Novel clinical outcome models using heterogeneous Electronic Healthcare Record (EHR) data

by

Shaodong Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:
Qing Li, Major Professor
Wenli Zhang
Hantang Qin
Lizhi Wang
Cameron MacKenzie

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me throughout the course of this research. First and foremost, I would like to express my deepest appreciation to my advisor Dr. Qing Li, who is always professional, patient, and supportive. I would also like to thank Dr. Wenli Zhang for her detailed guidance on the research and career. I want to thank the rest of my dissertation committee: Dr. Qin, Dr. Wang, and Dr. MacKenzie, for being on my committee and providing insightful suggestions.

I am also grateful to my collaborators for their technical help and moral support. I would be remiss in not mentioning my family, especially my parents and spouse. I could not have undertaken this journey without their encouragement and emotional support.

ABSTRACT

Clinical outcome models estimate the probability of developing a future adverse outcome for patients. In practice, the clinical outcome models help assess the severity of illness, evaluate the value of new treatments, provide expected outcomes, and promote clinical resource allocation. Meanwhile, the rich information in the EHR data provides great opportunities to build more accurate and reliable models for various clinical outcome tasks. However, there are still many challenges when developing clinical outcome models on EHRs, including the hierarchical structure of high-dimensional medical concepts, the pattern extraction of vital signs, and the data insufficiency of some lab tests. To address these challenges, we present three research designs in this dissertation, including a new framework of low-dimensional representations for medical concepts, a new feature extraction scheme for vital sign data, and a new transfer learning for lab outcome prediction with limited samples. By tackling the abovementioned issues in EHR data, our work has great potential to enlarge the social impact of clinical outcome models.

CHAPTER 1. GENERAL INTRODUCTION

Clinical outcome models estimate the probability of developing a future adverse outcome for patients (Whittemore, 2010). The outcome estimation is usually made by summarizing the available patient information, such as the demographics and the vital signs. Existing research has predicted various clinical outcomes, such as readmission, mortality, and the disease onsite (Fang et al., 2021; Whittemore, 2010). Such models are becoming more and more popular and important in clinical practice, as the volumes of clinical data keep growing and inference models become increasingly accurate (Mahmoudi et al., 2020).

The clinical outcome models have various applications in clinical practice. First, healthcare practitioners can assess the severity of illness based on the estimated clinical outcomes (Pirracchio, 2016). If the estimated risk of adverse outcomes keeps going up, then the patient illness probability is deteriorating. Second, the change of the clinical outcomes describes the value of new treatments, interventions and healthcare policies (Xu et al., 2017). If the risk of adverse outcomes drops after the patient receive a new treatment, this could be evidence that the new treatment is taking effect. Third, the clinical outcome models potentially inform rationing decisions on patients' medical needs, and help doctors communicate the expected outcomes to patients (Truog et al., 2008). Fourth, the outcome model can help promote effective resource allocation and alleviate healthcare burdens (Knaus et al., 1991). It is reasonable to allocate more clinical resources to patients with higher risk. The need for clinical outcome predictions has been magnified during public health threats like the COVID-19 pandemic because hospitals have been overwhelmed by the influx of patients and the clinical resource need to be optimized.

In the past, researchers constructed outcome models using data from epidemiologic cohorts or clinical trials (Goldstein et al., 2016). However, the external generalizability of

constructed models was questioned due to the narrow and unrepresentative data source. First, the cohorts in the clinical trials are usually strictly defined, which do not represent patients in practice well. Second, the cohort studies and real-life clinical settings may have different data collection processes.

Then electronic health records (EHRs) provide opportunities to build more accurate and more generalizable clinical outcome models. An EHR is a digital version of a patient's medical history (Goldstein et al., 2016). Specifically, EHRs contain patients' demographics, diagnoses, treatments, vital signs, radiology images, laboratory test results, etc. EHRs include a large number of clinical cases from daily hospital admissions that represents the real world (Goldstein et al., 2016). Many clinical outcome models have been proposed based on the EHR data. For example, some researchers use historical visits in EHRs to build readmission prediction models (Allam et al., 2019; Ashfaq et al., 2019; Min et al., 2019). Many scientists prognosticate the patients' mortality risk at intensive care units based on the available features in EHRs, such as vital signs and demographics (Davoodi & Moradi, 2018; Hsieh et al., 2018; Kong et al., 2020; Zhai et al., 2020).

However, there are still a lot of challenges when developing clinical outcome models on EHRs, including the hierarchical structure of high-dimensional medical concepts, the pattern extraction of vital signs, and the data insufficiency of some lab tests. 1) Specifically, EHR data usually contain high-dimensional medical concepts, such as 17,000 World Health Organization's (WHO) International Classification of Diseases (ICD). These medical concepts are usually of hierarchical structure according to their clinical relationships. The high-dimensionality and the hierarchical structure make the medical concepts difficult to be represented in the clinical outcome models. 2) Additionally, patients' vital sign data have opened new possibilities to

propose more reliable clinical outcome models. The vital signs are real-time time series of body measurements, such as heart rates and respirations. Researchers have recently demonstrated that the vital signs contain rich dynamic patterns that can be helpful in informing prognosis, provide early forecasts of life-threatening conditions, and predict clinical outcomes (Hong et al., 2013; Lehman et al., 2015). However, it is challenging to extract the helpful patterns effectively from the time series of vital signs. 3) The laboratory test is a key resource in ICU to inspect the patients' health condition. However, the laboratory resources are not always sufficient, and some laboratory tests are unnecessary without providing helpful information. Therefore, researchers propose to predict the risk of abnormal outcome for lab tests, which quantifies the expected information and helps allocate the lab resources. During the development of the lab anomaly prediction models, a new challenge comes out. Some lab tests are not popular in hospitals, which strictly limits the available samples for the model training. The insufficient training samples could strongly harm the model's accuracy and reliability. It is challenging to construct lab anomaly models for the lab types that do not have a large quantity of training data.

These limitations or characteristics of the current EHRs hinder the outcome models from broader implementations. To address these challenges, we conducted three studies in Chapters 2-4.

In Chapter 2, we propose a new framework to generate low-dimensional representations with Manifold Learning for sets of hierarchical medical concepts in EHR data. The framework is essential for healthcare-related classifications because it solves the high-dimensional problem of the complicated medical concepts in the EHR. To demonstrate the efficacy of our proposed framework, we generate low-dimensional representations for hospital visits of heart failure patients, which are further used for a 30-day readmission prediction. The proposed framework

can not only boost the performance of readmission prediction as shown in this work but can also be easily generalized to various healthcare-related prediction tasks, such as mortality prediction, length-of-stay prediction, etc.

In Chapter 3, we propose a new ICU mortality prediction model capable of effectively extracting valid and interpretable patterns from the readily-available vital sign data with improved accuracy, by combining stochastic signal analysis and machine learning techniques. To illustrate the efficacy of our model, we evaluate it on a large real-world multi-center ICU dataset. The proposed model outperforms baseline methods, including APACHE IV (the "golden standard" in ICU outcome predictions), deep learning-based models (i.e., LSTM, GRU, CNN), statistical feature classification, and time series forecasting methods (i.e., ARMA, ARIMA) by a large margin. The innovative artifacts obtained from this study are salient to both the data science and healthcare communities.

In Chapter 4, we propose a new transfer learning method for lab anomaly prediction with limited training data. Specifically, we develop a novel distance to select the optimal source domain from multiple high-frequency lab tests. We design a recurrent neural network to estimate the probability of obtaining an abnormal lab outcome. We transfer knowledge from the selected source domain to improve the model performance on the target domains (low-frequency lab tests). We evaluate the proposed method on five low-frequency lab types that are related to heart failure and five high-frequency lab types that are most common in the hospital. The experiments show that the designed neural network outperforms all traditional machine learning models by a large margin. In the experiments, the transfer learning and the proposed domain distance further improve the model performance for all selected low-frequency lab types (e.g., AUC scores increase from 0.729 to 0.795 for Brain natriuretic peptide tests). The new transfer learning

method address the data insufficiency problem for lab outcome prediction, which provides a more reliable way to optimize clinical resource allocation.

The contribution of this dissertation is to fill the research gap and address the three challenges in EHR data, including the hierarchical structure of high-dimensional medical concepts, the pattern extraction of vital signs, and the data insufficiency of some lab tests. Specifically, we invent a new method to generate low-dimensional representations for clinical records with hierarchically structured and high-dimensional medical concepts. We create a novel framework that effectively extracts powerful and meaningful patterns from the time series of vital signs. We design a lab anomaly model that does not require many training samples. Overall, we facilitate the clinical outcome modeling on EHR data by addressing these challenges. The proposed works increase the predictive performance and reliability of clinical outcome models. Besides, this dissertation also provides great opportunities for other researchers to build more powerful clinical outcome models. For example, the generated representations in Chapter 2 and the extracted vital sign features in Chapter 3 can be used as inputs for other clinical models. And the methodology in Chapter 4 can also be used on other clinical tasks that do not have sufficient training samples.

This dissertation also has great potential to enlarge the social impact of clinical outcome models. We facilitate the model adoption and implementation in clinical practice by improving the model's accuracy and reliability. For patients, our work helps to decrease healthcare costs and adverse impacts. For healthcare providers, we provide a more accurate and reliable way to assess the severity of illness, the value of new treatments, and expected outcomes, which enables more scientific clinical decisions. For society, we enable the clinical outcome models to alleviate more healthcare expenditure burden and optimize better resource allocation.

References

Allam, A., Nagy, M., Thoma, G., & Krauthammer, M. (2019). Neural networks versus Logistic regression for 30 days all-cause readmission prediction. *Scientific Reports*, *9*(1), 9277. https://doi.org/10.1038/s41598-019-45685-z

Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics*, *97*, 103256. https://doi.org/10.1016/j.jbi.2019.103256

Davoodi, R., & Moradi, M. H. (2018). Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. *Journal of Biomedical Informatics*, *79*, 48–59.

Fang, H. S. A., Tan, N. C., Tan, W. Y., Oei, R. W., Lee, M. L., & Hsu, W. (2021). Patient similarity analytics for explainable clinical risk prediction. *BMC Medical Informatics and Decision Making*, *21*(1), 207. https://doi.org/10.1186/s12911-021-01566-y

Goldstein, B. A., Navar, A. M., & Pencina, M. J. (2016). Risk Prediction With Electronic Health Records: The Importance of Model Validation and Clinical Context. *JAMA Cardiology*, *1*(9), 976–977. https://doi.org/10.1001/jamacardio.2016.3826

Hong, W., Earnest, A., Sultana, P., Koh, Z., Shahidah, N., & Ong, M. E. H. (2013). How accurate are vital signs in predicting clinical outcomes in critically ill emergency department patients: *European Journal of Emergency Medicine*, *20*(1), 27–32. https://doi.org/10.1097/MEJ.0b013e32834fdcf3

Hsieh, M. H., Hsieh, M. J., Chen, C.-M., Hsieh, C.-C., Chao, C.-M., & Lai, C.-C. (2018). Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. *Scientific Reports*, *8*(1), 1–7.

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G., Sirio, C. A., Murphy, D. J., Lotring, T., & Damiano, A. (1991). The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults. *Chest*, *100*(6), 1619–1636.

Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Medical Informatics and Decision Making*, *20*(1), 1–10.

Lehman, L. H., Adams, R. P., Mayaud, L., Moody, G. B., Malhotra, A., Mark, R. G., & Nemati, S. (2015). A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction. *IEEE Journal of Biomedical and Health Informatics*, *19*(3), 1068–1076. https://doi.org/10.1109/JBHI.2014.2330827

Mahmoudi, E., Kamdar, N., Kim, N., Gonzales, G., Singh, K., & Waljee, A. K. (2020). Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review. *BMJ*, m958. https://doi.org/10.1136/bmj.m958

Min, X., Yu, B., & Wang, F. (2019). Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. *Scientific Reports*, *9*(1), 2362. https://doi.org/10.1038/s41598-019-39071-y

Pirracchio, R. (2016). Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 295–313). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_20

Truog, R. D., Campbell, M. L., Curtis, J. R., Haas, C. E., Luce, J. M., Rubenfeld, G. D., Rushton, C. H., & Kaufman, D. C. (2008). Recommendations for end-of-life care in the intensive care unit: A consensus statement by the American College of Critical Care Medicine. *Critical Care Medicine*, *36*(3), 953–963.

Whittemore, A. S. (2010). Evaluating health risk models. *Statistics in Medicine*, *29*(23), 2438–2452. https://doi.org/10.1002/sim.3991

Xu, J., Zhang, Y., Zhang, P., Mahmood, A., Li, Y., & Khatoon, S. (2017). Data mining on ICU mortality prediction using early temporal data: A survey. *International Journal of Information Technology & Decision Making*, *16*(01), 117–159.

Zhai, Q., Lin, Z., Ge, H., Liang, Y., Li, N., Ma, Q., & Ye, C. (2020). Using machine learning tools to predict outcomes for emergency department intensive care unit patients. *Scientific Reports*, *10*(1), 1–10.

CHAPTER 2. MD-MANIFOLD: A MEDICAL-DISTANCE-BASED MANIFOLD LEARNING APPROACH FOR MEDICAL CONCEPT REPRESENTATION AND DIMENSION REDUCTION

Shaodong Wang[1], Qing Li[1], and Wenli Zhang[2]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University

[2] Department of Information Systems & Business Analytics, Iowa State University

Modified from a manuscript submitted to *Information Systems Research*

Abstract

Investigating electronic healthcare records (EHR) using machine learning techniques has brought significant opportunities for healthcare predictive analytics. EHR data include meaningful, well-structured, yet extremely complicated medical concepts (e.g., diagnosis codes). Representing these categorical concepts for machine learning models often leads to high-dimensional, thereby, computationally-intensive problems. Therefore, dimension reduction is considered necessary in EHR preprocessing. However, current approaches have two major shortcomings. First, few methods are available to generate sensible low-dimensional representations for sets of medical concepts (i.e., the medical concepts in a medical record that contains a patient's clinical information in one medical practice), which most prediction models require. Second, few studies have leveraged the medical domain knowledge contained in medical concepts' properties (i.e., hierarchical structure and co-occurrences) while generating the representations for an improved prediction performance. This study proposes a new method to generate low-dimensional representations for sets of medical concepts. We first propose a new medical-concept level distance metric to incorporate medical domain knowledge into patient-patient networks. Then, using patient-patient networks as input, we generate low-dimensional

representations for medical records using manifold learning techniques. Next, we fuse the generated representations with other data modalities for better performance. To demonstrate the efficacy of our method, we evaluate it in two prediction tasks, readmission and in-hospital mortality predictions, over different patient cohorts using two large real-world medical databases, and compare it with 14 state-of-the-art baselines. The experimental results show that our method is very effective; it exceeds all baselines in 70% of the cases. (top-1 AUC scores) and reaches 80% of state-of-the-art performance (top-3 AUC scores). The low-dimensional representations generated by our method can be pre-trained and task-agnostic, therefore providing a computationally efficient solution for various healthcare prediction tasks. The innovative artifacts obtained from this study are salient to both the information system and the healthcare communities.

Introduction

Electronic healthcare records (EHRs) contain a wealth of information gleaned through diagnoses, treatment plans, laboratory and test results, etc (Jamie L. Habib, 2010). They are created by healthcare providers for specific encounters in hospitals and ambulatory environments. Normally, one record contains a patient's medical and treatment information in a single medical practice[1]. EHRs serve as the data source for electronic health records (EHRs), giving patients, physicians, and insurers access to a patient's medical records across time and facilities (Jamie L. Habib, 2010). Since the mid-2000s, the rapid increase in the use of health IT has made a vast amount of EHRs available to healthcare practitioners and researchers. EHR data have been leveraged to improve healthcare outcomes by providing new ways to approach evidence-based medicine.

[1] https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference

In the past few years, the rush to unlock the power of EHR data with the help of machine learning techniques has led to significant opportunities in various ways. However, (1) how to effectively represent medical concepts in EHR remains a challenge, and (2) high dimensionality is a curse when using EHR data for healthcare predictive analysis. These challenges are posed by a large number of medical concepts, such as medical codes/terms, drug names, and billing codes, in the medical language systems. These medical concepts are widely used in patient care, health services billing, public health statistics, and health services research. They are the primary cause of EHR data's high dimensionality in prediction models' feature spaces. For example, the World Health Organization's (WHO) International Classification of Diseases (ICD) is a comprehensive disease classification system that is widely used among healthcare organizations[2]. There are more than 17,000 ICD-9 (ICD Ninth Revision) codes (i.e., medical concepts) and often multiple ICD-9 codes in each medical record that stand for a patient's medical condition. In healthcare predictive tasks (e.g., predicting the readmission rate for a patient based on his/her current hospital visit), representing ICD-9 codes in each medical record as categorical data is challenging. Dimension reduction, which transforms data from a high-dimensional space into a low-dimensional space and still retains the original data's meaningful properties, is considered a necessary technique in EHR data processing.

Currently, there are three methods for generating low-dimensional representations for medical concepts in EHR data. (1) The first choice is to derive summary measures (e.g. Charlson Comorbidity Index (CCI)). However, the generalizability of such methods is constrained by their designed uses and applications (Charlson et al., 1987; Elixhauser, A. et al., 1998). (2) The second option is to map medical concepts to standard medical terminologies with acceptable dimensions

[2] https://www.who.int/standards/classifications/classification-of-diseases

(e.g. map the 5-digit ICD-9 codes to 3-digit ICD-9 categories or Clinical Classifications Software (CCS) codes). Although widely adopted, this option suffers from information loss and inferior performance in healthcare prediction tasks (Jung et al., 2019; Rasmy et al., 2020; Xiang et al., 2019). (3) The third strategy is to develop embeddings for individual medical concepts (i.e., low-dimensional representations of the individual medical concepts in the form of real-valued vectors). Such methods dramatically decrease the required dimensionality to represent a medical concept. However, limitation still exists due to the way we can use such embeddings for healthcare predictive analysis. Current embedding methods focus on representing individual medical concepts. Nonetheless, in the EHR data, a medical record contains multiple medical concepts. Researchers are further required to concatenate or aggregate these medical concept embeddings to obtain fixed-size inputs for prediction models. The crude aggregation and concatenation are likely to cause information loss and deteriorate machine learning models' predictive performance.

Therefore, there is a research gap in medical concept representation. To cope with the shortcomings of previous approaches, we propose to generate a low-dimensional representation of a set of medical concepts. In other words, we suggest generating a low-dimensional representation for all the medical concepts in a medical record. The generated representations are ready to be used by various machine learning algorithms as the input for different healthcare prediction tasks. Hence, our first research question is how we can generate sensible low-dimensional representations for sets of medical concepts.

Besides, medical concepts in EHR data have other non-negligible characteristics. (1) The first important property of medical concepts is the well-organized hierarchical structure that is determined by healthcare domain knowledge. For example, the ICD-9 system is designed to map

diseases to corresponding generic categories. Thus, major categories of ICD-9 codes include a set of similar medical conditions. For a more specific example, heart disease is one of the circulatory system diseases. Therefore, the ICD-9 code of heart disease ("420-429") belongs to the circulatory system disease ("390-459") in the ICD-9 hierarchy. Medical records with sets of similar medical concepts are likely to reflect patients' similar health conditions. When generating low-dimensional representations for sets of medical concepts, it is important to take medical concepts' hierarchical structure as domain knowledge, so that the generated low-dimensional representations align well with medical knowledge and help machine learning models reach better performance. (2) The second critical attribute of medical concepts lies in the co-occurrence of two medical concepts in the same medical record. Such co-occurrences indicate the propensity of the simultaneous presence of two diseases in a patient. The disease co-occurrences also form patient-patient networks, which are commonly used to connect and evaluate diseases that frequently co-occur. The medical concepts' co-occurring properties and the associated patient-patient network are important features for healthcare predictive models. However, few studies have investigated the co-occurrence properties of medical concepts during generating representations for sets of medical concepts. Therefore, our second research question is how we can incorporate the well-organized hierarchical structure and co-occurring properties of medical concepts when generating low-dimensional representations for sets of medical concepts.

To answer these two research questions, we propose a new framework, Medical-Distance-manifold (MD-manifold), which leverages the knowledge of both the hierarchical structures and co-occurrence properties of medical concepts to generate low-dimensional representations for sets of medical concepts in medical records. Various healthcare predictive tasks, such as readmission, mortality, and length-of-stay predictions, are expected to benefit from

the low-dimensional representations generated by our method. The proposed method consists of three steps. We first develop a new medical concept-distance metric for medical concept-distance calculation in the high-dimensional manifold feature space formed by medical concepts in EHRs. The new metric is knowledge-driven as well as data-driven, which preserves the medical domain knowledge in both medical concepts' hierarchical structure and co-occurrence properties. Second, we generate the patient-patient network using the proposed distance metric so that the generated network also has medical domain knowledge embedded in it. Then we introduce the patient-patient network to manifold learning algorithms and produce low-dimensional representations for sets of medical concepts. Last, we fuse multimodal data (i.e., the generated representations of medical concepts and patients' demographic information) for healthcare predictive analyses. To evaluate the effectiveness of the proposed method, we follow the design science paradigm (Hevner et al., 2004) and perform two healthcare prediction tasks (i.e., readmission and in-hospital mortality prediction) on two large real-world EHR databases.

The contributions of this study are twofold. (1) From the perspective of design science, we propose a novel method to generate low-dimensional representations for medical concepts that takes advantage of medical domain knowledge in the medical-concept hierarchy and disease interconnections embedded in the patient-patient network. Using our low-dimensional representations, the prediction models outperform multiple state-of-the-art methods in predicting hospital readmission and in-hospital mortality rates. The proposed method's generalization ability is strong in that other healthcare prediction tasks, whose prediction granularities are medical records-level (i.e., using a set of medical concepts, e.g., one hospital visit with a medical record or several hospital visits during a period with multiple medical records), can benefit from our work. (2) From the perspective of healthcare analytics, the proposed method can help

healthcare practitioners decide whether a patient should be considered for any intervention program to avoid readmission or reduce in-hospital mortality by providing an accurate forecast of the mortality rate. Because of the strong generalizability, our method has the potential to foster the actual use of healthcare prediction models in clinical practice, hence eventually improving healthcare outcomes and curbing healthcare costs.

In the remainder of this paper, we review the related literature in Section 2. We introduce the proposed framework in Section 3. Experiment results and discussions are presented in Section 4. We discuss the limitations and future direction of this work in Section 5.

Related work

This section provides a critical review of the literature on (1) medical concepts representation and dimension reduction, (2) manifold learning, (3) patient-patient network and medical concept-distance measuring, and (4) healthcare predictive analytics.

Medical concepts representation and dimension reduction

The rich information in EHR data is important for healthcare predictive analysis. In medical records, there are millions[3] of medical concepts, such as 17,000 ICD-9 codes (Song et al., 2020), 140,000 ICD-10 codes (Quan et al., 2005), 7,000 International Classification of Health Intervention[4], and 360,000 National Drug Codes[5]. Meanwhile, medical concepts are usually organized in a hierarchical structure based on their relationships, determined by healthcare domain knowledge. Besides, the simultaneous existence of two or more diseases in a patient is indicated by the co-occurrence of two or more medical concepts in the same medical record. In healthcare predictive analytics, it is crucial to represent and use these medical concepts

[3] UMLS contains over five million medical concepts: https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_002.html
[4] https://www.who.int/standards/classifications/international-classification-of-health-interventions
[5] https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory

and their relationships to improve prediction performance. This, however, is challenging for two reasons. (1) High dimensionality: many unique medical concepts form a high-dimensional manifold feature space, which is a primary cause of the deficiencies of many healthcare predictive models. As a result, dimension reduction is essential in processing EHR data. (2) The difficulty to leverage medical domain knowledge: during medical concepts representation, it is difficult to incorporate in medical concepts' hierarchical structure and co-occurrences. It is critical to represent such medical domain knowledge and aid predictive models in improving their effectiveness. To overcome these two challenges, researchers have done considerable work to develop representations for medical concepts.

Table 2.1. Existing methods in representing medical concepts.

Type	Algorithm	Literature	Limitations
Deriving summary measures	CCI	(Charlson et al., 1987)	Limited generalizability, inferior predictions results
	ECI	(Elixhauser, A. et al., 1998; van Walraven et al., 2009)	
	RSI	(Sacco Casamassima et al., 2014; Sessler et al., 2010)	
Mapping to higher level code hierarchy or standard terminologies	3-digit ICD-9	(H.-H. Wang et al., 2019; X. Wang et al., 2014)	Information loss, limited generalizability, ambiguity of optimal level of granularity and mapping sources
	CCS, CUI, SNOMED, Read 2, OPCS 4	(Deschepper et al., 2019; Melton et al., 2006; Min et al., 2019; Rasmy et al., 2020; Williams et al., 2017)	
Generating embedding for individual medical concept[6]	FastText	(Si et al., 2021; Tang et al., 2018; Youngduck Choi et al., 2016)	Inferior performance, variable-sized input data for machine learning prediction models, large training data size
	GloVe		
	Word2Vec		
	Supervised deep learning model	(Choi, Bahadori, Searles, et al., 2016; Choi et al., 2017, 2018)	

Typically, researchers have three strategies for resolving the high dimensionality problem, as summarized in Table 2.1. (1) The first choice is to derive summary measures.

[6] To clarify, these models can generate medical record level representations, which can be the aggregation of the embeddings for individual medical concepts.

Charlson et al. (1987) and Elixhauser et al. (1998) first propose Charlson Comorbidity Index (CCI) and Elixhauser Comorbidity Index (ECI) to categorize comorbidities of patients based on ICD-9 diagnosis codes (Schneeweiss et al., 2003). Such comorbidity indexes give each patient a single comorbidity score by summing the weighted values of predefined comorbidity categories. Although both CCI and ECI can predict in-hospital mortality, the weights and the predefined comorbidity categories need to be adjusted based on comorbidities' mortality risk or resource use. Risk Stratification Indices (RSI) are then proposed to overcome such limitations and provide more reproducible summary indices (Sacco Casamassima et al., 2014). However, the intended applications of these summary indexes limit their use in other healthcare predictive analyses. For example, CCI and ECI are used for mortality prediction while RSI is designed to predict length-of-stay and mortality for surgical patients. It is difficult to adjust them for other predictive tasks, thus, the generalizability is weak. Other limitations of summary measures include producing inferior predictions resulting from coding errors (particularly when relying on ICD-9 codes) and difficulties in determining if a diagnosis should be included to calculate these measures (Taneja, 2010). (2) Another common practice is to map the medical concepts to their higher-level code hierarchy or standard terminologies with acceptable dimensions, such as 3-digit ICD-9 categories, concept unique identifier (CUI) codes, and clinical classifications software (CCS) codes, to facilitate data aggregation and dimension reduction. For example, the ICD-9 code is designed to use the first three letters to capture the group-level disease information. Wang et al. (2019) reduce the feature dimension of their prediction tasks from 13,000 to 942 using 3-digit ICD-9 codes. Moreover, the UMLS' CUI codes provide mappings to almost all clinical terminologies at different hierarchical levels. Rasmy et al. (2020) selected it as one of their mapping terminologies. Furthermore, Min et al. (2019) investigated three different grouping

strategies: 3-digit ICD-9 codes, CCS codes, and Hierarchical Condition Category codes. These methods have several practical limitations as well. First, the terminology mappings are normally in many-to-many styles, which use inexact approximations and partial data discarding to represent the original medical concepts. Information loss is inevitable during the mapping process, which may hinder the accuracy and interpretability of the predictive models (Jung et al., 2019; Rasmy et al., 2020; Xiang et al., 2019). Second, terminology standards evolve constantly, and newer versions introduce additional levels of data redundancy. This method may restrict the generalizability of predictive models to vendor-specific solutions or even to a single hospital if the mappings are different between sites (Rasmy et al., 2020). (3) The third option is to construct low-dimensional representations (a.k.a., embeddings) for individual medical concepts with techniques borrowed from natural language processing (NLP) (Kowsari et al., 2019; Mikolov et al., 2013; Pennington et al., 2014). These techniques treat medical concepts in each patient's medical record as tokens in a natural language document (Table 2.2.). The embedding methods significantly reduce the number of dimensions necessary to represent a single medical concept (Choi, Bahadori, Searles, et al., 2016; Choi et al., 2017, 2018; De Vine et al., 2014; Si et al., 2021; Tang et al., 2018). (De Freitas et al., 2020) and Tang et al. (2018), for example, use unsupervised embedding models Word2Vec, GloVe, and FastText to build individual medical concept embeddings. Choi et al. (2016), Choi et al. (2017), and Choi et al. (2018) employ one-hot encodings as inputs and learn individual medical concept embeddings using supervised deep learning models such as a fully connected network, attention model, and multilevel network. The learned low-dimensional representations can be used in the clustering and classification of medical concepts. However, they are not suitable for healthcare prediction tasks due to the fact that these methods focus on representing individual medical concepts. A medical record, on the

other hand, usually contains many medical concepts. While most machine learning models require input vectors of the same dimensionality, concatenating embeddings of various medical concepts results in variable-sized input data. To use these embeddings for healthcare prediction tasks, researchers are required to further combine multiple medical concept embeddings in medical records. The usual practices include: averaging (or weighted averaging) multiple embeddings, summing over multiple embedding vectors, or concatenating a fixed number of embeddings (Figure 2.1. shows examples of these practices). For example, De Freitas et al. (2020) (Phe2Vec) use the weighted (i.e., weights are the frequencies of the medical concepts) summation of medical concepts' embeddings to build phenotype definitions for patient cohorts identification. Cui et al. (2018) concatenate the embeddings of the two most important medical codes in each hospital admission for the prediction of cost and length of stay. These approaches, which perform crude aggregations of embeddings, have a number of drawbacks as well. By averaging/summing embedding vectors, the information in the individual embeddings is partially lost during the aggregation process. The aggregated embeddings deteriorate predictive performance because of, for instance, being sensitive to extreme elements in the original embeddings or losing the encoded order of medical concepts (Stiebellehner et al., 2018). By concatenating the embedding vectors, the generated representations (with variable-sized) are inappropriate to use as input for predictive models (such as Linear Regression, Random Forest, AdaBoost, etc). To address it, researchers have to pick a fixed number of medical concepts, which lead to information loss as well. Besides, these methods suffer from the training data insufficiency problem. Most of these embedding methods are adapted from NLP-related deep learning techniques, which typically require training data volume that exceeds the capacity of most medical records systems (Choi et al., 2018; Mikolov et al., 2013).

Therefore, a research gap emerges. The representation of a set of medical concepts (i.e., all medical concepts in one or multiple medical records), instead of each individual medical concept, is required by most machine learning tasks. So far, representing sets of medical concepts in low-dimensional space is an important, yet largely untouched, research area.

Table 2.2. Examples of regarding medical records as documents and ICD-9 codes as tokens.

	Token 1: the	Token 2: train	Token 3: was	Token 4: late	Token 5: Mary	Token 6: waited	Token 7: for	…
Document 1: The train was late.	1	1	1	1				
Document 2: Mary waited for the train.	1	1			1	1	1	
…	…	…	…	…	…	…	…	…

(a) Documents and word tokens in NLP

	ICD-9 code 1: 42820	ICD-9 code 2: 42833	ICD-9 code 3: V066	ICD-9 code 4: 4019	ICD-9 code 5: 311	ICD-9 code 6: 490	…
Medical record 1:	1	1	1	1			
Medical record 2:	1	1			1	1	
…	…	…	…	…	…	…	…

(b) Medical records and ICD-9 codes in EHR

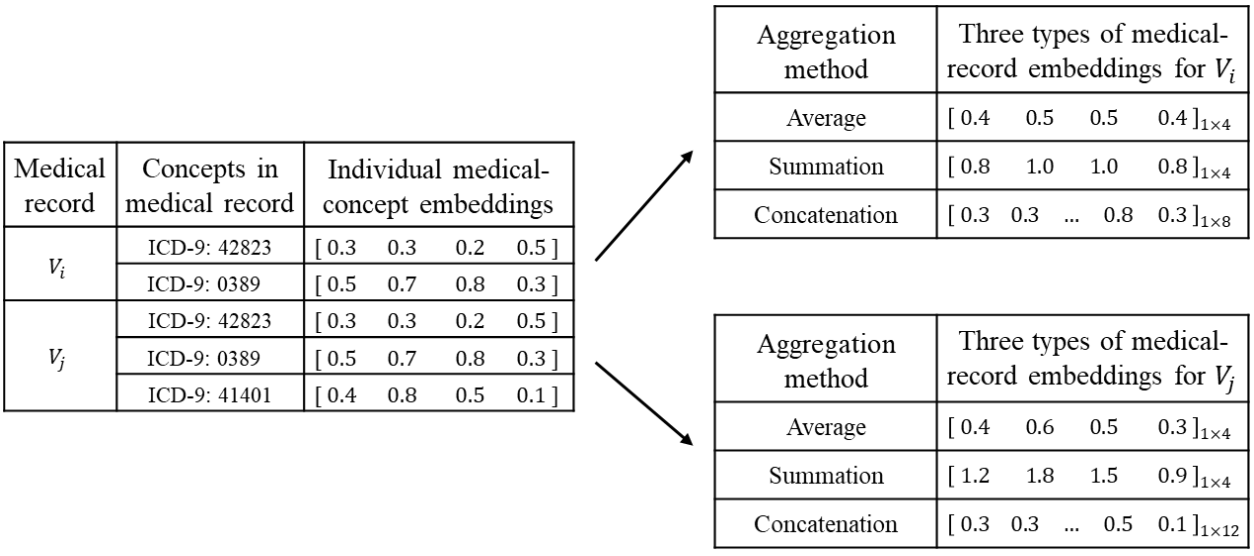


Figure 2.1. Example of medical concept embedding aggregation.

Furthermore, the medical concepts in the EHR data have two significant features that indicate important medical domain knowledge. First, medical concepts are usually organized in a hierarchical structure based on their relationships, which are determined by healthcare domain knowledge. For example (Figure 2.3. (a)), ICD-9 codes, which are widely used among healthcare organizations, present a hierarchical structure that the lower-level (child) code is a subtype of its upper-level (parent) code, e.g., 42823 (acute on chronic systolic heart failure) is a subtype of 4282 (systolic heart failure). Naturally, sibling codes that have the same parent code have similar clinical implications. As a result, it's logical to infer that medical concepts with short distances in the hierarchical structure have higher possibilities to reflect patients' similar health conditions. Thus, it is important to take the hierarchical structure of medical concepts as domain knowledge during the concept representation process. Then, the generated representations can help machine learning models reach better performance because medical domain knowledge is meaningfully incorporated. Second, the co-occurring of two or more medical concepts in the same medical record shows the propensity of the simultaneous presence of two or more medical conditions in a patient. Thus, the medical concepts' co-occurring frequencies are important features for healthcare prediction models.

The three above-mentioned known techniques, however, have limitations in preserving the medical domain knowledge contained in medical concepts' properties. The first option (i.e., the summary measures) does not consider the medical concepts' hierarchy or co-occurrences as domain knowledge. The second option (i.e., code mappings) makes use of knowledge about the hierarchical structure of medical concepts but ignores the co-occurring frequencies or subtle differences between sibling medical concepts in the hierarchy. The third option (i.e., embeddings for individual medical concepts) can incorporate medical domain knowledge into the individual

medical concept's embedding (Choi et al., 2017) but there is information loss during the aggregation of multiple embeddings for prediction models. Therefore, our second research question is how can we incorporate the well-organized hierarchical structure and co-occurring frequencies of medical concepts as domain knowledge when generating low-dimensional representations for sets of medical concepts in EHR.

Manifold learning for dimension reduction

We believe that manifold learning (Talwalkar et al., 2008) is an excellent method for addressing the first research question - generating representations for sets of medical concepts. In healthcare prediction tasks, medical concepts in the EHR create a high-dimensional manifold space. Each data point in the manifold space represents a medical record with various medical concepts (Figure 2.5. (f)) (a data point can alternatively be all medical records in a period). Manifold learning is an approach for non-linear dimensionality reduction which learns the high-dimensional inherent structure of the data and is capable of capturing non-linear structures in the data (Silva & Tenenbaum, 2003).

The way manifold learning algorithms depict the manifold space varies, but they all follow a similar pattern. First, manifold learning algorithms construct the nearest neighbor network to create a representation of all the data points. Second, by keeping the topology of the nearest neighbor network (i.e., the geometry of the original data points), manifold learning algorithms provide a low-dimensional representation for each data point. Formally, given n data points (e.g., n medical records, each contains multiple medial concepts), $X = \{x_i\}_{i=1}^{n}$ and $x_i \in R^d$, the goal of manifold learning is to find corresponding outputs, $Y = \{y_i\}_{i=1}^{n}$ where $y_i \in R^k$, and $k \ll d$. Manifold learning algorithms generate low-dimensional space Y to represent high-dimensional space X while keeping data points' internal structure in X.

There are two types of manifold learning algorithms: local and global approaches (Silva & Tenenbaum, 2003). Local approaches, for example, Local Linear Embedding (LLE) (Roweis & Saul, 2000), Laplacian Eigenmaps (Belkin & Niyogi, 2003), Hessian Eigenmaps (Donoho & Grimes, 2003), and Semidefinite Embedding (Weinberger et al., 2005), attempt to map adjacent points in the original high-dimensional space to nearby locations in low-dimensional space. The global approaches, such as Isomap (Tenenbaum et al., 2000) and Structural Preserving Embedding (Shaw & Jebara, 2009), like the local approaches, aim to keep the distance among adjacent points in the original high-dimensional space. In the meantime, they map distant points in the original high-dimensional space to distant locations in low-dimensional space. Because the local approaches focus on the relationship between nearby points, they are more computationally efficient than the global approaches (Silva & Tenenbaum, 2003). However, they may not retain the global topography of the original feature space (Shaw & Jebara, 2009). The global approaches tend to provide more reliable representations by preserving the global structure of the original manifold space (Silva & Tenenbaum, 2003).

In this study, we adapt two of the most widely used manifold learning algorithms, Laplacian Eigenmap (i.e., a local method) and Isomap (i.e., a global approach), to generate representations for sets of medical concepts. We then compare their performance in a variety of healthcare prediction tasks. (1) Laplacian Eigenmap first computes nearest neighbors for each data point and creates the weighted nearest neighbor network (i.e., node: each data point, edge: nearest neighbor relationship, edge weight: proportional to the reverse distance between nearest neighbors). The larger the edge weight, the more similar the nodes, and the closer they are in the manifold space. Laplacian Eigenmap computes a low-dimensional representation of the data point and optimally preserves the local neighborhood information. (2) Isomap also calculates the

nearest neighbors for each data point and develops the nearest neighbor network (i.e., node: each data point, edge: nearest neighbor relationship, edge length: distance between nearest neighbors). It then computes the shortest path distances between all pairs of points in the network. Isomap preserves the global structure of the original manifold space and finds the optimum low-dimensional representation for data points by retaining the geodesic distance between each pair of nodes on the constructed nearest neighbor network.

Manifold learning can be an excellent way to generate low-dimensional representations for sets of medical concepts in a medical record. Since the first and essential step in manifold learning (i.e., both the local approach LE and the global approach Isomap) is to build the nearest neighbor network, it is crucial that we find the proper network to represent the medical records data and compute the distance between medical records for developing the nearest neighbor network. Such a network should also have medical domain knowledge embedded, including the hierarchical structure and co-occurrences of medical concepts, so that the generated representations from manifold learning naturally inherit the medical domain knowledge.

Patient-patient network

To address the second research question, i.e., incorporating medical domain knowledge when generating low-dimensional representation for a set of medical concepts, we introduce the patient-patient network as manifold learning's nearest neighbor network.

In our research, each data point in a manifold space (i.e., formed with numerous medical concepts) represents a medical record with various medical concepts. We attempt to generate low-dimensional representations for medical records and preserve the similarities (i.e., distances) among medical records based on two assumptions. Assumption 1: if two patients' medical records contain similar medical concepts, their health conditions are similar. Assumption 2: patients with similar health conditions may have similar healthcare outcomes.

We find that the patient-patient network is an ideal data structure for constructing medical records' nearest neighbor network while preserving medical domain knowledge in medical concepts' properties. Patient-patient network is a sub-research area of the human disease network, which aims to understand human diseases through network theory. Instead of viewing disease as an independent entity, the human disease network provides a powerful way for uncovering hidden links between diseases and other biomedical entities like genes and disease pathologies (García del Valle et al., 2019). The human disease network is an important and fast growing research area. Since its inception, various human disease networks have been developed based on the similarities and relationships between diseases at biological level (e.g., genes, proteins, or compounds) or phenotypic level (e.g., comorbidity or side-effects). For example, Goh et al. (2007) and Barrenas et al. (2009) first propose the disease-gene networks, which are used to understand disease-gene associations. Then, Campillos et al. (2008) and Yıldırım et al. (2007) create disease-drug networks and use them to identify the new use of old drugs. Later, a disease-metabolic network and a molecular interaction network are proposed to investigate the disease phenotypes and genetic defects (D.-S. Lee et al., 2008). Besides, Zhou et al. (2014) develop disease-symptom networks, which quantify the symptom similarity of diseases.

Recently, patient-patient networks, which are constructed based on patients' similarities (e.g., similarities in their medical records), have become an important research direction. The patient-patient network provides a way of describing disease interconnections from an epidemiological perspective. For example, Li et al. (2015) cluster similar patients through the patient-patient network to identify type 2 diabetes. (Pai et al., 2019) propose a patient-patient network to classify patients for precision medicine (Pai & Bader, 2018). To analyze cardiac risk

factors, Hou et al. (2021) construct a patient-patient network in which patients with similar health conditions are connected.

The nodes in patient-patient networks are usually patients (e.g., patients' medical records and demographic information), while the edges represent the similarities between patients (e.g., co-occurrence of disease). The existing patient-patient networks do not take into account the well-organized hierarchy of medical concepts as medical domain knowledge. Therefore, we propose to define a new patient-patient network (i.e., nodes, edges, and edge weights) with both medical concepts' hierarchy and co-occurrences embedded. To build such a patient-patient network, we need a new distance metric for medical concepts and medical records. Using such a metric, the patient-patient network's edge weights represent the medical domain knowledge and the similarities between medical records. As long as we have the proper metric to calculate the distance between medical records, we can create the nearest neighbor network for manifold learning algorithms and generate low-dimensional representations for sets of medical concepts, and further use the generated low-dimensional representations for healthcare predictive analysis.

In the below subsection, we review the existing literature on medical concepts' distance metrics.

Distance metrics of medical concepts

There are two steps to construct the distance between two medical records that contain multiple medical concepts: concept-level distance and record-level distance calculations (Jia et al., 2019). The concept-level distance measures the distance between medical concepts, whereas the record-level distance measures the distance between medical records based on concept-level distance.

As summarized by Jia et al. (2019), two existing metrics have incorporated the medical-concept hierarchy for concept-level distance calculation. Wu and Palmer (1994) propose a metric

that considers the two concepts as close in the concept hierarchy if their least common ancestor is close. Yuhua Li et al. (2003) introduce two parameters in their metric to assign weights to the compared concepts and their least common ancestor. We adopt Wu and Palmer (1994)'s metric CD_{WP} as the baseline (please see Appendix A for more details) on account of its simple design and powerful performance (Jia et al., 2019).

Though powerful, CD_{WP} has its limitations – the distance is fully pre-determined by the medical concepts' hierarchical structure regardless of the concepts co-occurring frequencies in the EHR. To make it more concrete, two medical concepts that are distant in the hierarchic structure can frequently co-occur in the real-world EHR data, which means they tend to relate closely to each other. However, such co-occurrence relationships cannot be reflected in the medical concepts' hierarchy. Moreover, it is likely that one particular medical concept occurs more frequently than its siblings (see the example in Section 3.2.2). Thus, it is reasonable to believe that this medical concept may have a closer relationship with its parent than its siblings. Nevertheless, using CD_{WP}, the distances between a parent node and its child nodes are equal. For example, in Figure 2.3. (a), $CD_{WP}(42820{,}4282) = CD_{WP}(42823{,}4282)$, regardless of the frequencies of 42820 and 42823 in the real-world dataset. To address this limitation, we propose a concept-level distance metric that is both knowledge-driven and data-driven (see more details in Section 3.2.2).

Meanwhile, there are four popular medical record-level distance metrics that measure distances among medical records. Yuhua Li et al., (2003) calculate the distance between two medical records based on the most similar concept pairs' average distance. Girardi et al. (2016) measure the distance between medical records based upon the average distance of all concept pairs that are not in the intersection of the two medical records, focusing on the difference

between two medical records. Jia et al. (2019) propose two medical record-level distance metrics. The first one calculates the average of the distances of all concept pairs, and the second one computes the average distance of the minimum weighted bipartite matching. Each distance metric has its own significance, and the results vary according to applications.

We can use suitable medical concept distance metrics to solve our second research question, i.e., retaining medical domain knowledge in medical concepts' attributes. Therefore, we propose a new medical concept distance metric in this work, and we develop a new patient-patient network to represent the relationships of sets of medical concepts in medical records using the new metric. The new patient-patient network is then embedded in manifold learning algorithms to represent sets of medical concepts. The obtained low-dimensional representations are ready to be used for further analysis, such as readmission and mortality predictions.

Healthcare predictive analytics

Predictive analytics, including empirical methods for generating and evaluating predictions, is an important research area in information systems (Shmueli & Koppius, 2011). The primary goal of predictive analytics is to create models with practical applications, i.e., generating accurate and robust prediction results for new observations. While researchers may use predictive analytics for explanatory modeling (i.e., explain whether a factor contributes to an outcome), explanatory power does not imply predictive power. This critical difference drives distinctive principles for predictive model development and evaluation - predictive analytics focuses on building empirical models that predict well (Shmueli & Koppius, 2011).

