellinger, R. P., Levy, M. M., et al. “Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock: 2008”. *Intensive care medicine* (2008).

[Doe16] Doersch, C. “Tutorial on variational autoencoders”. *arXiv preprint arXiv:1606.05908* (2016).

[Dor07] Dorr, D. et al. “Informatics systems to promote improved care for chronic illness: a literature review”. *Journal of the American Medical Informatics Association* **14**.2 (2007), pp. 156–163.

[Eli11] Elixhauser, A. et al. *Septicemia in U.S. Hospitals, 2009*. 2011.

[Elm90] Elman, J. L. “Finding structure in time”. *Cognitive science* **14**.2 (1990), pp. 179–211.

[Est] Esteban, C. et al. “Combining Static and Dynamic Information for Clinical Event Prediction” ().

[Est16] Esteban, C. et al. “Predicting clinical events by combining static and dynamic information using recurrent neural networks”. *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. IEEE. 2016, pp. 93–101.

[GL14] Ganin, Y. & Lempitsky, V. “Unsupervised domain adaptation by backpropagation”. *arXiv preprint arXiv:1409.7495* (2014).

[Gan16] Ganin, Y. et al. “Domain-adversarial training of neural networks”. *The Journal of Machine Learning Research* **17**.1 (2016), pp. 2096–2030.

[Gat12] Gatti, E et al. “A continuous time Bayesian network model for cardiogenic heart failure”. *Flexible Services and Manufacturing Journal* **24**.4 (2012), pp. 496–515.

[GZ11] Ghanem-Zoubi, N. O. et al. “Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments”. *Critical Care* (2011).

[Gha15] Ghassemi, M. et al. “A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data”. *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[Gha14] Ghassempour, S. et al. “Clustering multivariate time series using hidden Markov models”. *International journal of environmental research and public health* **11**.3 (2014), pp. 2741–2763.

[Giu07] Giuliano, K. K. “Physiological monitoring for critically ill patients: testing a predictive model for the early detection of sepsis”. *American Journal of Critical Care* (2007).

[Gon12] Gong, B. et al. “Geodesic flow kernel for unsupervised domain adaptation”. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2066–2073.

[Gop14] Gopakumar, S. et al. “Stabilizing Sparse Cox Model using Clinical Structures in Electronic Medical Records”. *arXiv preprint arXiv:1407.6094* (2014).

[Gop11] Gopalan, R. et al. “Domain adaptation for object recognition: An unsupervised approach”. *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 999–1006.

[Gra] Graves, A. “Supervised sequence labelling with recurrent neural networks. 2012”. *ISBN 9783642212703. URL http://books. google. com/books* ().

[Gre07] Gretton, A. et al. “A kernel method for the two-sample-problem”. *Advances in neural information processing systems*. 2007, pp. 513–520.

[GG14] Gultepe, E., Green, J., et al. “From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system”. *JAMIA* (2014).

[Ham15] Hammerla, N. Y. et al. “PD Disease State Assessment in Naturalistic Environments Using Deep Learning.” *AAAI*. 2015, pp. 1742–1748.

[Has16] Hashimoto, K. et al. “Domain adaptation and attention-based unknown word replacement in chinese-to-japanese neural machine translation”. *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*. 2016, pp. 75–83.

[HS17] Health, U. D. of & Services, H. *Health, United States, 2016: With Chartbook on Long-term Trends in Health*. 2017.

[Hen14] Henry, K. et al. “63: REWS: REAL-TIME EARLY WARNING SCORE FOR SEPTIC SHOCK”. *Critical Care Medicine* (2014).

[Hen15] Henry, K. E. et al. “A targeted real-time early warning score (TREWScore) for septic shock”. *STM* (2015).

[Her11] Herasevich, V. et al. “Enrollment into a time sensitive clinical study in the critical care setting: results from computerized septic shock sniffer implementation”. *JAMIA* (2011).

[Her13] Herasevich, V. et al. “Connecting the dots: rule-based decision support systems in the modern EMR era”. *Journal of clinical monitoring and computing* **27**.4 (2013), pp. 443–448.

