The Pennsylvania State University
The Graduate School

SEPSIS DATA ANALYTICS

A Thesis in
Industrial Engineering and Operations Research
by
Sida Shen

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2020

The thesis of Sida Shen was reviewed and approved* by the following:

Soundar Kumara
Allen E. Pearce and Allen M. Pearce Professor of Industrial Engineering
Thesis Advisor

Kamesh Madduri
Associate Professor of Computer Science and Engineering
Operations Research Program Faculty

Robert Voigt
Professor of Industrial and Manufacturing Engineering
Professor and Graduate Program Coordinator

Abstract

Sepsis is a potentially life-threatening condition caused by the body's response to infection. Body releases chemicals into the blood stream to fight infection. However, sepsis occurs when the body's response to the chemicals go out of balance. Despite the use of antibiotics and modern treatments, sepsis is still one of the main causes of ICU mortality rate. The current broad definition of sepsis is not suitable for the heterogeneous nature of this disease; it is necessary to discover novel phenotypes of sepsis and design custom treatment plans. In this thesis, two novel phenotype discovery methods have been successfully developed and tested on MIMIC-III database. The first method utilizes first lab result for each patient, after feature imputation to resolve missing values, 11 features are included (heart rate, respiratory rate, systolic blood pressure noninvasive (sbp-noninvasive), temperature, sodium, white blood cells (WBC), creatinine, glucose and all 3 scores on the Glasgow Coma Scale (GCS)). With dimensionality reduction using Principal Component Analysis and clustering using K-means algorithm, three phenotypes are discovered; the first group patients (population: 44.9%, mortality: 14.88%) have high possibility of respiratory and renal failures; the second group patients (population: 23.8%, mortality: 9.15%) have high possibility of liver and coagulation failures; the third group patients (population: 31.4%, mortality: 20.9%) have high possibility of cardiovascular and central nervous system (CNS) failures. In the second model, we adapted [1] to cluster sepsis patient into novel phenotypes. We included 7 measurements (heart rate, respiratory rate, hemoglobin, white blood cell, creatinine, glucose and sodium) combined with 12 time-steps with 4-hour intervals (48 hours span). For each patient a sample with 84 features is constructed. A multi-layer fully connected auto-encoder is trained with 20 latent units; after 300 epochs, auto-encoder reconstruction loss (mean square error) converges. The encoder and soft assignment clustering layer are trained jointly using stochastic gradient descent; the auxiliary target distribution is updated every 200 steps. The network converges after 7800 steps with Kullback-Leibler divergence loss of 0.028. The derived 4 phenotypes present a clearly separated patient outcome with mortality standard deviation of 4.97%. However, by comparing bio-markers' statistics with patient outcome across the derived phenotypes, we can not see reasonable pattern that connect the two. There are some high dimensional features that the deep clustering model has captured, which we believe can lead to the discovery of the true cause of sepsis mortality.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would first like to thank my thesis advisor Dr. Kumara. He consistently allowed this thesis to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to acknowledge Dr. Madduri as the reader of this thesis, and I am gratefully indebted to his very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents and to my partner for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Chapter 1
Motivation and Problem Statement

1.1 Motivation

Every three to four seconds, there is some one in the world dying of sepsis; sepsis is the final common pathway to death from the vast majority of infectious diseases world-wide. Sepsis arises when, in combating infection, immune system damages a person's own tissues and organs, which can lead to organ failure and ultimately death. Current management of sepsis greatly relies on the usage of antibiotics; when a patient is admitted, aerobic and anaerobic blood cultures are taken to initialize the usage of antibiotics. Despite the use of antibiotics and modern treatments, sepsis is still one of the main contributors to ICU mortality rate. According to the current definition, sepsis is categorized into different levels, including sepsis, severe sepsis and septic shock [4]. However, there is little progress made in terms of customized treatments for patients. Although severeness (patient outcome) of the disease is identified, the current categorization protocol contributes little to customized treatments. Currently, the state-of-the-art sepsis definition is the Sepsis-3 criteria [5], classifying sepsis into sepsis and severe sepsis and present no customized treatment for individual patients; this broad definition is not suitable for the heterogeneous nature of this disease. It is necessary to discover novel phenotypes of sepsis and design customized treatment plans.

Modern medicine relies heavily on biology and clinical research; with the introduction of machine learning, it has shown great promise of using artificial intelligence to assist doctor's decision or research. There are research efforts [6] that use computer vision machine learning algorithms to assist identification of CT scans, the use of reinforcement learning [7] to assist doctor decision processes like dosing control, the use of electronic health record for fast disease identification and diagnosis, etc.

Electronic health record (EHR) is an electronic version of a patient's medical history;

it often includes all of the vital clinical data relevant to that patient under a particular medical provider. EHRs are the next stage in the progress of health that can strengthen the bond between patients and clinicians. The data and accessibility of it, will empower medical providers to settle on better choices and give better cares.

1.2 Problem Statement and Discussion

1.2.1 Problem Statement

In this study, the objective is to utilize state-of-the art data-driven machine learning methods to discover novel phenotypes of sepsis based on Electronic Health Record of Intensive Care Unit patients. Our objective is to identify the phenotypes of sepsis based on EHR from ICU patients. We uses Medical Information Mart for Intensive Care (MIMIC-III) [8], a freely accessible EHR for ICU patients, as the main source of the data. Extensive data pre-processing techniques based on medical domain knowledge and state-of-the-art unsupervised learning methods are used to derive the phenotype. Two models are introduced in this study;

- Principal Component Analysis (PCA) with K-means Clustering: K-means clustering is a simple yet powerful unsupervised machine learning algorithm which is known to fail with a large feature space (curse of dimensionality); PCA performs an orthogonal transformation on the data to achieve feature reduction. In the this model, we propose to use PCA as a feature reduction technique to enhance K-means' performance. For the model's data-driven nature, source of data is vital. In this case, the period of time that is included for each patient is an important variable to consider; the period of time is the time span with the patient's ICU admit timestamp as time-zero. In order to consider such an impact, two experiments are run with this model, the first one including data from all periods and the second one including only first-time lab result.

- Times series clustering with Deep Embedding Clustering [1]: In this model, we propose to use multivariate time series to represent patients. We think that the introduction of temporal features can capture features and feature interactions that are hidden deeper; the model can take advantage of information that is only discoverable in time domain such as the progression of patient response to medication.

1.2.2 Uniqueness of Research

- We present an end-to-end model including data pre-processing pipeline and clustering algorithm.
- We present the first data-driven novel phenotypes discovery study that includes time domain information in the data.

Our work is different from the previous effort [9].

- Previous work used the most abnormal value recorded within the first 6 hours of hospital representation; with using only the first value upon ICU admission (our first model), we achieved similar degree of separation in terms of patient's organ dysfunction and outcome.
- We have presented a comparison between the period of time that is included for each patient (our first model), which is not considered by previous work.
- We have presented a model that includes time domain information in the data (our second model), multiple values in time domain are stored for each patient and feature combination, which is vastly difference from previous work.

1.2.3 Organization of the Thesis

In chapter 2, we present the background domain knowledge that is associated with this study; background in medical domain knowledge and general process of data-driven pattern discovery procedure will be introduced in detail. In chapter 3, we present related literature and studies both in medical and machine learning domain. In chapter 4, we present the patient diagnosis process; in specific, the logic behind why International Statistical Classification of Diseases version 9 (ICD-9) is not used although it is provided by MIMIC-III. In chapter 5 we present the traditional machine learning model (k-means), 2 experiments with different data extraction methods are presented and discussed. In chapter 6 we present a deep clustering model that learns time domain information. Chapter 7 concludes this thesis by highlights in results from chapters 5 and 6.

Chapter 2
Background

This chapter will address the background literature and methodological foundations of the proposed research work.