In healthcare, predictive analytics examine historical and real-time medical data and make diagnostic or prognostic risk predictions (Shmueli & Koppius, 2011; Van Calster et al., 2019). The research results of healthcare predictive analytics have a wide range of implications. (1) Identify patients at risk and support clinical decision making on individual patient level. For

example, Bardhan et al. (2015) propose a model to predict the propensity, frequency, and timing of readmissions of patients diagnosed with congestive heart failure. Ben-Assuli & Padman (2020) investigate the impact of time-stable and time-varying covariates in predicting recurrent, unplanned readmissions for patients with chronic diseases. Lin et al. (2017) propose a Bayesian multitask learning approach that allows healthcare providers to achieve multifaceted risk profiling and model an arbitrary number of patient's risk of future adverse health events. (2) Track the health of populations and inform interventions on a population level. For example, Zhang & Ram (2020) develop a machine learning-based framework to predict asthma risk factors that can provide guidance for developing interventions for specific subpopulations. (3) Monitor healthcare practitioners' performance and provide insight into hospitals' administrative challenges. For example, Meyer et al. (2014) develop and illustrate a machine learning approach to improve dynamic decision making for the treatment of patients with type 2 diabetes mellitus.

In this study, we use two healthcare prediction tasks as research cases to evaluate the effectiveness of our proposed method: readmission prediction and in-hospital mortality prediction. (1) Readmission prediction is critical in curbing the cost of healthcare and improving patient outcomes. First, LACE index is developed to evaluate the likelihood of patient readmission (van Walraven et al., 2010) based on the length of stay (L), acuity of the admission (A), comorbidity of the patient (C), and emergency department use in the duration of 6 months before admission (E). Later, van Walraven et al. (2012) improve the performance of LACE by adjusting its parameters. Later, machine learning models are widely implemented for accurate readmission prediction (Baillie et al., 2013; Cotter et al., 2012; Yu et al., 2015). With the recent development of deep learning, various sequential models are used in readmission prediction (Allam et al., 2019; Ashfaq et al., 2019; Min et al., 2019). (2) In-hospital mortality prediction is

of paramount importance for assessing the severity of disease (Becker & Zimmerman, 1996), adjudicating new treatments (Pirracchio, 2016), comparing patients' cohorts treated across different hospitals (Becker & Zimmerman, 1996), allocating resources and determining levels of care (J. Lee et al., 2016), and discussing expected outcomes with the hospitalized patients (J. Lee et al., 2016). Healthcare practitioners develop severity scoring systems for in-hospital mortality prediction, such as Acute Physiology and Chronic Health Evaluation (APACHE) IV (Zimmerman et al., 2006) and Simplified Acute Physiology Score (SAPS) III (Moreno et al., 2005). To improve the performance of mortality prediction models, researchers explore the value of EHR data using various machine learning and deep learning models (Altibi et al., 2021; Kong et al., 2020). Though promising, the existing models fail to make full use of the information contained in EHR data because of the limitations of current medical concepts representation techniques.

Using different healthcare prediction tasks as research cases, we aim to show that the low-dimensional representations generated by our method have many uses because they can be (1) pre-generated and task-agnostic, (2) concatenated with other patients' information, and (3) incorporated into other healthcare prediction models.

Summary

The deficiencies of existing medical concept representations, coupled with the challenges in incorporating medical domain knowledge while generating low-dimensional representation for sets of medical concepts, motivate us to develop a new method for medical concepts representation and dimension reduction. The design science paradigm provides a good foundation for our work. Design science is an outcome-based research methodology (Nunamaker et al. 1990). According to its definition, a design is both a product and a process (Hevner et al. 2004). The product is an artifact that can be broadly defined as construct, method, model, or

instantiation (Simon 1996). The process is a sequence of expert activities composed of developing and evaluating the artifact (March and Smith 1995). In this study, the artifact we intend to deliver is a framework consisting of methods and instantiations that is capable of (1) incorporating the medical domain knowledge in the well-organized hierarchical structure and co-occurrences of medical concepts, and (2) generating low-dimensional representations for sets of medical concepts.

Research design: Medical Distance-Manifold (MD-Manifold)

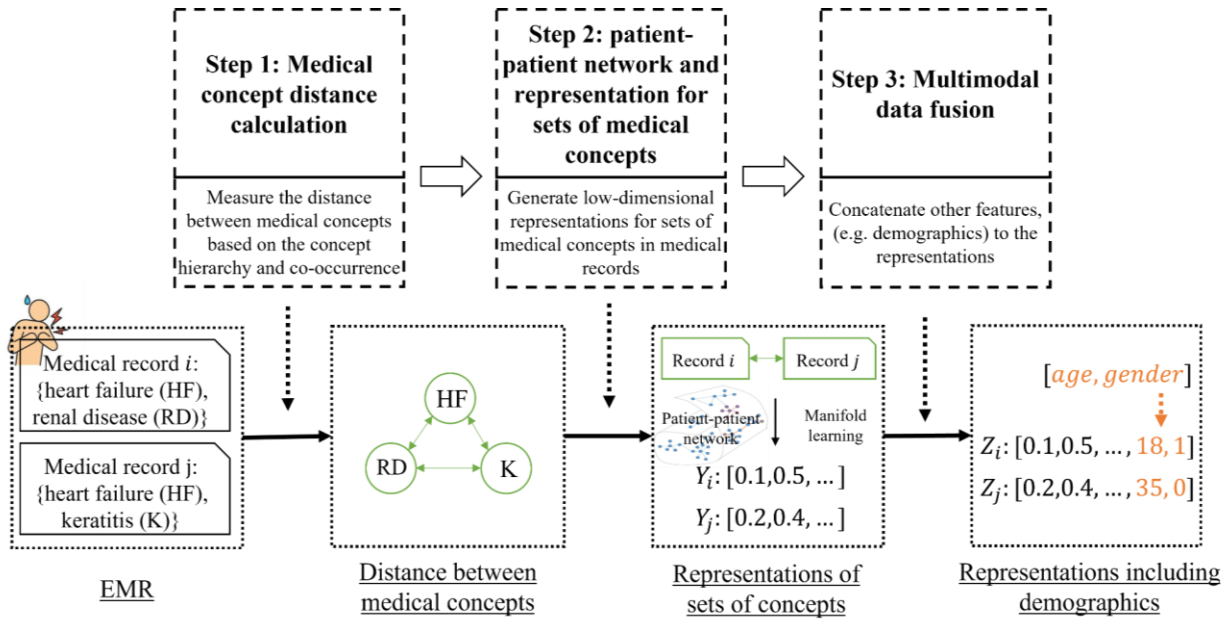


Figure 2.2. The proposed MD-Manifold framework.

We introduce our research design of MD-Manifold in this section. As shown in Figure 2.2., the proposed method consists of three steps. Step 1: Medical concept distance calculation. We first create a new medical concept-distance metric that is both knowledge-driven and data-driven to preserve medical domain knowledge in medical concepts' properties. Step 2: Patient-patient network construction and sets of medical concepts' representation generation. We construct patient-patient networks as the nearest neighbor networks for manifold learning algorithms. Then, we adapt manifold learning algorithms to represent sets of medical concepts in

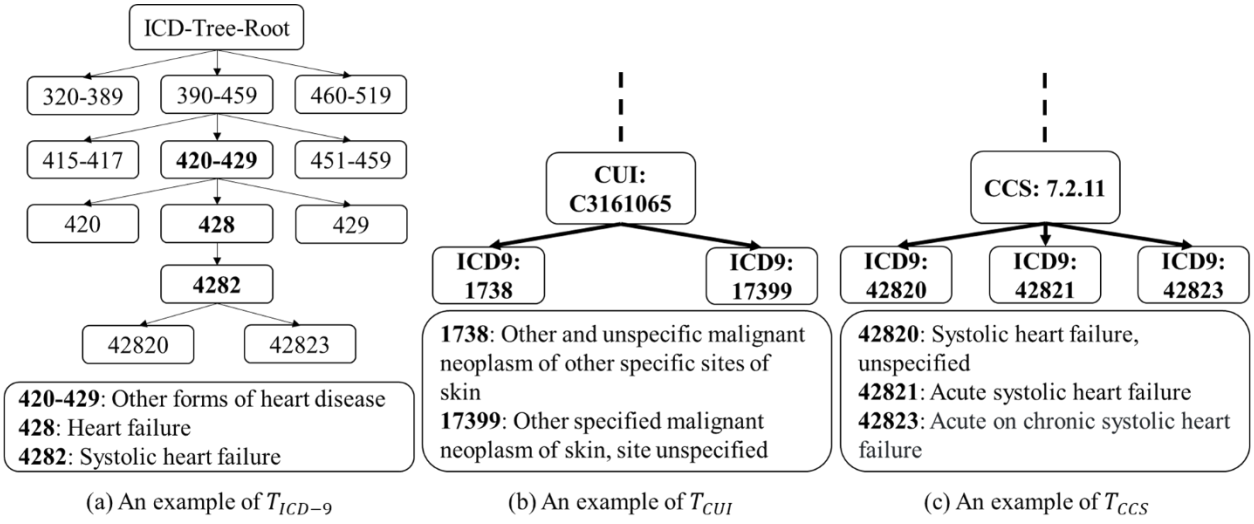
a low-dimensional space. Step 3: Multi-mode data fusion. Furthermore, we fuse multimodal embeddings (i.e., medical record embeddings and patients' demographic embeddings) as the input of diverse healthcare predictive analytical tasks.

Terminology

We denote an EHR dataset as D with n medical records V_i $(i = 1, 2, \dots, n)$, where a medical record $V_i$contains a set of medical concepts M_j's to describe a patient's health condition in a single medical practice, $V_i = (M_1, M_2, \dots, M_{h_i})$. M_j $(j = 1, 2, \dots, h_i)$ is a medical concept with the number of concepts to be $h_i \in [1, m]$, where m is the maximum number of medical concepts for a medical record. Each medical record contains different number of medical concepts, therefore $m = max_i(h_i)$. We then define the medical-concept hierarchy structure as a prefix tree T (Fredkin, 1960) derived from the medical domain knowledge. A tree T has a root node N_{root}, the internal nodes N_{branch}'s (i.e., branch nodes), and the terminal nodes N_{leaf}'s (i.e., leaf nodes, N_{leaf}'s are equivalent to M_j's in D). The relations between the root, branch, and terminal nodes are represented as a set of linked nodes. In this study we explore the performance of three prefix trees, i.e., T_{ICD9}, T_{CUI}, and T_{CCS}. Specifically, (1) T_{ICD9} represents the relationship between a medical concept and its higher-level ICD-9 disease diagnosis categories, as shown in Figure 2.3. (a). ICD-9 diagnosis codes are composed of codes with 3, 4, or 5 digits[7], which are medical concepts in our research setting. Three-digit ICD-9 codes stand for the categorical information of diseases. Three-digit ICD-9 codes are further divided by the use of fourth and/or fifth digits, which provides greater details of diseases. Hence, the medical domain knowledge is contained in the ICD-9 diagnosis codes' hierarchical structure. (2) T_{CUI} exhibits the relationship between the medical concepts and the corresponding Concept Unique Identifiers (CUI) from the

[7] https://www.cdc.gov/nchs/data/icd/icd9cm_guidelines_2011.pdf

UMLS. As shown in Figure 2.3. (b), two medical concepts may indicate similar diagnoses, and CUI links these medical concepts in D that mean exactly or nearly the same. Therefore, UMLS Metathesaurus structure, which represents the properties of diseases and their relations to other diseases, serves as the source of medical domain knowledge (Bodenreider & Stevens, 2006). (3) T_{CCS} reflects the projection of medical concepts M_j's onto the CCS categorization scheme. As the example in Figure 2.3. (c) shows, a group of medical concepts at the bottom can be collapsed into a smaller number of clinically meaningful categories (CCS codes[8]) that are sometimes more useful for presenting descriptive statistics than the individual medical concept.



(a) An example of T_{ICD-9} (b) An example of T_{CUI} (c) An example of T_{CCS}

Figure 2.3. Examples of medical-concept hierarchy structures.

Step 1: Medical concept-distance calculation: a new medical concept-distance metric that is both knowledge-driven and data-driven.

Deriving suitable distances among medical concepts is crucial for incorporating medical domain knowledge contained in medical concepts' hierarchy and co-occurrences. Nevertheless, the most state-of-the-art medical concept distance metric, like CD_{WP}, has limitations. As discussed in Section 2.3.1, high co-occurrence frequencies of medical concepts in the real-world

[8] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

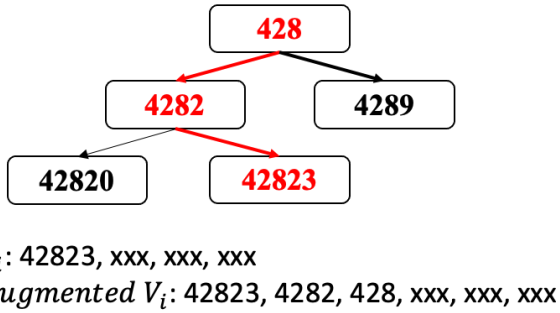
EHR data indicate their close relationships. However, using CD_{WP}, the distance of two medical concepts is solely determined by their relative positions in the concept hierarchy T, which does not reflect their co-occurring frequencies in the real-world observational data. To overcome this limitation, we define a new type of medical-concept distance metric CD_{new} that considers both the medical concepts' hierarchical structure and co-occurrences in the dataset D. The proposed new metric is both knowledge-driven and data-driven.

Knowledge-driven occurrence matrix construction

We first construct an occurrence matrix $O_{n\times N}$ for all medical records in D, where n is the number of medical records, and N is the total number of medical concepts in the data. Denote each element of O as O_{ij}, where i is the index of the medical record, and j is the index of the medical concept. First, for each medical record V_i, we augment it by adding all the ancestors (N_{branch}'s) of its medical concepts. Then, we set $O_{ij} = 1$ if the j^{th} concept occurs in the augmented V_i, otherwise, $O_{ij} = 0$.

Figure 2.4. presents an example of the augmented V_i. If V_i contains an ICD-9 code 42823, the augmented V_i contains ICD-9 codes 4282 and 428, which are the ancestors of the ICD-9 code 42823. The red line shows the path in the prefix tree T_{ICD9} from 42823 to 4282 and 428. Figure 2.5. (a) and (b) presents an example of the occurrence matrix, O. Suppose we have a dataset, the second column in Figure 2.5. (a) shows the medical concepts (e.g., ICD-9 codes) that belong to each medical record, and the third column shows the corresponding frequencies. The first row indicates that there are ten medical records in the dataset that contain both medical concepts 4289 and 42823. We insert their ancestors into V_i to obtain the augmented V_i (Figure 2.5. (b)). Then we obtain the occurrence matrix O in Figure 2.5. (c).

The purpose and advantage of constructing the occurrence matrix, O, are to keep the medical domain knowledge in the tree T for later use when generating low-dimensional representations. The construction of the occurrence matrix O is a knowledge-driven process because (1) it keeps the path information in the prefix tree T, and (2) by adding all the ancestors that one medical concept belongs to into the corresponding medical record V_i, we add the disease categorical information (e.g., if use T_{ICD9} as the knowledge source) or higher-level medical concept ontologies (e.g., if use T_{CUI} as knowledge source) into the augmented V_i, depending on how the medical domain knowledge tree T was constructed.



Note: (1) The augmented V_i that keeps the path in the prefix tree T_{ICD9} from 42823 to 4282 and 428.
(2) xxx in V_i represents the other ICD-9 codes in the same medical record. The ICD-9 code 428 is not at the highest level in the hierarchy. We do not show the parents of 428 for simplicity.

Figure 2.4. An example of an augmented V_i.

Data-driven co-occurrence matrix construction

In the next step, we construct a co-occurrence matrix $C_{N\times N}$ by computing the co-occurrences of medical concepts in D, where N is the total number of medical concepts. We calculate the co-occurrence matrix using the occurrence matrix O and $C = O^T O$. C is symmetric, and the non-diagonal elements in C are the co-occurrences of medical concepts. The co-occurrence matrix C is used to calculate the distance among medical concepts in the next step.

The construction of the co-occurrence matrix C is a data-driven process because the elements (i.e., co-occurring frequencies) in C is derived from the medical record dataset D. The

co-occurrence matrix C has important implications in healthcare prediction tasks: first, the co-occurrence of diseases in a medical record is often referred to as the comorbidity or multimorbidity in clinical practice, which is highly related to many negative health conditions, such as anxiety or depressive symptoms, functional impairment, and mortality (John et al., 2003); second, comorbidity usually associates with the linked diseases at the molecular level (Barabási et al., 2011; Hidalgo et al., 2009), which provides implicit information for healthcare predictive analysis (Emmert-Streib et al., 2013).

Using the augmented V_i in matrix O is beneficial to the construction of matrix C due to two reasons. (1) The augmented V_i contains disease categorical information or higher level medical concept ontologies. The co-occurrence matrix C, therefore, contains both the co-occurring relationships between diseases and the co-occurring relationships between disease categories. (2) Assume two medical concepts: M_j and $M_{j'}$ (i.e., both are leaf nodes in T) are rare in D and frequently co-occur. Such a relationship is difficult to capture due to the rarity of M_j and $M_{j'}$, as well as the fact that they may have varied co-occurrences with other medical concepts in D. On the other hand, their parent nodes N_{branch_j} and $N_{branch\_j'}$ in the augmented V_i have a significantly higher possibility of co-occurring, because M_j's and $M_{j'}$'s sibling concepts (sibling concepts indicate the nuanced difference in the medical knowledge) have the same parent nodes. The co-occurring relationship between N_{branch_j} and $N_{branch\_j'}$ is considerably easier to capture and be represented for M_j and $M_{j'}$ in the co-occurrence matrix C. Such a relationship is important for medical-concept distance calculation since the vector representations of M_j and $M_{j'}$ should have a relatively short distance in the manifold feature space because they tend to co-occur. The co-occurring relationships are important medical

domain knowledge that we strive to preserve when obtaining the representation for sets of medical concepts.

Calculating the medical-concept distance

In the last step of the medical concepts' distance calculation, we consider each row of the co-occurrence matrix as a vector that represents a medical concept. All the medical concepts form a manifold space. We define a new type of medical-concept distance metric (note the metric induces the node topology in the manifold space): $CD_{new} = distance(C_a, C_b)$, where a and b are two medical concepts, and C_a and C_b are row a and row b of the co-occurrence matrix C, respectively. The $distance(\cdot)$ functions can be defined in different ways (Table 2.3.). In this work, we compare four distance formulas: $Cosine(C_a, C_b)$, $Manhattan(C_a, C_b)$, $Euclidean\ (C_a, C_b)$, and $eHDN(C_a, C_b)$. In the evaluations, we compare and identify which ones are better for different healthcare prediction tasks.

Table 2.3. $CD_{new} = distance(C_a, C_b)$.

CD_{new}	$distance(\cdot)$	$CD_{new} = distance(C_a, C_b)$	Notes
$CD_{new-Cosine}$	$Cosine(C_a, C_b)$	$1 - \frac{C_a \cdot C_b}{\sqrt{C_a \cdot C_a}\sqrt{C_b \cdot C_b}}$	Measures inner product of two normalized vectors, which is also the same as the angle between vectors.
$CD_{new-Manhattan}$	$Manhattan(C_a, C_b)$	$\|\|C_a - C_b\|\|_1$	Measures the sum of the absolute differences between C_a and C_b, which is also known as the *l1* norm.
$CD_{new-Euclidean}$	$Euclidean\ (C_a, C_b)$	$\|\|C_a - C_b\|\|_2$	Measures the sum of the squared differences between C_a and C_b, which is also known as the *l2* norm.
$CD_{new-eHDN}$	$eHDN(C_a, C_b)$	$1 - \frac{C_{a,b}N - \sum C_a \sum C_b}{\sqrt{\sum C_a \sum C_b (N - \sum C_a)(N - \sum C_b)}}$	Measures similarity between diseases using the observed probability co-occurrence of these diseases (Jiang et al., 2018).

Note: $C_{a,b}$: the number of co-occurrence between concepts a and b.
$\sum C_j, j \in (a, b)$: the summation of all elements in row j of C.
N: the total number of medical records in the data set.

	Medical concepts	Frequency
Medical record V_1	{4289, 42823}	10
Medical record V_2	{4289}	2
Medical record V_3	{42823}	10
Medical record V_4	{42820}	5

(a) An example of medical concepts in medical records.

	Medical concepts	Frequency
Augmented V_1	{4289, 42823, 4282, 428}	10
Augmented V_2	{4289, 428}	2
Augmented V_3	{42823, 4282, 428}	10
Augmented V_4	{42820, 4282, 428}	5

(b) An example of augmented V_i.

4289	42823	428	4282	42820	Frequency
1	1	1	1	0	10
1	0	1	0	0	2
0	1	1	1	0	10
0	0	1	1	1	5

(c) Knowledge-driven occurrence matrix, O.

	4289	42823	428	4282	42820
4289	12	10	12	10	0
42823	10	20	20	20	0
428	12	20	27	25	5
4282	10	20	25	25	5
42820	0	0	5	5	5

(d) Data-driven co-occurrence matrix, C.

Figure 2.5. A run-through example of the medical concept distance calculation and the patient-patient network construction for manifold learning.

Figure 2.5. (e) shows an example of $CD_{new-Cosine}$ given the co-occurrence in Figure 2.5. (d). Notice that the concept 42823 occurs more frequently than 42820 in Figure 2.5. (a). It is reasonable to believe that patients with an upper-level concept 4282 are more likely to have 42823 than 42820 as a specified disease, which indicates that the concept 4282 is more related to 42823 than 42820. By using our method $CD_{new-Cosine}$, as we expected, (42823,4282) has a

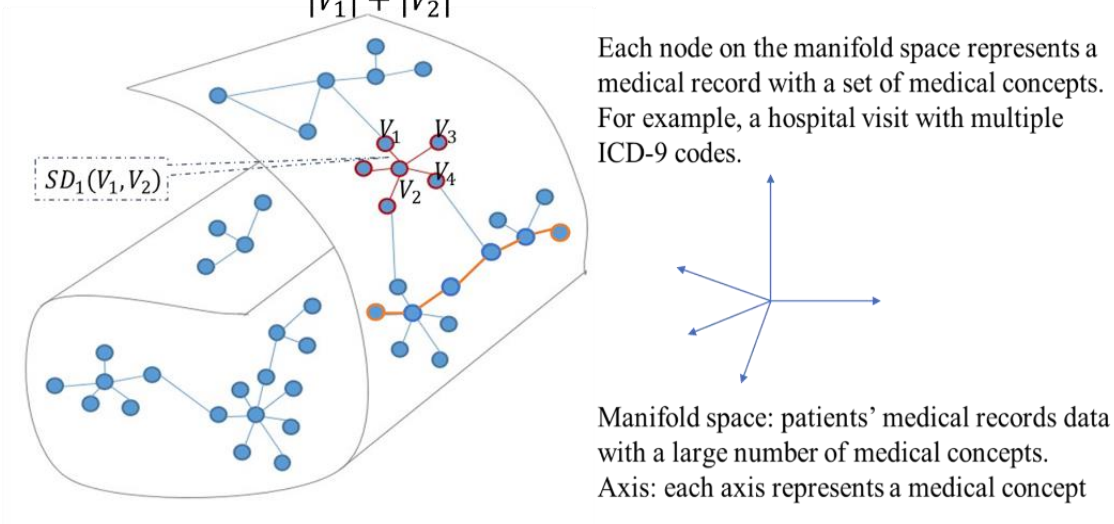
smaller distance than (42820,4282) with $CD_{Cosine}(42823,4282) = 0.0125$ and $CD_{Cosine}(42820,4282) = 0.2463$.

	4289	42823	428	4282	42820
4289	0	**0.0458**	0.0523	0.0652	**0.4250**
42823	0.0458	0	0.0133	**0.0125**	0.3595
428	0.0523	0.0133	0	0.0014	0.2495
4282	0.0652	0.0125	0.0014	0	**0.2463**
42820	0.4250	0.3595	0.2495	0.2463	0

(e) Medical concept distance calculation,

e.g. $CD_{new-Cosine}(4289,42823) = 1 - \frac{C_{4289} \cdot C_{42823}}{\sqrt{C_{4289} \cdot C_{4289}}\sqrt{C_{42823} \cdot C_{42823}}} = 0.0458$

$$SD_1(V_1, V_2) = \frac{1}{|V_1| + |V_2|} CD_{new-Cosine}(42823,4289) = 0.0153$$



(f) Patient-patient network for manifold learning

Figure 2.5.Continued.

Moreover, due to the higher co-occurrence frequency of (4289,42823) than (4289,42820), $CD_{Cosine}(4289,42823) = 0.0458$ is smaller than $CD_{Cosine}(4289,42820) = 0.425$, in spite of the equal-distance relationship in the medical concept hierarchy T_{ICD9} , which overcomes the limitation of the baseline distance metric CD_{WP} (i.e., without taking into account

their co-occurring frequencies, the distance between two medical concepts is exclusively determined by their positions in T).

Step 2: Patient-patient network construction and the representation for sets of medical concepts generation

In the second step, we generate medical record-level representations by constructing a patient-patient network and introducing it to manifold learning algorithms. To construct a patient-patient network, we first measure distances among medical records using the medical concept-level distance. Then we construct the patient-patient network by connecting similar medical records based on the medical record distance. Last, using manifold learning, we map the constructed patient-patient network into a low-dimensional latent space, where a vertex (a row of D, i.e., a medical record) in the patient-patient network is represented as a low-dimensional vector.

Table 2.4. $SD = distance(V_i, V_j)$.

SD	$distance(V_i, V_j)$	Notes
SD_1	$\frac{1}{\lvert V_i\rvert + \lvert V_j\rvert}(\sum_{a\in V_i} \underset{b\in V_j}{minCD}(a,b) + \sum_{b\in V_j} \underset{a\in V_i}{minCD}(b,a)$	Uses the average distance of the most similar concept pairs to calculate the medical-record distance.
SD_2	$\frac{1}{\lvert V_i \cup V_j\rvert}(\sum_{a\in V_i \setminus V_j} \frac{1}{\lvert V_j\rvert}\sum_{b\in V_j} CD(a,b) + \sum_{b\in V_j \setminus V_i} \frac{1}{\lvert V_i\rvert}\sum_{a\in V_i} CD(b,a))$	Defines the medical-record distance as the average distance of all concept pairs that are not in the union of two sets.

Based on the medical concept distance in Step 1, we are able to calculate the distance between medical records $SD = distance(V_i, V_j)$, where each record contains a set of medical concepts. We adopt and compare four widely used metrics (Table 2.4.) for sets of medical concepts distance calculation (Jia et al., 2019). SD_1 and SD_4 are designed to capture the similarities of the most similar medical-concept pairs from two medical records. SD_2 does not include the overlapping medical concept but focuses on the difference between two medical

records. SD_3 is widely used in clustering analysis in measuring the distance between each cluster (also known as "Average Linkage"). Because each metric has its merit in finding the distance between medical records, we compare them in our experimental evaluations.

Table 2.4. Continued.

SD	$distance(V_i, V_j)$	Notes
SD_3	$\frac{1}{\lvert V_i\rvert \cdot \lvert V_j\rvert} \sum_{a\in V_i, b\in V_j} CD(a,b)$	Defines the medical-record distance as the average distance of all concept pairs.
SD_4	$\frac{1}{\lvert MWBM\rvert} \sum_{(a,b)\in MWBM} CD(a,b)$	The average of all weights (i.e., code distances) in the MWBM .

Note: (1) The cardinality of the two sets of concepts, V_i and V_j , is denoted as $|V_i|$ and $|V_j|$, respectively.
(2) MWBM: minimum weighted bipartite matching. We view the two sets of concepts V_i and V_j as a bipartite undirected graph $G = (V_i, V_j)$. Then, we use the Kuhn-Munkres algorithm (Kuhn 1955) to find the minimum weighted bipartite matching (MWBM). The MWBM is a subset of edges with a minimum sum of weights and at most one edge is incident to a vertex in V_i or V_j . Specifically, in our research, given a bipartite undirected graph $G = (V_i, V_j)$ and a weight function $w = CD(a,b)$, where a, b are medical concepts, the MWBM is a group of edges representing the most similar medical-concept pairs from medical record V_i to medical record V_j .

Then, given the distance, SD, between the medical records, we are able to construct the patient-patient network G_{SD}. Specifically, after we compute the distance for each pair of medical records in the dataset, we find k neighbors with the shortest distance for each medical record. We construct the network by regarding each medical record as a node and connecting each pair of neighbors as an edge. Next, the manifold learning algorithms take the constructed network, G_{SD}, as the input to generate the low-dimensional representations that preserve the topology of the original patient-patient network. Mathematically, $Y = ML(G_{SD})$, where ML is the manifold learning algorithm, and $Y = \{Y_i\}_{i=1}^{n}$ is the low-dimensional representation. Formally, we denote a medical record as V_i, and the corresponding low dimensional representations as Y_i. We connect two vertices V_i and V_j with an edge E_{ij} if V_i is the k-nearest neighbor of V_j or vice versa. The k-nearest neighbor is determined by the distance SD. The vertices and connected edges make up the network $G_{SD}(V, E)$. Please note that different combinations of CD and SD can lead to different patient-patient networks. We compare the performance of different G_{SD}'s in our

experimental evaluation. Then manifold learning algorithms, $Y = ML(G_{SD})$, are applied to find the low-dimensional representations of V_i's. Laplacian Eigenmap minimizes the objective function, $\Phi(Y) = \sum_{i,j} ||Y_i - Y_j||$, the total distance between connected vertices (i.e. k-nearest neighbors of each other) in the low-dimensional space. Isomap solves the objective function $\Phi(Y) = \sum_{i \neq j}(d_{ij} - ||Y_i - Y_j||)^2$, where d_{ij} is the shortest distance between two records V_i and V_j in the network $G_{SD}(V, E)$. Laplacian Eigenmap and Isomap have different strategies to optimize the representations. The Laplacian Eigenmap preserves the relationships of close neighboring nodes, while the Isomap keeps the shortest distance between each pair of nodes. This explains why Laplacian Eigenmap is a local approach and Isomap is a global approach. Both Laplacian Eigenmap and Isomap have advantages and disadvantages, as discussed in Section 2.2. In the experiments, we examine and compare their efficiency for healthcare prediction tasks.

Step 3: Multimodal data fusion

Step 3 is multimodal data fusion. Besides medical concepts, there is an abundance of multimodal data in medical systems, such as demographic data, pre-ICU conditions, etc. (Lopez et al., 2020). Such data provide different perspectives to describe patients and their health conditions. Combining such information with the generated low-dimensional representations of medical concepts in Step 2 is expected to create more comprehensive vector representations of patients, which are more informative and effective for healthcare predictive modeling (Lahat et al., 2015). The data fusion step can be expressed as $Z = CONCAT(Y, U)$, where Z is the concatenated representation, Y is the representation of a set of medical concepts from Step 2, and U represents other available features. Note that both Z and Y can be used as the input of healthcare prediction tasks.

The benefits of this step are twofold. First, there are interactions between medical concepts (e.g., diagnosis or treatments) and other features (e.g., demographics). Multimodal data fusion helps prediction models describe the clinical risk more accurately than considering individual features separately (Oh, 2019). For example, the readmission and mortality risks of senior (represented in demographic features) and diabetic (represented in medical concepts' low-dimensional representation) patients are higher than senior patients without diabetes or junior patients with diabetes (Sejong Oh, 2021). Second, the healthcare prediction models can obtain higher reliability by using data from multiple sources with different modalities (Castanedo, 2013; Y. Li et al., 2020; Xu et al., 2018).

Evaluations, Results, and Discussions

In the design science paradigm, the evaluation of an artifact provides feedback information and a better understanding of the problem in order to improve both the quality of the design product and the design process (Hevner et al., 2004). Our evaluation plan and procedures are summarized in Figure 2.6. In this study, to show the effectiveness of the proposed method, we presented two research cases, readmission prediction and in-hospital mortality prediction, on two large real-world medical databases, NRD 2014[9] and MIMIC III (Johnson et al., 2016). Both databases contained de-identified hospital visit records (i.e., EHR). Please note, in the experiments, a hospital visit record was defined as a medical record that contains the medical information of the patient in one medical practice[10]. In each medical record, the diseases (i.e., medical concepts) were encoded by the ICD-9-CM system[11]. For each prediction task on each dataset, we first generated low-dimensional representations for patients' medical records using

[9] https://www.hcup-us.ahrq.gov/db/nation/nrd/nrddbdocumentation.jsp
[10] https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference
[11] https://www.cdc.gov/nchs/icd/icd9cm.htm

the proposed MD-manifold method and state-of-the-art benchmark methods. Then the generated low-dimensional representations were used in the two prediction tasks.

	Experiment part 1	Experiment part 2	Experiment part 3
Evaluation goal	Select classifiers, manifold learning algorithms, and distance metrics	Compare with benchmarks (w/o data fusion)	Compare with benchmarks (w/ data fusion); Compare the generated representation w/o data fusion and w/ data fusion
Datasets	MIMIC III – 0389, MIMIC III – 428, MIMIC III 41401, MIMIC III – 41071, NRD - 428		
Prediction tasks	30-day readmission, In-hospital mortality		
Classifier	Logistic regression Random forest AdaBoost	Logistic regression	Logistic regression
Medical concept distance	CD_{WP}, $CD_{new-Cosine}$, $CD_{new-Euclidean}$, $CD_{new-Manhattan}$, $CD_{new-eHDN}$	$CD_{new-Cosine}$, $CD_{new-eHDN}$	
Medical record distance	SD_1, SD_2, SD_3, SD_4	SD_1, SD_2, SD_3	
Manifold learning algorithms	Isomap, Laplacian Eigenmap	Isomap	Isomap
Data Fusion	No	No	Yes
Evaluation method	Cross validation		

Figure 2.6. Evaluation plan.

Datasets description

Our experiments were conducted on five datasets extracted from two large real-world databases (i.e., NRD2014 and MIMIC III). Four diseases (i.e., primary diagnosis ICD-9 code: 428, 41401, 0389, 41071) were chosen as the focus of this research. We chose heart failure (ICD-9 code 428) because it was one of the leading causes of medical institution admission in the US (Gheorghiade et al., 2013). Predicting readmission and in-hospital mortality in heart failure patients was challenging with substantial implications (O'Connor, 2017). Therefore, we extracted heart failure patients from both databases. ICD-9 codes 41401, 0389, and 41071 were the most common diagnosis codes in the MIMIC III database, which account for 7.1%, 4.2%,

and 3.6% of all hospital admissions (Johnson et al., 2016). The diseases chosen were all common and important, but their distributions in readmissions and in-hospital mortalities were widely different. Using these diseases as research cases allowed us to evaluate our findings in a more general and realistic setting. The basic statistics of each dataset were shown in Table 2.5.

Table 2.5. Datasets description.

Database		NRD2014	MIMIC III			
Primary diagnosis code (ICD-9)		428	428	41401	0389	41071
Number of medical records i.e., hospital visits		27661	1488	3498	2069	1751
Age	Range	18 - 90	18 - 89	18 - 89	18 - 89	18 - 89
	Mean	68	72	67	69	71
Gender	Male	14975	808	2654	1086	1092
	Female	12686	680	844	983	659
Ethnicity	White	--	1044	2386	1515	1219
	African American		208	99	209	61
	Hispanic		35	60	44	24
	Asian		12	49	41	13
Insurance	Medicare	19127	1150	1914	1467	1200
	Private	3188	218	1359	405	436
	Self-pay	1221	5	7	8	16
Marital status	Married	--	685	2329	887	959
	Single		295	439	527	232
	Widowed		368	376	384	348
Readmission		6710 (24.97%)	165 (12.63%)	111 (3.20%)	136 (9.61 %)	77 (4.78%)
Mortality		793 (2.87%)	182 (12.23%)	31 (8.86%)	654 (31.61%)	140 (8.00%)

Note: Primary diagnosis code (ICD-9): 428: heart failure; 41401: coronary atherosclerosis of native coronary artery; 0389: unspecified septicemia; 41071: subendocardial infarction, initial episode of care

NRD2014 We extracted the first dataset from the Healthcare Cost and Utilization Project (HCUP), the Nationwide Readmission Database (NRD). The NRD was a database specifically for national readmission analysis[12]. The NRD collected patients' admission and discharge dates for various kinds of diseases in a year. We identified patients whose primary

[12] https://www.hcup-us.ahrq.gov/db/nation/nrd/nrddbdocumentation.jsp

diagnoses were heart failure (i.e., ICD-9 code 428) in 2014. Each patient could have multiple hospital visits in the dataset. For consistency, we extracted data records from the large, private, non-profit, and teaching hospitals in one large metropolitan area (NRD STRATUM = 109).

We labeled a hospital visit as a readmission visit if the patient was readmitted within 30 days of the discharge from his/her last hospitalization. The in-hospital death labels were used as provided in the original database. The data records in December 2014 were removed due to the lack of data in the next 30 days for readmission predictions. The extracted dataset included 27,661 hospital visits of adult patients (age >= 18) with heart failure as their primary diagnosis. The incidence of in-hospital death was 2.87% (793 records). After excluding in-hospital death cases, there were 6710 (24.97%) readmission cases.

MIMIC III We evaluated our proposed method on the MIMIC-III database as well. MIMIC-III was a freely-accessible clinical database of over 40,000 patients in the Beth Israel Deaconess Medical Center's critical care units (Johnson et al., 2016). We extracted four datasets from the MIMIC-III database by setting the primary diagnosis ICD-9 codes to be 428 (heart failure), 41401 (coronary atherosclerosis of native coronary artery), 0389 (unspecified septicemia), and 41071 (subendocardial infarction, initial episode of care).

We labeled 30-day readmissions as described above. The in-hospital death labels were used as provided in the database. The four datasets included 1,488, 3,498, 2,069, and 1,751 medical records, among which there were 182 (12.23%), 31 (8.86%), 654 (31.61%), and 140 (8.00%) in-hospital death cases, respectively. Excluding in-hospital death cases, there were 165 (12.63%), 111 (3.20%), 136 (9.61%), and 77 (4.78%) readmission cases.

Healthcare prediction tasks

In the evaluation, two healthcare prediction tasks were used to test the proposed method: 30-day readmission and in-hospital mortality predictions. In NRD2014, we implemented

readmission and in-hospital mortality prediction tasks for patients with heart failure (ICD-9 code 428) as their primary diagnosis. In MIMIC III, we performed the two prediction tasks in patient cohorts with four different diseases: (1) heart failure (ICD-9 code 428), (2) coronary atherosclerosis of native coronary artery (ICD-9 code 41401), (3) unspecified septicemia (ICD-9 code 0389), and (4) subendocardial infarction, initial episode of care (ICD-9 code 41071).

Thirty-day readmissions were defined as any hospital admission within 30 days of discharge from the last hospitalization (Bardhan et al., 2015). We included the 30-day readmission prediction task in the evaluation for three reasons. (1) Hospital readmissions cost around $27 billion every year in the US (Kauffman 2016). It was a major economic burden on medical systems (Allam et al. 2019). (2) According to Leppin et al. (2014), proper post-discharge interventions could reduce hospital readmissions by improving "patient capacity to enact burdensome self-care." The effective post-discharge interventions included discharge planning, telephone follow-up, patient education, etc (Leppin et al., 2014). (3) Readmission predictions could identify subgroups of patients with a high risk of readmissions, which helped doctors provide more accurate and effective interventions (Teo et al., 2021). NRD and MIMIC III were two widely-used databases for readmission prediction studies (Allam et al., 2019; Y.-W. Lin et al., 2019; Mumtaz et al., 2019).

In-hospital mortality was defined as a death occurring within the primary admission and before discharge (Altibi et al., 2021). We incorporated in-hospital mortality prediction in the evaluation because in-hospital mortalities could be reduced using various techniques, including multidisciplinary rounds, rapid response teams, and ventilator bundles (Whittington et al., 2005). There were 23.4% of in-hospital mortalities that have opportunities for improvement (Kobewka et al., 2017). A mortality prediction model could help healthcare providers identify patients with

a high risk of mortality in the early stages, allowing them to deploy more effective measures. Many mortality prediction models were tested on NRD and MIMIC III databases (Altibi et al., 2021; Kong et al., 2020).

Benchmark methods

We considered 14 baseline methods (Table 2.6.) in two prediction tasks to evaluate our proposed method. These baseline methods could be categorized into three types. (1) Summary measures: we included three medical summary measures as benchmarks, including Charlson Comorbidity Index (CCI) (Sundararajan et al., 2004), Elixhauser Comorbidity Index (ECI) (Mehta et al., 2018), and Risk Stratification Index (RSI) (Verdecchia, 2003). CCI and ECI were two of the well-known comorbidity severity measures, commonly used to predict the mortality risk in patients with comorbidities (Mehta et al., 2018; Sundararajan et al., 2004). RSI was frequently used to predict the length of stay and mortality (Sigakis et al., 2013). (2) Code mappings: we mapped the 4 or 5-digit ICD-9 codes (i.e., medical concepts with high dimensions) in the datasets to four standard medical terminologies (with lower dimensions) as benchmarks (Deschepper et al., 2019; Rasmy et al., 2020), including 3-digit ICD-9 codes, CCS codes, CUI codes, and SNOMED. We further represented the mapped codes in each medical record (i.e., hospital visit) with one-hot encoding. This strategy reduced the dimension of medical concepts in a prediction model's feature space. We used official resources for code mappings. Specifically, the 4 or 5-digit ICD-9 codes were mapped to 3-digit ICD-9 codes according to the latest ICD-9 hierarchy. We used the latest version of ICD-9 to CCS single-level mapping provided by HCUP. The mappings from ICD-9 codes to SNOMED[13] and CUI[14] were available in the UMLS. (3) Embeddings of individual medical concepts: we adopted three state-of-the-art healthcare

[13] https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html
[14] https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

prediction studies as benchmarks. The first method represented medical records using the element-wise sum of individual medical concepts' embeddings generated from Word2Vec, GloVe, and FastText (Tang et al., 2018; Youngduck Choi et al., 2016). The second method was Phe2Vec, a state-of-the-art unsupervised embedding framework for disease phenotyping. Phe2Vec was based on pre-computing embeddings of medical concepts generated using Word2Vec, GloVe, and FastText (De Freitas et al., 2020). In Phe2Vec, medical records were represented as a weighted sum of embeddings of individual medical concepts, with the weights determined by the frequency of the medical concepts in the dataset. The last was Med2Vec, a supervised deep-learning method that used a two-layer neural network to generate representations for medical concepts and records (Choi, Bahadori, Searles, et al., 2016).

Table 2.6. Benchmarks in the evaluation.

	Benchmarks		Description
Summary measures	CCI		Charlson Comorbidity Index
	ECI		Elixhauser Comorbidity Index
	RSI		Risk Stratification Index
Code mappings	ICD-9 3-digit code		Map ICD-9 4 or 5-digit codes to ICD-9 3-digit codes
	CCS		Map ICD-9 4 or 5-digit codes to CCS codes
	CUI		Map ICD-9 4 or 5-digit codes to CUI codes
	SNOMED		Map ICD-9 4 or 5-digit codes to SNOMED codes
Embeddings of individual medical concepts	Sum	Word2Vec	Element-wise sum of individual medical concepts' embeddings generated from Word2Vec
		GloVe	Element-wise sum of individual medical concepts' embeddings generated from GloVe
		FastText	Element-wise sum of individual medical concepts' embeddings generated from FastText
	Phe2vec	Word2Vec	Representations of Phe2vec with Word2Vec as the individual medical concept' embedding
		GloVe	Representations of Phe2vec with GloVe as the individual medical concept' embedding
		FastText	Representations of Phe2vec with FastText as the individual medical concept' embedding
	Med2Vec		Supervised two-layer neural network

Experimental settings

In the experiments, we generated representations with the dimensions of 16, 32, 64, 128, 256, and 512 for each medical record. The generated representations were used as the input of

three classifiers, including logistic regression (LR), random forest (RF), and AdaBoost, to predict the 30-day hospital readmissions and in-hospital mortalities. The parameters of the classifiers were listed in Table 2.7. We used grid-search to find the best parameters for the classifiers using the MIMIC III - 428 dataset. These classifiers were implemented for all representations generated using the proposed method and baseline methods except the three summary measures (i.e., CCI, ECI, and RSI, which already indicated prediction probabilities) and Med2Vec (which already contained a classifier in their research design). We evaluated two manifold learning algorithms, Laplacian Eigenmap (i.e., a local approach) and Isomap (i.e., a global approach), both of which were widely used in various applications (Huang et al., 2019; Park, 2012; Tu et al., 2012). We grid-searched n_neighbors, a parameter in the patient-patient network for the manifold learning algorithms.

We evaluated all classifiers' performance through five-fold cross-validation, where the original dataset was randomly split into five equal-sized sub-samples without replacement. The process was repeated in five rounds (i.e., folds). In each round, one single sub-sample was retained as the testing set, and the other four sub-samples were used for classifier training. The classifiers were trained from only the training data of the current round, and the testing data were not seen by the model during the training stage. Please note that the cross-validation was not employed for selecting optimal parameters. All classifiers for different datasets used the same parameters selected using the MIMIC III - 428 dataset. The main reasons for adopting this validation technique were that it achieved a lower bias towards estimating the generalization performance by averaging the individual classifier's estimates (Hastie et al., 2009) and it estimated how the model's performance can be generalized to an independent dataset.

Table 2.7. Algorithms and parameters.