[Ho14a] Ho, J. et al. “Septic shock prediction for patients with missing data”. *Management Information Systems* **5**.1 (2014).

[Ho12] Ho, J. C. et al. “Imputation-enhanced prediction of septic shock in ICU patients”. *Proceedings of the ACM SIGKDD workshop on health informatics (HI-KDD12)*. 2012, p. 18.

[Ho14b] Ho, J. C. et al. “Septic shock prediction for patients with missing data”. *ACM Transactions on Management Information Systems (TMIS)* (2014).

[HS97] Hochreiter, S. & Schmidhuber, J. “Long short-term memory”. *Neural computation* **9**.8 (1997), pp. 1735–1780.

[Hou16] Hou, C.-A. et al. “Unsupervised domain adaptation with label and structural consistency”. *IEEE Transactions on Image Processing* **25**.12 (2016), pp. 5552–5562.

[Hri15] Hripcsak, G. et al. “Parameterizing time in electronic health record studies”. *Journal of the American Medical Informatics Association* **22**.4 (2015), pp. 794–804.

[HY09] Huang, F. & Yates, A. “Distributional representations for handling sparsity in supervised sequence-labeling”. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics. 2009, pp. 495–503.

[Hua14] Huang, S. H. et al. “Toward personalizing treatment for depression: predicting diagnosis and severity”. *Journal of the American Medical Informatics Association* **21**.6 (2014), pp. 1069–1075.

[Ike91] Ikeda, K. et al. “Effect of repeated transcatheter arterial embolization on the survival time in patients with hepatocellular carcinoma. An analysis by the Cox proportional hazard model”. *Cancer* **68**.10 (1991), pp. 2150–2154.

[Ito10] Ito, K. et al. “Disease progression meta-analysis model in Alzheimer’s disease”. *Alzheimer’s & Dementia* **6**.1 (2010), pp. 39–53.

[Jae16] Jaech, A. et al. “Domain adaptation of recurrent neural networks for natural language understanding”. *arXiv preprint arXiv:1604.00117* (2016).

[Jia17a] Jia, X. et al. “Incremental Dual-memory LSTM in Land Cover Prediction”. *SIGKDD*. ACM. 2017.

[Jia17b] Jia, X. et al. “Predict land covers with transition modeling and incremental learning”. *Proceedings of the 2017 SDM*. 2017.

[JZ07] Jiang, J. & Zhai, C. “Instance weighting for domain adaptation in NLP”. *Proceedings of the 45th annual meeting of the association of computational linguistics*. 2007, pp. 264–271.

[Kan18] Kang, G. et al. “Deep Adversarial Attention Alignment for Unsupervised Domain Adaptation: the Benefit of Target Expectation Maximization”. *arXiv preprint arXiv:1801.10068* (2018).

[Kar12] Karnik, S. et al. “Predicting atrial fibrillation and flutter using electronic health records”. *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE. 2012, pp. 5562–5565.

[Kaw12] Kawaler, E. et al. “Learning to predict post-hospitalization VTE risk from EHR data.” *AMIA... Annual Symposium proceedings. AMIA Symposium*. Vol. 2012. American Medical Informatics Association. 2012, pp. 436–445.

[KW13] Kingma, D. P. & Welling, M. “Auto-encoding variational bayes”. *arXiv preprint* (2013).

[Kum06] Kumar, A. et al. “Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock”. *Critical care medicine* (2006).

[Lai17] Lai, G. et al. “Modeling Long-and Short-Term Temporal Patterns with Deep Neural Networks”. *arXiv preprint arXiv:1703.07015* (2017).

[Lao04] Lao, Z. et al. “Morphological classification of brains via high-dimensional shape transformations and machine learning methods”. *Neuroimage* **21**.1 (2004), pp. 46–57.

[Las13] Lasko, T. A. et al. “Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data”. *PloS one* **8**.6 (2013), e66341.

[LG93] Le Gall, J.-R. et al. “A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study”. *Jama* **270**.24 (1993), pp. 2957–2963.