2.1 Medical Background

2.1.1 Sepsis

Sepsis is a potentially life-threatening condition caused by the body's response to infection [4]. Body releases chemicals into the blood stream to fight infection. However, sepsis occurs when the body's response to the chemicals goes out of balance to cause damage in the host's own organ and tissues. Multiple end-organ dysfunction can be triggered, including lungs, brain, liver, etc. Most common infections can easily lead to sepsis. Among these common infections such as pneumonia, urinary infections, infections in the abdomen, skin or wound infections, or meningitis. Seasonal flu, malaria, dengue, yellow fever and Ebola may all also result in sepsis. More than 80 percent of infections leading to sepsis are contracted outside of a hospital. People with a weakened immune system are especially at risk. This includes adults over 60, infants under one year-old, people with chronic diseases of the lung, liver, or heart, people with diabetes or AIDS, and people without a spleen. With the recent global pandemic of the novel coronavirus (COVID-19), study [10] has found an increased odds of COVID-19 in-hospital death when associated with sepsis related diseases; it is also discovered that sepsis is the most frequently observed complication with COVID-19 patients. All the above reasons makes sepsis both an epidemic and very deadly condition. Sepsis had been categorised into different levels [5]. Patients first show symptoms of sepsis including:

- Change in mental state

- Low reading of Systolic Blood Pressure
- High Reading of Respiratory rate

There is a possibility that a sepsis patient progresses into septic shock [5]; normally it is caused by some abnormal change in the patient's circulatory system. As the symptom of septic shock is more severe and aggressive, it is more likely to cause death than sepsis. Septic shock manifests in:

- Extremely low Systolic Blood Pressure (typically lower than 65), present the need of specific medication (ex. Norepinephrine) to maintain at normal level.
- Body is unable to use oxygen properly or effectively; resulting in high lactic acid level.

2.1.2 Organ Failure/Dysfunction

Organ dysfunction is when the organ is not in a normal state, either malfunctioning or not performing at its expected state. Organ failure, although sounds similar, describes a state that normal homeostasis (steady state of a living system in terms of it's internal, physical and chemical conditions) is not able to be maintained without assistance from external clinical interventions. Organ dysfunction and organ failure are not diagnosis since the causes of such condition are often unknown; however, based on the symptoms and outcomes of the failure/dysfunction, it can be acute or chronic organ failure/dysfunction. There are different types of organ dysfunction/failure:

- Respiratory failure: When the body has insufficient amount of Oxygen or excessive amounts of Carbon dioxide. Acute respiratory distress syndrome (ARDS) [11] is a common end organ failure associated with sepsis. It is characterised by a rapid onset of widespread inflammation in the lungs. Studies has shown that severe sepsis is the most common cause for ARDS. [12].
- Heart failure: a condition when heart is unable to pump enough blood into the blood flow to maintain normal function of a living body. [13].
- Renal failure: also known as kidney failure or end-stage kidney disease is a condition that kidney's function is at less than 85 percent of it's normal level. There are two ways that kidney can be harmed by sepsis; first, the infection that initiated sepsis starts in the kidney; second, a sequential organ failure caused by sepsis creates a cascade event that can lead to kidney damage.

- Hematologic failure: A condition when blood and blood forming organs are harmed. It can be an indicator of severe sepsis; hematologic organ system dysfunction is an early manifestation of severe sepsis and is seen in virtually all patients with this disease. [14]
- Neurologic failure: A condition when the nervous system presents certain disorder [15].
- Hepatic failure: a condition where liver function is lost or damaged. Sepsis can lead to acute hepatic failure, which can cause serious complications including excessive bleeding and increasing pressure in the brain.

Multiple organ failure often is associated with sepsis and has a high mortality rate. Although often fatal, clinical practitioners have tools at hand for evaluating and making prognosis of multiple organ failure's outcome. One of the most common evaluations is through the sequential organ failure assessment.

2.1.2.1 Sequential Organ Failure Assessment

Sequential Organ Failure Assessment (SOFA) score [2] can be used to determine the level of organ dysfunction and to predict the likelihood of mortality for Intensive Care Unit patients. It measures the severity of each organ system's condition using some combinations of clinical measurements that represent the organ system's functionality. After the condition of each organ system is calculated, the severity of organ dysfunction of all organ systems are summed as the SOFA score. Detailed organ systems with associated clinical measurements are presented in Table 2.1.

Table 2.1. Sequential Organ Failure Measurements [2]

Organ System	SOFA
Respiratory	Pao2/FIO2 ratio Mechanical ventilation
Cardiovascular	Mean arterial pressure Use of vasoactive agents
Renal	Creatinine Urine output
Hematologic	Platelets
Neurologic	Glasgow Coma Scale Score
Hepatic	Bilirubin

Literature reports studies to predict patient outcome using SOFA score. Table 2.2 presents predicted mortality rate for each level of SOFA score.

Table 2.2. Sequential Organ Failure Mortality Rate [3]

SOFA	Mortality (Initial Score) (%)	Mortality (Highest Score) (%)
0-1	0.0	0.0
2-3	6.4	1.5
4-5	20.2	6.7
6-7	21.5	18.2
8-9	33.3	26.3
10-11	50	45.8
12-14	95.2	80.0
>14	95.2	89.7

2.2 Data Mining

Data mining is the process of drawing and discovering patterns in data with methods including machine learning, statistics and database systems. In the early years of data mining, manual extraction of features and patterns from data was the main focus of research and application. Data mining has become more of an interdisciplinary field with the development of both computer hardware and algorithm, combining different fields of science.

2.2.1 Machine Learning

During the introduction of computers, simple tasks like logic operation, numerical computation can be done at incredible speed with consistent accuracy. From that time, scientists have been finding ways to teach computers to perform tasks on their own, exhibiting human levels of intelligence; this process is commonly known as artificial intelligence or machine learning. Intelligence is coded in the form of rule-based systems; Rules usually consist of a set of if/else statements. Identification and utilization of those rules can collectively represent knowledge. In the early years, rule-based system relied heavily on domain knowledge as the rules are human-crafted. As the availability of data increased exponentially over time, handcrafting rules has become computationally infeasible and data driven machine learning algorithm became popular to draw statistical inferences. Deep learning was commonly known to be discovered in 1985, when Hinton et al [16] demonstrated the use of backpropagation to optimize neural networks. Deep neural

networks are known as a universal non-linear approximators. With the introduction of deep learning (deep neural network), more exciting research and applications have been initiated. fold like computer vision, natural language processing, recommender system, etc. have experienced revolutionary changes. Learning is three fold: supervised, unsupervised and semi-supervised. We discuss the first two here:

2.2.1.1 Supervised Learning

Supervised learning is the process of finding statistical inference from given data which consists of input and output (sample and label). It is typically done by some learning algorithm that approximates the mapping from the input space to the output space. The ultimate goal in supervised learning task is to establish a mapping that can generalize the pattern for both seen and unseen samples. One difficulty of supervised learning task is the balance of bias-variance trade off. When the model suffers from high bias, the model is not able to establish a complex enough mapping to explain the actual pattern going from the input space into the output space; one example would be using a linear approximator such as linear regression to fit a data that presents feature interaction or other kind of non-linear properties. The model suffers from high variance when it draws too much unnecessary information from the data which causes high model sensitivity. High model variance is often caused by model being too complex (i.e. excessive parameters) or being trained for excessive amount of time. It is easy to see that the model is at it's best performance when it's variance and bias are minimized; however, in practice, there is a trade off between them. It is the analyst's job to find the optimal model and its hyper-parameters to best suit the task at hand.

2.2.1.2 Unsupervised Learning

Unsupervised learning is a collection of machine learning algorithms that inference only based on the data without the need for ground truth (label). Clustering analysis, being an unsupervised classification technique, is suitable for exploratory analysis. Clustering Analysis is a task of partitioning a set of samples into groups in a way that samples in each group are like each other based on some criteria and distant from other groups. Deep learning has been shown its capabilities in unsupervised learning tasks. Deep Representation Learning, also called Deep embedding learning is the procedure of finding an abstract representation, often done with dimension reduction techniques with deep neural network as a universal nonlinear approximator. One difficulty for exploratory clustering analysis is the difficulty of evaluating the model. Since a label is not given,

performance metrics like accuracy, precision, recall, etc. (classification) and mean square error, etc. (regression) cannot be used. However, some unsupervised learning specific performance metrics can be used. For example, Internal validity is a performance metric that increases when similar samples are grouped together which decreases when different samples are grouped together; the similarity metric can be selected based on the specific task; Euclidean distance, Manhattan distance, Minkowski distance are all used extensively.

2.2.2 Knowledge discovery in databases (KDD)

Knowledge discovery in databases (KDD) is a high level protocol of finding knowledge in data [17]. The goal of KDD process is to infer patterns from large databases. The overall process can be boiled down to data selection, pre-processing, transformation, data mining and interpretation/evaluation.

2.2.2.1 Domain Understanding

The first step of the KDD process is to build a general understanding of the domain knowledge. As an initial step, it is necessary to develop the ultimate goal with a road-map with decisions like algorithm, transformation and final representation. With prior knowledge, one has to have a deep understanding of the end-user, the environment in which the algorithm would be deployed and the problem that is to be solved.