Algorithm		Parameters
Classifier	Logistic Regression (LR)	L1 penalty: [0.1, **1**, 10]
	Random Forest (RF)	Max_depth: [1, **3**, 5, 7] n_estimators: [50, **100**, 200] criterion: [**gini**]
	AdaBoost	Max_depth: [**1**, 3, 5, 7] n_estimators: [**50**, 100, 200]
Manifold learning	Laplacian Eigenmap	n_neighbors: [8, 16, 32, 64, 128, 256, 512]
	IsoMap	

Note: Optimal parameters were bold. The optimal parameters for the classifiers were determined through grid search on the dataset of MIMIC III - 428

Experimental results

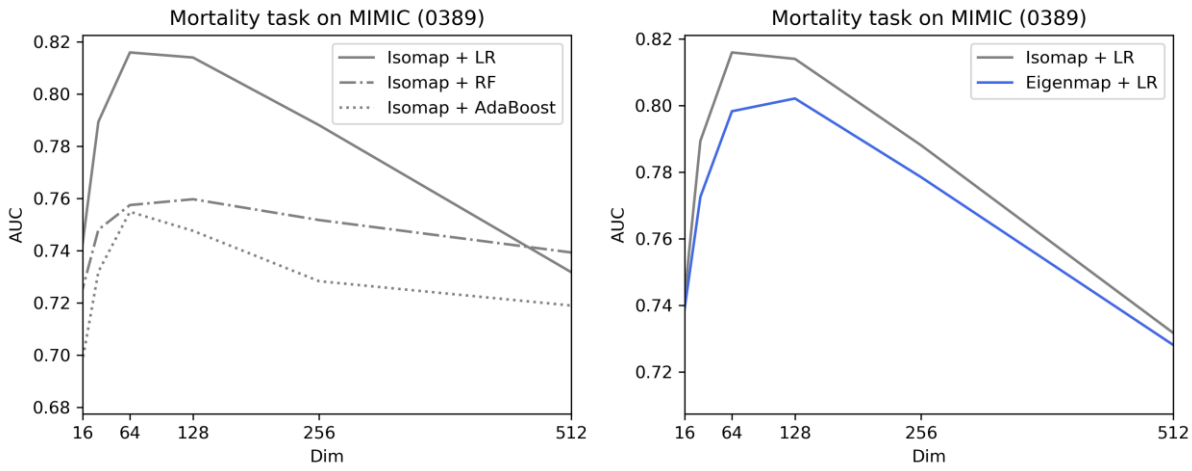
In the following section, we reported the performance of the proposed method and the benchmarks in readmission and in-hospital mortality predictions. We first showed that the proposed medical distance metrics CD_{new} led to more accurate predictions than the CD_{WP}. Then we showed that the proposed method was more effective than three types of medical concepts' dimension reduction methods, including summary measures, code mapping, and individual medical concepts' embeddings. Finally, we concatenated the low-dimensional representations of sets of medical concepts and other patients' data, demonstrating the applicability and compatibility of our method.

Part 1: The selection of classifiers, manifold learning algorithms, and distance metrics

We first selected the classifiers used in the two healthcare prediction tasks. According to our evaluations, LR outperformed both RF and AdaBoost in all the experiments. Please see Table 2.12. in Appendix B for experimental results on all five datasets. For example, Figure 2.7. (left) showed the performance of LR, RF, and AdaBoost in predicting mortality for patients with subendocardial infarction (ICD-9: 0389) in the MIMIC III database. The x-axis was the dimension of the representations generated using our method, while the y-axis was the mean of

the AUC scores in 5-fold cross-validation. The solid line (LR) was always above the other two lines at different dimensions, which showed using the same representations, LR made more accurate predictions with higher AUC scores. For conciseness, we only presented the prediction performance of LR in the following discussion.

We also compared the global and local manifold learning algorithms (i.e., Isomap and Laplacian Eigenmap) in our experiments. Generally, the Isomap had a better performance. For example, in the MIMIC III - 0389 dataset (Figure 2.7.), the AUC scores of the representations from Isomap were always higher than that of Laplacian Eigenmap. In other experiments (see Table 2.13. in Appendix B), the highest AUC scores of Isomap at different dimensions were usually higher than that of Laplacian Eigenmap. The possible reason was that Isomap was more robust to noise than Laplacian Eigenmap; similar findings were also reported by Mysling et al. (2011) and Talwalkar et al. (2008). We only report Isomap's prediction results in the following sections for brevity.



Note: Prediction task: in-hospital mortality; Dataset: MIMIC III - 0389 dataset; Distance metrics: CD_{eHDN} and SD_3; Patient-patient network: n_neighbors=256; Medical domain knowledge: $T_{ICD\text{-}9}$

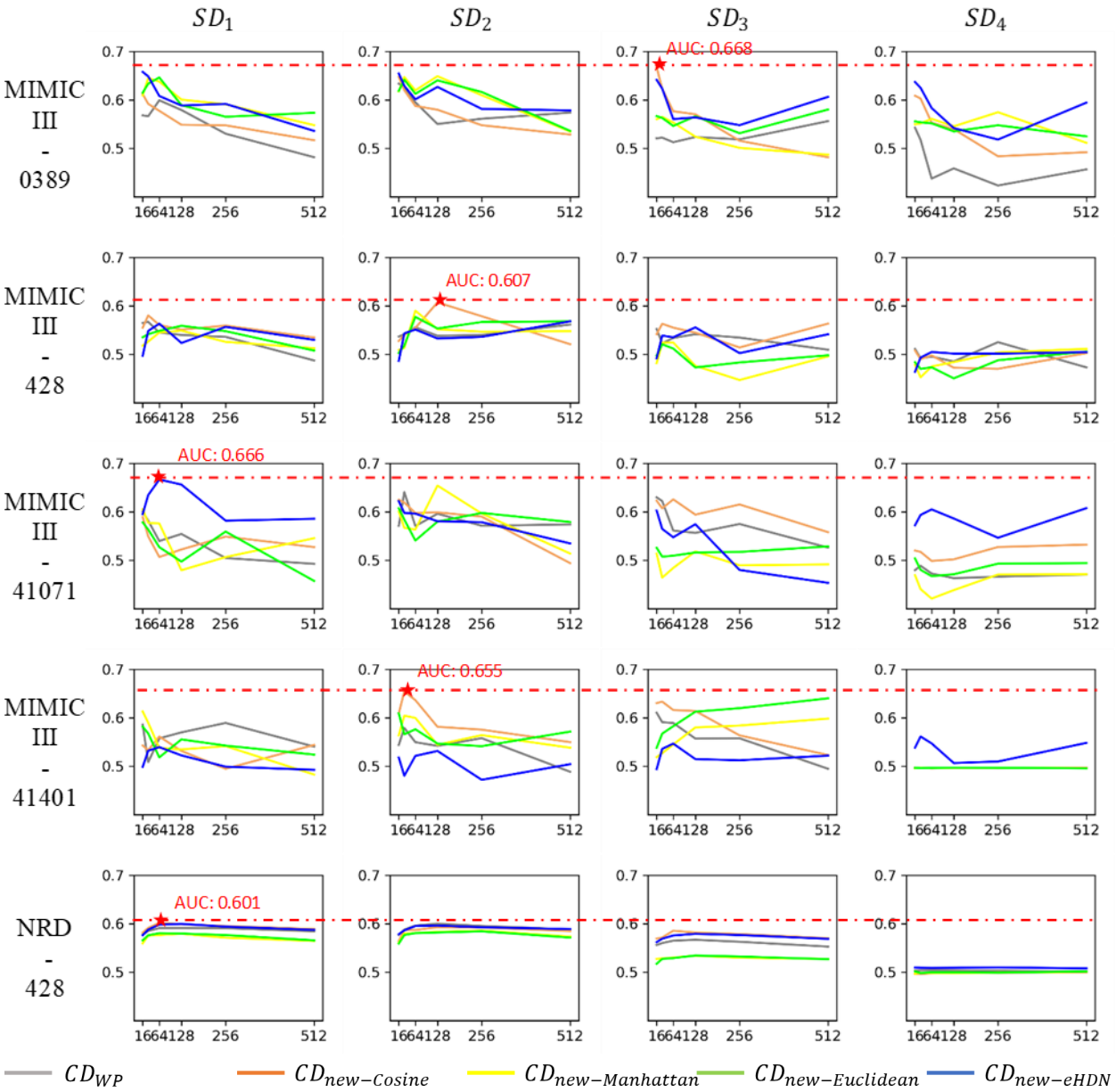
Figure 2.7. Prediction results using Isomap with three classifiers (left), and prediction results using two manifold algorithms with the LR classifier (right).

Next, we evaluated the performance of different patient-patient networks, G, generated using different combinations of medical concept-distance metrics (i.e., benchmark metric CD_{WP} and our proposed metrics CD_{new}, including $CD_{new-Cosine}$, $CD_{new-Manhattan}$, $CD_{new-Euclidean}$, and $CD_{new-eHDN}$) and medical record-distance metrics (i.e., four widely used metrics SD_1, SD_2, SD_3, and SD_4). n_neighbors was a hyperparameter of G, determined through grid search. All the results were reported in Figure 2.8.

The experimental results revealed several interesting findings.

(1) The red dotted lines in Figure 2.8. represented the best performance on each dataset. On all five datasets, the performance of CD_{WP} (gray lines) never achieved the top AUC scores in both prediction tasks. Therefore, our metrics CD_{new} were more effective at measuring the distances between medical concepts, resulting in higher AUC scores in healthcare prediction tasks.

(2) Figure 2.8. showed that most representations (80% of the results) reached their highest AUC scores at low dimensions (dimensions = 16 - 64). The possible reason was that manifold learning algorithms were good at representing high-dimensional data in extremely low dimensions (J. Zhang et al., 2010). In many scenarios, manifold learning algorithms were used for visualizations that required two or three dimensions (Patwari et al., 2005; Shier et al., 2021). Higher dimensions may introduce more noise in the prediction model's feature space. Since prediction models were known to converge fast on a low-dimensional feature space, this was an advantage of the proposed method.

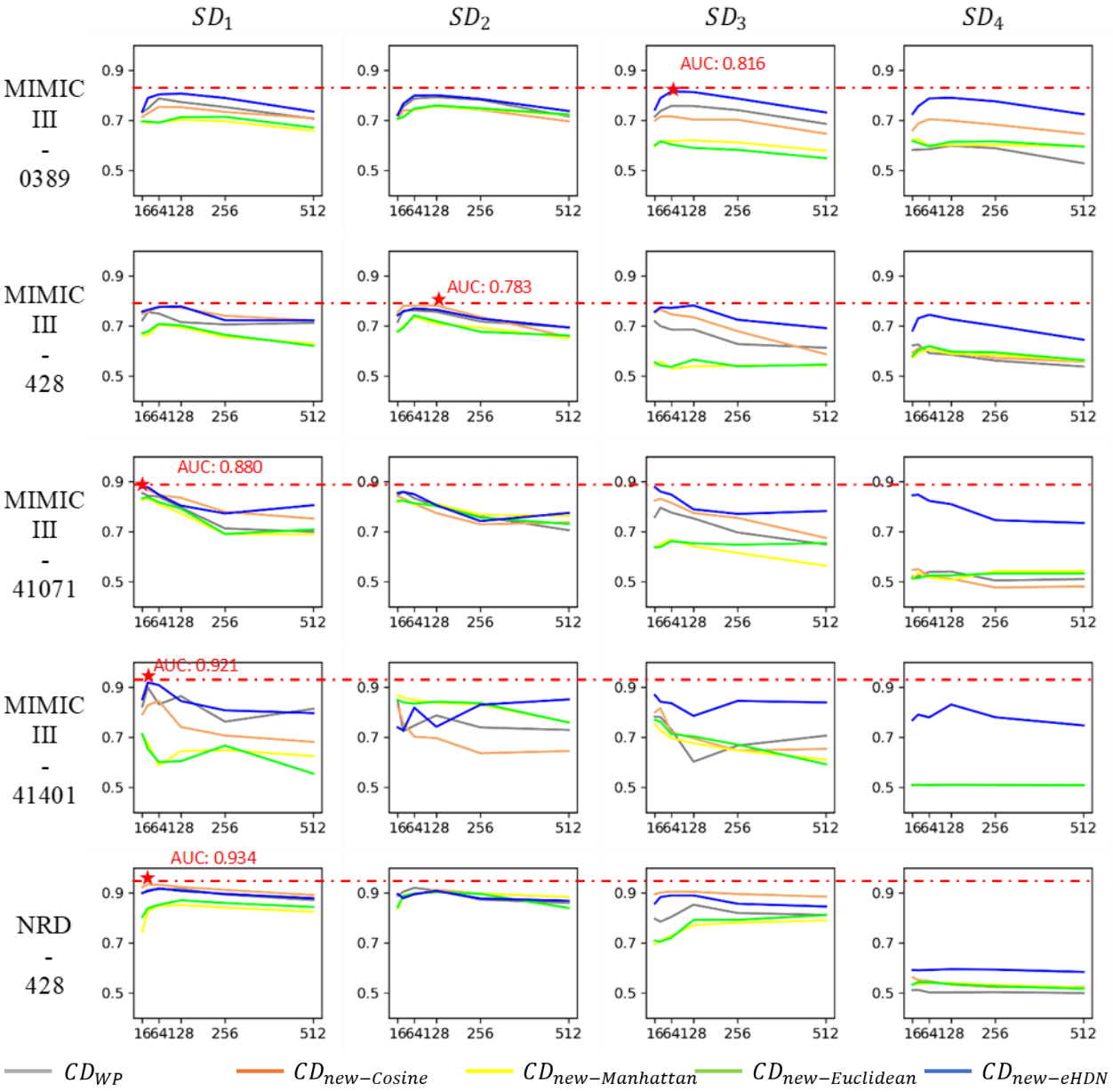


(a) Readmission prediction

Figure 2.8. Performance of different patient-patient networks.

(3) We proposed a new medical-concept distance metric CD_{new} with four distance formulas: $Cosine$, $Manhattan$, $Euclidean$, and $eHDN$. In our evaluation, $CD_{new-Cosine}$ and $CD_{new-eHDN}$ outperformed the $CD_{new-Euclidean}$ and $CD_{new-Manhattan}$. $CD_{new-Euclidean}$ and $CD_{new-Manhattan}$ never achieved the highest AUC scores on all five datasets and two prediction tasks. Especially, combined with SD_1, $CD_{new-eHDN}$ achieved the highest AUC score (0.666) in

the readmission prediction on the MIMIC III - 41071 dataset (Figure 2.9. (a)). In addition, when paired with SD_2, $CD_{new-Cosine}$ was the best CD metric (AUC = 0.783) for the in-hospital mortality prediction on the MIMIC III - 428 dataset (Figure 2.9. (b)).



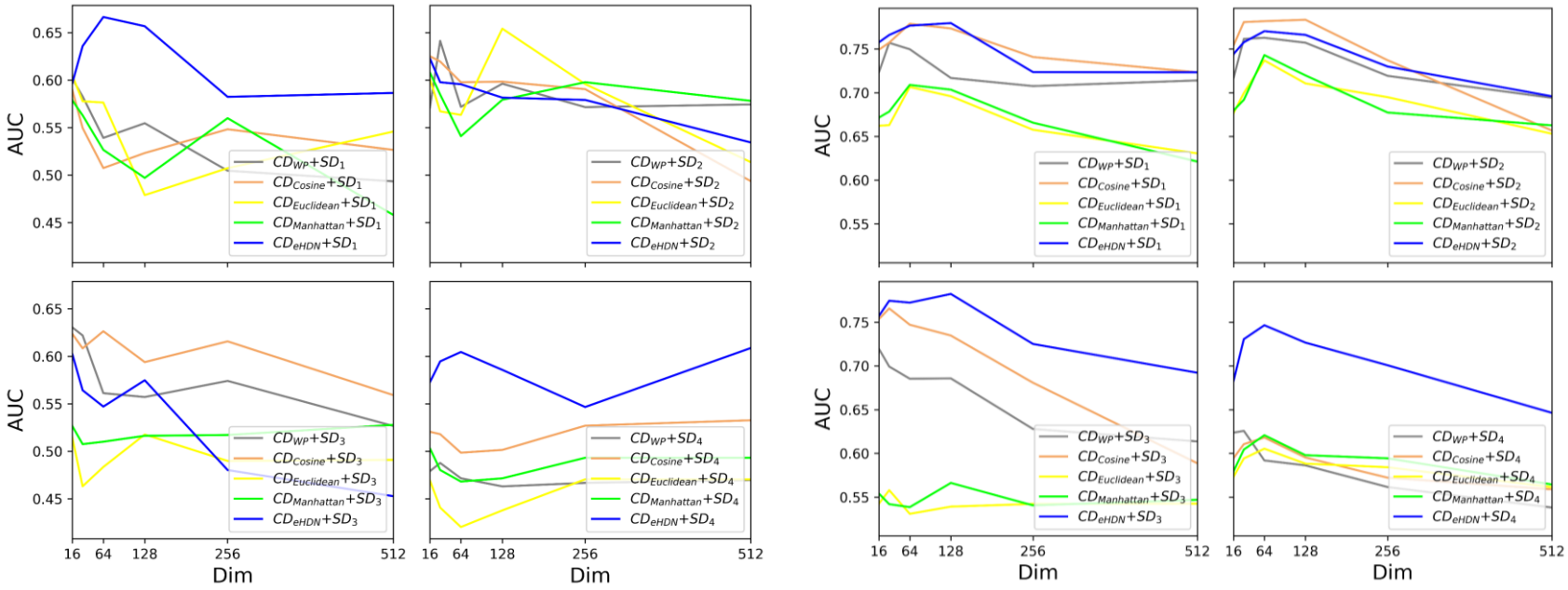
(b) Mortality prediction

Figure 2.8. Continued.

This was an interesting finding because it indicated that it was important to normalize the co-occurrences of medical concepts for medical-concept distance calculation when considering disease co-occurrences as medical domain knowledge. Both $CD_{new-Cosine}$ and $CD_{new-eHDN}$ included normalization terms in the distance formulas (i.e., $\sqrt{C_a \cdot C_a}\sqrt{C_b \cdot C_b}$ and $\sqrt{\sum C_a \sum C_b (N - \sum C_a)(N - \sum C_b)}$). The significance of the normalization terms was that they eliminated the impact of very popular diseases across all patient cohorts.

For example, a very popular medical concept M_j co-occurred with most other medical concepts. Therefore, most of the elements in row j in the co-occurrence matrix C were large values. By contrast, there were two rare medical concepts M_a and M_b. The elements in both rows a and b were small numbers in the co-occurrence matrix C. If M_a and M_b co-occurred frequently, we expected the medical-concept distance to reflect such a co-occurring relationship. However, $CD_{new-Euclidean}$ and $CD_{new-Manhattan}$ might fail to capture such a pattern, leading to an undesired performance in healthcare prediction tasks. This finding echoed other studies which showed the significance of co-occurrence normalization (Heidary Moghadam et al., 2019; Kumar et al., 2015).

(4) We also looked at the performance of four distance metrics, $SD = distance(V_i, V_j)$, for measuring the distances between medical records. SD_1 showed better performance compared to other medical record distance metrics; in the ten combinations of two prediction tasks and five datasets, SD_1 achieved the best AUC scores 50% of the time. The possible explanation could be that SD_1 was designed to capture the similarities of the most similar medical-concept pairs from two medical records, which were essential features for the two healthcare prediction tasks.



(a) Readmission prediction (MIMIC III - 41071) (b) In-hospital mortality prediction (MIMIC III - 428)

Figure 2.9. Prediction performance using different combinations of distance metrics

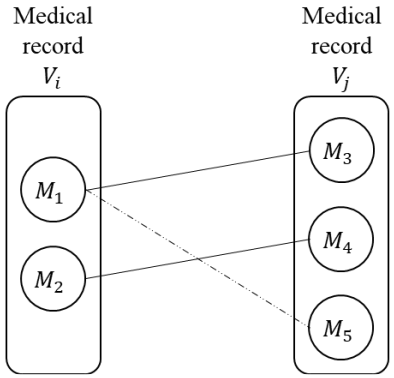
An interesting finding was that SD_4 was also developed to compare the most similar medical-concept pairs from two medical records. However, SD_4 performed poorly in the prediction tasks on all five datasets. As shown in Figure 2.8., SD_4 never achieved the best AUC scores. This result differed from the finding of Jia et al. (2019). The difference between SD_1 and SD_4 lay in how they defined the most similar medical-concept pairs (see Figure 2.10.). In SD_1, every medical concept could be paired with another medical concept. Such a pair formed a set of "most similar pairs" for medical-record distance calculation. However, in SD_4, it was possible that a medical concept could not be paired with other medical concepts. Hence, SD_4 excluded such a medical concept from medical-record distance calculation, which compromised the accuracy of prediction models. The experimental results suggested that every medical concept contained important information like disease diagnosis and was important for healthcare prediction tasks.

To summarize, we developed a new medical concept-distance metric CD_{new} that was both knowledge-driven and data-driven to preserve medical domain knowledge in medical concepts' properties, i.e., the hierarchical structure and co-occurrences. Extant metric CD_{WP} did not consider the co-occurrences of medical concepts, hence was outperformed by our metric CD_{new}. Since CD_{new} took medical concepts' co-occurrences into consideration, it was important to use distance formulas with normalization terms that normalize the co-occurrences of medical concepts.

Part 2: Compare the prediction performance with benchmarks

We compared the proposed method with the state-of-the-art baseline methods (three types, 14 in total, Table 2.6.) and reported the major findings below.

(1) In the two prediction tasks, our method exceeded all baselines in 70% of cases (top-1 AUC scores) on five datasets; and reached 100% of state-of-the-art performance (top-3 AUC scores). As shown in Tables 2.8. - 2.9., we highlighted the top three AUC scores in bold in each prediction task and on each dataset, with the highest AUC score bolded and underlined. The superior performance in the experimental results demonstrated the effectiveness of our method.



Most similar pairs in SD_1 (solid and dash lines): $< M_1, M_3 >, < M_2, M_4 >, < M_1, M_5 >$.
Most similar pairs in SD_4 (solid lines): $< M_1, M_3 >, < M_2, M_4 >$.

Figure 2.10. Example of most similar medical-concept pairs in SD_1 and SD_4.

(2) We conducted two healthcare prediction tasks: readmission and in-hospital mortality predictions. According to our findings, the overall performance of all models (i.e., the proposed and baseline methods) was better in mortality prediction (average AUC = 0.753) than in readmission prediction (average AUC = 0.552). The experimental results suggested that readmission prediction was a more challenging task on the five datasets we adopted.

We noticed that Med2Vec, as a representative state-of-the-art for individual medical concept embedding techniques, performed well in mortality prediction (average AUC=0.765, 1.2% better than the overall average). Especially on the datasets MIMIC III - 0389 and MIMIC III - 428, Med2Vec's AUC scores for mortality prediction were 0.818 and 0.809 respectively, the highest among all methods. However, Med2Vec performed poorly in readmission prediction (average AUC=0.487, 6.5% below the overall average) and never achieved the best AUC on any

of the five datasets. The explanation for this could be that readmission prediction was a more difficult task, and the Med2Vec model (i.e., a deep learning model) was under training. This finding echoed the observation by E. Choi et al. (2018) that many healthcare prediction tasks benefited from deep learning models, however, these models often required a large amount of training data, which was beyond the capacity of most healthcare systems.

By contrast, our method exhibited stable and superior performance in both prediction tasks. In the readmission prediction task, our method had an average AUC of 0.593, which was 4.1% higher than the overall average (average AUC_T_{ICD9} = 0.588, 3.6% better than the overall average; average AUC_T_{CUI} = 0.589, 3.7% better than the overall average; average AUC_T_{CCS} = 0.601, 4.9% better than the overall average). In the in-hospital mortality prediction task, our method had an average AUC of 0.823, which was 7.0% higher than the overall average (average AUC_T_{ICD9} = 0.826, 7.3% better than the overall average; average AUC_T_{CUI} = 0.819, 6.6% better than the overall average; average AUC_T_{CCS} = 0.825, 7.2% better than the overall average).

The possible reason was that most of these embedding methods are adapted from NLP-related deep learning techniques. The embedding for the individual medical concept was trained discriminatively, i.e., the model learned a conditional distribution of outputs given inputs. For example, in word2vec, the model predicted a medical concept given other medical concepts in the same medical record (i.e., CBOW) or predicted other medical concepts in the same medical record given a medical concept (i.e., skip-gram). In contrast, manifold learning algorithms acted as generative models which strived to depict the actual distribution of the data. We would expect algorithms that acted like generative models to do better with less training data, but for methods

that acted like discriminative models to catch up with sufficient training data[15] (Ng & Jordan, 2001). Since a large amount of training data was not always available for healthcare prediction tasks, using manifold learning algorithms was an advantage of the proposed method.

(3) We included three types of benchmarks, i.e., summary measures, code mapping, and embeddings of individual medical concepts, among which embeddings of individual medical concepts performed the best.

Summary measures were one of the traditional strategies used to address medical concepts' high dimensionality issues. Some of the summary measures worked extremely well on the in-hospital prediction task. For example, RSI's AUC score on the MIMIC III - 41401 dataset was 0.882, 12.9% higher than the overall average. However, its performance on readmission prediction was not impressive. This was because RSI was designed to predict in-hospital mortality rather than readmission (Mehta et al., 2018; Sundararajan et al., 2004; Verdecchia, 2003). The findings corroborated previous research in that one of the disadvantages of summary measures was that they were difficult to adapt for prediction tasks other than the designed purpose.

Code mapping strategies were frequently employed by healthcare predictive analytics researchers. The performance, on the other hand, was not particularly outstanding. In our tests, such methods never received the highest AUC score. The findings contradicted some previous research that found code mapping to be beneficial (Choi, Bahadori, Kulas, et al., 2016; Min et al., 2019). Our findings, however, backed up Xiang et al., (2019), Jung et al. (2019), and Rasmy

[15] To clarify, generative and discriminative models are two types of statistical classification models. Both our proposed method and baseline methods were not developed for classifications. We borrow the concepts of "generative" and "discriminative" to describe the different learning processes of deep-learning-based and manifold-learning-based algorithms.

et al. (2020)'s conclusions that code translations were not always useful in healthcare prediction tasks.

(4) Our evaluation included three types of medical concept hierarchies, T_{ICD9}, T_{CCS}, and T_{CUI}. We found that the medical concept hierarchy that could help the MD-Manifold achieve the best performance in different cases was not the same. For example, in the readmission task on MIMIC III - 41071, T_{ICD9} helped the proposed method achieve the highest AUC score (0.666). Similarly, using T_{CCS} as domain knowledge, our method's AUC score (0.678) is higher than all other methods in the readmission prediction for septicemia patients (MIMIC III - 0389). Also, T_{CUI} helped the proposed method reach the highest AUC score (0.887) to predict in-hospital mortality for the subendocardial infarction patients (MIMIC III - 41071). Our results suggested that using proper medical domain knowledge could provide useful information for medical concept distance calculation, which was consistent with Melton et al. (2006)'s finding.

Part 3: Experiments of multimodal data fusion

In this subsection, we reported the experimental results of multimodal data fusion. For multimodal data, we concatenated the demographic features with medical records' low-dimensional representations generated from the proposed method. Then, we evaluated the concatenated representations using the two prediction tasks on the five datasets. For the MIMIC III datasets, the demographics included age, gender, insurance, religion, marital status, and ethnicity. For the NRD2014 dataset, the features included age, gender, and insurance (no marital status and ethnicity information).

Table 2.8. Performance of MD-Manifold and baselines in the readmission prediction task.

Database			MIMIC III						MIMIC III						MIMIC III						MIMIC III						NRD					
Disease (ICD-9 code)			0389						428						41071						41401						428					
Dimensions			16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512
Summary measures	CCI		53.0						57.8						63.1						62.5						55.7					
	ECI		56.3						56.3						64.7						62.0						54.4					
	RSI		54.0						49.6						63.1						61.9						54.1					
Code mapping methods	ICD9-3digit		49.1						53.3						55.7						50.8						58.6					
	CCS		49.6						57.8						56.2						50.9						57.8					
	CUI		56.4						54.5						57.6						47.3						58.4					
	SNOMED		55.3						54.8						52.1						47.0						58.0					
Individual embeddings	Sum	GloVe	52.7	54.2	57.9	61.0	58.3	56.8	52.8	57.6	50.1	52.9	51.7	52.0	57.9	59.9	58.1	61.2	58.3	62.1	57.2	59.1	57.4	58.9	60.9	60.1	57.6	58.8	59.8	60.1	**60.2**	60.0
		Word2Vec	60.6	60.4	59.5	55.6	56.7	55.9	54.8	55.1	53.4	56.9	53.7	51.2	56.2	57.3	55.1	53.8	51.7	50.2	42.5	51.0	48.5	45.7	41.0	38.4	53.8	54.9	56.1	57.8	57.1	55.2
		FastText	59.3	59.7	59.2	57.8	57.0	53.8	55.9	55.3	55.6	54.2	54.9	56.8	62.3	63.6	63.3	58.7	61.0	58.6	64.7	63.9	65.6	**66.7**	65.3	62.3	57.3	58.4	58.8	59.2	59.0	58.8
	Phe2Vec	GloVe	47.4	52.4	47.1	52.5	52.7	48.4	54.7	55.9	49.5	51.1	47.4	49.2	55.4	42.4	49.9	45.1	47.6	51.5	58.3	53.2	54.7	50.1	56.5	51.5	53.0	52.5	52.6	50.6	50.9	50.6
		Word2Vec	52.2	56.1	55.3	55.1	55.5	50.9	55.6	54.3	53.1	52.1	48.5	50.5	50.0	49.4	49.0	51.5	49.2	46.0	53.5	44.4	48.4	45.3	47.1	48.8	50.0	50.0	50.0	50.0	50.0	50.0
		FastText	49.4	53.0	45.6	47.9	51.0	50.6	54.1	56.1	51.8	54.0	50.7	53.6	48.5	47.6	52.0	49.7	43.1	41.9	65.0	58.7	60.4	61.0	60.0	60.9	54.5	54.4	54.4	54.4	54.4	54.3
	Med2Vec		45.3	45.1	44.5	50.7	50.8	45.3	45.3	44.7	49.9	47.6	45.9	53.3	39.3	41.7	41.6	48.2	47.0	49.2	45.5	51.5	50.3	45.4	49.1	49.5	55.6	56.0	55.5	55.0	55.9	55.8
MD-Manifold	T_{ICD9} as domain knowledge		**66.8**	62.3	57.6	57.1	51.5	48.2	52.8	53.9	55.4	**60.7**	57.8	52.2	59.7	63.6	**66.6**	65.7	58.2	58.6	60.9	**65.5**	63.7	58.2	57.6	55.0	58.2	59.2	60.1	59.9	59.4	58.9
	T_{CCS} as domain knowledge		61.7	63.5	**67.8**	62.2	55.3	55.0	53.8	58.0	58.1	**62.0**	58.5	54.1	60.4	60.1	**66.5**	64.7	61.0	54.5	64.5	**65.3**	63.9	62.1	55.7	56.7	58.2	58.7	59.8	**60.2**	59.9	59.6
	T_{CUI} as domain knowledge		**66.6**	64.2	54.8	56.1	47.4	46.1	52.7	56.9	60.5	**62.5**	55.6	49.0	60.5	60.3	**66.5**	64.8	61.4	55.5	62.0	65.0	65.1	62.8	59.2	54.5	58.3	58.5	60.0	**60.3**	60.0	59.5

Note: (1) Performance: AUC (%)

(2) The top-3 AUC scores in each dataset were marked in bold, among which the highest AUC scores were underlined.

(3) The dimensions of summary measures were all 1. The dimensions of code mappings varied across different methods and datasets. The dimensions of ICD9-3 digit (1018 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 632, 503, 485, 499, and 953, respectively. The dimensions of CCS (367 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 340, 311, 305, 325, and 256, respectively. The dimensions of CUI (16150 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 2035, 1448, 1363, 1467, and 3768, respectively. The dimensions of SNOMED (7561 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 1277, 908, 838, 916, and 2314, respectively.

Table 2.9. Performance of MD-Manifold and baselines in the mortality prediction task.

Database			MIMIC III						MIMIC III						MIMIC III						MIMIC III						NRD					
Disease (ICD-9 code)			0389						428						41071						41401						428					
Dimensions			16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512
Summary measures	CCI		58.6						53.8						62.2						77.6						57.5					
	ECI		54.0						50.1						62.0						72.2						62.8					
	RSI		67.5						72.8						82.9						88.2						88.8					
Code mapping methods	ICD9-3digit		78.7						78.4						81.5						83.4						86.4					
	CCS		79.0						76.2						86.4						66.0						85.2					
	CUI		81.4						79.9						84.0						82.8						86.6					
	SNOMED		78.3						77.6						83.8						75.8						82.7					
Individual embeddings	Sum	GloVe	69.2	71.9	75.2	75.1	76.4	75.4	69.5	72.4	73.8	73.6	74.0	74.3	85.6	83.1	84.5	85.2	84.9	86.4	83.0	90.2	**92.6**	91.3	88.8	83.0	92.5	93.5	92.0	**93.6**	81.6	**93.9**
		Word2Vec	66.8	70.5	72.0	74.1	74.2	74.9	52.9	67.5	74.2	71.8	73.9	73.3	70.1	74.5	79.1	77.5	76.6	75.5	69.7	77.5	79.8	77.2	80.9	83.2	62.4	69.4	78.8	76.2	91.1	71.8
		FastText	66.6	71.6	69.1	70.3	71.7	71.9	70.4	70.7	71.0	71.0	69.0	71.9	81.2	81.0	82.8	82.2	84.5	83.4	90.3	83.8	89.6	90.3	90.3	89.3	88.8	91.5	93.1	93.0	90.6	**93.4**
	Phe2Vec	GloVe	57.9	58.6	60.2	63.3	64.3	63.4	62.0	57.0	64.0	61.6	63.3	60.4	71.5	72.8	72.2	72.6	70.6	74.7	71.5	75.9	60.1	66.2	60.9	70.3	72.4	73.5	78.4	63.5	77.3	75.7
		Word2Vec	51.3	52.9	50.9	50.4	51.4	54.1	54.4	51.6	56.1	59.8	57.9	59.5	56.1	58.3	60.2	59.0	59.3	62.1	58.0	67.7	53.8	57.7	61.8	68.9	50.3	50.2	50.2	50.0	50.1	50.1
		FastText	60.2	62.4	60.4	60.9	61.6	62.0	59.8	60.7	59.4	59.6	58.6	60.3	66.9	68.6	70.3	70.0	68.7	72.6	81.2	81.1	75.4	74.0	75.3	69.6	79.3	79.7	80.2	79.9	80.0	79.7
	Med2Vec		**81.8**	81.7	79.7	79.8	78.6	78.8	63.3	72.1	71.6	**80.9**	74.6	79.7	65.3	73.6	82.0	79.8	76.6	77.5	61.5	55.4	72.6	73.9	74.6	81.2	82.7	81.1	87.0	83.7	81.8	82.7
MD-Manifold	T_{ICD9} as domain knowledge		74.3	78.9	81.6	81.4	78.8	73.2	75.4	78.1	78.2	78.3	73.7	65.7	**88.0**	86.8	84.7	83.6	78.1	75.2	85.3	**92.1**	**91.0**	84.6	80.9	79.7	92.3	**93.4**	93.2	91.2	92.5	89.2
	T_{CCS} as domain knowledge		74.4	78.8	**81.8**	81.5	78.4	73.7	76.5	76.5	78.0	**79.1**	74.3	71.3	**87.9**	86.6	85.9	82.7	78.8	74.9	82.0	90.3	**91.0**	86.6	79.3	78.0	92.1	92.1	91.5	92.2	90.3	88.2
	T_{CUI} as domain knowledge		74.4	78.8	**81.8**	81.4	78.5	73.7	77.1	77.0	78.4	**79.1**	75.9	69.8	**88.7**	86.9	84.7	81.5	74.1	70.6	73.9	90.3	**91.0**	86.6	79.3	78.0	91.3	92.1	92.1	90.9	90.4	89.3

Note: (1) Performance: AUC (%)
(2) The top-3 AUC scores in each dataset were marked in bold, among which the highest AUC scores were underlined.
(3) The dimensions of summary measures were all 1. The dimensions of code mappings varied across different methods and datasets. The dimensions of ICD9-3 digit (1018 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 632, 503, 485, 499, and 953, respectively. The dimensions of CCS (367 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 340, 311, 305, 325, and 256, respectively. The dimensions of CUI (16150 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 2035, 1448, 1363, 1467, and 3768, respectively. The dimensions of SNOMED (7561 in total) in MIMIC III - 0389, 428, 41071, 41401, and NRD - 428 were 1277, 908, 838, 916, and 2314, respectively.

Table 2.10. Performance of MD-Manifold and baselines with multimodal data fusion in the readmission prediction task.

Database			MIMIC III						MIMIC III						MIMIC III						MIMIC III						NRD					
Disease (ICD-9 code)			0389						428						41071						41401						428					
Dimensions			16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512
Summary measures	CCI		55.6						64.4						55.8						58.0						58.2					
	ECI		57.1						63.6						55.9						57.1						57.5					
	RSI		57.4						63.1						56.0						55.3						57.8					
Code mapping methods	ICD9-3digit		52.1						59.7						52.7						53.5						59.8					
	CCS		54.2						62.6						51.1						52.7						59.3					
	CUI		57.8						60.2						56.4						51.6						59.4					
	SNOMED		57.1						61.1						54.0						49.9						59.3					
Individual embeddings	Sum	GloVe	57.9	58.3	59.6	62.1	60.7	60.5	61.2	61.5	59.8	60.1	59.3	59.2	58.7	58.5	59.1	58.7	58.0	59.9	58.6	60.5	59.1	60.3	60.2	60.2	57.5	57.5	57.4	57.4	57.4	57.4
		Word2Vec	61.6	**62.7**	61.8	59.7	58.9	58.8	62.9	61.1	59.1	60.0	56.1	55.2	54.7	57.7	56.7	50.2	54.6	54.5	55.0	56.1	54.3	47.5	44.5	42.4	57.4	57.4	57.4	57.4	57.4	57.4
		FastText	58.5	59.5	58.7	57.8	59.1	57.1	61.9	61.7	61.3	61.3	61.7	62.0	57.4	60.5	60.1	56.8	59.6	57.0	62.8	62.0	63.1	**64.8**	63.4	60.8	57.6	57.7	57.6	57.6	57.7	57.6
	Phe2Vec	GloVe	55.7	56.6	53.3	54.9	54.7	54.5	**63.2**	62.3	58.9	59.5	54.0	55.4	56.0	48.0	50.9	46.6	51.3	53.3	61.1	56.8	56.3	54.9	57.2	55.3	59.5	60.1	60.8	61.0	61.0	61.0
		Word2Vec	53.7	56.0	55.7	57.0	56.9	53.7	61.9	60.5	59.8	57.6	52.9	54.0	50.1	48.7	50.5	49.3	47.3	49.9	55.5	50.3	54.6	50.1	50.7	48.8	58.0	58.5	58.7	58.7	57.9	57.4
		FastText	55.1	55.4	54.9	54.7	55.1	55.1	62.1	61.8	61.3	61.0	61.7	61.4	52.6	52.6	53.7	53.6	51.6	50.1	62.1	61.4	60.1	60.4	61.3	61.6	59.2	59.9	60.1	60.4	60.2	60.2
	Med2Vec		52.2	43.5	44.2	55.1	50.9	47.9	56.5	56.0	52.6	50.4	52.6	50.7	54.8	45.4	49.8	54.3	46.3	43.7	51.4	50.9	51.9	45.1	43.8	43.9	56.6	56.4	56.3	56.1	56.1	55.7
MD-Manifold	T_{ICD9} as domain knowledge		**<u>63.6</u>**	62.1	58.2	57.9	53.4	48.0	60.8	60.8	60.8	62.5	59.1	53.7	62.5	63.8	**66.8**	64.4	61.0	58.5	60.6	63.8	63.8	59.2	58.8	53.9	60.0	60.6	**<u>61.3</u>**	61.1	60.7	60.0
	T_{CCS} as domain knowledge		62.1	62.6	65.0	61.0	55.4	55.3	59.9	61.2	61.7	**63.2**	58.9	56.6	62.3	63.3	**66.8**	65.0	62.0	55.7	63.1	63.7	60.9	60.3	56.1	58.1	59.5	60.6	**61.2**	61.1	60.7	59.8
	T_{CUI} as domain knowledge		**62.4**	64.0	63.1	57.8	50.6	46.6	59.3	61.0	62.6	**<u>64.1</u>**	56.8	52.0	62.5	63.3	**<u>66.9</u>**	64.4	61.0	55.7	61.3	**64.3**	**<u>66.1</u>**	64.0	60.9	55.4	59.7	60.6	61.0	**61.2**	60.6	59.7

Note: (1) Performance: AUC (%)
(2) The top-3 AUC scores in each dataset were marked in bold, among which the highest AUC scores were underlined.

Table 2.11. Performance of MD-Manifold and baselines with multimodal data fusion in the mortality prediction task.

Database			MIMIC III						MIMIC III						MIMIC III						MIMIC III						NRD					
Disease (ICD-9 code)			0389						428						41071						41401						428					
Dimensions			16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512	16	32	64	128	256	512
Summary measures	CCI		64.0						68.6						68.7						64.1						82.3					
	ECI		61.9						68.2						69.2						64.4						78.7					
	RSI		70.7						75.5						83.8						72.8						82.6					
Code mapping methods	ICD9-3digit		81.3						**80.5**						85.1						84.7						81.8					
	CCS		81.3						**80.4**						**<u>88.9</u>**						70.4						77.0					
	CUI		**<u>84.0</u>**						**83.6**						**87.9**						85.0						90.0					
	SNOMED		81.2						79.8						**88.0**						81.8						88.0					
Individual embeddings	Sum	GloVe	69.9	72.4	75.6	75.4	76.3	76.1	73.2	72.9	74.4	73.9	74.0	74.0	82.9	82.4	83.3	84.3	84.1	83.9	78.2	83.9	82.5	83.1	83.4	80.0	90.1	88.6	82.3	93.2	86.2	93.9
		Word2Vec	68.2	71.4	72.3	74.5	75.7	76.4	64.6	70.3	75.0	74.6	76.8	76.7	71.7	75.2	80.1	81.7	79.6	80.0	72.6	79.9	83.8	81.4	82.5	85.3	81.7	87.1	89.0	77.9	82.4	84.7
		FastText	68.0	71.0	69.8	70.8	72.5	72.7	72.6	72.9	73.9	73.0	72.5	73.0	76.2	79.1	79.8	79.8	81.1	81.5	74.1	77.7	76.3	79.0	75.8	80.0	88.7	91.2	87.6	90.6	90.4	82.4
	Phe2Vec	GloVe	60.8	62.1	63.1	64.5	65.9	65.0	66.7	64.2	69.4	64.5	66.7	64.9	72.5	75.7	73.2	75.1	73.6	77.2	64.6	74.8	64.7	70.3	67.9	78.9	75.1	75.9	73.5	79.2	74.1	72.6
		Word2Vec	60.5	60.0	58.8	56.7	55.6	57.2	65.6	64.5	65.7	66.3	63.3	64.9	62.6	65.1	63.3	65.3	65.7	68.2	59.0	67.1	66.2	62.9	65.2	66.9	67.4	81.0	74.8	71.3	74.1	75.0
		FastText	61.4	63.0	61.5	61.9	62.9	63.5	66.6	66.0	67.1	67.5	66.1	67.3	65.7	69.1	71.0	69.6	69.6	71.0	67.5	74.7	69.2	67.8	69.6	68.5	75.5	78.7	80.3	77.0	77.4	83.4
	Med2Vec		**82.4**	80.4	80.9	81.6	76.6	76.4	78.5	76.2	74.9	73.7	78.6	75.2	78.8	85.8	79.2	83.0	78.0	73.6	69.5	78.6	80.2	76.0	81.0	73.1	79.4	78.0	76.0	74.3	74.2	75.7
MD-Manifold	T_{ICD9} as domain knowledge		75.4	78.9	81.2	81.4	80.0	74.0	74.0	75.7	76.9	77.3	73.9	70.7	84.2	85.3	84.9	83.3	81.8	78.9	81.3	86.4	**<u>90.3</u>**	83.1	83.3	82.5	92.9	**<u>93.7</u>**	**93.2**	92.9	91.5	89.3
	T_{CCS} as domain knowledge		75.4	78.5	81.2	**81.7**	79.3	75.2	75.9	75.7	77.5	77.5	73.7	74.4	84.8	84.5	84.8	83.0	79.7	77.3	81.1	87.0	**89.7**	82.3	81.3	82.3	92.1	92.7	92.6	91.6	90.8	89.5
	T_{CUI} as domain knowledge		75.4	78.5	81.2	**81.7**	79.3	75.2	76.6	75.9	77.4	77.1	76.4	72.3	86.0	85.0	83.5	81.2	77.0	76.1	81.1	87.0	**89.7**	82.3	81.2	82.2	92.1	**93.2**	92.2	91.4	91.1	89.6

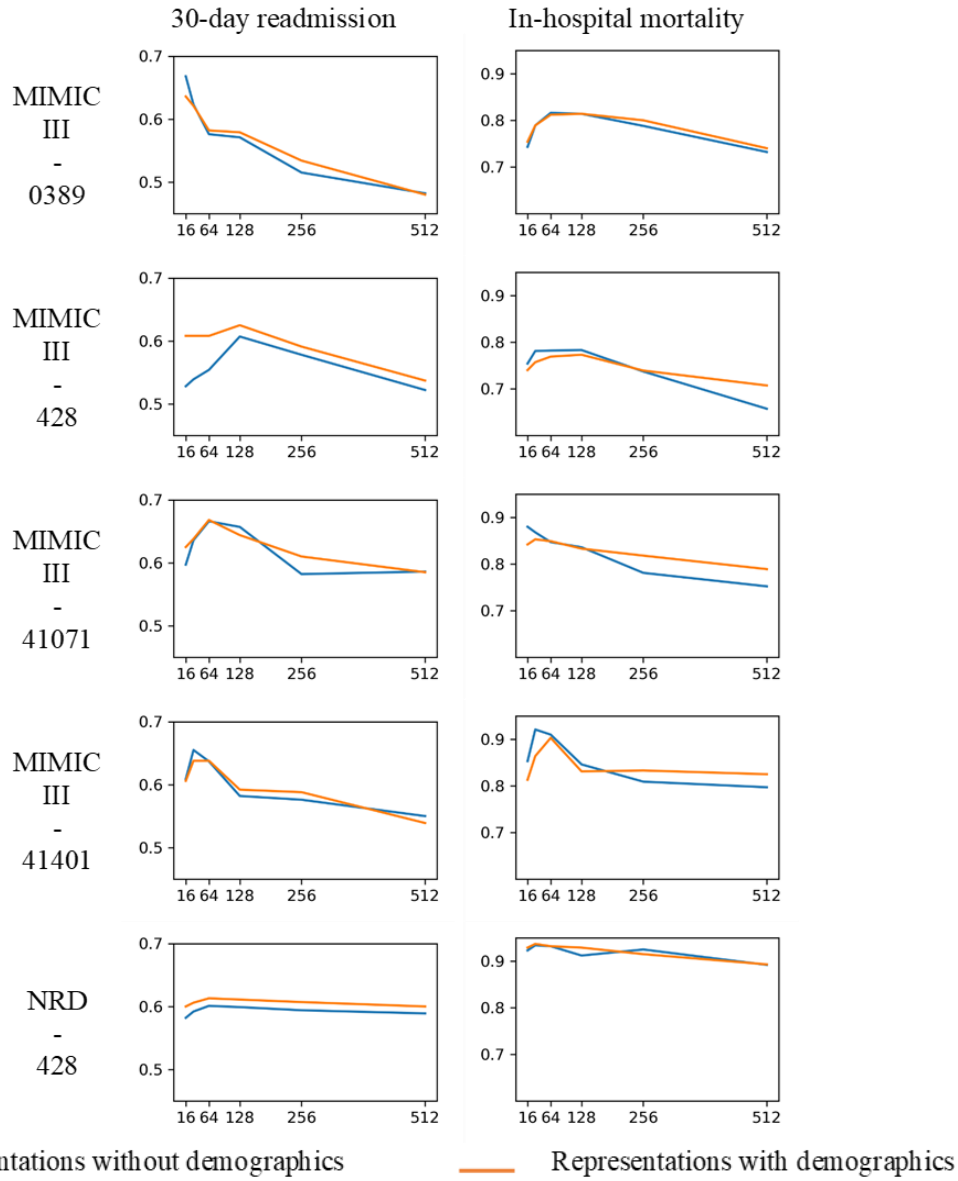
Note: (1) Performance: AUC (%)

(2) The top-3 AUC scores in each dataset were marked in bold, among which the highest AUC scores were underlined.