[Li17] Li, Z. et al. “End-to-end adversarial memory network for cross-domain sentiment classification”. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2017, p. 2237.

[Lin18] Lin, C. et al. “Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM”. *ICHI*. IEEE. 2018, pp. 219–228.

[Lip15] Lipton, Z. C. et al. “Learning to diagnose with LSTM recurrent neural networks”. *arXiv preprint arXiv:1511.03677* (2015).

[Lip16] Lipton, Z. C. et al. “Modeling missing data in clinical time series with rnns”. *Machine Learning for Healthcare* (2016).

[Liu13] Liu, Y.-Y. et al. “Longitudinal modeling of glaucoma progression using 2-dimensional continuous-time hidden markov model”. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 444–451.

[Lon15] Long, M. et al. “Learning transferable features with deep adaptation networks”. *arXiv preprint arXiv:1502.02791* (2015).

[Lum02] Lumley, T. et al. “A stroke prediction score in the elderly: validation and Web-based application”. *Journal of clinical epidemiology* **55**.2 (2002), pp. 129–136.

[Luo18] Luo, L. et al. “Discriminative Label Consistent Domain Adaptation”. *arXiv preprint arXiv:1802.08077* (2018).

[Ma17] Ma, F. et al. “Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks”. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 1903–1911.

[Ma18] Ma, T. et al. “Health-ATM: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction”. *SDM*. SIAM. 2018, pp. 261–269.

[Man14] Mani, S. et al. “Medical decision support using machine learning for early detection of late-onset neonatal sepsis”. *Journal of the American Medical Informatics Association* **21**.2 (2014), pp. 326–336.

[Mar03] Martin, G. S. et al. “The epidemiology of sepsis in the United States from 1979 through 2000”. *New England Journal of Medicine* (2003).

[MHH15] Ming Harry Hsu, T. et al. “Unsupervised domain adaptation with imbalanced cross-domain data”. *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4121–4129.

[Mio16] Miotto, R. et al. “Deep patient: an unsupervised representation to predict the future of patients from the electronic health records”. *Scientific reports* **6** (2016), p. 26094.

[NH12] Nachimuthu, S. K. & Haug, P. J. “Early detection of sepsis in the emergency department using Dynamic Bayesian Networks”. *AMIA Annual Symposium Proceedings*. 2012.

[Nac10] Nachimuthu, S. K. et al. “Modeling glucose homeostasis and insulin dosing in an intensive care unit using dynamic Bayesian networks”. *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association. 2010, p. 532.

[Nas15] Nasir, N. et al. “Mortality in Sepsis and its relationship with Gender”. *Pakistan journal of medical sciences* **31**.5 (2015), p. 1201.

[Nel11] Nelson, J. L. et al. “Prospective trial of real-time electronic surveillance to expedite early care of severe sepsis”. *Annals of emergency medicine* (2011).

[NM14] Nguyen, S. Q., Mwakalindile, E., et al. “Automated electronic medical record sepsis detection in the emergency department”. *PeerJ* (2014).

[Nor17] Nori, N. et al. “Learning implicit tasks for patient-specific risk modeling in ICU”. *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[Ols04] Olsson, T. et al. “Rapid Emergency Medicine Score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients”. *Journal of internal medicine* **255**.5 (2004), pp. 579–587.

[Pas12] Pascanu, R. et al. “Understanding the exploding gradient problem”. *CoRR, abs/1211.5063* (2012).

[Pee10] Peelen, L. et al. “Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit”. *Journal of biomedical informatics* (2010).

[Pha16] Pham, T. et al. “Deepcare: A deep dynamic memory model for predictive medicine”. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2016, pp. 30–41.

[Pur16] Purushotham, S. et al. “Variational recurrent adversarial deep domain adaptation” (2016).

[RM06] Raghavan, M. & Marik, P. E. “Management of sepsis during the early âĂIJgolden hoursâĂİ”. *The Journal of emergency medicine* **31**.2 (2006), pp. 185–199.