2.2.2.2 Data Selection and Cleansing

With the objective well defined, the next step is determining the data to be used for the knowledge discovery process in the downstream analysis. It is important for researchers and practitioners to evaluate what data might be useful. This process is vital since as the evidence base for the system, data is the only source that data mining learns and discovers the knowledge. After selecting potential data samples and features based on domain knowledge, it is important to evaluate the quality of the data; if one feature is containing too many missing or wrong values, it creates a high risk that the learnt pattern is tampered with contaminated data.

Though there is no perfect way to deal with missing values, we explore few methods. First is to address the reason behind the missing data. To begin we assume that the samples are missing at random (MAR), which means that the probability of data missing is only consistent within some group defined by the observed data; for example, a student

having a higher grade point average is more likely to put it on his resume than a student having a lower one; however, the probability of grade point average missing is the same within group A (high grade point average) and group B (low grade point average). Second, the samples are missing completely at random (MCAR), which means that there is no relationship between the missing sample and any part of the observed data; there is nothing in the data that can suggest one sample is more likely to be missing than the other. Typically, as a natural assumption, MCAR is the most frequent case. Lastly, the sample does not fall into neither MCAR or MAR, it is missing not at random (MNAR) or non-ignorable non-response; it happens when the reason of the value missing is caused by its value itself. On the other hand, MCAR and MAR are considered ignorable since the missing value does not contain any unknown information. For example, a MNAR case would be missing body weight is caused by body weight exceeding the upper limit of scale.

There are possible ways to deal with missing values. The easiest way would be imputation; If the missing value is MCAR, it is possible to use List-wise imputation (remove the sample if it contains one or more missing value) without introducing any bias. Sometimes, it is better to keep samples instead of dropping them; in some cases, an insignificant variable can be dropped if majority of the samples are missing it.

There are time-series specific methods; Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB) are common statistical approaches to missing data in a time series. LOCF assumes that the missing value is the same as the previous value in time and NOCB assumes that the missing value is the same as the next value in time. If the time series presents trend and seasonality, one can use linear interpolation with seasonal adjustment.

2.2.2.3 Transformation and Modeling

Data transformation is the process of transforming data into a form that is suitable for the downstream data mining techniques; a common way is dimensionality reduction. Principle Component Analysis (PCA) [18] is the most used statistical method to perform an orthogonal transformation on the data, resulting in a set of linearly uncorrelated variables. The transformation of PCA is linear which means that PCA is able to capture linear relationship between the features. If there are nonlinear relationships between the features, PCA tend to fail often; in this case, kernel-PCA can use a high dimensional function to convert the linear transformation of PCA into the reproducing kernel space, providing PCA the ability to perform nonlinear mapping.

Data mining is a collection of techniques that extracts patterns from data. It can be done using collection of supervised learning methods if label is available or collection of unsupervised learning methods if label is unavailable.

2.2.2.4 Interpretation and Evaluation

Pattern Evaluation is the process of identify patterns that represent knowledge based on some measures. If the data mining model is a prediction model, one can develop a performance metric to evaluate the performance of the model. For regression problems, one can use mean square error (MSE), mean absolute error (MAE), Huber loss, etc. For classification problem, one can use accuracy as performance metric; if the training labels are not balanced, one can look at the confusion matrix and F-score to evaluate the performance without bias from unbalanced label. The performance metric can be designed to fit the need of the data mining task (ex. high recall). The model can then be fine tuned with an objective of maximizing the performance metric.

Chapter 3
Literature Review

In this chapter we review the literature specific to the sepsis problem and electronic health record (EHR).

3.1 Electronic Health Record

Electronic Health Record (EHR) is a collection of systems that digitally stores health related information both for patients and general population [19]. The digital nature of this collection of systems not only simplifies and systematizes one's health history, but also allows convenient sharing of data among medical practitioners and researchers. The existence of EHR makes care transition between care givers faster and smoother [20] , which can greatly improves patient outcome. [21]. The availability of digital information provides what data-driven research and projects needs; it initiates and makes possible a number of exciting research opportunities; with the complete historical data that is available, research like patient outcome prediction [22] [23], dosing control prediction and control [24], pathophysiology discovery and research [25] and much more are made possible.

MIMIC-III (Medical Information Mart for Intensive Care III) [8] is a large scale, open access intensive care unit database containing over 60,000 patients from Beth Israel Deaconess Medical Center's Intensive Care Unit between 2001 and 2012. The database contains almost every detail possible for every patient, which includes laboratory test results, procedures, medications, caregiver notes, imaging reports and mortality. MIMIC-III has been used for numerous research efforts, some of them focusing on predicting a certain type of disease; [26] proposed an interpretable machine learning model to predict sepsis among ICU patients. In this research, patients in MIMIC-III are used as a cohort to validate the result. Some research efforts focus on general techniques

for data-driven models; [27] proposed a recurrent neural network based model to fill in missing data in multivariate time series. Some other efforts focused on assisting doctor's decision; [28] developed a reinforcement learning based model that predicts time-to-extubation readiness and recommend customized regime of sedation dosage and ventilator support for each patient.

3.2 Sepsis

During the last decade, there are numerous studies that provided valuable insight into sepsis from both analytical and health care prospective. In this section, some highlights of contribution to sepsis research will be addressed.

3.2.1 Sepsis Definition and Diagnosis

Sepsis is defined as life threatening organ dysfunction caused by a dysregulated host response to infection [5]. The word sepsis was invented by the Hippocrates in the 4th century BC, having a meaning similar to decomposition and decay. However, it was not until the 19th century people started using the specific word sepsis to describe the condition as we know by. Abu Ali Sina (980 - 1037 A.D.), a Persian who was know as a physician, an astronomer, a singer and a writer, used the term 'blood rot' for severe infection and it was kept that way for hundreds of years. Moving into the modern period, septicemia and sepsis were included in the International Statistical Classification of Diseases and Related Health Problems (ICD) version 9 [29]; sepsis is known as the blood poisoning disease; septicemia is the condition that bacteria is present in the blood stream, which would often lead to sepsis. ICD version 9 is used in the United States of America prior to 2012 until ICD version 10 [30] was introduced. ICD version 10 removed septicemia and suggested to use sepsis to indicate the underlying infection.

Early treatment is the key to curing and managing outcomes of patients; hence, diagnosis of sepsis in both a timely and accurate manner is important. Sepsis had been defined into different levels (sepsis, severe sepsis and sepsis shock) using System inflammatory response syndrome (SIRS) criteria [31]; SIRS is defined as fulfilled if at least two of the four criteria are met: fever > 38.0°C or hypothermia < 36.0°C, tachycardia $>$ 90 beats/minute, tachypnea > 20 breaths/minute, leukocytosis $> 12 * 10^9$ or leucopoenia $< 4 * 10^9$. Although the patient is very unlikely to have sepsis if his SIRS is negative, a patient with positive result only have a moderate possibility of having sepsis. In order

to improve upon this, a consensus is achieved in 2016 to use quick Sequential Organ Failure Assessment (qSOFA) score to replace SIRS [5]. qSOFA is a simplified version of Sequential Organ Failure Assessment (SOFA) as it only includes 3 clinical criteria namely mental state, respiration rate and systolic blood pressure, which makes it relatively easy and quick to compute.

3.2.2 Sepsis phenotype study

Sepsis phenotype attracted attention in 2019. Previous work [9] used feature space Euclidean Distance to partition sepsis patients into 4 novel phenotypes. In this study, 16552 unique patients who met the Sepsis-3 criteria were used as samples to derive novel sepsis phenotype; from the clustered cohort's lab variable statistics and patient outcome, this work successfully derived 4 novel phenotypes; the simulation also suggests that the phenotypes may help understanding the heterogeneity of the treatment effect, which would be very useful in future downstream analysis.

3.3 Machine Learning

3.3.1 Representation Learning

In Pearson's [18] own words, "in many physical, statistical and biological investigations, it is desirable to represent a system of points in plane, three or higher dimension space by the best-fitting straight line or plane." This illustrates the earliest idea of dimensionality reduction, also called representation learning or feature embedding in future studies.

Representation learning is the process of finding an abstract representation of the input data. This process often replaces the manual feature engineering process which involves high labor intensive work. In the current state-of-the-art, representation learning are mostly done with Deep Neural Network, however, there are traditional approaches that present comparable performance at certain task while offering faster speed.

