GENERALIZABLE MODELS FOR PREDICTION OF PHYSIOLOGICAL DECOMPENSATION FROM MULTIVARIATE AND MULTISCALE PHYSIOLOGICAL TIME SERIES USING DEEP LEARNING AND TRANSFER LEARNING TECHNIQUES

A Dissertation
Presented to
The Academic Faculty

By

Supreeth Prajwal Shashikumar

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2020

GENERALIZABLE MODELS FOR PREDICTION OF PHYSIOLOGICAL DECOMPENSATION FROM MULTIVARIATE AND MULTISCALE PHYSIOLOGICAL TIME SERIES USING DEEP LEARNING AND TRANSFER LEARNING TECHNIQUES

Approved by:

Dr. Shamim Nemati, Advisor
Department of Biomedical Informatics
University of California San Diego

Dr. David Anderson, Co-Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Omer Inan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Zsolt Kira
School of Interactive Computing
Georgia Institute of Technology

Dr. Ravindra Mehta
Department of Medicine
University of California San Diego

Dr. Atul Malhotra
Department of Medicine
University of California San Diego

Date Approved: March 25, 2020

To Appaji and Amma

ACKNOWLEDGEMENTS

I would first like to thank my parents, Dr. **D.R. Shashikumar** and Dr. **K.N. Pushpalatha**, for all the love, encouragement and support that they have provided throughout my life. Thank you for understanding when I have been unresponsive or distracted, and for always advocating the importance of education.

Similarly, it is hard to find the right words to express how grateful I am to my advisor Professor **Shamim Nemati**. During the last four years Shamim has been the one presence in my life that I have found easy to look up to, as a sounding board for when things have gone wrong, and to be the voice of logic and reason in any conversation. Thank you for taking the time to teach me the skills required for pursuing interdisciplinary research in the field of healthcare. Thank you for teaching me first hand the importance of pursuing curiosity driven, original and creative research, and for giving me the freedom and flexibility to explore what I was passionate about. As I look back to the person I was four years ago, and the person I am today, it is hard to find any aspect of my life that has not been changed, in some profound way from our conversations and interactions. I look forward to many more years of working and learning with you as we solve the many pieces of the puzzle that is healthcare predictive analytics.

I would like to thank Professor **David Anderson** who provided me with a lab space to work at during my first semester at Georgia Tech. All the conversations and the mentoring I received from him and his lab members came during a pivotal phase during which I was exploring the field of research I wanted to work in.

Professor **Zsolt Kira**'s deep learning class helped me to learn and understand the fundamentals of neural networks which was essential in conducting research related to this thesis.

I would also like to thank Professor **Omer Inan**, Professor **Ravindra Mehta** and Professor **Atul Malhotra** for agreeing to serve on my committee and for providing crucial

guidance when needed.

I would like to thank the many teachers and professors who have kindly provided me with their guidance and wisdom throughout my life: I will forever be grateful to Professor **Narayan Desai** who provided a second home for me during the preparation of my college entrance exams during high school. Professor **S.R.M. Prasanna**, at the Indian Institute of Technology Guwahati, was my first true academic role model. It was through him that I was first introduced to the world of Digital Signal Processing. His dedication towards pursuing original academic research left a deep impression on me, and further motivated me to pursue a PhD. I would also like to thank Professor **Deepu Vijayasenan**, National Institute of Technology Karnataka, for providing critical support and mentoring during the final years of my undergrad. His courses were one of the few courses that I enjoyed the most attending wherein he laid an emphasis on teaching concepts intuitively. I would also like to thank Professor **Gari Clifford**, Emory University, who was instrumental in providing guidance to me during the first few years of my PhD. Professor **Ashish Sharma**, Emory University, provided useful advice and guidance when it came to implementing the predictive models (developed in this thesis) in real-time at Emory.

I would also like to acknowledge those colleagues without whose help this work could not have been possible: I am grateful to all the members of Department of Biomedical Informatics at Emory University - **Cathy Keeler**, **Jim Kinney**, **Robert Tweedy**, **Julie Schneider**, **Barbara Birt** (Rest in Peace). I would like to thank my labmates at Nematilab for all the concersations and brainstorming sessions we had over the years - Dr. **Chris Josef**, Dr. **Russell Jeter**, **Chad Robichaux**, **Fatemeh Amrollahi** and **Sajad Mousavi**. I'm also grateful to all the current and former members of the Clifford lab at Emory - Dr. **Erik Reinertsen**, Dr. **Giulia Da Poian**, **Camillo**, Dr. **Adriana Vest**, **Chris Aaron**, Dr. **Matthew Reyna** and Dr. **Qiao Li**. I'm also grateful to the academic office of ECE department at Georgia Tech and in particular Dr. **Daniela Staiculescu** whose office doors have also been open to graduate students. Thank you for being considerate and understanding

with all of my requests in these past few years, especially during my move to San Diego. I would also like to thank the Department of Biomedical Informatics at UCSD for graciously welcoming us to San Diego - **Elizabeth Santillanez** and **Paulina Paul**. Special thanks to **Nancy Herbst** who has been very kind in helping process hiring paperworks many number of times ever since I've arrived at UCSD. I would also like to thank Dr. **Gabriel Wardi** and Dr. **Venktesh Ramnath** for being fantastic clinical collaborators at UCSD, and for being very responsive to all the questions I've had with regards to processing the UCSD data.

My friendships have defined the very person that I am today, and for that I am grateful to: Special thanks go to my dear friends **Girish Kini** and **Vidyashankar Lakshman**. The energy and effort Girish puts into caring for and loving the people around him has been a precious source of inspiration for me! Thanks to him, I could not have asked for a better roommate to spend my years at Georgia Tech. It is hard to find words to express my gratitude to Shankar who has been a friend for more than 15 years. Thank you for always being there to pull me out from when I've been at the lowest points in my life. Special thanks go out to the other 7103 folks - **Pradyumna**, **Niteesha** and **Jayati**. 7103 folks have been my pillars of strength through all times, good and the bad. I would like to thank Erik for all the conversations about work and life we had in evenings and weekends at the lab. Additionally, I would also like to thank my friends at Georgia Tech - **Subhrajit**, **Rakshith**, **Supriya Nagesh**, **Yashas**, **Rahul**, **Madhuri**, **Nikhil Sorabha**, **Pooja**, **Shreya** and **Suraj**. It would be a crime to not acknowledge the critical role that **2D boys** have played in my life. A bunch of 19 misfits, we've grown onto become a closely knit family supporting each other through good and bad times. Finally I would like to thank **Amogh**, **Tejas** and **Srinidhi**.

I would also like to thank my aunts **Sudha** and **Preeti** chikkamma and my uncle **Kumar** chikkappa, and my grandmothers for their unconditional love and support they've provided throughout my life. Finally I would like to thank my dear sister **Neha** who has tolerated all my mischief's all throughout these years and has been a great source of support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The goal of this thesis is to develop generalizable machine learning models for early prediction of physiological decomposition from multivariate and multiscale physiological time series data. A combination of recent advances in machine learning and the increased availability of more granular physiological time series data (due to increased adoption of electronic medical records in US hospitals) has encouraged the development of more accurate prediction models for the critically ill patients. One such physiological decomposition prediction task we consider in our work is the early prediction of onset of sepsis. Sepsis is a syndromic, life-threatening condition that arises when the body's response to infection injures its own internal organs. While there are effective protocols for treating sepsis (e.g. administration of broad-spectrum antibiotics, Intravenous fluids, and vasopressors) once it has been diagnosed, there still exists challenges in reliably identifying septic patients early in their course. The purpose of this work is to explore the feasibility of utilizing low-resolution electronic medical record data and high-resolution physiological time series data to develop accurate prediction models for onset of sepsis in critically ill patients.

We first investigate the connection between heart rate (HR) and blood pressure (MAP) time series - as captured through quantification of the structure of their corresponding network representation - for early signs of sepsis. We show that the topological features extracted from these network representations in combination with commonly available EHR features have a better predictive power compared to traditional (such as multiscale entropy) indices that are used to capture physiological time series variability.

We then explore the utility of recurrent neural network models for accurate prediction of onset of sepsis. We present a recurrent neural survival model called Deep Artificial Intelligence Sepsis Expert (DeepAISE). DeepAISE automatically learns predictive features related to higher-order interactions and temporal patterns among clinical risk factors that maximize the data likelihood of observed time to septic events. DeepAISE has been in-

corporated into a clinical workflow, which provides real-time hourly sepsis risk scores. Additionally, we focus on making DeepAISE predictions interpretable wherein relevance scores (inspired from the concept of saliency maps for convolutional neural networks) are used to determine the top contributing factors of the output risk score at every point of time during patient's ICU stay.

Finally, we introduce the Conformal Multidimensional Prediction of Sepsis Risk (COMPOSER) model wherein the focus is on developing a generalizable model that accounts for care-level and institution specific patterns in data generation. We combine ideas from adversarial domain adaptation, representation learning and conformal prediction to develop a model that can adapt well to new target populations (without the requirement of obtaining gold-standard labels). We show the generalizability of COMPOSER on five different patient cohorts comprising of data from three different academic medical centers and two different levels of care (Intensive Care Unit and Emergency Department).

Sepsis survivors often suffer from high rates of readmission and many survivors of sepsis face life-long, debilitating sequelae as a result of the disease. The findings presented in this thesis provide significant clinical evidence for a radical change to the sepsis treatment paradigm that has real-time high-dimensional data analysis and model transparency at its center. The techniques developed in this thesis are general, and can be extended to other physiological decompensations involving multi-dimensional cohort time-series.

CHAPTER 1
INTRODUCTION

1.1 Need for developing prediction models for onset of sepsis

Sepsis is known to be one of the oldest and most elusive syndromes in medicine. It is as a life-threatening organ dysfunction caused by a dysregulated host response to infection [1]. The pathophysiology of sepsis suggests that an infection in human body triggers a complex, variable, and prolonged host response wherein both proinflammatory (directed at eliminating invading pathogens) and anitiinflammatory mechanisms contribute to clearance of infection and tissue recovery as well as organ injury and secondary infections [2]. Additionally, the exaggerated inflammation results in impaired tissue oxygenation and as a result of which organ damage occurs.

Sepsis is a major public health concern accounting for more than $20 billion (5.2%) of total US hospital costs in 2011 [3]. It has also been estimated that the total inpatient admission sepsis cost for Medicare patients was $41.5 billion in 2018 [4]. Sepsis affects over 1 million patients in United States (US) alone, killing roughly a quarter of those affected [5, 6]. Though the condition lacks the same public notoriety as other conditions like heart attacks, 6% of all hospitalized patients in US carry a primary diagnosis of sepsis as compared to 2.5% for the latter. When all hospital deaths are ultimately considered, nearly 35% are attributable to sepsis [6]. This condition stands in stark contrast to heart attacks which have a mortality rate of 2.7-9.6% and only cost the US $12 billion annually, roughly half of the cost of sepsis [3]. Additionally, studies have shown that survivors of sepsis often suffer from long-term physical, psychological, and cognitive disabilities with significant health care and social implications [7].

Starting in 2004 the Surviving Sepsis Campaign (SSC), an international consortium of

professional societies involved in critical care, began addressing the variations in clinical treatment regimens for sepsis and septic shock (a more severe form of sepsis) through the promulgation of evidence based practice guidelines called "sepsis care bundles" [8]. These bundles consolidate the results of numerous investigations that have repeatedly demonstrated improvement in sepsis outcomes after the timely administration of broad-spectrum antibiotics, Intravenous (IV) fluids, and vasopressors when indicated [9, 10, 11, 12]. The most recent recommendation from the SSC is a 1-hr bundle that in addition to obtaining diagnostic tests like cultures and lactate levels, prescribes standard treatment with broad spectrum antibiotics, IV fluid, and vasoactive drugs if necessary, all within an hour of a sepsis diagnosis [13]; Although effective treatment guidelines for sepsis treatment exist, identifying true cases of sepsis before they are clinically apparent is categorically one of the most important needs for modern medicine to address.

A recent study from Seymour et al. [14] observed that for every hour that the administration of antibiotics were delayed in septic patients, the risk of mortality increased by 4-8%. The above point in addition to the aforementioned factors highlight the need for early recognition of sepsis in hospitalized patients. Estimates suggest that if septic patients admitted to US hospitals were identified and appropriately treated with life-saving antibiotics as soon as organ failure is detected, there would be fewer deaths, fewer hospital days, and a reduction in hospital expenditures by about $1.5 billion [15]. The main focus of this thesis is to utilize data that are commonly available in the Electronic Health Records (EHRs) to develop machine learning models that can assist in early recognition of onset of sepsis.

1.2 Defining onset time of sepsis

Consistently identifying the onset time for sepsis presents unique challenges because the condition manifests as a constellation of signs and symptoms with significant variability in presentation and timing. The Third International Consensus Definitions for Sepsis (Sepsis-

3) guidelines have provided two primary criteria for making a formal diagnosis of sepsis: (i) there must be a suspicion for infection (indicated by the administration of antibiotics for at least 72 hours with the concomitant collection of cultures) (ii) there must be a two-point increase in the SOFA (Sequential Organ Failure Assessment) score [16, 1]. These criteria have associated time points and from these time points, sepsis (which we will henceforth identify as $t_{sepsis-3}$) can be consistently labeled (see Table 1.1 for a description of all the timepoints used in our study). While the Sepsis-3 criterion is considered the current standard for labeling sepsis onset time, previous consensus criteria for sepsis (based on Sepsis-1 and Sepsis-2 definitions) [8, 17] remain in wide use. Additionally, there are other sepsis criteria developed by the Center for Disease Control (CDC) and Center for Medicare and Medicaid Services (CMS) for use in surveillance studies [18, 19].

In our work, we use the Sepsis-3 clinical criterion i.e. $t_{sepsis-3}$ for labeling the onset time of sepsis.

Table 1.1: Description of defined time points utilized in the study.

Time point	**Criteria**
$t_{suspicion}$	Clinical suspicion of infection identified as the earlier timestamp of antibiotics and blood cultures within a specified duration. (If antibiotics were given first, the cultures must have been obtained within 24 hours. If cultures were obtained first, then antibiotic must have been subsequently ordered within 72 hours)
t_{SOFA}	The occurrence of end organ damage as identified by a two-point deterioration in SOFA score within a 6-hour period
$t_{sepsis-3}$	The onset time of sepsis-3 is marked when both $t_{suspicion}$ and t_{SOFA} have happened within close proximity to each other. Specifically, t_{SOFA} must occur 24 hours before $t_{suspicion}$ or up to 12 hours after the $t_{suspicion}$ (t_{SOFA} + 24 hours $> t_{suspicion} > t_{SOFA}$ - 12 hours). The earlier of the t_{SOFA} or $t_{suspicion}$ was assigned to $t_{sepsis-3}$.

1.3 Machine learning models for predicting onset of sepsis

In recent years, the increased adoption of Electronic Health Records (EHRs) in hospitals has led to the emergence of Big data in healthcare [20]. The increased opportunities for access to these big data, in parallel with advancements in machine learning has provided an impetus to the development of machine learning based models to analyze these large amounts of data. As a result of this, there has been increased interest to utilize machine learning (ML) techniques in the area of early recognition of sepsis [21].

In this section, we will discuss some of the published works that focus on early prediction of sepsis that are of relevance to this thesis. We direct the reader to the paper from Fleuren et al. [21] for a more comprehensive analysis of all ML based sepsis predictions models that have been published until 2019. The Physionet 2019 challenge focused on development of algorithms for the early prediction of sepsis using routinely available clinical data from three different hospitals in the US. A total of 104 groups from academia and industry participated, submitting entries based on various deep learning approaches to predict the onset of sepsis six hours in advance [22]. Desautels et al. [23, 24] used a proprietary machine learning system called *InSight* based on commonly available patient data (vitals, Oxygen saturation, Glasgow Coma Score and Age) to predict onset of sepsis (level of care: ICU, Sepsis-3 criterion) four hours in advance with Area Under the Curve (AUC) in the range of 0.78-0.85. Nemati et al. [25] used a modified Weibull-Cox model on a combination of low-resolution Electronic Medical Record (EHR) data and high-resolution vital signs time series data (a total of 65 features) to predict onset of sepsis (level of care: ICU, Sepsis-3 criterion) four hours in advance with an AUC of 0.85. Futoma et al. [26] proposed a multiple-output gaussian process based recurrent neural network model for classifying patient encounters (level of care: ICU, modified SIRS criterion) as being septic or not. Lukaszewski et al. [27] demonstrated that a neural network using only cytokine data predicted sepsis (level of care: ICU, Sepsis-1 criterion) better than a similar algorithm using

clinical EMR data. However, cytokines are not routinely measured, making it an impractical tool for contemporary practice. Giannini et al. [28] used a random-forest classifier on a total of 587 features (consisting of demographics, vital signs and laboratory results) to predict onset of sepsis (level of care: ICU, International Classification of Diseases 9th Edition codes for severe sepsis) with an AUC of 0.88. Khojandi et al. [29] used a random-forest classifier on a total of 57 features (vital signs and demographics) to classify patients to be septic or not (level of care: All in patient admissions, SIRS criterion). Wyk et al. [30] used a random-forest classifier on a dataset of high-frequency physiological signals to discriminate between septic and non-septic patients (level of care: ICU, Sepsis-2 criterion). Henry et al. [31] used a Cox proportional hazards model to predict the onset of septic shock in patients admitted to the ICU. Septic shock is a more severe form of sepsis, wherein sepsis leads to low blood pressure that persists despite treatment with intravenous fluids. It should be noted that a direct comparison of these methods is not possible for several reasons: 1) utilization of different labels for sepsis and septic shock, 2) variations in prediction horizon (finite prediction vs infinite horizon prediction), 3) differences in frequency of prediction (single event classification vs sequential prediction), and 4) variations in study design, disease prevalence (case-control design vs calibrated real-world prevalence models), and evaluation methods (e.g., classification versus sequential prediction with varying window sizes).

1.4 Limitations of predictive models in healthcare

Although there has been a growing body of literature on ML models for prediction of sepsis, very few of these have been successfully integrated into a Clinical Decision Support (CDS) system. There are various reasons for this including the lack of interoperability and generalizability (across different institutions or across different levels of care within the same institution) of these models, poor understanding of CDS requirements and bedside workflow processes for a predictive model etc. The work presented in this thesis attempts

to identify key problems in developing ML based generalizable models for predicting onset of sepsis, and provides frameworks and solutions that may help with the deployment and successful integration of these predictive models as a CDS system. In the following three sections, we will briefly discuss and introduce the main focus areas of the thesis.

1.4.1 Capturing interactions in high resolution vital signs time series data

Much of the existing literature on application of predictive analytics to early prediction of sepsis has focused around the EHR and lab results as features for a supervised ML model. However, the resolution of EHR data is far too coarse to produce individually specific predictions. In fact, often availability of certain lab results in themselves may be indicative of clinical suspicion of sepsis. Conversely, without having a clinician in the loop to recognize early signs of decompensation and order the necessary lab values, an EHR-based predictive analytic algorithm may not have access to the required features to predict sepsis in a timely manner. A possible solution to this caveat would be to build predictors based on continuously measured high resolution data such as Heart Rate (HR) and Blood Pressure (BP) time series. It has been studied that the time-series of HR and BP exhibit rich dynamical patterns of interactions and coupling prior to onset of sepsis [32, 33]. According to the anti-inflammatory reflex model [34], pathogen-induced inflammation increases the activity of vagus nerve which controls the production of pro-inflammatory cytokines and prevents tissue damage. Although, the relationship amongst inflammation, vagus nerve activity and heart rate variability (HRV) and Baroreflex control of BP and HR is complex, this model suggests that monitoring indices of heart rate variability and complexity (as markers of vagus nerve activity) may provide useful surrogate markers of the inflammatory reflexes in healthy and diseased populations.

In recent years, one of the novel advances in time series representation and quantification has been the mapping of time series to network, based on ideas such as transition probabilities [35, 36], visibility [37, 38], and correlations [39, 40]. Each of these studies

demonstrated that many characteristics of time series can be extracted from the properties of the corresponding network. Moreover, network-based representations are capable of extracting more nuanced characteristics of time series and could therefore help in building accurate prediction models for onset of sepsis.

1.4.2 Interpretable predictive models in healthcare

While performance characteristics of machine learning algorithms are important, providing interpretable data to the bedside clinicians that can guide diagnosis and therapeutic interventions is a critical requirement of CDS systems. A major barrier to wide adoption of modern machine learning based CDS tools in clinical practise has been their "black box" nature [41]. Thus, there has been increased focus on not only developing high accuracy models but also ensuring that such models provide interpretable explanations of their predictions.

Traditional computational techniques such as logistic regression models and decision trees are popular amongst researchers as they have explainability built into the model itself. However, this should be treated with caution as these explanations represent a "global" notion of interpretability wherein features that contribute to the outcome, for the cohort at large are identified as the top contributing factors. It is often the case that the top contributing factors vary from patient to patient (i.e. local interpretability), and employing models that provide globally important factors would serve limited purpose to the bedside clinicians. While delving into the analysis of top factors contributing to an outcome risk score, there are two types of factors involved: 1) multiplicative interactions where the degree of risk associated with a factor (e.g. temperature) is a function of other factors in a multiplicative sense (e.g. hypothermia and old age are together a greater risk factor than either by itself), 2) temporal contexts of a risk factor can alter its contribution to a given risk score calculation (e.g. leukocytosis immediately after surgery may not be unexpected and contribute differently to the risk for sepsis). These multiplicative and temporal factors result

in variations in the importance of risk factors when viewed from a local, hourly prediction perspective for each patient. Thus, there is a need to develop models which provide locally interpretable predictions for each patient.

1.4.3 Generalizability of machine learning models in healthcare

With advances in tools for collecting and analyzing healthcare data, there have been an increasing number of machine learning algorithms applied to CDS tasks. Sendak et al. provide a comprehensive review of such algorithms, with an emphasis on models that have been integrated into clinical workflow. In addition to the examples mentioned in Section 1.3, other examples of CDS systems include early prediction of AKI [42, 43], as well as algorithms for ED triage [44], prediction of cardiac arrest [45], 30 days readmission prediction [46], and inpatient fall risk assessment [47], among others. However, very few of these CDS systems have been adopted into clinical practise and a prominent reason for this is due to the fact that they suffer from lack of generalizability across institutions and performance degradation within the same institution [48]. This lack of generalizability is due to number factors including differences in local populations, EHR systems, coding definitions, laboratory equipment and assays, as well as variations in clinical and administrative practices. For instance, most existing published clinical ML and predictive analytics models are either based on data from a single hospital [26, 31] or multiple hospitals from the same healthcare system [49] where the processes of care are mostly standardized. Although less common, clinical ML models that have been validated across different healthcare systems are either re-trained from scratch or are fine-tuned (via transfer learning) on every new patient cohort [24, 25, 50], or when applied out-of-the-box often exhibit significant degradation in performance [51]. It should be noted that re-training of models is typically expensive and impractical (namely due to the difficulty of obtaining gold-standard labels).

As noted by Agniel et al. [52], while without careful considerations of context EHR data may be unsuitable for answering many research questions, when healthcare processes

are adequately addressed and incorporated into ML models through introduction of inductive biases (i.e., necessary and appropriate assumptions built into model architecture, learning process, and application/deployment) such data can be leveraged to gain insight into patients' state of health.

1.5 Document Outline and Thesis Contributions

In this thesis, we present techniques to analyse multivariate physiological time series for prediction of early onset of sepsis:

- In **Chapter 2**, we focus on developing a technique to capture interactions between multiscale HR and BP time series - through quantification of the structure of their corresponding network representations - for early prediction of sepsis. The goal of the predictive model is to define a set of physiological states that are represented by the nodes of a network. Transitions among these physiological states are captured by the network edges, and the network structure thereby captures the state trajectory through time. We utilize a non-parametric adaptive partitioning method called Darbellay-Vajda partitioning algorithm to obtain state-space representation of the HR-BP time series. We show that topological features extracted from these constructed networks in combination with commonly available EHR features have a better predictive power compared to traditional indices (such as multiscale entropy) that are used to capture physiological time series variability.

- **Chapter 3** discusses DeepAISE (Deep Artificial Intelligence Sepsis Expert), a recurrent neural survival model for the early prediction of sepsis. DeepAISE utilizes: 1) a class of deep learning models called Recurrent neural networks (RNNs), and 2) the Weibull Cox survival model on a combination of low-resolution EHR data and high resolution vital signs time series data for early prediction of onset of sepsis. This architecture was chosen in the context of predicting sepsis onset time as a time-

to-event analysis and considering that temporal changes in patients' physiology are important for prediction of sepsis. DeepAISE is an externally evaluated sepsis model developed using over 25,000 patient admissions to the Intensive Care Units (ICUs) at two Emory University hospitals, over 18,000 ICU admissions to the UC San Diego Health system and over 40,000 ICU admissions from the Medical Information Mart for Intensive Care-III (MIMIC-III) ICU database. This model has been incorporated into a clinical workflow, which provides real-time hourly sepsis risk scores.

Furthermore, we focus on making DeepAISE predictions interpretable wherein we propose to use *relevance scores* (inspired from the concept of saliency maps for convolutional neural networks) to determine the top contributing factors of the output risk score at every point of time during an ICU stay. Additionally, the hidden representations learned by the recurrent neural network are used to construct a lower dimensional view of a patients' trajectory. These two attributes allow the bedside clinician to identify pathologic deviations from expected physiology early and in real-time throughout the duration of patients' hospital admission. Moreover, we show that the top causes can be broken down into two categories of positively and negatively contributing factors to the risk score. Notably, this analysis has shed insight on the input features contributing significantly to the sensitivity (positive contributors) and specificity (negative contributors) of DeepAISE.

- With the goal of developing generalizable predictive models, **Chapter 4** introduces the **Co**nformal **M**ultidimensional **P**rediction **o**f **Se**psis **R**isk (COMPOSER) model. We first propose a weighted input layer that is designed to handle missing data and variations in data measurement frequency across various levels of care (Emergency Departments, ICUs, Wards etc.) and across different institutions. We then utilize the technique of Adversarial Domain adaptation to learn representations that minimize healthcare system specific variations. A key importance of utilizing ADA training procedure is the design of a predictive model that can adapt to new unlabeled tar-

get patient population; therefore, gold-standard labels which are often expensive to obtain, are not required to deploy the model at a new center. We finally utilize the framework of conformal prediction for establishing the 'conditions for use' of the COMPOSER model. This enables one to explicitly determine at what level of data covariance shift one may still trust a clinical risk score.

We show the generalizability of COMPOSER model by utilizing data from over 480,000 patients collected between 2016-2019 from three different academic medical centers in the US, including data from Emergency Departments (EDs), Intensive Care Units (ICUs), and general wards. Additionally, the COMPOSER model predictions were validated against a cohort of 400 patients who were manually chart reviewed to determine the onset of sepsis.

CHAPTER 2
MULTISCALE NETWORK REPRESENTATION OF PHYSIOLOGICAL TIME SERIES FOR EARLY PREDICTION OF SEPSIS

2.1 Introduction

Sepsis is known as a dysregulated immune-mediated host response to infection. Alteration in heart rate (HR) and blood pressure (BP) variability and coupling prior to onset of sepsis has been reported in the literature [32, 33], and potential links to neuro-immune system interactions have been established. According to the anti-inflammatory reflex model [34], pathogen-induced inflammation increases the activity of vagus nerve which controls the production of pro-inflammatory cytokines and prevents tissue damage. Although, the relationship amongst inflammation, vagus nerve activity and heart rate variability (HRV) and Baroreflex control of BP and HR is complex, this model suggests that monitoring indices of heart rate variability and complexity (as markers of vagus nerve activity) may provide useful surrogate markers of the inflammatory reflexes in health and disease.

Entropy is a measure of unpredictability of the state of a system, or equivalently, of its average information content. Information can be thought of as a measure of "surprise" and entropy can be thought of as a measure of "average surprise". In recent years, one of the novel advances in time series representation and quantification has been the mapping of time series to network, based on ideas such as transition probabilities [35, 36], visibility [37, 38], and correlations [39, 40]. Each of these studies demonstrated that many characteristics of time series can be extracted from the properties of the corresponding network. Moreover, network-based representations are capable of extracting more nuanced characteristics of time series.

In particular, the concept of modularity has been used to characterize time series [53].

By modularity, we mean a set of densely connected nodes within a network. Other authors have used the terms "cluster" or "communities" [54, 55, 56] to denote such constellation of nodes. Networks with high modularity have dense connections between the nodes within modules, but sparse connections between nodes in different modules. An interpretation of what these modules represent is in terms of "set points" of a system. Classical physiology is grounded on the principle of homeostasis in which regulatory mechanisms act to maintain a steady state, i.e., "set point". However, as argued by Ary Goldberger et al. in his editorial [57], many physiological systems tend to operate out of equilibrium and in locally stable regimes (several set points versus a single set point), hence the observation of modularity in the resulting networks of joint HR and BP time series.

Therefore an aim of this study was to investigate the connection between HR and BP time series structure, as captured through quantification of the structure of their corresponding network representation, and early signs of sepsis. However, physiological time series can often exhibit complex patterns of variability over multiple time scales [58, 59]. For instance, time series of BP can exhibit oscillations on the order of seconds (e.g., due to the variations in sympathovagal balance), to minutes (e.g., as a consequence of fever, blood loss, or behavioral factors), to hours (e.g., due to humoral variations, sleep-wake cycle, or circadian effects) [60, 61]. It should also be noted that interactions (or coupling) between physiological systems are often caused by distinct physiological mechanisms that operate across different time scales [62]. We therefore investigate the multiscale structure of vital signs network and their utility for early prediction of sepsis.

2.2 A high resolution dataset of critically ill patients in the Intensive Care Unit

Heart rate (HR) and mean arterial blood pressure (MAP) time series at 2 seconds resolution were collected from bedside monitors in an Emory affiliated Intensive Care Unit (ICU), using the BedMaster system (Excel Medical Electronics, Jupiter FL, USA). All adult ICU units were included in this study, including Medical and Surgical, Cardiac Care, and Neuro-

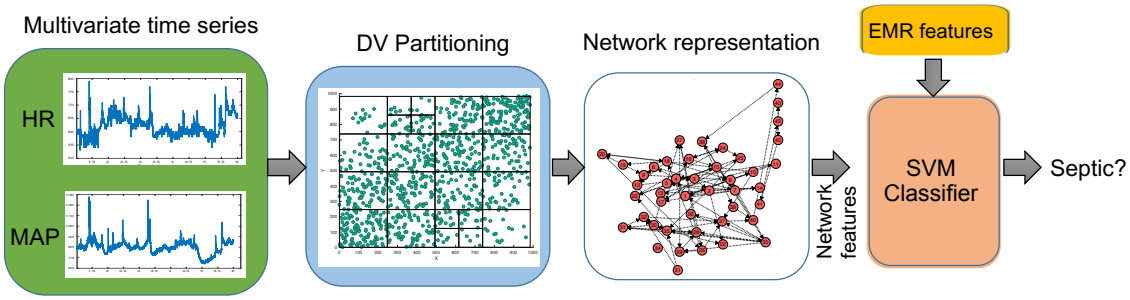


Figure 2.1: **Schematic diagram of the proposed algorithm**. The DV partitions obtained from the space of time-lagged HR and MAP time series are transformed to a network g - which consists of a set of nodes and an Adjacency matrix. Every time scale will have a corresponding network. Various topological attributes and features derived from the constructed networks are used as inputs to the SVM classifier. In addition to the network attributes, EMR features are also fed into the SVM classifier

intensive care units. The bedside monitor data was then matched and time synchronized to each patient's EMR data. A total of 100 patients (around 22%) met the definition of sepsis by Seymour et al. [16] at some time point during their ICU stay. The average length of hospital stay (LOS) among the septic patients was 137.6 [68.2-295.7] hours, and the percentages of in-hospital mortality and in-patient hospice were 15.2% and 13.5%, respectively. The septic patients exhibited a higher average SOFA score compared to non-septic patients (4.8 [3.1-6.8] versus 1.6 [0.6-3.4]).

2.3 Development of the multiscale network model

The goal in this work is to define a set of physiological states, that are represented by the nodes of a network. Transition among these physiological states are captured by the network edges, and therefore the network structure will capture the state trajectory through time.

Dynamic Bayesian networks have been used to model the trajectory of the state of physiological systems [63], where a system's *state* refers to a set of (observed or latent) attributes of the system that summarize all one needs to know about the system to predict its evolution through time [64]. Parametric approaches such as the switching dynamical sys-

tems [65] assume the states transition dynamics to follow a Markov Chain. The approach taken in this work is non-parametric and we extract a set of system states via adaptive partitioning of the state-space. The partitions define the nodes in the corresponding network representation of the time series, and the transition probabilities are captured by the edges.

2.3.1 Defining the state-space

Time-lagged embedding provides information on the underlying dynamical system without having direct access to all the system variables [66]. As a first step to defining the state-space we applied timed-lagged embedding (of order l) to each time series dimensions. Next, the embedded time series samples were replaced by their rank orders (via rank order transformation) to achieve robustness to outliers. This step exploited the fact that mutual information between a set of random variables is invariant to invertible transformations such as the rank order transformation. Next, we partitioned the resulting state-space using an adaptive partitioning algorithm as described next.

2.3.2 The Darbellay-Vajda (DV) partitioning algorithm

As shown in Figure 2.1 the DV partitioning algorithm allows us to partition the state-space associated with a multivariate time series into varying size bins (or hypercubes) for the purpose of density estimation [67]. The DV partitioning was previously shown to be effective in calculating transfer entropy [68, 69], a statistical measure of the amount of directed entropy transfer between two random processes, and it was shown to have lower computational cost than the competing methods. Similar to the method of variable-bandwidth kernel density estimation [70], the DV partitioning algorithm automatically adjusts the bin (partition) size, depending on the density and local distribution of the data points, but requires no *a priori* assumption on the Kernel bandwidth and is computationally more efficient to evaluate [68]. This is in contrast to the equipartitioning scheme (aka, a multidimensional histogram) where the entire state-space is split into equal partitions, which is an inefficient

method to represent non-uniformly distributed data (see Figure 2.2).

The DV partitioning algorithm involves recursively dividing the state-space into more refined partitions, based on chi-squared test statistic that checks whether the data in the proposed partitioned cells are uniformly distributed. Let us consider a bivariate (2-dimensional) time series, $X = [x_1, x_2, ..., x_T]$ and $Y = [y_1, y_2, ..., y_T]$ where T is the length of the time series. First, a non-linear transformation is applied to X and Y, wherein the data in each time series are replaced by their rank orders (also called rank-order transformation). Let the rank order transformed time series be denoted by U and V respectively. We perform partitioning in the UV space as follows:

1. At every iteration, a bin (parent cell) is partitioned into smaller blocks (child cells) and we use the chi-squared test of independence to decide on the need for partitioning to child cells or not. The null hypothesis for the chi-squared test is that the sample distribution in the parent cell is uniform.

2. The chi-squared test statistic $\mathcal{S}$ is given by

$$\mathcal{S} = \left(\sum_{i=1}^{M} \left(\frac{\sum N_i}{M} - N_i \right)^2 \right) / \sum_{i=1}^{M} \left(\frac{N_i}{M} \right) \tag{2.1}$$

   where, M is the total number of child cells for a parent cell, and N_i $(i = 1, ...N)$ are the sample numbers.

3. For a 5% significance level with 3 degrees of freedom, if $\mathcal{S}$ is greater than $\chi^2_{95\%}(3)$, then the distribution of data is not uniform and partitioning is continued. If not, the partitioning is stopped at that level. The level of statistical significance is a parameter that can be tuned.

4. At first, the observation space is partitioned at the medians of U, V margins to generate 4 child cells. And the Chi-square test of independence is performed, if partitioning condition holds, the child cells are split into further smaller blocks (partitioned

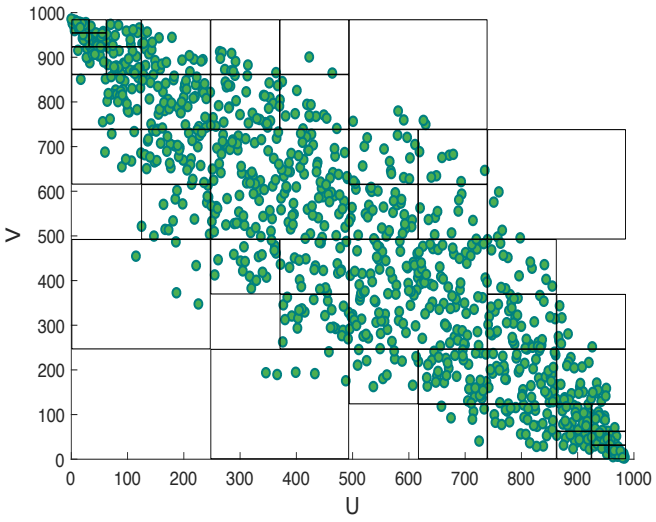


Figure 2.2: **A two-dimensional visualization of DV partitioning.** The observation space consists of 1,000 data points sampled from a bivariate Gaussian distribution with σ_{xy} = -0.9, σ_x^2 = 1, and σ_y^2 = 1. The figure shows the observation space after ordinal sampling. It can be observed that densely populated regions in the space have smaller partitions, in comparison to fewer partitions created in sparser areas.

at the medians of their respective margins), and this continues recursively until the Chi-square test statistic is no more satisfied across all cells.