[Rel09] Rello, J. et al. “PIRO score for community-acquired pneumonia: a new prediction rule for assessment of severity in intensive care unit patients with community-acquired pneumonia”. *Critical care medicine* **37**.2 (2009), pp. 456–462.

[Rob14] Robert, C. *Machine learning, a probabilistic perspective*. 2014.

[Roq11] Roque, F. S. et al. “Using electronic patient records to discover disease correlations and stratify patient cohorts”. *PLoS computational biology* **7**.8 (2011), e1002141.

[Ros05] Rose, C. et al. “A dynamic Bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis”. *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on*. IEEE. 2005, 5–pp.

[SS13] Sarkar, C. & Srivastava, J. “Impact of density of lab data in ehr for prediction of potentially preventable events”. *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*. IEEE. 2013, pp. 529–534.

[Sar10] Sarvestani, A. S. et al. “Predicting Breast Cancer Survivability using data mining techniques”. *Software technology and Engineering (ICSTE), 2010 2nd international Conference on*. Vol. 2. IEEE. 2010, pp. V2–227.

[Sch98] Schröder, J. et al. “Gender differences in human sepsis”. *Archives of Surgery* **133**.11 (1998), pp. 1200–1205.

[SN13] Sengupta, D. & Naik, P. K. “SN algorithm: analysis of temporal clinical data for mining periodic patterns and impending augury”. *Journal of clinical bioinformatics* **3**.1 (2013), p. 24.

[SG04] Seshamani, M. & Gray, A. M. “A longitudinal study of the effects of age and time to death on hospital costs”. *Journal of health economics* **23**.2 (2004), pp. 217–235.

[Sey16] Seymour, C. W. et al. “Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)”. *Jama* **315**.8 (2016), pp. 762–774.

[SW17] Sha, Y. & Wang, M. D. “Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network”. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM. 2017, pp. 233–240.

[Sha97] Shahar, Y. “A framework for knowledge-based temporal abstraction”. *Artificial intelligence* **90**.1-2 (1997), pp. 79–133.

[Sha03] Shapiro, N. I. et al. “Mortality in Emergency Department Sepsis (MEDS) score: a prospectively derived and validated clinical prediction rule”. *Critical care medicine* **31**.3 (2003), pp. 670–675.

[Sha07] Shavdia, D. “Septic shock: Providing early warnings through multivariate logistic regression models”. PhD thesis. Massachusetts Institute of Technology, 2007.

[Shi10] Shin, A. M. et al. “Diagnostic analysis of patients with essential hypertension using association rule mining”. *Healthcare informatics research* **16**.2 (2010), pp. 77–81.

[Sho09] Shortell, S. M. et al. “Improving chronic illness care: a longitudinal cohort analysis of large physician organizations”. *Medical care* **47**.9 (2009), pp. 932–939.

[Sim11] Simon, N. et al. “Regularization paths for CoxâĂŹs proportional hazards model via coordinate descent”. *Journal of statistical software* **39**.5 (2011), p. 1.

[Sin16] Singer, M. et al. “The third international consensus definitions for sepsis and septic shock (sepsis-3)” (2016).

[SG10] Singh, H. & Graber, M. “Reducing diagnostic error through medical home-based primary care reform”. *Jama* **304**.4 (2010), pp. 463–464.

[Sin12] Singh, H. et al. “System-related interventions to reduce diagnostic errors: a narrative review”. *BMJ Qual Saf* **21**.2 (2012), pp. 160–170.

[Soc11] Socher, R. et al. “Parsing natural scenes and natural language with recursive neural networks”. *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 129–136.

[Sta14] Stanculescu, I. et al. “Autoregressive hidden Markov models for the early detection of neonatal sepsis”. *IEEE journal of biomedical and health informatics* **18**.5 (2014), pp. 1560–1570.

[Sub01] Subbe, C. et al. “Validation of a modified Early Warning Score in medical admissions”. *Qjm* **94**.10 (2001), pp. 521–526.

[Sub10] Subbe, C. et al. “The Simple Clinical Score: a tool for benchmarking of emergency admissions in acute internal medicine”. *Clinical medicine* **10**.4 (2010), pp. 352–357.