We compared the prediction power of the original representations to the concatenated representations in Figure 2.11. (all using T_{ICD9} as domain knowledge for brevity). The multimodal data fusion yielded varying degrees of performance improvement. (1) For certain diseases, such as septicemia (ICD-9: 0389), subendocardial infarction (ICD-9: 41071), and coronary atherosclerosis of native coronary artery (ICD-9: 41401), adding demographics did not improve the AUC scores significantly for either readmission prediction or in-hospital mortality prediction. This result was in line with previous medical research in that age and gender were not significant in factor analyses for patient cohorts with these diseases (Lam et al., 2019; Singh et al., 2019). (2) The demographics information improved the prediction accuracy significantly for heart failure patients (MIMIC III - 428 and NRD - 428) in the readmission tasks. After data fusion, the AUC scores increased from 0.528 to 0.608 (at the dimension of 16) in MIMIC III - 428, and increased from 0.582 to 0.600 (at the dimension of 16) in NRD - 428. The findings were consistent with medical observations that heart failure readmissions were strongly linked to demographic characteristics such as age, gender, and race (Mirkin et al., 2017). Another interesting finding was the decrease in AUC as the dimension of the representations for heart failure patients grew larger. The noise in the representations became stronger as the dimension increased, negatively impacting the prediction model's performance.

Additionally, we concatenated the demographic vectors to the representations generated from baseline methods and compared them with our concatenated representations. For Med2Vec (a deep learning method), we inserted demographics into its hidden layer as Choi et al. (2016) suggested. Tables 2.10. – 2.11. showed the AUC scores of the concatenated representations from both proposed and baseline methods. We highlighted the top-3 AUC scores in bold in each dataset and prediction task, with the highest AUC score bolded and underlined. We found that

the proposed representations were still the most effective. On all the five datasets and two prediction tasks, our method exceeded all the baselines in 70% of the cases (top-1 AUC scores), and reached 80% of state-of-the-art performance (top-3 AUC scores).



Note: domain knowledge: T_{ICD9}

Figure 2.11. Performance comparison: with or without multimodal data fusion. .

To summarize, we proposed a new medical concept distance metric CD_{new} to incorporate medical domain knowledge into the patient-patient network. Then, we generated low-

dimensional representations for medical records using manifold learning algorithms with the patient-patient network as the input. The generated medical records' representations were used for two healthcare prediction tasks on various patient cohorts. (1) The proposed metrics $CD_{new-eHDN}$ and $CD_{new-Cosine}$ showed superior performance over the other metrics, including the existing metric CD_{WP}. (2) Our proposed method, MD-Manifold, generated more effective low-dimensional representations for medical records (i.e., sets of medical concepts) than various state-of-the-art baseline methods. (3) Multimodal data fusion could create substantial added value for healthcare prediction analyses.

Conclusion and future work

To sum up, this study proposes a new method, Medical-Distance-manifold (MD-manifold), to generate low-dimensional representations for medical records (i.e., each record contains a set of medical concepts) in EHR.

The technical contributions of this study are significant. (1) In the proposed method, we develop a new medical-concept distance metric that considers both the medical concepts' hierarchy (i.e., knowledge-driven) and their co-occurrences (i.e., data-driven) as medical domain knowledge. The experimental results show the proposed metric is better than the existing metric, CD_{WP}, for measuring the distances between medical concepts in EHR. (2) Using the proposed medical-concept distance metric, we create a new patient-patient network with medical domain knowledge embedded. Using the patient-patient network as the input of manifold learning algorithms, we generate the low-dimensional representations of medical records. Using our representations, prediction models outperform the state-of-the-art methods on two large real-world databases and two prediction tasks. (3) Patients and their health conditions can be described by multimodal data derived from various sources. The medical record's low-

dimensional representation generated by our method can be combined with other data modalities and enhance the performance of healthcare prediction models.

The advantages of our method are as follows. (1) Computation efficiency. The proposed method is based on manifold learning algorithms, which are known for representing high-dimensional data in low dimensions. Our method provides a computationally efficient solution for medical record-level healthcare prediction tasks (e.g., readmission and in-hospital mortality predictions) since it dramatically decreases the dimensions needed to represent a medical record (i.e., sets of medical concepts). (2) Less demanding on the training data size. Most existing medical-concept embedding approaches are based on NLP-related deep learning techniques, which likely necessitate large training datasets. For example, Pennington et al. (2014) train their representations on 42 billion word tokens which are more than most medical record systems can handle (Johnson et al., 2016). On the contrary, the proposed method is based on manifold learning algorithms (e.g., Laplacian Eigenmaps and Isomap), which generally optimize representations using matrix factorization techniques such as singular value decomposition (SVD) (Klema & Laub, 1980) and do not need huge training datasets. Our low-dimensional representations are generated using thousands of data records, and their performance is consistent in different prediction tasks and patient cohorts. (3) Pre-trained and task-agnostic medical-record-level representations. The basic idea of this work is to preprocess the high-dimensional EHR data and derive the low-dimensional representations offline. Most preprocessing computations can be completed offline during machine idle time. In addition to readmission and death prediction, the derived representations can be used for other medical-record-level prediction tasks, such as length-of-stay prediction, healthcare cost prediction, and equipment maintenance needs identification. (4) Strong generalizability. First, we evaluate our

method on five datasets and two healthcare prediction tasks, and compare the results with 14 baselines. The experimental results indicate that our method leads to improved performance in different healthcare prediction tasks on different patient cohorts. Second, the generated low-dimensional representations are ready to be used as input to any machine learning models for supervised or unsupervised tasks. Although we use simple classification models in the experiments to evaluate our framework's efficacy, more advanced deep learning models (e.g., sequence-to-sequence models and attention-based models) can benefit from using our low-dimensional representations. Third, we propose an open framework, and its performance can be further boosted by incorporating more advanced medical knowledge trees, medical distance metrics, or manifold learning algorithms.

The managerial implications of this study are twofold.

(1) Our method has great potential to alleviate the inaccurate prediction problem caused by medical coding errors. In medical systems, accurate coding of medical concepts (e.g., diagnosis and services codes) has become increasingly important. Normally, coding is a manual process that involves the human evaluation of clinical documentation to identify applicable codes. The code assignment may be carried out by physicians, but it is often performed by other personnel, such as coding professionals. They need to extract key information from medical records and assign correct codes based on category, anatomic site, laterality, severity, and etiology (Quan et al. 2005). The coding process is labor-intensive and error-prone (Stanfill et al., 2010). According to Horsky et al. (2017), only a half (56%) of the issued diagnostic codes are rated as appropriate in the US, and about one-quarter are omitted.

Our method is more robust to low-quality EHR data because it is based on the medical-concept hierarchical structure. Substituting a medical concept (e.g., ICD-9 code) with another

similar medical concept (i.e., inaccurate coding) in a medical record does not significantly affect the concept-level distance calculation. Thus, the generated low-dimensional representations still preserve valuable information from the EHR, resulting in better performance in healthcare prediction tasks. By contrast, other widely used embedding methods do not consider the medical-concept hierarchy, leading to less favorable results. For example, even two sibling ICD-9 codes are represented by entirely different vectors using Word2Vec embeddings, if they do not have a close frequency in the EHR. Therefore, such methods are sensitive to coding errors and the quality of the EHR data.

(2) The proposed method is likely to increase the use of healthcare prediction models in actual clinical practice. There is no doubt that healthcare predictive analytics can support clinical decisions. However, the use of healthcare prediction models in real-world clinical practice is still limited (Moons et al., 2009). One of the barriers is that most extant healthcare prediction models focused on one specific prediction task over specific patient cohorts (Y.-K. Lin et al., 2017). It is difficult for healthcare practitioners to adapt the models from the intended purpose to other predictions or the study population to the local population (Moons et al., 2009).

On the other hand, the low-dimensional representations for medical records generated by our method can be pre-generated and task-agnostic. Different healthcare prediction models can use the generated low-dimensional representations for multiple purposes, lowering the barriers to the wide use of prediction models in healthcare practice.

Although promising, the proposed method is not without limitations, and our method can be extended in the following ways in the future. (1) The proposed framework does not distinguish between the sequential medical records of same patients from other medical records. As in previous research (Choi, Bahadori, Searles, et al., 2016), all medical records are treated

equally and independently. However, the proposed method can be further improved by considering the connection among the successive medical records of same patients (De Freitas et al., 2020). It is reasonable to assume that the sequential records are related because patients' health conditions change continuously. To some extent, the representations of these medical records should show similarities. It would be potentially beneficial if we capture the sequentiality in the future. (2) Manifold learning is an increasing research area with many new algorithms and applications. In this study, we adapt Laplacian Eigenmap and Isomap as examples of the local and global approaches of the manifold learning algorithms, which are not thorough. We intend to explore more manifold learning algorithms in the future. (3) Moreover, we can evaluate the generalization ability of our method by inspecting other medical record-level prediction tasks, such as length-of-stay prediction, healthcare cost prediction, and equipment maintenance needs identification. (4) The databases we used for our empirical evaluation (i.e., NRD and MIMIC III) do not have the timestamps of the ICD-9 codes. We cannot build healthcare prediction models a few days before patients' discharges/deaths, limiting the prediction models' practical significance when early interventions are needed. The proposed method can be evaluated using early prediction models when databases with medical concepts' timestamps are available. (5) We only include the ICD-9 diagnosis codes in our evaluation due to data limitations. The proposed method can be extended when other medical concepts (e.g., procedure codes or drug codes) are available. (6) We evaluate the proposed method empirically. In the future, our method can be evaluated by medical experts (Choi et al., 2016).

References

Allam, A., Nagy, M., Thoma, G., & Krauthammer, M. (2019). Neural networks versus Logistic regression for 30 days all-cause readmission prediction. *Scientific Reports*, *9*(1), 9277. https://doi.org/10.1038/s41598-019-45685-z

Altibi, A. M., Prousi, G., Agarwal, M., Shah, M., Tripathi, B., Ram, P., & Patel, B. (2021). Readmission-free period and in-hospital mortality at the time of first readmission in acute heart failure patients—NRD-based analysis of 40,000 heart failure readmissions. *Heart Failure Reviews*, *26*(1), 57–64. https://doi.org/10.1007/s10741-019-09912-z

Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics*, *97*, 103256. https://doi.org/10.1016/j.jbi.2019.103256

Baillie, C. A., VanZandbergen, C., Tait, G., Hanish, A., Leas, B., French, B., William Hanson, C., Behta, M., & Umscheid, C. A. (2013). The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission: Patients at Risk for Readmission. *Journal of Hospital Medicine*, *8*(12), 689–695. https://doi.org/10.1002/jhm.2106

Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, *12*(1), 56–68. https://doi.org/10.1038/nrg2918

Bardhan, I., Oh, J. (Cath), Zheng, Z. (Eric), & Kirksey, K. (2015). Predictive Analytics for Readmission of Patients with Congestive Heart Failure. *Information Systems Research*, *26*(1), 19–39. https://doi.org/10.1287/isre.2014.0553

Barrenas, F., Chavali, S., Holme, P., Mobini, R., & Benson, M. (2009). Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS ONE*, *4*(11), e8090. https://doi.org/10.1371/journal.pone.0008090

Becker, R. B., & Zimmerman, J. E. (1996). ICU scoring systems allow prediction of patient outcomes and comparison of ICU performance. *Critical Care Clinics*, *12*(3), 503–514. https://doi.org/10.1016/s0749-0704(05)70258-x

Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317

Ben-Assuli, O., & Padman, R. (2020). Trajectories of Repeated Readmissions of Chronic Disease Patients: Risk Stratification, Profiling, and Prediction. *MIS Quarterly*, *44*(1), 201–226. https://doi.org/10.25300/MISQ/2020/15101

Bodenreider, O., & Stevens, R. D. (2006). Bio-ontologies: Current trends and future directions. *Briefings in Bioinformatics*, *7*(3), 256–274. https://doi.org/10.1093/bib/bbl027

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., & Bork, P. (2008). Drug Target Identification Using Side-Effect Similarity. *Science*, *321*(5886), 263–266. https://doi.org/10.1126/science.1158140

Castanedo, F. (2013). A Review of Data Fusion Techniques. *The Scientific World Journal*, *2013*, 1–19. https://doi.org/10.1155/2013/704504

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, *40*(5), 373–383. https://doi.org/10.1016/0021-9681(87)90171-8

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). *RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism*. https://doi.org/10.48550/ARXIV.1608.05745

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., & Sun, J. (2016). Multi-layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1495–1504. https://doi.org/10.1145/2939672.2939823

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: Graph-based Attention Model for Healthcare Representation Learning. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 787–795. https://doi.org/10.1145/3097983.3098126

Choi, E., Xiao, C., Stewart, W. F., & Sun, J. (2018). MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. *ArXiv:1810.09593 [Cs, Stat]*. http://arxiv.org/abs/1810.09593

Cotter, P. E., Bhalla, V. K., Wallis, S. J., & Biram, R. W. S. (2012). Predicting readmissions: Poor performance of the LACE index in an older UK population. *Age and Ageing*, *41*(6), 784–789. https://doi.org/10.1093/ageing/afs073

Cui, L., Xie, X., & Shen, Z. (2018). Prediction task guided representation learning of medical codes in EHR. *Journal of Biomedical Informatics*, *84*, 1–10. https://doi.org/10.1016/j.jbi.2018.06.013

De Freitas, J. K., Johnson, K. W., Golden, E., Nadkarni, G. N., Dudley, J. T., Bottinger, E. P., Glicksberg, B. S., & Miotto, R. (2020). *Phe2vec: Automated Disease Phenotyping based on Unsupervised Embeddings from Electronic Health Records* [Preprint]. Health Informatics. https://doi.org/10.1101/2020.11.14.20231894

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., & Bruza, P. (2014). Medical Semantic Similarity with a Neural Language Model. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1819–1822. https://doi.org/10.1145/2661829.2661974

Deschepper, M., Eeckloo, K., Vogelaers, D., & Waegeman, W. (2019). A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Computer Methods and Programs in Biomedicine*, *173*, 177–183. https://doi.org/10.1016/j.cmpb.2019.02.007

Donoho, D. L., & Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, *100*(10), 5591–5596. https://doi.org/10.1073/pnas.1031596100

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). *Comorbidity Measures for Use with Administrative Data*. *36*(1), 8–27.

Emmert-Streib, F., Tripathi, S., Simoes, R. de M., Hawwa, A. F., & Dehmer, M. (2013). The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes. *Systems Biomedicine*, *1*(1), 20–28. https://doi.org/10.4161/sysb.22816

Fredkin, E. (1960). Trie memory. *Communications of the ACM*, *3*(9), 490–499. https://doi.org/10.1145/367390.367400

García del Valle, E. P., Lagunes García, G., Prieto Santamaría, L., Zanin, M., Menasalvas Ruiz, E., & Rodríguez-González, A. (2019). Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources. *Journal of Biomedical Informatics*, *94*, 103206. https://doi.org/10.1016/j.jbi.2019.103206

Gheorghiade, M., Vaduganathan, M., Fonarow, G. C., & Bonow, R. O. (2013). Rehospitalization for Heart Failure. *Journal of the American College of Cardiology*, *61*(4), 391–403. https://doi.org/10.1016/j.jacc.2012.09.038

Girardi, D., Wartner, S., Halmerbauer, G., Ehrenmüller, M., Kosorus, H., & Dreiseitl, S. (2016). Using concept hierarchies to improve calculation of patient similarity. *Journal of Biomedical Informatics*, *63*, 66–73. https://doi.org/10.1016/j.jbi.2016.07.021

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, *104*(21), 8685–8690. https://doi.org/10.1073/pnas.0701361104

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Heidary Moghadam, P., Heidari, V., Moeini, A., & Kamandi, A. (2019). An exponential similarity measure for collaborative filtering. *SN Applied Sciences*, *1*(10), 1172. https://doi.org/10.1007/s42452-019-1142-8

Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Hidalgo, C. A., Blumm, N., Barabási, A.-L., & Christakis, N. A. (2009). A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology*, *5*(4), e1000353. https://doi.org/10.1371/journal.pcbi.1000353

Horsky, J., Drucker, E. A., & Ramelson, H. Z. (2017). Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2017*, 912–920.

Hou, Y., Zhou, Y., Hussain, M., Budd, G. T., Tang, W. H. W., Abraham, J., Xu, B., Shah, C., Moudgil, R., Popovic, Z., Watson, C., Cho, L., Chung, M., Kanj, M., Kapadia, S., Griffin, B., Svensson, L., Collier, P., & Cheng, F. (2021). Cardiac risk stratification in cancer patients: A longitudinal patient–patient network analysis. *PLOS Medicine*, *18*(8), e1003736. https://doi.org/10.1371/journal.pmed.1003736

Huang, R., Zhang, G., & Chen, J. (2019). Semi-supervised discriminant Isomap with application to visualization, image retrieval and classification. *International Journal of Machine Learning and Cybernetics*, *10*(6), 1269–1278. https://doi.org/10.1007/s13042-018-0809-6

Jamie L. Habib. (2010). *EHRs, Meaningful Use, and a Model EMR*. *22*(4).

Jia, Z., Lu, X., Duan, H., & Li, H. (2019). Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Medical Informatics and Decision Making*, *19*(1), 91. https://doi.org/10.1186/s12911-019-0807-y

Jiang, Y., Ma, S., Shia, B.-C., & Lee, T.-S. (2018). An Epidemiological Human Disease Network Derived from Disease Co-occurrence in Taiwan. *Scientific Reports*, *8*(1), 4557. https://doi.org/10.1038/s41598-018-21779-y

John, R., Kerby, D. S., & Hagan Hennessy, C. (2003). Patterns and Impact of Comorbidity and Multimorbidity Among Community-Resident American Indian Elders. *The Gerontologist*, *43*(5), 649–660. https://doi.org/10.1093/geront/43.5.649

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 160035. https://doi.org/10.1038/sdata.2016.35

Jung, K., Sudat, S. E. K., Kwon, N., Stewart, W. F., & Shah, N. H. (2019). Predicting need for advanced illness or palliative care in a primary care population using electronic health record data. *Journal of Biomedical Informatics*, *92*, 103115. https://doi.org/10.1016/j.jbi.2019.103115

Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, *25*(2), 164–176. https://doi.org/10.1109/TAC.1980.1102314

Kobewka, D. M., van Walraven, C., Turnbull, J., Worthington, J., Calder, L., & Forster, A. (2017). Quality gaps identified through mortality review. *BMJ Quality & Safety*, *26*(2), 141–149. https://doi.org/10.1136/bmjqs-2015-004735

Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Medical Informatics and Decision Making*, *20*(1), 1–10.

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Kumar, A., Gupta, S., Singh, S. K., & Shukla, K. K. (2015). Comparison of various metrics used in collaborative filtering for recommendation system. *2015 Eighth International Conference on Contemporary Computing (IC3)*, 150–154. https://doi.org/10.1109/IC3.2015.7346670

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*, *103*(9), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

Lam, L., Ahn, H. J., Okajima, K., Schoenman, K., Seto, T. B., Shohet, R. V., Miyamura, J., Sentell, T. L., & Nakagawa, K. (2019). Gender Differences in the Rate of 30-Day Readmissions after Percutaneous Coronary Intervention for Acute Coronary Syndrome. *Women's Health Issues*, *29*(1), 17–22. https://doi.org/10.1016/j.whi.2018.09.002

Lee, D.-S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N., & Barabási, A.-L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, *105*(29), 9880–9885. https://doi.org/10.1073/pnas.0802208105

Lee, J., Dubin, J. A., & Maslove, D. M. (2016). Mortality Prediction in the ICU. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 315–324). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_21

Leppin, A. L., Gionfriddo, M. R., Kessler, M., Brito, J. P., Mair, F. S., Gallacher, K., Wang, Z., Erwin, P. J., Sylvester, T., Boehmer, K., Ting, H. H., Murad, M. H., Shippee, N. D., & Montori, V. M. (2014). Preventing 30-day hospital readmissions: A systematic review and meta-analysis of randomized trials. *JAMA Internal Medicine*, *174*(7), 1095–1107. https://doi.org/10.1001/jamainternmed.2014.1608

Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, *7*(311), 311ra174. https://doi.org/10.1126/scitranslmed.aaa9364

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, *10*(1), 7155. https://doi.org/10.1038/s41598-020-62922-y

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., & Yang, H.-J. (2017). Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach. *MIS Quarterly*, *41*(2), 473–495. https://doi.org/10.25300/MISQ/2017/41.2.07

Lin, Y.-W., Zhou, Y., Faghri, F., Shaw, M. J., & Campbell, R. H. (2019). Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLOS ONE*, *14*(7), e0218942. https://doi.org/10.1371/journal.pone.0218942

Lopez, K., Fodeh, S. J., Allam, A., Brandt, C. A., & Krauthammer, M. (2020). Reducing Annotation Burden Through Multimodal Learning. *Frontiers in Big Data*, *3*, 19. https://doi.org/10.3389/fdata.2020.00019

Mehta, H. B., Sura, S. D., Adhikari, D., Andersen, C. R., Williams, S. B., Senagore, A. J., Kuo, Y.-F., & Goodwin, J. S. (2018). Adapting the Elixhauser comorbidity index for cancer patients: Comparison of Comorbidity Scores in Surgery. *Cancer*, *124*(9), 2018–2025. https://doi.org/10.1002/cncr.31269

Melton, G. B., Parsons, S., Morrison, F. P., Rothschild, A. S., Markatou, M., & Hripcsak, G. (2006). Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*, *39*(6), 697–705. https://doi.org/10.1016/j.jbi.2006.01.004

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A Machine Learning Approach to Improving Dynamic Decision Making. *Information Systems Research*, *25*(2), 239–263. https://doi.org/10.1287/isre.2014.0513

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. https://doi.org/10.48550/ARXIV.1310.4546

Min, X., Yu, B., & Wang, F. (2019). Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. *Scientific Reports*, *9*(1), 2362. https://doi.org/10.1038/s41598-019-39071-y

Mirkin, K. A., Enomoto, L. M., Caputo, G. M., & Hollenbeak, C. S. (2017). Risk factors for 30-day readmission in patients with congestive heart failure. *Heart & Lung*, *46*(5), 357–362. https://doi.org/10.1016/j.hrtlng.2017.06.005

Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *BMJ*, *338*(feb23 1), b375–b375. https://doi.org/10.1136/bmj.b375

Moreno, R. P., Metnitz, P. G., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., & Le Gall, J.-R. (2005). SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, *31*(10), 1345–1355.

Mumtaz, K., Issak, A., Porter, K., Kelly, S., Hanje, J., Michaels, A. J., Conteh, L. F., El-Hinnawi, A., Black, S. M., & Abougergi, M. S. (2019). Validation of Risk Score in Predicting Early Readmissions in Decompensated Cirrhotic Patients: A Model Based on the Administrative Database: Hepatology. *Hepatology*, *70*(2), 630–639. https://doi.org/10.1002/hep.30274

Mysling, P., Hauberg, S., & Pedersen, K. S. (2011). An Empirical Study on the Performance of Spectral Manifold Learning Techniques. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2011* (Vol. 6791, pp. 347–354). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21735-7_43

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 841–848.

O'Connor, C. M. (2017). High Heart Failure Readmission Rates. *JACC: Heart Failure*, *5*(5), 393. https://doi.org/10.1016/j.jchf.2017.03.011

Oh, S. (2019). Feature Interaction in Terms of Prediction Performance. *Applied Sciences*, *9*(23), 5191. https://doi.org/10.3390/app9235191

Pai, S., & Bader, G. D. (2018). Patient Similarity Networks for Precision Medicine. *Journal of Molecular Biology*, *430*(18), 2924–2938. https://doi.org/10.1016/j.jmb.2018.05.037

Pai, S., Hui, S., Isserlin, R., Shah, M. A., Kaka, H., & Bader, G. D. (2019). netDx: Interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology*, *15*(3). https://doi.org/10.15252/msb.20188497

Park, H. (2012). ISOMAP induced manifold embedding and its application to Alzheimer's disease and mild cognitive impairment. *Neuroscience Letters*, *513*(2), 141–145. https://doi.org/10.1016/j.neulet.2012.02.016

Patwari, N., Hero, A. O., & Pacholski, A. (2005). Manifold learning visualization of network traffic data. *Proceeding of the 2005 ACM SIGCOMM Workshop on Mining Network Data - MineNet '05*, 191. https://doi.org/10.1145/1080173.1080182

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pirracchio, R. (2016). Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 295–313). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_20

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E., & Ghali, W. A. (2005). Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data: *Medical Care*, *43*(11), 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83

Rasmy, L., Tiryaki, F., Zhou, Y., Xiang, Y., Tao, C., Xu, H., & Zhi, D. (2020). Representation of EHR data for predictive modeling: A comparison between UMLS and other terminologies. *Journal of the American Medical Informatics Association*, *27*(10), 1593–1599. https://doi.org/10.1093/jamia/ocaa180

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Sacco Casamassima, M. G., Salazar, J. H., Papandria, D., Fackler, J., Chrouser, K., Boss, E. F., & Abdullah, F. (2014). Use of risk stratification indices to predict mortality in critically ill children. *European Journal of Pediatrics*, *173*(1), 1–13. https://doi.org/10.1007/s00431-013-1987-6

Schneeweiss, S., Wang, P. S., Avorn, J., & Glynn, R. J. (2003). Improved Comorbidity Adjustment for Predicting Mortality in Medicare Populations: Improved Comorbidity Adjustment in Medicare Populations. *Health Services Research*, *38*(4), 1103–1120. https://doi.org/10.1111/1475-6773.00165

Sejong Oh. (2021). Age dependent associations of risk factors with heart failure: Pooled population based cohort study. *BMJ*, n880. https://doi.org/10.1136/bmj.n880

Sessler, D. I., Sigl, J. C., Manberg, P. J., Kelley, S. D., Schubert, A., & Chamoun, N. G. (2010). Broadly Applicable Risk Stratification System for Predicting Duration of Hospitalization and Mortality. *Anesthesiology*, *113*(5), 1026–1037. https://doi.org/10.1097/ALN.0b013e3181f79a8d

Shaw, B., & Jebara, T. (2009). Structure preserving embedding. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 1–8. https://doi.org/10.1145/1553374.1553494

Shier, J., McNally, K., Tzanetakis, G., & Brooks, K. G. (2021). Manifold Learning Methods for Visualization and Browsing of Drum Machine Samples. *Journal of the Audio Engineering Society*, *69*(1/2), 40–53. https://doi.org/10.17743/jaes.2020.0064

Shmueli & Koppius. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, *35*(3), 553. https://doi.org/10.2307/23042796

Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Jim Zheng, W., & Roberts, K. (2021). Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *Journal of Biomedical Informatics*, *115*, 103671. https://doi.org/10.1016/j.jbi.2020.103671

Sigakis, M. J. G., Bittner, E. A., & Wanderer, J. P. (2013). Validation of a Risk Stratification Index and Risk Quantification Index for Predicting Patient Outcomes. *Anesthesiology*, *119*(3), 525–540. https://doi.org/10.1097/ALN.0b013e31829ce6e6

Silva, V. D., & Tenenbaum, J. B. (2003). Global Versus Local Methods in Nonlinear Dimensionality Reduction. In *Advances in Neural Information Processing Systems 15* (pp. 705–712). MIT Press.

Singh, A., Bhagat, M., George, S. V., Gorthi, R., & Chaturvedula, C. (2019). Factors Associated with 30-day Unplanned Readmissions of Sepsis Patients: A Retrospective Analysis of Patients Admitted with Sepsis at a Community Hospital. *Cureus*. https://doi.org/10.7759/cureus.5118

Song, C., Zhang, S., Sadoughi, N., Xie, P., & Xing, E. (2020). Generalized Zero-Shot Text Classification for ICD Coding. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4018–4024. https://doi.org/10.24963/ijcai.2020/556

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association: JAMIA*, *17*(6), 646–651. https://doi.org/10.1136/jamia.2009.001024

Stiebellehner, S., Wang, J., & Yuan, S. (2018). *Learning Continuous User Representations through Hybrid Filtering with doc2vec*. https://doi.org/10.48550/ARXIV.1801.00215

Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H., & Ghali, W. A. (2004). New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of Clinical Epidemiology*, *57*(12), 1288–1294. https://doi.org/10.1016/j.jclinepi.2004.03.012

Talwalkar, A., Kumar, S., & Rowley, H. (2008). Large-scale manifold learning. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2008.4587670

Taneja, S. S. (2010). *Complications of urologic surgery: Prevention and management* (4th ed). Saunders/Elsevier.

Tang, F., Xiao, C., Wang, F., & Zhou, J. (2018). Predictive modeling in urgent care: A comparative study of machine learning approaches. *JAMIA Open*, *1*(1), 87–98. https://doi.org/10.1093/jamiaopen/ooy011

Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Teo, K., Yong, C. W., Chuah, J. H., Hum, Y. C., Tee, Y. K., Xia, K., & Lai, K. W. (2021). Current Trends in Readmission Prediction: An Overview of Approaches. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-021-06040-5

Tu, S. T., Chen, J. Y., Yang, W., & Sun, H. (2012). Laplacian Eigenmaps-Based Polarimetric Dimensionality Reduction for SAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(1), 170–179. https://doi.org/10.1109/TGRS.2011.2168532

Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: How can we know it works? *Journal of the American Medical Informatics Association*, *26*(12), 1651–1654. https://doi.org/10.1093/jamia/ocz130

van Walraven, C., Austin, P. C., Jennings, A., Quan, H., & Forster, A. J. (2009). A Modification of the Elixhauser Comorbidity Measures Into a Point System for Hospital Death Using Administrative Data. *Medical Care*, *47*(6), 626–633. https://doi.org/10.1097/MLR.0b013e31819432e5

van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., & Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, *182*(6), 551–557. https://doi.org/10.1503/cmaj.091117

van Walraven, C., Wong, J., & Forster, A. J. (2012). LACE+ index: Extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open Medicine: A Peer-Reviewed, Independent, Open-Access Journal*, *6*(3), e80-90.

Verdecchia, P. (2003). Improved cardiovascular risk stratification by a simple ECG index in hypertension. *American Journal of Hypertension*, *16*(8), 646–652. https://doi.org/10.1016/S0895-7061(03)00912-9

Wang, H.-H., Wang, Y.-H., Liang, C.-W., & Li, Y.-C. (2019). Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer. *JAMA Dermatology*, *155*(11), 1277. https://doi.org/10.1001/jamadermatol.2019.2335

Wang, X., Wang, F., Hu, J., & Sorrentino, R. (2014). Exploring joint disease risk prediction. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2014*, 1180–1187.

Weinberger, K. Q., Packer, B. D., & Saul, L. K. (2005). Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *International Workshop on Artificial Intelligence and Statistics* (pp. 381–388). PMLR.

Whittington, J., Simmonds, T., & Jacobsen, D. (2005). *Reducing Hospital Mortality Rates (Part 2)*. IHI Innovation Series white paper.

Williams, R., Kontopantelis, E., Buchan, I., & Peek, N. (2017). Clinical code set engineering for reusing EHR data for research: A review. *Journal of Biomedical Informatics*, *70*, 1–13. https://doi.org/10.1016/j.jbi.2017.04.010

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* -, 133–138. https://doi.org/10.3115/981732.981751

Xiang, Y., Xu, J., Si, Y., Li, Z., Rasmy, L., Zhou, Y., Tiryaki, F., Li, F., Zhang, Y., Wu, Y., Jiang, X., Zheng, W. J., Zhi, D., Tao, C., & Xu, H. (2019). Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Medical Informatics and Decision Making*, *19*(S2), 58. https://doi.org/10.1186/s12911-019-0766-3

Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., & Sun, J. (2018). *RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data*. https://doi.org/10.48550/ARXIV.1807.08820

Yıldırım, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., & Vidal, M. (2007). Drug—Target network. *Nature Biotechnology*, *25*(10), 1119–1126. https://doi.org/10.1038/nbt1338

Youngduck Choi, Chiu, C. Y.-I., & Sontag, D. (2016). Learning Low-Dimensional Representations of Medical Concepts. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2016*, 41–50.

Yu, S., Farooq, F., van Esbroeck, A., Fung, G., Anand, V., & Krishnapuram, B. (2015). Predicting readmission risk with institution-specific prediction models. *Artificial Intelligence in Medicine*, *65*(2), 89–96. https://doi.org/10.1016/j.artmed.2015.08.005

Yuhua Li, Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, *15*(4), 871–882. https://doi.org/10.1109/TKDE.2003.1209005

Zhang, J., Huang, H., & Wang, J. (2010). Manifold Learning for Visualizing and Analyzing High-dimensional Data. *IEEE Intelligent Systems*, 5401149. https://doi.org/10.1109/MIS.2010.8

Zhang, W., & Ram, S. (2020). A Comprehensive Analysis of Triggers and Risk Factors for Asthma Based on Machine Learning and Large Heterogeneous Data Sources. *MIS Quarterly*, *44*(1), 305–349. https://doi.org/10.25300/MISQ/2020/15106

Zhou, X., Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms–disease network. *Nature Communications*, *5*(1), 4212. https://doi.org/10.1038/ncomms5212

Zimmerman, J. E., Kramer, A. A., McNair, D. S., & Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, *34*(5), 1297–1310.

Appendix A: The baseline medical concept distance metric CD_{WP}

We introduce a widely used medical concept distance metric CD_{WP} (Wu and Palmer 1994), as our baseline. It is also the foundation upon which we develop our new distance metric CD_{new}.

Given the concept structure shown in Figure 2.4., if two medical concepts are connected by an edge, then the medical concept in the upper level is called a parent, and the one in the lower level is called a child. For example, in Figure 2.4., 428 is the parent of 4282; and 4282 is the child of 428.

Intuitively, CD_{WP} considers concept a and b as distant if their least common ancestor (LCA) is much closer to the root of the concept tree compared with a and b. Specifically, $CD_{WP}(a,b) = 1 - \frac{2IC(c)}{IC(a)+IC(b)}$, where c is the LCA, and Information Content (IC) is defined as the number of the level in the concept tree. A concept has more IC if it is far from the root because it is more specific. Particularly, the IC of the root is 1, the IC of the concept that is connected with the root is defined as 2, and so on. If $IC(c)$ is much smaller than $IC(a)$ and $IC(b)$, this indicates that c is far from a and b; consequently, a and b are also distant with a large $CD_{WP}(a,b)$, and vice versa. For example, as shown in Figure 2.4., the distance between 4289 and 42820 is $1-(2\times 3)/(4\times 5) = 1/3$, if 428 is connected to the root.

As we mentioned in Sections 2 and 3, though straight forward and powerful, CD_{WP} has its limitations. First, the distances among medical concepts are fully determined by the concept hierarchy without considering the concept co-occurrences or frequencies in practice. For example, two distant medical concepts co-occurring frequently tend to relate closely with each other, which is not reflected in the medical concepts' hierarchical structure. Moreover, it is likely that a medical concept occurs more frequently than its siblings, thus, it is possible that such a

medical concept may have a closer relationship with its parent than its siblings. Nevertheless, the distance between a parent and each child is equal in CD_{WP} . For example, in Figure 2.4., $CD_{WP}(4282, 42820) = CD_{WP}(4282, 42823)$, regardless of the frequency of 42820 and 42823 in practice.

To address the above-mentioned limitations, we propose a new data-driven concept-level distance metric, CD_{new}, that considers both the structure of the medical concept hierarchy and the medical concepts' cooccurrences and frequencies in the EHR dataset.

Appendix B: Supplementary experimental results

Table 2.12. Prediction performance using different classifiers.

Dataset	Classifier	Dimension					
		16	32	64	128	256	512
MIMIC - 0389	LR	**0.668**	0.623	0.576	0.571	0.515	0.482
	AdaBoost	0.556	0.571	0.555	0.557	0.520	0.513
	RF	0.569	0.571	0.582	0.542	0.546	0.529
MIMIC - 428	LR	0.528	0.539	0.554	**0.607**	0.578	0.522
	AdaBoost	0.457	0.484	0.510	0.502	0.520	0.523
	RF	0.506	0.516	0.529	0.534	0.491	0.553
MIMIC - 41071	LR	0.597	0.636	**0.666**	0.657	0.582	0.586
	AdaBoost	0.577	0.501	0.567	0.552	0.529	0.523
	RF	0.485	0.596	0.597	0.556	0.511	0.519
MIMIC - 41401	LR	0.609	**0.655**	0.637	0.582	0.576	0.550
	AdaBoost	0.609	0.607	0.581	0.559	0.522	0.574
	RF	0.646	0.619	0.603	0.602	0.561	0.607
NRD - 428	LR	0.582	0.592	**0.601**	0.599	0.594	0.589
	AdaBoost	0.566	0.573	0.571	0.569	0.563	0.563
	RF	0.577	0.583	0.585	0.583	0.582	0.576

(a) Readmission prediction

Note: (1) The best performance on each dataset is bold. (2) Medical knowledge: T_{ICD9}. (3) RF: random forest. LR: logistic regression.

Table 2.12. Continued.

Dataset	Classifier	Dimension					
		16	32	64	128	256	512
MIMIC - 0389	LR	0.743	0.789	**0.816**	0.814	0.788	0.732
	AdaBoost	0.698	0.732	0.757	0.748	0.728	0.719
	RF	0.725	0.748	0.755	0.760	0.752	0.739
MIMIC - 428	LR	0.754	0.781	0.782	**0.783**	0.737	0.657
	AdaBoost	0.674	0.680	0.705	0.766	0.699	0.689
	RF	0.753	0.762	0.766	0.715	0.730	0.706
MIMIC - 41071	LR	**0.880**	0.868	0.847	0.836	0.781	0.752
	AdaBoost	0.765	0.753	0.737	0.742	0.754	0.753
	RF	0.840	0.843	0.837	0.841	0.830	0.822
MIMIC - 41401	LR	0.853	**0.921**	0.910	0.846	0.809	0.797
	AdaBoost	0.709	0.661	0.771	0.759	0.682	0.648
	RF	0.887	0.896	0.894	0.919	0.874	0.899
NRD - 428	LR	0.923	**0.934**	0.932	0.912	0.925	0.892
	AdaBoost	0.908	0.905	0.911	0.900	0.895	0.889
	RF	0.906	0.908	0.907	0.907	0.904	0.900

(b) In-hospital mortality prediction

Note: (1) The best performance on each dataset is bold. (2) Medical knowledge: T_{ICD9}. (3) RF: random forest. LR: logistic regression.

Table 2.13. Prediction performance using different manifold learning algorithms

Dataset	Classifier	Dimension					
		16	32	64	128	256	512
MIMIC - 0389	Isomap	**0.668**	0.623	0.576	0.571	0.515	0.482
	Eigenmap	0.660	0.616	0.575	0.556	0.554	0.468
MIMIC - 428	Isomap	0.528	0.539	0.554	**0.607**	0.578	0.522
	Eigenmap	0.526	0.512	0.543	0.580	0.528	0.534
MIMIC - 41071	Isomap	0.597	0.636	0.666	0.657	0.582	0.586
	Eigenmap	0.617	0.561	0.618	0.622	**0.670**	0.541
MIMIC - 41401	Isomap	0.609	**0.655**	0.637	0.582	0.576	0.550
	Eigenmap	0.588	0.583	0.567	0.589	0.600	0.576
NRD - 428	Isomap	0.582	0.592	**0.601**	0.599	0.594	0.589
	Eigenmap	0.573	0.588	**0.601**	**0.601**	0.600	0.594

(a) Readmission prediction

Table 2.13. Continued.

Dataset	Classifier	Dimension					
		16	32	64	128	256	512
MIMIC - 0389	Isomap	0.743	0.789	**0.816**	0.814	0.788	0.732
	Eigenmap	0.739	0.773	0.798	0.802	0.778	0.728
MIMIC - 428	Isomap	0.754	0.781	0.782	**0.783**	0.737	0.657
	Eigenmap	0.752	0.763	0.765	0.778	0.731	0.668
MIMIC - 41071	Isomap	**0.880**	0.868	0.847	0.836	0.781	0.752
	Eigenmap	0.877	0.872	0.839	0.827	0.783	0.769
MIMIC - 41401	Isomap	0.853	**0.921**	0.910	0.846	0.809	0.797
	Eigenmap	0.803	0.859	0.865	0.835	0.855	0.802
NRD - 428	Isomap	0.923	**0.934**	0.932	0.912	0.925	0.892
	Eigenmap	0.908	0.927	0.926	0.920	0.900	0.895

(b) In-hospital mortality prediction

Note: (1) The best performance on each dataset is bold. (2) Medical knowledge: T_{ICD9}.

CHAPTER 3. ICU MORTALITY PREDICTION: CAN WE DO BETTER? A NEW MODEL BASED ON MACHINE LEARNING AND STOCHASTIC SIGNAL ANALYSIS TECHNIQUES

Shaodong Wang[1], Yiqun Jiang[1], Qing Li[1], and Wenli Zhang[2]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University

[2] Department of Information Systems & Business Analytics, Iowa State University

Modified from a manuscript submitted to *Journal of the Association for Information Systems*

Abstract

Intensive care units (ICU) provide critical care to patients with life-threatening illnesses and injuries. Worldwide, the high demand for intensive care imposes severe challenges on patient management and resource allocation. ICU mortality prediction can potentially inform rationing decisions on patients' medical needs and promote effective resource allocation. Thus, it can play an essential role in reducing the social burden of intensive care needs. Researchers have developed severity score systems, machine learning, and deep learning-based models for ICU mortality prediction. However, current methods have two major shortcomings: (1) The severity score systems depend upon laboratory tests and intensivists' assessments as the predicting variables and are questioned by their unsatisfactory performance, and (2) the machine learning and deep learning models suffer from feature extraction or interpretability issues that prevent them from having superior performance or wide application. In this work, by combining stochastic signal analysis and machine learning techniques, we propose a new ICU mortality prediction model capable of effectively extracting valid and interpretable patterns from the readily-available ICU bedside monitoring data with improved accuracy. To illustrate the efficacy of our model, we evaluate it on a large real-world multi-center ICU dataset. The proposed model outperforms baseline methods, including APACHE IV (the "golden standard" in ICU outcome

predictions), deep learning-based models (i.e., LSTM, GRU, CNN), statistical feature classification, and time series forecasting methods (i.e., ARMA, ARIMA) by a large margin. The innovative artifacts obtained from this study are salient to both the data science and healthcare communities.

Introduction

The intensive care unit (ICU) is a special department of a hospital that provides critical care medicine to patients who are at risk of, have, or are recovering from life-threatening illnesses or injuries. The ICU can provide patients with intensive monitoring, life support (e.g., airway, breathing, or circulation), resuscitation services, and end-of-life care (Medicine, 2021; Nates et al., 2016).

The burden of ICU-requiring care is massive. In the United States (US), there are 4 million ICU admissions every year, which accounts for 13.7% of hospital costs, 4.1% of national health expenditures, and 0.66% of the gross domestic product (Halpern & Pastores, 2010). ICU patients are extremely vulnerable to adverse outcomes due to their severe medical conditions (Pronovost et al., 2002). ICUs are the highest mortality units (8% to 19%, depending on patients' age, the number of comorbidities, and the severity of illness) in almost all healthcare institutions (Halpern & Pastores, 2010).

Researchers have long recognized the importance of ICU mortality predictions in alleviating the economic and healthcare burdens imposed by critical care needs (Knaus et al., 1991). It is crucial for assessing severity of illness and adjudicating the value of new treatments, interventions and health care policies (Pirracchio et al., 2015). The need for ICU mortality predictions has been magnified during public health threats like the COVID-19 pandemic because hospitals have been overwhelmed by an influx of patients, many requiring critical care.

To date[16], more than 308 million people have been tested positive worldwide. Five percent to twelve percent of diagnosed patients and up to thirty-three percent of hospitalized patients require intensive care (CDC, 2020; Myers et al., 2020; Richardson et al., 2020). Such public health threats pose serious challenges in ICU resource allocation, distribution, and optimization.

Due to the importance of ICU mortality predictions, considerable efforts have been invested in this research area over the past two decades. Generally, the extant methods can be categorized into three major types (see Table 3.4.): severity scoring systems (e.g., Acute Physiology and Chronic Health Evaluation Model (APACHE) (Knaus et al., 1985), Simplified Acute Physiology Score (SAPS) (Le Gall et al., 1993)), machine learning models, and deep learning models. However, current ICU outcome prediction models are not without limitations. For the traditional severity score systems, researchers question the score systems' accuracy and reliability (Becker & Zimmerman, 1996), healthcare practitioners complain about the overlong waiting on laboratory results as the predicting variables (Goswami et al., 2010; Winkelman et al., 1997), and critical care clinicians are dissatisfied with the fact that it relies on expert assessments as input (Reith et al., 2017). Hence, our *first research question* is how we can develop a new ICU mortality prediction model leveraging readily-available data with minimized requirements on the intensivists' expertise and having improved accuracy.