3.3.1.1 Traditional Approach

Principal Component analysis (PCA) [18] generates a set of vectors that present the highest variance often at much lower dimension of the original feature space. Though PCA is fast in feature embedding, it fails when data is linear inseparable. Kernel Principal Component Analysis (KPCA) [32] is the nonlinear version of the linear PCA, which is

capable of extracting non-linear relationships between high-dimension features. It is done by projecting the original data into a higher dimension with a kernel function. In theory, part of the features that are linear inseparable would be linear separable in the projected higher dimensional space.

3.3.1.2 State-of-the-art Representation Learning Applications

With advancements in computer vision, natural language processing and signal processing, representation learning has developed rapidly over the past decade. In this section, some important applications of presentation learning in different machine learning fields would be addressed. As the field is vast we only give a brief introduction.

In natural language processing, text are often presented as an one-hot encoded vector, if the vocabulary grows, the dimension of the vector grows with it. There are two common approaches,

- Continuous bag-of-word: predict the word by the context.
- Skip-gram: Given the word, predict the context.

Word-to-vector (word2vec) [33] introduces a feature embedding methods for words and phrases. They introduce a methodology of representing some sparse matrix in a dense quantitative form; in other words, substituting the words into fixed length numeric vectors.

Graph is a data structure that is heavily used in social networks, computer vision, etc. Graph embedding transfers the component of a graph (nodes, edges and features) into a lower dimensional vector space, where the information and structure of the graph are preserved. DeepWalk [34] uses random walk to learn the local information of a graph as latent representation. Graph Convolutional Network (GCN) [35] is a kind of neural network that operates directly on graphs; with a convolutional architecture achieved by spectral graph convolutions, local structure and information of graph can be obtained. GCN surpassed previous state-of-the-art models by a great margin on numbers of classification tasks. Graph Attention Network (GAT) [36] is an extremely efficient convolution-style neural network structure that uses attention layer to implicitly assign different importance to different nodes within the same neighborhood without operating on the whole graph; although very efficient, the performance of GAT in terms of classification task is also comparable to other state-of-the-art models.

3.3.2 Unsupervised Learning

There are different types of traditional clustering algorithms, each having its own advantages and drawbacks; Centroid-based clustering often requires known number of clusters and prefers them having similar size; Connectivity-based Clustering does not require number of clusters, but the computational burden is quite large. Though the traditional clustering algorithms are popular, they present a number of limitations including curse of dimensionality, difficulty to introduce non-linearity, etc. With recent developments in deep learning, there has been some promise using deep learning to perform clustering analysis. There are models that train on the loss of the cluster [37] [38] [39]. These algorithms use loss of clusters to directly obtain assignments of the samples to clusters. Auto-encoder is a kind of neural network structure that reconstructs its input; it is often done by projecting the input data to the latent space (often lower dimension space, for representation learning), and then map it back to its original dimension. If the reconstruction loss (often use mean square error) converges, it is safe to say that the latent space contains accurate abstract information of the input data. Auto-encoder based clustering is the major part of deep clustering network families and has been studied the most. [37] proposes a method to jointly train an auto-encoder and a k-means cluster; [1] first pre-trains an auto-encoder and the latent space vectors are fed into a clustering layer; the clustering layer refines using the KL-divergence between the distribution of the soft label and the selected target distribution.

Chapter 4
Patient Selection and Diagnosis

In this section, source of data and patient extraction logic would be explained in detail. Source of data, MIMIC-III will be introduced first, then patient selection logic will be presented with explanation of the criteria. Finally, the logic behind using Sepsis-3 criteria for diagnosis instead of ICD-9 codes will be explained.

4.1 Electronic Health Record

MIMIC-III (Medical Information Mart for Intensive Care III) [8] is a large scale, free access Intensive Care Unit database containing over 40,000 patients from Beth Israel Deaconess Medical Center's intensive care unit between 2001 and 2012. The database contains almost every detail possible for every patient, which includes laboratory test results, procedures, medications, caregiver notes, imaging reports and mortality. There are two data sources, Carevue and Metavision. Carevue records patient's information from 2001 to 2008; Metavision records patient's information from 2008 to 2012. In this study, we use patient data from Metavision.

4.2 Data Preparation

4.2.1 Patient Selection

There are over 40,000 ICU patients in the MIMIC-III database; in this study, ICU stays are used to represent patients. The total ICU stays in the database is 61,532; Following previous studies [40], some ICU stays that are outliers are eliminated to minimize model variance. The detailed number of cases eliminated is displayed in Table 4.1

- It is possible that one patient is admitted into the ICU multiple times; in this case, we only keep their first ICU stay.
- Non-adult patients often present different measurements such as higher heart rate. In this study, we only include patients whose age >= 18
- We only include patients that are admitted to the non-cardiac surgical ICU.
- Since sepsis is caused by infection, we screen all patients who are not suspected of infection.
- There are patients that has no data stored in the EHR (possible result from system error), we do not consider those patients.

Table 4.1. ICU stays Exclusion results

Number Of ICU Stays	Exclusion
3	Non-adult
7536	SecondStay
2298	CardiacSurgery
1974	NonInfection
18	MissingData

4.2.2 Sepsis Diagnosis

In MIMIC-III database, ICD-9 codes are used to record diagnosis of patients; ICD codes are alphanumeric codes that are used by doctors, public health agents and various health related practitioners to represent a systematic diagnosis. Every disease and complication that is known by the medical world has its own ICD code with detailed descriptions. ICD codes version 9 was used in the United States starting around 1975 and discontinued on October 1st, 2015. Transitioning from ICD version 9 to ICD version 10, numerous changes have been made;

- There are 18 times as many procedure codes in ICD version 10 (71924 codes) than ICD version 9 (3824 codes).
- There are 5 times as many diagnosis codes in ICD version 10 (69823 codes) than ICD version 9 (14025 codes).
- Data quality improved vastly from ICD version 9 to version 10.

- Changes in diagnosis and recommended procedure.

As the medical community learns more about the disease, the perspective changes, so does the diagnosis process, sepsis as a disease also underwent major changes transitioning from ICD version 9 to ICD version 10.

- Septicemia, an infection in blood that likely causes sepsis is eliminated.
- Emphasis is laid on the linkage between sepsis and acute organ dysfunction.

There are several concerns we have regarding ICD code version 9 with this research.

- ICD code version 9 is relatively outdated with the current state-of-the-art and understanding of sepsis
- ICD codes are all coded by doctors, there exists possible human error or different opinions leading to different conclusions based on symptom/lab results.
- ICD codes are being refreshed and updated from time to time, diagnosis at different times may not be the same.

With the above considerations, we suspect the use of ICD code version 9 would cause both inaccurate and inconsistent diagnosis of patients. With the availability of the patients' full medical history in the ICU (EHR), it is decided to diagnose the patient purely using data from the EHR. We decide to use the latest Sepsis-3 criteria [5] to diagnosis and extract patients using EHR (MIMIC-III). We have referenced this study [40] for the patient diagnosis process.

The patient diagnosis process strictly follows the current sepsis 3 criteria.

- Fever $> 38.0°C$ or hypothermia $< 36.0°C$
- Tachycardia > 90 beats/minute
- Tachypnea > 20 breaths/minute
- Leukocytosis $> 12 * 10^9$ or Leucopoenia $< 4 * 10^9$.

Chapter 5
Traditional Approach

In this chapter, we aim to build a naive K-means clustering model using the ICU patients filtered from chapter 4. For the model's data-driven nature, source of data is vital. In this case, the period of time that is included for each patient is an important variable to consider; the period of time is the time span with the patient's ICU admit timestamp as time-zero. In order to consider such impact, two experiments were run with this model, the first one including data from all periods and the second one including only first-time lab result.

5.1 Data Pre-Processing

K means is known to fail if the data pre-processing is not done correctly. In this section, data pre-processing including feature engineering and outlier detection and elimination will be explained in detail. This flow chart summarizes the pre-processing pipeline that is shown in Figure 5.1

5.1.1 Feature Engineering

5.1.1.1 Domain Knowledge Feature Selection

Based on domain knowledge and the availability of the MIMIC-III database, we have selected some features that are expected to be important to sepsis diagnosis: heart rate, respiratory rate, systolic blood pressure, systolic blood pressure non-invasive, temperature, GCS, GCS MV, Oxygen Saturation, Bands, White Blood Cells (WBC), total Bilirubin, unrine Creatinine, Creatinine, INR, Troponin, Glucose, Sodium, Hemoglobin, Chloride, Bicarbonate, Lactate, Albumin, C Reactive protein, ALT and ALS.