[Suk12] Sukkar, R. et al. “Disease progression modeling using hidden Markov models”. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2012, pp. 2845–2848.

[Sun12] Sundermeyer, M. et al. “LSTM neural networks for language modeling”. *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[TM10] Tang, C. H., Middleton, P. M., et al. “Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study”. *Physiological measurement* (2010).

[Tan11] Tangri, N. et al. “A predictive model for progression of chronic kidney disease to kidney failure”. *Jama* **305**.15 (2011), pp. 1553–1559.

[Thi10] Thiel, S. W. et al. “Early prediction of septic shock in hospitalized patients”. *JHM* (2010).

[Tib97] Tibshirani, R. “The lasso method for variable selection in the Cox model”. *Statistics in medicine* **16**.4 (1997), pp. 385–395.

[TA16] Torio, C. & Andrews, R. *National inpatient hospital costs: the most expensive conditions by payer, 2013. Rockville, MD: Agency for Healthcare Research and Quality; 2013*. 2016.

[Tze17] Tzeng, E. et al. “Adversarial discriminative domain adaptation”. *Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2. 2017, p. 4.

[Ums15] Umscheid, C. A. et al. “Development, implementation, and impact of an automated early warning and response system for sepsis”. *Journal of hospital medicine* (2015).

[Ver07] Verduijn, M. et al. “Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data”. *Artificial intelligence in medicine* **41**.1 (2007), pp. 1–12.

[Vin96] Vincent, J.-L. et al. “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure”. *Intensive care medicine* **22**.7 (1996), pp. 707–710.

[VR13] Vinzamuri, B. & Reddy, C. K. “Cox regression with correlation based regularization for electronic health records”. *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE. 2013, pp. 757–766.

[WJ08] Wainwright, M. J., Jordan, M. I., et al. “Graphical models, exponential families, and variational inference”. *Foundations and Trends® in Machine Learning* **1**.1–2 (2008), pp. 1–305.

[Wan10] Wang, S.-L. et al. “Prediction of severe sepsis using SVM model”. *Advances in computational biology*. Springer, 2010.

[Wan14] Wang, X. et al. “Unsupervised learning of disease progression models”. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 85–94.

[Wri10] Wright, A. et al. “An automated technique for identifying associations between medications, laboratory results and problems”. *Journal of biomedical informatics* **43**.6 (2010), pp. 891–901.

[Xu07] Xu, R. et al. “Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization”. *IEEE TCBB* (2007).

[Xun16] Xun, G. et al. “Detecting epileptic seizures with electroencephalogram via a context-learning model”. *BMC medical informatics and decision making* (2016).

[Yan18] Yang, X. et al. “Time-aware Subgroup Matrix Decomposition: Imputing Missing Data Using Forecasting Events”. *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 1524–1533.

[Zha17] Zhang, Y. et al. “LSTM for septic shock: Adding unreliable labels to reliable predictions”. *Big Data*. IEEE. 2017, pp. 1233–1242.

[Zho11] Zhou, J. et al. “A multi-task learning formulation for predicting disease progression”. *SIGKDD*. ACM. 2011, pp. 814–822.

[Zho12] Zhou, J. et al. “Modeling disease progression via fused sparse group lasso”. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, pp. 1095–1103.

APPENDIX

APPENDIX

A

RESULTS INTEGRATING THRUST 1 TO THRUST 4

A.1 Overview

From Chapter 4 - Chapter 7, we have thoroughly validate thrust 1 to thrust 4 separately in the proposed framework as shown in Fig. 1.1. Here we integrated all the thrusts together to evaluate the whole framework.

Different from the description order in this dissertation, i.e. Thrust 1 to Thrust 4, the framework is actually implemented in the following process:

- **Thrust 4: VRNN imputation.** We adopt the method of VRNN-MI imputation to generate missing values for EHRs, and feed the complete data to LSTM;
- **Thrust 2: Time-aware Attention Networks.** We add time-aware and attention modules (i.e. ATTAIN) to LSTM in each epoch of the training process and output the probabilities and the predicted labels.
- **Thrust 1: Multi-scale Label Learning.** At the end of each epoch, we jointly use both visit-level

and event-level labels to revise the labels of the training data according to the outputs in the last step.