The output of the partitioning algorithm is thus a list of partitions P, with each partition defined by a lower and upper bound in the observation space. An illustration of the DV partitioning algorithm for bivariate data is shown in Figure 2.2 with the scatter plot of the data and the corresponding partitions obtained. It should be noted that the aforementioned procedure can be easily extended to any arbitrary N dimensional observation space.

2.3.3 Construction of network from partitions

Here we describe the process of construction of a network from a multivariate time series X. An example of a multivariate time series would be the HR and MAP time series recorded from a single subject. Given a list of partitions P, a map $M : T \Rightarrow G$ can be defined from the time domain T to the network domain G. More formally, let us define a map M from time domain $X \in T$ to a network $g \in G$, where $X = \{X_1, X_2, ..., X_k\}$, k is the

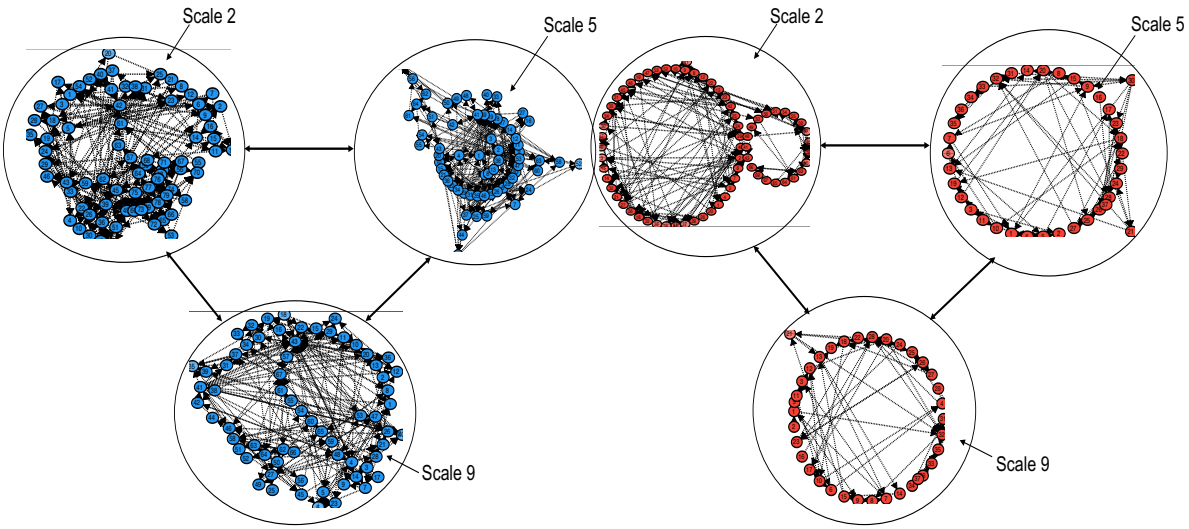


Figure 2.3: **Examples of networks constructed from bivariate time series (HR and MAP) of a control (left panel) and a pre-septic (right panel) patient at different time scales.** Within each of the networks, the arrows represent the transition from one node to another.

total number of time series recorded for each subject (in the above example, since HR and MAP are recorded for every subject, k = 2), and $X_i \in \mathbb{R}^L$, with L being the length of the time series, and $g = \{S, A\}$ consisting of a set of nodes S and adjacency matrix A. The total number of nodes N correspond to the total number of partitions obtained from the DV partitioning algorithm. Therefore each partition p_i $(i = 1, ...N)$ is assigned to a node $n_i \in N$ in the graph g. Every multidimensional data point in X is assigned to one of the partitions. The adjacency matrix A is a NxN matrix where a_{ij} corresponds to the transition from node n_i to node n_j. Two nodes n_i and n_j are connected in the network with a weight a_{ij}, with a_{ij} representing the total number of transitions from node n_i to n_j. Each partition p_i can be thought of as a dynamical state in a physiological system and the a_{ij} of the adjacency matrix represent the probability of transition between the dynamical states of the system. In the above example, we would thus construct one network from the bivariate time series (HR and MAP time series) recorded from the subject.

Multiscale Network representation

Interactions in biological systems manifest on multiple time scales [62], and the interactions may change in different ways at these different time scales. It may therefore be important to capture this multiscale nature of the interactions to help differentiate between healthy and unhealthy individuals. For a one dimensional time series $[x_1, x_2, ..., x_N]$, a coarse grained time series $\{y^{(\tau)}\}$, corresponding to the scale factor τ was constructed as follows: First, the original times series was divided into non-overlapping windows of length τ; second, the data points inside each window were averaged. In our experiments, we coarse grained both HR and MAP according to the scale factor τ. Thus, for every scale factor τ_i $(i = 1, ...M)$, where M is the total number of scale factors, a network G_i $(i = 1, ...M)$ was constructed. Figure 2.3 provides a visualization of the networks constructed from bivariate time series (HR and MAP) of a control and a pre-septic patient at different time scales.

2.3.4 Network attributes for classification

In our proposed algorithm, for the network that we obtain as described in the previous sections, we compute many topological attributes and use the derived features for classification. The following network attributes were computed for every network in the dataset: **number of nodes** (total number of nodes in the network), **number of edges** (total number of edges in the network), **Link density** (defined as the total number of edges divided by the maximum possible edges in the network), **average degree** (the average value of the degree of all nodes in the network, where the degree of a node is defined as the total number of it's neighboring edges), **number of loops** (the total number of independent loops in the network, also know as the "cyclomatic number" or the number of edges that need to be removed so that the network cannot have cycles), **Loop3** (the total number of loops of size 3 in the network), **Loop4** (the total number of loops of size 4 in the network), **average clustering coefficient** (the clustering coefficient $c(u)$ for node u can be defined as the ratio of the number of actual edges between the neighbors of u to the number of possible edges

between them, and the average clustering coefficient $C(G)$ of a network is the average of $c(u)$ taken over all the nodes in the network), **Pearson coefficient** (the pearson correlation coefficient for a degree sequence, also known as the assortativity coefficient [71]), **Algebraic connectivity** (the second smallest Eigen value of the Laplacian matrix of a network, where the Laplacian matrix of a network is the difference between the sum of degrees of the diagonal elements in adjacency matrix and the adjacency matrix), **Closeness** (the closeness centrality, $cc(u)$ for node u is the inverse of sum of distance from node u to all other nodes in the network, where the closeness centrality of a graph is the average mean of the above is the average of $cc(u)$ taken over all the nodes in the network, **Average eccentricity** (eccentricity of a node u is defined as $e(u) = max\{d(u, v) : v \in V\}$, where the distance $d(u, v)$ is the length of the shortest path from u to v, and V is the set of all nodes. The average effective eccentricity is the average of effective eccentricities over all nodes in the network), **Maximum effective eccentricity** (Also known as the effective diameter, is defined as the maximum value of effective eccentricity over all nodes in the graph), **Spectral radius** (defined as the largest magnitude eigenvalue of the adjacency matrix of the network), **Trace** (sum of the eigenvalues of the adjacency matrix, i.e., $\sum \lambda$, and **Energy** (squared sum of the eigenvalues of the adjacency matrix A. More formally, the energy of a network G is: $E(G) = \sum_i^n \lambda_i^2$).

2.3.5 Entropy and other EMR features

For every subject, their socio-demographics features (Age, Gender, Weight, Race) were collected. We also included features that were commonly recorded by the bedside nurses including, Mean Arterial Pressure (MAP), Heart Rate (HR), Peripheral capillary Oxygen Saturation (SpO_2), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiration Rate ($Resp$), Glasgow Coma Score (GCS), and Temperature ($Temp$). Each of the above mentioned features were quantized into 8 levels, and each level was encoded into dummy binary representations. And these discretized representations were used in the

classification model. We also extracted a few features that capture history, comorbidity, and the clinical context of the patient, including Charlson Comorbidity Index, Mechanical Ventilation, Unit Information (surgical, cardiac care, or neuro-intensive care), as well as Surgical Speciality (cardiovascular, neuro, ortho-spine, oncology, urology, etc) and Wound Type (clean, contaminated, dirty, or infected) if the patient had a surgery in past 12 hours.

We also calculated the following features from the HR and MAP time series (2 second resolution) derived from the bedside monitor's proprietary software from the ECG and BP waveforms: standard Deviation of HR (HR_{STD}), Standard Deviation of MAP (MAP_{STD}), Multiscale Entropy [59] of (60/HR or RR intervals) and MAP (Over 17 Scales; RR_{MSE}, and MAP_{MSE} respectively)

2.4 Feature selection and classification

For every subject in the dataset, networks were constructed for time scales 1 through 10. A total of 16 network attributes were extracted from every constructed network. It is to be noted that the HR and MAP were each processed with a lag of order l. In addition to the network attributes, the Entropy and EMR features as described in Section 2.3.5 were extracted. All the features were then used to train a Support Vector Machine (SVM) classifier to predict onset of sepsis four hours ahead of time, based on the data from preceding six hours. The output of the SVM was the probability of membership in the Sepsis class. Hyper-parameters of the model including the time scale factors, and lag order l were optimized using Bayesian Optimization technique [72].

For all continuous variables, we have reported the medians with Inter-Quartile range (IQR), and utilized a two-sided Wilcoxon ranksum test when comparing the septic and control populations. For binary features, we have reported the percentages, and utilized a two-sided Chi-square test to assess differences in proportions between the septic and control populations. To assess the performance of the proposed algorithm on out-of-sample data, we performed a 10 fold cross validation study. The features in training set were trans-

Table 2.1: Patient characteristics in the dataset

	Control	**Septic**	**p-value**
N	100	150	–
Age	59.5 [46.0 68.0]	63.0 [47.5 72.5]	0.15
Male(%)	56%	48%	0.21
MAP	81.7 [75.0 90.1]	78.5 [70.3 91.3]	0.22
HR	84.8 [73.2 97.6]	92.5 [75.1 110.0]	<0.01
SpO_2	97.6 [96.3 99.3]	97.9 [95.1 99.5]	0.32
SBP	126.0 [111.7 143.7]	121.2 [103.3 143.3]	0.20
DBP	60.0 [55.0 66.7]	58.3 [52.5 67.2]	0.25
Respiration Rate	16.8 [14.2 18.7]	16.2 [2.25 20.4]	0.3
GCS	14.5 [10.0 15.0]	9.7 [6.0 14.3]	<0.01

formed to have Gaussian distributions using either the identity, square root or logarithmic transformations. The transformation which provided the lowest k-statistic using the Lilliefors test was used on both training and test sets. The transformed data (both training and test data) was then normalized by subtracting the mean computed from the training set and dividing by the standard deviation computed from the training set. Feature transformation, training, and classifier evaluation was performed separately for all the ten folds. Area Under the Receiver Operating Characteristic (AUROC) curve, accuracy, and specificity were calculated for training and test sets for all the folds. The sensitivity level was fixed at 0.85. We combined all the predictions (probability of being septic) across all the 10 folds to report a single pooled AUROC [73].

2.5 Results

A total of 250 subjects were considered for this study. The median [IQR] age for the septic and control subjects was 63 [47.5 72.5] and 59.5 [46.0 68.0] respectively. The patient characteristics of the entire dataset have been tabulated in Table 2.1. It can be observed that the onset of sepsis is associated with a drop in MAP as well as SBP, DBP, and a significant increase in HR (92.5 vs. 84.8) and a significant decrease GCS (9.7 vs. 14.5), reflecting a moderate loss of consciousness or alertness.

2.5.1 Construction of network based on HR and MAP

The most commonly selected scales and embedding dimension by the Bayesian Optimization were scales 2, 3, 5, 6, 7, 9, and 10, lag order of 3. We therefore fixed these parameters across all experiments and model comparisons. We employed feature selection to find a minimum of set of relevant features. The most commonly selected features across all scales included the average clustering coefficient, pearson correlation coefficient, spectral radius, energy of graph, Trace, and number of loops of size 4.

In the following experiments we used the graph attributes alone as features for the classifier. First, we constructed multiscale networks from HR alone, and the pooled testing AUROC was 0.61. Next, we constructed multiscale networks from MAP alone, and the pooled testing AUROC was 0.61. By combining HR and MAP, and constructing multiscale networks achieved a pooled testing AUROC of 0.78.

2.5.2 Classifier trained on combination of Network, entropy and EMR features

Table 2.2: Performance summary of classifier trained on combinations of network, entropy and EMR features. Values shown are pooled AUROCs

Model	**Training AUROC**	**Testing AUROC**
MSE	0.72	0.66
EMR	0.79	0.70
MSNR (MAP + HR)	0.85	0.78
MSE + EMR	0.83	0.73
MSNR + EMR	0.89	0.79
MSNR + MSE	0.85	0.75
MSNR + MSE + EMR	0.89	0.80

Seven separate models were constructed, based on, 1) multiscale entropy (MSE) features calculated from the HR and MAP time series, 2) EMR features including patient demographics, and other features described in Section 2.3.5, 3) features extracted from Multiscale Network Representation ($MSNR$), 4) combining the EMR features and Entropy features ($MSE + EMR$), 5) combining $MSNR$ and EMR features, 6) combining

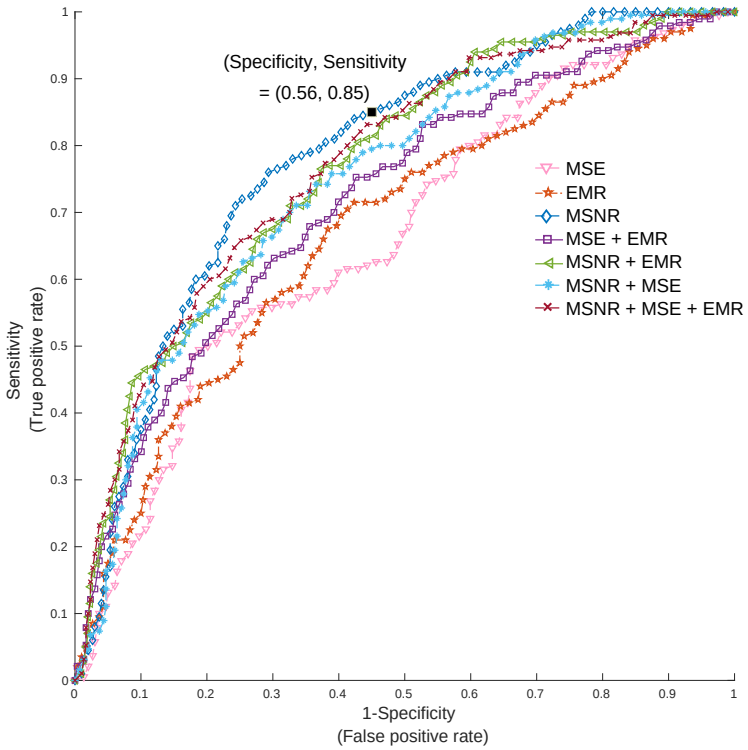


Figure 2.4: **ROC curves for models based on combinations of network, entropy and EMR features.** For the model corresponding to $MSNR$ + MSE + EMR features, the AUROC on test set was 0.80, with a specificity of 0.57 at 0.85 sensitivity level. Notably, $MSNR$ features alone achieved an AUC of 0.78, with the corresponding sensitivity (0.85) and specificity (0.56) marker on the plot

$MSNR$ and MSE features, and 7) combining the EMR, $MSNR$ and MSE features. The performance of each of the above models have been tabulated in Table 2.2. The model based on $MSNR$ features alone achieved a pooled testing AUROC of 0.78, with a corresponding sensitivity of 0.85 and specificity of 0.56. Combining the MSE features and $MSNR$ features did not result in any improvement of AUROC (statistically insignificant). Combining EMR features and the $MSNR$ features resulted in an improvement in AUC from 0.78 to 0.79 (statistically significant). For the model corresponding to $MSNR$ + MSE + EMR features, the pooled AUROC on test set was 0.80 (statistically significant), with a specificity of 0.57 at 0.85 sensitivity level. The Receiver Operating Characteristic (ROC) curves for the above models have been plotted in Figure 2.4.

2.6 Discussion

We have shown that features derived from a multiscale HR and MAP time series network provide approximately 20% improvement in the area under the receiver operating characteristic (AUROC) for four hours ahead prediction of sepsis over traditional indices of heart rate entropy. This improvement is attributable to the information embedded in the higher order interaction of HR and MAP time series, as well as the proposed novel approach to network construction that utilizes adaptive partitioning of the state-space to define a set of discrete states. This discretization method naturally trades off uncertainty in defining an event (a unique state) for a more accurate estimation of the probability of the event. The resulting algorithm is quick to implement and readily extensible to multiscale analysis of the time series networks. Our final model, which includes the most commonly available clinical measurements in patients' electronic medical record (EMR), multiscale entropy features as well as the proposed network-based features, achieved an AUROC of 0.80 on the testing set.

The proposed network construction technique takes advantage of the fact that the mutual information between a set of random variables is invariant to invertible transformations such as the rank order transformation (replacing the data by their ranks). The rank order transformation makes the proposed technique robust to time series outliers samples with high amplitudes. Moreover, time-lagged embedding provides information on the underlying dynamical system without having direct access to all the system variables [66]. By applying the DV partitioning algorithm on the space of time-lagged embedded HR and MAP time series we arrive at states that capture the nonlinear dynamics of HR and MAP. Similar to the method of variable-bandwidth kernel density estimation [70], the DV partitioning algorithm automatically adjusts the bin size (hypercubes), depending on the density and local distribution of the data points, but requires no *a priori* assumption on the Kernel bandwidth and is computationally more efficient to evaluate [68].

Some of the most important features including the average clustering coefficient are reflective of modularity of the network; networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Further study is needed to assess the correlation between the network features considered in this work and other commonly used predictive features within the literature. However, we hypothesize that the proposed framework provides a more generalizable set of features that are highly descriptive of the break down in autoregulatory mechanisms, and predictive of the eventual physiological decompensations, as in the case of sepsis. Notably, the multiscale nature of the proposed features provides robustness to the varying durations and time-scales of physiological deterioration in critically care patients.

Many methods have been proposed in the literature to study human physiology as a complex network of interactions among body organs and processes. Much of the effort have been concentrated on identification and quantification of the interactions between these physiological processes [74]. Bashan et al. [75] proposed the concept of time delay stability (TDS) to quantify the dynamic interactions among physiological processes, such as sleep and cardio-respiratory coupling. Building upon the concept of TDS, interactions across time scales and frequency bands have been explored to reveal dynamic interactions across body organs [76, 77, 78]. Utilizing the concept of "information dynamics", entropy-based approaches have been proposed to quantify the information transfer between physiological processes [79, 68]. Our proposed MSNR approach complements other pioneering works in "Network Physiology" by introducing a non-parametric approach to partitioning the state-space, and taking advantage of network analysis to quantify the non-linear interactions among multiple physiological time series.

Clinical decision support tools can help identify those at the highest risk for future sepsis. Although, the existing works on utilizing EMR and laboratory data for prediction of

sepsis seem promising [27, 80, 23], they are limited by low-frequency, and often inconsistent data collected for purposes other than timely and accurate representation of patients' physiology. Highly predictive features extracted directly extracted from the high-resolution vital signs time series can improve sepsis prediction over low-resolution clinical data in the ICU patients, and a high-performance prediction model can be derived from a combination of EMR and high-frequency physiologic data. A real-time system capable of early prediction of sepsis, followed by appropriate antibiotics therapy, will have a significant impact on the overall mortality and cost burden of this deadly disease [14].

CHAPTER 3

DEEPAISE - AN END-TO-END DEVELOPMENT AND DEPLOYMENT OF A RECURRENT NEURAL SURVIVAL MODEL FOR EARLY PREDICTION OF SEPSIS

The part I of this chapter discusses DeepAISE (Deep Artificial Intelligence Sepsis Expert) and presents results showing its performance on ICU cohorts from three different academic medical centers in US. The part II of this chapter extends the analysis of DeepAISE to study the interpretability of its predictions and its integration into clinical workflow.

3.1 PART I: Developing a sequential model architecture for early prediction of sepsis

In recent years, the increased adoption of electronic medical records (EHRs) in hospitals has motivated the development of machine learning based surveillance tools for detection or classification [26, 23, 81, 82, 27] and prediction [23, 25, 31] of patients with sepsis or septic shock. For prediction of sepsis in particular, Desautels et al. [23] used a proprietary machine learning algorithm called InSight to achieve an Area Under the Curve (AUC) of 0.78 in predicting onset of sepsis four hours in advance. Lukaszewski et al [27] demonstrated that a neural network using only cytokine data predicted sepsis better than a similar algorithm using clinical EMR data. However, cytokines are not routinely measured, making it an impractical tool for contemporary practice. Nemati et al. [25] used a modified Weibull-Cox model on a combination of low-resolution Electronic Medical Record (EHR) data and high-resolution vital signs time series data to predict onset of sepsis four hours in advance with an AUC of 0.85. Additionally, Futoma et al. [26] proposed a multiple-output gaussian process based recurrent neural network models for classifying patient encounters as being septic or not. Other works have focused on developing models to predict septic shock, which occurs when sepsis leads to low blood pressure that persists despite treat-

ment with intravenous fluids. In particular, Henry et al. [31] used the cox proportional hazards model to predict the onset of septic shock in patients admitted to the ICU. They developed their model based on the publicly available MIMIC-II clinical database. A direct comparison of these methods is not possible for several reasons: 1) utilization of different labels for sepsis and septic shock, 2) variations in prediction horizon (finite horizon prediction vs infinite horizon prediction, 3) differences in frequency of prediction (single event classification vs sequential prediction), and 4) variations in study design and disease prevalence (case-control design vs calibrated real-world prevalence models). To date most sepsis prediction research has failed to make the transition into viable Clinical Decision Support (CDS) systems owing to the relatively low clinical tolerance for false-alarms [83], as well as the interpretability and workflow integration requirements for CDS systems [84, 85]. False clinical alarms not only increase the cognitive load on clinicians but can also expose patients to unnecessary antibiotics and may contribute to emergence of antibiotic resistance pathogens [86]. Nevertheless, identifying and treating true cases of sepsis before they are clinically apparent is categorically one of the most important needs for modern medicine to address.

The primary contribution of this work is a deep learning framework for prediction of sepsis (called DeepAISE) that reduces incidents of false alarms by automatically learning predictive features related to higher-order interactions and temporal patterns among clinical risk factors for sepsis. Unlike comparable models, this algorithm maintains interpretability by tracking the top relevant features contributing to the sepsis score as a function of time, providing clinicians with rationale for alerts. Most importantly, DeepAISE is an externally evaluated sepsis model developed using over 25,000 patient admissions to the Intensive Care Units (ICUs) at two Emory University hospitals, over 18,000 ICU admissions to the UC San Diego Health system and over 40,000 ICU admissions from the Medical Information Mart for Intensive Care-III (MIMIC-III) ICU database.

3.1.1 A multicenter dataset of critically ill patients in the ICU

We performed a retrospective study of all patients admitted to the ICUs at two hospitals within the Emory Healthcare system in Atlanta, Georgia from 2014 to 2018. This investigation was conducted according to Emory University Institutional Review Board (IRB) approved protocol #33069 and the UCSD IRB approved protocol #191098X. External evaluation of the DeepAISE algorithm was performed on two separate cohorts: 1) the UCSD cohort, which consisted of all patients admitted to the ICUs at two hospitals within the UC San Diego Health system in San Diego, California from 2016 to 2019, and 2) the Medical Information Mart for Intensive Care-III (MIMIC-III) ICU database [87], which is a publicly available database containing de-identified clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts from June 2001 to October 2012. Patients 18 years or older were followed throughout their ICU stay until discharge or development of sepsis, according to Sepsis-3 guidelines [1, 16]. During the external evaluation step, the DeepAISE model (trained on the Emory cohort) was fine-tuned and tested on the UCSD and MIMIC-III cohorts separately, and a comparison with baseline models was performed.

For the Emory cohort, data from the EHR (Cerner, Kansas City, MO) were extracted through a clinical data warehouse (MicroStrategy, Tysons Corner, VA). High-resolution heart rate and Mean arterial pressure time series at 2 seconds resolution were collected from select ICUs, through the BedMaster system (Excel Medical Electronics, Jupiter, FL), which is a third-party software connected to the hospital's General Electric monitors for the purpose of electronic data extraction and storage of high-resolution waveforms. Patients were excluded if they developed sepsis within or prior to the first 4 hours of ICU admission (by analyzing pre-ICU IV antibiotic administration and culture acquisition) or if their length of ICU stay was less than 8 hours or more than 20 days.

The Emory cohort contained a total of 25,820 patients, 1,445 of whom met the Sepsis-3 criterion four hours or later after ICU admission. Out of the 25,820 patients, 70% of them

were used for developing the model (training set), 10% were used for hyper-parameter optimization, and the remaining 20% formed the testing set (see Table 3.1 for a description of the various holdout datasets that have been used for analysis in this chapter). The Emory training set contained a total of 18,074 patients out of which 1,003 patients met the Sepsis-3 criterion, and the Emory testing set contained a total of 5,165 patients out of which 287 patients met the Sepsis-3 criterion during their stay in the ICU. Those who developed sepsis tended to have a slightly higher percentage of male patients compared to non-septic patients (55.2% vs. 53.2%) and had more comorbidities (Charlson Comorbidity Index [CCI] 3 vs. 2). Septic patients had longer median lengths of ICU stay (5.9 vs. 1.9 days), higher median SOFA scores (5.0 vs. 1.7), and higher hospital mortality (15.2% vs. 3.5%). The median [interquartile range (IQR)] time from ICU admission to $t_{sepsis-3}$ in the Emory cohort was 24 [9, 63] hours. Similar patterns were observed for the UCSD cohort (Table 3.6) and the MIMIC-III cohort (Table 3.7).

The UCSD cohort contained a total of 18,752 patients, 1073 of whom met the Sepsis-3 criterion four hours or later after ICU admission. Out of the 18,752 patients, 80% of them were used for developing the model (training set), and the remaining 20% formed the testing set. The MIMIC-III cohort contained a total of 40,474 patients, 2276 of whom met the Sepsis-3 criterion four hours or later after ICU admission. Out of the 40,474 patients, 80% of them were used for developing the model (training set), and the remaining 20% form the testing set. The patient characteristics of Emory, UCSD and MIMIC-III cohorts are tabulated in Tables 3.5, 3.6 and 3.7.

The complete set of patient features (65 in total, see Section 3.4) was grouped into three categories: clinical features (e.g. heart rate, mean arterial pressure, etc.), laboratory test results (e.g. hemoglobin, creatinine, etc.) and demographic/history/context features (e.g. age, care unit type, etc.). Some of the clinical or laboratory features that were unavailable in the UCSD and MIMIC-III cohorts were treated as 'missing' features during fine-tuning of DeepAISE.

Table 3.1: Description of the various datasets used in the analysis of DeepAISE

Dataset	Description
Emory training	70% of patients from the entire Emory cohort
Emory testing	20% of patients from the entire Emory cohort
Emory hyperparameter optimization	10% of patients from the entire Emory cohort
Emory year-based training	Patients in Emory cohort from the year 2014 through 2017
Emory year-based holdout	Patients in Emory cohort from the year 2017 through 2018
MIMIC training	80% of patients from the entire MIMIC-III cohort
MIMIC testing	20% of patients from the entire MIMIC-III cohort
UCSD training	80% of patients from the entire UCSD cohort
UCSD testing	20% of patients from the entire UCSD cohort

3.1.2 Development of the DeepAISE model

DeepAISE began producing scores four hours after ICU admission, and was designed to predict (on an hourly basis) the probability of onset of sepsis within the next four hours. The two distinct characteristics of the model were: 1) utilization of a class of deep learning algorithms for multivariate time series data known as the Gated Recurrent Unit (GRU) [88] that allowed for modeling the clinical trajectory of a patient over time, and 2) deployment of a parametric survival model called the Weibull Cox proportional hazards (WCPH) [89], which cast the problem of sepsis prediction to a time-to-event prediction framework and allows for handling of right censored outcomes [25], using features learned from the underlying GRU model. The parametric survival model allowed for efficient end-to-end learning of the GRU and the WCPH parameters using standard deep learning optimization techniques.

For each patient admitted to the Intensive Care Unit (ICU), the goal of the proposed DeepAISE model was to predict (at a regular interval of 1 hour) the probability of onset of sepsis, using all data available for the patient up until the time of prediction. Figure 3.1 provides an overview of the proposed DeepAISE model.

The notations that are followed throughout the rest of the chapter is described as follows: For a total of N patients admitted to the ICU, we considered a dataset $D = \{D_i\}_{i=1}^{N}$

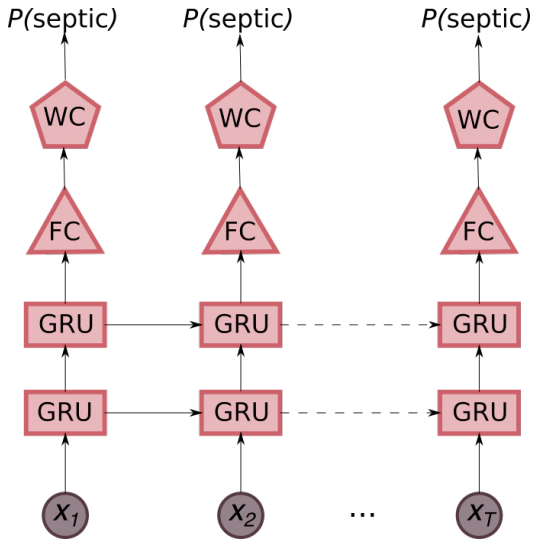


Figure 3.1: **Schematic diagram of the Deep Artificial Intelligence Sepsis Expert (DeepAISE) model.** The 65 features that are measure/computed every hour are fed sequentially into a 2 layer stacked GRU framework, the output from the stacked GRU layer is then fed into a fully connected layer, and a modified Weibull Cox Proportional Hazards Model (WCPH) is employed to compute the probability of occurrence of sepsis within the proceeding m hours (denoted by $F_t(m)$, with $t = [1, 2, \ldots T]$). In our work, we are interested in the prediction of onset of sepsis 4 hours in advance.

with $D_i = \{ \mathbf{X}_i, T_i, \mathbf{e}_i, \mathcal{T}_i \}$. The length of time series for patient i is denoted by T_i. A total of 65 features are measured/computed every hour for every patient in the ICU. The features for patient i is represented by $\mathbf{X}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}]$, where $\mathbf{x}_{it} \in \mathbb{R}^{65}$ is the feature vector at time step t. The sepsis event indicator for patient i is denoted by $\mathbf{e}_i = [e_{i1}, e_{i2}, \ldots, e_{iT_i}]$ where at time step t, $e_{it} = 0$ if onset of sepsis does not occur within the prediction horizon otherwise $e_{ij} = 1$. $\mathcal{T}_i = [\tau_{i1}, \tau_{i2}, \ldots, \tau_{iT_i}]$ represents the time to sepsis event for patient i.

We considered the prediction of onset of sepsis as a sequential prediction task in our study. To achieve this, the proposed DeepAISE model, shown in Figure 3.1, employed a combination of a 2 layer stacked Gated Recurrent Unit (GRU) framework and a modified weibull-cox proportional hazards model (WCPH) to predict the onset of sepsis at a regular interval of 1 hour.

Let us consider a sequence of data of length T_i belonging to patient i. At each timestep t the stacked GRU model takes in as input, the feature vector $\mathbf{x}_{it}$ and stores the temporal

information within its hidden layers.

The GRU model used in our study is composed of 3 components at every timestep t: the $reset$ gate $\mathbf{r}_t$, the $update$ gate $\mathbf{z}_t$, and the hidden layer $\mathbf{h}_t$. The hidden layer $\mathbf{h}_t$ is computed as follows :

$$\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z) \\
\mathbf{r}_t &= \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r) \\
\widetilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h\mathbf{x}_t + \mathbf{r}_t \odot \mathbf{U}_h\mathbf{h}_{t-1} + \mathbf{b}_h) \\
\mathbf{h}_t &= \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \widetilde{\mathbf{h}}_t
\end{aligned} \tag{3.1}$$

where $\sigma()$ is the logistic sigmoid function, $\odot$ denotes element-wise multiplication (or Hadamard product) and $\mathbf{W}_{\{z,r,h\}}$, $\mathbf{U}_{\{z,r,h\}}$, $\mathbf{b}_{\{z,r,h\}}$ are the weight matrices and bias terms associated with the calculation of $update$ gate, $reset$ gate, and hidden unit respectively. In this formulation of the GRU, if the $reset$ gate is close to 0 the intermediate hidden layer $\widetilde{\mathbf{h}}_t$ will ignore the previous hidden state $\mathbf{h}_{t-1}$ and reset with the current input $\mathbf{x}_t$, essentially allowing the hidden state to ignore any information that would be irrelevant later. The $update$ gate controls the extent of information carried over from the previous hidden state $\mathbf{h}_{t-1}$ to the current hidden state $\mathbf{h}_t$. This behavior of the GRU helps the DeepAISE model to remember long-term dependencies that are present in the sequential data, and aids the model in identifying and capturing information that is only necessary for prediction of onset of sepsis.

In our proposed DeepAISE model, we stack 2 layers of GRU on top of each other to increase the representational power of the model. We represent the GRUs in the stacked layer as $\mathrm{G}^{(1)}$ and $\mathrm{G}^{(2)}$, with $\mathrm{G}^{(l)} = \{\mathbf{W}^{(l)}_{\{z,r,h\}}, \mathbf{U}^{(l)}_{\{z,r,h\}}, \mathbf{b}^{(l)}_{\{z,r,h\}}\}\ \forall\ l = 1, 2$. Correspondingly $\mathbf{r}^{(l)}_t$, $\mathbf{z}^{(l)}_t$, and $\mathbf{h}^{(l)}_t$ would be the output of $reset$ gate, $update$ gate and the hidden state of the GRU in layer l at timestep t.

The output $\mathbf{h}^{(2)}_t$ from the stacked GRU layer is then fed into a fully connected layer before being fed into the modified Weibull-cox proportional hazards model, for predicting

onset of sepsis in the proceeding m hours (where m = 2, 4, 6, 8, 10 or 12 hours). The weibull-cox proportional hazards is a more robust parametric counterpart to the more familiar cox proportional hazards model [89] . The Weibull-Cox model defines the baseline hazard function as $H_0(m|\lambda,\nu) = (\nu/\lambda)(m/\lambda)^{\nu-1}$, where $\lambda > 0$ is a scale parameter and $\nu > 0$ is a shape parameter. For a patient i, the hazard function $H(m)$ at time step t is then defined as -

$$H_{it}(m|x_{it},\theta,\beta,\lambda,\nu) = H_0(m|\lambda,\nu)\exp(\beta^T f(x_{it})) \tag{3.2}$$

where $f()$ is the output from the fully connected layer in the deep learning pipeline, β is a L dimensional weight vector ($\beta \in \mathbb{R}^L$), and θ = $\{\mathrm{G}^{(1)}, \mathrm{G}^{(2)}, \mathbf{W}_{fc}, \mathbf{b}_{fc}\}$. $\mathbf{W}_{fc}$ and $\mathbf{b}_{fc}$ denote the weights and bias term of the fully connected layer respectively. The survival function (i.e. probability of not getting sepsis in the proceeding m hours from current time step t) is given by -

$$S_t(m|x_{it},\theta,\beta,\lambda,\nu) = \exp(-\Lambda_0(m)\exp(\beta^T f(x_{it})) \tag{3.3}$$

where $\Lambda_0(m) = (m/\lambda)^{\nu}$ is the cumulative base hazard rate. The probability that onset of sepsis occurs with the proceeding m hours is then given by $F_t(m) = 1 - S_t(m)$.

Learning model parameters

Given the dataset D, we would like to compute the posterior probability over the parameters of the model which is defined as $p(\theta,\beta,\lambda,\nu|D) \propto p(D|\theta,\beta,\lambda,\nu)p(\theta)p(\beta)p(\lambda)p(\nu)$. We assume that $p(\theta)$, $p(\beta)$, $p(\lambda)$ and $p(\nu)$ are constant, and therefore maximization of the posterior probability is nothing but maximization of the likelihood of the data. The parameters of the proposed model is then learned through the maximum likelihood approach, wherein the log likelihood of the data is maximized (or the negative log likelihood of the

data is minimized). The data likelihood is given in Equation 3.4.

$$
\begin{aligned}
P(D|\theta,\beta,\lambda,\nu) &= \prod_{i=1}^{N}\prod_{t=1}^{T_i}[H_0(\tau_{it}|\lambda,\nu)\exp(\beta^T f(x_{it}))]^{e_{it}} S(\tau_{it}|x_{it},\lambda,\nu,\beta,\theta) \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T_i}[H_0(\tau_{it}|\lambda,\nu)\exp(\beta^T f(x_{it}))]^{e_{it}} \exp(-\Lambda_0(\tau_{it})\exp(\beta^T f(x_{it}))
\end{aligned}
\tag{3.4}
$$

where for patient i at time step t, $\mathbf{x}_{it} \in \mathbb{R}^{65}$ is the feature vector, e_{it} is the sepsis event indicator, and τ_{it} represents the time to sepsis event.