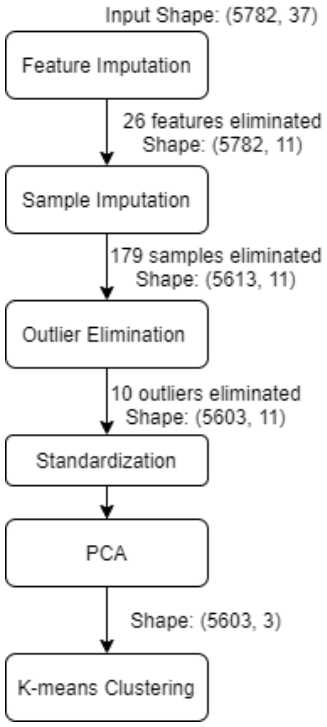

Figure 5.1. K means model data pipeline

5.1.1.2 Categorical Features

Most of the features selected are numerical; for categorical features, we can either use one hot encoding or domain knowledge (only if there is a numerical difference between different levels). The first categorical feature is the Glasgow Coma Scale [41], the most common scoring system that is used to describe a person's level of consciousness. The scoring system, shown in Table 5.1 is divided into three parts: response, verbal response and motor response. Since the scoring system has been validated by the medical community for over 40 years, we would use the score as numeric value.

The second categorical value that is been treated is Oxygen Saturation with three levels: Needs O_2 inhalation to maintain O_2 sat $> 90\%$, Able to maintain O_2 sat $>92\%$ on RA, O_2 sat $< 90\%$ even with supplemental O_2. It is clear that there are numerical differences between the three levels; we decided to assign values to them in terms of severity, where details are in Table 5.2.

5.1.2 Missing Value Treatment

There are a lot of missing values present in the extracted data; it is important for us to either fill in the missing value or impute features/instances for minimum bias introduced.

Table 5.1. GCS score

Response	Score
Eye Opening	
Spontaneous	4
Response to verbal command	3
Response to pain	2
No eye opening	1
Best verbal response	
Oriented	5
Confused	4
Inappropriate words	3
Incomprehensible sounds	2
No verbal response	1
Best motor response	
Obey commands	6
Localizing response to pain	5
Withdrawal response to pain	4
Flexion to pain	3
Extension to pain	2
No motor response	1

Table 5.2. Oxygen Saturation Scoring

Levels	Score
O2 sat $< 90\%$ even with supplemental O2	1
Needs O2 inhalation to maintain O2 sat $> 90\%$	2
Able to maintain O2 sat $>92\%$ on RA	3

First it is necessary to evaluate why the missing values are present in MIMIC-III.

- Some patients may not afford some of the premium services financially (i.e. not having an insurance plan).
- Variance in the tests performed by doctors.
- Human errors.

It is concluded that the cause of missing value is very complicated; the cause of data goes missing does not appear to be related to the observed data; we conclude that the data is missing at random (MAR). Hence, we can impute and fill the null values freely without introducing bias.

There are some features that are only available for a few patients; for those features, we decided to discard them all together. A threshold = 90% has been set to eliminate all

features that contain >10% null value; similar process is used for samples (ICU stays), with threshold = 95%. Outliers are also treated; all samples that present Z score higher than 3 are dropped. The prepared data has 5603 ICU stays and 11 features. Statistics on the resulting training data are shown in Table 5.3.

Table 5.3. Data after feature engineering

Feature	Mean	Standard Deviation	25% Q	75% Q
HeartRate	86.33	14.09	76.09	96.07
RespiratoryRate	22.75	3.734	16.96	19.27
SBPNoninvasive	120.6	16.04	107.6	129.8
Temperature	98.17	1.320	97.65	98.97
Sodium	143.5	4.460	136.7	141.6
WBC	11.89	9.24	7.922	14.02
Creatinine	1.440	1.414	0.7333	1.529
Glucose	135.22	43.36	109.00	150.64
GCS - Eye Opening	3.409	0.707	3.160	3.958
GCS - Motor Response	5.388	1.020	5.179	6.000
GCS - Verbal Response	3.571	1.433	2.281	5.000

5.1.3 Principal Component Analysis

K means algorithm uses Euclidean distance to measure the difference between samples; it is known to fail when the feature space is too large. It is a common practice to perform feature embedding/dimensionality reduction techniques on the data before using the K means clustering algorithm.

In this research, we chose Principal Component Analysis (PCA) as the dimensionality reduction technique. PCA is a commonly used statistical method to perform an orthogonal transformation on the data, resulting in a set of linearly uncorrelated variables. PCA is displayed in Algorithm 1

Algorithm 1 PCA Algorithm

1: Compute dot product matrix: $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu})$
2: Do $\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$
3: Compute Eigen-vectors $\mathbf{U} = \mathbf{X}\mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}$
4: Keep desired number of components
5: Compute n features:$\mathbf{Y} = \mathbf{U}_\mathbf{n}^T\mathbf{X}$

For PCA, one important hyper-parameter to choose is the number of components. It is a trade off between curse of dimensionality and information loss; higher number the

components would cause low information loss and high output dimension; low number of components would cause high information loss and low output dimension. In this study, since we are feeding the data to a K means clustering algorithm, we chose the number of components = 3 to keep the feature space as small as possible.

5.2 K-means Clustering

5.2.1 Background

K-means is a very well known algorithm for unsupervised learning; the algorithm aims to partition n samples in to k (k < n) clusters that each sample belongs to the cluster with the closest mean. The algorithm starts with k randomized centroids; then each sample is assigned to the closest centroid in terms of Euclidean distance; the centroid is then recalculated based on the samples that are assigned to that centroid. The algorithm is a NP-complete, but some local optimum can be met. Usually the algorithm runs until some criteria is met (i.e. cluster loss is less than a certain threshold). Since the starting point of the clusters is random, although converges, k-means algorithm may not give the exact cluster assignment on each run. The cluster loss is defined as sum of square error between samples and cluster centroid. The formulation is shown with S_k as samples in kth cluster, $\bar{x}_{k,j}$ as the jth variable of the cluster center for the kth cluster.

$$SSE = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \tag{5.1}$$

The only parameter for K-means clustering algorithm is the the number of clusters k. Higher number of clusters would result in low clustering loss. One way to select optimal number of clusters is to perform a k-means for $k = 1, ..., n$, plot and observe cluster loss vs k, the best k appears at the point where keep increasing k would not lower the clustering loss as much; this method of finding the best k is also called the elbow method, which is displayed in algorithm 2. The K-means algorithm is displayed in Algorithm 3.

5.3 Validation

A good clustering algorithm groups similar samples into the same group and dissimilar samples into different groups. In order to define a way to validate the performance of the clustering model, one first needs to derive a a metric for a patient which enables

Algorithm 2 Elbow Method

1: Initialize an ArrayList a
2: **for** k=0; k<n; i++ **do**
3: Perform K-means cluster with k clusters
4: Compute cluster loss L = $\sum_{i=1}^{n} (x_i - \bar{x})^2$
5: Append L into a
6: **end for**
7: Plot L vs k, Return chosen k

Algorithm 3 K-means Algorithm

Randomly Initialize Cluster Centroid $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$
while No Convergence **do**
for every sample i **do**
set sample assignment $c^{(i)} := \arg\min \left\| x^{(i)} - \mu \right\|^2$
end for
for every centroid k **do**
set centroid $\mu := \frac{\sum_{i=1}^{m} 1\{c^{(i)}=\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=\}}$
end for
end while

comparison between patients and between clusters. Most of the sepsis mortality is caused by sequential organ failure; sequential organ failure assessment score (SOFA score) is the assessment that measures the severity of the organ dysfunction [2]. We decided to use first day SOFA score to represent the state of the patient; the performance of the model would then be validated through the statistical comparisons between clusters.

5.4 Experiments, Results and Discussion

5.4.1 Source of the data

The MIMIC-III data base contains all the details of all patients throughout his/her ICU stay at Beth Israel Deaconess Medical Center. One important thing to consider is to control the period that is included for each patient.

The first way would be take all the data and take a mean of that for every feature and patient; since patients are assumed to be treated as soon as they are in the ICU, taking all the data would allow the model to partition not only the initial status of the patients, but also the response of the patients to the treatments. However, this would make the model only useful in studying the disease since there is no way to group an

actual future patients before he/she is discharged from the ICU. The second way would instead take a period of time of the patient's stay, which enables early cluster assignment of newly admitted patients.

Due to limited amount of time and resource availability, we proposed to include two experiments in this study.

- First experiment that includes all the data (from all periods), mean is taken for each feature/patient combination.
- Second experiment includes only the first reading of the first day lab result.

5.4.2 Experiment 1 with All periods

This experiment utilizes data from all periods for each patient.