- **Thrust 3: Deep Domain Adaptation.** Based on the new set of labels, we calculate the total loss for the model and domain classifier, and update the parameters accordingly. Then we proceed to the next training epoch until the overall model converges.

A.2 Experiments

A.2.1 Experimental Setup

Both *consistent* and *inconsistent* visits from CCHS dataset (described in Chapter 3.3) are used in the following experiments. Note that *inconsistent* visits are only used in the thrusts 1 and 3.

We compare the performance of the standard LSTM against that of LSTM integrated with all the four thrusts, LSTM (th1 $-$ th4). In addition, we compose an intermediate model, LSTM integrated with only thrust 2 and 4, LSTM (th2 + th4), since the comparison between the models of LSTM (th2 + th4) and LSTM (th1 $-$ th4) would reflect how much improvement from *inconsistent* data. Given that *consistent* data are rare and difficult to collect in real-world applications, we would like to explore if we can efficiently make use of the *inconsistent* data to improve the model performance.

A.2.2 Evaluation and Setting

We adopt the same evaluation metrics and settings as discussed in Chapter 4.4.3. In addition, the best model and parameters from each thrust are applied to the following experiments. Specifically, we apply *VRNN*-MI imputation for thrust 4, ATTAIN with local attention for thrust 3 and we perform the domain adaptation task with race groups, where WA is the source domain and AA is the target domain. We report the model performance on both all the test data and the target domain data only.

A.3 Results

A.3.1 Results of Overall Prediction

Table A.1 shows the predictive performance of the standard LSTM and two integrated models, LSTM (th2 + th4) and LSTM (th1 $-$ th4). The first sub-table shows the results of the three models on the whole testing set and the second sub-table shows the results of the three models on target domain test data only. In each sub-session of a sub-table, the first row reports the values of evaluation

metrics and the second row indicates the improvement of F1 and AUC compared to the standard LSTM.

When comparing between the standard LSTM and LSTM (th2 + th4), we can observe that all the metrics are improved comprehensively on both *All* and *AA only* testing sets. For example, the AUC got increased 12.5% and 7.4% on *All* and *AA only* respectively. We can conclude that the time-aware attention networks is able to better capture the informative previous events with the complete data without missing values.

The similar improvement patterns reproduce when we add thrust 1 and 3 to the second model, which composing LSTM (th1 − th4). The AUC and F1 score got increased 5.6% and 7.3% on *All* testing set, and 12.9% and 10.9% on *AA only* testing set from LSTM (th2 + th4). Although the absolute performance on *AA only* testing set is not as good as on *All* testing set, but the improvement on *AA only* testing set surpasses that on all the testing set, which validates the effectiveness of DA. Also, the comparison between the two models shows that using additional *inconsistent* data without reliable labels can significantly improve the prediction results.

Overall, when integrating all the four thrusts, we get the performance of the standard LSTM greatly improved, i.e. AUC with an increase **18.1**% and F1 score with an increase **18.7**% on the *All* testing set, and AUC with an increase **20.3**% and F1 score with an increase **16.3**% on the *AA* testing set.

Table A.1 Performance (±std) of LSTM and two integrated models with on all the test data or target domain data only.

Test Set	Model	Sensitivity /Recall	Specificity	Precision	F1	AUC
All	LSTM	0.616(±0.010)	0.709(±0.013)	0.625(±0.015)	0.620(±0.013) **(0%)**	0.713(±0.011) **(0%)**
	LSTM (th2 + th4)	0.704(±0.009)	0.745(±0.017)	0.766(±0.015)	0.734(±0.012) **(11.4%)**	0.838(±0.016) **(12.5%)**
	LSTM (th1 − th4)	0.824(±0.018)	0.814(±0.015)	0.791(±0.014)	0.807(±0.016) **(18.7%)**	0.894(±0.012) **(18.1%)**
AA only	LSTM	0.570(±0.017)	0.728(±0.010)	0.605(±0.019)	0.587(±0.018) **(0%)**	0.673(±0.012) **(0%)**
	LSTM (th2 + th4)	0.613(±0.012)	0.737(±0.019)	0.672(±0.015)	0.641(±0.014) **(5.4%)**	0.747(±0.016) **(7.4%)**
	LSTM (th1 − th4)	0.717(±0.006)	0.822(±0.012)	0.787(±0.009)	0.750(±0.007) **(16.3%)**	0.876(±0.018) **(20.3%)**