Further, the negative log-likelihood of data is then denoted by -

$$
\mathcal{L}(\theta,\beta,\lambda,\nu) = -\frac{1}{N}\log P(D|\theta,\beta,\lambda,\nu) \tag{3.5}
$$

We then follow a mini-batch stochastic gradient descent approach to learn the optimal parameters of the model, by minimizing $\mathcal{L}(\theta,\beta,\lambda,\nu)$. Intuitively, maximizing the data likelihood (or minimizing the negative log-likelihood) will correspond to: 1) maximizing the probability that sepsis does not occur before time τ_{it} and 2) maximizing the probability of actual sepsis events, when the events are not censored (i.e. $e_{it} = 1$).

3.1.3 Data processing, model evaluation and statistical analysis

First, features in the Emory training set were normalized by subtracting the mean and dividing by the standard deviation (both of which were computed on the Emory training set). Next, all the remaining datasets were normalized using the mean and standard deviation computed from the Emory training set. For handling missing data, we used a simple sample-and-hold approach in all the datasets. For all continuous variables, we have reported median ([25th - 75th percentile]). For binary variables, we have reported percentages. The area under receiver operating characteristic (AUC) curves statistics, specificity

(1-false alarm rate) and accuracy at a fixed 85% sensitivity level were calculated to measure the performance of the models. We have reported the DeepAISE performance results of four hours ahead prediction on the training and testing sets of the Emory cohort. External evaluation of DeepAISE was performed on the UCSD and MIMIC-III cohorts separately. During the external evaluation step, the DeepAISE model was fine-tuned on the training set of each evaluation cohort and was evaluated on the corresponding test set. Additionally, we have also reported the performance results for 2, 4, 6, 8, 10 and 12 hours ahead prediction of onset of sepsis. Statistical comparison of all AUC curves was performed using the method of DeLong et al [90].

3.1.4 Hyperparameter optimization

We trained each model for a total of 200 epochs using the Adam optimizer [91], with a learning rate fixed at 1e-2. The mini-batch size was fixed at a total of 1,000 patients (90% control patients, 10% septic patients), with data randomly sampled (with replacement) in every epoch. To minimize overfitting and to improve the generalizability of the model, L1-L2 regularization was used with L2 regularization parameter set to 1e-3 and L1 regularization parameter set to 1e-5. Our final model had 2 GRU layers stacked on top of each other with the size of hidden state being 100 per layer, followed by 1 fully connected layer with the size of the hidden state being 100, and the output of which was fed into a modified Weibull-Cox proportional hazards model for prediction of onset of sepsis. All of the hyper-parameters of the model: Number of GRU layers, the size of hidden state in each of GRU layers, Number of fully connected layers, the size of hidden state in each of the fully connected layers, learning rate, mini-batch size, L1 regularization parameter, and L2 regularization parameter were optimized using Bayesian optimization [92]. All pre-processing of the data was performed using Numpy [93], with the rest of the pipeline implemented using TensorFlow [94].

3.1.5 Results

DeepAISE prediction performance for sepsis onset

The DeepAISE model was trained to predict the early onset of sepsis ($t_{sepsis-3}$). DeepAISE made hourly predictions, starting four hours after ICU admission, and considered a total of 65 features that were commonly available in the EHR. The Emory training and testing sets contained a total of approximately 500,000 and 125,000 hourly prediction windows, respectively. External evaluation of the DeepAISE algorithm was performed on two separate cohorts: 1) the UCSD cohort, and 2) MIMIC-III cohort. During the external evaluation step, the DeepAISE model (trained on the Emory cohort) was fine-tuned and tested on the UCSD and MIMIC-III cohorts separately, and a comparison with baseline models was performed.

Table 3.2: Summary of DeepAISE performance for prediction horizons of 4, 6 and 12 hours.

Performance metric	**4 hours** *testing(training)*	**6 hours** *testing(training)*	**12 hours** *testing(training)*
Emory cohort			
AUC*	0.90 (0.94)	0.89 (0.90)	0.88 (0.89)
Specificity	0.80 (0.89)	0.78 (0.84)	0.73 (0.78)
UCSD cohort^{+}			
AUC*	0.88 (0.90)	0.87 (0.89)	0.85 (0.86)
Specificity	0.77 (0.78)	0.74 (0.77)	0.68 (0.70)
MIMIC-III cohort$^{\#}$			
AUC*	0.87 (0.90)	0.86 (0.87)	0.83 (0.86)
Specificity	0.75 (0.78)	0.72 (0.75)	0.69 (0.73)

* *AUC = Aread Under the Curve; Sensitivity was fixed at 0.85*

$^{+}$ *DeepAISE model (trained on Emory cohort) fine-tuned to the UCSD cohort*

$^{\#}$ *DeepAISE model (trained on Emory cohort) fine-tuned to the MIMIC-III cohort*

The DeepAISE model reliably predicted $t_{sepsis-3}$ four hours in advance with an AUC of 0.90 (specificity of 0.80 at sensitivity of 0.85) on the Emory testing set. Slightly lower performances were observed for the UCSD and MIMIC testing sets, with the fine-tuned models achieving an AUC of 0.88 and 0.87 respectively for predicting $t_{sepsis-3}$ four hours

in advance (see Table 3.2 for more details).

To assess the impact of changes in institutional practices and patient populations over time (*temporal validation*) we performed an experiment in which a model trained on Emory year-based training set (patients in Emory cohort from the year 2014 through 2017) was applied to a holdout test set collected from 2017 to 2018 (Emory year-based holdout set). The DeepAISE model achieved an AUC of 0.88 (Specificity of 0.75 at sensitivity of 0.85) on the Emory year-based holdout set (see Figure 3.2 for more details), which was comparable to the model performance on external evaluation cohorts.

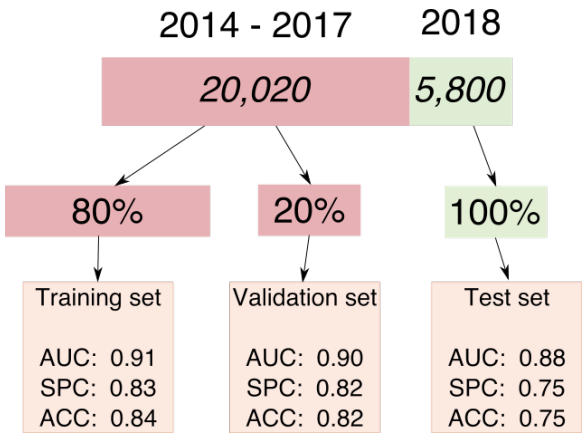


Figure 3.2: **Performance of DeepAISE that was first trained on Emory year-based training set (patients in Emory cohort from the year 2014 through 2017) and then applied to a heldout test set collected from 2017 to 2018 (Emory year-based holdout set).**

DeepAISE performance against other baselines

The DeepAISE model is a composite model made of a class of Recurrent Neural Network (RNN) models known as Gated Recurrent Units (GRUs) [88] that feeds into a Weibull Cox Proportional Hazards (WCPH) model. This architecture was chosen in the context of predicting sepsis onset time as a time-to-event analysis and considering that temporal changes in patients' physiology are important for the early prediction of sepsis. We assessed the utility of this model architecture by comparing performance of DeepAISE against three different baseline models: 1) a Logistic Regression (LR) model, 2) a Weibull-Cox propor-

tional hazards (WCPH) model, and 3) a Feedforward Neural Network (FFNN) with two layers of 100 hidden units and a final WCPH layer for prediction of onset of sepsis (See Figure 3.3). A more comprehensive comparison of DeepAISE with other baseline models can be found in Figure 3.16.

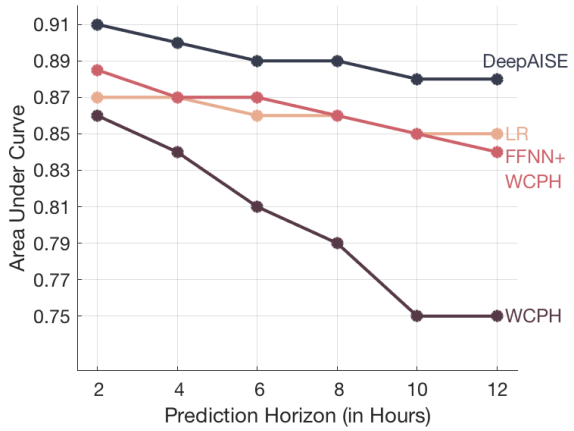
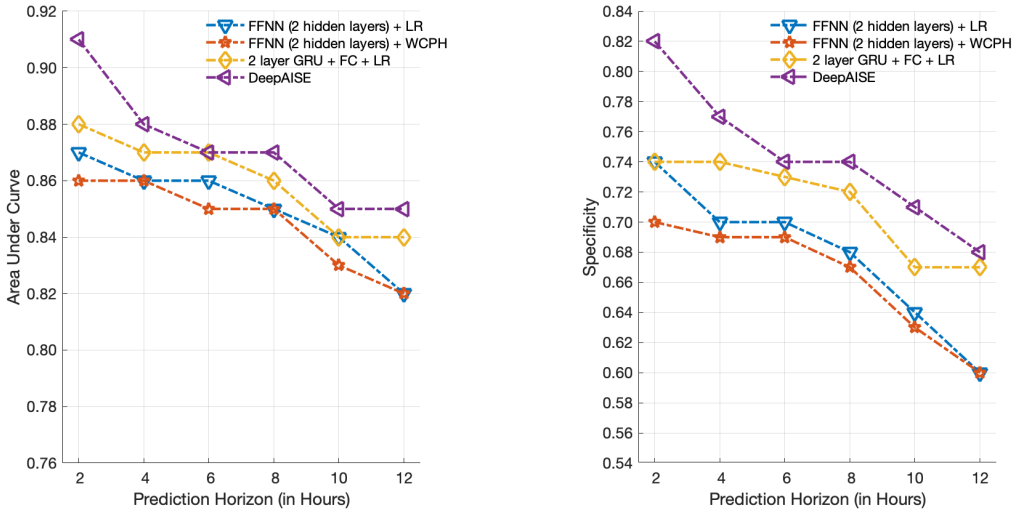


Figure 3.3: **Comparison of DeepAISE performance on Emory testing set for prediction horizons of 2, 4, 6, 8, 10 and 12 hours.**

Across all prediction windows, DeepAISE consistently outperformed all other baseline classifiers ($p < 0.001$; when AUC of DeepAISE was compared with AUC of other baseline methods) for prediction of $t_{sepsis-3}$ (See Figure 3.3, Figure 3.16 and Table 3.4) and across all prediction windows, indicating that capturing temporal trends and interactions among the risk factors is important for accurate prediction of sepsis. The performance of all the models decreased with the increase in prediction horizon. For DeepAISE, the AUC on the Emory testing set decreased from 0.90 at 4-hour prediction horizon to 0.88 at 12-hour prediction horizon. We also observed that these findings were consistent with the performance of the fine-tuned model on the UCSD and MIMIC testing sets (Figure 3.4 and Figure 3.5).

In addition, a FFNN trained to predict $t_{sepsis-3}$ with delta change in SOFA score as input achieved 0.54 AUC on Emory testing set, and a FFNN trained to predict $t_{sepsis-3}$ with delta change in SOFA score and static covariates (such as age, gender, weight etc.) as inputs achieved 0.68 AUC on the Emory testing set. The above results show that DeepAISE scores were not simply recapitulations of the SOFA scores.

(a) (b)



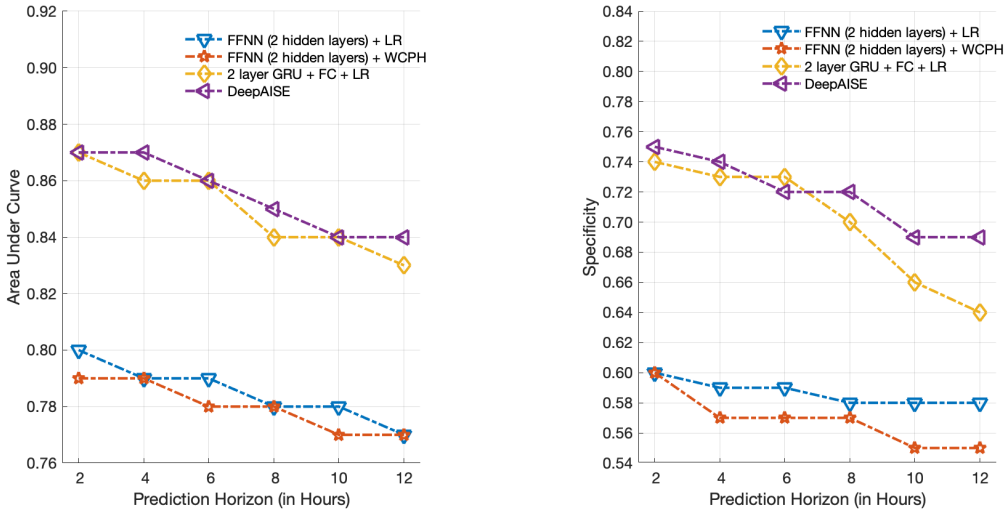
** LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figure 3.4: **Comparison of performance of baseline models and DeepAISE on the UCSD cohort to predict $t_{sepsis-3}$ for prediction horizons of 2, 4, 6, 8, 10, and 12 hours.** a) The Area Under the Curve (AUC) is shown in the left panel. b) The Specificity (SPC) is shown in the right panel.

Performance of DeepAISE under different levels of missingness

We evaluated the performance of DeepAISE algorithm (when fine tuned to a new cohort) under different levels of missingness of input features (specifically laboratory measurements). For every patient, we computed the percentage of missing laboratory measurements (the percentage of missing measurements was computed over a rolling 24 hour window, and were then averaged across all the windows). This number represented the percentage of laboratory measurements missing on average for a patient over a 24 hour window. The patients were then split into 3 different groups based on the percentiles of their missingness. Group 1 consisted of patients whose percentage of missingness fell below 33 percentile of the overall cohort. Group 2 consisted of patients whose percentage of missingness was above 33 percentile and below 66 percentile of the overall cohort. Group 3 consisted of patients whose percentage of missingness was above 66 percentile and below

(a) (b)



** LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figure 3.5: **Comparison of performance of baseline models and DeepAISE on the MIMIC-III cohort to predict $t_{sepsis-3}$ for prediction horizons of 2, 4, 6, 8, 10, and 12 hours.** a) The Area Under the Curve (AUC) is shown in the left panel. b) The Specificity (SPC) is shown in the right panel.

100 percentile of the overall cohort. The performance of DeepAISE (on the entire UCSD cohort) for each of the above groups has been tabulated in Table 3.3. The Area under the Receiver operating characteristic curves and Area under precision recall curves for Groups 1, 2 and 3 are shown in Figure 3.7 and Figure 3.8.

Table 3.3: Performance of DeepAISE on the entire UCSD cohort for differing levels of missingness of input features. For reference, the percentage of missingness in Group 1 $<$ Group 2 $<$ Group3. (AUC: Area under the receiver operating characteristic curve. AUCpr: Area under the precision recall curve)

	Total patients	**Septic patients**	**AUC**	**AUCpr**
Group 1	6188	672 (10.85%)	0.871	0.278
Group 2	6188	561 (9.06%)	0.906	0.242
Group 3	6376	140 (2.19%)	0.916	0.179

Note: The Positive Predictive Value (or Precision) is defined as the ratio of number of true positives to the sum of the true positives and false positives. In the scenario where the class labels are highly imbalanced (in our case very low prevalence of positive labels

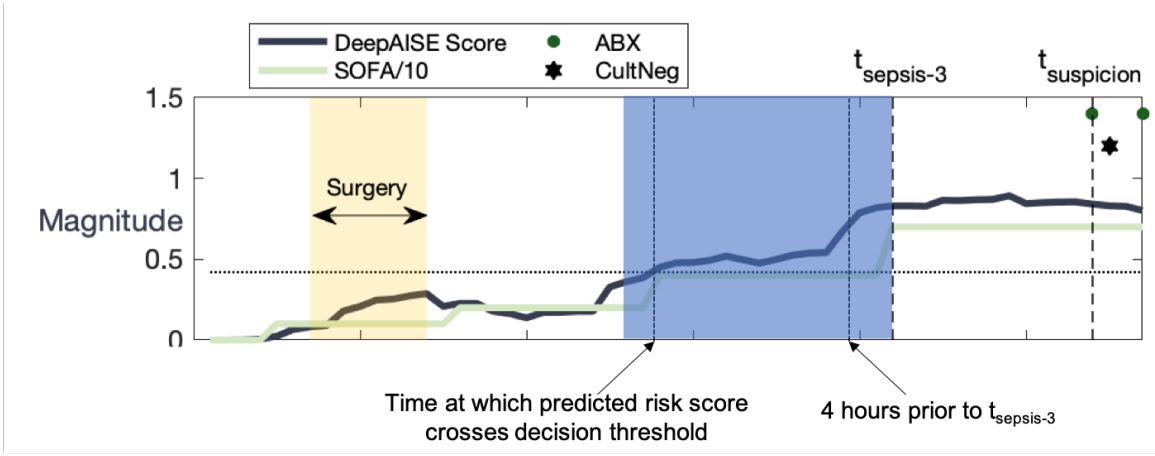


Figure 3.6: **The DeepAISE risk score crosses the decision threshold about 12 hours prior to $t_{sepsis-3}$.** In this case, according to the definition of positive predictive value all the positive predictions up until 4 hours prior to $t_{sepsis-3}$ would be counted as false positives. This is not clinically optimal, as earlier warnings are still relevant. In order to not penalize the algorithm for making positive predictions before the expected 4 hours prediction horizon, during the computation of PPV we considered any positive predictions that occurred upto 24 hours prior to $t_{sepsis-3}$ as true positives (the blue shaded region)

compared to negative labels), the positive predictive value (PPV) can get penalized by the false positives to a very large extent. It is also often the case that the predicted risk scores from a sequential prediction algorithm like DeepAISE can cross the decision threshold earlier than the 4-hours prediction horizon. For example, the DeepAISE risk score crossed the decision threshold about 12 hours prior to $t_{sepsis-3}$ for the patient shown in Figure 3.6. In this case, according to the definition of positive predictive value all the positive predictions up until 4 hours prior to $t_{sepsis-3}$ would be counted as false positives. This is not clinically optimal, as earlier warnings are still relevant. In order to not penalize the algorithm for making positive predictions before the expected 4 hours prediction horizon, during the computation of PPV we considered any positive predictions that occurred upto 24 hours prior to $t_{sepsis-3}$ as true positives (the blue shaded region in Figure 3.6). The AUCpr tabulated in Table 3.3 consists of PPV computed as described above.

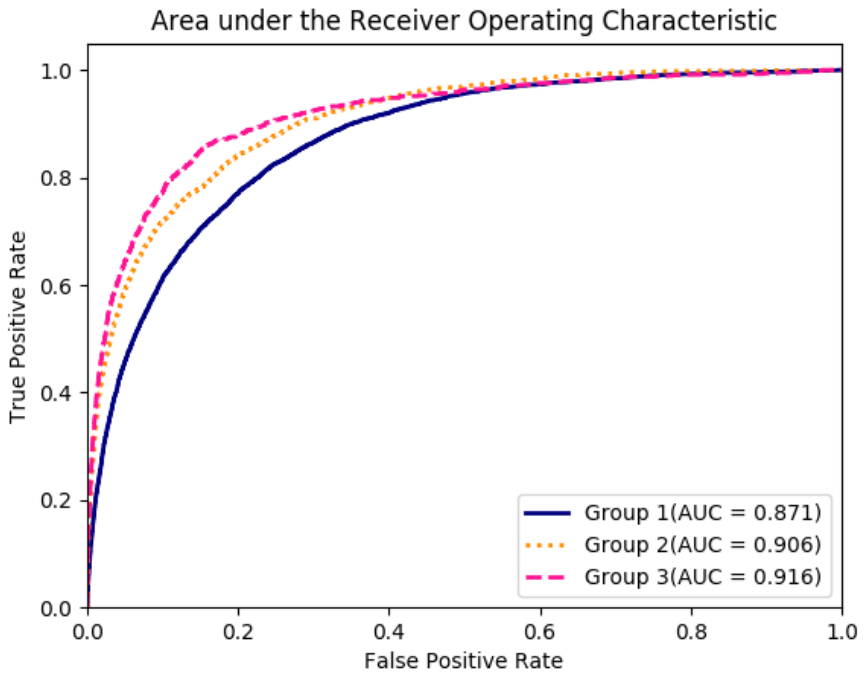


Figure 3.7: **Area under the receiver operating characteristic curves for Groups 1, 2 and 3.**

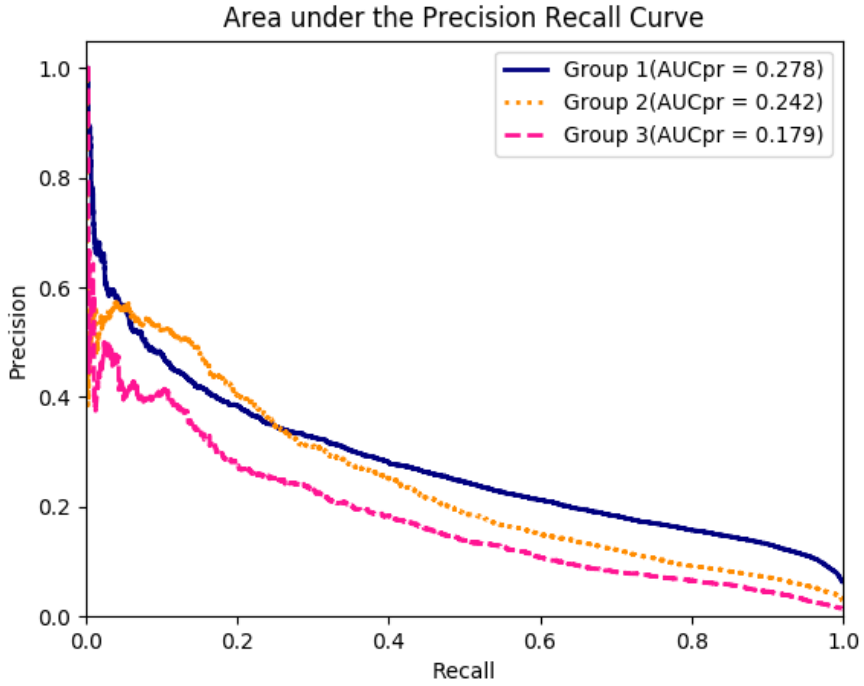


Figure 3.8: **Area under the precision recall curves for Groups 1, 2 and 3.**

3.2 PART II: Clinical interpretation of DeepAISE predictions and tele-ICU workflow integration

While performance characteristics of machine learning algorithms are important, providing interpretable data to the bedside clinicians that can guide diagnosis and therapeutic interventions is a critical requirement of CDS systems. To date many sepsis models have failed to demonstrate which physiologic aberrations contributed to the model's prediction, compelling many to refer to them as "black boxes". In this section, we focus on the clinical utility of DeepAISE by studying in detail the interpretability of its predictions. The goal of any CDS system is to improve patient outcomes and reduce hospitalization costs; however, actualization of these goals is incumbent upon clinical teams embracing and actually employing the technology. In this section, we also discuss the integration of DeepAISE into a clinical workflow system and the stakeholders involved in developing the DeepAISE UI (User Interface).

3.2.1 Testing the validity of the relevance scores and model interpretability

Unlike many other sepsis prediction algorithms, DeepAISE is uniquely interpretable wherein apart from computing the sepsis risk score, the model identifies the most relevant features contributing to the sepsis risk score as well. For a given risk score calculation the contribution of the individual input features was calculated using the associated relevance scores, by calculating the gradient of the sepsis risk score with respect to all input features and element-wise multiplication by the corresponding input features. The resultant scores, also known as the *relevance scores* were then z-scored and the top 10 features (i.e., most frequently observed features across patients and across time) with a z-score of larger than 1.96 (corresponding to a 95% confidence interval) were reported for analysis of the overall importance of input risk factors (or *global interpretability*). To test the hypothesis that the individual relevance scores capture meaningful information about the contribution of each

input feature to the risk scores, we performed a series of studies in which we systematically masked (or treated as missing data): 1) the top 10 features with the largest positive and negative relevance scores in a global sense (*global feature replacement analysis*), 2) the top 10 locally important features for each individual risk scores (*local feature replacement analysis*), and 3) a random set of 10 features at each point in time (repeated 100 times), and calculated the resulting risk scores produced by DeepAISE and the corresponding AUCs. More specifically, for the *global feature replacement analysis*, we identified 10 features that appeared the most common as a top 10 relevant feature starting 10 hours prior to and until $t_{sepsis-3}$ (two separate analysis were run for the positive relevance scores and the negative relevance scores). We then replaced these 10 global relevant features with the population mean, for the entire cohort. For the *local feature replacement analysis*, for each hour that a patient was in the ICU we computed the relevance score of all the input features, and replaced 10 of them with the highest positive (or negative) relevance score by their population mean. This analysis allowed us to systematically compare the contribution of the locally important features against the globally important predictors and against the 95-percentile of a randomly selected set of features.

3.2.2 Understanding the effect of masking important features

In a nonlinear sequential model such as DeepAISE the relationship between the input features and the model output is by no means obvious. As such, untangling this relationship requires careful analysis.

In our analysis, we were interested in understanding the importance of features that were positively and negatively contributing to the sepsis risk score. **Features with positive relevance score:** These were the features for which: 1) $\frac{dY}{dX}$ was positive and X was positive (first quadrant in Figure 3.9a) or 2) $\frac{dY}{dX}$ was negative and X was negative (second quadrant in Figure 3.9b). In both the above cases, when X is replaced with 0 (Note: All the features are normalized to a standard normal distribution, hence population mean is 0),

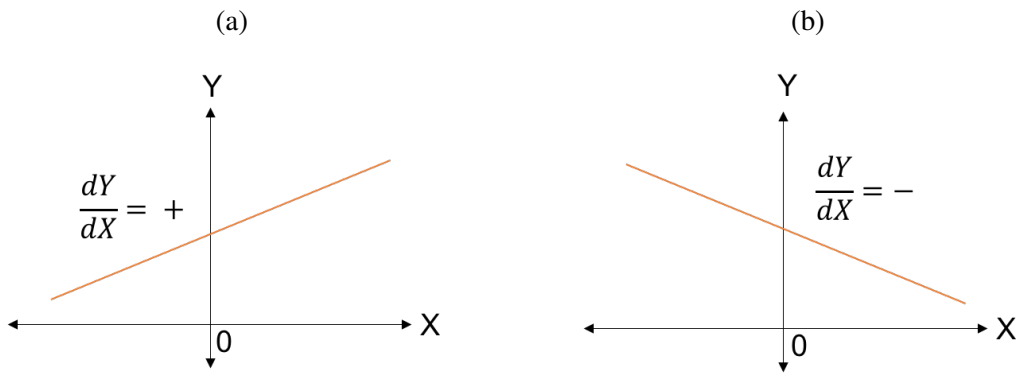


Figure 3.9: **Relationship between the output sepsis risk score and input.** a) Plot of sepsis risk score (Y) against a single variable X, when the slope is positive i.e. $\frac{dY}{dX} = +$, and b) Plot of Y against a single variable X, when the slope is negative i.e. $\frac{dY}{dX} = -$

the sepsis risk score Y decreases. Thus, when features that have a positive relevance score are replaced with the population mean, we would expect the sepsis risk score Y to drop. This should result in decreased sensitivity (reduction in true positive rate) and increased specificity (reduction in false alarm rate). **Features with negative relevance score:** These were the features for which: 1) $\frac{dY}{dX}$ was positive and X was negative (second quadrant in Figure 3.9a) or 2) $\frac{dY}{dX}$ was negative and X was positive (first quadrant in Figure 3.9b). In both the above cases, when X is replaced with 0 (Note: All the features are normalized to a standard normal distribution, hence population mean is 0), the sepsis risk score Y increases. Thus, when features that have a negative relevance score are replaced with the population mean, we would expect the sepsis risk score Y to increase. This should result in increased sensitivity (increase in true positive rate) and decreased specificity (increase in false alarm rate).

3.2.3 A case study of DeepAISE predictions

Note that the DeepAISE model used in the analysis henceforth was the model trained on Emory cohort. All the results shown are for the Emory testing set.

Unlike many other algorithms, DeepAISE is uniquely interpretable as evidenced in Figure 3.10 in which the trajectory of a septic patient who developed ventilator associated

pneumonia in the ICU is displayed. In this visualization, the sepsis risk score predicted by the model is shown along with vital sign trends, and most notably, the most relevant features contributing to the risk score. In this example, early deterioration of the patient's respiratory status was detected by the model. The model identified aberrations in PaO_2, $PaCO_2$, $bloodpH$ and Glasgow coma score (GCS) as some of the top features relevant to its prediction. The importance of each feature was calculated using the magnitude and sign of the associated relevance scores, in a fashion similar to saliency maps for convolutional neural networks [95].

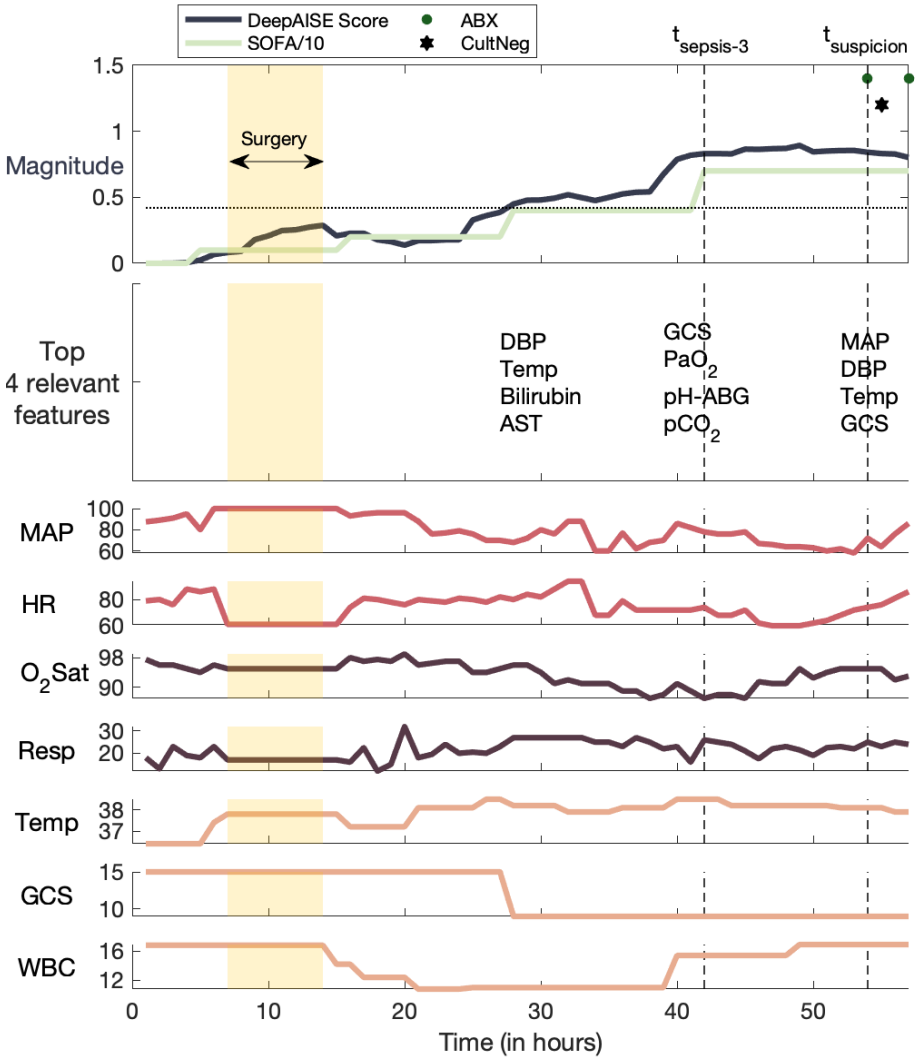


Figure 3.10: **A clinically interpretable example of DeepAISE.** The DeepAISE score is shown for a septic patient according to the Sepsis-3 guidelines. The horizontal dashed line indicates the prediction threshold corresponding to a sensitivity of 0.85. Commonly recorded hourly vital signs of the patient, including heart rate (HR), mean arterial blood pressure (MAP), respiratory rate ($RESP$), temperature ($TEMP$), oxygen saturation (O_2Sat) are shown. The most significant features contributing to the DeepAISE score are listed immediately below the DeepAISE Scores (for clarity of presentation, only selected time points are shown). The horizontal dashed line indicates the prediction threshold corresponding to a sensitivity of 0.85. Refer to Section 3.4 for more details on the abbreviated features

3.2.4 Visualizing the most relevant features for sepsis prediction

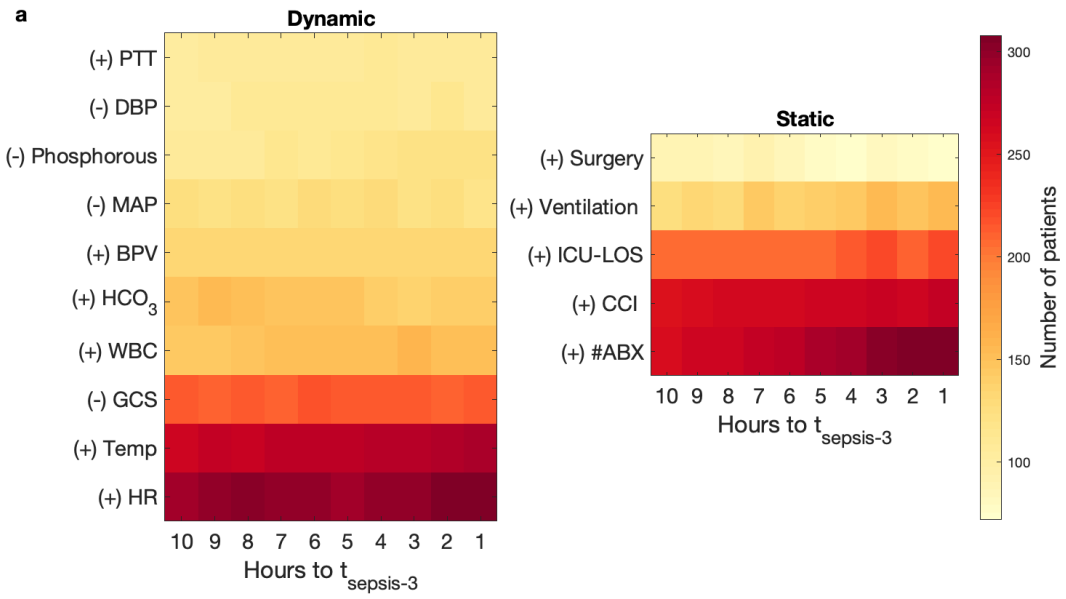


Figure 3.11: **Every hour DeepAISE identifies the top features contributing to an individual septic patient's risk score**. The left subfigure demonstrates the frequency of the top ten dynamic features (ordered according to the magnitude of the relevance score) across the septic patient population (in the Emory cohort) preceding $t_{sepsis-3}$ and the right subfigure demonstrates the frequency of the top five static features that are seen preceding $t_{sepsis-3}$. Features with positive gradient with respect to the sepsis risk score are identified by '(+)'. Features with negative gradient with respect to the sepsis risk score are identified by '(-)'.

To validate the clinical interpretability of the DeepAISE model, analysis of the most relevant features starting 10 hours prior to and ending at $t_{sepsis-3}$ was conducted (see Figure 3.11). This investigation revealed that DeepAISE ascribed importance to several features that have already been identified as risk factors for sepsis such as recent surgery, length of ICU stay, heart rate, GCS, white blood cell count, and temperature, and some less appreciated but known factors such as Phosphorus (or hypophosphatemia) [96]. This analysis provides a global view of model interpretability, whereas the individual relevance scores provide a local view of interpretability by listing the top features contributing to the risk scores for each hourly prediction window. Perturbation analysis revealed that the globally important features may not provide an accurate view of the top contributing factors to the individual risk scores. We observed that treating the top locally important features as miss-

ing values yielded a significantly lower AUC compared with a similar analysis replacing the globally most important features (See Figure 3.12).

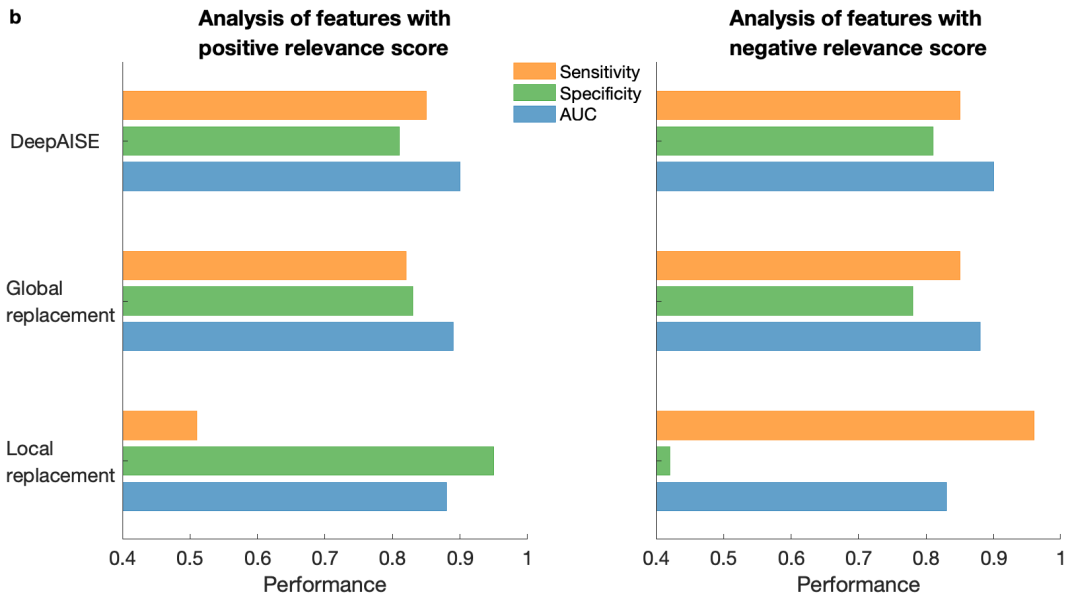


Figure 3.12: **Summary of performance of DeepAISE (on the Emory testing set) when global feature replacement analysis and local feature replacement analysis were performed for features with positive relevance score (left subfigure) and negative relavance score (right subfigure).** Note that the performance (AUC) of DeepAISE when a random set of 10 features at each point in time were masked (repeated 100 times) was 0.899 [0.886, 0.901]. The sensitivity and specificity values reported for global feature replacement analysis and local feature replacement analysis were measured at threshold corresponding to 0.85 sensitivity for the original model.