5.4.2.1 Experiment 1 Setup

The k-means clustering algorithm is run with the pre-processed data (all periods); in order to find the best number of clusters k, the algorithm is run for 9 iterations, with the number of clusters k from (1,9]; sum of squared error is used as cluster loss. The resulting graph is displayed in Figure 5.2. It is decided to use cluster number k=3; as increasing k=4 does not lower the cluster loss as much.

5.4.2.2 Results and Discussion

The patient outcome grouped by cluster is displayed in Table 5.4. Cluster 1 is the most popular group with 3241 patients (57.8%), Cluster 0 with 1822 patients (32.5%) and Cluster 2 with 9.64%. It is clear that across the 3 clusters, there is a huge difference in mortality rate; cluster 2 has the most severe outcome with mortality rate at 68.5% while cluster 1 has the least severe outcome with mortality rate at only 5.8%. Cluster 2 with only 9.64% of the patients, contributed 67.8% of the mortality. It is safe to say that the designed algorithm has very successfully partitioned the patients into three groups.

Table 5.4. Patient Outcome K-means Experiment 1

Cluster	0	1	2
Group Size	1822	3241	540
Mortality (%)	13.1	5.8	68.5

SOFA score is divided into 6 parts (respiration, coagulation, liver, cardiovascular, CNS and renal) and each part is evaluated separately (detailed explanation of SOFA can be

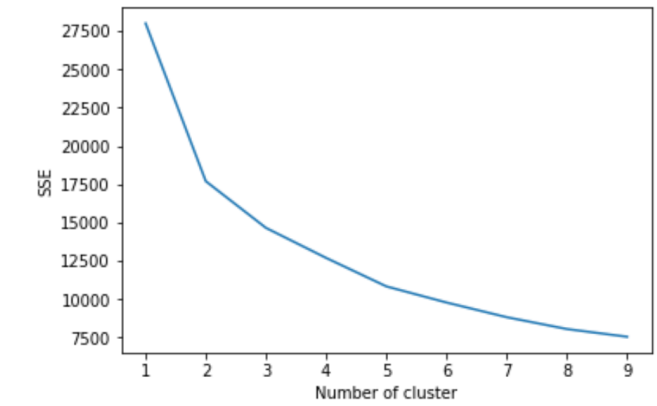

Figure 5.2. Elbow Method for Experiment 1

found in Chapter 2). Table 5.5 displays the patients' first SOFA score breakdown statistics grouped by cluster. It is clear that the group that suffers from the highest mortality rate has a SOFA score significantly higher than the other two groups. Cluster 2, where all of the most severe patients are, suffer the most from respiration and cardiovascular organ failure, which is the most frequently seen and caused the most fatality for sepsis patients.

5.4.2.3 Experiment 1 Conclusions

In this experiment, using all the data, three phenotypes are derived; the clusters show distinct characteristic in terms of patient outcome, with the highest group at 68.5% and lowest group at 5.8%.

5.4.3 Experiment 2 with First observation

In this experiment, for each patient and feature combination, we only include the first observation. In other words, when a patient is admitted into the ICU, for each lab measurement (feature), only his/her first observation is included in this experiment.

5.4.3.1 Experiment 2 Setup

The design of the experiment is the same as experiment 1. The k-means clustering algorithm is run with the pre-processed data (first day lab); in order to find the best number of clusters k, the algorithm is run for 9 iterations, with number of clusters k from (1,9]; sum of squared error is used as cluster loss. The resulting graph is displayed in Figure 5.3. Same as experiment 1, cluster number k=3 is used.

Table 5.5. Patient First day SOFA score statistics Experiment 1

				Cluster 0			
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	6.055	1.873	0.5798	0.628	1.581	1.139	0.9088
std	3.238	1.492	0.8630	1.001	1.308	1.330	1.162
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	4.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	5.000	2.000	0.000	0.000	1.000	1.000	0.000
75%	8.000	3.000	1.000	1.000	3.000	2.000	1.000
max	20.000	4.000	4.000	4.000	4.000	4.000	4.000
				Cluster 1			
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	4.3625	1.638	0.5472	0.7671	1.267	0.6602	1.029
std	2.309	1.183	0.8728	1.102	1.014	0.7755	1.240
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	2.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	4.000	2.000	0.000	0.000	1.000	1.000	1.000
75%	6.000	2.000	1.000	2.000	1.000	1.000	2.000
max	15.000	4.000	4.000	4.000	4.000	4.000	4.000
				Cluster 2			
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	8.390	2.249	0.7012	0.8214	2.188	1.329	1.562
std	4.314	1.572	1.019	1.151	1.545	1.698	1.468
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	5.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	8.000	3.000	0.000	0.000	1.000	0.000	1.000
75%	12.000	4.000	1.000	2.000	4.000	3.000	3.000
max	21.000	4.000	4.000	4.000	4.000	4.000	4.000

Table 5.6. Patient Outcome K-means Experiment 2

Cluster	0	1	2
Group Size	1331	2513	1759
Mortality (%)	14.88	9.15	20.9

5.4.3.2 Experiment 2 Results and Discussion

The result group's outcome is displayed in Table 5.6. Cluster 1 is the most popular group with 2513 patients (44.9%), Cluster 0 with 1822 patients (23.8%) and Cluster 2 with 1759 patients (31.4%). It is clear that across the 3 clusters, there is a difference in mortality rate, although not as clear as experiment 1; cluster 2 has the most severe outcome with mortality rate at 20.9% while cluster 1 has the least severe outcome with

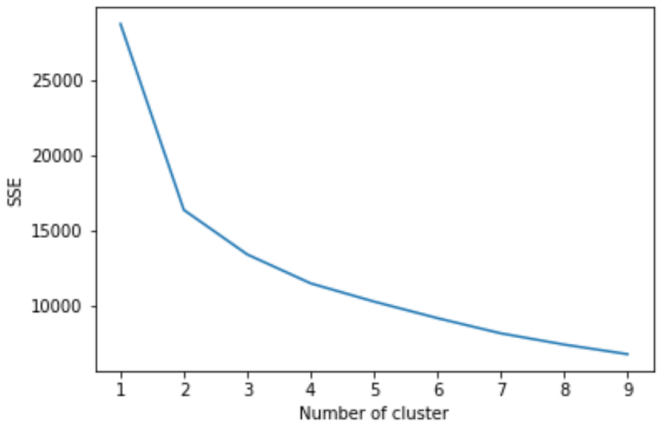

Figure 5.3. Elbow Method for Experiment 2

Table 5.7. Patient First SOFA score break down Experiment 2

	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
Cluster 0							
mean	5.380	2.238	0.530	0.730	1.424	0.814	1.292
std	3.214	1.278	0.907	1.074	1.232	0.976	1.354
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	2.000	0.000	0.000	1.000	0.000	0.000
50%	5.000	2.000	0.000	0.000	1.000	1.000	1.000
75%	7.000	3.000	1.000	1.000	1.000	1.000	2.000
max	21.000	4.000	4.000	4.000	4.000	4.000	4.000
Cluster 1							
mean	4.548	1.646	0.596	0.836	1.306	0.726	0.998
std	2.629	1.328	0.899	1.143	1.036	0.842	1.217
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	4.000	2.000	0.000	0.000	1.000	1.000	1.000
75%	6.000	3.000	1.000	2.000	1.000	1.000	2.000
max	19.000	4.000	4.000	4.000	4.000	4.000	4.000
Cluster 2							
mean	6.318	1.848	0.571	0.572	1.702	1.153	0.915
std	3.466	1.505	0.849	0.963	1.368	1.484	1.198
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	4.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	5.000	2.000	0.000	0.000	1.000	0.000	0.000
75%	8.000	3.000	1.000	1.000	3.000	3.000	1.000
max	21.000	4.000	4.000	4.000	4.000	4.000	4.000

mortality rate at only 9.15%. It is safe to say that with only patient's first lab result, the designed algorithm can also partitioned the patients into three groups that present different characteristics.

Table 5.7 displays the patients' first SOFA score break down statistics grouped by cluster. Similar to experiment 1, it is clear that the group that suffers from the highest mortality rate has a SOFA score significantly higher than the other two groups. One thing to notice is that although cluster 0 presents lower mortality rate than cluster 2 (14.88% vs 20.9%), cluster 0 presents the most severe respiratory and renal failure out of the three clusters; cluster 1, being the least fatal, has the highest coagulation and liver failure. This observation shows that the result not only presents good separation on patient outcomes, but also partitions different symptoms into different groups.