Over the past ten years, the implementation of electronic ICU technology has allowed large amounts of ICU patients' vital sign data[17] to be collected and streamed for real-time monitoring (Pollard et al., 2018), which has opened up new possibilities to propose more sophisticated ICU outcome prediction models. Rich dynamical patterns have been demonstrated

[16] WHO Coronavirus (COVID-19) Dashboard: https://covid19.who.int/ Accessed January 2022.

[17] Vital signs are a group of important medical signs (e.g.,body temperature, blood pressure, heart rate, and respiratory rate) that indicate the status of the body's life-sustaining functions. These measurements are taken to help assess the general physical health of a person, give clues to possible diseases, and show progress toward recovery. https://www.emergencyphysicians.org/ (Accessed April 2021)

in the time series of the monitoring data (Lehman et al., 2015). These dynamical patterns can be used to inform prognosis, provide early forecasts of life-threatening conditions, and predict patients' ICU outcomes (Lehman et al., 2015). However, these dynamic patterns are difficult to trace because of their large quantity and diversity of distribution. An increasing number of data scientists turn to machine learning and deep learning models. Although promising, both the machine learning- and deep learning-based models suffer from either feature extraction issues that hinder them from having superior performance, or low interpretability issues that prevent them from wide applications. Researchers and practitioners urge the next generation of ICU outcome prediction models to be more accurate, timely, and interpretable (Zimmerman & Kramer, 2014). Therefore, our *second research question* is how we can effectively extract valid and interpretable features from the readily-available ICU bedside monitoring data.

We find stochastic signal processing (Gray & Davisson, 2004) (e.g., Fourier transform and wavelet transform), a field of science concerned with processing and analyzing time-series data, a great tool to fill the gap. Stochastic signal processing techniques have been found to be particularly useful for extracting patterns from time-series signals which are normally described as aperiodic, noisy, intermittent, and transient (Addison, 2017). Signal processing techniques differ from other feature extraction methods in that (1) they examine the signal simultaneously in both time and frequency domains, so they have a strong feature extraction ability, (2) they have computational algorithms that reduce the computing time and complexity of large transformations, so the time-series data can be processed almost instantaneously and in real-time, and (3) the extracted patterns can be transformed back to and located on the original time-serial vital sign data, so the extracted features are far more interpretable than those extracted from black-box models.

In this study, following the design science paradigm (Hevner et al. 2004) and the recent information systems research on healthcare analytics (Aaron Baird et al., 2018), we propose a novel model which combines stochastic signal processing and machine learning techniques for ICU mortality predictions. The proposed method has three steps: (1) calculating frequency spectrums (a frequency spectrum of a signal is the range of frequencies contained by a signal) through various stochastic signal processing techniques, (2) extracting features from both frequency spectrums and time-series of vital signs, and (3) making ICU mortality predictions using machine learning classifiers. To demonstrate the effectiveness of our model, we evaluate it on a large multi-center ICU database - eICU (Pollard et al., 2018). The proposed model outperforms baseline methods, including APACHE IV, deep learning-based models (i.e., CNN, LSTM, GRU), statistical feature classification, and time series forecasting methods (i.e., ARMA, ARIMA) by a large margin. Meanwhile, as early as 3 hours after patients' admission to ICUs, the proposed method is capable of making more accurate predictions (AUC = 0.815) than APACHE IV (AUC = 0.750), while AHACHE IV makes predictions 24 hours after patients' admission. In addition, unlike traditional severity score systems, the proposed framework can make real-time predictions. The proposed model makes increasingly accurate predictions with patients' increasing length of stay. Moreover, the features extracted by the proposed model can be used to increase prediction accuracy in existing ICU outcome prediction models (e.g., APACHE IV).

The contributions of this work are two-fold. First, from the perspective of design science, we propose a new ICU mortality prediction model by combining stochastic signal processing and machine learning techniques, which provide a new perspective for temporal medical data analysis. Our method can add value to the literature for three reasons. (1) To our knowledge, this

is the first study to convert ICU patients' time-series of vital signs to frequency domains for mortality prediction. (2) We are among the first to explore how we can effectively extract features from the frequency domains for patient outcome prediction. (3) Our method is one of the first to use features from both frequency domain and time series as inputs to machine learning algorithms. The advantages of the proposed method include, (1) it requires only time series of vital sign data from ICU bedside monitors (rather than laboratory results and intensivists' assessments), allowing real-time predictions; (2) it significantly improves the performance of ICU mortality predictions; and (3) the extracted features are highly interpretable, which facilitates model adoption by critical care practitioners. Our model's superior performance and interpretability are difficult to achieve by traditional severity score systems and extant machine learning- or deep learning-based models. Second, from the perspective of data science for social good, the proposed model is expected to enlarge the social impact of ICU outcome prediction studies. As specified by previous studies (Barnato & Angus, 2004; Zimmerman & Kramer, 2014), the social impact and the value of an ICU mortality prediction model can be weighted by its reliability, availability, relevance, and resistance received from intensive care practitioners. (1) Reliability. An ICU outcome prediction model's reliability depends on its accuracy, generalizability, interpretability, and other factors. The proposed model improves the prediction accuracy substantially compared to the existing state-of-the-art baselines. It can be generalized to various ICU admission diagnoses and broader patient groups. Its prediction results are highly interpretable. Accordingly, the reliability of the model has been improved. (2) Availability. We use readily-available ICU bedside monitoring data for real-time mortality predictions. Thus, the prediction results are available in a timely fashion both to initiate and to continue intensive care. (3) Relevance. The predicted mortality probabilities are of the interests of ICU patients and are

expected to have an influence on clinician behaviors. Therefore, the prediction results are relevant. (4) Resistance. The predicting features of the proposed model are interpretable in human terms. As a result, the predictive model is expected to receive less resistance from practitioners. Overall, we believe the value of the proposed model is significant and measurable.

We follow the design science paradigm (Hevner et al. 2004) to structure the remainder of the paper. We first review the related literature and the shortcomings of existing studies on ICU mortality prediction. Second, we create and describe a new stochastic signal analysis and machine learning-based ICU mortality prediction model. Next, we implement and evaluate the proposed model and demonstrate its feasibility and implications. Finally, we conclude our work with a summary and directions for future research.

Related Work

The significance of ICU mortality prediction

The intensive care resources are limited and expensive. In the US, there are 4 million ICU admissions every year, whereas the number of ICU beds per 100,000 population is only 20.0 - 31.7 (Table 3.1). This number is much lower in other countries, especially in developing countries like China, Sri Lanka, and Zambia, ranging from 1.6 to 4.6 (Table 3.2.) (Prin & Wunsch, 2012). The care provided in the ICU is expensive. It is one of the largest cost drivers in the healthcare system in the US. Although the number of ICU beds only accounts for less than 10% of the hospital beds, the cost of the ICU explains nearly one-third of the total inpatient costs (Dasta et al., 2005; Kalb & Miller, 1989; Shorr, 2002; Sirio et al., 1994). Critical care treatments taking place in the ICU remain the most expensive healthcare interventions, with an estimated $80 billion spending every year, which consumes approximately 3% of all health care spending and nearly 1% of the gross domestic product (Halpern & Pastores, 2010).

Table 3.1. ICU Bed Availability in the US (Halpern & Pastores, 2010; Population Clock, 2021; Prin & Wunsch, 2012)

Year	ICU bed number	Population	ICU beds per 100,000 people
2000	88,252	281,421,906	31.4
2005	93,955	295,516,599	31.8
2012	62,564-99,164	312,818,676	20.0-31.7

Table 3.2. Selected ICU Bed Availability by Country with Per Capita Healthcare and Life Expectancy at Birth (Adapted from Prin & Wunsch, 2012)

Country	ICU beds per 100,000 people	Per capita healthcare cost	Life expectancy at birth
United States	20.0-31.7	$7,164	79
Canada	13.5	$3,867	81
Denmark	6.7-8.9	$3,814	79
Australia	8.0-8.9	$3,365	82
South Africa	8.9	$843	54
Sweden	5.8-8.7	$3,622	81
Spain	8.2-9.7	$2,941	82
Japan	7.9	$2,817	83
United Kingdom	3.5-7.4	$3,222	80
New Zealand	4.8-5.5	$2,655	81
China	2.8-4.6	$265	74
Trinidad & Tobago	2.1	$1,237	70
Sri Lanka	1.6	$187	71
Zambia	--	$80	48

Note: Estimates are pooled from multiple sources and involve different definitions of ICU beds, and different years of data. The per capita healthcare cost includes all public and private expenditures, not limited to critical care.

Due to the scarcity of intensive care resources, it is crucial to conduct ICU outcome evaluation and promote efficient use of such resources (Halpern & Pastores, 2010). The benefits of ICU mortality prediction include, first, ICU mortality prediction lays the scientific foundation for assessing the severity of illness (Becker & Zimmerman, 1996); second, it gives a standard for

adjudicating new treatments and policies (Pirracchio, 2016); third, it provides a way for comparing cohorts of ICU patients treated across different hospitals and countries (Becker & Zimmerman, 1996); next, it is an effective measure in allocating resources and determining levels of care (Lee et al., 2016); and last, it is helpful when discussing expected outcomes with ICU patients and families (Lee et al., 2016).

ICU mortality predictions and limitations of current studies

Critical care researchers have developed several severity scoring systems for ICU mortality prediction. The most reputable ones are the acute physiology and chronic health evaluation model (APACHE), simplified acute physiology score (SAPS), and mortality probability model (MPM) (Keegan et al., 2011). Major revisions of these models, APACHE IV (Zimmerman et al., 2006), SAPS III (Moreno et al., 2005), and MPM III (Higgins et al., 2007), have been published in 2006, 2005, and 2007, respectively.

Table 3.3. Apache IV Variables

Type	Factors	Advantages and disadvantages
Vital signs	temperature (C), mean arterial pressure (mmHg), heart rate (/min), respiratory rate	The vital signs can be obtained directly from the monitors, providing real-time information.
GCS results	motor, eyes, verbal	The GCS results strongly rely on manual examination from intensivists.
Lab test results	pO2 (mmHg), fiO 2(%), arterial pH, pCO2 (mmHg), Na+ (mEq/L), Ht (%), Bilirubin (mg/dL), creatinine (mg/dL), Urea (mEq/L), BSL (mg/dL), Albumin (g/L), WBC (x1000/mm3), sodium (mEq/L), hematocrit (Hct lab value), albumin (g/dL), glucose (mg/dL)	The sample analysis for these tests usually take hours to complete. It can take longer due to lab capacity and demand.
Chronic health condition (indicator)	CRF / HD, Lymphoma, Cirrhosis, Leukemia / Myeloma, Hepatic Failure, Immunosuppression, Metastatic Carcinoma, AIDS	The indicator function checks the conditions of chronic health issues. Chronic diseases such as heart disease, cancer, and diabetes are the leading causes of death and disability in the US. Thus the chronic health condition provides essential information in mortality prediction.

Table 3.3. Continued

Type	Factors	Advantages and disadvantages
Patients' information	age	The patients' information is usually available from multiple resources.
Admission information and diagnosis	pre-ICU length of stay (days), origin, readmission, emergency surgery, non-operative/post-operative	The admission information and diagnosis classifies patients to a more specific category, which helps the doctors adopt further instructions.
Others	mechanical ventilation, urine output (mL/24h)	

Among the extant severity scoring systems, APACHE IV is considered the "golden standard" in ICU outcome predictions (Keegan et al., 2012). It gives a patient a 0 - 286 score. Higher scores correspond to more severe diseases and higher risks of death. Recently, the reliability of these severity scoring systems, including APACHE IV, has been questioned by researchers and critical care practitioners (Zimmerman & Kramer, 2014). Meanwhile, independent investigators highlight the concerns about the prolonged waiting time of data collection and the assessments needed from subject-matter experts for calculating the severity scores (Knaus et al., 1991). Take APACHE IV as an example. The predicting variables include laboratory results and Glasgow Coma Scale (GCS) measures (see Table 3.3.). The lab results can take hours to days to obtain depending on the complexity of the tests (Goswami et al., 2010; Winkelman et al., 1997). The GCS scores require professional medical judgment, and their reproducibility has been questioned by researchers (Jain & Iverson, 2021; Teasdale et al., 2014). It takes APACHE IV 24 hours to collect all needed information for predicting (Sasaki et al., 2020). The limitation of the extant severity score systems motivates our *first research question*: can we propose a new ICU mortality prediction model leveraging readily-available data, with minimized requirements on intensivists' expertise and having improved accuracy?

Recently, the emergence of electronic ICU (eICU or tele-ICU) (Celi et al., 2001; Pollard et al., 2018) has opened up new possibilities for better ICU outcome prediction. The advantages

are twofold. (1) First, large quantities of time-series data that indicate the status of the body's vital signs (i.e., vital functions, such as heart rate, respiration rate, O_2 saturation, intracranial pressure, etc.) are continuously recorded via bedside monitors for ICU patients (Pollard et al., 2018). The vital sign data are real-time and commonly available, which neither require laboratory testing nor assessment from medical professionals. (2) Meanwhile, vital signs show patients' pathological states (e.g., the onset of sepsis) as well as their response to treatments. A growing body of literature has shown that many shared dynamical patterns can be identified across heterogeneous patients' cohorts (Lehman et al., 2015). These dynamical patterns can be used to inform prognosis, provide early forecasts of life-threatening conditions, and predict patients' ICU outcomes.

Many data scientists who work on ICU mortality predictions have explored the value of the eICU data by incorporating the real-time vital sign data as the input features of machine learning and deep learning models. S. Kim et al. (2011) are among the first to implement machine learning classifiers, such as support vector machines (SVM) and logistic regression, to predict the ICU mortality rate. Such models use patients' demographic data, lab results, and vital measurements in the first 24 hours of ICU admission. These models force the prediction to be a linear combination of features which leads to high interpretability. Then, to achieve higher accuracy, Davoodi & Moradi, (2018); Hsieh et al., (2018); Kong et al., (2020); Zhai et al., (2020) introduce ensemble models to the research area of mortality prognostication, including random forest, gradient boosting, and extreme gradient boosting. Although researchers keep introducing various classification models, the features extracted from the time series of vital signs are elementary. Only simple statistics of vital signs are included in these classification models, such as the minimum and maximum respiration rate or blood pressure. However, there is increasing

evidence that superior accuracy in ICU outcome prediction requires complex modeling with effective feature extraction methods (Zimmerman & Kramer, 2014). The properties of the vital sign data pose challenges to capturing meaningful dynamical patterns and revealing the relationships among these patterns. Later on, Caicedo-Torres & Gutierrez, (2019); S. Y. Kim et al., (2019) take the entire time series of vital signs as the input of a convolutional neural network (CNN, deep learning). CNN models can summarize useful patterns from the vital sign data. Recently, Thorsen-Meyer et al., (2020) take the vital sign data into long short term memory (LSTM, deep learning) to infer ICU outcome, which takes advantage of the temporal information of vital signs. However, the deep learning models work as a non-interpretable black-box because their behavior cannot be comprehended, even if we know their structures and weights. Although existing deep learning-based research improves the performance of ICU mortality prediction to a certain extent, such models are not favorable for medical professionals. Extant machine learning and deep learning models' imperfections motivate us to answer the *second research question*: how to effectively extract useful and interpretable features from the time series of vital signs?

Table 3.4. Literature Summary of Mortality Prediction at ICU

Literature	**Types of ICU mortality prediction models**	**Models**	**Pre-admission character-istics**	**Lab results**	**Glasgow Coma Scale (GCS)**	**Vital signs**			**Hours of data needed**
						Statistics features	**Time series**	**Features extracted by stochastic signal processing**	
Moreno et al., (2005)	Traditional severity scoring system	SAPS III		Y	Y	Y			24 H
Higgins et al., (2007)		MPM III	Y		Y	Y			1 H
Zimmerman et al., (2006)		APACHE IV		Y	Y	Y			24 H

Table 3.4. Continued.

Literature	Types of ICU mortality prediction models	Models	Pre-admission character-istics	Lab results	Glasgow Coma Scale (GCS)	Vital signs			Hours of data needed
						Statistics features	Time series	Features extracted by stochastic signal processing	
Kong et al., (2020)	Machine learning model	RF, GBM, LR		Y	Y	Y			24 H
S. Kim et al., (2011)		DT, SVM, NN, LR		Y	Y	Y			24 H
Zhai et al., (2020)		SVM, GBDT, XGBoost, LR		Y	Y	Y			6 H
Hsieh et al., (2018)		RF, LR, NN, SVM		Y	Y	Y			24 H
Davoodi & Moradi, (2018)		NB, DT, GB, DBN, D-TSK-FC		Y		Y			48 H
Thorsen-Meyer et al., (2020)	Deep learning model	LSTM	Y		Y		Y		Real-time
S. Y. Kim et al., (2019)		CNN					Y		24 H
Caicedo-Torres & Gutierrez, (2019)		CNN		Y	Y		Y		48 H
Proposed model		SVM, NN, LR, RF				Y	Y	Y	Real-time*

Note:
Lab results: laboratory tests to measure vital body functions after ICU admission, e.g., arterial blood gas test (ABG) to measure oxygen and carbon dioxide levels in the blood (Frassica, 2005).
*: The proposed model can make more accurate predictions than APACHE IV with only 3 hours of ICU bedside monitoring data, whereas APACHE IV makes predictions after 24 hours of admission.

There are multiple statistical forecasting models to analyze the time-series data, such as moving average (AR), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) (Wei, 2006). However, these time series models are not created for classification tasks or probability estimations. In order to make ICU outcome predictions using time series data, researchers regard the coefficients of the time series models as input features and train machine learning classifiers (Carden & Brownjohn, 2008; Zhang, Ji, et al., 2017). Nevertheless, these methods are not ideal for the time series of vital sign data from ICUs. The

order of a time series model has to be determined by the statistical characteristics for a specific time series (e.g., one time series of vital signs for one patient) (Wei, 2006). In healthcare-related studies, researchers usually treat model orders as hyperparameters and determine them through experiments and Akaike information criterion (Anderson et al., 1998; Zhang, Ji, et al., 2017; Zhang, Liu, et al., 2017). But a fixed order of time series model is required for all patients to ensure input features have the same dimension for the classification task, which limits the predictive power of the time series forecasting models (Z. Liu & Hauskrecht, 2017).

Stochastic signal analysis techniques for feature extraction

We find the combination of stochastic signal processing techniques (Little, 2019) and interpretable machine learning models a great strategy to overcome the above-mentioned shortcomings. Stochastic signal analysis, which treats time-series signals as a stochastic process, is used to process and analyze the time-series data using their statistical properties. In signal processing, a signal is a function that conveys information about a phenomenon (Little, 2019). Patients' vital signs can be mathematically defined as signals (i.e., functions) that indicate the status of the body's vital functions. Stochastic signal processing provides extremely efficient ways to extract meaningful features for vital signs by transforming time-series data into frequency spectrums. The frequency spectrum is the frequency-domain representation of the signal, which shows "how much of the signal is present among each given frequency band" (DeepAI, 2019). Specifically, vital signs that are continuously collected by the eICU systems are in the time domain, while their strengths of the periodicity (also known as frequency spectrum) are in the frequency domain. In the frequency domain, the amplitude (i.e., the maximum extent of an oscillation) for a specific frequency is a number, which reflects the magnitude of the change in the signal (Zhou, 2013). As shown in Figure 3.1., the left-side plots show Signals (a), (b), and (c) with three different frequencies, and the right-side plots show the signals' frequency

spectrums, which reflect the strength of oscillations of the signals in the left-side plots. In the time domain, it is difficult to find Signal (c) is a mixer of Signals (a) and (b), but their relationship can be revealed in frequency spectrums obviously. Additionally, it is intuitive that stochastic signal processing can disclose a patient's health conditions. For example, if a patient's heart rate undergoes a rapid change (e.g., rapid heart rate increase after a treatment (Bergfeldt et al., 2017)), it may indicate adverse cardiac events (Kannel et al., 1987). This kind of pattern can be easily detected by signal processing in frequency spectrums, whereas it is not easy to detect by simple statistics and classical machine learning algorithms. Deep learning models may have the ability to identify such patterns. However, the identified features cannot be traced back to the original signals, which significantly decreases the interpretability of the deep learning models.

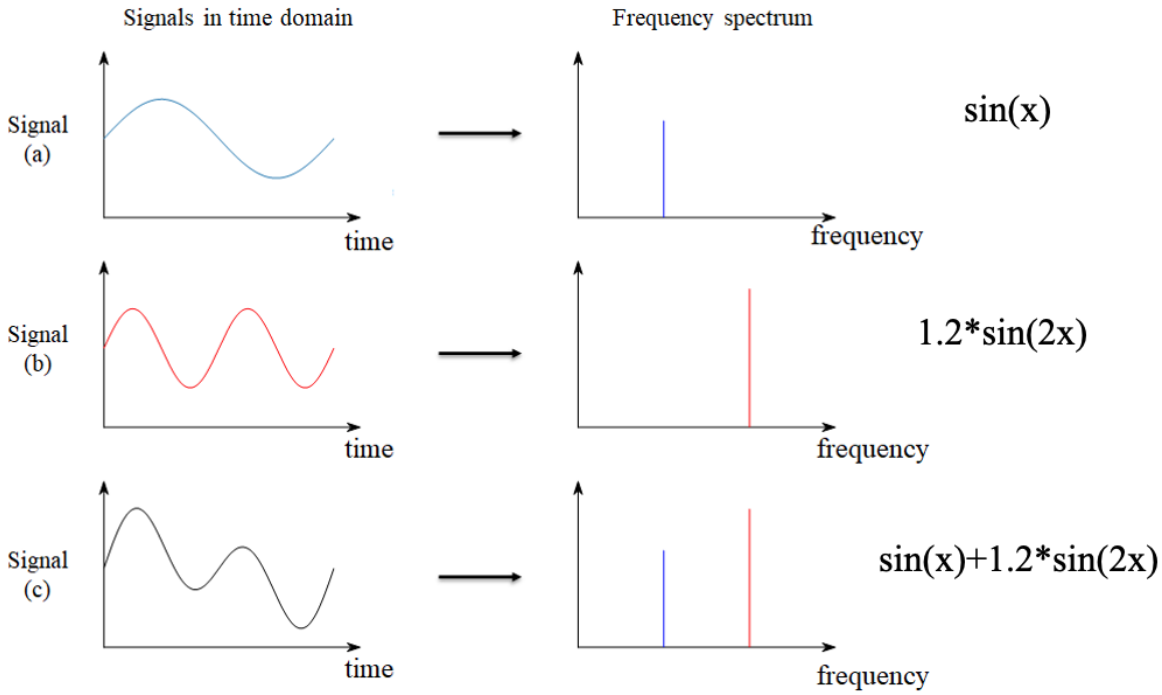


Figure 3.1. An Example of Frequency Spectrum.

In healthcare-related research, stochastic signal analysis techniques have been applied for different purposes and in different settings. Medical diagnosis signals, such as

electrocardiograms (ECG), electroencephalograms (EEG), and photoplethysmogram (PPG), can be processed and analyzed by signal processing techniques. For example, Prasad & Parthasarathy (2018) propose an algorithm to detect cardiovascular abnormalities from ECG data with signal analysis technique. Wang et al., (2014) extract frequency power features from ECG for the classification of obstructive sleep apnoea. Signal analysis technique has also been used to analyze EEG on human movements and compare the results with EEG recording during a resting state condition (Wairagkar et al., 2019). Besides, Kumar et al., (2017) have classified normal and epileptic patients by extracting signal processing features from EEG and training machine learning models. Additionally, some researchers developed signal processing-based methods to determine oxygen saturation with PPG (Addison, 2017). These studies show the potential of adapting stochastic signal processing techniques in healthcare analytics research. However, to our best knowledge, we are the first to combine machine learning and stochastic signal processing techniques for ICU mortality prediction. Meanwhile, the ICU bedside monitoring data have never been systematically analyzed by signal processing techniques. The proposed method provides a new way to extract predictive variables for ICU outcome prediction for improved prediction results.

To summarize, the deficiencies of existing ICU mortality prediction methods, coupled with the challenges associated with developing more accurate, timely, and interpretable ICU outcome prediction models, motivate us to propose a new signal processing- and machine learning-based model. The design science paradigm provides a good foundation for our study. Design science is an outcome-based research methodology (Nunamaker et al., 1990). According to its definition, a design is both a product and a process (Hevner et al., 2004). The product is an artifact that can be broadly defined as design logic, models, methods, constructs, instantiations,

new design and developments models, and implementation processes or methods (Gregor, 2002; Gregor & Hevner, 2013; J. Ellis & Levy, 2010; March & Smith, 1995; March & Storey, 2008). The process is a sequence of expert activities composed of the procedures taken to develop and evaluate the artifact (March & Smith, 1995). In this study, the artifact we intend to deliver is a model consisting of methods and instantiations that can be used to (1) effectively extract valid and interpretable features from readily available ICU bedside monitoring data, and (2) predict ICU mortality with improved accuracy. Nomenclature (Optional Section, if Included Abide by the Following)

Research Design

We propose a novel model to effectively extract interpretable features with strong predictive power from the time series of vital signs for ICU mortality prediction. As shown in Figure 3.2., the proposed model includes three steps: (1) frequency spectrum extraction, (2) feature extraction, and (3) mortality prediction. For each vital sign collected from ICU bedside monitors, v_t, $t = 1, 2, \ldots, N$, we first extract the frequency spectrums with various signal processing techniques. Then we extract signal processing features from the frequency spectrums. Meanwhile, we take statistical features from the time series of vital sign data. Finally, both signal processing features and statistical features are used to predict ICU mortality using machine learning algorithms.

Frequency spectrum extraction

In the first step, we calculate the frequency spectrums for patients' vital sign v_t, using fast Fourier transformation (FFT), power spectrum density (PSD), auto-correlation (AC), and wavelet transformation (WT).

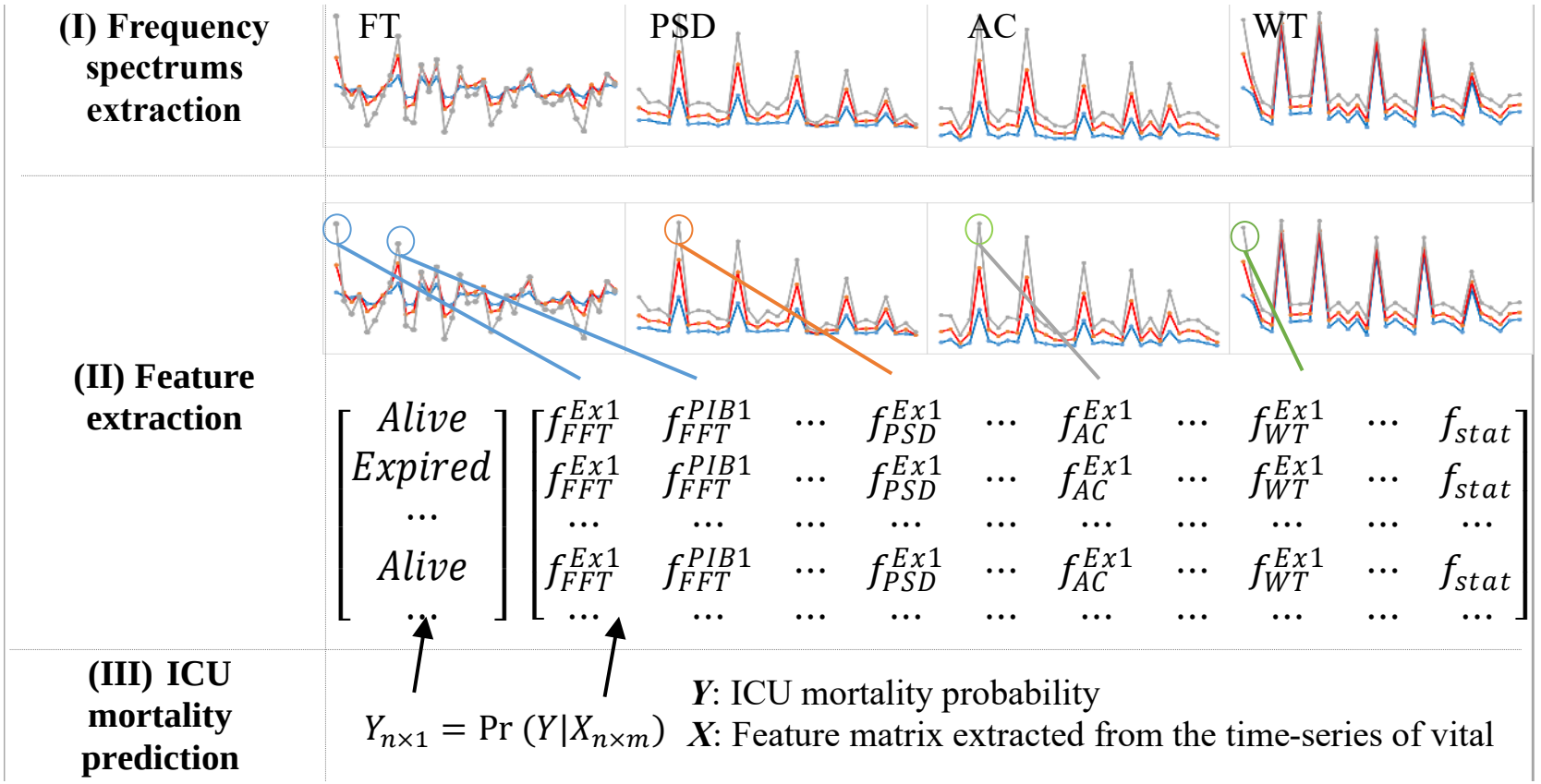


Figure 3.2. Flow Chart of the Proposed Method

Using FFT (Bloomfield, 2004), any time series can be decomposed into a series of simple sinusoids of different frequencies. The FFT estimates the coefficients of each sinusoid for a given time series. The PSD describes the distribution of the power of a time series over frequency (Woyczyński, 2019). We include PSD because researchers believe that FFT is great at analyzing vibration when there are a finite number of dominant frequency components, but PSDs can be used to characterize random vibration signals (Little, 2019). AC is the correlation of a time series with the lagged version of itself over successive time intervals (Broersen, 2006). As a signal processing tool, AC is usually used to detect repeating patterns, such as periodic signals hidden in noisy data (Broersen, 2006). The outputs of the above three signal processing techniques (FFT, PSD, AC) provide abundant information about the frequencies (frequency domain) in time series data. Still, the information of frequencies' time location (time domain) is absent. To overcome this problem, we include WT as well. The WT decomposes a time series into a series of wavelets with different scales at different time points (Addison, 2017). Thus, the outputs of WT present both the strength and location of frequencies (i.e., patterns from both the frequency and the time domains) in the time series.

A frequency spectrum is denoted as $F(\omega)$, where ω is the parameter of the signal processing. Specifically, ω indicates frequency in FFT and PSD, scale and shift parameters in WT, and time difference in AC. For FFT, PSD, and AC, the frequency spectrum of a vital sign v_t is a vector, $[F(\omega_1), F(\omega_2), \ldots, F(\omega_t)]$. For WT, the frequency spectrum is a matrix, $[[F(\omega_{1,1}), F(\omega_{1,2}), \ldots, F(\omega_{1,t})], [F(\omega_{2,1}), F(\omega_{2,2}), \ldots, F(\omega_{2,t})], \ldots, [F(\omega_{s,1}), F(\omega_{s,2}), \ldots, F(\omega_{s,t})]]$. The frequency spectrum depicts the periodicity strength of the vital sign, which offers important information about the patient's health condition. All frequency spectrums converted from patients' time series of vital signs form a space $X_{n\times w}$ (n: number of patients, w: the number of

frequency spectrums). The following subsections introduce the signal processing transfer functions in our research setting.

Fast Fourier Transform (FFT)

The FFT is an algorithm that computes the Fourier transformation of a signal. It uses sinusoids of different frequencies to represent signals in the time domain, which reveals periodicity in time-series data and indicates the frequencies of these periodical components. The resulting signals after the FFT are frequency spectrums $F_{FFT}(\omega) = \sum_{t=1}^{N} v_t \cdot e^{-i2\pi t\omega}$, where v_t is the vital sign, and ω is the frequency at which a complex sinusoid is computed. A major advantage of FFT to other frequency domain transform methods is its computational efficiency.

Power Spectral Density (PSD)

The PSD characterizes the average power[18] at a frequency ω in the signal, providing useful information in a signal's frequency domain (Stoica & Moses, 2005). The PSD $F_{PSD}(\omega)$ is calculated by $F_{PSD}(\omega) = \sum_{k=-\infty}^{\infty} r(k)e^{-i\omega k}$, where $r(k)$ is the autocovariance sequence of v_t and $r(k) = E\{v(t)v^*(t-k)\} = \sum_{t=k+1}^{N} v_t\, v_{t-k}$. Here $v^*(t-k)$ denotes the complex-conjugate transpose of $v(t-k)$. The PSD of a signal analyzes the distribution of the power over the frequencies composing the signal. Specifically, for time series data, the PSD uses the signal's ACs to measure the power. Compared to FT, which obtains the amplitudes of a signal's frequency components, the PSD of the signal describes the power present in the signal as a function of frequency, per unit frequency (Grami, 2016).

Auto-Correlation (AC)

An AC coefficient A_k measures the correlation between a signal and its delayed version with lag k, which can be calculated by $F_{AC}(\omega) = A_k = \sum_{t=1}^{N-k} v_t v_{t+k}$. It reveals the influence of

[18] The power of a signal is the sum of the absolute squares of its time-domain samples divided by the signal length (Grami, 2016). It is a measure of signal strength.

the previous signal on the following signal in the sequence. When the signal does not repeat the sequence of values regularly after a fixed length of time, the AC coefficients tend to be small, which indicates the fluctuation of a vital sign. Otherwise, the AC coefficients tend to be large, which represents the stable health status of the patient.

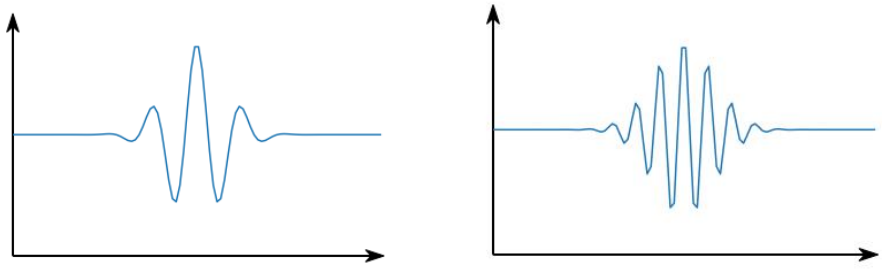
Wavelet Transform (WT)

The WT analyzes signals with a dynamic frequency spectrum, providing a high resolution in both the frequency domain and the time domain. The WT of the vital sign signal v_t is expressed by $F_{WT}(\omega) = F(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} v(t)\psi(\frac{t-b}{a})dt$, where $\psi(\cdot)$ is the mother wavelet (Addison, 2017). A wavelet is a wave-like oscillation. Parameters a, b define the scale and the time location of the wavelet, correspondingly. The scale defines how stretched a wavelet is, and the location defines where the wavelet is positioned in time. One widely used wavelet is the real Morlet wavelet, which is defined as $\psi(\frac{t-b}{a}) = \frac{1}{\pi^{1/4}} e^{i2\pi f_0[(t-b)/a]^2}$. Figure 3.3. shows examples of the Morlet wavelet function with the scales of a=4 and a=8, respectively. In this research, we use different types of wavelets to generate various frequency spectrums, including the Morlet wavelets, the complex Morlet wavelets, and the Mexican wavelets. We adopt the Morlet wavelets and the complex Morlet wavelets because they are closely related to human perception of hearing and vision (Daugman, 1985). We use the Mexican hat wavelets because they are frequently employed as broad-spectrum source terms in WT analysis (Addison, 2017).

Feature extraction

Although the frequency spectrums, $X_{n\times w}$, extracted from time-series data represent useful periodicity information of patients' vital signs, they are not ideal to use as input features of machine learning classifiers. This is because the frequency spectrums are of high dimensionality, which increases the complexity of machine learning models and decreases

classifiers' predictive power (R. Liu & Gillies, 2016). Therefore, we extract the most representative periodicity information as classifiers' input features by taking the (1) relative extrema (i.e., local maxima and minima), and (2) power in band from the frequency spectrums. Besides, we take various statistical features from the time series of vital signs. The extracted features form a new feature space, $X_{n\times l}$ (n: number of patients, l: the number of features), where $l \ll w$. We systematically evaluate the relative importance of extracted features and select the most important features for ICU mortality predictions. The selected features, $X_{n\times m}$ (n: number of patients, m: the number of selected features), are the input of the proposed ICU mortality prediction model.



(a) Morlet wavelet with the scale of 4 (b) Morlet wavelet with the scale of 8

Figure 3.3. Morlet Wavelets with Different Scales

Relative extrema

Based on the extracted frequency spectrums from the previous step, we extract the frequency spectrums' positions and values of the local maxima and local minima as the ICU mortality predicting features. More formally, we extract the (1) value of the frequencies where the oscillations, ω_i^*, occur and (2) their corresponding amplitudes, $F(\omega_i^*)$, as predictive features (see examples in Figure 3.4.). Specifically, the relative extrema, $(\omega_i^*, F(\omega_i^*))$, is the local maximum (or local minimum). Namely, $F(\omega_i^*) \geq F(\omega)$ (or $F(\omega_i^*) \leq F(\omega)$), for all ω within a threshold distance ε on the frequency spectrum, where ε is a small positive value. We extract one

relative extrema point ω^* within each distance range $(-\varepsilon, \varepsilon)$. Note there are multiple ω^*on the entire frequency spectrum, $[\omega_1^*, \omega_2^*, \ldots, \omega_i^*, \ldots, \omega_t^*]$, where t is the number of extrema. After we find all relative extrema $F(\omega^*)$ satisfying the requirement, we obtain a vector $u = [F(\omega_1^*), F(\omega_2^*), \ldots\ldots, F(\omega_m^*)]$. The top n maximums are defined as the largest n values on u (accordingly, the top n minimums are defined as the smallest n values on u). When there are less than n elements in u, we adopt all available relative extrema (i.e., m in total) as features and include $n - m$ missing values. The maxima reflects the periodicities that the time series has, while the minima reflect the periodicities that the time series lacks. The relative extrema, both local maxima and local minima, on the frequency spectrums are critical characteristics of patients' vital signs and reflect ICU patients' health conditions.

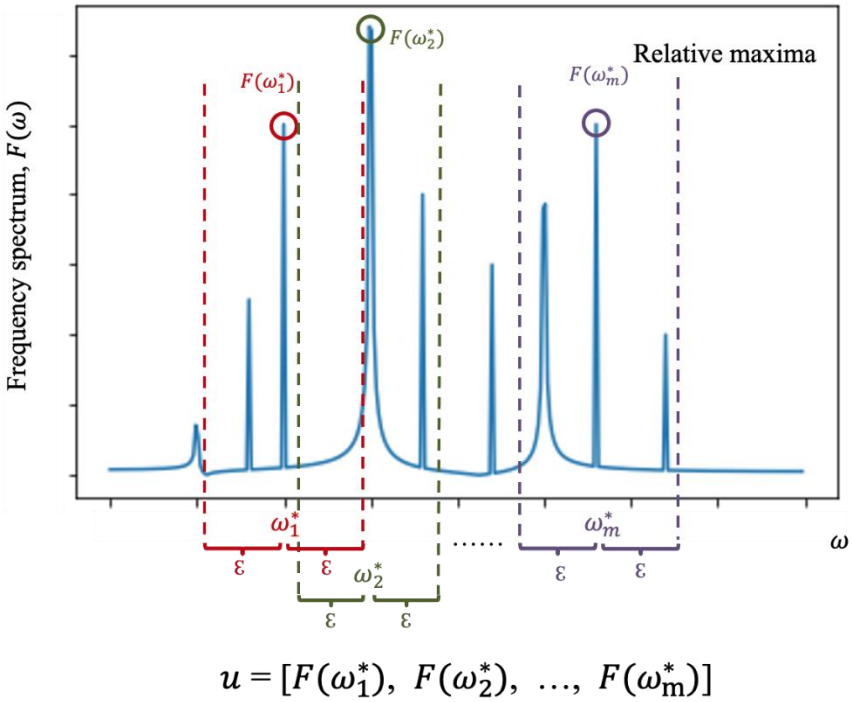


Figure 3.4. An Example of Relative Maxima of the Frequency Spectrum

Power-in-band

The power-in-band feature is the sum of the total power (please see Appendix A for the definition of power) within a frequency band (i.e., frequency range). With a specified center frequency ω_c and bandwidth ω_{bw}, we can derive the low and high bounds, $\omega_c - \omega_{bw}$ and $\omega_c + \omega_{bw}$, respectively, of the frequency band. The power-in-band feature, denoted by *PIB*, is $\sum_{\omega=\omega_c-\omega_{bw}}^{\omega_c+\omega_{bw}} F(\omega)$ (see Figure 3.5.).

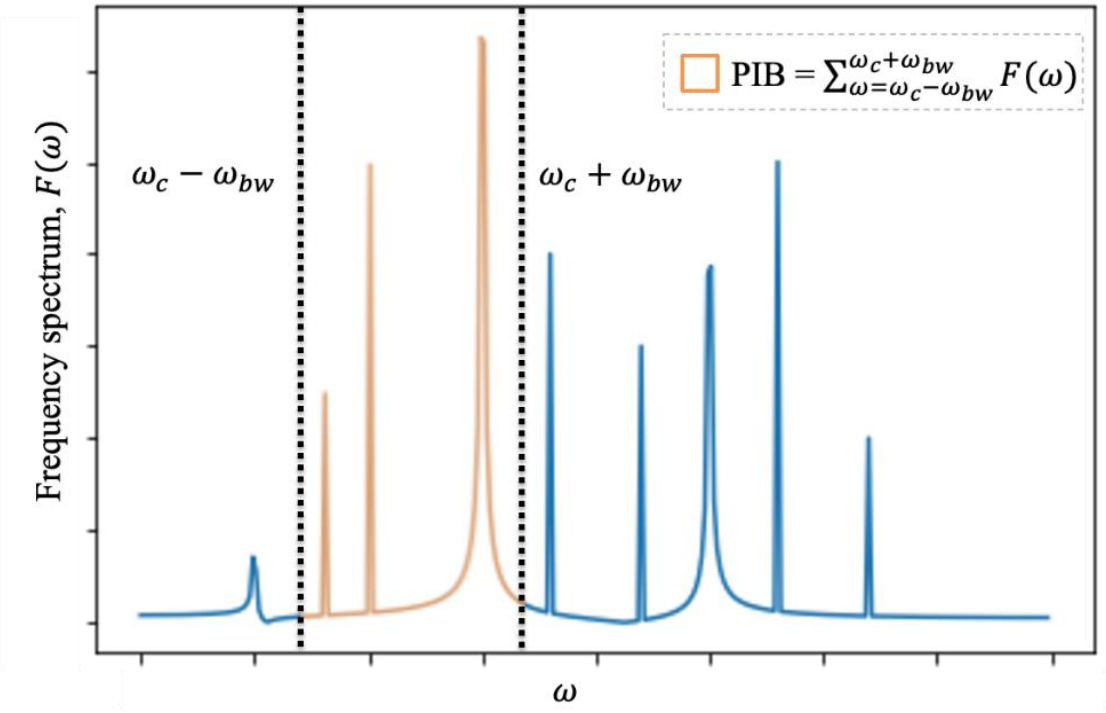


Figure 3.5. An Example of the Power-in-Band Feature of the Frequency Spectrum

The power-in-band summarizes the strength of the signal in the frequency band by computing a single number. The benefits of using power in band features are two-fold. First, the power-in-band feature summarizes the contribution of the given frequency band to the overall strength of the signal, which contains important patterns of patients' vital signs (please see the example in Appendix A). Second, power-in-band is a simple yet powerful feature extraction

method for ICU mortality prediction. Power-in-band features are easy to extract and use. In practice, we compute the summation over the different segments of a vector $[F(\omega_1), F(\omega_2), \ldots, F(\omega_t)]$ (i.e., the vector represents the frequency spectrums transformed from a vital sign). Power-in-band features are effective in machine learning prediction as well. The frequency/time representation converted from time-series of vital signs are usually of high dimensionality, making them unsuitable for use as classifiers' inputs. Power-in-band extract the key characteristics from the frequency spectrums, resulting in a single number that describes a specific aspect of the frequency spectrums.

Statistical features

In signal processing, summary statistics are used to outline and provide information on signals. For example, the mean of a signal is an estimate of the center of the whole signal. The standard deviation and variance measure the spread extent of the signal from its average value. Take a patient's heart rate as a simple example. A normal resting heart rate of an adult is between 60 and 100 beats per minute (Kannel et al., 1987). Hence the mean of the normal heart rate should also be in this range, and the standard deviation should be less than 7. Abnormal heart rates can be an indicator of a deteriorating health condition. In this research, we calculate the standard deviation, variance, mean, median, quantiles, the first and the last of ICU patients' vital signs as features for ICU mortality prediction.

Meanwhile, the extreme values of the time series of vital signs usually indicate unfavorable health conditions as well. To detect the extreme values on the time series data, we create a series of moving windows. Each window has k observations. With the k observations, we calculate the mean and standard deviation (see Figure 3.6.). The observations that are not within three standard deviations of the mean are treated as extreme values (M. X. Cohen, 2008). Intuitively, the observations above and below the three standard deviations can be considered as

sudden rise and sudden drop of the vital signs, respectively. Then we take the "top n" extrema from the moving windows of vital signs, denoted as $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, where the x^* is the event time and the y^* is the value of the extrema. When the vital sign in the moving window has less than n relative extrema (i.e., less than n data points are above and below the three standard deviations of the data in a given moving window), we set all available extreme points (i.e., m data points) as features and include $n - m$ missing values. Parameter selection for the parameter n has been discussed in Appendix A.