3.2.5 Inferring significance of individual patient trajectories

Clinicians have long appreciated that trends in patient metrics are often more telling than discrete point values. The high dimensional nature of the data used to represent a patient is challenging to represent. Display of patient trajectories as they pass from states of sickness to health provides yet another opportunity to inform the clinician about a patient's expected clinical course.

Each point on the manifold shown in Figure 3.13 is a 3D representation (projection) of the patient's 65 features, constructed via first learning a 100-dimensional representation

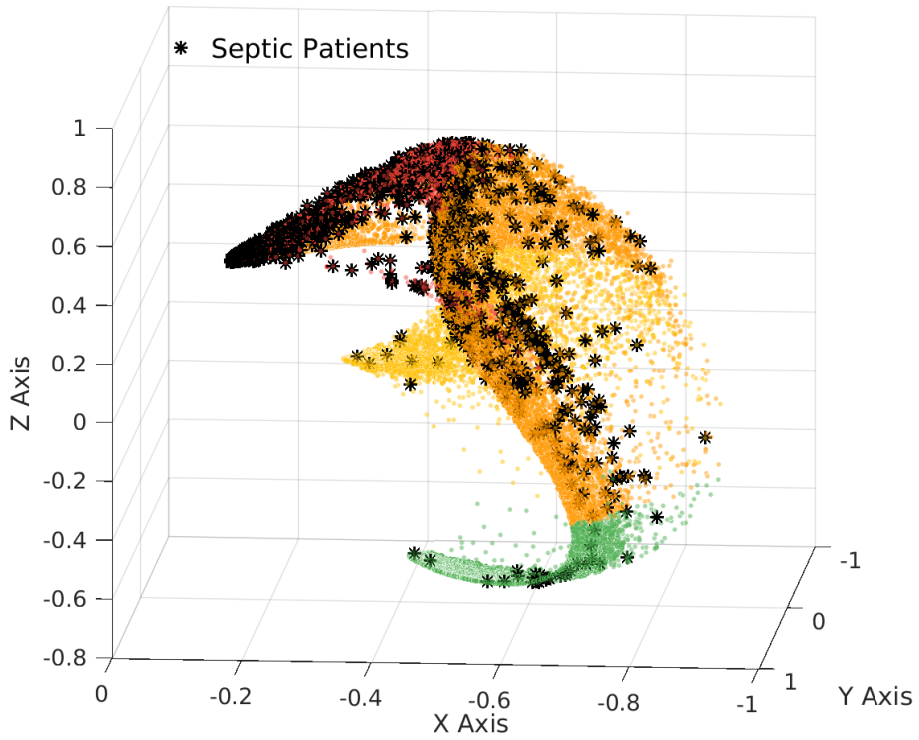


Figure 3.13: **Visualization of DeepAISE time series covariates performed by spectral clustering, with septic patients represented by asterisk.** The colors for the patients in the plots were chosen based on the predicted sepsis risk score (green represents the lowest predicted sepsis risk score, and red represents the highest predicted sepsis risk score).

(last layer of the DeepAISE model) followed by dimensionality reduction via Spectral clustering [97]. The individual axes in Figure 3.13 denote a unique weighted combination of the learned 100-dimensional representations, designed to construct a 3D space that preserves the distance among the original data points as much as possible. Two exemplar 3D patient trajectories are presented in Figure 3.14. Patient 2 was in a state of good health (specifically no suspicion for infection) prior to developing a subdural hemorrhage which prompted admission. This patient went on to be diagnosed with a ventilator associated pneumonia two days after an emergent craniectomy. In contrast Patient 1, who was several weeks status post craniectomy for stroke, was readmitted with a culture positive pneumonia present on admission. The manifold in Figure 3.14 shows that trajectories for patient 1 and 2 follow similar terminal patterns; however, correctly assigns them different starting positions with

patient 2 starting from a comparatively higher risk cluster. The specific trajectory of an ICU patient may be useful in categorizing infectious phenotypes and detecting anomalous physiological dynamics.

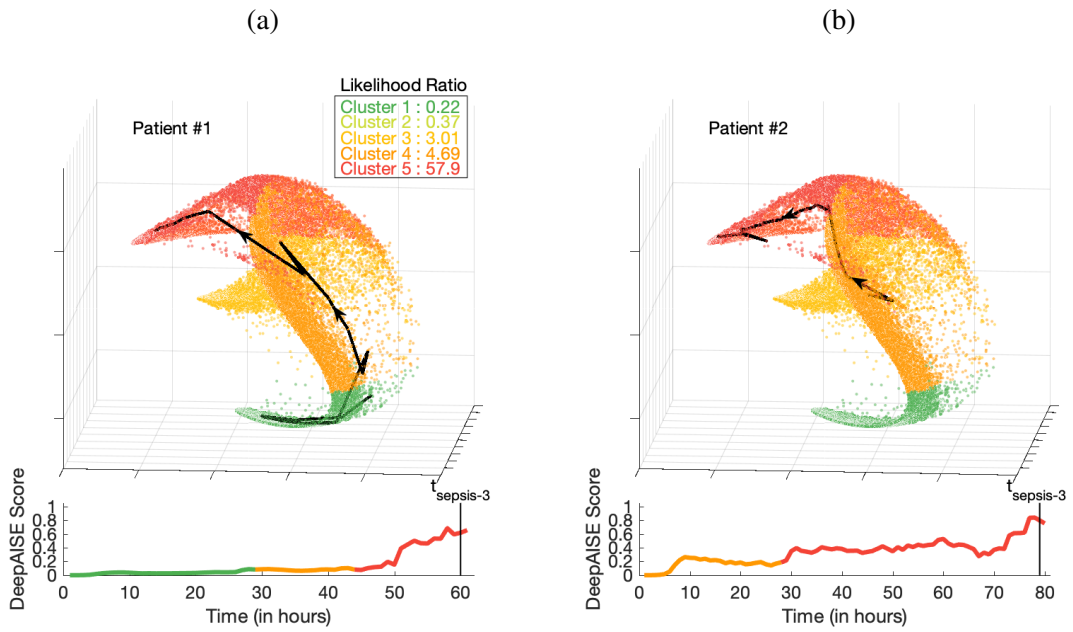


Figure 3.14: **Visualization of DeepAISE time series covariates performed by spectral clustering.** The trajectory of the DeepAISE score for 2 septic patients from ICU admission until sepsis diagnosis is displayed below a larger manifold that makes use of spectral clustering to visually display a patient's physiologic journey through their illness (each point on the graph represents one hour of data from one patient). The colors for both patients in the plots are chosen based on the predicted sepsis risk score (green represents the lowest predicted sepsis risk score, and red represents the highest predicted sepsis risk score). (A) Patient #1 (P1) was a 63-year-old female admitted for a left sided subdural hemorrhage who underwent a craniectomy on hospital day zero. This patient remained intubated after surgery and began receiving treatment for a culture proven ventilator associated pneumonia the afternoon of hospital day number three. DeepAISE identified this patient as being septic nearly 24hrs before clinical treatment was implemented (See Figure 3.17). (B) Patient #2 (P2) was a 70 year old male who was admitted for altered mental status and seizures after vascular coiling of a middle cerebral artery (MCA) aneurysm. P2 was intubated on admission and began treatment for a culture proven ventilator associated pneumonia on hospital day five however DeepAISE made its sepsis prediction nearly 36 hours prior to this time, after demonstrating a steadily worsening score since admission (See Figure 3.18).

3.2.6 DeepAISE user interface and tele-ICU workflow integration

The goal of any CDS system is to improve patient outcomes and reduce hospitalization costs; however, actualization of these goals is incumbent upon clinical teams embracing and actually employing the technology. The integration of a CDS system into clinical workflows depends on many factors, and therefore the development of the DeepAISE UI (User Interface) involved nursing stakeholders in our tele-ICU center. Appreciating the workflow of the tele-ICU staff was a critical component of ensuring that the developed UI was both useable and interpretable. Nursing stakeholders identified the key tasks in the tele-ICU as consisting of the following: routine patient assessment, sepsis risk assessment, communication with the bedside clinical team, and physician initiation of therapeutic interventions. A minimal user interface (UI) that enhanced workflow awareness, provided easy actionability, and ensured data integrity was built after soliciting requirements from the aforementioned stakeholders in early 2017. The resultant UI shown in Figure 3.15 was designed to present a list of patients sorted by DeepAISE risk score for predicting $t_{sepsis-3}$ four hours in advance. Square cards that include the sepsis risk score as well as the change in score over the past hour are used to represent a single patient. Individual cards can be flipped via a single mouse-click to reveal the top factors contributing to the presented score. To improve individual and unit situational awareness regarding patient interventions and assessments, users can drag-and-drop patient cards into columns representing different treatment categories.

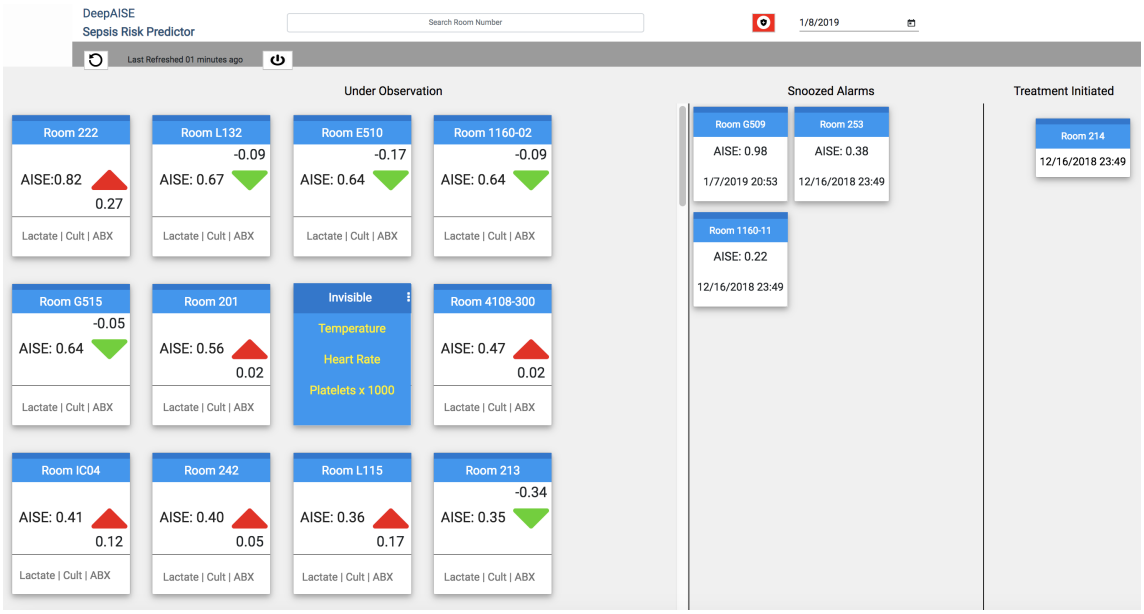


Figure 3.15: **Screenshot of the clinician facing DeepAISE UI.** In the left column, patients are ranked in decreasing severity of illness. An individual patient card shows DeepAISE score on the front, and upon a single mouse click the card is turned displaying the top causes contributing to the risk score (e.g. Temperature, Heart Rate, Platelets). The middle column displays patients that have undergone review by a clinician. The right most column displays patients for whom treatment has been initiated.

3.3 Discussion

In this work, a deep learning model was used to automatically learn complex features, including temporal trends and higher-order interactions among the risk factors, to accurately predict the likelihood of sepsis in patients admitted to the ICU up to 12 hours in advance. Additionally, satisfaction amongst users of our CDS system was most greatly impacted by the enhancement of clinical interpretability of the findings through a workflow-aware UI that incorporated a patient's trajectory and the key factors contributing to their risk score. DeepAISE was developed to predict $t_{sepsis-3}$ (Sepsis-3 criterion) in this study. We observed that the four hours ahead prediction AUC of DeepAISE (on Emory testing set) was 0.90 for $t_{sepsis-3}$. Additionally, the performance of DeepAISE expectedly dropped as the prediction window increased from 2 to 12 hours. All these findings were externally reproduced with the UCSD and MIMIC-III patient cohorts, providing supporting evidence that the DeepAISE algorithm can be tailored and applied to a geographically diverse patient population.

Another advantage of the proposed deep learning approach is in its ability to provide the top factors contributing to the risk score for every point in time for each patient (i.e., local interpretability). The distinction between the global and local notions of interpretability (i.e., what features are contributing to the sepsis risk score for the cohort at large versus an individual patient's hourly prediction window) is most notable when dealing with models capable of capturing higher order interactions and temporal trends in the data. As a result, the degree of risk associated with a factor (e.g., temperature) is a function of other factors in a multiplicative sense (e.g., hypothermia and old age are together a greater risk factor than either by itself). Similarly, the temporal context of a risk factor can alter its contribution to a given risk score calculation (e.g., leukocytosis immediately after surgery may not be unexpected and may contribute differently to the risk for sepsis). These multiplicative and temporal factors (which are captured by the DeepAISE model) result in variations in the

importance of risk factors when viewed from a local, hourly prediction perspective for each patient. Note that traditional logistic regression models and decision trees are not capable of making such inferences unless the relevant features are hand-crafted by the experts and included in the model.

The DeepAISE algorithm was implemented in a real-time setting and a decision support user interface (UI) was designed to communicate the risk scores to the clinical team. A major barrier to wide adoption of modern machine learning based CDS tools in clinical practice has been their "black box" nature [41]. While it is important to design deep learning models with high performance, it is imperative to build models that provide interpretable data to bedside clinicians that can augment their understanding of the disease process and can contribute to the selection and initiation of appropriate treatments. DeepAISE was designed to be transparent by: 1) continually revealing the top causes contributing to the sepsis risk score (see Figure 3.10), 2) providing a lower dimensional view of the patients' trajectory (see Figure 3.11), and 3) providing a prioritized list of patients at risk for sepsis (see Figure 3.15). These three attributes allow the bedside clinician to identify pathologic deviations from expected physiology early and in real-time throughout the duration of patients' hospital admission. Incorporation of these top causes in the DeepAISE UI contributed to user satisfaction in our preliminary study. Further longitudinal usability studies are required to validate the utility of this feature to improve situational awareness and assist clinicians with independent evaluation of patient's risk for sepsis prior to initiation of interventions.

We have shown that the top causes can be broken down into two categories of positively and negatively contributing factors to the risk score. Notably, this analysis shed insight on the input features contributing significantly to the sensitivity (positive contributors) and specificity (negative contributors) of DeepAISE (see Figure 3.12). Since one of the key limitations of using EHR data is the intermittent nature of laboratory measurements, we hypothesize that one may use the knowledge of the top contributing factors to protocolize

the ordering of laboratory tests, to ensure specific updated measurements of these factors are available to the algorithm, thus improving model sensitivity and specificity. Our results indicate that the degree of data missingness was inversely correlated with the prevalence of sepsis within a given subgroup of patients (see Table 3.3). DeepAISE had the highest AUC on the sub-group with the highest level of missingness and the lowest prevalence of sepsis; since data missingness pushes the risk score to lower values which translates to more specific predictions at the cost of reduced sensitivity and positive predictive value. The net effect was a higher AUC due to the higher prevalence of negative labels within this subgroup. Further work is required to assess the performance of DeepAISE under varying degrees of missingness of the top contributing factors to the risk score. This is particularly important as one extends such algorithms to non-ICU units where patients are not as frequently monitored.

In recent years, several machine learning-based models for early prediction of sepsis and septic shock have been proposed; although variations in experimental design and definitions of sepsis make a direct comparison of these methods impractical. Desautels et al. [23] proposed a proprietary machine learning model called InSight to predict sepsis in ICU patients. Their model used a combination of vital signs, pulse oximetry, GCS, and age as input features. An earlier version of this algorithm relied on the Sepsis-3 definition (specifically t_{SOFA}) to train its model and was able to reliably identify (detect) patients at the time they had met the Sepsis-3 criteria, with a 4-hours ahead prediction AUC of 0.74, which is comparable to performance reported by Amland et al. [98]. Following the Sepsis-3 definition, Nemati et al. [25] achieved an AUC of 0.85 for 4 hours ahead prediction of sepsis, by combining 65 data points including low-resolution data from the EHR and high-resolution vital sign time series features from the bedside monitors. The superior performance of DeepAISE in comparison to the abovementioned models can be attributed to employment of an RNN-based model that captures patients' clinical trajectory.

Experimental design can have a pronounced effect on the reported AUC of machine

learning algorithms. A commonly utilized method known as the 'case-control' design (which includes the majority of studies involving biomarkers) significantly overestimates the true prevalence of positive labels and can result in highly optimistic reported performances in the literature when compared to a 'sequential prediction' design [25]. Assuming a sepsis prevalence of 8% in the ICU population (after excluding all cases of sepsis developed before ICU admission), a median time-to-sepsis of 23 hours, and a 4-hours ahead prediction window, typically only 1-2% of the observed windows include a positive label for sepsis. The case-control study design assumes the timing of sepsis is known a priori and seeks to show that certain physiological or biomarker signatures preceding this time are significantly different than that of the non-septic control patients. The resulting algorithms, which are tuned to a 50% prevalence of positive labels, tend to produce high rates of false alarms when deployed prospectively.

In general, statistical evaluation methods (such as the AUC) have a limited applicability when evaluating the clinical utility of such algorithms, although they can provide quantitative metrics for the comparison of various algorithms. In practice, performance metrics are only meaningful when coupled with appropriate clinical protocols that describe the course of action in response to the associated risk alerts. Simple clinical actions (such as 'snoozing' the alarm for X hours if the patient did not meet the clinical threshold to initiate therapy) can significantly alter the false-alarm rate (defined as 1-Specificity) and the associated AUC of an algorithm in practice.

While we have strictly adhered to the Sepsis-3 criteria for defining septic labels in our study, it has been noted that this criterion is too stringent and the sensitivity of early detection is lost to an increased specificity [99, 100]. The Sepsis-3 criterion utilized in this study is an acausal clinical construct for demarcating the onset time of sepsis, and as such cannot be used in a clinical setting for early detection of sepsis; however, a predictive analytic risk score when trained to predict the associated onset-time can combine the specificity advantages of Sepsis-3 with the benefits of early recognition. Moreover, it is critical to appreciate

that making a clinical diagnosis of sepsis carries much greater value than simply identifying 'poor health' or general decompensation. True cases of sepsis can be positively impacted by the rapid administration of broad-spectrum antibiotics, IV fluids, and vasopressors if indicated [10, 13] where as "decompensated" patients still need further assessment to ascertain the etiology of the deterioration and to determine appropriate intervention.

A potential use case of DeepAISE is to facilitate compliance with standardized care bundles such as SEP-1 [19], which advocates for obtaining blood cultures, administering broad spectrum antibiotics, measuring lactate, and starting appropriate fluid resuscitation if clinically indicated, all within 3 hours of clinical recognition of sepsis. We anticipate that a likely clinical workflow may include: 1) flagging of a patient by the DeepAISE risk score with a prediction horizon of 4 hours, 2) independent evaluation of the patient by a bedside caregiver (this may include ordering of additional labs such as lactate), 3) followed by ordering of cultures prior to ordering of antibiotics, and 4) completion of the SEP-1 bundle components. The predictive ability of DeepAISE and enumeration of top causes contributing to the risk score is remarkable because clinicians will be able to independently evaluate the algorithm's rationale for flagging a patient, and when clinically appropriate begin implementing components of the sepsis bundle much earlier. In fact, a recent study provided critical evidence [101] that longer intervals from antibiotic order to infusion are associated with higher mortality rates in septic and septic shock patients, thus emphasizing the importance of improving workflow related factors to the care of this patient population.

Sepsis survivors often suffer from high rates of readmission [102] and many survivors of sepsis face life-long, debilitating sequelae as a result of the disease [103]. Future extensions of this work will involve performing prospective clinical trials to validate DeepAISE's real-time predictions in a clinical setting; however, our findings provide significant clinical evidence for a radical change to the sepsis treatment paradigm that has real-time high-dimensional data analysis and model transparency at its center.

3.4 Appendix

Input features

The complete list of the input features to the model is as follows -

1. High-resolution dynamical features (calculated using 6 hours sliding windows, with 5 hours overlap; 6 features)

    - standard deviation of RR intervals and Mean Arterial Blood Pressure (RR_{STD} and MAP_{STD}), average multiscale entropy of RR and MAP (HRV_1 and BPV_1) and average multiscale conditional entropy of RR and MAP (HRV_2 and BPV_2).

2. Clinical features (10 features)

    - Mean Arterial Blood Pressure (MAP), Heart Rate (HR), Oxygen Saturation (O_{2Sat}), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiratory Rate ($RESP$), Temperature ($Temp$), Glasgow Coma Scale (GCS), Partial Pressure of Arterial Oxygen (PaO_2), Fraction of Inspired O2 (FiO_2).

3. Laboratory (General; 25 features)

    - White Blood Count (WBC), Hemoglobin, Hematocrit, Creatinine, Bilirubin and Bilirubin Direct, Platelets, International Normalized Ratio (INR), Partial Prothrombin Time (PTT), Aspartate Aminotransferase (AST), Alkaline Phosphatase, Lactate, Glucose, Potassium, Calcium, Blood rea nitrogen (BUN), Phosphorus, Magnesium, Chloride, B-type Natriuretic Peptide (BNP), Troponin, Fibrinogen, CRP, Sedimentation Rate, Ammonia.

4. Laboratory (Arterial Blood Gas or ABG; 5 features):

    - pH, pCO_2, HCO_3, Base Excess, SaO_2.

5. Demographics/History/Context (19 features)

- Care Unit (Surgical, Cardiac Care, or Neurointensive care), Surgery in the past 12 hours, Wound Class (clean, contaminated, dirty, or infected), Surgical Specialty (Cardiovascular, Neuro, Ortho-Spine, Oncology, Urology, etc.), Number of antibiotics in the past 12, 24, and 48 hours, Age, Charleston Comorbidity Index (CCI), Mechanical Ventilation, maximum change in SOFA score over the past 6 hours

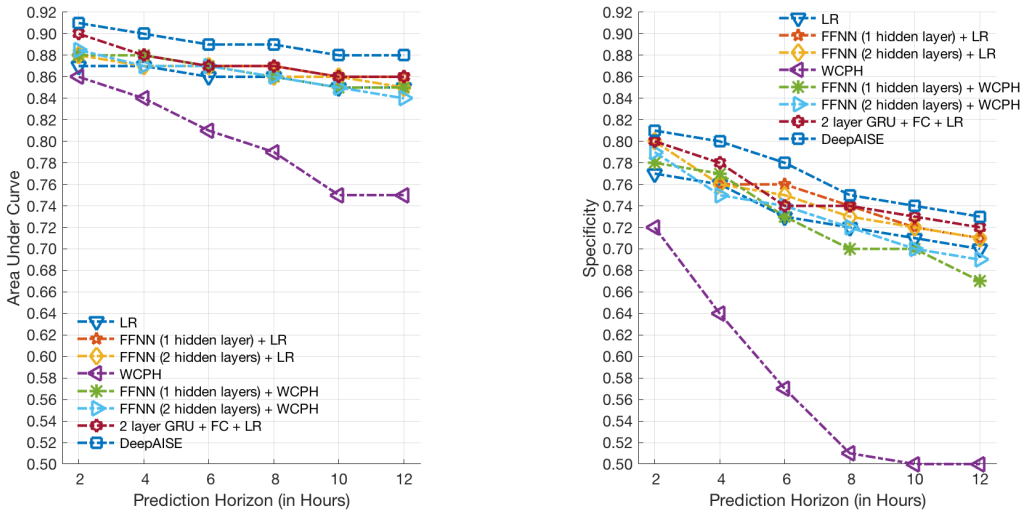
Table 3.4: Summary of Emory testing set prediction performance of DeepAISE model in predicting $t_{sepsis-3}$ four hours in advance. The DeepAISE model consists of a 2 layer GRU, a fully connected layer and WCPH model. The Area Under the Curve (AUC), Specificity (SPC) and Accuracy (ACC) are reported for both training set and testing set

	Testing set (Training Set)		
Model	**AUC**	**SPC**	**ACC**
LR	0.87 (0.89)	0.76 (0.79)	0.76 (0.79)
FFNN (1 hidden layer) + LR	0.88 (0.92)	0.76 (0.85)	0.77 (0.85)
FFNN (2 hidden layers) + LR	0.87 (0.92)	0.76 (0.85)	0.78 (0.85)
WCPH	0.84 (0.86)	0.64 (0.71)	0.64 (0.71)
FFNN (1 hidden layer) + WCPH	0.88 (0.92)	0.77 (0.85)	0.77 (0.85)
FFNN (2 hidden layers) + WCPH	0.87 (0.93)	0.75 (0.88)	0.75 (0.88)
2 layer GRU + FC + LR	0.88 (0.90)	0.78 (0.88)	0.78 (0.88)
DeepAISE	**0.90 (0.94)**	**0.80 (0.89)**	**0.78 (0.82)**

* *LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figures

(a) (b)



** LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figure 3.16: **Comparison of Emory testing set performance of all baseline models and DeepAISE to predict $t_{sepsis-3}$ for prediction horizons of 2, 4, 6, 8, 10, and 12 hours.** The Area Under the Curve (AUC) is shown in the left panel. The Specificity (SPC) is shown in the right panel.

Table 3.5: Summary of patient characteristics of Emory ICU cohort

Model	All Patients	Non-Septic	Septic
Patients, (#)	25820	24375	1445
			6%
Male, no. (%)	53.3	53.2	55.2
Age, median (IQR) y	61	61	61.5
	[49 - 71]	[49 - 71]	[50.5 - 72]
Race, no. (%)			
Caucasian	48.6	48.9	45.0
Black	43.3	43.1	45.4
Asian	1.3	1.3	1.3
Hispanic	0.024	0.02	0.08
Surgery (%)			
Cardiovascular	13.1	-	-
Neuro	6.1	-	-
Ortho-spine	1.8	-	-
Oncology/General Surgery	3.6	-	-
Urology	0.4	-	-
ICU LOS, median (IQR) h	48	46	141
	[28 - 90]	[27 - 77]	[77 - 258]
Inpatient Mortality, %	4.1	3.5	15.2
Inpatient Hospice, %	3.8	3.3	12.5
SOFA, median (IQR)	1.9	1.7	5.0
	[0.6 - 4.0]	[0.5 - 3.6]	[3.1 - 7.4]
CCI, median (IQR)	2	2	3
	[1 - 4]	[1 - 4]	[2 - 5]
ICU Admission to $t_{sepsis-3}$,	-	-	24
median (IQR) h			[9 - 63]

Table 3.6: Summary of patient characteristics of UCSD ICU cohort

Model	All Patients	Non-Septic	Septic
Patients, no.	18752	17679	1073
			5.7%
Male, no. (%)	61.1	60.8	66.5
Age, median (IQR) y	59.8	59.7	60.8
	[46.4 - 70.8]	[46.3 - 70.8]	[46.9 - 70.3]
Race, no. (%)			
Caucasian	52.2	52.5	47.4
Black	7.9	7.9	6.2
Asian	5.4	5.4	6.4
ICU LOS, median (IQR) h	44.8	43.3	143.8
	[24.3 - 79.6]	[23.8 - 72.9]	[78.5 - 241.2]
Mortality, %	4.7	3.7	21.1
SOFA, median (IQR)	2	1	4
	[0 - 4]	[0 - 3]	[2 - 6]
CCI, median (IQR)	3	2	3
	[1 - 6]	[1 - 3]	[2 - 6]
ICU Admission to $t_{sepsis-3}$,	-	-	38
median (IQR) h			[16 - 74]

Table 3.7: Summary of patient characteristics of MIMIC-III ICU cohort

Model	All Patients	Non-Septic	Septic
Patients, no.	40474	38198	2276
			5.6%
Male, no. (%)	56.5	56.3	58.8
Age, median (IQR) y	66	65	66
	[52 - 77]	[52 - 77]	[50 - 71]
Race, no. (%)			
Caucasian	71.6	71.5	72.4
Black	9.2	9.2	8.6
Asian	2.2	2.2	2.4
Hispanic	3.4	3.4	3.4
ICU LOS, median (IQR) h	47	45	158
	[27 - 88]	[26 - 74]	[83 - 266]
Mortality, %	8.9	8.1	22.0
SOFA, median (IQR)	1.6	1.5	3.3
	[0.65 - 3.1]	[0.6 - 2.9]	[2.0 - 5.1]
CCI, median (IQR)	2	2	3
	[1 - 3]	[1 - 3]	[1 - 4]
ICU Admission to $t_{sepsis-3}$,	-	-	31.2
median (IQR) h			[13.3 - 70.2]

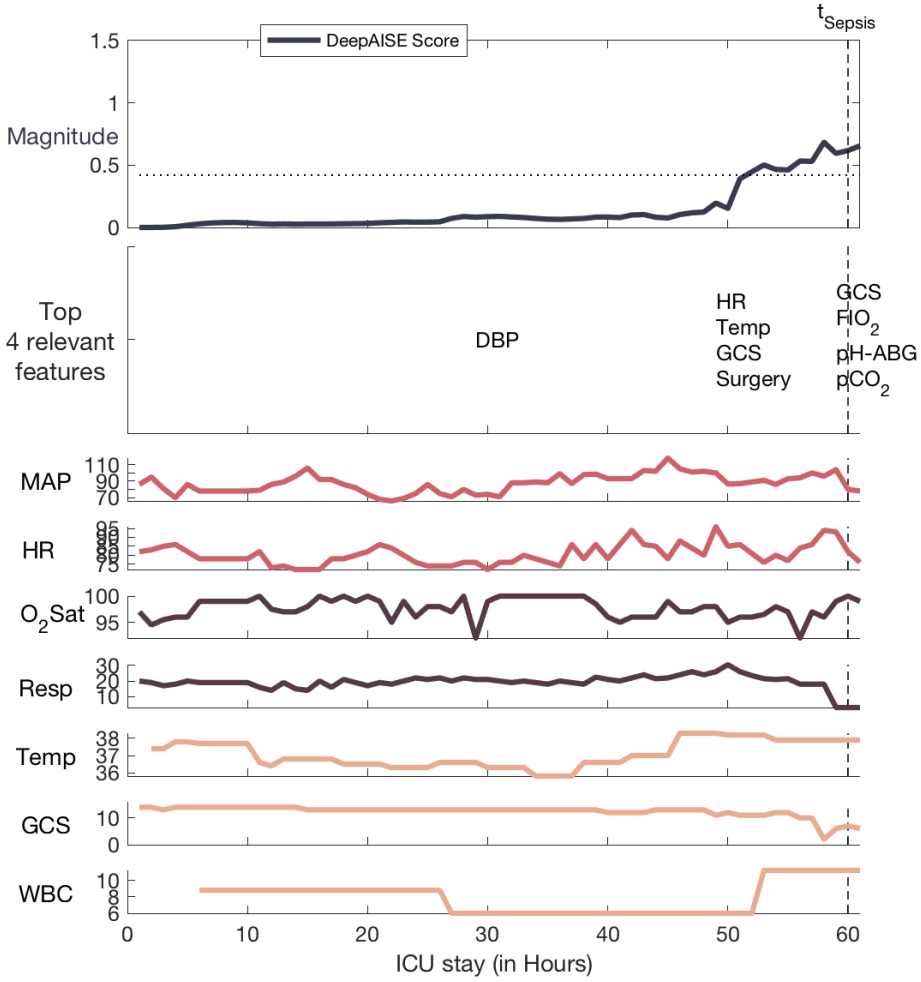


Figure 3.17: **DeepAISE score shown for Patient #1 (P1).** Commonly recorded hourly vital signs of the patient, including heart rate (HR), mean arterial blood pressure (MAP), respiratory rate ($RESP$), temperature ($TEMP$), oxygen saturation (O_2Sat) are shown. The most significant features contributing to the DeepAISE score are listed immediately below the DeepAISE Scores (for clarity of presentation, only selected time points are shown). The horizontal dashed line indicates the prediction threshold corresponding to a sensitivity of 0.85. Refer to Appendix C of Supplementary Material for more details on the abbreviated features.

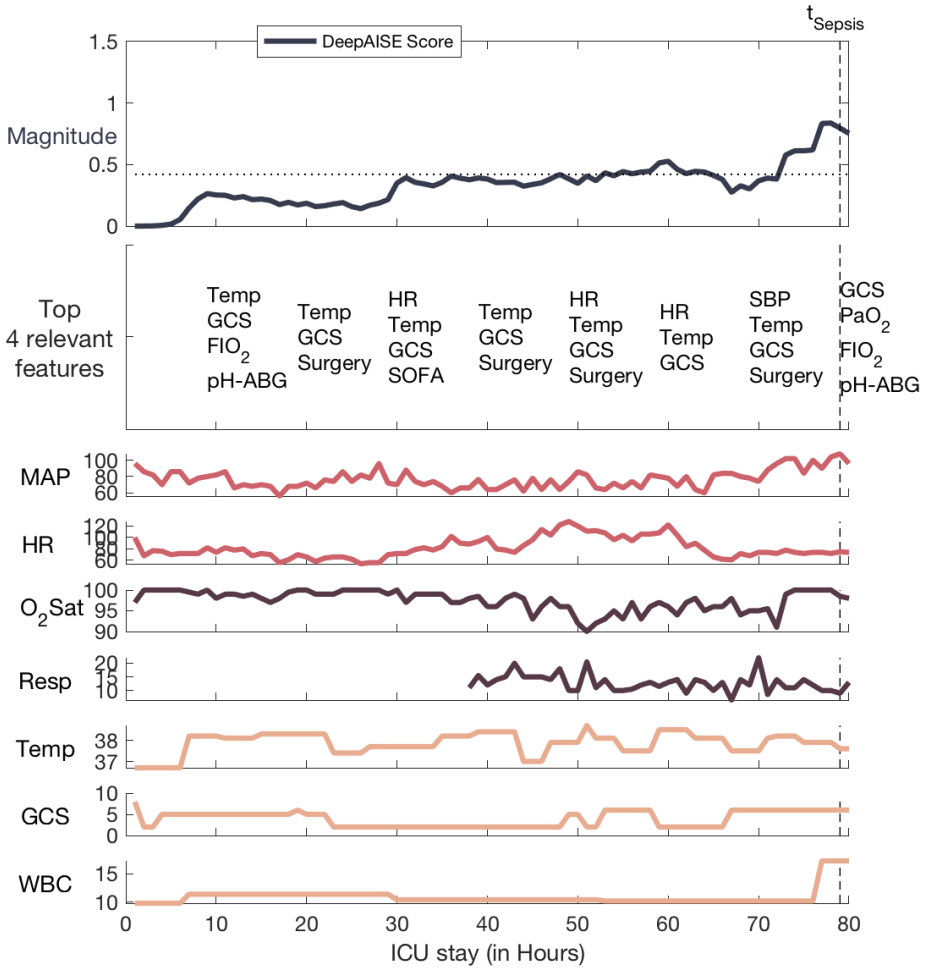


Figure 3.18: **DeepAISE score shown for Patient #2 (P2).** Commonly recorded hourly vital signs of the patient, including heart rate (HR), mean arterial blood pressure (MAP), respiratory rate ($RESP$), temperature ($TEMP$), oxygen saturation (O_2Sat) are shown. The most significant features contributing to the DeepAISE score are listed immediately below the DeepAISE Scores (for clarity of presentation, only selected time points are shown). The horizontal dashed line indicates the prediction threshold corresponding to a sensitivity of 0.85. Refer to Appendix C of Supplementary Material for more details on the abbreviated features.

CHAPTER 4

COMPOSER - DEVELOPMENT AND VALIDATION OF A GENERALIZABLE MODEL FOR EARLY PREDICTION OF SEPSIS USING CONFORMAL METHODS AND DOMAIN ADAPTATION

4.1 Introduction

Acute care facilities and in particular intensive care units (ICUs) provide an environment where an immense amount of data is acquired, and it is expected that with the advent of wearables and biometric patches even more continuously recorded data will be available in such settings [104, 105]. But at present, very little of these data are used in a real-time setting to prognosticate effectively, and the existing machine learning (ML) and predictive analytics risk scores (many of them only tested retrospectively) suffer from lack of generalizability across institutions and performance degradation within the same institution across time [48]. This limited generalizability is due to number factors including differences in local populations, EHR systems, coding definitions, laboratory equipment and assays, as well as variations in clinical and administrative practices. For instance, in a recent study aimed at detecting abnormal chest radiographs, the specificity of an ML model at a fixed operating point varied widely, from 0.566 to 1.000, across five independent datasets [106, 107]. Similarly, most existing published clinical ML and predictive analytic models are either based on data from a single hospital [26, 31] or multiple hospitals from the same healthcare system [49] where the processes of care are mostly standardized. Although less common, clinical ML models that have been validated across different healthcare systems are either re-trained from scratch or are fine-tuned (via transfer learning) on every new patient cohort [25, 50], or when applied out-of-the-box often exhibit significant degradation in performance [51].