A.3.2 Results of Early Prediction

Here we investigate the effectiveness of the integrated models for early prediction, i.e. to detect the patients who are developing septic shock in advance. The method to perform early prediction using shifting labels is described in Chapter 4.5.4.

We conduct experiments on the standard LSTM and the two integrated models, LSTM (th2 + th4) and LSTM (th1 — th4). We track the performance of each model on F1-score and AUC from one-event early up to 24-event early. Since our data is aggregated on an hourly basis, η-event early guarantees at least η-hour early. Most of prior works report the performance up to 24-hour early for the task of septic shock prediction [Lin18]. The variation of F1-score and AUC with shifting labels are shown in Fig. A.1.

As in each thrust, as η increases, the early prediction tasks become more and more challenging in that the performance of all three models has gradually degraded. Despite this, the integrated model LSTM (th2 + th4) keeps being far higher than the standard LSTM and the model equipped with four thrusts, LSTM (th1 — th4), keeps being more effective. In specific, the AUC and F1-score of LSTM (th1 — th4) is hanging around 0.75 and 0.7 even for 24-event early prediction. We can conclude that the whole system can effectively and stably detect patients who are developing septic shock at an early state with high performance.

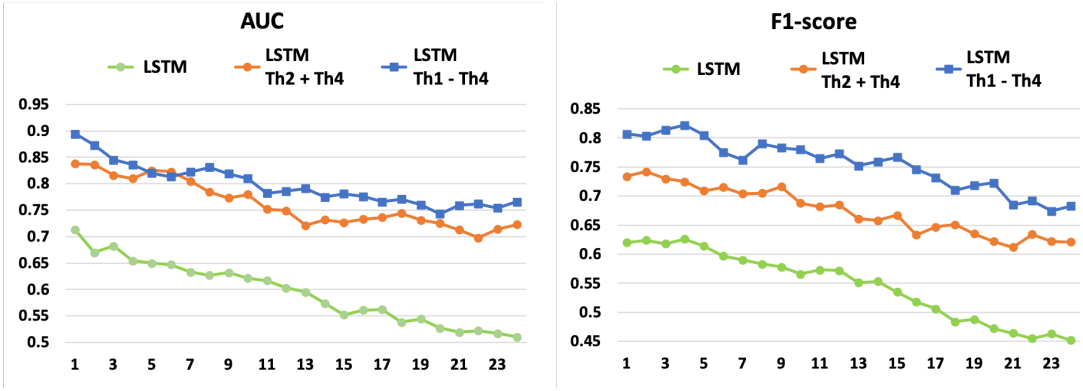


Figure A.1 (a) AUC of the early prediction on LSTM and two integrated models. (b) F1-score of the early prediction on LSTM and two integrated models.

A.4 Summary

In this chapter, we evaluate the whole framework by integrating all the four thrusts into the standard LSTM, and results demonstrate that the proposed framework achieved great success for identifying high-risk patients at an early stage.

As a brief summary, each thrust in the framework contributes to a more reasonable, realistic and informative representation of patients' health status individually and in a combined manner. Multi-scale label learning opens the ways of learning from uncertainty. Time-aware Attention Networks provides a natural and hence cogent way for clinicians to trust machine learning methods. Domain Adaptation delivers a specific care targeting on different patient groups. VRNN approaches a deep understanding with respect to each patient. As a whole, the framework is applied to an extremely challenging disease, e.g., septic shock, and help doctors give immediate care for patients, which would be a huge social good.