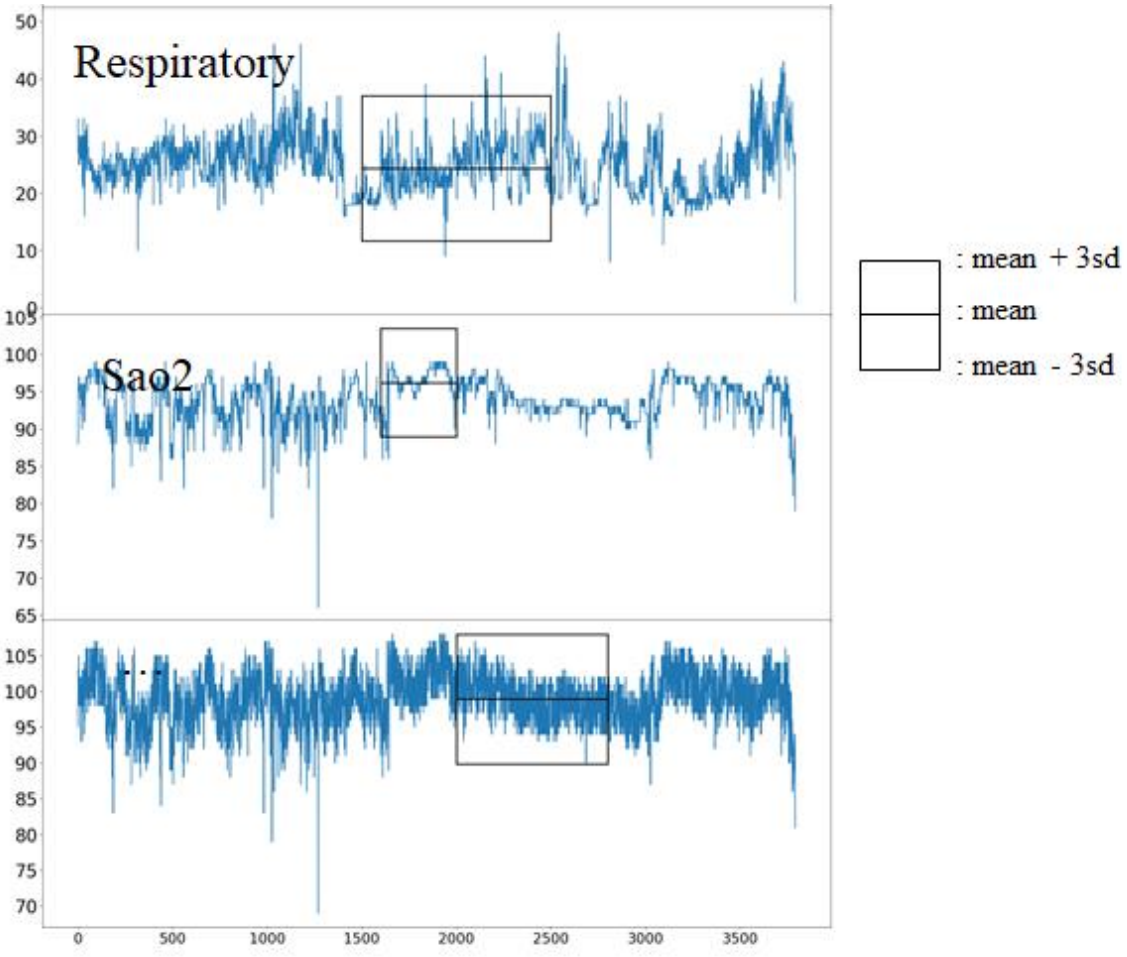
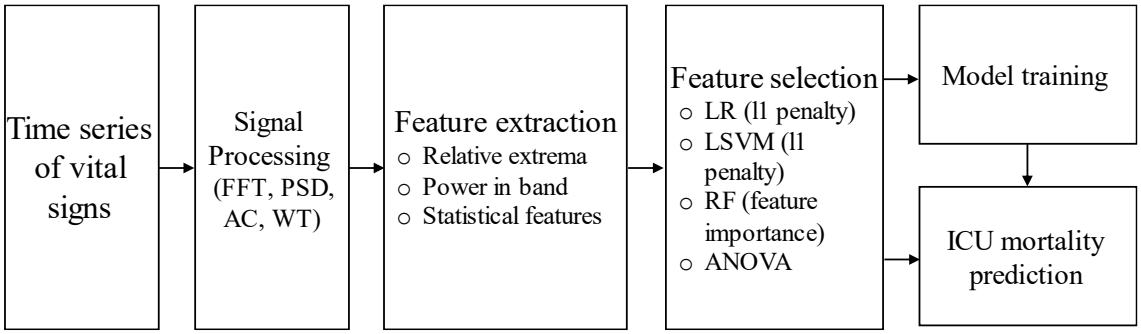


Figure 3.6. Moving Window on the Time Series of Vital Sign

ICU mortality prediction



Note: Fast Fourier Transform (FFT), Power Spectral Density (PSD), Auto-Correlation (AC), Wavelet Transform (WT), Logistic Regression (LR), Linear SVM (LSVM), Random Forest (RF), and Neural Networks (NN)

Figure 3.7. ICU Mortality Prediction Framework

We define the mortality prediction as a probabilistic classification problem $Y = Pr(Y|X)$. X denotes the input space, where

$$X = \begin{bmatrix} f_{FFT}^{Ex1} & f_{FFT}^{PIB1} & \cdots & f_{PSD}^{Ex1} & \cdots & f_{AC}^{Ex1} & \cdots & f_{WT}^{Ex1} & \cdots & f_{stat} \\ f_{FFT}^{Ex1} & f_{FFT}^{PIB1} & \cdots & f_{PSD}^{Ex1} & \cdots & f_{AC}^{Ex1} & \cdots & f_{WT}^{Ex1} & \cdots & f_{stat} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ f_{FFT}^{Ex1} & f_{FFT}^{PIB1} & \cdots & f_{PSD}^{Ex1} & \cdots & f_{AC}^{Ex1} & \cdots & f_{WT}^{Ex1} & \cdots & f_{stat} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

The input space X includes signal processing features and statistical features obtained from the previous steps. The output space is defined as $Y = \{1: expired, 0: alive\}$. The machine learning models provide a mapping function, $f(x)$, which maps the input data X to the out space Y and produces the ICU mortality prediction results $Pr(Y|X)$. Figure 3.7. shows the model training and prediction process.

In machine learning, there are different classifiers. Usually, no single classifier always outperforms the others in different research settings. In this study, we selected four classifiers that are widely used in healthcare analytics, including linear support vector machine (LinearSVM), logistic regression (LR), random forest (RF), and fully connected neural networks

(NN). The training process of these machine learning models aims to minimize the loss function defined as $R(f) = E[L(Y, f(X))] + \lambda J(f)$ by finding an appropriate $f \in F$, where $F = \{f|Y = f(X)\}$is the mapping function space, $E[L(Y, f(X))]$ is the expected loss obtained from the data, and $J(f)$ represents the complexity of the model. $\lambda J(f)$ adds a penalty to complex models to avoid overfitting. The solution of $min_{f \in F}\, E[L(Y, f(X))] + \lambda J(f)$ offers an optimal parameterization of the chosen model. We discuss the specifics of how the classifiers are implemented in Appendix A.

To obtain the best prediction results, we also conduct feature selection techniques before feeding the entire input feature space $X_{n \times l}$ to machine learning classifiers. Since there are more than one methods for feature importance assessment, and there is no consensus on which method works best in a given situation, we compare the feature selection results using various feature selection methods, including logistic regression (LR, $l1$ penalty), linear support vector machine (LinearSVM, $l1$ penalty), random forest (RF, average feature impurity as feature importance measures), and ANOVA test. LR and LinearSVM select features automatically with the $l1$ penalty (J. Cohen et al., 2013; Guyon et al., 2002). Classifiers with $l1$ penalties are useful for feature selection because many of the estimated coefficients are zero, and the important features are those with non-zero coefficients. RF provides an importance score (i.e., average feature impurity, e.g., entropy or Gini index) for each feature, based on which we select the top features with the highest scores. ANOVA is a statistical test that selects features based on the F-value (Bejani et al., 2014). The selected features aim to help the machine learning classifiers to achieve improved accuracy. Formally, for a set of features S, the feature selection method finds the optimal subset s of S by minimizing the loss function: $min\ _{s \subseteq S}||Y - Pr(Y|X, X \in s)||$, where $||\cdot||$ is the error estimation function. The classification mapping function Pr is decided by the

feature selection methods. The selected features $X_{n\times m}$ are the input of the ICU mortality prediction model, where $m < l$.

Implementation and evaluation

In the design science paradigm, the evaluation of an artifact provides feedback information and a better understanding of the problem in order to improve both the quality of the design product and the design process. Our evaluation plan and procedures are summarized in Figure 3.8.

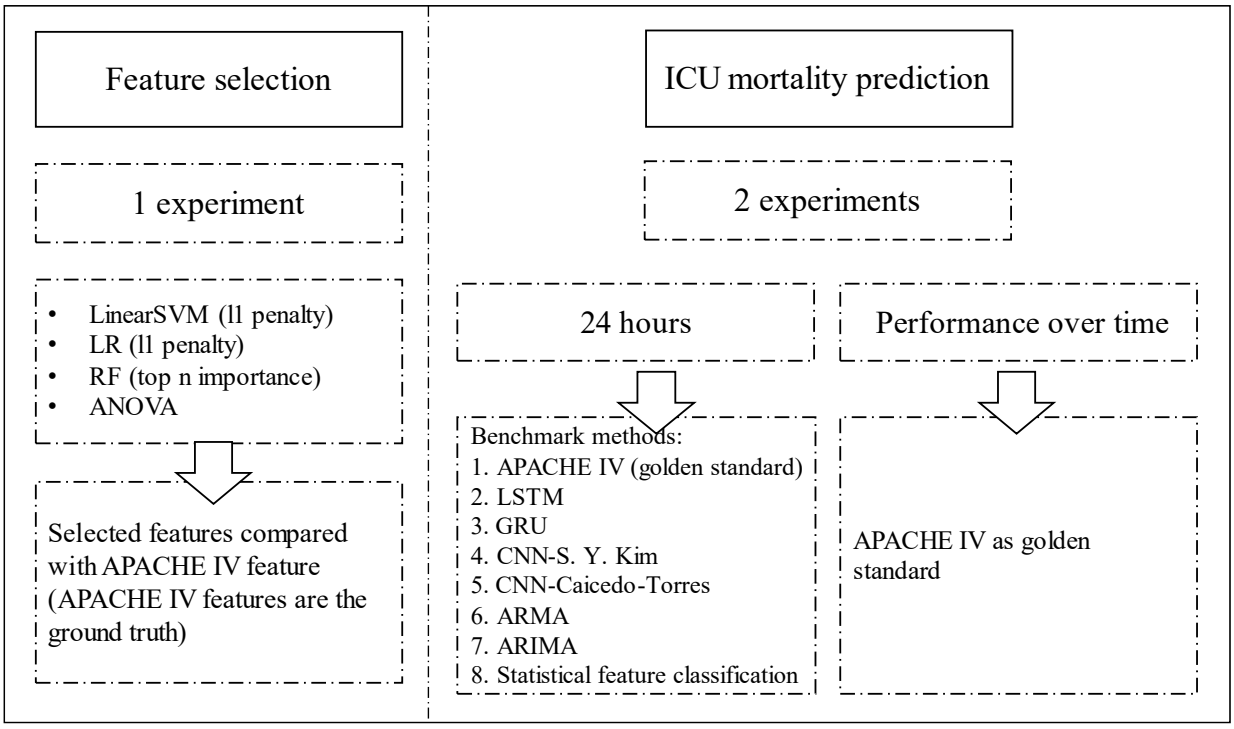


Figure 3.8. Evaluation Plan

In this study, two main parts need to be evaluated: (1) ICU mortality predicting features, and (2) ICU mortality prediction results. For feature assessment, we identify important features that have a significant contribution to the ICU mortality prediction. The features generated and selected by the proposed method are the informative representations of patients' health conditions in the ICU, which can be utilized in the machine learning classifiers for ICU mortality

prediction. In the predicting performance assessment, we evaluate the machine learning models. We select the classifier that performs the best and compare the results to state-of-the-art baselines. As stated earlier (see Section 3.3), the machine learning classifiers we adopted for this study are Linear Support Vector Machine (LinearSVM), Logistic Regression (LR), Random Forest (RF), and Neural Network (NN).

To show the effectiveness of the proposed method, we take the ICU-admitted congestive heart failure patients (refer to them as heart failure patients in the rest of the paper) as a research case. We choose heart failure as a research case because heart failure patients make up a significant portion of patients at risk in the ICU every year (Poppas & Rounds, 2002), and it is the 4th prevalent disease in the eICU database (Pollard et al., 2018). To test the model's generalizability, we also evaluate our method on two other diseases, sepsis pulmonary (SP) and sepsis renal/UTI (including bladder) (SI), which are the most prevalent admission diagnoses in the eICU database. The results are reported in Appendix B.

Analysis, Results, and Discussion

In this section, we discuss the implementation process, the evaluation results of our methods, and the interpretation of the results.

We conducted three experiments. (1) In the first experiment, we assessed the features generated by our method. (2) The second experiment used patients' first 24 hours (H) of vital sign data after their ICU admission to evaluate the proposed method's performance. We compared our methods with eight baselines. Twenty-four hours was a critical time point because APACHE IV, the "golden standard" in ICU outcome prediction, made forecasts based on patients' first 24 hours of data. (3) In the third experiment, to examine our method's real-time prediction ability, we used data from different time periods, including 3H, 6H, 12H, 24H, 48H,

and 72H. Our method was capable of making real-time predictions without laboratory results or intensivist assessment because it only required ICU bedside monitoring data.

Data description

The experiment data was from the eICU database (Pollard et al., 2018). It was a multi-center ICU database with minute granularity data for more than 200,000 ICU admissions. These patients were monitored by the eICU program[19] from a large number of hospitals in the US. The eICU database made it possible to evaluate ICU outcome prediction models from a multi-center perspective (Pollard et al., 2018). We extracted records with both APACHE IV scores and congestive heart failure as the ICU admission diagnosis. The extracted dataset contained 4,801 patients and 5,282 admissions (including re-admissions), among which 5.66% of the patients expired in the ICU, and 94.34 % survived. Patients' demographic information was shown in Table 3.5.

The proposed method was established on the time series of vital signs. The vital signs in the eICU database were consistently interfaced from bedside monitors, which were readily-available and updated in real-time. The vital signs were generally interfaced as one-minute averages, and archived as five-minute median values (Pollard et al., 2018). The vital signs we adopted were periodic time-series data with a time interval of five minutes. To reduce the impact of missing values, we mainly considered the following vital signs that were measured for more than 50% of patients in our dataset: sao2 (oxygen saturation), heart rate, respiration, temperature, st1, st2, and st3 (estimated ST-segment level x of the ECG $x \in \{1,2,3\}$), as shown in Table 3.5. Sao2 was useful in understanding the oxygen-carrying capacity of hemoglobin. It is particularly important in patients' care and management because low oxygen saturation can lead to many acute adverse effects on individual organ systems. Heart rate and respiration were commonly

[19] Philip eICU program: https://www.usa.philips.com/healthcare/resources/landing/teleicu (accessed April 2021).

used vital signs, which were indicators of the body's basic functions. St1, st2, and st3 were estimated ST segment levels of the ECG (Pollard et al., 2018). We included body temperature, despite the fact that it had a high missing rate. The reasons are twofold. (1) First, because body temperature did not need to be collected frequently (it was usually stable and normally monitored routinely), a high missing rate was considered normal. Replacing the missing body temperature by the mean (i.e., the mean of the previous and next available records) was a common practice (Chacko & Peter, 2018). (2) Second, body temperature is extremely important for ICU mortality prediction (Achaiah & Ak, 2022). ICU patient's abnormal body temperature (e.g., fever) is linked to life-threatening medical emergencies (e.g., sepsis) (Kushimoto et al., 2014) and higher mortality. It forms part of ICU mortality prediction scores, such as SAPS III and APACHE IV. Note the adopted vital signs were all readily-available real-time time-series data, so our model did not require laboratory results and intensivists' assessment.

Table 3.5. Descriptive Statistics of the Patients' Demographics and Vital Signs

	Overall	**Expired patients**	**Alive patients**
ICU stay (hours)	75.15 (4.03-1232.98)	113.75 (4.26-1009.48)	72.87 (4.03-1232.98)
Age	70.16 (19-90), missing rate 0.019%*(1)	72.90 (20-90), missing rate 0	69.97 (19-90), missing rate 0.020%
Gender	51.01 % male, 48.97% female, the rest are unknown	53.51 % male, 46.49 % female, the rest are unknown	50.86 % male, 49.12 % female, the rest are unknown
Ethnicity	71.54% Caucasian, 15.71% African American, 5.26% Hispanic, 1.78% Asian, 0.30% Native American, the rest are unknown or other races	80.27% Caucasian, 11.71% African American, 3.68% Hispanic, 1.00% Asian, 0.33% Native American, the rest are unknown or other races	71.02% Caucasian, 15.95% African American, 5.35% Hispanic, 1.83% Asian, 0.30% Native American, the rest are unknown or other races
Height	168.03 (63.50-210.80)*(2)	168.07 (91.90-193.00)	168.03 (63.50-210.80)
Weight	89.09 (28.70-771.00)	89.16 (31.80-275.00)	89.09 (28.70-771.00)

Table 3.5. Continued.

			Overall	Expired patients	Alive patients
Periodic vital signs	Sao2 (0.95 % missing)	oxygen saturation, the percentage of available binding sites on hemoglobin that are bound with oxygen in arterial blood	96.17 (0-100)	95.27 (0-100)	96.25 (0-100)
	Heart rate (0.02% missing)	the number of times the heart beats per minute	84.47 (0-300)	88.65 (0-271)	84.07 (0-300)
	Temperature (92.39% missing)	the body temperature of a person	38.33 ((-132.35)-102.40)	36.89 ((-132.35)-42.90)	38.53 (0.20-102.40)
	Respiration (9.08% missing)	the number of breaths a person takes per minute	21.09 (0-185)	21.77 (0-171)	21.03 (0-185)
	St1 (46.44 % missing)	estimated ST segment level 1 of the ECG	1.05 ((-17.30)-570)	1.12 ((-10.16)-220)	1.04 ((-17.30)-570)
	St2 (45.02% missing)	estimated ST segment level 2 of the EC	1.78 ((-14.20)-830)	1.40 ((-14.15)-500)	1.81 ((-14.20)-830)
	St3 (47.67% missing)	estimated ST segment level 3 of the ECG	1.91 ((-24.75)-1040)	1.55 ((-24.75)-330)	1.94 ((-16.90)-1040)

Note: The total admissions are 5282. The total number of patients is 4801. (1) The variable age in the eICU dataset is set to ">89" if the patients are older than 89. To calculate the mean, we simply set the ages of the patients older than 89 to be 90. (2) When calculating the demographic statistics, we removed 11 records with irregular height (e.g., height: 772) and/or irregular weight (e.g., weight: 974).

Data preprocessing and experiment setting

After implementing Step 1 and Step 2 of the proposed method (see Sections 3.1 - 3.2), some missing values were detected in the extracted feature set. The missing values were introduced for two reasons. First, not all vital signs were constantly measured for all patients. Second, the missing values were generated during the feature extraction process. For example, for the relative extrema features, if the output of the frequency spectrum extraction was smooth without any local minima and maxima, then the corresponding extrema feature was a missing value. Before feeding the extracted feature set to the machine learning classifiers, we imputed missing values with the mean over all patients in the dataset. For more discussions about missing value imputation, please see Appendix A.

Due to the imbalance of the dataset (94.34% expired, 5.66% alive), we implemented a stratified 5-fold cross-validation in the evaluation process (Refaeilzadeh et al., 2009). The stratified k-fold cross-validation was a variation of k-fold cross-validation, which ensured each fold was representative of the class proportions in the training dataset. In our research setting, it yielded better bias and variance estimates in cases of unequal class proportions. Different machine learning classifiers were adopted and compared in our experiments. We grid-searched hyper-parameters for each machine learning algorithm, as shown in Appendix A (Table 3.12.). Meanwhile, to alleviate the influence of the data imbalance issue during classifier training, we assigned different weights to both the majority and minority classes according to the skewed distribution of the classes. The purpose was to penalize minority class misclassification by assigning a greater class weight while decreasing weight for the majority class (Pedregosa et al., 2018). During model training, the larger weight of the minority class in the algorithm's cost function delivered a stronger penalty to the minority class misclassification, thus the algorithm can focus on reducing errors for the minority class. In this research, $Y = \{1: expired\}$ was the minority class and $Y = \{0: alive\}$ was the majority. The assigned weights of two classes were inversely proportional to their relative frequencies, namely, the weight $w_{y=1} = \frac{\#samples}{2\times\#samples_{y=1}}$ and $w_{y=0} = \frac{\#samples}{2\times\#samples_{y=0}}$. Such weights' ratio was first introduced by King & Zeng (2001) and had since been adopted as the built-in parameter of class weights by most popular machine learning toolkits, such as Python's scikit-learn[20], LightGBM[21], and CatBoost[22], to help users optimize the prediction for the minority class.

[20] https://scikit-learn.org/stable/
[21] https://lightgbm.readthedocs.io/en/latest/
[22] https://catboost.ai/

In the second experiment, we compared our method with state-of-the-art baselines. The ARIMA- and ARMA-based methods were evaluated through 5-fold cross-validation. CNN (S. Y. Kim et al., 2019), CNN (Caicedo-Torres & Gutierrez, 2019), LSTM (Thorsen-Meyer et al., 2020), and GRU (Che et al., 2016) were evaluated using holdout evaluation (70% training, 30% testing). For CNNs, we followed the network architectures and parameters of S. Y. Kim et al. (2019) and Caicedo-Torres & Gutierrez (2019). For LSTM (Thorsen-Meyer et al., 2020) and GRU (Che et al., 2016), we tested different parameters as shown in Table 3.12. since the authors did not provide their parameters.

ICU mortality predictive feature assessment

We first assessed the relative feature importance and selected the most important features for ICU mortality prediction. As we stated earlier, there were a number of methods for feature importance assessment, and there was no consensus on which model works best in a given situation. We provided comparison results with various feature importance assessment algorithms, including logistic regression (LR, $l1$ penalty), linear support vector machine (LinearSVM, $l1$ penalty), random forest (RF), and ANOVA test. The selected features were then evaluated with four classifiers via cross-validation. The test set's AUC scores were utilized to see if the selected features improved prediction performance. As shown in Table 3.6., LinearSVM with $l1$ penalty selected the 397 most powerful features (using the first 24H vital sign data after ICU admission), which led to the best prediction results (AUC = 0.849). In the following mortality predicting experiments, we adopted these features for further evaluation.

Moreover, to gain insights into the contributions of different feature groups, we reported the most important 30 features that were selected by LinearSVM with the $l1$ penalty (the feature importance was measured by LinearSVM coefficients), as shown in Table 3.7. Both statistical

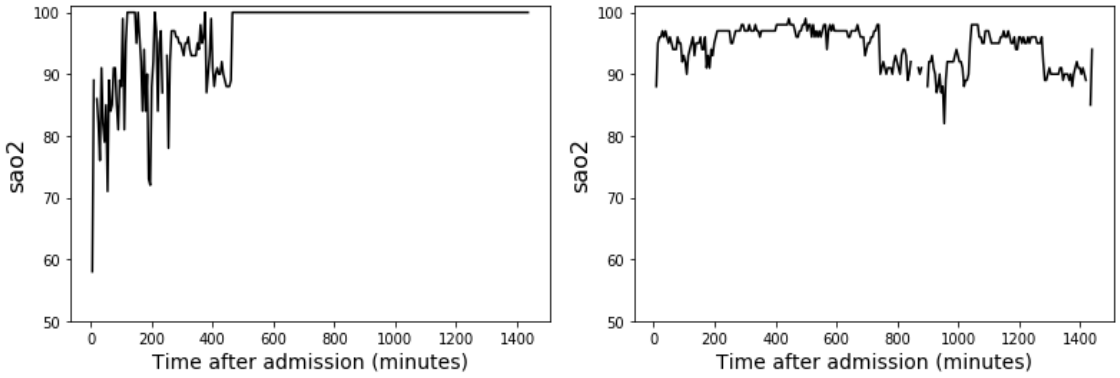
features and signal processing features contribute a lot to mortality prediction. (1) The top 3 most important features were statistical features. The first one was the rapid, sudden drop of sao2 (extracted by using a moving window on the time series (see Section 3.2 Statistical features). The sao2 was the oxygen saturation in blood, and the drop of sao2 was usually related to the most dangerous situation of patients (Vold et al., 2015). The second feature was the last value of sao2 in the first 24 hours (represented by statistical feature "last" indicating the last data point in the time series (see Section 3.2 Statistical features), which reflected the latest health condition (Vold et al., 2015). The third feature was the median value of heart rate. Usually, a healthy heart rate ranged from 60 to 100 (Kannel et al., 1987); hence a median value of heart rate outside of this range indicated high risk. (2) The extracted signal processing features had great predicting power as well. For instance, the location of the local minima of the WT with the mother wavelet = "cmor" of sao2 (the 4th most important feature) reflected the time points that lack the time-frequency relationship characterized by "cmor" wavelets. In other words, this feature indicated the time when the fluctuation of sao2 slowed down. Existing research (Bhogal & Mani, 2017) showed that the fluctuation of sao2 carried critical information about patients' health conditions. The local maxima of FFT of respiration (the 7th most important feature) revealed the specific frequency ranges of the vital sign. The frequency of respiration was an important characteristic of patients (Fadel et al., 2004); thus, it was useful for mortality prediction.

To investigate the contribution of each vital sign and each signal processing type, we summed the feature importance for each group. As shown in Table 3.8., the heart rate, respiration, and sao2 contributed more compared to other vital signs in the mortality prediction task. These three vital signs were not difficult or expensive to measure in ICUs. In the eICU database, the heart rate, respiration, and sao2 were constantly measured for more than 90% of

patients. Moreover, WT provides the most informative features among all signal processing techniques, while AC has the least impact. As a signal processing technique, AC is conceptually close to the time series forecasting algorithms like ARMA and ARIMA. The unsatisfactory performance of AC revealed the fact that time series forecasting algorithms can hardly capture sufficient information for ICU mortality prediction. This observation was also supported by our experiment using coefficients of ARMA and ARIMA for prediction (see Table 3.10.).

To demonstrate the feasibility and implications of our model, we selected and compared two patients from our dataset as an illustrating example, as shown in Table 3.9. (1) According to the oxygen saturation (sao2) signal (Figure 3.9. (a)), Patient 1 was at risk at the beginning, but went out of risk later. The sao2 of Patient 1 dropped to 58%, but after 24 hours (1440 minutes), the sao2 stayed at 100%. The oxygen saturation of Patient 1 fluctuated first, but stabilized around 840 minutes. APACHE IV only considered the worst measurements of the first 24 hours. It ignored the important healthy signal that the second half part of the sao2 conveyed, which caused APACHE IV to make a wrong inference (86.4% death probability). Our model captured the stable and healthy sao2 signal after 840 minutes, thus offering a lower risk score (31.8% death probability). (2) The sao2 of Patient 2 (Figure 3.9. (b)) dropped to 82% at first, and stayed at 85% later. Her oxygen saturation kept fluctuating during the first 24 hours. Compared to that at other times, sao2 was more stable at 610 minutes, which was detected by the WT technique. Although the lowest sao2 of Patient 2 was higher than Patient 1, her healthy status had not been stabilized after 24 hours of stay in ICU, which was not considered by APACHE IV. Therefore, our method gave a higher mortality probability (59.8%) than APACHE IV (8.3%) for Patient 2, who would expire later. This example showed that our model can capture the hidden patterns in the time series of vital signs which were difficult to detect by other methods, such as APACHE

IV. These captured patterns provided valid and interpretable feature sets for the ICU outcome prediction.



(a) sao2 of an alive patient (b) sao2 of an expired patient

Figure 3.9. Sao2 Comparison Between an Alive Patient and an Expired Patient (First 24 Hours of ICU Admission)

Table 3.6: Prediction Performance of Selected Features, 24H

Feature selection method	**Feature selection parameter**	**Classifier**	**Number of selected features**	**AUC**
ANOVA	-	LinearSVM	818	0.777
		LogisticRegression	818	0.776
		NeuralNetwork	818	0.716
		RandomForest	818	0.790
Logistic regression with $l1$ penalty	C = 0.005	LinearSVM	50	0.778
		LogisticRegression	50	0.780
		NeuralNetwork	50	0.688
		RandomForest	50	0.792
	C = 0.008	LinearSVM	131	0.817
		LogisticRegression	131	0.820
		NeuralNetwork	131	0.736
		RandomForest	131	0.809
	C = 0.010	LinearSVM	193	0.820
		LogisticRegression	193	0.825
		NeuralNetwork	193	0.792
		RandomForest	193	0.811

Table 3.6. Continued.

Feature selection method	**Feature selection parameter**	**Classifier**	**Number of selected features**	**AUC**
Logistic regression with $l1$ penalty	C = 0.020	LinearSVM	516	0.832
		LogisticRegression	516	0.828
		NeuralNetwork	516	0.824
		RandomForest	516	0.809
	C = 0.030	LinearSVM	756	0.823
		LogisticRegression	756	0.800
		NeuralNetwork	756	0.831
		RandomForest	756	0.801
Random Forest	criterion = entropy	LinearSVM	50	0.709
		LogisticRegression	50	0.708
		NeuralNetwork	50	0.721
		RandomForest	50	0.726
	criterion = gini	LinearSVM	50	0.715
		LogisticRegression	50	0.714
		NeuralNetwork	50	0.700
		RandomForest	50	0.719
	criterion = entropy	LinearSVM	100	0.745
		LogisticRegression	100	0.750
		NeuralNetwork	100	0.684
		RandomForest	100	0.747
	criterion = gini	LinearSVM	100	0.740
		LogisticRegression	100	0.741
		NeuralNetwork	100	0.671
		RandomForest	100	0.750
	criterion = entropy	LinearSVM	200	0.747
		LogisticRegression	200	0.748
		NeuralNetwork	200	0.640
		RandomForest	200	0.770
	criterion = gini	LinearSVM	200	0.745
		LogisticRegression	200	0.744
		NeuralNetwork	200	0.682
		RandomForest	200	0.782

Table 3.6. Continued.

Feature selection method	Feature selection parameter	Classifier	Number of selected features	AUC
	criterion = entropy	LinearSVM	400	0.760
		LogisticRegression	400	0.758
		NeuralNetwork	400	0.703
		RandomForest	400	0.805
	criterion = gini	LinearSVM	400	0.753
		LogisticRegression	400	0.747
		NeuralNetwork	400	0.691
		RandomForest	400	0.800
Linear Support Vector Machine with $l1$ penalty	C = 0.001	LinearSVM	20	0.762
		LogisticRegression	20	0.763
		NeuralNetwork	20	0.721
		RandomForest	20	0.777
	C = 0.005	LinearSVM	**397**	**0.849**
		LogisticRegression	397	0.843
		NeuralNetwork	397	0.827
		RandomForest	397	0.806
	C = 0.010	LinearSVM	678	0.846
		LogisticRegression	678	0.810
		NeuralNetwork	678	0.841
		RandomForest	678	0.794

We listed the groups of features adopted in APACHE IV in Table 3.3. (see the Related Work section) to emphasize the intuition of adopting features extracted from the vital signs for ICU mortality prediction. We noticed that among all the features used by APACHE IV, the vital signs were the only real-time updated information, which can be acquired directly and automatically from the ICU bedside monitors without any input from the experts. Tracking laboratory results can be time-consuming. On average, it took 35 minutes - 5.5 hours for routine inpatient or clinical biochemistry tests (Goswami et al., 2010; Winkelman et al., 1997). In practice, if we considered the labs' availability, the waiting time for obtaining laboratory test results was even longer. This was why researchers believed the next generation of ICU mortality

predictive models should use an automated electronic system for data gathering and prediction generating (Zimmerman & Kramer, 2014).

Table 3.7: Feature Importance and Selected Features from LinearSVM with $l1$ Penalty (Top 30)

Feature name	**Importance**	**Vital sign**	**Feature type**
sao2_sudden-drop_value_1	1.57E-01	sao2	statistics
stat_sao2_last	1.18E-01	sao2	statistics
stat_heartrate_median	8.43E-02	heart rate	statistics
sao2_wt_cmor2-8_dist2_local-minima_shift_4	7.61E-02	sao2	wt
respiration_wt_morl_dist4_local-minima_shift_5	7.12E-02	respiration	wt
stat_respiration_median	6.61E-02	respiration	statistics
respiration_fft_dist8_local-maxima_x_4	5.89E-02	respiration	fft
stat_heartrate_last	5.88E-02	heart rate	statistics
heartrate_wt_mexh_length5_power-in-band_2	5.65E-02	heart rate	wt
heartrate_psd_dist16_local-minima_x_4	5.43E-02	heart rate	psd
st3_psd_dist2_local-minima_x_2	4.95E-02	st3	psd
st1_fft_dist2_local-minima_x_1	4.47E-02	st1	fft
heartrate_wt_cmor1-4_dist4_local-minima_shift_2	4.43E-02	heart rate	wt
temperature_psd_dist16_local-minima_x_4	4.40E-02	temperature	psd
stat_respiration_last	4.39E-02	respiration	statistics
respiration_sudden-drop_value_5	4.38E-02	respiration	statistics
sao2_ac_length5_power-in-band_1	4.33E-02	sao2	ac
heartrate_sudden-drop_x_2	4.29E-02	heart rate	statistics
st2_psd_dist4_local-minima_x_5	4.21E-02	st2	psd
respiration_wt_mexh_dist2_local-minima_freq_3	4.10E-02	respiration	wt
sao2_psd_dist32_local-minima_x_1	4.07E-02	sao2	psd
stat_respiration_min	4.04E-02	respiration	statistics
st1_fft_dist32_local-maxima_x_5	3.98E-02	st1	fft
st2_fft_dist8_local-minima_x_5	3.88E-02	st2	fft
st2_fft_dist32_local-maxima_x_4	3.82E-02	st2	fft
st1_sudden-drop_x_1	3.80E-02	st1	statistics
sao2_wt_cmor2-4_dist2_local-minima_shift_1	3.78E-02	sao2	wt
st1_psd_dist32_local-minima_x_1	3.77E-02	st1	psd
heartrate_psd_length5_power-in-band_4	3.74E-02	heart rate	psd
heartrate_wt_mexh_dist4_local-minima_shift_4	3.72E-02	heart rate	wt

Table 3.8: Summary of the Importance of Different Feature Types

	AC	WT	FFT	PSD	Statistics	SUM
heart rate	0.034	0.742	0.150	0.303	0.290	1.520
respiration	0.125	0.666	0.165	0.247	0.246	1.449
sao2	0.043	0.486	0.118	0.125	0.358	1.131
st1	0.061	0.119	0.153	0.111	0.114	0.559
st2	0.057	0.228	0.139	0.100	0.018	0.542
st3	0.071	0.134	0.073	0.09	0.011	0.379
temperature	0	0.091	0.038	0.104	0	0.233
SUM	0.393	2.467	0.836	1.079	1.036	

Table 3.9: Examples of two different patients

	Feature name	**Patient 1 (alive)**	**Patient 2 (expired)**
Proposed features	sao2_sudden-drop_value_1	58	82
	stat_sao2_last	100	85
	stat_heartrate_median	83	88
	sao2_wt_cmor2-8_dist2_local-minima_shift_4	840	610
	respiration_wt_morl_dist4_local-minima_shift_5	1,050	965
	stat_respiration_median	15	24
	respiration_fft_dist8_local-maxima_x_4	0.021	0.012
	stat_heartrate_last	93	86
	heartrate_wt_mexh_length5_power-in-band_2	688,626.9	716,518.4
	heartrate_psd_dist16_local-minima_x_4	0.032	0.084
	...		
APACHE IV features (not used by the proposed framework)	age	79	83
	gender	Female	Female
	If the patient has active treatment	yes	yes
	If the patient has diabetes	yes	no
	If the patient was intubated at 24 hours	yes	no
	...		
Probability of death at 24 H	Our proposed model	31.8 %	59.8 %
	APACHE IV	86.4 %	8.3 %

ICU mortality prediction results

In the following experiments, we evaluated the proposed model using the first 24 hours (H) ICU time series of vital signs, and compared our method with eight baselines. (1) APACHE IV (Zimmerman et al., 2006), a widely accepted ICU scoring system. APACHE IV had already

been applied and stored in the eICU database. (2) CNN-(S. Y. Kim et al., 2019) and (3) CNN-(Caicedo-Torres & Gutierrez, 2019), two CNN models previously used for ICU mortality prediction and achieved high performance. (4) LSTM (Hochreiter & Schmidhuber, 1997; Thorsen-Meyer et al., 2020) and (5) GRU (Che et al., 2016), two RNN models that took vital signs in time sequence to estimate mortality rate. (6) ARMA and (7) ARIMA classification (Carden & Brownjohn, 2008), the classical statistical time-series forecasting methods: fitted the ARMA/ARIMA models on vital signs, and took the estimated coefficients as the inputs of machine learning classifiers to predict mortality probabilities. (8) Statistical feature classification (S. Kim et al., 2011; Davoodi & Moradi, 2018; Hsieh et al., 2018; Kong et al., 2020; Zhai et al., 2020): simple statistics of vital signs including mean, median and standard deviation were calculated as features for ICU mortality prediction.

The results were reported in Table 3.10. and Table 3.11. Our model (signal processing features and statistical features + LinearSVM $l1$feature selection (top 397 features) + LinearSVM) achieved higher AUC (0.849) and F1 (0.316) than all the 8 baseline models: APACHE IV, LSTM, GRU, CNN-S. Y. Kim, CNN-Caicedo-Torres & Gutierrez, ARMA, ARIMA, and statistical feature classification. Although deep learning models dominate the world of data science, both CNN and RNN were far outperformed by our signal processing and machine learning-based model. This was because we explicitly extracted valid information from the vital sign data, which made it easier for classifiers to find the relationship between the input space (vital signs) and the prediction out space (mortality).

Unlike our model, the APACHE IV required laboratory results (which can be time-consuming to obtain) and intensivists' assessment (which may not always be available) as input variables. For example, APACHE IV required fiO2 value from the worst ABG data and GCS

verbal score (Pollard et al., 2018) (see Table 3.3. for more information). APACHE IV used more resource-demanding variables to make predictions. Nevertheless, it only achieved an AUC of 0.750 and an F1 of 0.124, both of which were much lower than our proposed method. Therefore, the proposed method can achieve better performance even using fewer resources than APACHE IV.

To examine if the proposed method can add value to existing systems like APACHE IV, we merged the features generated by our model (based on 24H vital signs) with APACHE IV variables. The APACHE variables include patients' demographics and other attributes available at ICU admission, which can have great value for ICU mortality prediction. All APACHE IV variables were available in the eICU database[23] (Pollard et al., 2018). We excluded APACHE variables that require lab resources (e.g., ABG) or intensivists' assessment (e.g., GCS) because we want to keep our prediction model resource efficient. The new feature set reaches an AUC of 0.869 (see Table 3.11., Feature type "Proposed feature set and APACHE IV variables"), which means the features generated by our model can be integrated into other ICU mortality prediction models. The resulting model can have improved prediction power and should reduce the need for time-consuming or resource-demanding human intervention.

The feature set we proposed included the signal processing features and the statistical features. We examined their effectiveness respectively. We excluded statistical features and signal processing features, respectively, and re-conducted the evaluation. As shown in Table 3.11. (Feature type "Signal processing features only" and Feature type "Statistical features only"), using only signal processing features versus only statistical features, the predictive model reaches AUC of 0.782 and 0.765, respectively. The results indicated both signal processing and

[23] APACHE IV variables are available in https://eicu-crd.mit.edu/eicutables/apachepredvar/.

statistical techniques extracted informative features, and the extracted features were more predictive than APACHE IV (AUC = 0.750) variables for ICU mortality prediction. Moreover, the obtained AUC scores of signal processing features (Table 3.11., Feature type "Signal processing features only") were greater than that of statistical features (Table 3.11., Feature type "Statistical features only"), which validated the necessity of the signal processing techniques for feature extraction.

Table 3.10: Baseline Methods Using 24H Vital Signs

	Feature type	**Classifier**	**AUC**	**Precision**	**Recall**	**F1**
Baselines	APACHE IV	-	**0.750**	0.404	0.074	0.124
	Original vital signs	CNN-1	**0.732**	1.00	0.123	0.218
	Original vital signs	CNN-2	**0.712**	0.857	0.057	0.106
	Original vital signs	GRU	**0.722**	1.00	0.132	0.233
	Original vital signs	LSTM	**0.698**	0.818	0.085	0.154
	ARMA coefficients	LinearSVM	0.660	0.090	0.545	0.155
		LogisticRegression	0.663	0.097	0.509	0.163
		NeuralNetwork	0.598	0.214	0.109	0.144
		RandomForest	**0.709**	0.142	0.500	0.222
	ARIMA coefficients	LinearSVM	0.611	0.069	0.555	0.123
		LogisticRegression	0.633	0.059	0.494	0.105
		NeuralNetwork	0.594	0.170	0.121	0.141
		RandomForest	**0.695**	0.117	0.464	0.187
	Statistical features	LinearSVM	0.745	0.134	0.562	0.216
		LogisticRegression	0.742	0.133	0.565	0.215
		NeuralNetwork	0.652	0.383	0.234	0.289
		RandomForest	**0.765**	0.971	0.138	0.241

Note: AUC: based on the probability of ICU mortality
Precision, recall, and F1: based on two-way classification results ($Y = \{1: expired, 0: alive\}$)
CNN-1: CNN-S. Y. Kim
CNN-2: CNN-Caicedo-Torres & Gutierrez

Table 3.11: Proposed Method Using First 24H Data after ICU Admission

	Feature type	**Classifier**	**AUC**	**Precision**	**Recall**	**F1**
Proposed method (Feature set comparison)	Statistical features only	LinearSVM	0.745	0.134	0.562	0.216
		LogisticRegression	0.742	0.133	0.565	0.215
		NeuralNetwork	0.652	0.383	0.234	0.289
		RandomForest	**0.765**	0.971	0.138	0.241
	Signal processing features only	LinearSVM	**0.782**	0.168	0.580	0.260
		LogisticRegression	0.777	0.163	0.555	0.252
		NeuralNetwork	0.758	0.303	0.111	0.161
		RandomForest	0.781	0.400	0.006	0.012
	Statistical features + Signal processing features (Proposed feature set)	LinearSVM	**0.849**	0.216	0.586	0.316
		LogisticRegression	0.843	0.217	0.571	0.315
		NeuralNetwork	0.827	0.561	0.241	0.335
		RandomForest	0.806	0.960	0.087	0.160
	Proposed feature set + APACHE IV variables*	LinearSVM	**0.869**	0.247	0.619	0.353
		LogisticRegression	0.865	0.244	0.616	0.349
		NeuralNetwork	0.852	0.551	0.246	0.338
		RandomForest	0.817	0.960	0.081	0.149

Note: AUC: based on the probability of ICU mortality
Precision, recall, and F1: based on two-way classification results ($Y = \{1: expired, 0: alive\}$)
* APACHE IV variables: contain patients' demographics and other information that are available before ICU admission, we have excluded APACHE variables that require lab resources (e.g., ABG) or intensivists' assessment (e.g., GCS) because we want to keep our prediction model resource efficient.

To evaluate the real-time predictive power of the proposed method, we conducted the experiments with data from different time spans: 3H, 6H, 12H, 24H, 32H, 72H, and the whole ICU stay. Patients may expire or leave the ICU after hours to days of stays. Figure 3.10. showed the percentage of patients that stayed in the ICUs over time. For patients whose length of stay was less than x hours, $x \in (3, 6, 12, 24, 32, 72)$, we included all these patients and the whole time series of their vital signs in the x-hour experiments. For patients whose length of stay was more than x hours, we only used their vital signs within the first x hours. The reasons for such a setup are twofold. (1) The proposed feature set did not represent patients' length of stay. The

properties of patients' vital signs were summarized and abstractly represented by our proposed feature set (i.e., the relative extrema and power-in-band from the time/frequency domain, the statistical features from the time series of vital signs). The extracted features had the same dimensionalities for patients with different lengths of stay. There was no way for machine learning classifiers to correspond to patients' length of stay and infer patients' mortalities. Hence, there was little overfitting and information leakage during the classifier's training process. The learned classifiers can effectively fit additional data and predict future observations reliably. (2) This setup was common for ICU outcome predictions. Well-acknowledged ICU outcome prediction scoring systems, such as APACHE IV, MPM III, and SAPS II, also included patients whose length of stays were less than 24 hours for their 24 hours predictions (Vasilevskis et al., 2009; Zimmerman et al., 2006).

Figure 3.11. presented the performance of the proposed model over time using different classifiers. Note that only the APACHE IV score at the time point of 24 hours after ICU admission was readily available and compared with our method. We cannot compare the performance of our method with APACHE IV at other time points, since the APACHE IV was not an open-source system and we did not have the resources to compute the APACHE IV scores at other time points. (1) As shown in Figure 3.11., the AUC of the proposed method went up over time using cumulative data from patients with prolonged ICU stays. (2) More importantly, except random forest, as early as 3H, all the other classifiers using the proposed features (AUC > 0.786) achieved better performance than APACHE IV. (3) The blue and red lines were above the others, which indicated the better prediction performance of LinearSVM and LR.