Finally, as the processes of care are often variable across different levels-of-care (Emergency Departments, ICUs, and step-down units/general wards) and the degree of data missingness is commonly a function of illness severity [52], ML algorithms typically have to be optimized for different levels-of-care and patient type. This observation has prompted some researchers to suggest that exploiting the patterns of data missingness (e.g., presence or absence of laboratory tests) might be useful for assessment of risk, with one single-center study suggesting 4-5 point improvement in AUC when including missing data indicators as features into a 30-day mortality prediction model [108]. In our experience, such models often perform extremely poorly when applied to data from other healthcare systems, where the frequency of measurement of labs is also tied to other workflow-related factors. Moreover, the re-training of models is typically expensive and impractical (namely due to the difficulty of obtaining gold-standard labels). Finally, regulatory guidelines surrounding the re-training of ML algorithms, categorized under Software as a Medical Device (SaMD), remain unclear [109]. In spite of tremendous interests and hypes around the application of ML algorithms, the above-mentioned factors have impeded the process of evidence generation, commercialization, and wide-adoption of these tools.

As noted by Agniel et al.[52], while without careful considerations of context, EHR data may be unsuitable for answering many research questions. However, when healthcare processes are adequately addressed and incorporated into ML models through introduction of inductive biases (i.e., necessary and appropriate assumptions built into model architecture, learning process, and application/deployment) such data can be leveraged to gain insight into patients' state of health. The work proposed in this chapter addresses these key gaps in generalizability, including 1) modeling of variations in frequency of laboratory measurements across levels-of-care and different healthcare systems (also related to data missingness), 2) handling of data/population drifts (aka, data distribution shifts), and 3) explicitly quantifying the *conditions for use* of an algorithm via design of algorithmic controls.

The primary contributions of the work proposed in this chapter are as follows:

- We propose a weighted input layer that is designed to handle *missing data* and *variations in data measurement frequency* across various levels of care (Emergency Department, ICU etc.) and different institutions.

- We utilize the technique of Adversarial Domain Adaptation (ADA) [110] to learn representations that minimize healthcare system specific variations. A key importance of utilizing ADA training procedure is the design of a predictive model that can adapt to new unlabeled target patient population; therefore, gold-standard labels which are often expensive to obtain, are not required to deploy the model at a new center.

- We also provide a framework that combines *representation learning* [111] and ideas from conformal methods [112, 113, 114, 115] for establishing the 'conditions for use' of a clinical model. This enables us to explicitly determine at what level of data covariance shift one may still trust a clinical risk score [116].

The proposed framework utilizes all three of the above contributions to make a **Co**nformal **M**ultidimensional **P**rediction **o**f **Se**psis **R**isk (COMPOSER) Score for the onset of sepsis within a four hour prediction horizon. We show the generalizability of COMPOSER model by utilizing data from over 480,000 patients collected between 2016-2019 from three different academic medical centers in the US, including data from Emergency Departments (EDs), Intensive Care Units (ICUs), and general wards. The rest of the chapter discusses in detail the datasets used in this study, the development of COMPOSER model and the results obtained by applying COMPOSER model on these datasets.

4.2 A multicenter dataset of patients in Emergency Department and ICU

In this study, we collected data from three different academic medical centers in the US to make up a total of five different patient cohorts (Refer to Table 4.17, Table 4.18 and Table

4.19 for a summary of patient characteristics for each of the five cohorts used in this study). This investigation was conducted according to UCSD IRB approved protocol #191098X. The Hospital-A ICU cohort was drawn from the Electronic Health Record (EHR) data of all patients admitted to the ICUs at two hospitals within the Hospital-A health system from 2016 to 2019. The Hospital-B ICU cohort was drawn from the EHR data of all patients admitted to the ICUs at two hospitals within the Hospital-B health system from 2014 to 2018. The Hospital-C ICU cohort was drawn from EHR data of all patients admitted to the ICUs at Hospital-C from 2016 to 2017. Patients 18 years or older were followed throughout their ICU stay until discharge or development of sepsis according to Sepsis-3 guidelines [1, 16]. For all the above ICU cohorts, we excluded patients if they developed sepsis within or prior to the first four hours of ICU admission (by analyzing pre-ICU IV antibiotic administration and culture acquisition) or if their length of ICU stay was less than 8 hours or more than 20 days.

The fourth cohort was the Hospital-A ED cohort which consisted of all patients admitted to the Emergency Departments (EDs) at two hospitals within the Hospital-A health system from 2016 to 2019. The fifth cohort was the Hospital-B ED cohort which consisted of all patients admitted to the Emergency Departments (EDs) at two hospitals within the Hospital-B health system from 2014 to 2018. For both the above ED cohorts, patients 18 years or older were followed throughout their ED stay until discharge or development of sepsis according to Sepsis-3 guidelines [16, 1]. Additionally, we excluded patients if they developed sepsis within or prior to the first 2 hours of ED admission or if their length of ward stay was less than 3 hours or more than 20 days.

The Hospital-A ICU cohort contained a total of 18990 patients, 1236 (6.5%) of whom met the Sepsis-3 criterion four hours or later after ICU admission. Out of the 18990 patients 70% of them were used for developing the COMPOSER model (training set), 10% were used as validation set and the remaining 20% formed the testing set. Note that the proportion of the splits were the same for both septic and non-septic patients. The Hospital-

B ICU cohort contained a total of 45679 patients, 2563 (5.6%) of whom met the Sepsis-3 criterion four hours or later after ICU admission. The Hospital-C ICU cohort contained a total of 7426 patients, 229 (3.1%) of whom met the Sepsis-3 criterion four hours or later after ICU admission. The Hospital-A ED cohort contained a total of 86869 patients, 1308 (1.5%) of whom met the Sepsis-3 criterion 2 hours or later after ED admission. The Hospital-B ED cohort contained a total of 325916 patients, 6236 (1.9%) of whom met the Sepsis-3 criterion 2 hours or later after ED admission. For the Hospital-A ICU/Hospital-A ED/Hospital-B ED cohorts, 80% of the patients were used for developing the COMPOSER model and the remaining 20% formed the testing set. For the Hospital-C ICU cohort, 70% of the patients were used for developing the COMPOSER model and the remaining 30% formed the testing set. Please refer to Table 4.17 for more details and comparison amongst all the five cohort used in this study.

Additionally, in terms of number of one-hour windows, the Hospital-A ICU dataset consisted of a total of 725,886 windows out of which 717,468 (98.9%) windows were non-septic and 8418 (1.1%) were septic. The Hospital-B ICU dataset consisted of a total of 1,747,139 windows out of which 1,729,275 (98.9%) windows were non-septic and 17,864 (1.1%) were septic. The Hospital-C ICU dataset consisted of a total of 284,570 windows out of which 283,006 (99.4%) windows were non-septic and 1564 (0.6%) were septic. The Hospital-A ED dataset consisted of a total of 762,103 windows out of which 755,432 (99.1%) windows were non-septic and 6671 (0.9%) were septic. The Hospital-B ED dataset consisted of a total of 1,938,829 windows out of which 1,909,181 (98.4%) windows were non-septic and 29,648 (1.6%) were septic.

4.3 Clinical variables for model development

A total of 40 clinical variables (34 dynamic and 6 demographic variables) were extracted based on their association with onset of sepsis and their availability in EHR across the different hospitals considered in our study [25, 117, 22]. These included vital signs mea-

surements (heart rate, pulse oximetry, temperature, Systolic blood pressure, mean arterial pressure, diastolic blood pressure, respiration rate and end tidal carbon dioxide), laboratory measurements (bicarbonate, measure of excess bicarbonate, fraction of inspired oxygen, pH, partial pressure of carbon dioxide from arterial blood, oxygen saturation from arterial blood, asparate transaminase, blood urea nitrogen, alkaline phosphatase, calcium, chloride, creatinine, bilirubin direct, serum glucose, lactic acid, magnesium, phosphate, potassium, total bilirubin, troponin, hematocrit, hemoglobin, partial thromboplastin time, leukocyte count, fibrinogen and platelets) and demographic variables (age, gender, identifier for medical ICU unit, identifier for surgical ICU unit, length of hospital stay, length of ICU stay). All vital signs and laboratory variables were organized into 1-hour non-overlapping time series bins to accommodate for different sampling frequencies of available data. The 1-hour time bin interval was selected as a balance between having short windows with too many missing data points (low frequency clinical data) and having time windows too long to make any meaningful prediction. All the variables with sampling frequencies higher than once every hour were uniformly resampled into 1-hour time bins, by taking the median values if multiple measurements were available. Variables were updated hourly when new data became available; otherwise, the old values were kept (sample-and-hold interpolation). Mean imputation was used to replace all remaining missing values (mainly at the start of each record). Additionally, for every vital signs and laboratory variable, the time since the variable was last measured (TSLM) was recorded.

In the later part of this chapter, we will refer to the 34 dynamical variables by $X_{dyamical}$, the TSLM features by X_{TSLM} and the 6 covariate features by X_{covar}. All of the above features together make up 74 features.

4.4 Development of the COMPOSER model

The COMPOSER model is designed to operate sequentially over the EHR data of a patient. Starting from beginning of a patient record, the predictive model is fed with the input

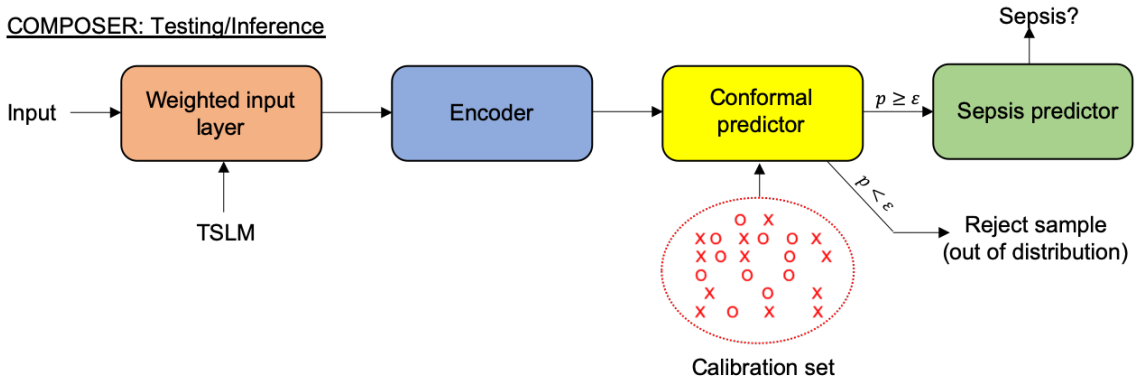


Figure 4.1: **Schematic diagram of the COMPOSER model during testing phase.** The test data point is first passed through the weighted input layer along with the TSLMs for each of the 34 dynamical variables. The output from the weighted input layer is then fed into the encoder to obtain a lower dimensional feature vector of the input. This feature vector is then fed into the conformal predictor, which compares the feature vector with other representations present in the calibration set to determine if the test data point belongs to the same probability distribution of the calibration or not. If yes ($p \geq \epsilon$), the feature vector is passed onto the sepsis predictor to obtain the probability of onset of sepsis.

features (as described in Section 4.3) to obtain hourly predictions of the probability of onset of sepsis ($t_{sepsis-3}$) within the next four hours. Whenever the predicted probability crosses a decision threshold, it is considered to be a positive prediction for onset of sepsis. The main contributions of this work are in the design of a generalizable model using a constrained deep learning architecture with domain adaptation, and the demonstration of its generalizability – on five different patient cohorts and across different levels of hospital care (ICU vs ED) – in making effective predictions for onset of sepsis.

Figure 4.1 provides the overall schematic diagram of the COMPOSER model during testing/inference phase. In this phase, the dynamic variables are first fed into the weighted input layer to obtain a scaled feature vector. A combination of the scaled feature vector and demographic variables are then fed into the encoder module. The encoder module transforms the high dimensional input feature vector into a lower dimensional representation that captures all information relevant to prediction of onset of sepsis. The lower dimensional vector obtained from the encoder module is then passed through the conformal predictor to detect any distributional shift. If no distribution shift is detected, it is then

passed through the sepsis predictor to obtain a probability score for onset of sepsis within the next four hours.

4.4.1 Weighted input layer

In this study, we make use of 34 laboratory and vital measurement data available in the EHR for helping distinguish healthy from septic patients. There are two challenges that exist when using dynamical variables and their contextual information for building predictive models. First, while a laboratory test's value can provide information about the state of health of a patient, it has been shown that laboratory test measurement patterns provide complementary information in addition to the values of the laboratory tests themselves [118]. This should however be treated with caution when the aim is to develop a generalizable predictive model since some of the tests (such as basic metabolic panel, calcium, phosphorous, magnesium) are ordered based on guidelines set by a hospital protocol, and these protocols can i) vary within a hospital across different levels of care (ICU vs Ward vs ED) and ii) vary from one hospital to another for the same level of care (Hospital A ICU vs Hospital B ICU). Due to the differing measurement patterns (and differing healthcare processes), a predictive model trained on data from Hospital A ICU can perform poorly on data from Hospital B ICU. Second, when a sample-and-hold approach is used for interpolation, there is no sense of the extent to which an interpolated value can be trusted. For example, heart rate is a highly variable feature and a measurement that was made more than 6 hours ago often does not represent the current physiological status of a patient and in such a scenario the interpolated value can misguide a predictive model. On the other hand, there are variables (such as lactate measurement) whose values are valid for longer durations after they have been measured. In order to address the above problems, we propose a weighted input layer that scales the latest measured value of a variable depending on the duration since it was measured. The extent of scaling is controlled by a parameter that is learned from the data. Let us consider X_t^n to be the 34 dimensional vector (consisting of all

dynamical variables) at time t for patient n. Henceforth, with a slight abuse of notation we will refer to X_t^n by X_t, wherein $X_t = [x_{1,t}, ; x_{2,t}; ...x_{34,t}]$, $x_{i,t} \in \mathbb{R}$. Next $\delta_{i,t}$ corresponds to the duration since variable i was last measured (in reference to the current time t). Each of the variables $x_{i,t}$ is then non-linearly weighted based on the duration since it was last measured, to obtain $x_{i,t}^w = x_{i,t} * f(\delta_{i,t}, \alpha_i)$.

The weighting function $f(.)$ is defined as follows:

$$f(\delta_{i,t}, \alpha_i) = 2 * \left(1 - \frac{1}{(1 + exp(-\alpha_i^2 * \delta_{i,t}))}\right) \tag{4.1}$$

Where α_i is a scaling factor for each of the dynamical variable and is learned during the training of the model. Thus, the output of the weighted input layer is a 34 dimensional feature vector $X_t^w = [x_{1,t}^w; x_{2,t}^w; ...x_{34,t}^w]$. The scaling factors α_i (for each dynamic variable) obtained at the end of training the COMPOSER model should reflect the extent of interpolation that is useful for predicting the onset of sepsis.

4.4.2 Adversarial domain adaptation

The arrival of Big Data in healthcare has provided an impetus to development of data-driven machine learning models (deep learning models) that are aimed at improving healthcare delivery to patients [119, 120, 121, 122, 123, 124]. A main concern with such data-driven deep learning models are that they might not generalize to unseen datasets that might have differing distributions of data, eg. A model trained using Hospital A ICU data (source domain) deployed at Hospital B ICU (target domain). This distributional differences in data (which in turn can lead to poor generalization in performance of such models) can arise due to wide variety of factors – shifts in demographic characteristics, shifts in baseline data measurements (due to change in measuring instruments), shifts in level of care, change in healthcare delivery protocol etc.

Domain adaptation could be a natural solution to this issue, wherein the focus is to

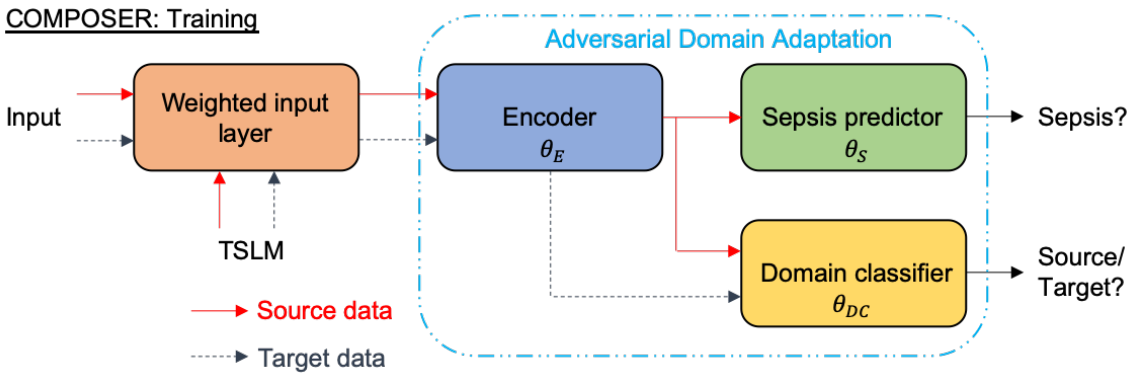


Figure 4.2: **Schematic diagram of the COMPOSER model during training phase.** The source dataset (labeled) is used to train the sepsis predictor, while both source and target (unlabeled) datasets are used to train the domain classifier. The encoder is trained to learn representations that remove institution-specific variations whilst retaining information useful for sepsis prediction.

learn representations that minimize the distributional shift between source and target domains [125, 126]. In this study, we use the technique of Adversarial Domain Adaptation (ADA) to find representations that maximize the performance of a classifier in the source domain (where ground truth labels are available) and minimize the domain shift that exists between source domain and application (or target) domain (where ground truth labels are not available) [110]. Note that while the focus of introducing weighted input layer was in improving generalizability of a model by addressing the problem of workflow and care level-specific patterns of data missingness, the focus of ADA is to learn representations that are robust to institution specific variability and noise. ADA training typically requires two datasets – the source dataset which contains ground truth labels (eg. Hospital-A ICU cohort) and the target dataset which does not contain any ground truth labels (eg. Hospital-B ICU cohort). The ADA component in COMPOSER model consists of three modules (three different Feedforward Neural Networks) – encoder, sepsis predictor, and domain classifier. During training, the encoder and the sepsis predictor are trained to maximize likelihood of source labels (septic labels) given the source inputs. Next, to ensure that the model generalizes well towards target domain, the encoder is trained to confuse domain classifier whose goal is to estimate the domain of a given input sample. Figure 4.2 shows a block diagram

of various components of ADA that are used during training of COMPOSER model.

Although we focus on using adversarial loss to minimize domain shift in this study, several works have used other types of losses to achieve the same purpose. The Maximum Mean Discrepancy (MMD) loss which computes the norm of the difference between two domain means has been used in recent studies [127, 128, 129]. Other studies have focused on proposing variants of the adversarial loss to obtain feature representations that are domain invariant [130, 128, 131, 132, 133]. More recently, the sliced Wasserstain discrepancy (SWD) metric was successfully applied to perform unsupervised domain adaptation [134].

Let us consider $\bar{X}^n = [X_1^{w;n}, X_2^{w;n}, \ldots . X_{Tn}^{w;n}]$ to be the set of data corresponding to a patient n whose total sequence length is T_n, note that $X_t^{w;n}$corresponds to the output of the weighed input layer (which has been described in Section 4.4.1). Additionally $\bar{Y}^n = [Y_1^n, Y_2^n, \ldots . Y_{Tn}^n]$ corresponds to the set of septic labels ($Y_t^n = \{0, 1\}$) for the same patient n. The ADA learning algorithm is then provided with labeled source data S drawn from the source domain, and unlabeled target data T drawn from the target domain. The following notations are for different components of the ADA algorithm: $G_E(*; \theta_E)$ is the L-dimensional encoder with parameters θ_E. $G_S(*; \theta_S)$ is the output of sepsis predictor with parameters θ_S which represents the probability of occurrence of sepsis within the next four hours, and $G_{DC}(*; \theta_{DC})$ corresponds to the output from domain classifier with parameters θ_{DC}. The ADA loss is given by:

$$L(\theta_E, \theta_S, \theta_{DC}) = L_S(\theta_E, \theta_S) + \lambda L_A(\theta_{DC}, \theta_E) \ , \tag{4.2}$$

where the adversarial loss (L_A) and the sepsis prediction loss (L_S) are given by:

$$L_A(\theta_E, \theta_{DC}) = L_A^{DC}(\theta_{DC}) + L_A^E(\theta_E) \ , \tag{4.3}$$

$$L_S(\theta_E, \theta_S) = \mathbb{E}_{i \sim S} \left[-\log \left(G_S \Big(G_E \big(\bar{X}^i; \theta_E \big); \theta_S \Big)_{\bar{Y}^n} \right) \right] \ . \tag{4.4}$$

Note, the adversarial loss is based on two separate classification losses, the first one (L_A^{DC}) focuses on cross-domain classification of source and target data (and a function of θ_{DC}) and the second one (L_A^E) focuses on learning features or representations that minimize cross-domain classification (and a function of θ_E):

$$L_A^{DC}(\theta_{DC}) = \mathbb{E}_{i\sim(S,T)}\left[-\log\left(G_{DC}\left(G_E\left(\bar{X}^i\right);\theta_{DC}\right)_{Y^i_{domain}}\right)\right] \tag{4.5}$$

where $G_{DC}(*;\theta_{DC})_{Y^i_{domain}}$ corresponds to the probability of actual domain from which example i was sampled from.

$$L_A^E(\theta_E) = -L_A^{DC} \tag{4.6}$$

In summary, the ADA model is trained to simultaneously minimize sepsis predictor (or supervised) loss L_S and adversarial loss L_A (Equation 4.2). The adversarial loss L_A is further broken down into two components – loss corresponding to the encoder L_A^E and loss corresponding to the domain classifier L_A^{DC} (Equation 4.3). The training objective for sepsis prediction module corresponds to minimizing the cross-entropy loss (supervised loss) as shown in Equation 4.4. The training objective for domain classifier module corresponds to minimizing the adversarial domain classifier loss L_A^{DC} as shown in Equation 4.5. Note that parameters of the encoder θ_f are updated during minimization of the supervised loss L_S and adversarial encoder loss L_A^E.

Additionally, the supervised loss L_S is computed only for data from source domain (since data in target domain are unlabeled). The weighting factor λ controls the relative strength between supervised and adversarial objectives. This procedure for training ADA model ensures that the encoder learns representations that balance relevance of prediction of sepsis and simultaneously maximize domain invariance.

Once ADA training has been completed, the trained encoder is used to create a calibration set which forms an integral part of the conformal prediction pipeline. More details

regarding the testing phase of COMPOSER model once ADA training has been completed is described in the next section.

4.4.3 Detecting distribution shift using conformal prediction

Deep neural networks such as COMPOSER are trained under the assumption that distribution of data at test time will be the same as that of the training distribution. This assumption might not hold true in the real world where data distributions could shift over time within the same institutions, or could vary across institutions. The model still attempts to make predictions even under the existence of such distribution shifts which can be potentially harmful. This issue has been formulated as a problem of detecting whether a given input data belongs to the training distribution or is out-of-distribution (OoD). There is a wide body of research that has focused on studying the problem of detecting out-of-distribution data [135, 136, 137, 138, 139] with a goal of providing robust predictions from an ML model.

In this section, we use the method of *conformal prediction* [112, 113, 114] to develop a pipeline that is designed to determine whether a given data sample belongs to the data distribution from which the ADA model was trained on. A sepsis prediction is made on the data sample only if it belongs to training distribution of the ADA model, else the data sample is rejected and no sepsis prediction is made.

Conformal prediction has been traditionally used in classification or regression, and can be used with any base classifier (eg. Support Vector Machine, Logistic Regression, Neural Networks). In a regular conformal prediction framework, each possible prediction that can be made for a given test data is evaluated based on its nonconformity score. The non-conformity scores (for each possible prediction) with a calibration set, are then used to compute a p-value, which are then thresholded to provide a guarantee on the error rate [114]. We will be explaining in detail below many of the aforementioned concepts. In our work, we modify the conformal prediction framework to help identify the presence of dis-

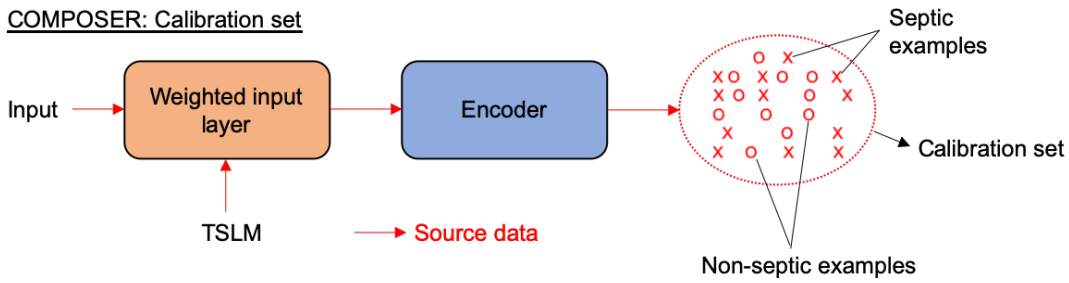


Figure 4.3: **Schematic diagram of the COMPOSER model being used to create calibration set.** Once training of COMPOSER is completed, a sub-sample of patients from source dataset are chosen and their encoder representations are used to form the calibration set. The calibration set is used by the conformal predictor to detect samples that are out of sepsis predictor training distribution.

tribution shift in data, and for which we will use as input a lower dimensional representation of the input data (the output from encoder module).

Notations: For simplicity, let us assume h_i to be a P dimensional output from the encoder for a data sample i. Let $H = [h_1, h_2, ..., h_U]$ represent the entire training set (containing both septic and non-septic examples). A subset of the training set is chosen as the calibration set, C, of size M. The calibration set can contain examples from both the septic and non-septic classes.

A brief outline of the 'conformal predictor' in COMPOSER model pipeline is as follows: We are given a calibration set $(c_1, c_2, ..., c_M)$, where each $c_i \in \mathbb{R}^P$ is the output from encoder corresponding to example i. We are then given a test example c_{M+1} for which: 1) the task of conformal prediction is to predict if the test example is drawn from the same probability distribution as that of other examples in the calibration set, and 2) if yes, the test example c_{M+1} is passed onto sepsis predictor to obtain the sepsis risk score.

First, we would like to measure how likely it is that a given sequence of examples were drawn from the same probability distribution. We will use the term $p-value$ to measure the typicalness of a sequence of examples wherein the $p-value$ is computed using a function $p : Z^* \rightarrow [0, 1]$. For a given input c_{M+1}, the $p-value$ of c_{M+1} denoted by $p(c_{M+1})$ refers to typicalness of the sequence $(c_1, c_2, ..., c_M, c_{M+1})$ (the sequence consists of all examples

in calibration set plus the given test example). If p-value of a given test example is under some very low threshold (e.g. 0.05), this would signify that the such a sequence would only be generated at most 5% of the time by any i.i.d process, and is unlikely to belong to the probability distribution of the calibration set. In other words, the hypothesis being tested says "All examples in the sequence $(c_1, c_2, ..., c_M, c_{M+1})$ belong to the same probability distribution", and the hypothesis is rejected if $p(c_{M+1}) \leq \epsilon$ for some predetermined ϵ.

The p-value function can be constructed by comparing how different each example in the sequence is from all the other examples. This is possible using the measure of nonconformity. The measure of nonconformity intuitively corresponds to how *atypical* a sequence is, and maps a bag of examples and one additional example to a scalar $\eta_i \in \mathbb{R}$:

$$\eta_i = A(\wr c_1, ..., c_{i-1}, c_{i+1}, ..., c_{M+1}\int, c_i) \tag{4.7}$$

for each example c_i, thereby measuring how different it is from other examples in the bag $\wr c_1, ..., c_{i-1}, c_{i+1}, ..., c_{M+1}\int$. We use $\wr .\int$ to denote a bag since the order in which examples appear in the sequence will not have any impact on the non-conformity score η_i. In this work, the non-conformity measure is computed as follows:

$$\eta_i = A(\wr c_1, ..., c_{i-1}, c_{i+1}, ..., c_{M+1}\int, c_i) = \sum_{j=1, j\neq i}^{M+1} -\frac{c_i.c_j}{||c_i||||c_j||} \tag{4.8}$$

The p-value of c_{M+1} can now be calculated as follows:

$$p(c_{M+1}) = \frac{\#\{i = 1, 2, 3, ..., M : \eta_i \geq \eta_{M+1}\}}{M} \tag{4.9}$$

Once p-value of the test example has been calculated, the decision rule is that a positive prediction (test example belongs to same probability distribution as that of calibration set) is made when $p(c_{M+1}) > \epsilon$, where $\epsilon \in [0, 1]$. We choose ϵ to be 0.10 in our analysis, this means that for i.i.d. data, we would expect with 90% confidence that the new test example

is from the same distribution as that of calibration set. Only those test examples that have a positive prediction (from the conformal predictor) are passed on to the sepsis predictor to obtain a sepsis risk score. The above procedure is repeated for all samples in the testing set, and the sepsis risk score is obtained for only those examples whose p-value is higher than ϵ.

Figure 4.3 shows the schematic diagram of creating the calibration set after ADA training has been completed. Figure 4.1 shows the schematic diagram of COMPOSER model during testing phase, wherein the conformal predictor module is used for detecting distribution shift.

4.4.4 Data processing, training and hyperparameters

First, the Hospital-A ICU training set was normalized by subtracting the mean and dividing by the standard deviation (both of which were computed on Hospital-A ICU training set). Next, all remaining datasets were normalized using the mean and standard deviation statistics computed from the Hospital-A ICU training set. For handling missing data, we used a simple sample-and-hold approach in all the datasets.

Weighted input layer: The scaling factors α_i were all initialized to 1. *ADA model:* The learning rates for encoder, sepsis predictor and domain classifier were set to 0.01. Weighting factor λ used in the ADA model was fixed at 1. To minimize overfitting and to improve generalizability of the model, L1-L2 regularization was used with L2 regularization parameter set to 1e-3 for encoder and sepsis predictor, 1e-4 for domain classifier and L1 regularization parameter set to 1e-3 for encoder and sepsis predictor, 1e-4 for domain classifier. Mini-batch size for the source dataset was fixed at a total of 10000 windows (50% septic windows, 50% non-septic windows). Mini-batch size for the target dataset was set at 5000 windows. The encoder was made up of 1 hidden layer of dimension 40. The sepsis predictor was made up of 1 hidden layer of dimension 25. The domain classifier was made up of 1 hidden layer of size 25. Both sepsis predictor and domain classifier were further

followed by a fully connected layer and a softmax layer. *Conformal predictor:* Threshold ϵ was set at 0.10. Calibration set consisted of equal proportion of septic windows and non-septic windows. For the chosen septic patients, 2 septic windows were sampled at random. For the chosen non-septic patients, 2 non-septic windows were sampled at random.

We trained the COMPOSER model for a total of 500 epochs using Adam optimizer [91], with early stopping. All hyper-parameters of the model: Number and size of layers for encoder-sepsis predictor-domain classifier, learning rate, mini-batch size, L1 regularization parameter, and L2 regularization parameter were optimized using Bayesian optimization [92]. All pre-processing of data was performed using Numpy [93], with the rest of pipeline implemented using TensorFlow [94].

4.5 Clinical workflow aware AUC (C-AUC): An improved performance evaluation metric for sequential predictive models in healthcare

Traditionally, the Area Under the Curve (AUC) and Area under the Precision-Recall Curve (AUCpr) have been used to measure performance of predictive models developed using healthcare data. We use the term healthcare data to encompass both longitudinal data (wherein data of every patient is available in the form of a sequence) and static data (wherein a snapshot of information is available of every patient in the dataset). In this study, we are specifically interested in evaluating performance of models on longitudinal/sequential data. A primary disadvantage of using AUC/AUCpr metrics is that they ignore the sequential nature of data and treat all data points equally without any dependence on time. The clinical workflow AUC (C-AUC) is designed taking into consideration: 1) COMPOSER model is intended to be used as a clinical decision support (CDS) system, and 2) Sequential nature of data.

First modification: A likely use-case scenario of COMPOSER model as a CDS would be in alerting clinical staff whenever the COMPOSER score for a patient goes above a predetermined decision threshold. Once an alert has been fired, a nurse/clinician will have

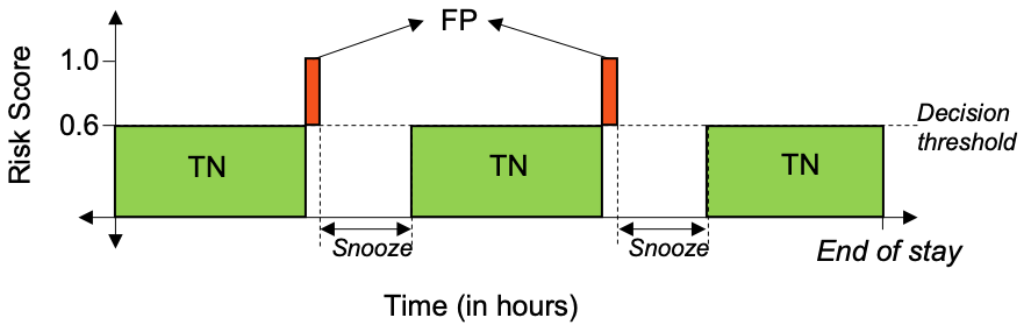


Figure 4.4: **Illustration of predictions being ignored during snooze period.** The first modification of C-AUC ignores predictions for the snooze period after predicted risk score crosses the decision threshold. The example shown in the above figure is for a patient who did not develop sepsis during the ICU stay.(FP = False Positive, TN = True Negative)

to evaluate the status of the patient to determine if initiation of treatment for sepsis is required. If the patient does not require treatment, the nurse/clinician can decide to re-evaluate after a few hours. As can be seen from this clinical workflow, once an alert has been issued by the COMPOSER model, any predictions from the model for next few hours will most likely be ignored by clinical staff - simply put, the model is snoozed for a few hours after an alert has been fired. The first modification that we make to the AUC/AUCpr metric is that any predictions after a positive prediction (onset of sepsis within then next four hours) has been made are ignored for a fixed duration (which we call as 'snooze duration'). Figure 4.4 shows an example of predictions being ignored during the snooze period.

Second modification: Sepsis is an illness that evolves progressively over time. It is difficult to exactly determine a discrete time-point as the time of onset of sepsis. Unfortunately, this fact is not considered in the case of AUC/AUCpr where ground truth labels (in case of four hour ahead prediction of sepsis) are defined to be non-septic until four hours prior to $t_{sepsis-3}$, and are defined to be septic between four hours prior to $t_{sepsis-3}$ and $t_{sepsis-3}$. Thus, there is a discrete transition from non-septic to septic labels which can unfairly penalize a predictive model if it makes a positive prediction for onset of sepsis prior to four hours from $t_{sepsis-3}$. All those time points at which the predictive model made a posi-

tive prediction, but fell outside the prediction horizon of four hours are counted as False Positives. In order to minimize penalizing the algorithm for making a positive prediction much earlier than the four hour prediction horizon, we modify the criterion for computing False Positives (FP), True Positives (TP), False Negatives (FN) and True Negatives (TN) as follows:

- FP: COMPOSER makes a positive prediction - M hours prior to $t_{sepsis-3}$ for septic patients or at any point of time for non-septic patients
- TP: COMPOSER makes a positive prediction - within M hours prior to $t_{sepsis-3}$
- FN: COMPOSER makes a negative prediction - between four hours prior to $t_{sepsis-3}$ and $t_{sepsis-3}$
- TN: COMPOSER makes a negative prediction four hours prior to $t_{sepsis-3}$ for septic patients or at any point of time for non-septic patients.

We consider M (which we will call 'positive prediction duration') to be 12 hours in our study, which means that we do no penalize the COMPOSER model for making a positive prediction of onset of sepsis up to 12 hours prior to $t_{sepsis-3}$. Note that we still penalize the COMPOSER model for making a negative prediction within four hours of $t_{sepsis-3}$. Figure 4.5 shows scenarios where a COMPOSER model prediction can fall into one of the four categories described above.

Finally, the C-AUC metric takes into consideration both the first modification and second modification described above.