5.4.3.3 Experiment 2 Conclusions

In this experiment using first lab result, three phenotypes are discovered, with the first group patients (population: 44.9%, mortality: 14.88%) having high possibility of respiratory and renal failure; the second group of patients (population: 23.8%, mortality: 9.15%) having high possibility of liver and coagulation failure and third group of patients (population: 31.4%, mortality: 20.9%) having high possibility of cardiovascular and CNS failure.

5.4.4 Future Study

From the previous experiments, it can be seen that experiment 1 (all period) partitions the patient much better than experiment 2 (first lab) in terms of patient outcome and first day SOFA score; we think that this might be caused by different response to medical treatment across patients given that the treatment in Beth Israel Deaconess Medical Center is standard. Future studies can focus on studying the response to medication of sepsis patients.

Experiment 2 showcased a result with each group of patients having a higher probability of having a certain kind of organ failure, which we believe can drive the development of a more customized treating plans for patients. Furthermore, by studying some common characteristics of elements within the groups, future studies can focus on the reason why some specific organ failure is developed and research ways to prevent such symptoms.

Chapter 6
Deep Clustering Model

As patients are admitted into the ICU, he/she is expected to be treated immediately. With the previous model (to use mean of all observations for each feature), from the data, we can infer information such as whether the patient positively reacted to the medical treatment. With such data preparation, two patients A and B are identical if the mean of each features are all the same. However, it is possible that patient A and B have completely different reactions to treatments and outcomes. For example, patient A could experience early recovery but worsen condition during later times, which would lead to an unsatisfying outcome; patient B could experience slow recovery throughout his/her stay at a ICU and has a satisfying outcome. In this model, in order to address this concern, we propose to include multiple observations for each patient and feature combination, resulting in a multivariate time series representing each patient. In this chapter, in order to discover whether temporal information is important in partitioning sepsis patients, a deep time series clustering algorithm is built. With the inclusion of temporal data, we aim to use only very easy-to-access lab measurements (features). A flowchart of this model is displayed in Figure 6.1.

6.1 Data Pre-processing

In practice, measurements are not taken at the same frequency; for example, heart rate is monitored regularly but measurements like bilirubin are only measured around once per day, or even just once, if the reading is normal. The time interval for the multivariate time series is 4 hours, which seems like long, but still cause null values because of the reasons above. We believe that null values caused by low monitoring frequencies cannot be treated as actual missing values, to be dropped, filled in with mean/median, etc. We performed Last Observation Carried Forward (LOCF) and Next Observation

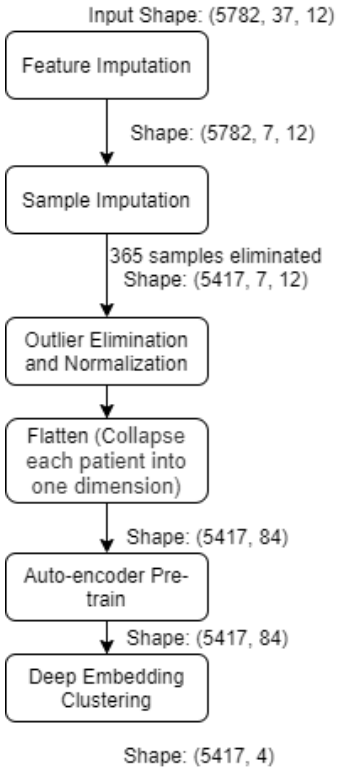

Figure 6.1. Deep clustering model data pipeline

Carried Backward (NOCB) on all patients. Here we made one assumption that if the measurement reading is missing, the most accurate value to fill is the non-missing value around that time point for that patient and feature.

After filling in null values, we have performed feature and sample elimination. A threshold = 90% has been set to eliminate all features that contain >10% null value; then all samples (patients) that contains null values are discarded. Outliers are also treated; all samples that present Z score higher than 3 are dropped. Further, there is a significant difference in mean and standard deviation between measurements; for example, a normal heart rate ranges from 60 to 100 beats per minute while normal creatinine ranges from 0.5 to 1.2 mg/dL. Hence, every measurement in the multivariate time series are normalized to (0, 1).

After pre-processing, 7 measurements (heart rate, respiratory rate, hemoglobin, white blood cell, creatinine, glucose and sodium) are used to train the model. All the 7 measurements should be common in any Intensive Care Unit. We are particularly interested in the progression of patients during the stay; we selected 48 hours of ICU stay with 4 hours interval, each patient's information is finally stored in a multivariate time series with 12 time-steps and 7 features.

6.2 Method

6.2.1 Auto-encoder

Auto-encoder is a kind of neural network structure that tries to reconstruct the input. Let us denote X as the input data and F as embedding space, this is the most basic architecture of an auto-encoder for image 6.4. Let us denote $\mathbb{Z}$ as the encoder that maps input data $\mathbb{X}$ to feature space $\mathbb{F}$; $\mathbb{Z}'$ as the decoder that maps feature space $\mathbb{F}$ back to $\mathbb{X}$. This architecture is trained on the mean square error between the input data and reconstructed data.

$$\mathbb{Z} : \mathbb{X} \to \mathbb{F} \tag{6.1}$$

$$\mathbb{Z}' : \mathbb{F} \to \mathbb{X} \tag{6.2}$$

$$\mathbb{Z}, \mathbb{Z}' = argmin||\mathbb{X} - (\mathbb{Z} \circ \mathbb{Z}'\mathbb{X})||^2 \tag{6.3}$$

Auto-encoder is well used for feature embedding. In this study, we first pre-trained a

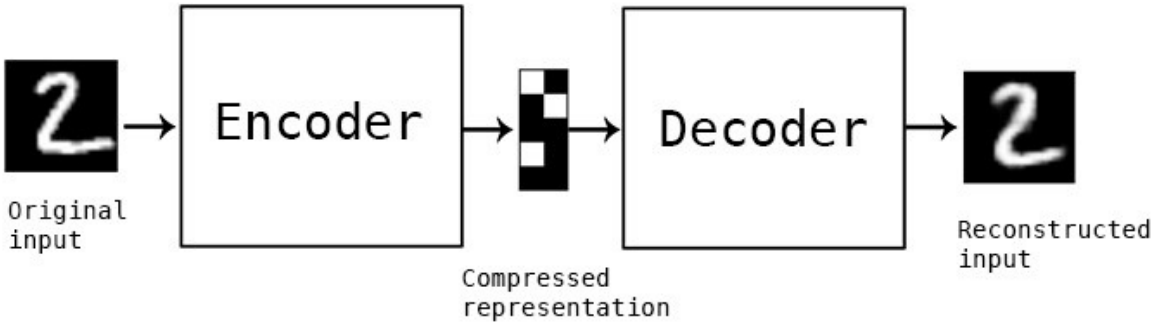

Figure 6.2. Auto-encoder architecture for a image

auto-encoder A to map X to X, intermediate latent space F can be interpreted as the feature embedding of input X. The decoder Z' of A is dropped after the auto-encoder converges; a soft assignment layer is connected to the encoder Z. t-distribution is used to measure the similarity between point z_i and centroid u_j. a is the degree of freedom of the t-distribution, which is set to 1 in this study.

$$q_{ij} = \frac{\left(1 + \|z_i - \mu_j\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'} \left(1 + \|z_i - \mu_{j'}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \tag{6.4}$$

Kullback-Leibler (KL) divergence measures the distance between the two distributions P and Q. It can be interpreted as the information gain going from prior distribution Q to posterior distribution P. KL Divergence loss is used to match soft assignment to the target distribution.

$$Loss = D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{6.5}$$

$$loss = D_{\mathrm{KL}}(P\|Q) = \sum_{i}\sum_{j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{6.6}$$

Following the original deep embedding clustering [1], we used an auxiliary distribution to update the cluster assignments.

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}} \tag{6.7}$$

The resulting cluster from the adapted deep embedding clustering architecture would be analyzed and explained using medical domain knowledge.

6.3 Implementation and Experiment

In this section, the experiment is explained in detail with network structure and hyper-parameter usage.