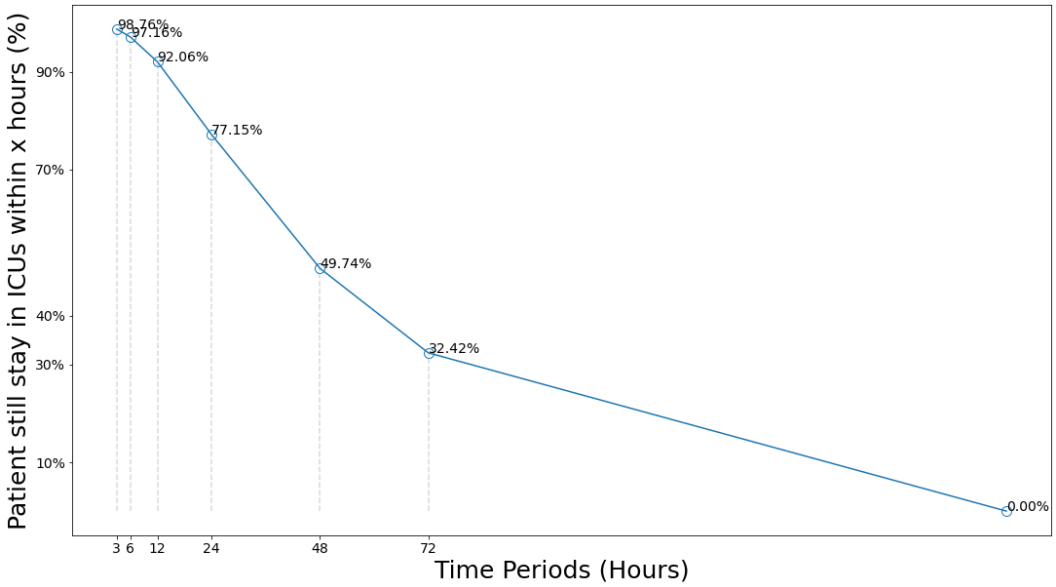
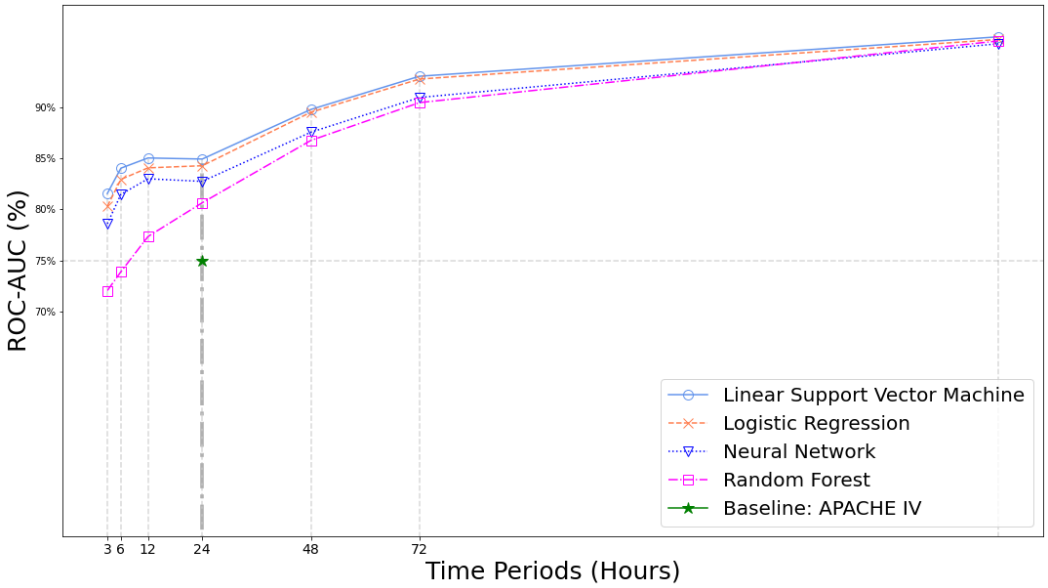


Figure 3.10: Percentage of Patients Still Staying in ICUs



Note: In the eICU dataset, only the APACHE IV score at the time point of 24 hours after the ICU admission is available.

Figure 3.11: ICU Mortality Prediction Performance over Time

According to our findings, the proposed method can provide real-time forecasts and make earlier predictions than APACHE IV without sacrificing accuracy. Practically, we suggest adopting LinearSVM or LR in Step 3, both of which are interpretable linear models. Note

APACHE IV has also adopted Logistic Regression as its predictive model. For application, the proposed framework is capable of supporting physicians at the bedside for patient management and resource allocation since our method continually calculated a risk score for the patient beyond the first 24 hours of ICU admission.

The interpretability of the proposed method

A model's interpretability was defined as the degree to which a medical practitioner can understand the reason behind a prediction made by the model (Dam et al., 2018). The goal of interpretability was to describe the internals of the prediction model (Gilpin et al., 2019).

The interpretability of our method existed in two parts, the machine learning algorithm and the proposed feature set. On one hand, we recommended linear models in practice because the learning function of such models can provide a weighting over the input features which was useful for the model explanation. On the other hand, the features obtained from the signal processing techniques were interpretable because they had practical meaning on the frequency/time domain and can be traced back to the original time series data.

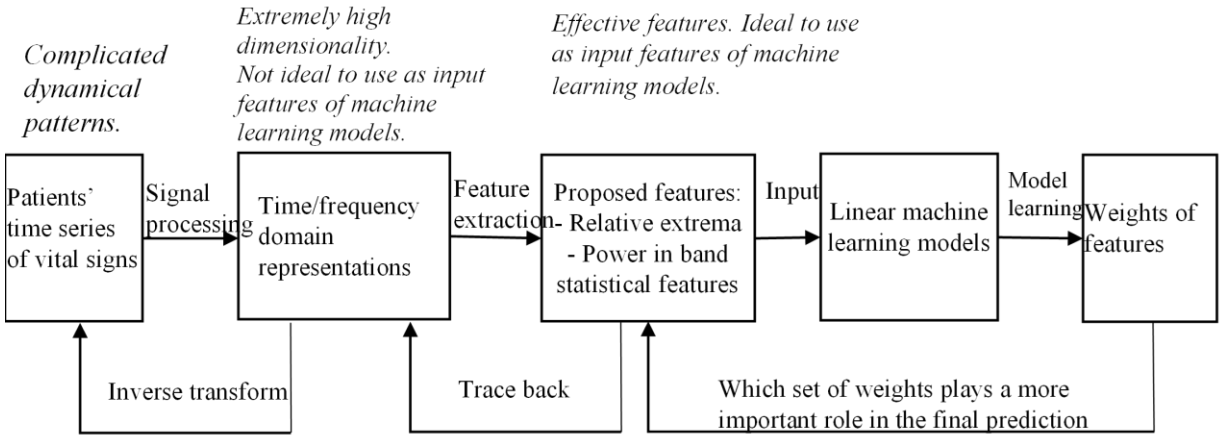


Figure 3.12 The Interpretability of the Proposed Method

Note that all proposed features, which revealed the properties of patients' vital signs, were explainable (just like simple statistics, e.g., mean and max), even though the input features

involved the frequency domain. As shown in Figure 3.12., the time/frequency domain we obtained in the proposed method were from stochastic signal processing techniques. All the stochastic signal processing techniques we adopted were mathematical transforms that decomposed the time series of patients' vital signs into functions depending on spatial or temporal frequency. We can apply inverse transforms that mathematically synthesize the original time-series data. Due to this, all the features were explainable or can be traced back to the original time-series data.

Conclusion and Future Work

In this study, we seek to answer two research questions. First, how can we develop a new ICU mortality prediction model leveraging readily-available data with minimized requirements on the intensivists' expertise and having improved accuracy? Second, how can we effectively extract valid and interpretable features from the time series of vital sign data? We propose a novel ICU mortality prediction method combining stochastic signal processing and machine learning techniques. We systematically evaluate the proposed method using a real-world multi-center ICU dataset. The proposed method outperforms state-of-the-art baselines by a large margin. In addition, the proposed method makes increasingly accurate predictions with patients' increasing length of stay. Our method makes accurate predictions with 3 hours' worth of data, whereas widely accepted methods like APACHE IV need 24 hours' worth of data for predictions. More importantly, we use stochastic signal processing, a novel technique in ICU outcome prediction, for feature extraction. The extracted features are both valid and interpretable. They can be incorporated into other extant ICU outcome prediction models for better prediction results.

Our proposed model is very promising for many reasons. Our work is the first study to (1) convert the time series of vital signs to the frequency domain, (2) effectively extract the

frequency domain's features, and (3) use features from both time series and frequency domain to predict ICU mortality. Our major contributions for ICU mortality predictions are: first, we offer a new model for real-time predictions that requires only ICU bedside monitoring data; second, the proposed method greatly advance the performance of ICU mortality prediction; third, the extracted features are highly interpretable compared to those extracted from black-box models, which facilitates model adoption and implementation. From the perspective of data science for social good, the proposed model enlarges the social impact of ICU outcome prediction studies. Our prediction result has improved accuracy and is more reliable. The prediction results are relevant to both ICU patients and critical care practitioners. The proposed methods can have wide applications because the extracted features are interpretable to healthcare professionals.

Future work

While the results are encouraging, the proposed method is not without limitations.

First, more vital-sign data and baselines can be examined. This study adopts the readily-available times series of vital signs, including heart rate, sao2, body temperature, respiration, st1, st2, and st3. Other vital sign data (e.g., the central venous pressure and pulmonary artery pressure) are not included due to the high missing rate. The predicting power of these vital sign data can be evaluated in the future. This study considers two RNNs (i.e., GRU, LSTM), and two CNNs, as deep learning-based baselines. However, recent studies show that the attention-based deep learning models are better at extracting patterns from time-series data (Tang et al., 2018; Vaswani et al., 2017). We can include and compare the attention-based models as baselines in the future as well.

Moreover, reliability is very important to consider in ICU outcome predictions. This work attempts to improve the reliability of our method by increasing its accuracy and interpretability. However, in order to apply our findings to a real-world ICU setting, the model's

trustworthiness and consistency must be verified in real-world intensive care units. A more overall understanding of the model's reliability can be achieved by a longer time series analysis in the future.

Furthermore, in ICU outcome predictions, generalizability is critical to investigate. This study uses heart failure as a research case. To evaluate its generalizability, we test our method on two other common ICU complications as well (please see the results in Appendix B). In our future work, we can expand our method to a more general model. We intend to make the disease type a predictive variable in the model. Specifically, the generalized model is $Model_{new} = Model_{current} + \alpha * Disease$, where α is the coefficient vector for different diseases. By taking different types of diseases into account, our method can be a general mortality predicting model like APACHE IV. Additionally, the new model will still keep its interpretability because the added variable does not break the linearity of the current model. The new model can be evaluated on multiple diseases as well. Meanwhile, more experiments can be conducted to examine the performance of other ICU outcome prediction tasks, such as length of stay and ICU cost.

Last, a model's interpretability is crucial to a healthcare-related predictive model. This work proposes a combination of linear machine learning models and interpretable features to increase the interpretability of our method. However, the frequency domain and the features extracted from them may not be straightforward to medical professionals, and may require more background knowledge. In future work, we can focus on the interpretability issue and create an interactive interpreting system to facilitate the process.

Reference

Aaron Baird, Corey Angst, & Eivor Oborn. (2018). Health Information Technology. https://doi.org/10.25300/06212018

Achaiah, N. C., & Ak, A. K. (2022). Fever In the Intensive Care Patient. In StatPearls. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK570583/

Addison, P. S. (2017). The illustrated wavelet transform handbook: Introductory theory and applications in science, engineering, medicine and finance. CRC press.

Anderson, C. W., Stolz, E. A., & Shamsunder, S. (1998). Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Transactions on Biomedical Engineering, 45(3), 277–286.

Barnato, A. E., & Angus, D. C. (2004). Value and role of intensive care unit outcome prediction models in end-of-life decision making. Critical Care Clinics, 20(3), 345–362.

Becker, R. B., & Zimmerman, J. E. (1996). ICU scoring systems allow prediction of patient outcomes and comparison of ICU performance. Critical Care Clinics, 12(3), 503–514. https://doi.org/10.1016/s0749-0704(05)70258-x

Bejani, M., Gharavian, D., & Charkari, N. M. (2014). Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. Neural Computing and Applications, 24(2), 399–412. https://doi.org/10.1007/s00521-012-1228-3

Bergfeldt, L., Lundahl, G., Bergqvist, G., Vahedi, F., & Gransberg, L. (2017). Ventricular repolarization duration and dispersion adaptation after atropine induced rapid heart rate increase in healthy adults. Journal of Electrocardiology, 50(4), 424–432.

Bhogal, A. S., & Mani, A. R. (2017). Pattern Analysis of Oxygen Saturation Variability in Healthy Individuals: Entropy of Pulse Oximetry Signals Carries Information about Mean Oxygen Saturation. Frontiers in Physiology, 8, 555. https://doi.org/10.3389/fphys.2017.00555

Bloomfield, P. (2004). Fourier analysis of time series: An introduction. John Wiley & Sons.

Broersen, P. M. T. (2006). Automatic Autocorrelation and Spectral Analysis. Springer Science & Business Media.

Caicedo-Torres, W., & Gutierrez, J. (2019). ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU. Journal of Biomedical Informatics, 98, 103269. https://doi.org/10.1016/j.jbi.2019.103269

Carden, E. P., & Brownjohn, J. M. (2008). ARMA modelled time-series classification for structural health monitoring of civil infrastructure. Mechanical Systems and Signal Processing, 22(2), 295–314.

CDC. (2020). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19)—United States, February 12–March 16, 2020. MMWR. Morbidity and Mortality Weekly Report, 69. https://doi.org/10.15585/mmwr.mm6912e2

Celi, L. A., Hassan, E., Marquardt, C., Breslow, M., & Rosenfeld, B. (2001). The eICU: It's not just telemedicine. Critical Care Medicine, 29(8), N183.

Chacko, B., & Peter, J. (2018). Temperature monitoring in the intensive care unit. Indian Journal of Respiratory Care, 7(1), 28. https://doi.org/10.4103/ijrc.ijrc_13_17

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2016). Recurrent Neural Networks for Multivariate Time Series with Missing Values. ArXiv:1606.01865 [Cs, Stat]. http://arxiv.org/abs/1606.01865

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge.

Cohen, M. X. (2008). Assessing transient cross-frequency coupling in EEG data. Journal of Neuroscience Methods, 168(2), 494–499. https://doi.org/10.1016/j.jneumeth.2007.10.012

Dam, H. K., Tran, T., & Ghose, A. (2018). Explainable software analytics. Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, 53–56. https://doi.org/10.1145/3183399.3183424

Dasta, J. F., McLaughlin, T. P., Mody, S. H., & Piech, C. T. (2005). Daily cost of an intensive care unit day: The contribution of mechanical ventilation. Critical Care Medicine, 33(6), 1266–1271.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. JOSA A, 2(7), 1160–1169. https://doi.org/10.1364/JOSAA.2.001160

Davoodi, R., & Moradi, M. H. (2018). Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. Journal of Biomedical Informatics, 79, 48–59.

DeepAI. (2019, November 1). Frequency Domain. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/frequency-domain

Fadel, P. J., Barman, S. M., Phillips, S. W., & Gebber, G. L. (2004). Fractal fluctuations in human respiration. Journal of Applied Physiology (Bethesda, Md.: 1985), 97(6), 2056–2064. https://doi.org/10.1152/japplphysiol.00657.2004

Frassica, J. J. (2005). Frequency of laboratory test utilization in the intensive care unit and its implications for large-scale data collection efforts. Journal of the American Medical Informatics Association, 12(2), 229–233.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. ArXiv:1806.00069 [Cs, Stat]. http://arxiv.org/abs/1806.00069

Goswami, B., Singh, B., Chawla, R., Gupta, V. K., & Mallika, V. (2010). Turn Around Time (TAT) as a Benchmark of Laboratory Performance. Indian Journal of Clinical Biochemistry, 25(4), 376–379. https://doi.org/10.1007/s12291-010-0056-4

Grami, A. (2016). Introduction to digital communications. Academic Press.

Gray, R. M., & Davisson, L. D. (2004). An introduction to statistical signal processing. Cambridge University Press.

Gregor, S. (2002). Design Theory in Information Systems. Australasian Journal of Information Systems, 10(1). https://doi.org/10.3127/ajis.v10i1.439

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. MIS Quarterly, 37(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1), 389–422. https://doi.org/10.1023/A:1012487302797

Halpern, N. A., & Pastores, S. M. (2010). Critical care medicine in the United States 2000–2005: An analysis of bed numbers, occupancy rates, payer mix, and costs. Critical Care Medicine, 38(1), 65–71.

Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75. https://doi.org/10.2307/25148625

Higgins, T. L., Teres, D., Copes, W. S., Nathanson, B. H., Stark, M., & Kramer, A. A. (2007). Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III). Critical Care Medicine, 35(3), 827–835.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Hsieh, M. H., Hsieh, M. J., Chen, C.-M., Hsieh, C.-C., Chao, C.-M., & Lai, C.-C. (2018). Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. Scientific Reports, 8(1), 1–7.

J. Ellis, T., & Levy, Y. (2010). A Guide for Novice Researchers: Design and Development Research Methods. 107–118. https://doi.org/10.28945/1237

Jain, S., & Iverson, L. M. (2021). Glasgow Coma Scale. In StatPearls. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK513298/

Kalb, P. E., & Miller, D. H. (1989). Utilization strategies for intensive care units. Jama, 261(16), 2389–2395.

Kannel, W. B., Kannel, C., Paffenbarger Jr, R. S., & Cupples, L. A. (1987). Heart rate and cardiovascular mortality: The Framingham Study. American Heart Journal, 113(6), 1489–1494.

Kay, S. M., & Marple, S. L. (1981). Spectrum analysis—A modern perspective. Proceedings of the IEEE, 69(11), 1380–1419. https://doi.org/10.1109/PROC.1981.12184

Keegan, M. T., Gajic, O., & Afessa, B. (2011). Severity of illness scoring systems in the intensive care unit. Critical Care Medicine, 39(1), 163–169.

Keegan, M. T., Gajic, O., & Afessa, B. (2012). Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. Chest, 142(4), 851–858.

Kim, S., Kim, W., & Park, R. W. (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthcare Informatics Research, 17(4), 232.

Kim, S. Y., Kim, S., Cho, J., Kim, Y. S., Sol, I. S., Sung, Y., Cho, I., Park, M., Jang, H., Kim, Y. H., Kim, K. W., & Sohn, M. H. (2019). A deep learning model for real-time mortality prediction in critically ill children. Critical Care, 23(1), 279. https://doi.org/10.1186/s13054-019-2561-z

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Knaus, W. A., Draper, E. A., Wagner, D. P., & Zimmerman, J. E. (1985). APACHE II: A severity of disease classification system. Critical Care Medicine, 13(10), 818–829.

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G., Sirio, C. A., Murphy, D. J., Lotring, T., & Damiano, A. (1991). The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults. Chest, 100(6), 1619–1636.

Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Medical Informatics and Decision Making, 20(1), 1–10.

Kumar, N., Alam, K., & Siddiqi, A. H. (2017). Wavelet Transform for Classification of EEG Signal using SVM and ANN. Biomedical and Pharmacology Journal, 10(4), 2061–2069.

Kushimoto, S., Yamanouchi, S., Endo, T., Sato, T., Nomura, R., Fujita, M., Kudo, D., Omura, T., Miyagawa, N., & Sato, T. (2014). Body temperature abnormalities in non-neurological critically ill patients: A review of the literature. Journal of Intensive Care, 2(1), 14. https://doi.org/10.1186/2052-0492-2-14

Le Gall, J.-R., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. Jama, 270(24), 2957–2963.

Lee, J., Dubin, J. A., & Maslove, D. M. (2016). Mortality Prediction in the ICU. In MIT Critical Data (Ed.), Secondary Analysis of Electronic Health Records (pp. 315–324). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_21

Lehman, L. H., Adams, R. P., Mayaud, L., Moody, G. B., Malhotra, A., Mark, R. G., & Nemati, S. (2015). A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction. IEEE Journal of Biomedical and Health Informatics, 19(3), 1068–1076. https://doi.org/10.1109/JBHI.2014.2330827

Little, M. A. (2019). Machine Learning for Signal Processing: Data Science, Algorithms, and Computational Statistics. Oxford University Press.

Liu, R., & Gillies, D. F. (2016). Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognition, 53, 73–86. https://doi.org/10.1016/j.patcog.2015.11.015

Liu, Z., & Hauskrecht, M. (2017). A personalized predictive framework for multivariate clinical time series via adaptive model selection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 1169–1177.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

March & Storey. (2008). Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research. MIS Quarterly, 32(4), 725. https://doi.org/10.2307/25148869

Medicine, I. C. (2021). About Intensive Care | The Faculty of Intensive Care Medicine. https://www.ficm.ac.uk/faculty-membership/about-intensive-care

Moreno, R. P., Metnitz, P. G., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., & Le Gall, J.-R. (2005). SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine, 31(10), 1345–1355.

Myers, L. C., Parodi, S. M., Escobar, G. J., & Liu, V. X. (2020). Characteristics of Hospitalized Adults With COVID-19 in an Integrated Health Care System in California. JAMA, 323(21), 2195. https://doi.org/10.1001/jama.2020.7202

Nates, J. L., Nunnally, M., Kleinpell, R., Blosser, S., Goldner, J., Birriel, B., Fowler, C. S., Byrum, D., Miles, W. S., Bailey, H., & Sprung, C. L. (2016). ICU Admission, Discharge, and Triage Guidelines: A Framework to Enhance Clinical Operations, Development of Institutional Policies, and Further Research. Critical Care Medicine, 44(8), 1553–1602. https://doi.org/10.1097/CCM.0000000000001856

Nunamaker, J. F., Chen, M., & Purdin, T. D. M. (1990). Systems Development in Information Systems Research. Journal of Management Information Systems, 7(3), 89–106. https://doi.org/10.1080/07421222.1990.11517898

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2018). Scikit-learn: Machine Learning in Python. ArXiv:1201.0490 [Cs]. http://arxiv.org/abs/1201.0490

Pirracchio, R. (2016). Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project. In MIT Critical Data (Ed.), Secondary Analysis of Electronic Health Records (pp. 295–313). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_20

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. The Lancet Respiratory Medicine, 3(1), 42–52. https://doi.org/10.1016/S2213-2600(14)70239-5

Platt, J. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv. Large Margin Classif., 10.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5(1), 180178. https://doi.org/10.1038/sdata.2018.178

Poppas, A., & Rounds, S. (2002). Congestive heart failure. American Journal of Respiratory and Critical Care Medicine, 165(1), 4–8. https://doi.org/10.1164/ajrccm.165.1.2102075

Population Clock. (2021). https://www.census.gov/popclock/

Prasad, B. V. P., & Parthasarathy, V. (2018). Detection and classification of cardiovascular abnormalities using FFT based multi-objective genetic algorithm. Biotechnology & Biotechnological Equipment, 32(1), 183–193. https://doi.org/10.1080/13102818.2017.1389303

Prin, M., & Wunsch, H. (2012). International comparisons of intensive care: Informing outcomes and improving standards. Current Opinion in Critical Care, 18(6), 700.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. Liu & M. T. Özsu (Eds.), Encyclopedia of Database Systems (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Reith, F. C., Synnot, A., van den Brande, R., Gruen, R. L., & Maas, A. I. (2017). Factors Influencing the Reliability of the Glasgow Coma Scale: A Systematic Review. Neurosurgery, 80(6), 829–839. https://doi.org/10.1093/neuros/nyw178

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., and the Northwell COVID-19 Research Consortium, Barnaby, D. P., Becker, L. B., Chelico, J. D., Cohen, S. L., Cookingham, J., Coppa, K., Diefenbach, M. A., Dominello, A. J., Duer-Hefele, J., Falzon, L., Gitlin, J., Hajizadeh, N., … Zanos, T. P. (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA, 323(20), 2052. https://doi.org/10.1001/jama.2020.6775

Sasaki, A., Shiraishi, A., Nozaki, S., & Takahashi, M. (2020). Acute Physiology and Chronic Health Evaluation IV Probability of Mortality Is an Intermediate Variable, Not a Confounder. Critical Care Medicine, 48(3), e252–e253.

Shorr, A. F. (2002). An update on cost-effectiveness analysis in critical care. Current Opinion in Critical Care, 8(4), 337–343.

Sirio, C. A., Angus, D. C., & Rosenthal, G. E. (1994). Cleveland Health Quality Choice (CHQC)–an ongoing collaborative, community-based outcomes assessment program. New Horizons (Baltimore, Md.), 2(3), 321–325.

Stoica, P., & Moses, R. L. (2005). Spectral analysis of signals. Pearson/Prentice Hall.

Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). Why self-attention? A targeted evaluation of neural machine translation architectures. ArXiv Preprint ArXiv:1808.08946.

Teasdale, G., Maas, A., Lecky, F., Manley, G., Stocchetti, N., & Murray, G. (2014). The Glasgow Coma Scale at 40 years: Standing the test of time. The Lancet Neurology, 13(8), 844–854.

Thorsen-Meyer, H.-C., Nielsen, A. B., Nielsen, A. P., Kaas-Hansen, B. S., Toft, P., Schierbeck, J., Strøm, T., Chmura, P. J., Heimann, M., Dybdahl, L., Spangsege, L., Hulsen, P., Belling, K., Brunak, S., & Perner, A. (2020). Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. The Lancet Digital Health, 2(4), e179–e191. https://doi.org/10.1016/S2589-7500(20)30018-2

Vallat, R. (2018). Bandpower of an EEG signal. https://raphaelvallat.com/bandpower.html

Vasilevskis, E. E., Kuzniewicz, M. W., Cason, B. A., Lane, R. K., Dean, M. L., Clay, T., Rennie, D. J., Vittinghoff, E., & Dudley, R. A. (2009). Mortality Probability Model III and Simplified Acute Physiology Score II. Chest, 136(1), 89–101. https://doi.org/10.1378/chest.08-2591

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. ArXiv:1706.03762 [Cs]. http://arxiv.org/abs/1706.03762

Viswanathan, M. (2017). Digital modulations using Matlab: Building simulation models from scratch (Black&White edition). publisher not identified.

Vold, M. L., Aasebø, U., Wilsgaard, T., & Melbye, H. (2015). Low oxygen saturation and mortality in an adult cohort: The Tromsø study. BMC Pulmonary Medicine, 15(1), 9. https://doi.org/10.1186/s12890-015-0003-5

Wairagkar, M., Hayashi, Y., & Nasuto, S. J. (2019). Modeling the Ongoing Dynamics of Short and Long-Range Temporal Correlations in Broadband EEG During Movement. Frontiers in Systems Neuroscience, 13. https://doi.org/10.3389/fnsys.2019.00066

Wang, X. L., Eklund, J. M., & McGregor, C. (2014). Parametric Power Spectrum Analysis of ECG Signals for Obstructive Sleep Apnoea Classification. 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, 8–13. https://doi.org/10.1109/CBMS.2014.37

Wei, W. W. (2006). Time series analysis. In The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2.

Winkelman, J. W., Tanasijevic, M. J., Wybenga, D. R., & Otten, J. (1997). How Fast Is Fast Enough for Clinical Laboratory Turnaround Time?: Measurement of the Interval Between Result Entry and Inquiries for Reports. American Journal of Clinical Pathology, 108(4), 400–405. https://doi.org/10.1093/ajcp/108.4.400

Woyczyński, W. A. (2019). Power Spectra of Stationary Signals. In A First Course in Statistics for Signal Analysis (pp. 175–191). Springer.

Zhai, Q., Lin, Z., Ge, H., Liang, Y., Li, N., Ma, Q., & Ye, C. (2020). Using machine learning tools to predict outcomes for emergency department intensive care unit patients. Scientific Reports, 10(1), 1–10.

Zhang, Y., Ji, X., Liu, B., Huang, D., Xie, F., & Zhang, Y. (2017). Combined feature extraction method for classification of EEG signals. Neural Computing and Applications, 28(11), 3153–3161.

Zhang, Y., Liu, B., Ji, X., & Huang, D. (2017). Classification of EEG signals based on autoregressive model and wavelet packet decomposition. Neural Processing Letters, 45(2), 365–378.

Zhou, T. (2013). Oscillation Amplitude. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), Encyclopedia of Systems Biology (pp. 1616–1616). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_523

Zimmerman, J. E., & Kramer, A. A. (2014). A history of outcome prediction in the ICU. Current Opinion in Critical Care, 20(5), 550–556.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., & Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. Critical Care Medicine, 34(5), 1297–1310.

Appendix A: Research Setup and Parameter Selection

Implementation of machine learning models

In this research, the probabilistic classification problem $Pr(Y|X)$ described in Section 3 and the four classifiers, including linear support vector machine (LinearSVM), logistic regression (LR), random forest (RF), and fully connected neural networks (NN) can be implemented as follows.

Linear SVM: Linear SVM defines a hyperplane function, f, to make predictions, $f(x_i) = 1$, if $w^T \cdot x + b \geq 0$; $f(x_i) = 0$ otherwise. The LinearSVM does not explicitly predict a probability. The probability of each class is estimated by cross-validation (Platt, 2000). The parameter $\theta = \{w, b\}$ can be obtained by optimizing $\underset{w,b}{min} \frac{1}{2}\Omega(w) + C\sum_i max[0, y_i(w^T \cdot x_i + b)]$, where C is a regularization parameter, and Ω is a penalty function of parameter w. Ω can be $\Omega(w) = |w|$, which is called the $l1$ penalty.

LR: LR is a linear classifier, whose mapping function f is defined as $f_\theta(x) = Pr(Y = 1|\, X = x) = \frac{1}{1+e^{-(w^T \cdot x+b)}}$, where $\theta = \{w, b\}$ is the parameter, and can be obtained by maximizing the likelihood function $L(\theta|Y, X) = \prod_i f_\theta(x_i)^{y_i}(1 - f_\theta(x_i))^{(1-y_i)}$.

RF: A decision tree classifier $h(\boldsymbol{x}|\boldsymbol{\theta})$ partitions the feature space on nodes to group the samples with the same labels, where θ determines the subset X_θ of the full set of the features X. The splitting on the node depends on the average feature impurity that can be measured by entropy or the Gini index. Entropy is the amount of information present in certain variables, defined as $Entropy = -\sum_{i=1}^{n} p_i * log(p_i)$, where p_i is the probability of class i in the samples on the node. The Gini index measures sample inequality, is defined as $Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2$. A random forest classifier with size k is based on k decision tree classifiers $h_i(\boldsymbol{x}|\boldsymbol{\theta_i})$, $i = 1, \ldots, k$, where θ_i is determined by bootstrap sampling. Each tree estimates $Pr(Y = 1|\, X_{\theta_i})$

and $Pr(Y = 0|\, X_{\theta_i})$ by taking the observed proportions of each class where the tree stops splitting. The random forest classifier is defined as $f(x|\theta) = arg\ max_y \frac{1}{k}\sum Pr(Y|X_{\theta_i}), Y \in \{0, 1\}$.

NN: NN is a non-linear classifier that includes at least a single hidden layer, an input layer and an output layer. Here we give the example of NN with one layer. Mathematically, $f_\theta(x) = g_{out}(w_{out}^T \cdot g_{hidden}(w_{hidden}^T \cdot x + b_{hidden}) + b_{out})$, where g_{hidden}, g_{out} are activation functions, and $\theta = \{w_{hidden}, b_{hidden}, w_{out}, b_{out}\}$ are parameters for hidden and output layers. The activation functions are defined as $g_{hidden}(z) = max(0, z)$ and $g_{out}(z) = \frac{1}{1+e^{-z}}$. The parameter θ can be optimized by minimizing the cross-entropy $H_\theta = \sum_i -y_i log(f_\theta(x_i)) + (1 - y_i) log(1 - f_\theta(x_i))$.

The parameters of these machine learning classifiers are grid-searched in the experiments. The parameters are shown in Table 3.12.

Table 3.12. Classifiers' Parameters.

Classifiers	**Parameters**
Logistic regression	Penalty: *l*1
	C: 0.005, 0.01, 0.05, 0.1
Linear support vector machine	Penalty: *l*1
	C: 0.005, 0.01, 0.05, 0.1
Neural network	hidden_layer_sizes: (100, 50, 30), (30, 20, 10), (100, 30), (50, 30), (30, 10), (30, 10), (50,), (30,)
Random forest	n_estimators: 400, 800, 1600
	max_depth: 2, 4, 8, 16, 32
	criterion: gini, entropy
LSTM (Thorsen-Meyer et al., 2020)	num_hidden_layers: 1, 2, 3
	hidden_layer_size: 5, 10, 15
	Optimizer: Adam (learning rate lr = 0.001)

Table 3.12. Continued

Classifiers	**Parameters**
GRU (Che et al., 2016)	num_hidden_layers: 1, 2, 3
	hidden_layer_size: 5, 10, 15
	Optimizer: Adam (learning rate lr = 0.001)
CNN (S. Y. Kim et al., 2019)	num_conv_layers: 2
	conv_filter_size: 1×5
	num_conv_channel: 256, 512
	num_pooling_layers: 2
	pooling_filter_size: 1×5
CNN (Caicedo-Torres & Gutierrez, 2019)	num_conv_layers: 1
	conv_filter_size: 1×3, 1×6, 1×12
	num_conv_channel: 16
	num_pooling_layers: 1
	pooling_filter_size: 1×3

Parameters for feature extraction

Relative extrema of frequency spectrums. We conduct a prior experiment to select the parameters (i.e., n and ε) for relative extrema (Section 3.2) for the frequency spectrums. The experiments predict the mortality for heart failure patients using 24 hours data. We perform a grid search of parameter pairs for the random forest algorithm with entropy as impurity measurements. The AUC scores are shown in Table 3.13. When $n = 5$, the classifier achieves the best AUC most of the time. Thus, n is set to 5. With different ε, we can extract different features. Our proposed feature set includes different relative extrema features extracted using various ε (i.e., $\varepsilon \in \{2,4,8\}$).

Table 3.13. Parameter Selection for n and ε.

AUC	ε = 2	ε = 4	ε = 8
n = 3	0.745	0.752	0.750
n = 5	**0.758**	**0.763**	0.764
n = 7	0.757	0.759	**0.774**

Extreme values on vital signs. As statistical features, we take the extreme values on the vital signs through the moving window method (Section 3.2). The window size is set to 10, which covers vital signs for 50 minutes. According to the medical professional's recommendations, the selected window size is large enough to estimate the local mean and variance of the vital signs. The parameter n is set to be 5 empirically because 43.43% of the vital signs have at least 5 extreme values in the moving windows. Such settings can include as many useful features as possible, while producing as few missing values as possible.

Power-in-band. Before we present the parameter setting of the power-in-band features, we first highlight the importance of power-in-band features from two aspects. First, power-in-band is the summary measure of the "strength" of a signal (i.e., vital signs are defined as signals in our research). (1) In signal processing, a signal is viewed as a function of time. "Power of a signal" is used to represent "strength of the signal". (2) In signal detection techniques, the strength of a signal (also known as energy) is often considered as the computation of the area under the square of the signal $E_{signal} = \sum_{n=-\infty}^{\infty} |signal(n)|^2$ (Viswanathan, 2017). Please see Figure 3.13.

According to Parseval's theorem (Figure 3.14.), the energy of the time domain signal is equal to the energy of the frequency domain transform (Kay & Marple, 1981). "Power" is also the measure of signal strength, which is defined as the amount of "energy" consumed *per unit time* $P_{signal} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{n=N} |signal(n)|^2$. Similarly, according to Parseval's theorem, the power in a signal when expressed in the time domain is equal to the power of that same signal

when expressed in the frequency domain. The power-in-band feature summarizes the contribution of the given frequency band (i.e., frequency range) to the overall power of the signal, which may contain important patterns of patients' vital signs.

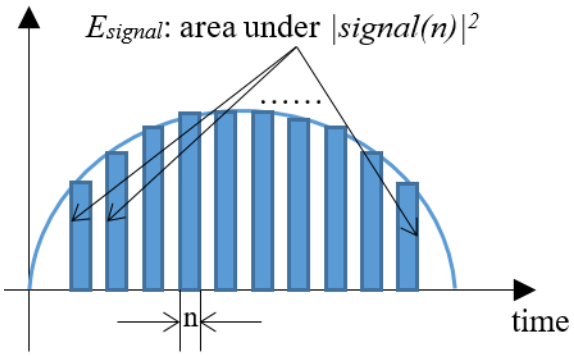


Figure 3.13. Strength (Energy) of the Signal.

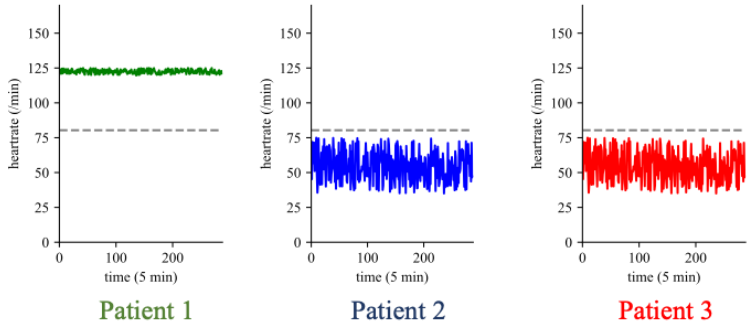
$$\sum_{n=-\infty}^{\infty} |signal(n)|^2 = \sum_{f=-\infty}^{\infty} |signal(f)|^2$$

energy of the time domain — energy of the frequency domain

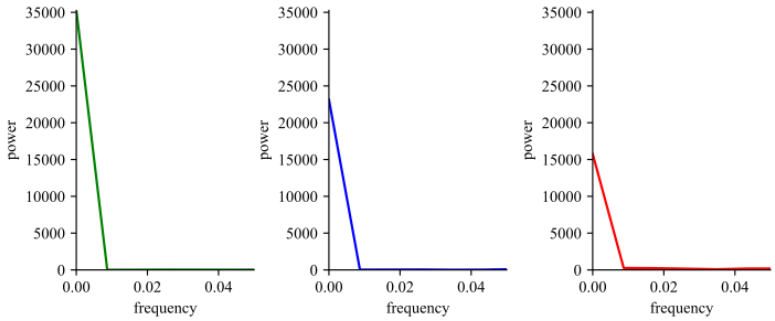
Figure 3.14. Parseval' Theorem.

Second, power-in-band is a simple yet powerful feature extraction method for ICU mortality prediction. (1) Power-in-band features are easy to extract and use. There are a variety of methods that can be used on the frequency spectrums once we convert patients' time series of vital signs to the frequency domain. Power-in-band is one of the easiest to analyze (Vallat, 2018). In practice, we compute the summation over different segments of a vector (i.e., the vector represents the frequency spectrums transformed from a vital sign). (2) Power-in-band features are useful for machine learning predictive analysis. The time/frequency representations (i.e. frequency spectrums generated by converting time series of vital signs to frequency domains) are of high dimensionality, making them unsuitable for use as classifier inputs. Power-

in-band extracts the key characteristics from the time/frequency representations, resulting in a single number that describes a specific aspect of the time/frequency representations.



(a) Time series of vital sign (heart rate)



(b) Power spectrum

Feature	Center frequency	Bandwidth	Power in band	Note
Patient 1	(2.08, …)	0.00208	(15000, …)	a large number in high frequency band
Patient 2	(2.08, …)	0.00208	(23000, …)	a number close to the center of the distribution of features in high frequency band
Patient 3	(2.08, …)	0.00208	(35000, …)	a small number in high frequency band
…	…	…	…	…

(c) Power in band features

Figure 3.15. An Example of Power-in-band Features for Different Heart Rates.

Let us use a commonly used vital sign, heart rate, as an example. Figure 3.15. depicts three patients with different heart rates. Abnormal heart rates can have causes that are due to underlying diseases. Different ranges of heart rates result in different powers in the frequency

domain. If a patient experiences tachycardia (i.e., fast heart rate, a heart rate over 100 beats per minute), the value of power-in-band feature in higher frequency bands is larger compared to the patients with normal heart rate. If a patient experiences bradycardia (i.e., slow heart rate, slower-than-expected heart rate, generally beating fewer than 60 beats per minute), the value of power-in-band feature in lower frequency bands is smaller compared to the patients with normal heart rate. To sum up, the power-in-band features are used to summarize the frequency spectrum on the frequency domain, in order to communicate a large amount of information in a simple way.

In order to select the proper center frequencies ω_c and bandwidth ω_{bw} for the power-in-band features, we conduct experiments on 24 hours ICU vital sign data of heart failure patients. The classifier is the random forest, and the impurity measurement is entropy. We obtain fixed-length vectors of frequency spectrums after we transform the time-series of vital signs to the frequency domain. We explore our model's performance by splitting the vectors into n bands of equal length ($n \in \{4, 5, 6\}$).

The center frequencies ω_c and bandwidth ω_{bw} are determined by the number of bands n. When the number of bands n is set to 4, there are four center frequencies (i.e., the center frequency is the middle point of each band). The width of each band ω_{bw} can be obtained accordingly. Similarly, when the number of bands n is set to 5 or 6, we calculate the center frequencies and bandwidth correspondingly. The values of the parameters and the prediction results are summarized in Table 3.14. The experimental results show that when the number of bands n is set to 5, the proposed method achieves the best empirical results. As a result, in this study, the center frequency ω_c is (1.67, 5.00, 8.33, 11.67, 15.00 ($* 10^{-4}$)) and the bandwidth ω_{bw} is $1.67 * 10^{-4}$.

Table 3.14. Power-in-band Parameter Selection.

Number of bands	Bandwidth ω_{bw}	Center frequencies $\omega_c(*\mathbf{10^{-4}})$	AUC
n = 4	$2.08 * 10^{-4}$	(2.08, 6.25, 10.41, 14.58)	0.706
n = 5	$1.67 * 10^{-4}$	(1.67, 5.00, 8.33, 11.67, 15.00)	**0.756**
n = 6	$1.38 * 10^{-4}$	(1.39, 4.17, 6.94, 9.72, 12.50, 15.26)	0.707

ARMA and ARIMA. For the baselines ARMA and ARIMA, we set the parameters $p = 4, d = 0, q = 0$, where p is the number of lag observations in the model, d is the number of times the raw observations are differentiated, and q is the size of the moving average window. The parameters are first generated for each patient's vital sign with the lowest Akaike information criterion (AIC). For the classification task (i.e., mortality prediction), a fixed order for the time series model is required for all patients to ensure the input features have the same dimensionality. We choose the most frequent p, d, and q. The ARIMA parameter selection process follows previous work (Carden & Brownjohn, 2008; Y. Zhang et al., 2017).

Missing values in the feature set

In the experiments, we impute missing values in the feature set, including the relative extrema on the frequency spectrums, with the population means. An alternative way of the imputation is filling zero, e.g., using 0 to fill in the relative extrema that are missing. Zero indicates no relative extrema, which is more close to the meaning of "no local minima and maxima". However, filling the missing value with 0 can be misleading and introduce noise to the model.

According to its definition, the frequency spectrum must be non-negative. Therefore, zero local extrema of the frequency spectrums can be interpreted as no oscillation in the original time series at a certain time point (i.e., local minima or maxima are zero) or a range of time (i.e., local maxima are zero), which is untrue for most vital sign data.

Addressing missing values without losing important information or adding noise is difficult for this dataset. We add additional experiments to empirically test which imputation method is better. The results are summarized in Table 3.15. When we fill missing relative extrema with the population mean, the AUC scores are much higher than filling the missing relative extrema with zero, regardless of the classifiers and features' combinations. It shows filling the population mean is more effective in dealing with missing values in this research.

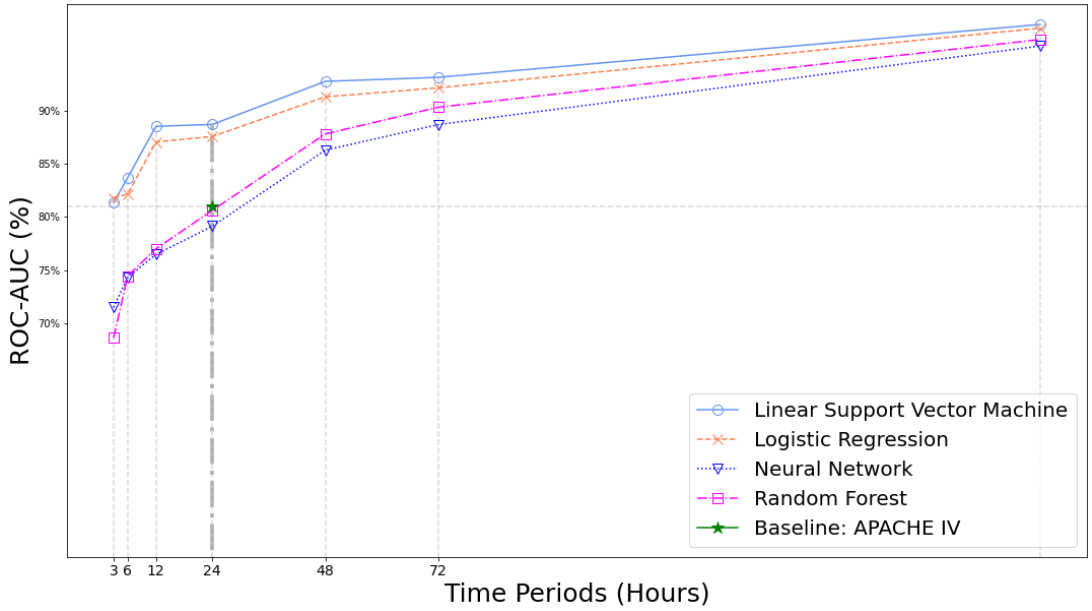
Table 3.15. Experiments to Test Different Imputation Methods.