4.6 Results

For training of all the models described in this section, ground truth labels (Sepsis-3 onset times) were required only for the Hospital-A ICU cohort. No ground truth labels were used from the Hospital-B ICU, Hospital-C ICU, Hospital-A ED and Hospital-B ED cohorts

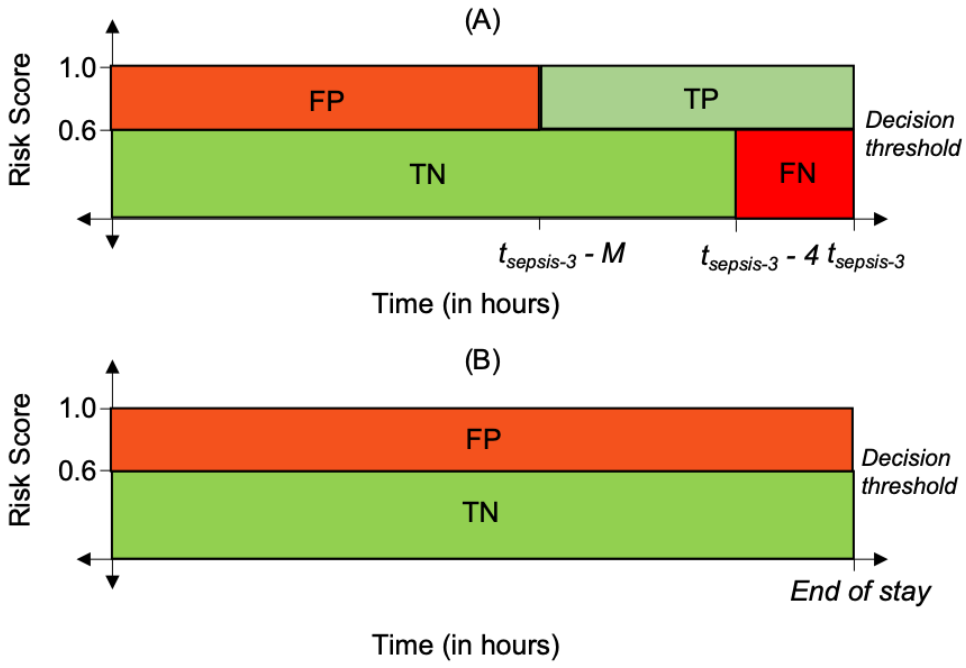


Figure 4.5: **Illustration of categorization of predictions from a predictive model according to the suggested second modification of C-AUC. (A)** The various scenarios where predictions could fall into one of four categories of TP, FP, TN or FN, for a septic patient. **(B)** The various scenarios where predictions could fall into one of two categories of FP or TN for a non-septic patient.

during training. The ground truth labels from all cohorts were used only to evaluate the performance of the trained models.

4.6.1 C-AUC as a performance evaluation metric

As described in Section 4.5, C-AUC consists of two hyperparameters: snooze duration and positive prediction duration. To study trends in C-AUC for various values of its hyperparameters, we performed an analysis wherein a 2 layered Feedfoward Neural Network (FFNN) was trained on the Hospital-A ICU dataset to predict onset of sepsis within four hours. The input to this model consisted of $[X_{dyamical}; X_{covar}]$ The resulting C-AUCs of the above model (performance evaluated on the Hospital-A ICU testing set) for Snooze durations of 0, 2, 4, 6, 8 and 12 hours; and for Positive prediction durations of 4, 6, 8 and 12 hours are shown in Table 4.1 and Table 4.2. We observed that with increasing

snooze duration and positive prediction duration, the C-AUC increased which could be attributable to the reduction of false positives. Consequently, we observed that the C-AUCpr decreased with increase in snooze duration and positive prediction duration which could be attributable to the reduction in total number of true positives (which leads to drop in precision). In our study, we fixed the snooze duration at 6 hours and positive prediction at 12 hours while computing C-AUC/C-AUCpr. Henceforth we will be reporting only the C-AUC/C-AUCpr for the aforementioned hyperparameters.

Table 4.1: Comparison of C-AUC values for various values of snooze duration and positive prediction duration. C-AUC values shown are for a trained FFNN model evaluated on Hospital-A ICU testing set.

	Prediction horizon before $t_{sepsis-3}$ where positive prediction not penalized *(positive prediction duration)*			
	4 hours	**6 hours**	**8 hours**	**12 hours**
C-AUC (Snooze = 0 hours)	0.858	0.876	0.889	0.906
C-AUC (Snooze = 2 hours)	0.865	0.883	0.895	0.911
C-AUC (Snooze = 4 hours)	0.870	0.887	0.899	0.915
C-AUC (Snooze = 6 hours)	0.878	0.894	0.905	**0.919**
C-AUC (Snooze = 8 hours)	0.880	0.896	0.909	0.922
C-AUC (Snooze = 12 hours)	0.876	0.895	0.908	0.922

Table 4.2: Comparison of C-AUCpr values for various values of snooze duration and positive prediction duration. C-AUCpr values shown are for a trained FFNN model evaluated on Hospital-A ICU testing set.

	Prediction horizon before $t_{sepsis-3}$ where positive prediction not penalized *(positive prediction duration)*			
	4 hours	**6 hours**	**8 hours**	**12 hours**
C-AUCpr (Snooze = 0 hours)	0.110	0.137	0.160	0.198
C-AUCpr (Snooze = 2 hours)	0.070	0.090	0.107	0.139
C-AUCpr (Snooze = 4 hours)	0.057	0.075	0.090	0.118
C-AUCpr (Snooze = 6 hours)	0.049	0.066	0.080	**0.102**
C-AUCpr (Snooze = 8 hours)	0.045	0.061	0.076	0.099
C-AUCpr (Snooze = 12 hours)	0.040	0.054	0.068	0.089

4.6.2 Model performance improves with weighted input layer

Due to workflow-related variations in frequency of laboratory measurements in ICUs and EDs, we hypothesized that when trained on ICU data and tested on ED data, a TSLM-weighted FFNN model (in this model X_{TSLM} is used to only scale $X_{dynamical}$, and not used as input feature to FFNN) would outperform a FFNN-T (input = $[X_{dyamical}; X_{covar}; X_{TSLM}]$) which has access to the missingness data directly as an input feature (and thus likely to overfit to the ICU workflow processes and patterns of data missingness). The data in Table 4.3 shows that the FFNN-T trained on Hospital-A ICU data (C-AUC=0.915) exhibited a significant drop in performance when applied to the Hospital-A ED population (C-AUC=0.692) and Hospital-B ED population (C-AUC=0.772). However the TSLM-weighted FFNN model was less likely to overfit to the ICU data and provided better performance on the Hospital-A ED population (C-AUC=0.763) and Hospital-B ED population (C-AUC=0.842) without a significant performance loss on the source Hospital-A ICU dataset (AUC=0.923). These results suggest that X_{TSLM} should not be used as a direct input feature to a FFNN, thus for the remainder of our analysis we will only be using it as part of the weighted input layer. In the case of training baseline FFNN models, the input features were $[X_{dyamical}; X_{covar}]$.

The TSLM-weighted FFNN (or FFNN with weighted input layer) model was also observed to perform better when trained on Hospital-A ICU data and tested on ICU data from other hospitals. Referring to Table 4.4, we observe that the TSLM-weighted FFNN performed slightly better compared to a FFNN (without any weighted input layer) on the Hospital-A ICU test data (C-AUC of 0.923 vs 0.919), Hospital-B ICU test data (C-AUC of 0.859 vs 0.842) and Hospital-C ICU test data (C-AUC of 0.871 vs 0.851). We also observed that the number of False Positives (measured at 80% sensitivity) were lesser in the case of TSLM-weighted FFNN as compared to a FFNN.

The above results indicate that a generic TSLM-weighted input layer can be used as the first layer in a deep learning model to capture the clinical 'intuition' about the validity

Table 4.3: Comparison of performance of models (FFNN-T vs FFNN with weighted input layer) trained on Hospital-A ICU dataset and tested on ED datasets. It can be observed that using missing data indicators (such as TSLM) as input features can lead to significant drop in performance (in our case FFNN-T) especially when evaluated on cohorts belonging to different level of care. However the extent of drop in performance is less significant in the case of FFNN with weighted input layer.

	FFNN-T[1]			**FFNN with weighted input layer**[2]		
Dataset	**C-AUC**	**C-AUCpr**	**FP**[#]	**C-AUC**	**C-AUCpr**	**FP**[#]
Hospital-A ICU test	0.915	0.110	10169	0.923	0.104	9878 ($\downarrow 2.9\%$)
Hospital-A ED test	0.692	0.031	27524	0.763	0.033	23550 ($\downarrow 14.4\%$)
Hospital-B ED test	0.772	0.068	72368	0.842	0.072	53854 ($\downarrow 25.6\%$)

[1] *Input* = $[X_{dyamical}; X_{covar}; X_{TSLM}]$

[2] X_{TSLM} *used only to scale* $X_{dyamical}$*, and not used as input feature to FFNN*

[#] *Number of False Positives measured at 80% sensitivity*

of a piece of imputed data. We hypothesized that imposing an exponentially decaying constraint on the scaling function with scaling factor α to be learned from the data would provide interpretable results. Figure 4.6 shows the scaling factors that were learnt after training a FFNN with weighted input layer on Hospital-A ICU dataset. It can be observed that the model learnt that an imputed heart rate (α=1.9) is supposed to have a shorter half-life than imputed lactate (α=0.078) or creatinine (α=0.75). A similar plot but for a FFNN model trained on Hospital-A ED dataset is shown in Figure 4.9. As expected, the scaling factors learnt by this model were lower compared to the model trained on ICU cohort.

Table 4.4: Comparison of performance of models (FFNN vs FFNN with weighted input layer) trained on Hospital-A ICU dataset and tested on Hospital-B and Hospital-C ICU datasets.

	FFNN[1]			**FFNN with weighted input layer**		
Dataset	**C-AUC**	**C-AUCpr**	**FP**[#]	**C-AUC**	**C-AUCpr**	**FP**[#]
Hospital-A ICU test	0.919	0.102	10059	0.923	0.104	9878 ($\downarrow 1.8\%$)
Hospital-B ICU test	0.842	0.064	37407	0.859	0.066	34726 ($\downarrow 7.2\%$)
Hospital-C ICU test	0.851	0.031	8690	0.871	0.033	7256 ($\downarrow 16.5\%$)

[1] *Input* = $[X_{dyamical}; X_{covar}]$

[#] *Number of False Positives measured at 80% sensitivity*

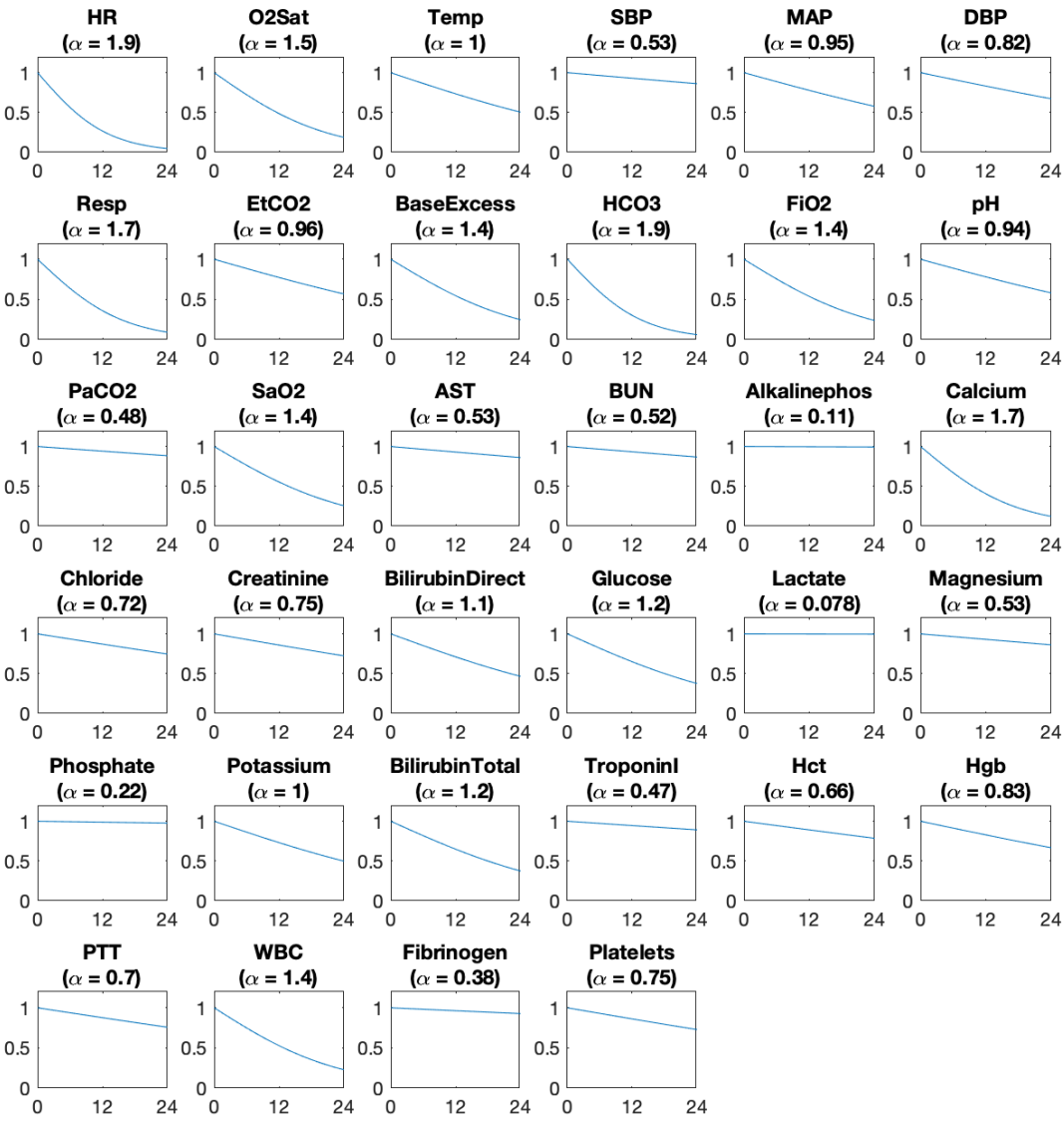


Figure 4.6: **Illustration of the weighting scheme learnt by a FFNN with weighted input layer model trained on Hospital-A ICU dataset** The plots shown depict the scaling function (varies from 0 to 1) imposed by the model for various values of Time Since Last Measurement (varies from 0 to 24 hours) of each of the 34 dynamical variables considered in our study.

4.6.3 Evaluating the performance of Adversarial domain adaptation with weighted input layer

Table 4.5: Adversarial Domain Adaptation with weighted input layer shows improved generalization performance over Adversarial Domain Adaptation. Performance of models on the target dataset is shown in this table. Performance of same models on the source dataset is shown in Table 4.15.

	ADA[1]			**ADA with weighted input layer**		
Source/Target	**C-AUC**	**C-AUCpr**	**FP**[#]	**C-AUC**	**C-AUCpr**	**FP**[#]
Hospital-A ICU / Hospital-B ICU	0.860	0.065	34802	0.870	0.069	33076 ($\downarrow 5.0\%$)
Hospital-A ICU / Hospital-C ICU	0.864	0.033	8252	0.879	0.030	7533 ($\downarrow 8.7\%$)
Hospital-A ICU / Hospital-A ED	0.821	0.039	19855	0.847	0.039	18325 ($\downarrow 7.7\%$)
Hospital-A ICU / Hospital-B ED	0.884	0.078	44589	0.893	0.083	40601 ($\downarrow 8.9\%$)

[1] *Input* = $[X_{dyamical}; X_{covar}]$
[#] *Number of False Positives measured at 80% sensitivity*

The focus of employing Adversarial Domain Adaptation (ADA) was to learn data representations that were robust to institution-specific variability and noise. We expected that for any given pair of cohorts - source dataset (labeled) and target dataset (unlabeled) - an ADA trained model would produce risk scores that generalize better to a target dataset without significant loss in predictive power on the source dataset. Results in Table 4.5 show that the ADA trained model performed better on the target datasets (when the source dataset was Hospital-A ICU) compared to a FFNN trained on Hospital-A ICU dataset (Table 4.4). The ADA trained model consisted of an encoder with 1 hidden layer (size of 40), sepsis predictor with 1 hidden layer (size of 25), and a fully connected layer at the end. When the source and target datasets had the same level of ICU care but were from different institutions, the ADA trained model performed better compared to the FFNN baseline - C-AUC of 0.860 on the Hospital-B ICU test set compared to C-AUC of 0.842. The ADA trained model was able to perform better on the target dataset without any significant loss in perfor-

mance on the source dataset - C-AUC of 0.913 on Hospital-A ICU test set. Similar patterns were observed for the Hospital-A ICU/Hospital-C ICU, Hospital-A ICU/Hospital-A ED and Hospital-A ICU/Hospital-B ED source/target combinations as well.

While the ADA trained model learned representations that are robust to institution-specific variability and noise, the addition of a weighted input layer provided further robustness to missingness-related deleterious factors. We observed this in results tabulated in Table 4.5. It can be observed that adding a weighted input layer improved the performance on target datasets in comparison to ADA model alone. The ADA with weighted input layer model performed especially well when the source and target datasets came from different levels of care. In the case of Hospital-A ICU/Hospital-B ED (source/target) datasets, the ADA with weighted input layer model (C-AUC=0.893) outperformed all the other baseline models[1]. Similar patterns in performance were observed for other source/target dataset combinations as well (Table 4.5). (The performance of ADA models discussed above on the source dataset is listed in Table 4.15)

The goal of employing domain adaptation was to adapt representations learned from the source domain to the target domain. We have tried to capture the alignment of the representations across both the domains during ADA training in Figures 4.7 and and 4.8. Each point on the clusters shown in the above figures is a 2D representation (projection) of the state of a patient, constructed via first learning a 25 dimensional representation (output of the encoder) followed by dimensionality reduction via the method of Uniform Manifold Approximation and Projection (UMAP) [140]. It can be seen that the ADA method is able to align the representations for the septic (and similarly non-septic) patients at the end of model training. The alignment is the most distinctive in the case of ADA involving Hospital-A ICU (source) and Hospital-B ED (target) datasets (Figure 4.8), wherein the representations of septic patients (similarly for non-septic patients) are misaligned during the beginning of training (this could mainly be attributed to the datasets belonging to different

[1]Baseline models include FFNN, FFNN with weighted input layer, ADA

levels of care) while at the end of training, the representations are aligned (but with some shift) in both the domains.

4.6.4 Evaluating the performance of COMPOSER model

Thus far we have shown that an ADA trained model with weighted input layer generalizes well to a target dataset, thereby learning representations that are both robust to institution-specific variability and level of care specific data missingness patterns. Although the goal of the ADA framework is to minimize distributional shifts that might potentially exist between datasets, it still lacks the ability to quantitatively determine the extent to which distributional shifts can be tolerated. By using a conformal prediction framework based on the above learned representations, we can explicitly determine the extent of distribution shifts that are tolerable to make a trustable prediction of sepsis risk score. The resulting pipeline consisting of the ADA trained model, weighted input layer and conformal predictor makes up the COMPOSER model. In this section, we consider scenarios where the source dataset is Hospital-A ICU cohort and study the performance of COMPOSER model on various target datasets - Hospital-B ICU/Hospital-C ICU/ Hospital-A ED/Hospital-B ED. The performance of COMPOSER model for the above four scenarios has been tabulated in Table 4.6, Table 4.7, Table 4.8 and Table 4.9.

For the conformal predictor, we chose ϵ to be 0.10 in our analysis which means that we could expect predictions on samples classified to be 'in-distribution' to be 90% correct. We would like to reiterate that first the ADA (with weighted input layer) model was trained on the source and target datasets, afterwhich a sub-sample of the source dataset was used to create the calibration set. This calibration set was used by the conformal predictor to determine whether to accept or reject a test sample for prediction of onset of sepsis. In our analysis, we found that the COMPOSER model performed better compared to a baseline C-FFNN[2] (FFNN with conformal predictor). The COMPOSER model trained using Hospital-

[2]C-FFNN: FFNN with conformal predictor: This was a 2 layer FFNN, wherein representations from the second layer of the network were used as input to the conformal predictor

Table 4.6: Performance of COMPOSER model trained using Hospital-A (source) and Hospital-B (target) ICU datasets. The scaling factors learnt by this COMPOSER model are shown in Figure 4.10. Also shown is the performance of a FFNN trained on Hospital-A ICU dataset and tested on Hospital-A and Hospital-B ICU datasets.

	C-FFNN[1]			**COMPOSER (Hospital-A ICU/Hospital-B ICU)**		
Dataset	**C-AUC**	**FP**[#]	**RW**[*]	**C-AUC**	**FP**[#]	**RW**[*]
Hospital-A ICU test (source)	0.925	9437	16102 (11.3%) (15912/190)	0.924	8347 ($\downarrow$ 11.6%)	15966 (11.2%) (15791/175)
Hospital-B ICU test (target)	0.869	32720	42649 (12.1%) (42267/382)	0.882	31558 ($\downarrow$ 3.6%)	42919 (12.2%) (42559/360)

[1] *Input* = $[X_{dyamical}; X_{covar}]$

[#] *Number of False Positives measured at 80% sensitivity*

[*] *Number of Rejected Windows (%) using conformal prediction. Also shown is - (number of rejected non-septic windows/rejected septic windows)*

A ICU as source data and Hospital-B ICU as the target data achieved a C-AUC of 0.882 on the Hospital-B ICU test set while still maintaining comparable performance on the source dataset (C-AUC=0.924). The rejection rate was found to be 12.2% and 11.2% for the Hospital-A and Hospital-B ICU datasets respectively. It can be observed that a majority of the rejected samples were from the non-septic class. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.10. A visualization of the alignment of representations of Hospital-A ICU and Hospital-B ICU datasets after ADA training is shown in Figure 4.7.

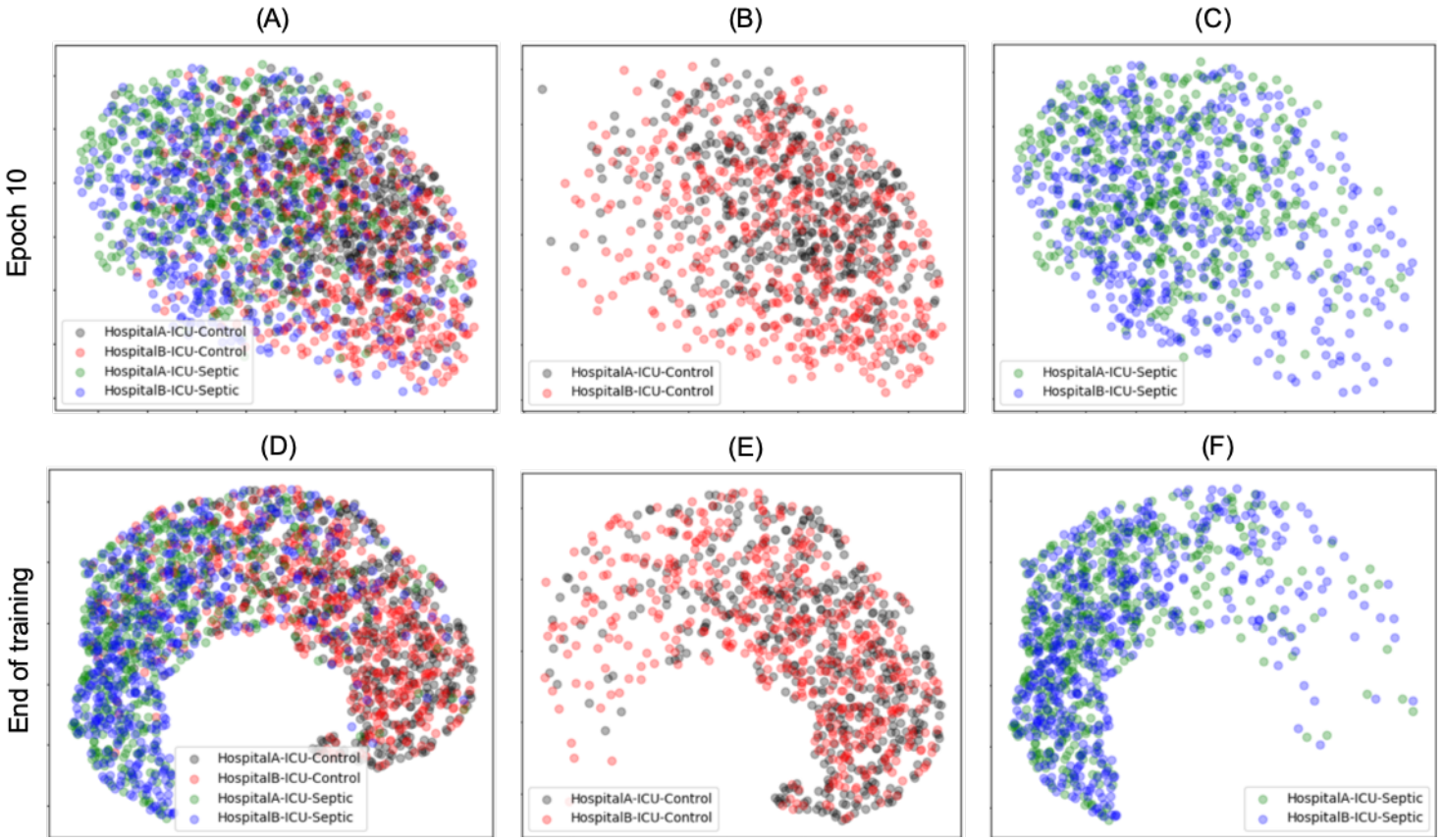


Figure 4.7: **Visualization of alignment of representations learned by COMPOSER model.** The plots shown are based on COMPOSER model trained on Hospital-A ICU (source dataset) and Hospital-B ICU (target dataset) cohorts. A random sample of septic and non-septic (control) patients were chosen from the source and target datasets and their encoder representations were projected onto a 2 dimensional space using the technique of Uniform Manifold Approximation and Projection (UMAP) [140]. **(A)-(C)** The distribution of representations at Epoch 10 during COMPOSER training. **(D)-(F)** The distribution of representations at the end of COMPOSER training. Representations corresponding to only the septic patients are shown in (B) and (E) while representations corresponding to control patients are shown in (C) and (F).

Table 4.7: Performance of COMPOSER model trained using Hospital-A (source) and Hospital-C (target) ICU datasets. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.11. Also shown is the performance of a FFNN trained on Hospital-A ICU dataset and tested on Hospital-A and Hospital-C ICU datasets.

	C-FFNN[1]			**COMPOSER (Hospital-A ICU/Hospital-C ICU)**		
Source/Target	**C-AUC**	**FP**[#]	**RW**[*]	**C-AUC**	**FP**[#]	**RW**[*]
Hospital-A ICU test (source)	0.925	9437	16102 (11.3%) (15912/190)	0.922	9572 ($\uparrow$ 1.4%)	15304 (10.7%) (15125/179)
Hospital-C ICU test (target)	0.856	8072	9895 (11.6%) (10557/53)	0.885	6743 ($\downarrow$ 16.5%)	10651 (12.5%) (10604/47)

[1] *Input* = $[X_{dyamical}; X_{covar}]$

[#] *Number of False Positives measured at 80% sensitivity*

[*] *Number of Rejected Windows (%) using conformal prediction. Also shown is - (number of rejected non-septic windows/rejected septic windows)*

A COMPOSER model trained using Hospital-A ICU data as the source and Hospital-C ICU as the target achieved a C-AUC of 0.885 on the Hospital-C ICU test set which was 0.031 points improvement in C-AUC compared to C-FFNN. The rejection rate was found to be 12.2% for the Hospital-C ICU test set. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.11.

The performance of COMPOSER remained consistent when the source and target datasets were from different levels of care. A COMPOSER model trained using Hospital-A ICU data as the source and Hospital-A ED as the target achieved a C-AUC of 0.851 on the Hospital-A ED test set compared to C-AUC 0.830 obtained by the C-FFNN. The rejection rate was found to be 8.4% for the Hospital-A ED test set. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.12.

Finally, a COMPOSER model trained using Hospital-A ICU data as the source and Hospital-B ED as the target achieved a C-AUC of 0.899 on the Hospital-B ED test set compared to C-AUC 0.890 obtained by the C-FFNN. The rejection rate was found to be 7.4% for the Hospital-B ED test set. Overall, it was observed that for ϵ of 0.10, the rejection rate was around 5-10% across all the cohorts. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.13. A visualization of the alignment of

Table 4.8: Performance of COMPOSER model trained using Hospital-A ICU (source) and Hospital-A (target) ED datasets. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.12. Also shown is the performance of a FFNN trained on Hospital-A ICU dataset and tested on Hospital-A ICU and Hospital-A ED datasets.

	C-FFNN[1]			**COMPOSER (Hospital-A ICU/Hospital-A ED)**		
	C-AUC	**FP**$^{\#}$	**RW***	**C-AUC**	**FP**$^{\#}$	**RW***
Hospital-A ICU test (source)	0.925	9437	16102 (11.3%) (15912/190)	0.925	9733 ($\uparrow$ 3.1%)	14896 (10.5%) (14727/169)
Hospital-A ED test (target)	0.830	19882	14093 (9.3%) (13972/121)	0.851	17868 ($\downarrow$ 10.3%)	12749 (8.4%) (12575/174)

[1] *Input* = $[X_{dyamical}; X_{covar}]$

[#] *Number of False Positives measured at 80% sensitivity*

[*] *Number of Rejected Windows (%) using conformal prediction. Also shown is - (number of rejected non-septic windows/rejected septic windows)*

representations of Hospital-A ICU and Hospital-B ED datasets after ADA training is shown in Figure 4.8.

Table 4.9: Performance of COMPOSER model trained using Hospital-A ICU (source) and Hospital-B (target) ED datasets. Additionally, the scaling factors learnt by this COMPOSER model are shown in Figure 4.13. Also shown is the performance of a FFNN trained on Hospital-A ICU dataset and tested on Hospital-A ICU and Hospital-B ED datasets.

	C-FFNN[1]			**COMPOSER (Hospital-A ICU/Hospital-B ED)**		
	C-AUC	**FP**$^{\#}$	**RW***	**C-AUC**	**FP**$^{\#}$	**RW***
Hospital-A ICU test (source)	0.925	9437	16102 (11.3%) (15912/190)	0.927	9689 ($\uparrow$ 3.1%)	15420 (10.8%) (15248/173)
Hospital-B ED test (target)	0.890	41106	23062 (5.9%) (22381/681)	0.899	39619 ($\downarrow$ 3.6%)	28774 (7.4%) (28145/629)

[1] *Input* = $[X_{dyamical}; X_{covar}]$

[#] *Number of False Positives measured at 80% sensitivity*

[*] *Number of Rejected Windows (%) using conformal prediction. Also shown is - (number of rejected non-septic windows/rejected septic windows)*

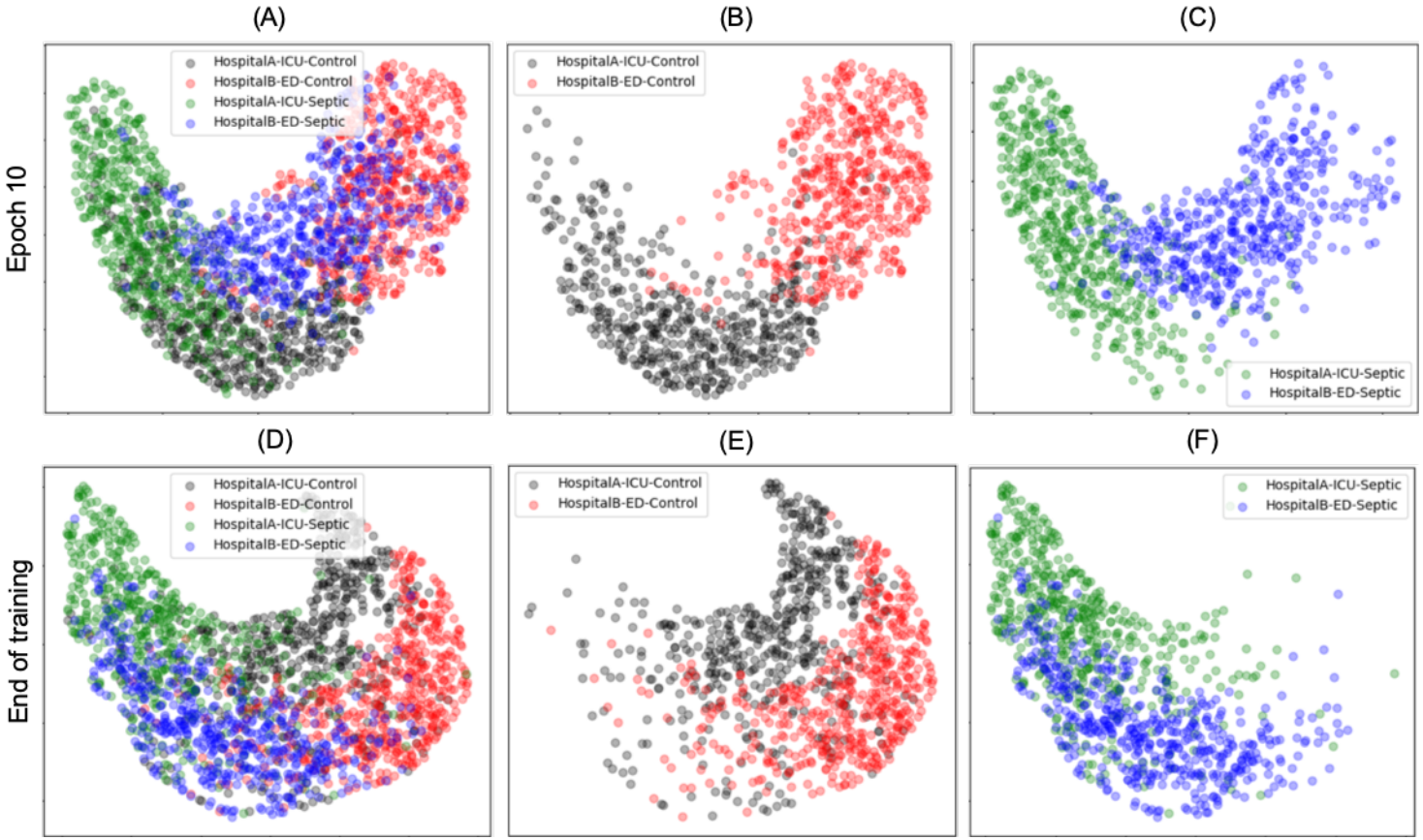


Figure 4.8: **Visualization of alignment of representations learned by COMPOSER model.** The plots shown are based on COMPOSER model trained on Hospital-A ICU (source dataset) and Hospital-B ED (target dataset) cohorts. A random sample of septic and non-septic (control) patients were chosen from the source and target datasets and their encoder representations were projected onto a 2 dimensional space using the technique of UMAP. **(A)-(C)** The distribution of representations at Epoch 10 during COMPOSER training. **(D)-(F)** The distribution of representations at the end of COMPOSER training. Representations corresponding to only the septic patients are shown in (B) and (E) while representations corresponding to control patients are shown in (C) and (F).

4.6.5 Validation of COMPOSER predictions with chart reviewed data

The guidelines recommended by Third International Consensus Definitions for Sepsis (Sepsis-3) has been used by the authors of this study to obtain labels for onset of sepsis ($t_{sepsis-3}$) in all the five patient cohorts. As mentioned in Section 1.2, there are numerous other criteria for sepsis that are still in wide use. One has to be cognizant of two points when employing these rule based criteria: 1) they are at best an approximate attempt to identify the onset time of this complex disease, and 2) they do not account for any confounding factors that might be contributing to symptoms manifesting as sepsis thereby leading to incorrect labelling of sepsis. In order to obtain a gold standard for onset of sepsis, one will have to manually go through the EHR record (with access to all notes from attending clinical staff, laboratory orders, culture orders etc.) of a patient and verify that there are no other confounding factors (eg. presence of other critical illness such as cancer) that are contributing to a patient displaying symptoms for sepsis.

As part of a Quality Improvement (QI) program at Hospital-A Health system, a quarterly review is conducted to evaluate the rate of compliance with SEP-1 care bundle [19]. The SEP-1 bundle advocates for obtaining blood cultures, administering broad spectrum antibiotics, measuring lactate, and starting appropriate fluid resuscitation if clinically indicated, all within 3 hours of clinical recognition of sepsis. A random sample of 60 patients who are admitted (to the two hospitals within the Hospital-A health system) in the quarter of interest are chosen, and are manually reviewed by the quality improvement team (consisting of clinical staff and hospital administrators) to identify the onset of sepsis (according to CMS guidelines) and if SEP-1 bundle compliance was met within three hours of onset of sepsis. We obtained access to this higher quality labels for the period between January 1^{st} 2016 and March 28^{th} 2019. Excluding all the transfer patients[3] and all patients whose onset time of sepsis was within the first hour of admission to ED/ICU, the cohort consisted

[3]Patients who were received from an inpatient, outpatient or emergency/observation department of an outside hospital or from an ambulatory surgery center

of 667 patients out of whom 524 were tagged septic (according to Sepsis CMS guidelines). Out of the 667 patients, 404 patients were present in our combined Hospital-A ICU and ED cohorts. The reduction in the number of patients was due to restricting the patients to only ICU and ED admits. We will be using this cohort (Hospital-A-QI cohort) of 404 patients for our analysis in this section. Out of the 404 patients in the Hospital-A-QI cohort, 311 patients were tagged septic (Sepsis CMS guidelines have been used to identify onset time of sepsis in Hospital-A-QI cohort).

Table 4.10: Evaluating accuracy of various rule based methods for identifying onset time of sepsis versus chart reviewed data. Also shown is the accuracy of predictions from COMPOSER model evaluated against chart reviewed data. For results shown in this table, the COMPOSER model was trained on the combined Hospital-A ICU and ED cohorts. A description of each of the sepsis criteria used for this analysis is available in Table 4.12.

	Sensitivity	**Specificity**	**PPV/Precision**	**F1 score**
Sepsis-2 [17]	0.778	0.355	0.801	0.790
Sepsis-3 [1]	0.559	0.688	0.857	0.677
Sepsis-CDC [18]	0.643	0.774	0.905	0.752
Sepsis-CMS[@] [19, 141]	0.707	0.774	0.913	0.797
Sepsis-CMS-Causal[$]	0.733	0.817	0.931	0.820
Any of definitions	0.875	0.258	0.798	0.834
COMPOSER[*]	0.852	0.548	0.863	0.858

[@] *Clinical suspicion of infection is determined using the same criterion specified in the Sepsis-3 guidelines (Ordering of Antibiotics and Cultures within a certain timeperiod). We do not look at documentation of suspected infection in EHR notes*

[$] *This is a causal implementation of the Sepsis-CMS criterion. Conditions for Suspicion of infection, SIRS and organ should be satisfied within the last six hours.*

[*] *Threshold for COMPOSER prediction set at 85% sensitivity level*

We evaluated the accuracy of Sepsis-3 guidelines by comparing the outcomes on patients from applying the Sepsis-3 criterion against the outcomes in the Hospital-A-QI cohort. We evaluated the outcomes on the patient level i.e. evaluating whether a patient was tagged septic or not, with the outcomes in Hospital-A-QI cohort considered as ground truth labels and outcomes using the sepsis-3 guidelines as predicted labels. The sensitivity, specificity, Positive predictive value (PPV) and F1 score for were computed based on the following measures:

- True Positives: Patient tagged septic in Hospital-A-QI cohort and the onset time of sepsis according to Sepsis-3 guidelines was within 12 hours prior to onset time of sepsis (as determined in Hospital-A-QI cohort) or three hours after.

- False Negatives: Patient tagged septic in Hospital-A-QI cohort and patient tagged non-septic by Sepsis-3 guidelines.

- True Negatives: Patient tagged non-septic in Hospital-A-QI cohort and patient tagged non-septic by Sepsis-3 guidelines.

- False Positives: Patient tagged non-septic in Hospital-A-QI cohort and patient tagged septic by Sepsis-3 guidelines.

The results obtained by applying Sepsis-3 guidelines on the Hospital-A-QI cohort is shown in Table 4.10. The F1 score for applying the sepsis-3 guidelines was found to be 0.677. The Sepsis-3 criterion had a higher F1 score when evaluated on only ICU admissions in the Hospital-A-QI cohort (F1 score=0.800). The lag between onset time of sepsis in the Hospital-A-QI cohort and the onset time identified by Sepsis-3 criterion was found to be 1.18 [0.24, 2.88] hours (Median [Interquartile range]). This means that the Sepsis-3 criterion was satisfied on an average 1.18 hours earlier to the criterion used in Hospital-A-QI cohort. Considering these statistics, we would expect that a predictive model such as COMPOSER trained to predict Sepsis-3 labels four hours in advance would have a higher F1-score on the Hospital-A-QI cohort.

For evaluating the performance of COMPOSER model on the Hospital-A-QI cohort, it was trained (using Sepsis-3 labels) using Hospital-A ICU cohort as the source dataset and Hospital-A ED cohort as the target dataset, with the exclusion of patients who were present in the Hospital-A-QI cohort. This made sure that the COMPOSER model had no prior access to patients present in the Hospital-A-QI cohort. The trained COMPOSER model was then applied sequentially on every patient in the Hospital-A-QI cohort to obtain prediction of onset of sepsis for every hour that the patient was in the ED or ICU. For all

Table 4.11: Comparison of COMPOSER model performance on the Hospital-A-QI cohort for various sensitivity levels threshold levels.

Sensitivity	Specificity	PPV/Precision	F1 score
0.502	0.882	0.934	0.653
0.598	0.796	0.907	0.721
0.698	0.731	0.897	0.785
0.797	0.581	0.864	0.829
0.852	0.548	0.863	0.858
0.897	0.430	0.840	0.868

septic patients in the Hospital-A-QI cohort, we considered a COMPOSER prediction to be septic if the predicted risk score crossed a decision threshold within 12 hours prior to onset time of sepsis (as determined in Hospital-A-QI cohort) or three hours after. The resulting performance of the COMPOSER model is shown in Table 4.11. The results shown in Table 4.11 are based on thresholds that were chosen at different sensitivity levels. It can be seen that the COMPOSER model achieved highest F1 score (0.858) for threshold level corresponding to 85% sensitivity.

We have additionally implemented some of the other sepsis guidelines that are in wide use currently and have shown their performance in matching with the labels of Hospital-A-QI cohort in Table 4.10.

4.7 Discussion

In this work, a deep learning model (COMPOSER) was used to learn representations that minimize variations in clinical workflow patterns in data across levels-of-care and different healthcare systems. COMPOSER was trained to accurately predict the likelihood of sepsis in patients admitted to the ED or ICU upto four hours in advance. The COMPOSER model consisted of three components: 1) a weighted input layer (with the weighting factors learned from data) that was designed to handle missing data and variations in data measurement frequency, 2) an Adversarial Domain Adaptation procedure to learn representations that minimize healthcare system specific variations, and 3) a conformal predictor that estab-

lished the level of data distribution shift that was tolerable to make a risk score prediction. A key advantage of the COMPOSER model was in its ability to adapt (or generalize) to new unlabeled target patient population; therefore, gold-standard labels which are often expensive to obtain are not required to deploy the model at a new center. COMPOSER was developed to predict $t_{sepsis-3}$ (Sepsis-3 criterion) in this study. Additionally, we introduced a new performance evaluation metric, C-AUC, that was designed taking into consideration the CDS requirements of a predictive model such as COMPOSER and the sequential nature of healthcare data.

The generalizability of COMPOSER was shown on data from over 480,000 patients comprising of ICU and ED cohorts from Hospital-A, ICU cohort from Hospital-C, and ICU and ED cohorts from Hospital-B. We observed that the COMPOSER model achieved a C-AUC of around 0.92 on the source Hospital-A ICU testing set, whilst achieving C-AUCs of around 0.88 on the two target ICU cohorts (Hospital-C and Hospital-B ICU testing sets) separately for four hour ahead prediction of $t_{sepsis-3}$. Similarly, the performance of COMPOSER was consistent when the target cohorts were from different level of care compared to the source cohort - C-AUC of 0.851 on Hospital-A ED testing set and C-AUC of 0.899 on the Hospital-B ED testing set when the source cohort was Hospital-A ICU. Finally, there was high concordance between the sepsis risk score predictions from COMPOSER and sepsis onset labels obtained through manual chart review.

The proposed weighted input layer showed improvement in performances by around 2 points in C-AUC on external testing cohorts over a baseline FFNN without the weighted input layer. The improvement was significant when a model trained on ICU cohort was tested on ED cohorts (Table 4.3). While the weighted input layer was designed to scale the value of a feature (with the scaling factor learned during training) based on the time since it was last measured (TSLM), a number of other studies have approached the problem of imputing missing data by concatenating a missing value indicator to the inferred values for missing labs [142, 143, 108], with the hope that complex models such as neural networks

can learn arbitrary features of the missing data that are helpful for downstream prediction tasks. Notably, a recent study [144] proposed a constrained network architecture to learn separate representations for masking (indicating observed versus missing/imputed values) and time interval (encapsulating the input observation patterns) related information in time series with missing data. While all these single-center studies report performance improvement with the incorporation of missing data indicators, there are little to no studies of the ability of such models to generalize well to new patient populations. In our work, we have shown the generalizability of using a weighted input layer on five different cohorts from three different hospital systems and across two different levels of care.

The Adversarial Domain Adaptation (ADA) technique used in this study allowed for successfully adapting representations learned from a source dataset (where labels were available) to a target dataset (where labels were not available, a likely scenario would include data from a new institution). The ADA training procedure was shown to perform well in scenarios where target datasets were from a different level of care compared to the source dataset eg. Hospital-A ICU (source) and Hospital-B ED (target). In literature, a majority of the clinical ML models that have been validated across different health systems are either trained from scratch or are fine-tuned (visa transfer learning) on every new patient cohort [24, 25, 50], or when applied out-of-the box often exhibit significant degradation in performance [51]. The advantage of employing a training procedure such as ADA was to transfer knowledge learned from the source dataset whilst overcoming the requirement of gold-standard labels on a target population.

The representations (output from the encoder) obtained by the combination of ADA and weighted input layer are robust to care-level and institution-specific patterns in data missingness. In order to establish the extent to which shifts in distribution of data was tolerable to make a prediction, the method of conformal prediction was employed on these representations. The rejection rate was found to be around 5-10% across all the cohorts, for ϵ of 0.10. Some of the advantages that arise from establishing a metric to quantify the

extent of tolerable distributional shift include: 1) it potentially satisfies a key requirement of FDA regulations on SaMD, by establishing the 'conditions for use' of the COMPOSER score, and 2) when deployed as a real-time system, one could track the average rejection rates over time and identify any potential drifts in data distribution.

In general, statistical evaluation methods (such as the AUC) have a limited applicability when evaluating the clinical utility of such algorithms, although they can provide quantitative metrics for the comparison of various algorithms. In practice, performance metrics are only meaningful when coupled with appropriate clinical protocols that describe the course of action in response to the associated risk alerts. The clinical workflow aware AUC (C-AUC) metric proposed in this study takes into account simple clinical actions such as 'snoozing' the alarm for X hours if the patient did not meet the clinical threshold to initiate therapy, not penalizing the algorithm for making a positive prediction earlier than the prediction horizon. By taking into consideration the above factors, the C-AUC metric brings us one step closer to evaluating the actual performance of COMPOSER when deployed as a CDS system.

In this chapter, we have shown that by carefully taking into consideration the processes that are involved in generation/collection of healthcare data and the likely workflow processes of CDS systems, we can develop predictive models that can generalize well to new target populations. Although the COMPOSER model was trained to predict the onset of sepsis in this study, the techniques used to construct COMPOSER can be utilized for building predictive models for other types of physiological decompensation such as AKI.

4.8 Appendix

Table 4.12: Description of defined time points utilized in this chapter.

Time point	**Criteria**
$t_{suspicion}$	Clinical suspicion of infection identified as the earlier timestamp of antibiotics and blood cultures within a specified duration. (If antibiotics were given first, the cultures must have been obtained within 24 hours. If cultures were obtained first, then antibiotic must have been subsequently ordered within 72 hours)
t_{SOFA}	The occurrence of end organ damage as identified by a two-point deterioration in SOFA score within a 6-hour period
$t_{sepsis-3}$	The onset time of sepsis-3 is marked when both $t_{suspicion}$ and t_{SOFA} have happened within close proximity to each other. Specifically, t_{SOFA} must occur 24 hours before $t_{suspicion}$ or up to 12 hours after the $t_{suspicion}$ (t_{SOFA} + 24 hours $> t_{suspicion} > t_{SOFA}$ - 12 hours). The earlier of the t_{SOFA} or $t_{suspicion}$ was assigned to $t_{sepsis-3}$.
t_{eSOFA}	The occurrence of end organ damage as identified by one point or higher eSOFA score within a 6-hour period [145]
$t_{sepsis-CDC}$	The onset time of $t_{sepsis-CDC}$ is marked when both $t_{suspicion}$ and t_{eSOFA} have happened within close proximity to each other. Specifically, t_{eSOFA} must occur 24 hours before $t_{suspicion}$ or up to 12 hours after the $t_{suspicion}$ (t_{eSOFA} + 24 hours $> t_{suspicion} > t_{eSOFA}$ - 12 hours). The earlier of the t_{eSOFA} or $t_{suspicion}$ was assigned to $t_{sepsis-CDC}$
$t_{cmsSOFA}$	The occurrence of end organ damage as specified in the SEP-1 CMS manual [141]
$t_{sepsis-CMS}$	All three of the signs of severe sepsis should occur within a 6-hour window: 1) $t_{suspicion}$, 2) $\geq$ 2 Systemic Inflammatory Response Syndrome criteria, and 3) $t_{cmsSOFA}$. The onset time of $t_{sepsis-CMS}$ is the time at which the last sign of severe sepsis within that 6-hour window is noted.
$t_{suspicion-Causal}$	Clinical suspicion of infection identified as the earlier timestamp of antibiotics and blood cultures within a specified duration. (Both antibiotics and cultures should have been ordered in the previous 6 hours)
$t_{sepsis-CMS-Causal}$	All three of the signs of severe sepsis should occur within a 6-hour window: 1) $t_{suspicion-Causal}$, 2) $\geq$ 2 Systemic Inflammatory Response Syndrome criteria, and 3) $t_{cmsSOFA}$. The onset time of $t_{sepsis-CMS}$ is the time at which the last sign of severe sepsis within that 6-hour window is noted.

Table 4.13: Comparison of performance of models (FFNN vs FFNN with weighted input layer) trained on Hospital-A ICU dataset. The performance of the models on the Hospital-A ICU, Hospital-C ICU and Hospital-B ICU training sets are shown.

	FFNN[1]		**FFNN with weighted input layer**	
Dataset	**C-AUC**	**C-AUCpr**	**C-AUC**	**C-AUCpr**
Hospital-A ICU train	0.927	0.107	0.927	0.111
Hospital-B ICU train	0.868	0.069	0.876	0.071
Hospital-C ICU train	0.843	0.028	0.861	0.031

[1] *Input* = $[X_{dyamical}; X_{covar}]$

Table 4.14: Performance of baseline FFNN[1] model (trained on Hospital-A ICU) on Hospital-A and Hospital-B ED datasets

Dataset	**C-AUC**	**C-AUCpr**
Hospital-A ICU test	0.919	0.102
Hospital-A ED test	0.829	0.037
Hospital-B ED test	0.881	0.072

[1] *Input* = $[X_{dyamical}; X_{covar}]$

Table 4.15: ADA with weighted input layer shows improved generalization performance over ADA. Performance of models on the source dataset is shown in this table. Performance of same models on the target dataset is shown in Table 4.5.

	ADA[1]		**ADA with weighted input layer**	
Source/Target	**C-AUC**	**C-AUCpr**	**C-AUC**	**C-AUCpr**
Hospital-A ICU / Hospital-B ICU	0.913	0.097	0.918	0.103
Hospital-A ICU / Hospital-C ICU	0.916	0.102	0.915	0.101
Hospital-A ICU / Hospital-A ED	0.913	0.102	0.921	0.103
Hospital-A ICU / Hospital-B ED	0.920	0.101	0.921	0.104

[1] *Input* = $[X_{dyamical}; X_{covar}]$

Table 4.16: List of clinical variables used in this study.

Variable	Measurement Unit
Vital Signs (Dynamical Features)	
Heart rate	beats/minute
Pulse oximetry	%
Temperature	°C
Systolic BP	mmHg
Mean Arterial Pressure	mmHg
Diastolic BP	mmHg
Respiration rate	breaths per minute
End tidal CO_2	mmHg
Laboratory values (Dynamical Features)	
Excess bicarbonate	mmol/L
Bicarbonate	mmol/L
Fraction of inspired Oxygen	%
pH	-
Partial pressure of CO_2 from arterial blood	
Oxygen saturation from arterial blood	%
Aspartate transaminase	IU/L
Blood Urea Nitrogen	mg/dL
Alkaline phosphate	IU/L
Calcium	mg/dL
Chloride	mmol/L
Creatinine	mg/dL
Bilirubin direct	mg/dL
Serum Glucose	mg/dL
Lactic acid	md/dL
Magnesium	mmol/dL
Phosphate	mg/dL
Potassium	mmol/L
Total Bilirubin	mg/dL
Troponin I	ng/mL
Hematocrit	%
Hemoglobin	g/dL
Partial Thromboplastin Time	seconds
White Blood Cell count	count*$10^3/\mu$L
Fibrinogen	mg/dL
Platelets	count*$10^3/\mu$L
Demographics	
Age	Years
Gender	-
Units	Medical/Surgical ICU unit
Hours between hospital admit and ICU admit	hours
ICU length of stay	hours

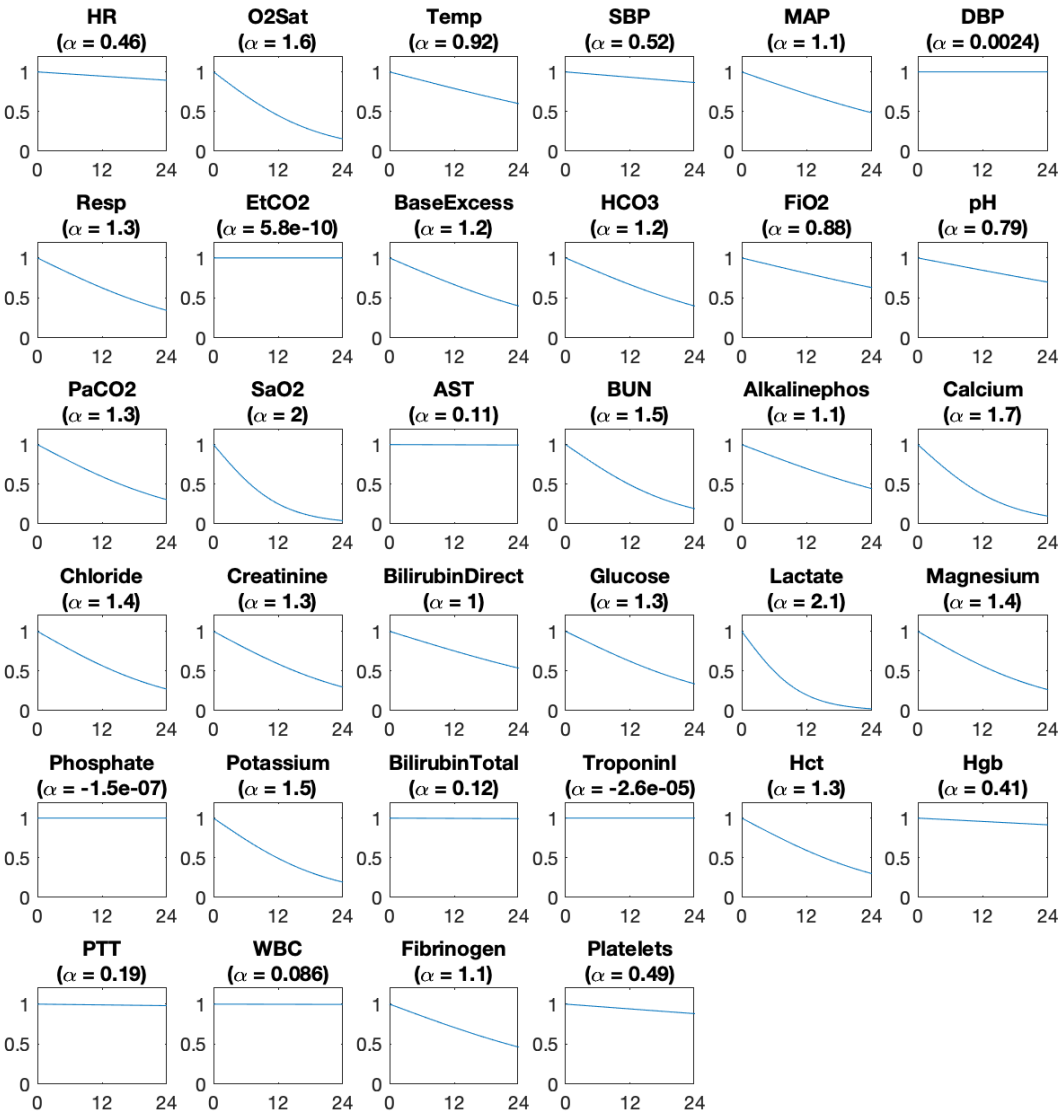


Figure 4.9: **Illustration of the weighting scheme learnt by a FFNN with weighted input layer model trained on Hospital-A ED dataset** The plots shown depict the scaling function (varies from 0 to 1) imposed by the model for various values of Time Since Last Measurement (varies from 0 to 24 hours) of each of the 34 dynamical variables considered in our study.

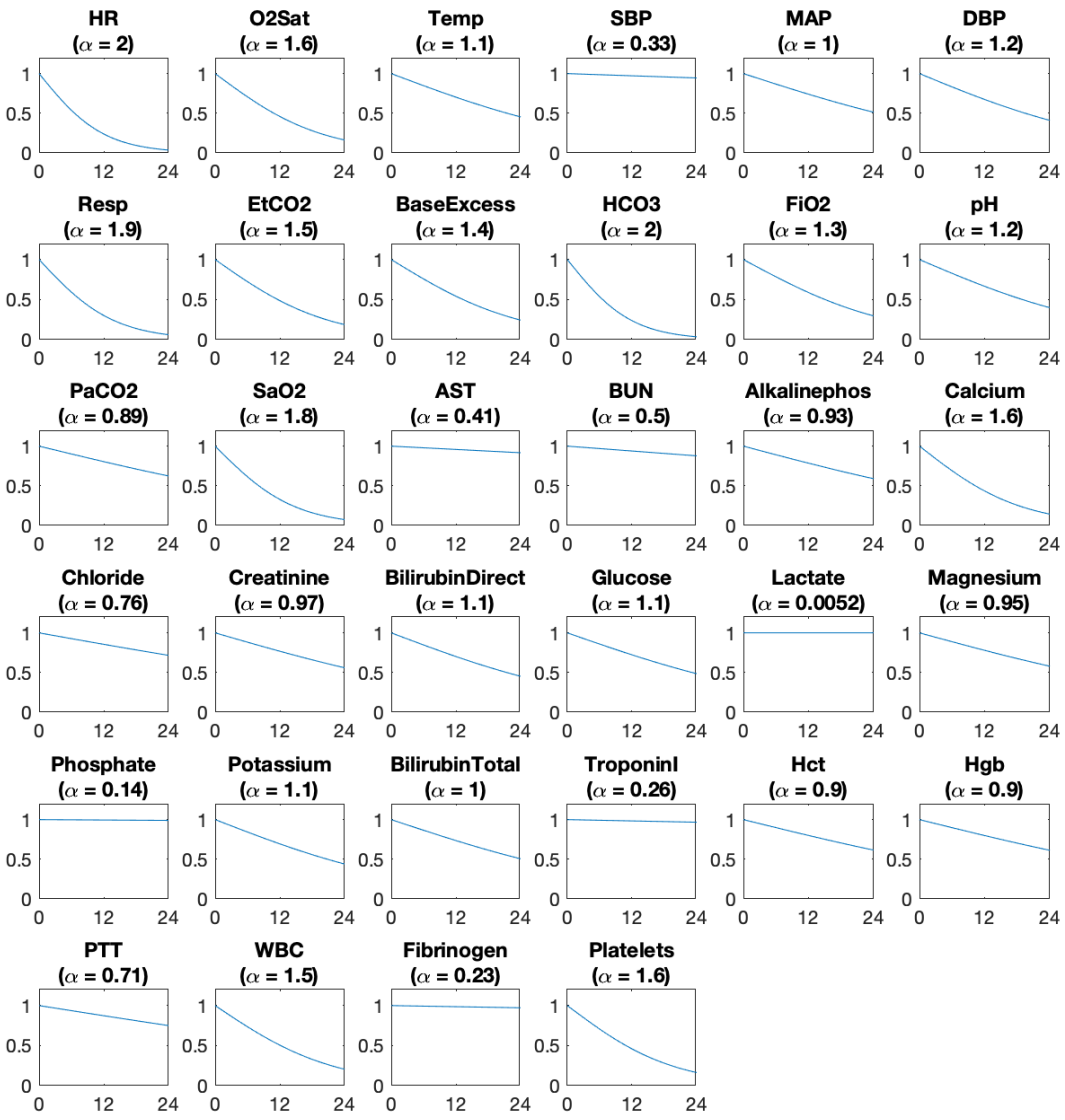


Figure 4.10: **Illustration of the weighting scheme learnt by a COMPOSER model trained on Hospital-A ICU (source dataset) and Hospital-B ICU (target dataset) cohorts.** The plots shown depict the scaling function (varies from 0 to 1) imposed by the model for various values of Time Since Last Measurement (varies from 0 to 24 hours) of each of the 34 dynamical variables considered in our study.

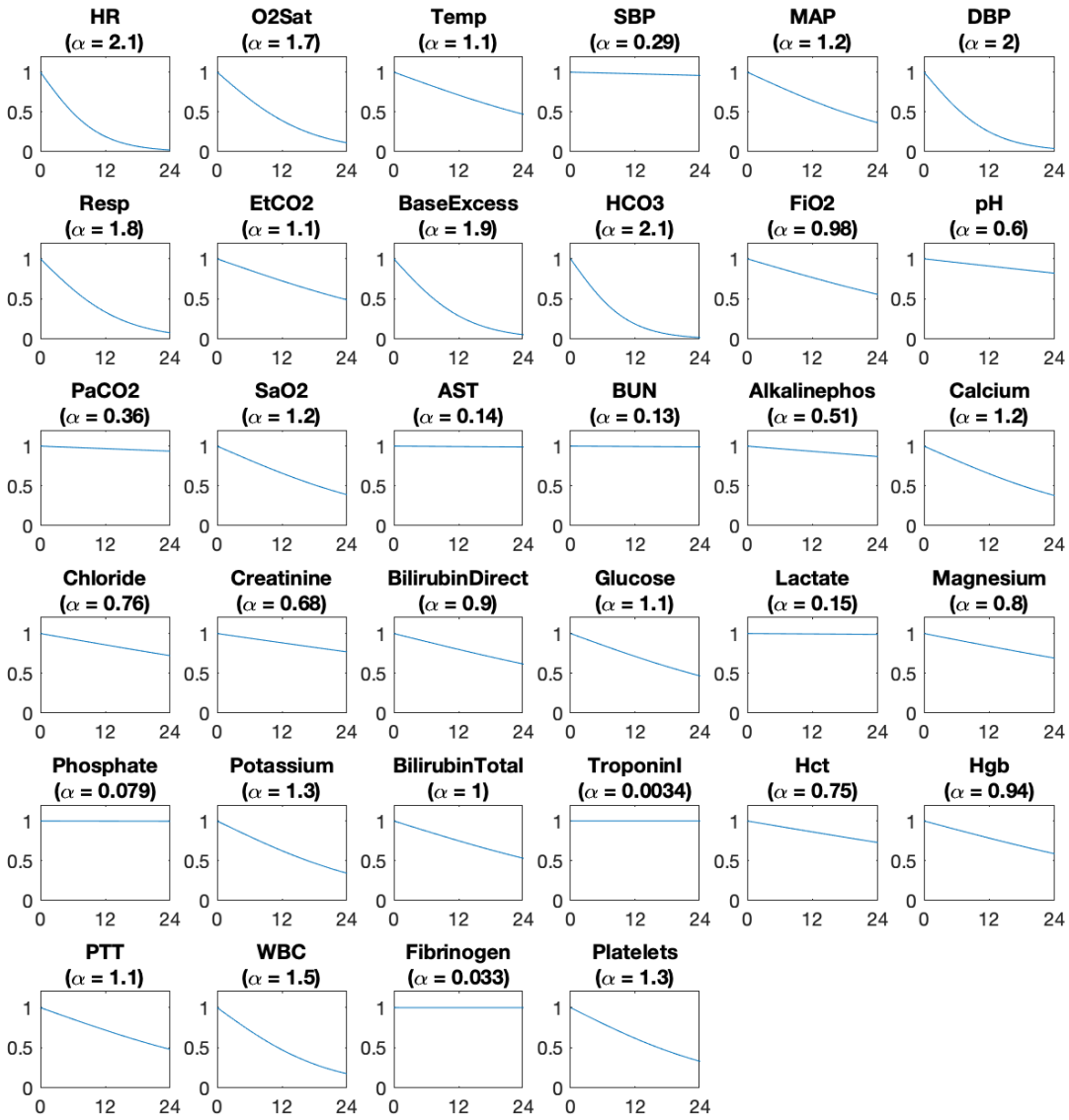


Figure 4.11: **Illustration of the weighting scheme learnt by a COMPOSER model trained on Hospital-A ICU (source dataset) and Hospital-C ICU (target dataset) cohorts.** The plots shown depict the scaling function (varies from 0 to 1) imposed by the model for various values of Time Since Last Measurement (varies from 0 to 24 hours) of each of the 34 dynamical variables considered in our study.

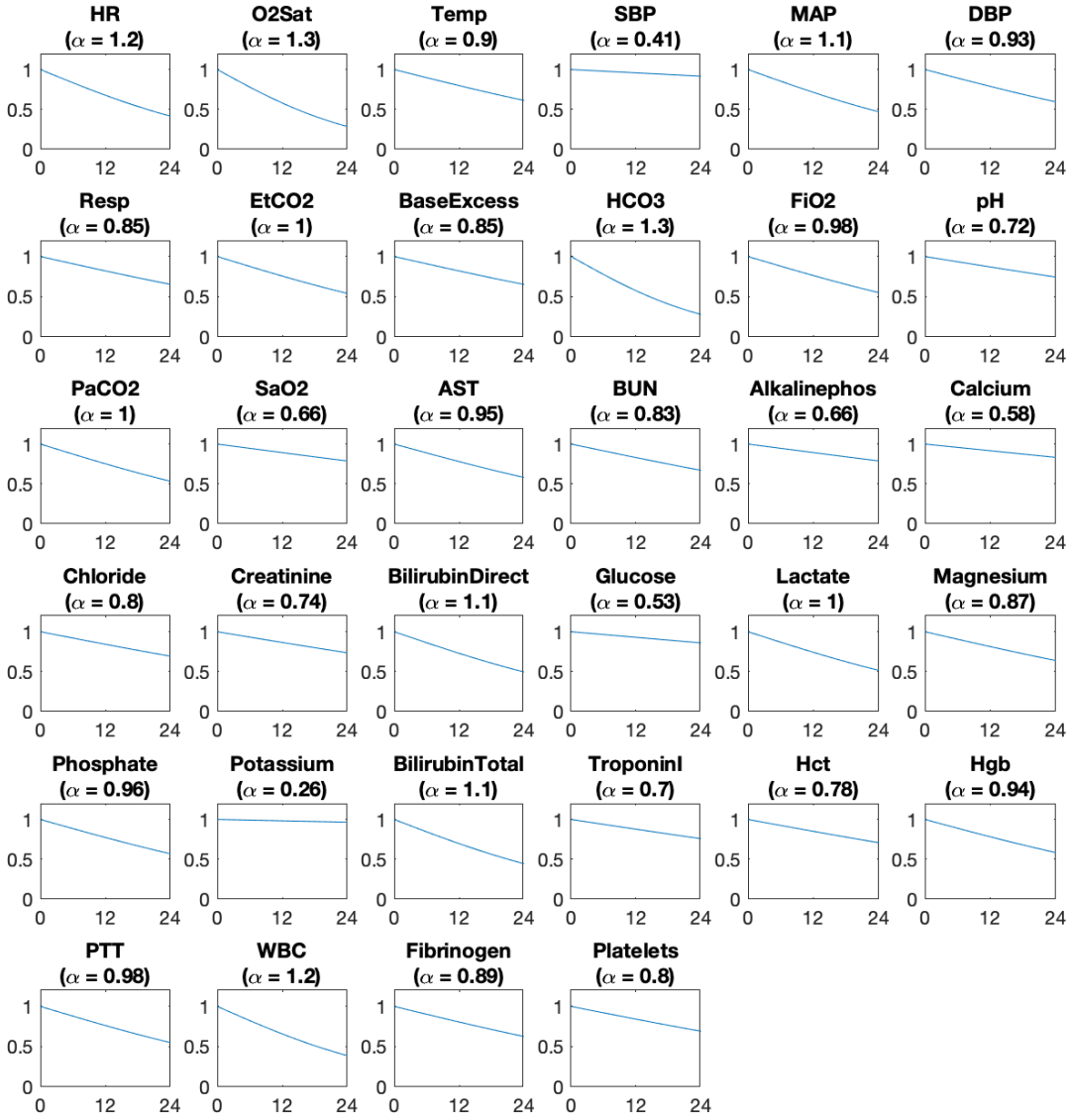


Figure 4.12: **Illustration of the weighting scheme learnt by a COMPOSER model trained on Hospital-A ICU (source dataset) and Hospital-A ED (target dataset) cohorts.** The plots shown depict the scaling function (varies from 0 to 1) imposed by the model for various values of Time Since Last Measurement (varies from 0 to 24 hours) of each of the 34 dynamical variables considered in our study.

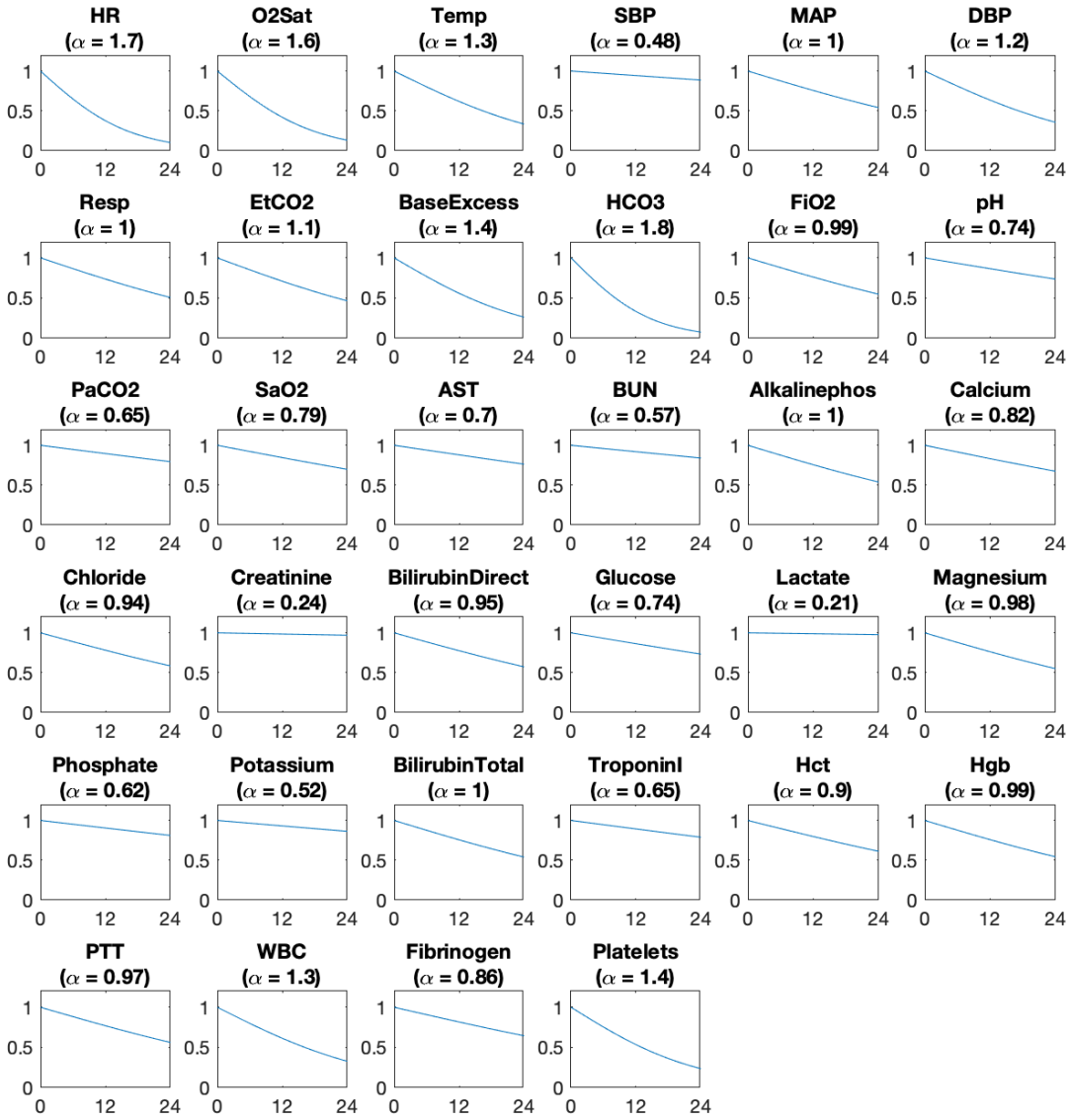


Figure 4.13: **Illustration of the weighting scheme learnt by a COMPOSER model trained on Hospital-A ICU (source dataset) and Hospital-B ED (target dataset) cohorts.** The plots shown depict the scaling function (varies from 0 to 1) imposed by the model for various values of Time Since Last Measurement (varies from 0 to 24 hours) of each of the 34 dynamical variables considered in our study.

Table 4.17: Characteristics of all five patients cohorts.

	Hospital-A ICU	**Hospital-B ICU**	**Hospital-C ICU**	**Hospital-A ED**	**Hospital-B ED**
No. Patients, *n*	18990	45679	7426	86869	325916
Septic Patients, *n (%)*	1236 (6.5%)	2563 (5.6%)	229 (3.1%)	1308 (1.5%)	6236 (1.9%)
Age (yrs), *median [IQR]*	59.8 [46.4 70.8]	62 [50 72]	63 [52 73]	55.4 [40.0 66.8]	51 [35 67]
Male, *n (%)*	11628 (61.2%)	24450 (53.5%)	4352(58.6%)	46537 (53.57%)	130251 (40%)
LOS (hrs), *median [IQR]*	45.0 [24.4 81.0]	48 [28 90]	46.9 [24.9 93.6]	8.38 [6.05 13.45]	-
CCI, *median [IQR]*	3 [1 6]	2 [1 4]	0 [0 1]	4.0 [2.0 7.0]	-
SOFA (yrs), *median [IQR]*	1.2 [0.4 2.9]	1.9 [0.6 4.0]	2.3 [0.7 4.6]	0 [0 0.88]	-
Inpatient mortality, *n (%)*	995 (5.2%)	1873 (4.1%)	851 (11.5%)	771 (0.9%)	-

Table 4.18: Characteristics of septic and non-septic population in the three ICU cohorts.