Feature types	**Classifier**	**AUC (Filling missing values with mean)**	**AUC (Filling missing values with 0)**
Statistical features + Signal processing features (Proposed feature set)	LinearSVM	0.849	0.761
	LogisticRegression	0.843	0.757
	NeuralNetwork	0.827	0.687
	RandomForest	0.806	0.792
Proposed feature set + apache variables (no GCS)	LinearSVM	0.869	0.788
	LogisticRegression	0.865	0.784
	NeuralNetwork	0.852	0.703
	RandomForest	0.817	0.798

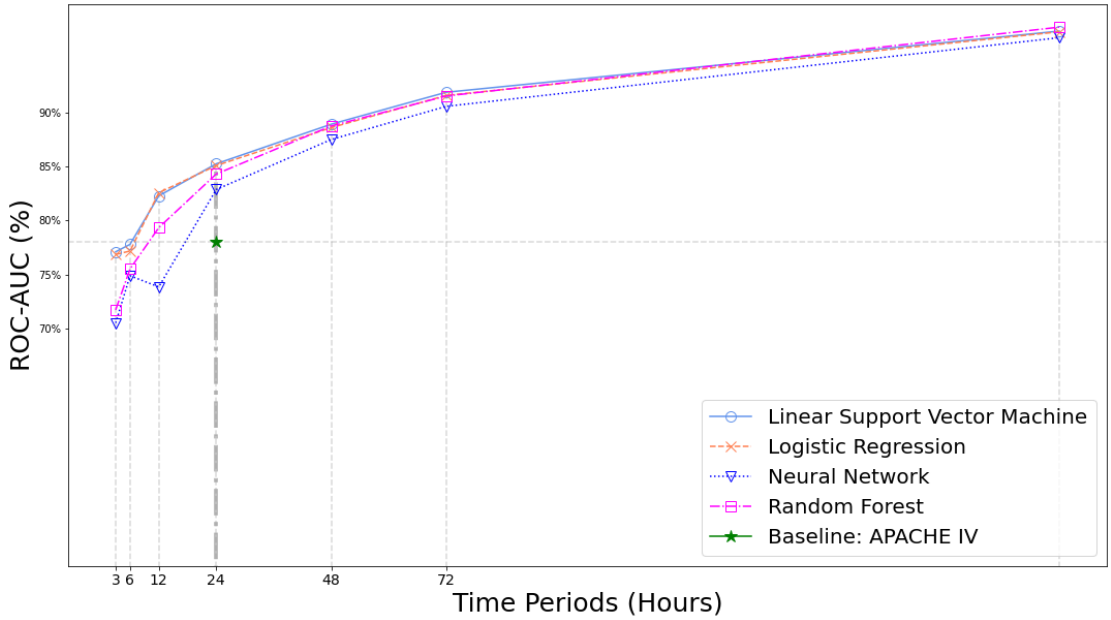
Appendix B: A Supplementary Experiment to Test the Generalizability of the Proposed Method

To further evaluate the reliability and generalizability of the proposed method, we apply our method on two other common ICU admission diagnoses, which are also of top frequent in eICU database - Sepsis pulmonary (SP) and Sepsis renal/UTI (including bladder) (SR). The results are shown in Figure 3.16. In both cases, the AUC for the proposed method increases over time. When applied on SP, the AUC of all machine learning classifiers using 24H data is significantly larger than APACHE IV. For SR, SVM and logistic regression show better

prediction performance than APACHE IV with only 3H data. Overall, the results show our method can be generalized to other diseases for ICU mortality prediction.



(a) Using Sepsis Pulmonary (SP) as a Research Case



(b) Using Sepsis Renal/UTI (Including Bladder) (SR) as a Research Case

Figure 3.16. ICU Mortality Prediction Performance.

CHAPTER 4. A NEW TRANSFER LEARNING METHOD FOR LAB OUTCOME PREDICTION WITH LIMITED TRAINING DATA

Shaodong Wang[1], Yiqun Jiang[1], Chao He[2], Qing Li[1], and Wenli Zhang[3]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University

[2] Department of Medicine, University of Alabama at Birmingham

[3] Department of Information Systems & Business Analytics, Iowa State University

Abstract

The laboratory test is a key resource in ICU to inspect the patient's health. Due to the limited lab resources and inappropriate lab utilization, researchers start to predict the lab outcomes. However, some lab tests do not have sufficient training data (low-frequency lab tests), which negatively impact the prediction models' accuracy. In this study, we propose a new transfer learning method for lab anomaly prediction with limited training data. Specifically, we develop a novel distance to select the optimal source domain from multiple high-frequency lab tests. We design a recurrent neural network to estimate the probability of obtaining an abnormal lab outcome. We transfer knowledge from the selected source domain to improve the model performance on the target domains (low-frequency lab tests). We evaluate the proposed method on five low-frequency lab types that are related to heart failure and five high-frequency lab types that are most common in the hospital. The experiments show that the designed neural network outperforms all traditional machine learning models by a large margin. In the experiments, the transfer learning and the proposed domain distance further improve the model performance for all selected low-frequency lab types (e.g., AUC scores increase from 0.729 to 0.795 for Brain natriuretic peptide tests). The new transfer learning method address the data insufficiency

problem for lab outcome prediction, which provide a more reliable way to optimize the clinical resource allocation.

Introduction

The laboratory test is a key resource in ICU to inspect the patient's health. 1) Doctors can select proper treatment according to the lab results. For example, calcium is usually given into veins (IV) to treat the muscle and heart effects of high potassium levels[24]. 2) The lab results are also important factors for clinical risk predictions, such as mortality prediction and length of stay. Many well-known ICU scoring systems include lab results as predictive attributes, including APACHE, and SAPS (Keegan et al., 2012).

However, there are some issues about limited lab resources and inappropriate lab utilization. 1) Laboratory resources are not always sufficient, especially in developing regions. Many countries and districts suffer from insufficient laboratory tests due to the lack of laboratory supplies, essential equipment, skilled personnel, educators and training programs, etc (Olmsted et al., 2010; Petti et al., 2006). During the recent Covid pandemic, lab resources are even more limited because hospitals receive more patients than usual (Moghadas et al., 2020). 2) Some lab tests are unnecessary, which wastes clinical resources. There are no standard definitions (Jha et al., 2009), but unnecessary lab tests are usually referred to as repeating tests without assessing the clinical necessity or the repeating tests in short time intervals that are unlikely to provide useful information (e.g., clinically significant change) (Baron & Dighe, 2014; Konger et al., 2016). According to the Institute of Medicine, unnecessary tests and procedures waste over $200 billion each year in the US (Smith et al., 2013; Xu et al., 2019). Up to 42% of lab tests are unnecessary or redundant (Konger et al., 2016), which could take 2.7% of the total inpatient

[24] https://www.heart.org/en/health-topics/heart-failure/treatment-options-for-heart-failure/hyperkalemia-high-potassium

costs (Jha et al., 2009). 3) Besides the financial consequences, unnecessary lab tests can also cause adverse impacts on patients, including lower patient satisfaction and higher mortality (Xu et al., 2019). For example, a high frequency of blood draws can negatively impact the patients' sleep (Konger et al., 2016), or even cause hospital-acquired anemia in some extreme cases (Huck & Lewandrowski, 2014). 4) Occasionally, inexperienced doctors may neglect some helpful lab tests when the tests are not listed in the protocol. This could impede the doctors from getting enough information for the diagnosis and treatments.

Due to the issues mentioned above, researchers propose to predict the lab results before the tests are ordered. The proposed work can be used in two ways to prioritize the lab resources, including financial cost, practitioners' workload, medical facilities, and lab materials. First, the lab anomaly prediction models can quantify the expected information (Xu et al., 2019). According to the probability of getting abnormal lab results, doctors can avoid unnecessary tests or increase the priority of important tests. Second, the lab anomaly prediction models can remind doctors to order necessary lab tests. If the models can continuously process the collected information and provide the real-time estimation of lab results, the model will be able to alarm the doctors when the predicted results change significantly.

Researchers have implemented various machine learning models for the lab anomaly prediction. Some propose to predict the results of a single test, such as ferritin tests and hemoglobin (Hgb) tests (Lobo et al., 2020; Luo et al., 2016). Others propose more general models to predict multiple lab results, including 12 common blood tests (Yu et al., 2020). However, researchers fail to notice that some tests' training data are insufficient, affecting the prediction models' accuracy. For example, in the study of Xu et al., (2019), among 1,000 patients, there are only one anti-hiv test, 26 glucose tests, and 40 iron tests, while the volume of

Magnesium tests and prothrombin time tests are 4,246 and 2,244, respectively. Consequently, the AUC scores of the tests with limited training samples in the prediction are significantly lower than that of the large-sample tests, as shown in Figure 4.1.

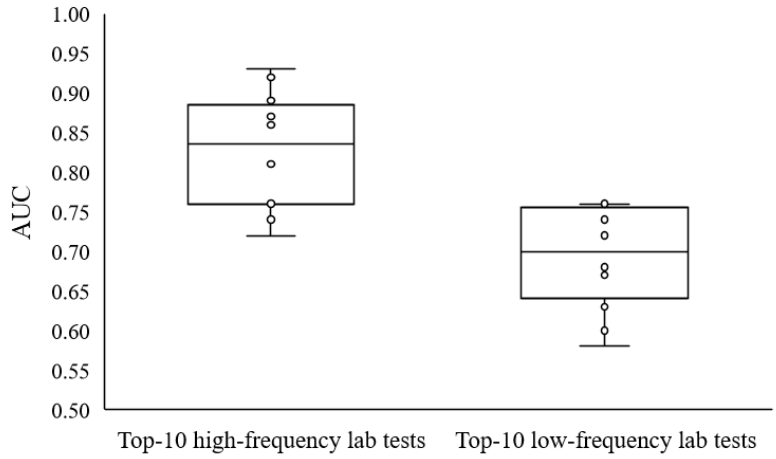


Figure 4.1: Prediction performance (AUC scores) of top-10 high-frequency & low-frequency lab tests in Xu et al., (2019).

Although some lab types do not have many cases in the hospital, they still need anomaly prediction models for resource optimization. In the rest of the article, we call the lab types that only have limited cases in the hospital low-frequency lab tests. For patients who need those low-frequency lab tests, avoiding a single unnecessary lab test can help. First, some lab tests are repeatedly taken for a small patient cohort. According to the eICU database, there are only 0.06% patients receiving Lidocaine tests. However, patients can receive 33 Lidocaine tests at most in a single stay. Second, low-frequency lab tests can also be expensive. For example, each iSTAT Cg4 costs $572.5 in the Chargemaster (Xu et al., 2019), which is more costly than most high-frequency tests. Third, some low-frequency lab tests are time-consuming. For example, the turnaround time of stool culture is 1 or 2 days. Last, some low-frequency tests can also make patients uncomfortable. For instance, patients are required to fast for at least 8 hours before taking the Folic acid blood test.

It is important to address the data insufficiency for low-frequency lab tests on the lab anomaly prediction model. In this study, we propose to fill the research gap by adopting transfer learning, which is the popular tool for small-size datasets in the machine learning fields. Transfer learning can increase the training efficiency and model accuracy for a target domain by transferring knowledge from a related source domain. In this study, the target domain is a low-frequency lab type, while the source domain is a high-frequency lab type with many samples. Transfer learning is promising because there are usually a lot of high-frequency lab types in hospitals and many lab tests are clinically related to each other (Bartsch et al., 2015). However, a not well-related source domain (a high-frequency lab type) can negatively impact the model performance on the target domain (Weiss et al., 2016). How to select the proper lab test as a source domain remains a challenge in this study.

In this study, we propose a new transfer learning method for low-frequency lab anomaly prediction. As the first step of the transfer learning, we develop a novel distance to select the optimal source domains, which measures the closeness between two domains. Then we design a neural network as a base model for lab anomaly prediction. The designed neural network is pre-trained and finetuned on the source domain (high-frequency lab tests) and the target domain (low-frequency lab tests), respectively. We evaluate the proposed method on five low-frequency lab types that are related to heart failure and five high-frequency lab types that are most common in the hospital. The experiments show that the designed neural network outperforms all traditional machine learning models by a large margin. In the experiments, the transfer learning and the proposed domain distance further improve the model performance for all selected low-frequency lab types (e.g., AUC scores increase from 0.729 to 0.795 for Brain natriuretic peptide tests).

Our important contributions are summarized as follows. 1) We are the first to develop the anomaly prediction model for low-frequency lab tests. We address the data insufficiency problem and improve the model accuracy for low-frequency lab tests. 2) The proposed transfer learning method provides a specific guideline to select the proper source domains from multiple high-frequency lab types. The selected source domain boosts the model performance efficiently. 3) Practically, the proposed work can optimize the clinical resource allocation by providing the expected lab outcome, especially the low-frequency lab tests.

In the remainder of this article, we review the related literature in Section 2. We introduced the proposed method in Section 3. We evaluate the proposed model in Section 4. We summarize the proposed work, limitations, and future work in Section 5.

Related Work

Lab Results Prediction

In order to optimize the strategy of clinical resources, many researchers proposed to predict the lab results using available features before the lab order, such as demographics, vital signs, and past lab results.

Most researchers built machine learning models for common or high-frequency lab tests with large sample sizes. For example, Xu et al., (2019) trained a recurrent neural network to identify hemoglobin (Hgb) levels using over 40K Hgb records. Yu et al. (2020) even fitted a Long Short Term Memory model on 598K laboratory observations.

However, only a few studies predict outcomes of low-frequency lab tests that have limited training samples. Voglis et al. (2020) identified the probability of hyponatremia for patients who underwent pituitary surgery. Due to the specialty of the cohort, only 207 samples were included in this study. Moreover, in the existing work, the small sample size limited the model performance (Xu et al., 2019). In the study of Xu et al., (2019) (Figure 4.1.), the model

performance (AUC scores) of top-10 high-frequency lab tests (>18K samples) was significantly higher than that of top-10 low-frequency lab tests (<1.5K samples). This observation motivated us to improve the lab prediction for low-frequency lab tests with limited samples.

Transfer Learning

Transfer learning is a promising technique to improve a machine learning model on one domain (target domain) by transferring knowledge from a different but related domain (source domain) (Weiss et al., 2016). As an intuitive example, people skilled at guitar (source domain) usually learn piano (target domain) more efficiently than those without any music background.

Transfer learning aims to solve the problem of data insufficiency (Zhuang et al., 2021). In many real-world scenarios, training data of the target domain are difficult and expensive to collect, which greatly limits the performance of the resultant machine learning models (Zhuang et al., 2021). By transferring information from the source domain, transfer learning decreases the required sample size in the target domain and improves learning performance (Weiss et al., 2016; Zhuang et al., 2021).

Transfer learning methods can be divided into four categories, including instance-based, mapping-based, network-based, and adversarial-based (Weiss et al., 2016). The instance-based method reweights the instances in the source domain to minimize the distribution difference between two domains. The feature-based method maps data in the source domain to the target domain or maps data in both domains into the same space. The parameter-based method reuses the shared parameters of source and target domains by reweighting multiple source learners. The relational-based method transfer knowledge from source domain to target domain through their defined relations.

In this study, we adopt feature-based transfer learning because it is widely used for deep learning models. In deep learning, the feature-based transfer learning is also called network-

based transfer learning (Tan et al., 2018). Specifically, we pre-train the network on the source domain, and then finetune the model on the target domain with the pre-trained model as a starting point. For example, Huang et al. (2020) utilized Google Inception-V3 convolutional neural network as a pre-trained model and identified the location of the anterior ethmoidal artery on sinus computed tomography scans. Shi et al. (2018) pre-trained a convolutional neural network on a huge image dataset and finetuned their model to predict occult invasive disease in ductal carcinoma.

Transfer learning is promising in our study because it fits our research goal for two reasons. First, the sample size of low-frequency lab tests is too small to support the training, while there are sufficient samples of high-frequency lab tests. Second, many lab tests are clinically related, especially when they examine the same body system or the same health condition (Hosten, 1990). The relatedness is important because a not well-related source domain can negatively impact on the target learner (Weiss et al., 2016). Therefore, transfer learning could be a promising way to address the data insufficiency problem and improve the model performance for low-frequency lab tests.

Domain Distance

Usually, there are many types of high-frequency lab tests in hospitals, which can be clinically related or non-related to the target low-frequency lab. On one hand, some lab tests examine the function of the same body system. For example, serum creatinine and blood urea nitrogen are both the common tests to measure the kidney function or damage (Hosten, 1990). These clinically related lab tests are more likely to improve the predictive power for the target lab type. On the other hand, some lab tests diagnose extremely different health conditions, such as HIV tests and diabetic tests. The data of unrelated lab tests are more likely to add noise to the target domain and negatively impact the target learner (Weiss et al., 2016).

Therefore, how to select the right lab samples remains a problem. Besides basic transfer learning, we propose selecting the right source domains (high-frequency lab tests) according to the closeness between high-frequency and low-frequency lab types. We assume the source domains similar to the target domains can most help the training on the target domain.

In the field of transfer learning, many distance metrics were used to measure the distance between two domains, such as Maximum Mean Discrepancy, KL-divergence, and H-divergence (Ben-David et al., 2010; Li et al., 2021). Most researchers calculated the distance on features in different domains or on the fully connected layers of neural networks (Li et al., 2021; Long et al., 2016; Zhuang et al., 2021).

Among the current methods, CORrelation ALignment (CORAL) distance is a widely used distance metric that measures the discrepancy between correlations of two feature sets (Sun & Saenko, 2016). Denote the feature sets of two domains as $X_1$and X_2. The CORAL distance is defined as $CORAL = \|C_1 - C_2\|_2^2$, where $C_i = Cor(X_i, X_i)$, $i = 1, 2$. C_i is a matrix, and each element of C_i is the Pearson Correlation Coefficient (Benesty et al., 2009) between two features in X_i.

CORAL is a general distance that can be implemented on datasets without labels. It didn't make use of the relationship between features and labels. However, this is important for classification problems, such as lab result predictions. Therefore, we propose a new distance metric by modifying the CORAL distance. Our distance measures the discrepancy between correlations of features and labels.

Neural Network

In this study, we designed a neural network as our model, including fully connected layers and long short-term memory (LSTM) layers. We select the neural network due to its superior performance on the classification tasks (Esteva et al., 2019).

Fully connected layers are a common form of neural network that takes a vector (tabular features) as input. The inputs are connected with each unit of the next layer in the network. Fully connected layers have been successfully implemented in many clinical tasks, such as predicting community-acquired pneumonia (Feng et al., 2021) and identifying substance use risk (Hassanpour et al., 2019).

Long Short-Term Memory networks (LSTM) are a special kind of neural network designed for long-term time series data (Hochreiter & Schmidhuber, 1997). Unlike fully connected networks, LSTM takes a sequence of data iteratively and stores the information in its hidden and cell states. LSTM controls the flow of a long data sequence with a series of gates, which enables it to remember long-term information. LSTM is promising in our work because long data sequences are available in many real-world cases. For example, hospitals commonly collect vital-sign sequences, especially in Intensive Care Units (ICU). Many researchers have adopted LSTMs on vital signs to address clinical problems, including identifying clinical deterioration (Naemi et al., 2020) and detecting influenza, dengue, and common cold (Nadda et al., 2022).

Research Design

In this section, we propose a new transfer learning method for low-frequency lab anomaly prediction. We first develop a domain distance to measure the closeness between two lab types, which can be used to select the proper source domain in transfer learning. Then we design a recurrent neural network as a base model of transfer learning. Lastly, we pre-train the

designed base model on the selected source domain and finetune the model on the target domain. Note the source domain and the target domain are the high-frequency lab tests and low-frequency lab tests, respectively.

Generally, we design the inputs and outputs of the anomaly prediction model as follows. The inputs are the features that can be collected before the lab tests. In this study, our input features include patient demographics, the existing lab tests, and the vital signs during the last 24 hours before the target lab tests. Given the input features, the model is supposed to detect or predict if the patients will have abnormal lab results.

Formally, we denote the input features as $X_i = \{x_0, x_t\},\ t = 1, 2, \dots, T, i = 1,2, \dots, n,$ where n is the number of samples in the dataset, x_0 is tabular features, and x_t is time series features. t is the time stamp of every hour, and T is the length of the time series, which is 24 in this study. Both x_0 and x_t are vectors. x_0 includes patient demographics and existing lab tests, and x_t is the vital signs at t^{th} hour. We denote the lab prediction model as f, and the output, $\hat{y} = f(X)$ indicates the probability of obtaining an abnormal lab result.

Domain Distance

Usually, there are more than one types of high-frequency lab tests in hospitals that can be our source domains. A similar source domain can improve the model performance on the target domain, while an extremely different source domain can impede the model performance (Weiss et al., 2016). Therefore, a scientific domain selection method is needed to select a proper source domain from multiple lab types.

Inspired by Sun & Saenko (2016), we propose a correlation-based distance metric to select the right lab type as our source domain in the transfer learning. Note that each domain here represents the dataset of a lab type. The distance measures the closeness between the source

domain and target domains, which are the high-frequency lab test and low-frequency lab test, respectively. The source domain (high-frequency lab tests) that are close to the target domain (low-frequency lab test) can improve the model.

Generally, we first calculate the correlation between features and the outcomes in each domain (i.e., a dataset of a lab type). Then we compare the distance between correlations from two domains as the domain distance. The distance measures the correlation difference between two domains. For example, if the outcomes of two lab types simultaneously positively (or negatively) correlate to the features, then their correlations between features and outcomes are similar and thus our distance metric gives them a small distance value.

Specifically, the distance between two types of labs is defined as

$$Distance(lab_1, lab_2) = \|C_{lab1} - C_{lab2}\|_2^2,$$

where $C_{labi} = Cor(X_{labi}, y_{labi}), i = 1, 2,$ and the Cor is the Pearson Correlation (Benesty et al., 2009). The X_{labi} and y_{labi} are the features and outcomes of each lab. The C_{lab1} and C_{lab2} are vectors whose elements are the Pearson Correlation Coefficients between features and the target labels.

Using the domain distance, we select the high-frequency lab type that is closest to the target domain as our source domain. The proposed distance metric helps us find the optimal source domains for two reasons. On one hand, the machine learning models try to find the relationship between input features and outcomes. On the other hand, our distance metric is based on the correlation between input features and outcomes as well. The selected high-frequency lab type and the target low-frequency lab type have the similar relationship between features and outcomes. Hence the models for the selected source domain and the target domain

should also be similar as well. Therefore, the knowledge in the selected source domain (a high-frequency lab type) can help the model training on the target domain (a low-frequency lab type).

Base Model

We design the following recurrent neural network as the base model of transfer learning (Figure 4.2.). We will train the model through the techniques of transfer learning in the next step. Generally, the model includes three parts, long short-term memory (LSTM) layers, fully connected layers (FCN), and a classification output layer.

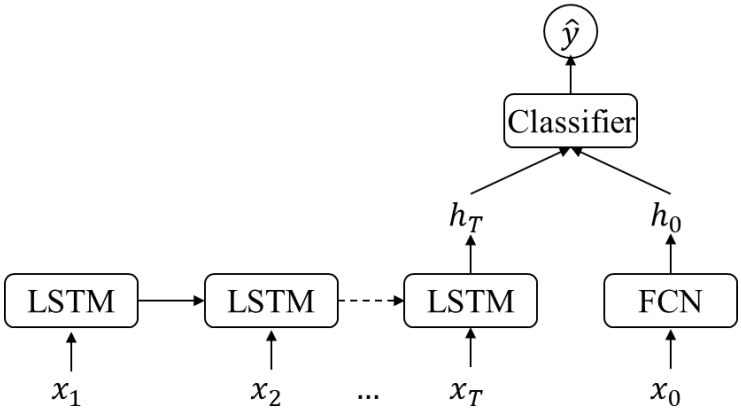


Figure 4.2: The proposed lab prediction model (recurrent neural network). The x_0 is tabular features, such as demographics. The x_t $(t = 1,2, \dots, T)$ is time series features, such as vital signs. $\hat{y}$ indicates the probability of obtaining an abnormal lab result.

We adopted both fully connected neural networks and LSTM because we have both tabular features and time series data as inputs. The model works as follows.

Firstly, the LSTM layers take the time series features x_t as inputs, iteratively. As the hidden features from x_t, the outputs of LSTM, h_T, will be used in the classification layer. Specifically, the LSTM is defined as follows.

$$i_t = \sigma(W_{ii}\, x_t + b_{ii} + W_{hi}\, h_{t-1} + b_{hi})$$

$$f_t = \sigma(W_{if}\, x_t + b_{if} + W_{hf}\, h_{t-1} + b_{hf})$$

$$g_t = tanh(W_{ig}\, x_t + b_{ig} + W_{hg}\, h_{t-1} + b_{hg})$$

$$o_t = \sigma(W_{io}\, x_t + b_{io} + W_{ho}\, h_{t-1} + b_{ho})$$

$$c_t = f_t \odot\ c_{t-1} + i_t \odot\ g_t$$

$$h_t = o_t \odot\ tanh\ (c_t)$$

where h_t and c_t are the hidden state and cell state at time t. The i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively. W and b are the weights and bias to learn. σ is the sigmoid function, $\odot$ is the Hadamard product.

Secondly, the fully connected layer takes the tabular features x_0 as inputs and generate the hidden features for the classification layer. Specifically, the fully connected layer is formulated as $h_0 = W_0 x_0 + b_0$, where h_0 is the generated hidden features. W_0 and b_0 are the learnable weights and bias, respectively.

Lastly, the generated hidden features h_0 and h_T are concatenated and fed into the classification output layer to estimate the predictive lab outcome. The classification layer is formulated as $\hat{y} = W_{classifier}[h_0, h_T] + b_{classifier}$, where the $W_{classifier}$ and $b_{classifier}$ are the learnable weights and bias, respectively.

Transfer Learning

Low-frequency lab tests usually do not have sufficient samples to train complex models, especially deep neural networks. In this case, we need transfer learning to take advantage of the large samples of high-frequency lab tests. Generally, we first pre-train the model on the high-frequency lab samples. Then we start from the pre-trained model and finetune the classification layer. During the finetuning, the fully connected layers and LSTM layers can be regarded as a feature extractor.

Denote θ as all learnable parameters (i.e., all weights W and bias b in the network), and $\theta_{classifier}$ as the learnable parameter in the classification layer (i.e., $W_{classifier}$ and $b_{classifier}$). The loss function is defined as $L(\hat{y}, y) = -(1-y)\, log\, log\, (1-\hat{y}) - y log(\hat{y})$. During the pre-training, we update all parameters in θ by $\theta \leftarrow \theta - \alpha \bigtriangledown_{\theta} L(f(X_{high}), y_{high})$, where η is the learning rate and $\{X_{high}, y_{high}\}$ are the training samples of high-frequency lab tests. During the finetuning, we update the classification parameter $\theta_{classifier}$ by $\theta_{classifier} \leftarrow \theta_{classifier} - \alpha \bigtriangledown_{\theta_{classifier}} L(f(X_{low}), y_{low})$, where η is the learning rate and $\{X_{low}, y_{low}\}$ are the training samples of low-frequency lab tests.

Evaluation

Data Description

We evaluated the proposed model on the eICU database, which contains clinical records in ICU, such as patients' demographics, diagnosis, laboratory tests, and vital signs (Pollard et al., 2018). We took heart failure patients as an example. Table 4.1. described the patient cohort.

We selected five low-frequency lab tests that are related to heart failure as our target low-frequency lab tests, including Brain Natriuretic Peptide (BNP), Iron (Fe), Low-Density Lipoprotein (LDL), Thyroid-Stimulating Hormone (TSH), and Troponin - T. Additionally, we selected top-5 high-frequency lab tests as the source domains for transfer learning, including Potassium, Sodium, Creatinine, Blood Urea Nitrogen (BUN), and Calcium. As shown in Table 4.2., the high-frequency lab tests had at least ten times more samples than the low-frequency lab tests in the eICU database. We defined a laboratory result as abnormal if it is out of the normal range. The normal ranges of lab tests are from clinical guides.

Table 4.1: Patient cohort and vital sign description

Length of stay at ICU (hours)			75.15 (4.03-1232.98)
Age			70.16 on average (19-90)
Gender			51.01 % male, 48.97% female
Ethnicity			71.54% Caucasian, 15.71% African American, 5.26% Hispanic, 1.78% Asian
Height			168.03 cm on average
Weight			89.09 kg on average
Vital signs	Sao2	oxygen saturation, the percentage of available binding sites on hemoglobin that are bound with oxygen in arterial blood	96.17 % on average
	Heart rate	the number of times the heart beats per minute	84.47 on average
	Respiration	the number of breaths a person takes per minute	21.09 on average
	St1	estimated ST segment level 1 of the ECG	1.05 on average
	St2	estimated ST segment level 2 of the ECG	1.78 on average
	St3	estimated ST segment level 3 of the ECG	1.91 on average

Table 4.2: Description of laboratory tests.

Laboratory Test	Normal Range	# Abnormal Outcomes	# Normal Outcomes	# Total Samples
Brain Natriuretic Peptide (BNP)	[0 pg/mL, 125 pg/mL] for patients aged 0-74 years. [0 pg/mL, 450 pg/mL] for age above 75	1255	172	1427
Iron (Fe)	[60 mcg/dL, 170 mcg/dL]	389	62	451

Table 4.2. Continued.

Laboratory Test	Normal Range	# Abnormal Outcomes	# Normal Outcomes	# Total Samples
Low-Density Lipoprotein (LDL)	[65 mg/dl, 180 mg/dl]	232	329	561
Thyroid-Stimulating Hormone (TSH)	[0.3 mIU/L, 3.04 mIU/L]	309	645	954
Troponin - T	[0 ng/mL, 0.01 ng/mL]	910	195	1105
Potassium	[3.5 mEq/L, 5.2 mEq/L]	4228	16266	20494
Sodium	[135 mEq/L, 145 mEq/L]	5532	12418	17950
Creatinine	[0.5 mg/dL, 1.4 mg/dL]	9324	7664	16988
BUN	[7 mg/dL, 20 mg/dL]	13597	3331	16928
Calcium	[8.8 mg/dL, 10.3 mg/dL]	10378	6137	16515

The objective of the model is to identify abnormal results of lab tests in the early stage. Therefore, we extracted input features that are available 24 hours before the target lab tests. The features included time series of vital signs (i.e., heart rate, respiration, sao2, st1, st2, st3), the most recent lab results, demographics (e.g., age and gender), and pre-admission chronic diseases (e.g., AIDS, cirrhosis, leukemia). Please see Table 4.2. for the detailed descriptions of the vital signs.

Experiment setting

We evaluated the model performance on group 5-fold cross-validation. Note that patients might take multiple lab tests during single ICU admission. In order to avoid possible information leakage, we grouped the lab samples in the same ICU admission of a patient on the cross-validation. The lab samples in the same ICU admission would be together either in the training set or testing set.

As baselines, we implemented Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost, and Feedforward Neural Network (FNN) on each low-frequency lab type separately. Consistent with the proposed model, the baselines were evaluated on the group 5-fold cross-validation as well.

Note that these machine learning models were not able to take time series data as input features. Therefore, we extracted statistical features from vital signs (min, max, mean, 25&75-quantile), and concatenated them with the tabular features before the baseline evaluation.

Table 4.3. showed the model parameters we searched. The best parameters were selected according to the Area Under the Curve (AUC) on the cross-validation.

Table 4.3: Parameters in the baseline models and the proposed model.

Model	Parameters
Random Forest (RF)	Max_depth: [1, 2, 4, 8, 16] N_estimators: [400, 800, 1600]
Logistic Regression (LR)	C: [10, 1, 0.1] Penalty: [l1, l2]
Support Vector Machine (SVM)	C: [10, 1, 0.1] Kernel: [rbf, poly]
AdaBoost	N_estimators: [50, 200, 400, 800] Learning_rate: [0.5, 1.0, 2.0]
Feedforward Neural Network (FNN)	Hidden_dim: [32, 64, 128]
Proposed neural network	LSTM_hidden_dim: [64, 128], Classification_hidden_dim: [32, 64, 128]

Prediction Performance

We first calculated the domain distance between high-frequency and low-frequency lab tests. As shown in Table 4.4., for BNP, Fe, LDL, and TSH, the closest source domain (high-frequency lab test) is potassium. The closest source domain of troponin -T is BUN.

For each low-frequency lab (i.e., target domain), we first pre-trained the model on its closest source domain and then finetuned the classification layer of the model on the target domain. Table 4.5. presents the AUC scores on cross-validation.

From Table 4.5., we had two observations. 1) The proposed network outperformed traditional machine learning models by a large margin. For example, without transfer learning, the proposed neural network achieved an AUC score of 0.667, while the AUC scores of all traditional machine learning models were below 0.629. 2) Transfer learning boosted the network greatly. Using the closest high-frequency lab samples as source domains for pre-training, the model increased AUC scores on all target domains. For example, the transfer learning increased the AUC score of the proposed network from 0.857 to 0.901. Generally, we could conclude that the model performance for the low-frequency lab tests could be improved by inducing the knowledge in similar domains. Therefore, the data insufficiency problem of the low-frequency lab tests could be addressed by transfer learning to some extent.

Table 4.4: domain distance between high-frequency lab tests (columns) and low-frequency lab tests (rows).

	BUN	calcium	creatinine	potassium	sodium
BNP	0.004834	0.006530	0.005934	0.002710	0.004796
Fe	0.008867	0.008923	0.013142	0.004017	0.008236
LDL	0.004773	0.006118	0.007612	0.003230	0.004609
TSH	0.004923	0.005260	0.007149	0.001822	0.004309
troponin - T	0.003813	0.005509	0.005682	0.004974	0.006754

Table 4.5: Prediction performance (AUC scores) of baselines and the proposed neural network with/without transfer learning.

	Baselines					Base model without transfer learning	Base model with transfer learning
	AdaBoost	LR	RF	SVM	FNN		
BNP	0.654	0.714	0.694	0.685	0.724	0.729	0.795
Fe	0.528	0.611	0.567	0.552	0.639	0.639	0.681
LDL	0.596	0.629	0.589	0.596	0.633	0.667	0.699
TSH	0.559	0.591	0.568	0.586	0.580	0.606	0.634
troponin - T	0.770	0.870	0.819	0.834	0.855	0.857	0.901

Domain Distance Evaluation

To comprehensively evaluate the proposed domain distance, we implemented the transfer learning with each high-frequency lab type as the source domain. Table 4.6. presented the model performance (AUC scores) of the proposed method with each high-frequency lab type as the source domain. For each target domain, we calculated Spearman's rank correlation coefficient (last column in Table 4.6.) between the AUC scores and its domain distances to the source domain. Spearman's rank correlation coefficient measures the strength and direction of association between the model performance and the proposed domain distance.

Table 6 and Table 4 showed that the closest source domain tended to obtain better predicting performance. For example, potassium was the source domain with the shortest distance to LDL and TSH. Using samples of potassium for pre-training, the model achieved the optimal AUC scores for LDL (0.699) and TSH (0.634). Also, BUN, as the closest source domain, helped the model make the second most accurate predictions for troponin – T (AUC score: 0.901).

Table 4.6: Prediction performance (AUC scores) of the proposed method with different high-frequency lab types as source domains. The last column presents Spearman's rank correlation coefficient between the AUC scores and the domain distances for each low-frequency lab type.

		High-frequency lab types (source domains)					Correlation between AUC and distance
		BUN	calcium	creatinine	potassium	sodium	
Low-frequency lab types (target domains)	BNP	0.802	0.777	0.804	0.795	0.788	-0.1
	Fe	0.607	0.658	0.651	0.681	0.686	-0.6
	LDL	0.630	0.688	0.675	0.699	0.682	-0.5
	TSH	0.583	0.618	0.611	0.634	0.631	-0.7
	troponin - T	0.901	0.886	0.885	0.900	0.911	0.0

Additionally, Spearman's rank correlation coefficients in the last column (Table 4.6.) were lower than or equal to 0. This meant that the closer source domain tended to improve the model performance more on the target domain. Such relationship could be very strong on some target domains. For example, the correlation coefficients on Fe and TSH were as high as 0.6 and 0.7.

To summarize, we could make three conclusions from the experiments. First, the designed neural network outperformed the traditional machine learning models by a large margin. Second, the transfer learning could address the data insufficiency problem and improve the model performance for the low-frequency lab tests. Third, the proposed domain distance could help to select the optimal source domains for the transfer learning in the lab result predictions.

Conclusions and Discussion

To conclude, we designed a neural network compatible with both time series and tabular features for the lab result prediction. Through transfer learning, we addressed the data insufficiency problem for low-frequency lab tests that did not have enough samples to support the model training. We proposed a domain distance that measures the closeness between two lab

tests' datasets. The proposed distance could help to select the optimal source domains (i.e., high-frequency lab tests) that would improve the model performance on the target domain (i.e., low-frequency lab tests).

We evaluated the proposed method on a real-world clinical dataset, eICU. In the experiments, the designed neural network outperformed all traditional machine learning models by a large margin. Using transfer learning and the domain selected by the proposed domain distance, the proposed neural network achieved higher AUC scores significantly.

Not limited to the lab result prediction, the proposed work also has great potential for other clinical models that do not have sufficient training samples. For example, some rare diseases may not have enough data to support the development of onsite prediction models. Potentially, our work could address the data insufficiency of rare diseases by selecting proper domains and transferring knowledge from other disease samples.

While the results are encouraging, the current study has certain drawbacks. First, although the experiments show that the model accuracy on the target domain is negatively correlated to its distance to the source domain, the model accuracy does not strictly increase as the distance decreases in all cases. This means other hidden factors affect the model efficiency, for example, the ratio of positive and negative samples in each domain. Therefore, in future work, we plan to investigate more factors to develop a more comprehensive domain selection process. Second, the current experiments only cover heart failure patients and heart failure-related lab tests. In the future, we plan to conduct experiments on other patient cohorts and lab types to evaluate the generalizability of the proposed model.

References

Baron, J. M., & Dighe, A. S. (2014). The role of informatics and decision support in utilization management. *Clinica Chimica Acta*, *427*, 196–201. https://doi.org/10.1016/j.cca.2013.09.027

Bartsch, R. P., Liu, K. K. L., Bashan, A., & Ivanov, P. Ch. (2015). Network Physiology: How Organ Systems Dynamically Interact. *PLOS ONE*, *10*(11), e0142143. https://doi.org/10.1371/journal.pone.0142143

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. In I. Cohen, Y. Huang, J. Chen, & J. Benesty, *Noise Reduction in Speech Processing* (Vol. 2, pp. 1–4). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Feng, D.-Y., Ren, Y., Zhou, M., Zou, X.-L., Wu, W.-B., Yang, H.-L., Zhou, Y.-Q., & Zhang, T.-T. (2021). Deep Learning-Based Available and Common Clinical-Related Feature Variables Robustly Predict Survival in Community-Acquired Pneumonia. *Risk Management and Healthcare Policy*, *Volume 14*, 3701–3709. https://doi.org/10.2147/RMHP.S317735

Hassanpour, S., Tomita, N., DeLise, T., Crosier, B., & Marsch, L. A. (2019). Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacology*, *44*(3), 487–494. https://doi.org/10.1038/s41386-018-0247-x

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hosten, A. O. (1990). BUN and Creatinine. In H. K. Walker, W. D. Hall, & J. W. Hurst (Eds.), *Clinical Methods: The History, Physical, and Laboratory Examinations* (3rd ed.). Butterworths. http://www.ncbi.nlm.nih.gov/books/NBK305/

Huang, J., Habib, A.-R., Mendis, D., Chong, J., Smith, M., Duvnjak, M., Chiu, C., Singh, N., & Wong, E. (2020). An artificial intelligence algorithm that differentiates anterior ethmoidal artery location on sinus computed tomography scans. *The Journal of Laryngology & Otology*, *134*(1), 52–55. https://doi.org/10.1017/S0022215119002536

Huck, A., & Lewandrowski, K. (2014). Utilization management in the clinical laboratory: An introduction and overview of the literature. *Clinica Chimica Acta*, *427*, 111–117. https://doi.org/10.1016/j.cca.2013.09.021

Jha, A. K., Chan, D. C., Ridgway, A. B., Franz, C., & Bates, D. W. (2009). Improving Safety And Eliminating Redundant Tests: Cutting Costs In U.S. Hospitals. *Health Affairs*, *28*(5), 1475–1484. https://doi.org/10.1377/hlthaff.28.5.1475

Keegan, M. T., Gajic, O., & Afessa, B. (2012). Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. *Chest*, *142*(4), 851–858.

Konger, R. L., Ndekwe, P., Jones, G., Schmidt, R. P., Trey, M., Baty, E. J., Wilhite, D., Munshi, I. A., Sutter, B. M., Rao, M., & Bashir, C. M. (2016). Reduction in Unnecessary Clinical Laboratory Testing Through Utilization Management at a US Government Veterans Affairs Hospital. *American Journal of Clinical Pathology*, *145*(3), 355–364. https://doi.org/10.1093/ajcp/aqv092

Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., & Shen, H. T. (2021). Maximum Density Divergence for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(11), 3918–3930. https://doi.org/10.1109/TPAMI.2020.2991050

Lobo, B., Abdel-Rahman, E., Brown, D., Dunn, L., & Bowman, B. (2020). A recurrent neural network approach to predicting hemoglobin trajectories in patients with End-Stage Renal Disease. *Artificial Intelligence in Medicine*, *104*, 101823. https://doi.org/10.1016/j.artmed.2020.101823

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2016). *Deep Transfer Learning with Joint Adaptation Networks*. https://doi.org/10.48550/ARXIV.1605.06636

Luo, Y., Szolovits, P., Dighe, A. S., & Baron, J. M. (2016). Using Machine Learning to Predict Laboratory Test Results. *American Journal of Clinical Pathology*, *145*(6), 778–788. https://doi.org/10.1093/ajcp/aqw064

Moghadas, S. M., Shoukat, A., Fitzpatrick, M. C., Wells, C. R., Sah, P., Pandey, A., Sachs, J. D., Wang, Z., Meyers, L. A., Singer, B. H., & Galvani, A. P. (2020). Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proceedings of the National Academy of Sciences*, *117*(16), 9122–9126. https://doi.org/10.1073/pnas.2004064117

Nadda, W., Boonchieng, W., & Boonchieng, E. (2022). Influenza, dengue and common cold detection using LSTM with fully connected neural network and keywords selection. *BioData Mining*, *15*(1), 5. https://doi.org/10.1186/s13040-022-00288-9

Naemi, A., Schmidt, T., Mansourvar, M., & Wiil, U. K. (2020). Personalized Predictive Models for Identifying Clinical Deterioration Using LSTM in Emergency Departments. In A. Värri, J. Delgado, P. Gallos, M. Hägglund, K. Häyrinen, U.-M. Kinnunen, L. B. Pape-Haugaard, L.-M. Peltonen, K. Saranto, & P. Scott (Eds.), *Studies in Health Technology and Informatics*. IOS Press. https://doi.org/10.3233/SHTI200713

Olmsted, S. S., Moore, M., Meili, R. C., Duber, H. C., Wasserman, J., Sama, P., Mundell, B., & Hilborne, L. H. (2010). Strengthening Laboratory Systems in Resource-Limited Settings. *American Journal of Clinical Pathology*, *134*(3), 374–380. https://doi.org/10.1309/AJCPDQOSB7QR5GLR

Petti, C. A., Polage, C. R., Quinn, T. C., Ronald, A. R., & Sande, M. A. (2006). Laboratory Medicine in Africa: A Barrier to Effective Health Care. *Clinical Infectious Diseases*, *42*(3), 377–382. https://doi.org/10.1086/499363

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, *5*(1), Article 1. https://doi.org/10.1038/sdata.2018.178

Shi, B., Grimm, L. J., Mazurowski, M. A., Baker, J. A., Marks, J. R., King, L. M., Maley, C. C., Hwang, E. S., & Lo, J. Y. (2018). Prediction of Occult Invasive Disease in Ductal Carcinoma in Situ Using Deep Learning Features. *Journal of the American College of Radiology*, *15*(3), 527–534. https://doi.org/10.1016/j.jacr.2017.11.036

Smith, M., Saunders, R., Stuckhardt, L., & McGinnis, J. M. (2013). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America* (p. 13444). National Academies Press. https://doi.org/10.17226/13444

Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In G. Hua & H. Jégou (Eds.), *Computer Vision – ECCV 2016 Workshops* (Vol. 9915, pp. 443–450). Springer International Publishing. https://doi.org/10.1007/978-3-319-49409-8_35

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A Survey on Deep Transfer Learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2018* (Vol. 11141, pp. 270–279). Springer International Publishing. https://doi.org/10.1007/978-3-030-01424-7_27

Voglis, S., van Niftrik, C. H. B., Staartjes, V. E., Brandi, G., Tschopp, O., Regli, L., & Serra, C. (2020). Feasibility of machine learning based predictive modelling of postoperative hyponatremia after pituitary surgery. *Pituitary*, *23*(5), 543–551. https://doi.org/10.1007/s11102-020-01056-w

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9. https://doi.org/10.1186/s40537-016-0043-6

Xu, S., Hom, J., Balasubramanian, S., Schroeder, L. F., Najafi, N., Roy, S., & Chen, J. H. (2019). Prevalence and Predictability of Low-Yield Inpatient Laboratory Diagnostic Tests. *JAMA Network Open*, *2*(9), e1910967. https://doi.org/10.1001/jamanetworkopen.2019.10967

Yu, L., Zhang, Q., Bernstam, E. V., & Jiang, X. (2020). Predict or draw blood: An integrated method to reduce lab tests. *Journal of Biomedical Informatics*, *104*, 103394. https://doi.org/10.1016/j.jbi.2020.103394

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

CHAPTER 5. GENERAL CONCLUSION

This dissertation is devoted to making clinical outcome predictions using statistical and machine learning. Lots of endeavors have been made in the field of clinical outcome modeling to improve the quality of healthcare. In practice, the clinical outcome models help assess the severity of illness, evaluate the value of new treatments, provide expected outcomes, and promote clinical resource allocation. Meanwhile, the rich information in the EHR data provides great opportunities to build more accurate and reliable models for various clinical outcome tasks.

However, there are still many challenges when developing clinical outcome models on EHRs, including the hierarchical structure of high-dimensional medical concepts, the pattern extraction of vital signs, and the data insufficiency of some lab tests. To address these challenges, we present three research designs in this dissertation.

Specifically, in Chapter 2, we propose a new framework to generate low-dimensional representations with Manifold Learning for sets of hierarchical medical concepts in EHR data. This work solves the high-dimensional problem of complicated medical concepts. In Chapter 3, we propose a new ICU mortality prediction model capable of effectively extracting valid and interpretable patterns from the readily-available vital sign data with improved accuracy, by combining stochastic signal analysis and machine learning techniques. This work solves the second challenge by providing an effective way to extract meaningful features from vital signs. In Chapter 4, we propose a new transfer learning method for lab anomaly prediction with limited training data. This work addresses the data insufficiency problem and enables the outcome modeling for laboratory tests with limited training data.

By tackling the abovementioned issues in EHR data, our work has great potential to enlarge the social impact of clinical outcome models. For patients, our work helps to decrease

healthcare costs and adverse impacts. For healthcare providers, we provide a more accurate and reliable way to assess the severity of illness, the value of new treatments, and expected outcomes, which enables more scientific clinical decisions. For society, we enable the clinical outcome models to alleviate more healthcare expenditure burden and optimize better resource allocation.