	Hospital-A ICU		**Hospital-B ICU**		**Hospital-C ICU**	
	Non-septic	Septic	Non-septic	Septic	Non-septic	Septic
No. Patients, *n (%)*	17754	1236 (6.5%)	43116	2563 (5.6%)	7197	229 (3.1%)
Age (yrs), *median [IQR]*	59.7 [46.3 7.9]	60.7 [47.0 70.2]	62 [50 72]	63 [51 72]	63 [52 73]	64 [53 73]
Male, *n (%)*	10815 (60.9%)	813 (65.8%)	23000 (53.3%)	1450 (56.5%)	4217 (58.6%)	135(58.9%)
ICU LOS (hrs), *median [IQR]*	43.2 [23.8 72.8]	141 [78.4 236.7]	46 [27 77]	141 [77 258]	46.1 [24.6 90.6]	150.1 [70.7 256.5]
CCI, *median [IQR]*	3 [1 6]	3 [2 6]	2 [1 4]	3 [2 5]	0 [0 1]	1 [0 2]
SOFA (yrs), *median [IQR]*	1.1 [0.31 2.6]	4 [2.6 5.4]	1.7 [0.5 3.6]	5.0 [3.1 7.4]	2.2 [0.6 4.5]	4.6 [2.8 7.1]
ICU Admission to $t_{sepsis-3}$ (hrs),	-	33.8 [16.2 70.4]	-	30 [13 74]	-	41.4 [18.6 95.9]
Inpatient mortality, *n (%)*	720 (4.1%)	275 (22.2%)	1509 (3.5%)	364 (15.2%)	787 (10.9%)	64 (27.9%)

Table 4.19: Characteristics of septic and non-septic population in the two ED cohorts.

	Hospital-A ED		**Hospital-B ED**	
	Non-septic	Septic	Non-septic	Septic
No. Patients, *n (%)*	85561	1308 (1.5%)	319680	6236 (1.9%)
Age (yrs), *median [IQR]*	55.3 [39.9 66.7]	61.4 [51.9 73.5]	51 [35 67]	64 [51 76]
Male, *n (%)*	45771 (53.5%)	766 (58.6%)	127059 (39.7%)	3192 (51.1%)
ED LOS (hrs), *median [IQR]*	8.4 [6.0 13.4]	10.1 [7.5 16.1]	-	-
CCI, *median [IQR]*	4.0 [2.0 7.0]	5.5 [3.0 8.0]	-	-
SOFA (yrs), *median [IQR]*	0 [0 0.85]	2.2 [1.4 3.4]	-	-
ED Admission to $t_{sepsis-3}$ (hrs),	-	4.9 [3.2 7.55]	-	3 [2 5]
Inpatient mortality, *n (%)*	666 (0.8%)	105 (8.1%)	-	-

CHAPTER 5
CONCLUSION AND FUTURE WORK

Developing predictive analytic models that can predict early onset of physiological decompensation may aid clinicians in initiating early treatment and can potentially save lives. The work presented in this thesis constitutes a step in bridging the gap between predictive models developed using retrospective data and the clinical utilization of such models, with the goal of improved patient monitoring and clinical decision support.

5.1 Summary of contributions

In **Chapter 2**, we focused on developing a technique to capture interactions between multiscale HR and BP time series - through quantification of the structure of their corresponding network representations - for early prediction of sepsis. It was shown that features derived from a multiscale HR and MAP time series network provide approximately 20% improvement over traditional indices of heart rate entropy in the AUROC for four-hour advance prediction of sepsis. This improvement was attributable to the information embedded in the higher order interaction of HR and MAP time series, as well as to the proposed novel approach to network construction, utilizing adaptive partitioning of the state-space to define a set of discrete states. The resulting algorithm was quick to implement, and readily extensible to multiscale analysis of the time series networks.

Chapter 3 discussed DeepAISE (Deep Artificial Intelligence Sepsis Expert), a recurrent neural survival model for the early prediction of sepsis. DeepAISE utilized: 1) a class of deep learning models called Recurrent neural networks (RNNs), and 2) the Weibull Cox survival model on a combination of low-resolution EHR data and high resolution vital signs time series data for early prediction of onset of sepsis. This architecture was chosen in the context of predicting sepsis onset time as a time-to-event analysis and considering that

temporal changes in patients' physiology are important for prediction of sepsis. DeepAISE was an externally evaluated sepsis model developed using over 25,000 patient admissions to the Intensive Care Units (ICUs) at two Emory University hospitals, over 18,000 ICU admissions to the UC San Diego Health system and over 40,000 ICU admissions from the Medical Information Mart for Intensive Care-III (MIMIC-III) ICU database.

Furthermore, we focused on making DeepAISE predictions interpretable by utilizing *relevance scores* (inspired from the concept of saliency maps for convolutional neural networks) to determine the top contributing factors of the output risk score at every point of time during an ICU stay. Additionally, the hidden representations learned by the recurrent neural network were used to construct a lower dimensional view of a patients' trajectory. These two attributes allow the bedside clinician to identify pathologic deviations from expected physiology early and in real-time throughout the duration of patients' hospital admission. Moreover, we showed that the top causes could be broken down into two categories of positively and negatively contributing factors to the risk score. Notably, this analysis has shed insight on the input features contributing significantly to the sensitivity (positive contributors) and specificity (negative contributors) of DeepAISE.

With the goal of developing generalizable predictive models, **Chapter 4** introduced the COMPOSER model. We first proposed a weighted input layer that was designed to handle missing data and variations in data measurement frequency across various levels of care (Emergency Departments, ICUs, Wards etc.) and across different institutions. We then utilized the technique of Adversarial Domain adaptation to learn representations that minimize healthcare system specific variations. A key importance of utilizing ADA training procedure was the design of a predictive model that could adapt to new unlabeled target patient population; therefore, gold-standard labels which are often expensive to obtain, were not required to deploy the model at a new center. We finally utilized the framework of conformal prediction for establishing the 'conditions for use' of the COMPOSER model. This enabled us explicitly determine at what level of data covariance shift we could still

trust a clinical risk score.

We showed the generalizability of COMPOSER model by utilizing data from over 480,000 patients collected between 2016-2019 from three different academic medical centers in the US, including data from Emergency Departments (EDs), Intensive Care Units (ICUs), and general wards. Additionally, the COMPOSER model predictions were validated against a cohort of 400 patients who were manually chart reviewed to determine the onset of sepsis.

5.2 Suggestions for future work

In this section, we discuss some of the potential future directions of the works presented in this thesis:

While we present a model that captured multiscale interactions in physiological time series data in Chapter 2, we hypothesize that the deep learning models discussed in Chapters 3 and 4 (DeepAISE and COMPOSER) can benefit from: 1) additional input features that capture multiscale interactions, and 2) designing hierarchical neural network architectures. In this thesis, we have specifically focused on utilizing structured EHR data (such as lab values, vital signs etc.) to build predictive models. However, EHRs additionally consists of unstructured clinical notes, which contain information about patients that could provide more detailed or complementary information in addition to structured data [146]. Incorporating clinical notes into the models proposed in this thesis could improve their prediction performance.

In Chapter 3, we have studied the validity of using *relevance scores* to identify the most relevant features (that are locally interpretable) contributing to the outcome risk score. In addition to relevance scores, other local interpretability techniques such as Local Interpretable Model-agnostic Explanations (LIME) [147] and Shapley Value Explanations [148] could be used to identify the most relevant features and consensus features identified amongst all the three techniques could be shown to the bedside clinician. A major

barrier in the progress of treatment therapies for sepsis is the overly broad definition of the syndrome, consisting of a wide array of clinical and biological features. Different combinations of these features could naturally cluster into phenotypes that may respond differently to treatments. While previous works have focused on deriving such phenotypes based on input features alone [149], we hypothesize that the representation learning and explainability-related features proposed in this thesis could further aid in discovering more meaningful phenotypes and likely provide promising directions for future sepsis research.

The technique of Adversarial Domain Adaptation involves adapting representations from a source dataset (labeled) to a target dataset (unlabeled). In our work, the target dataset typically consists of real-world data from a target population (eg. ICU population in a different hospital). In scenarios where access to large data from a target population is limited, one could train a generative model on data from the target population, and use the institution-specific generative model to augment real-world data for domain adaptation. Additionally, the sepsis predictor and the encoder modules in the COMPOSER model could be made more expressive by utilizing the technique of continual learning. For example, an encoder that was first trained on Hospital A ICU (source dataset) and Hospital B ICU (target dataset), could then be used as the initial starting point for performing domain adaptation on a new target dataset (Hospital C ICU) rather than initializing the encoder with random parameters for the new task. However, it has been observed that deep learning models trained using continual learning approaches are often susceptible to catastrophic forgetting and care must be taken to overcome catastrophic forgetting in such scenarios [150].

Our experiments have shown that false-positive cases are often attributable to patients who are very sick due to other types of decompensation [25]. Our preliminary data shows that incorporation of a mortality risk score as a feature can help reduce false alarms. Similarly, we hypothesize that utilizing other EHR-based risk scores such as those for AKI, Respiratory arrest, as input features in a sepsis prediction model can reduce false alarms.

Similarly, better modeling of co-morbid conditions and chronic illnesses and the associated treatments (such as chemotherapy) would be an effective approach to further reducing false positives. Ultimately, the three pillars of sepsis identification include pathogen detection and profiling, quantification of dysregulated immune system response to infection, and assessment of physiological decompensation/deterioration [2]. As such, incorporation of biomarkers for pathogen discovery, and host immune system response will improve the specificity of sepsis identification [151, 152, 153]. However, there is no effective method to determine the timing of ordering of biomarkers. The traditional biomarker-based scores can benefit from EHR-based continuous risk assessment models (such as COMPOSER) by providing guidance for timing of ordering of such biomarkers, which leads to improved recognition (improved specificity/reduce false positives) and effective treatment of septic patients.

REFERENCES

[1] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *The Journal of the American Medical Association*, vol. 315, no. 8, pp. 801–810, 2016.

[2] D. C. Angus and T. Van der Poll, "Severe sepsis and septic shock," *New England Journal of Medicine*, vol. 369, pp. 840–851, 2013.

[3] C. M. Torio and R. M. Andrews, "National inpatient hospital costs: The most expensive conditions by payer, 2011," *HCUP Statistical Brief*, vol. 160, p. 2013, 2013.

[4] T. G. Buchman, S. Q. Simpson, K. L. Sciarretta, K. P. Finne, N. Sowers, M. Collier, S. Chavan, I. Oke, M. E. Pennini, A. Santhosh, M. Wax, R. Woodbury, S. Chu, T. G. Merkeley, G. L. Disbrow, R. A. Bright, T. E. MaCurdy, and J. A. Kelman, "Sepsis among medicare beneficiaries: 1. the burdens of sepsis, 2012–2018*," *Critical Care Medicine*, vol. 48, no. 3, 2020.

[5] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, "Epidemiology and costs of sepsis in the united states—an analysis based on timing of diagnosis and severity level," *Critical Care Medicine*, vol. 46, no. 12, p. 1889, 2018.

[6] C. Rhee, R. Dantes, L. Epstein, D. J. Murphy, C. W. Seymour, T. J. Iwashyna, S. S. Kadri, D. C. Angus, R. L. Danner, A. E. Fiore, *et al.*, "Incidence and trends of sepsis in us hospitals using clinical vs claims data, 2009-2014," *The Journal of the American Medical Association*, vol. 318, no. 13, pp. 1241–1249, 2017.

[7] T. J. Iwashyna, E. W. Ely, D. M. Smith, and K. M. Langa, "Long-term cognitive impairment and functional disability among survivors of severe sepsis," *The Journal of the American Medical Association*, vol. 304, no. 16, pp. 1787–1794, 2010.

[8] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke, *et al.*, "Surviving sepsis campaign: International guidelines for management of severe sepsis and septic shock, 2012," *Intensive Care Medicine*, vol. 39, no. 2, pp. 165–228, 2013.

[9] A. K. Venkatesh, U. Avula, H. Bartimus, J. Reif, M. J. Schmidt, and E. S. Powell, "Time to antibiotics for septic shock: Evaluating a proposed performance measure," *The American Journal of Emergency Medicine*, vol. 31, no. 4, pp. 680–683, 2013.

[10] S. A. Sterling, W. R. Miller, J. Pryor, M. A. Puskarich, and A. E. Jones, "The impact of timing of antibiotics on outcomes in severe sepsis and septic shock: A systematic review and meta-analysis," *Critical Care Medicine*, vol. 43, no. 9, p. 1907, 2015.

[11] A. Rhodes, G. Phillips, R. Beale, M. Cecconi, J. D. Chiche, D. De Backer, J. Divatia, B. Du, L. Evans, R. Ferrer, *et al.*, "The surviving sepsis campaign bundles and outcome: Results from the international multicentre prevalence study on sepsis (the impress study)," *Intensive Care Medicine*, vol. 41, no. 9, pp. 1620–1628, 2015.

[12] R. Ferrer, I. Martin-Loeches, G. Phillips, T. M. Osborn, S. Townsend, R. P. Dellinger, A. Artigas, C. Schorr, and M. M. Levy, "Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: Results from a guideline-based performance improvement program," *Critical Care Medicine*, vol. 42, no. 8, pp. 1749–1755, 2014.

[13] M. M. Levy, L. E. Evans, and A. Rhodes, "The surviving sepsis campaign bundle: 2018 update," *Intensive Care Medicine*, vol. 44, no. 6, pp. 925–928, 2018.

[14] C. W. Seymour, F. Gesten, H. C. Prescott, M. E. Friedrich, T. J. Iwashyna, G. S. Phillips, S. Lemeshow, T. Osborn, K. M. Terry, and M. M. Levy, "Time to treatment and mortality during mandated emergency care for sepsis," *New England Journal of Medicine*, 2017.

[15] A. F. Shorr, S. T. Micek, W. L. Jackson Jr, and M. H. Kollef, "Economic implications of an evidence-based sepsis protocol: Can we improve outcomes and lower costs?" *Critical Care Medicine*, vol. 35, no. 5, pp. 1257–1262, 2007.

[16] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer, *et al.*, "Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3)," *The Journal of the American Medical Association*, vol. 315, no. 8, pp. 762–774, 2016.

[17] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[18] Centers for Disease Control & Prevention, "Hospital toolkit for adult sepsis surveillance," *Atlanta, US Department of Health and Human Services*, 2018.

[19] Centers for Medicare & Medicaid Services, *Qualitynet — inpatient hospitals specifications manual.* Quality website. https://www.qualitynet.org/inpatient/specifications-manuals. Accessed November 11, 2019.

[20] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *The Journal of the American Medical Association*, vol. 319, no. 13, pp. 1317–1318, 2018.

[21] L. M. Fleuren, T. L. T. Klausch, C. L. Zwager, L. J. Schoonmade, T. Guo, L. F. Roggeveen, E. L. Swart, A. R. J. Girbes, P. Thoral, A. Ercole, M. Hoogendoorn, and P. W. G. Elbers, "Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy," *Intensive Care Medicine*, 2020.

[22] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019," *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, 2020.

[23] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, *et al.*, "Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach," *Journal of Medical Internet Research Medical Informatics*, vol. 4, no. 3, 2016.

[24] Q. Mao, M. Jay, J. L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem, *et al.*, "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu," *BMJ open*, vol. 8, no. 1, e017833, 2018.

[25] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the icu.," *Critical Care Medicine*, vol. 46, no. 4, pp. 547–553, 2018.

[26] J. Futoma, S. Hariharan, M. Sendak, N. Brajer, M. Clement, A. Bedoya, C. O'Brien, and K. Heller, "An improved multi-output gaussian process rnn with real-time validation for early sepsis detection," *arXiv preprint arXiv:1708.05894*, 2017.

[27] R. A. Lukaszewski, A. M. Yates, M. C. Jackson, K. Swingler, J. M. Scherer, A. Simpson, P. Sadler, P. McQuillan, R. W. Titball, T. J. Brooks, *et al.*, "Presymptomatic prediction of sepsis in intensive care unit patients," *Clinical and Vaccine Immunology*, vol. 15, no. 7, pp. 1089–1094, 2008.

[28] H. M. Giannini, J. C. Ginestra, C. Chivers, M. Draugelis, A. Hanish, W. D. Schweickert, B. D. Fuchs, L. Meadows, M. Lynch, P. J. Donnelly, *et al.*, "A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice," *Critical Care Medicine*, vol. 47, no. 11, pp. 1485–1492, 2019.

[29] A. Khojandi, V. Tansakul, X. Li, R. S. Koszalinski, and W. Paiva, "Prediction of sepsis and in-hospital mortality using electronic health records," *Methods of Information in Medicine*, vol. 57, no. 04, pp. 185–193, 2018.

[30] F. van Wyk, A. Khojandi, A. Mohammed, E. Begoli, R. L. Davis, and R. Kamaleswaran, "A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier," *International Journal of Medical Informatics*, vol. 122, pp. 55–62, 2019.

[31] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, 299ra122–299ra122, 2015.

[32] T. G. Buchman, "Nonlinear dynamics, complex systems, and the pathobiology of critical illness," *Current Opinion in Critical Care*, vol. 10, no. 5, pp. 378–382, 2004.

[33] J. R. Moorman, J. B. Delos, A. A. Flower, H. Cao, B. P. Kovatchev, J. S. Richman, and D. E. Lake, "Cardiovascular oscillations at the bedside: Early diagnosis of neonatal sepsis using heart rate characteristics monitoring," *Physiological Measurement*, vol. 32, no. 11, p. 1821, 2011.

[34] J. M. Huston and K. J. Tracey, "The pulse of inflammation: Heart rate variability, the cholinergic anti-inflammatory pathway and implications for therapy," *Journal of Internal Medicine*, vol. 269, no. 1, pp. 45–53, 2011.

[35] A. S. Campanharo, M. I. Sirer, R. D. Malmgren, F. M. Ramos, and L. A. N. Amaral, "Duality between time series and networks," *PLOS ONE*, vol. 6, no. 8, e23378, 2011.

[36] G. Nicolis, A. G. Cantu, and C. Nicolis, "Dynamical aspects of interaction networks," *International Journal of Bifurcation and Chaos*, vol. 15, no. 11, pp. 3467–3480, 2005.

[37] L. Lacasa, V. Nicosia, and V. Latora, "Network structure of multivariate time series," *Scientific Reports*, vol. 5, 2015.

[38] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, "Horizontal visibility graphs: Exact results for random time series," *Physical Review E*, vol. 80, no. 4, p. 046 103, 2009.

[39] D. Steinhauser, L. Krall, C. Müssig, D. Büssis, and B. Usadel, "Correlation networks," *Analysis of Biological Networks*, pp. 305–333, 2008.

[40] Y. Yang and H. Yang, "Complex network-based time series analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 5, pp. 1381–1386, 2008.

[41] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[42] A. Connell, H. Montgomery, P. Martin, C. Nightingale, O. Sadeghi-Alavijeh, D. King, A. Karthikesalingam, C. Hughes, T. Back, K. Ayoub, *et al.*, "Evaluation of a digitally-enabled care pathway for acute kidney injury management in hospital emergency admissions," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[43] N. Tangri, L. A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, and A. S. Levey, "A predictive model for progression of chronic kidney disease to kidney failure," *The Journal of the American Medical Association*, vol. 305, no. 15, pp. 1553–1559, 2011.

[44] S. Levin, M. Toerper, E. Hamrock, J. S. Hinson, S. Barnes, H. Gardner, A. Dugas, B. Linton, T. Kirsch, and G. Kelen, "Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index," *Annals of Emergency Medicine*, vol. 71, no. 5, pp. 565–574, 2018.

[45] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, "Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards," *Critical Care Medicine*, vol. 44, no. 2, p. 368, 2016.

[46] S. Hao, Y. Wang, B. Jin, A. Y. Shin, C. Zhu, M. Huang, L. Zheng, J. Luo, Z. Hu, C. Fu, *et al.*, "Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the maine healthcare information exchange," *PloS One*, vol. 10, no. 10, 2015.

[47] I. Cho, E.-H. Boo, E. Chung, D. W. Bates, and P. Dykes, "Novel approach to inpatient fall risk prediction and its cross-site validation using time-variant data," *Journal of Medical Internet Research*, vol. 21, no. 2, e11505, 2019.

[48] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019.

[49] C. S. Josef, V. Ramnath, A. Malhotra, and S. Nemati, "Performance comparison of unit specific and generalizable sepsis prediction models across intensive care units," in *SHOCK*, vol. 51, 2019, pp. 85–86.

[50] J. Oh, M. Makar, C. Fusco, R. McCaffrey, K. Rao, E. E. Ryan, L. Washer, L. R. West, V. B. Young, J. Guttag, *et al.*, "A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health

centers," *Infection Control & Hospital Epidemiology*, vol. 39, no. 4, pp. 425–433, 2018.

[51] T. Bennett, S. Russell, J. King, L. Schilling, C. Voong, N. Rogers, B. Adrian, N. Bruce, and D. Ghosh, "Accuracy of the epic sepsis prediction model in a regional health system," *arXiv preprint arXiv:1902.07276*, 2019.

[52] D. Agniel, I. S. Kohane, and G. M. Weber, "Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study," *BMJ*, vol. 361, k1479, 2018.

[53] J. Sun and M. W. Deem, "Spontaneous emergence of modularity in a model of evolving individuals," *Physical Review Letters*, vol. 99, no. 22, p. 228 107, 2007.

[54] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[55] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical Review E*, vol. 72, no. 2, p. 027 104, 2005.

[56] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[57] A. L. Goldberger, "Heartbeats, hormones, and health: Is variability the spice of life?" *American Journal of Respiratory and Critical Care Medicine*, vol. 163, no. 6, pp. 1289–1290, 2001.

[58] P. C. Ivanov, L. A. N. Amaral, A. L. Goldberger, S. Havlin, M. G. Rosenblum, Z. R. Struzik, and H. E. Stanley, "Multifractality in human heartbeat dynamics," *Nature*, vol. 399, no. 6735, pp. 461–465, 1999.

[59] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, p. 068 102, 2002.

[60] G. Mancia, "Short-and long-term blood pressure variability," *Hypertension*, vol. 60, no. 2, pp. 512–517, 2012.

[61] G. Parati, J. E. Ochoa, C. Lombardi, and G. Bilo, "Blood pressure variability: Assessment, predictive value, and potential as a therapeutic target," *Current Hypertension Reports*, vol. 17, no. 4, p. 23, 2015.

[62] R. P. Bartsch, K. K. Liu, Q. D. Ma, and P. C. Ivanov, "Three independent forms of cardio-respiratory coupling: Transitions across sleep stages," in *Computing in Cardiology Conference (CinC), 2014*, IEEE, 2014, pp. 781–784.

[63] L.-w. H. Lehman, R. P. Adams, L. Mayaud, G. B. Moody, A. Malhotra, R. G. Mark, and S. Nemati, "A physiological time series dynamics-based approach to patient monitoring and outcome prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1068–1076, 2015.

[64] T. G. Buchman, "Physiologic stability and physiologic state," *Journal of Trauma and Acute Care Surgery*, vol. 41, no. 4, pp. 599–605, 1996.

[65] J. A. Quinn, C. K. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1537–1551, 2009.

[66] F. Takens *et al.*, "Detecting strange attractors in turbulence," *Lecture notes in Mathematics*, vol. 898, no. 1, pp. 366–381, 1981.

[67] J. E. Hudson, "Signal processing using mutual information," *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 50–54, 2006.

[68] J. Lee, S. Nemati, I. Silva, B. A. Edwards, J. P. Butler, and A. Malhotra, "Transfer entropy estimation and directional coupling change detection in biomedical time series," *Biomedical Engineering Online*, vol. 11, no. 1, p. 19, 2012.

[69] S. Nemati, B. A. Edwards, J. Lee, B. Pittman-Polletta, J. P. Butler, and A. Malhotra, "Respiration and heart rate complexity: Effects of age and gender assessed by band-limited transfer entropy," *Respiratory Physiology & Neurobiology*, vol. 189, no. 1, pp. 27–33, 2013.

[70] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *The Annals of Statistics*, pp. 1236–1265, 1992.

[71] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208 701, 2002.

[72] M. Ghassemi, L.-w. Lehman, J. Snoek, and S. Nemati, "Global optimization approaches for parameter tuning in biomedical signal processing: A focus on multi-scale entropy," in *Computing in Cardiology Conference (CinC), 2014*, IEEE, 2014, pp. 993–996.

[73] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski, "A comparison of auc estimators in small-sample studies," in *Machine Learning in Systems Biology*, 2009, pp. 3–13.

[74] P. C. Ivanov and R. P. Bartsch, "Network physiology: Mapping interactions between networks of physiologic networks," in *Networks of Networks: the last Frontier of Complexity*, Springer, 2014, pp. 203–222.

[75] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, "Network physiology reveals relations between network topology and physiological function," *Nature Communications*, vol. 3, p. 702, 2012.

[76] R. P. Bartsch, K. K. Liu, A. Bashan, and P. C. Ivanov, "Network physiology: How organ systems dynamically interact," *PLOS ONE*, vol. 10, no. 11, e0142143, 2015.

[77] K. K. Liu, R. P. Bartsch, A. Lin, R. N. Mantegna, and P. C. Ivanov, "Plasticity of brain wave network interactions and evolution across physiologic states," *Frontiers in Neural Circuits*, vol. 9, 2015.

[78] A. Lin, K. K. Liu, R. P. Bartsch, and P. C. Ivanov, "Delay-correlation landscape reveals characteristic time delays of brain rhythms and heart interactions," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2067, p. 20 150 182, 2016.

[79] L. Faes, G. Nollo, F. Jurysta, and D. Marinazzo, "Information dynamics of brain–heart physiological networks during sleep," *New Journal of Physics*, vol. 16, no. 10, p. 105 005, 2014.

[80] S.-L. Wang, F. Wu, and B.-H. Wang, "Prediction of severe sepsis using svm model," in *Advances in Computational Biology*, Springer, 2010, pp. 75–81.

[81] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, and L. A. Nathanson, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PloS one*, vol. 12, no. 4, e0174708, 2017.

[82] S. P. Shashikumar, M. D. Stanley, I. Sadiq, Q. Li, A. Holder, G. D. Clifford, and S. Nemati, "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," *Journal of Electrocardiology*, vol. 50, no. 6, pp. 739–743, 2017.

[83] R. Postelnicu, S. M. Pastores, D. H. Chong, and L. Evans, "Sepsis early warning scoring systems: The ideal tool remains elusive!" *Journal of Critical Care*, vol. 52, pp. 251 –253, 2019.

[84] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," *The Journal of the American Medical Association*, vol. 320, no. 21, pp. 2199–2200, 2018.

[85] J. Norrie, "The challenge of implementing ai models in the icu," *The Lancet Respiratory Medicine*, vol. 6, no. 12, pp. 886–888, 2018.

[86] C. L. Ventola, "The antibiotic resistance crisis: Part 1: Causes and threats," *Pharmacy and Therapeutics*, vol. 40, no. 4, p. 277, 2015.

[87] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160 035, 2016.

[88] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[89] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[90] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[92] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012, pp. 2951–2959.

[93] T. E. Oliphant, *A guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.

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

[95] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint*, 2013.

[96] V. Barak, A. Schwartz, I. Kalickman, B. Nisman, G. Gurman, and Y. Shoenfeld, "Prevalence of hypophosphatemia in sepsis and infection: The role of cytokines," *The American Journal of Medicine*, vol. 104, no. 1, pp. 40–47, 1998.

[97] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 568–586, 2010.

[98] R. C. Amland and B. B. Sutariya, "Quick sequential [sepsis-related] organ failure assessment (qsofa) and st. john sepsis surveillance agent to detect patients at risk of sepsis: An observational cohort study," *American Journal of Medical Quality*, vol. 33, no. 1, pp. 50–57, 2018.

[99] I. Cortés-Puch and C. S. Hartog, *Opening the debate on the new sepsis definition change is not necessarily progress: Revision of the sepsis definition should be based on new scientific insights*, 2016.

[100] A. H. Carneiro, P. Póvoa, and J. A. Gomes, "Dear sepsis-3, we are sorry to say that we don't like you," *Revista Brasileira de terapia intensiva*, vol. 29, no. 1, pp. 4–8, 2017.

[101] M. G. Kashiouris, Z. Zemore, Z. Kimball, C. Stefanou, B. Fisher, M. de Wit, S. Pedram, C. N. Sessler, *et al.*, "Supply chain delays in antimicrobial administration after the initial clinician order and mortality in patients with sepsis," *Critical Care Medicine*, vol. 47, no. 10, pp. 1388–1395, 2019.

[102] S. K. Gadre, M. Shah, E. Mireles-Cabodevila, B. Patel, and A. Duggal, "Epidemiology and predictors of 30-day readmission in patients with sepsis," *Chest*, vol. 155, no. 3, pp. 483–490, 2019.

[103] M. Shankar-Hari and G. D. Rubenfeld, "Understanding long-term outcomes following sepsis: Implications and challenges," *Current Infectious Disease Reports*, vol. 18, no. 11, p. 37, 2016.

[104] A. E. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, 2016.

[105] D. Shillan, J. A. Sterne, A. Champneys, and B. Gibbison, "Use of machine learning to analyse routinely collected intensive care unit data: A systematic review," *Critical Care*, vol. 23, no. 1, p. 284, 2019.

[106] E. J. Hwang, S. Park, K.-N. Jin, J. Im Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, J. G. Cohen, *et al.*, "Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs," *JAMA network open*, vol. 2, no. 3, e191095–e191095, 2019.

[107] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Medicine*, vol. 15, no. 11, 2018.

[108] A. Sharafoddini, J. A. Dubin, D. M. Maslove, and J. Lee, "A new insight into missing data in intensive care unit patient profiles: Observational study," *JMIR Medical Informatics*, vol. 7, no. 1, e11605, 2019.

[109] United States Food & Drug Administration, "Proposed regulatory framework for modifications to artificial intelligence," *Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). Discussion Paper and Request for Feedback. Available at: https://www. fda. gov/media/122535/download. Accessed June*, vol. 12, 2019.

[110] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[111] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[112] C. Saunders, A. Gammerman, and V. Vovk, "Transduction with confidence and credibility," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'99, Stockholm, Sweden: Morgan Kaufmann Publishers Inc., 1999, 722–726.

[113] V. Vovk, A. Gammerman, and C. Saunders, "Machine-learning applications of algorithmic randomness," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, 444–453, ISBN: 1558606122.

[114] H. Papadopoulos, V. Vovk, and A. Gammermam, "Conformal prediction with neural networks," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, IEEE, vol. 2, 2007, pp. 388–395.

[115] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. Mar, pp. 371–421, 2008.

[116] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," in *Advances in Neural Information Processing Systems*, 2018, pp. 5541–5552.

[117] S. P. Shashikumar, C. Josef, A. Sharma, and S. Nemati, "Deepaise–an end-to-end development and deployment of a recurrent neural survival model for early prediction of sepsis," *arXiv preprint arXiv:1908.04759*, 2019.

[118] R. Pivovarov, D. J. Albers, J. L. Sepulveda, and N. Elhadad, "Identifying and mitigating biases in ehr laboratory tests," *Journal of Biomedical Informatics*, vol. 51, pp. 24–34, 2014.

[119] I. Yelin, O. Snitser, G. Novich, R. Katz, O. Tal, M. Parizade, G. Chodick, G. Koren, V. Shalev, and R. Kishony, "Personal clinical history predicts antibiotic resistance of urinary tract infections," *Nature Medicine*, vol. 25, no. 7, pp. 1143–1152, 2019.

[120] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.

[121] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, no. 11, pp. 1716–1720, 2018.

[122] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, p. 65, 2019.

[123] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[124] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *The Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, 2016.

[125] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems*, 2007, pp. 137–144.

[126] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[127] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N Lawrence, *Covariate shift and local learning by distribution matching*, 2008.

[128] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[129] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.

[130] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[131] H. Huang, Q. Huang, and P. Krahenbuhl, "Domain transfer through deep activation matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 590–605.

[132] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3752–3761.

[133] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481.

[134] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 285–10 295.

[135] M. E. Hellman, "The nearest neighbor classification rule with a reject option," *IEEE Transactions on Systems Science and Cybernetics*, vol. 6, no. 3, pp. 179–185, 1970.

[136] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," *arXiv preprint arXiv:2002.11297*, 2020.

[137] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[138] A. Shafaei, M. Schmidt, and J. J. Little, "A less biased evaluation of out-of-distribution sample detectors," *arXiv preprint arXiv:1809.04729*, 2018.

[139] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[140] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[141] C. Rhee, S. R. Brown, T. M. Jones, C. O'Brien, A. Pande, Y. Hamad, A. L. Bulger, K. A. Tobin, A. F. Massaro, D. J. Anderson, *et al.*, "Variability in determining sepsis time zero and bundle compliance rates for the centers for medicare and medicaid services sep-1 measure," *Infection Control & Hospital Epidemiology*, vol. 39, no. 8, pp. 994–996, 2018.

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

[143] Z. C. Lipton, D. Kale, and R. Wetzel, "Directly modeling missing data in sequences with rnns: Improved classification of clinical time series," in *Machine Learning for Healthcare Conference*, 2016, pp. 253–270.

[144] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018.

[145] C. Rhee, Z. Zhang, S. S. Kadri, D. J. Murphy, G. S. Martin, E. Overton, C. W. Seymour, D. C. Angus, R. Dantes, L. Epstein, *et al.*, "Sepsis surveillance using adult sepsis events simplified esofa criteria versus sepsis-3 sequential organ failure assessment criteria," *Critical Care Medicine*, vol. 47, no. 3, pp. 307–314, 2019.

[146] W. Boag, D. Doss, T. Naumann, and P. Szolovits, "What's in a note? unpacking predictive value in clinical note representations," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 26, 2018.

[147] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[148] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[149] C. W. Seymour, J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez, *et al.*, "Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis," *The Journal of the American Medical Association*, vol. 321, no. 20, pp. 2003–2017, 2019.

[150] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[151] M. Reyes, M. R. Filbin, R. P. Bhattacharyya, K. Billman, T. Eisenhaure, D. T. Hung, B. D. Levy, R. M. Baron, P. C. Blainey, M. B. Goldberg, and N. Hacohen, "An immune-cell signature of bacterial sepsis," *Nature Medicine*, Feb. 2020.

[152] T. E. Sweeney and P. Khatri, "Benchmarking sepsis gene expression diagnostics using public data," *Critical Care Medicine*, vol. 45, no. 1, p. 1, 2017.

[153] T. E. Sweeney, A. Shidham, H. R. Wong, and P. Khatri, "A comprehensive time-course–based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set," *Science Translational Medicine*, vol. 7, no. 287, 287ra71–287ra71, 2015.

LIST OF PUBLICATIONS

1. **Early Prediction Of Sepsis From Clinical Data: The PhysioNet/Computing In Cardiology Challenge 2019:** Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. *Critical Care Medicine, 2020*

2. **DeepAISE–An End-to-End Development and Deployment of a Recurrent Neural Survival Model for Early Prediction of Sepsis:** Supreeth P. Shashikumar, Christopher Josef, Ashish Sharma, and Shamim Nemati. *arXiv preprint, 2019*

3. **Multiscale Network Dynamics Between Heart Rate and Locomotor Activity Are Altered In Schizophrenia:** Erik Reinertsen, Supreeth P. Shashikumar, Amit J. Shah, Shamim Nemati, and Gari D. Clifford. *Physiological Measurement, 2018*

4. **DeepAISE on FHIR—An Interoperable Real-Time Predictive Analytic Platform for Early Prediction of Sepsis:** Vidyashankar Lakshman, Fatemeh Amrollahi, Veera Supraja Koppisetty, Supreeth P. Shashikumar, Ashish Sharma, and Shamim Nemati. *AMIA Annual Symposium proceedings, 2018*

5. **A FHIR-enabled Streaming Sepsis Prediction System for ICUs:** Joel R. Henry, Dennis Lynch, Jeff Mals, Supreeth P. Shashikumar, Andre Holder, Ashish Sharma, and Shamim Nemati. *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018*

6. **Detection of Paroxysmal Atrial Fibrillation Using Attention-based Bidirectional Recurrent Neural Networks:** Supreeth P. Shashikumar, Amit J. Shah, Gari D. Clifford, and Shamim Nemati. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018*

7. **Multiscale Network Representation of Physiological Time Series for Early Prediction of Sepsis:** Supreeth P. Shashikumar, Qiao Li, Gari D. Clifford, and Shamim Nemati. *Physiological Measurement, 2017*

8. **Early Sepsis Detection in Critical Care Patients using Multiscale Blood Pressure and Heart Rate Dynamics:** Supreeth P. Shashikumar, Matthew D. Stanley, Ismail Sadiq, Qiao Li, Andre Holder, Gari D. Clifford, and Shamim Nemati. *Journal of Electrocardiology, 2017*

9. **A Deep Learning Approach to Monitoring and Detecting Atrial Fibrillation using Wearable Technology:** Supreeth P. Shashikumar, Amit J. Shah, Qiao Li, Gari D. Clifford, and Shamim Nemati. *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2017*