6.3.1 Network Structure and training

Table 6.1. Encoder structure

. layer	unit	activation
Dense	500	ReLu
Dense	500	ReLu
Dense	2000	ReLu
Dense	20	None

A multi-layer fully connected auto-encoder is trained, the encoder structure is displayed in Table 6.1, the decoder structure is displayed in Table 6.2; the encoder and decoder are symmetrical with 500, 500 and 2000 units for each layer in the encoder; note the output of the decoder is 84, which is the same as the input dimension of encoder. The auto-encoder

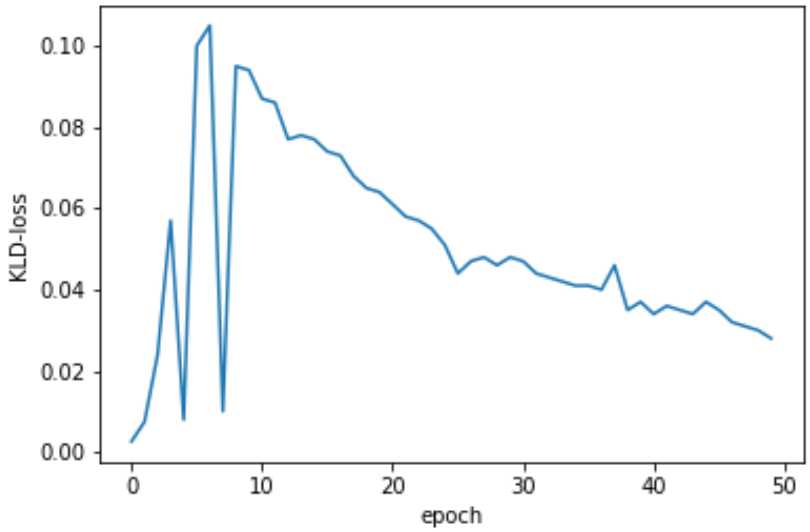

Figure 6.3. KL-Divergence Loss

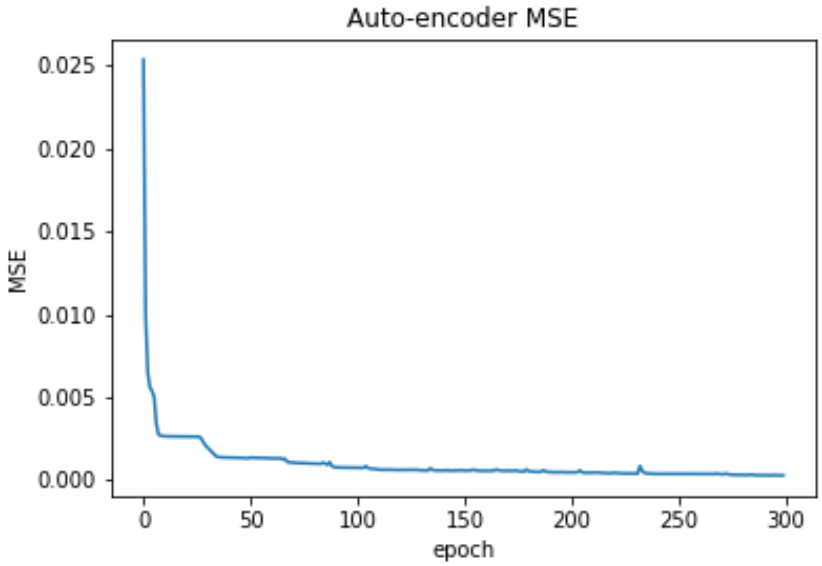

Figure 6.4. Auto-encoder loss

is trained for 300 epochs and its mean square error of input and output converged to $2.86 * 10^{-4}$ as shown in Figure 6.4. The encoder and soft assignment clustering layer are trained jointly with stochastic gradient descent (learning rate: 0.01, momentum: 0.9, batch size: 64); the auxiliary target distribution is updated every 200 steps. The model converges after 7800 steps with KLd loss of 0.028 (Figure 6.3).

Table 6.2. Decoder structure

layer	unit	activation
Dense	2000	ReLu
Dense	500	ReLu
Dense	500	ReLu
Dense	84	None

6.4 Discussion

Table 6.3 contains summary of patient SOFA score statistics of each cluster and Table 6.4 displays the average mortality rate of each cluster and its group size (how many people are in that cluster). It is clear that the outcome (mortality) of the patient is well separated with group 1 the highest at 20.8% and group 0 being the lowest at 9.1%; SOFA scores are also distributed differently among clusters. We can conclude the following. Patients from Cluster 1 (mortality: 9.1%) are likely to develop CNS failure; patients from cluster 1 (mortality: 20.8%) are more likely to develop respiratory failure; patients from cluster 2 (mortality: 15.6%) are more likely to develop renal failure; patients from cluster 3 (mortality: 12.5%) do not have a high chance of developing a specific organ dysfunction.

It is very interesting comparing Table 6.3 with Table 5.7 (experiment 2 chapter 5), with the patient outcome similarly separated among the two experiments, SOFA score distribution presents a noticeable difference. We can see that Table 5.7 presents a good separation on the SOFA score between groups:

- The higher the SOFA score, the higher the mortality rate.
- Each cluster has a higher chance to develop some organ dysfunction over other clusters, despite significant differences between clusters in terms of mean SOFA score and patient outcome.

However, in Table 6.3, SOFA score is not as separated as Table 5.7. In Table 6.3, the relationship between mortality rate and SOFA score is not as clear; cluster 2 has a higher SOFA score than cluster 1, despite having much lower (-5.2%) mortality rate. We suspect the reason behind this is because our deep clustering model captured some interactions between features that are hidden deeper. SOFA score is calculated through bio-markers of patients and most of them are used as features in both models. Bio-markers of patients

Table 6.3. Patient first day SOFA Score statistics grouped by cluster from Deep Embedding Clustering model

Cluster 0							
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	4.403	1.648	0.468	0.573	1.307	0.919	0.700
std	2.442	1.393	0.771	0.945	1.028	1.108	0.935
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	2.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	4.000	2.000	0.000	0.000	1.000	1.000	0.000
75%	6.000	3.000	1.000	1.000	1.000	1.000	1.000
max	21.000	4.000	4.000	4.000	4.000	4.000	4.000
Cluster 1							
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	5.603	2.213	0.663	0.807	1.459	0.896	0.877
std	3.474	1.407	0.981	1.075	1.282	1.154	1.101
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	1.000	0.000	0.000	1.000	0.000	0.000
50%	5.000	3.000	0.000	0.000	1.000	0.000	1.000
75%	7.000	3.000	1.000	2.000	1.000	1.000	1.000
max	20.000	4.000	4.000	4.000	4.000	4.000	4.000
Cluster 2							
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	5.978	1.807	0.630	0.838	1.545	0.801	1.666
std	3.354	1.448	0.935	1.224	1.245	1.065	1.523
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	5.000	2.000	0.000	0.000	1.000	0.000	1.000
75%	8.000	3.000	1.000	2.000	1.000	1.000	3.000
max	21.000	4.000	4.000	4.000	4.000	4.000	4.000
Cluster 3							
	SOFA	respiration	coagulation	liver	cardiovascular	cns	renal
mean	5.148	1.741	0.552	0.726	1.447	0.917	0.910
std	3.004	1.432	0.871	1.049	1.227	1.149	1.137
min	2.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	0.000	0.000	0.000	1.000	0.000	0.000
50%	4.000	2.000	0.000	0.000	1.000	1.000	1.000
75%	7.000	3.000	1.000	1.000	1.000	1.000	1.000
max	20.000	4.000	4.000	4.000	4.000	4.000	4.000

grouped by cluster is shown in Table 6.5; patients from cluster 1 (mortality: 20.8%) suffer from high heart rate; patients from cluster 0 (mortality: 9.1%) suffer from high

respiratory rate; patients from cluster 2 (mortality: 15.6%) suffer from high Creatinine level (risk of renal failure). By comparing bio-markers' statistics with patient outcome, we cannot see reasonable pattern that connect the two. With the patients well separated between clusters in terms of outcome (mortality), it means that there are some high dimensional features that the deep clustering model has captured.

Table 6.4. Patient Outcome from Deep Embedding Clustering model

Cluster	Mortality (%)	Group Size
0	9.1	1694
1	20.8	1147
2	15.6	1483
3	12.5	1093

Table 6.5. Bio-markers of patients grouped by cluster from Deep Embedding Clustering model

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Heart Rate (bpm)	72.13	105.0	82.55	95.77
Respiratory Rate (bpm)	41.85	23.66	18.36	19.22
WBC (kilo count/ml)	11.26	13.46	12.76	12.64
Creatinine (mg/dL)	1.11	1.28	2.57	1.36
Glucose (serum) (mg/dL)	138.4	148.3	136.3	146.2
Sodium (serum) (mg/dL)	140.9	140.5	135.6	139.2
Hemoglobin (g/dL)	8.96	9.19	9.62	11.80

6.5 Future Work

Although initial results look promising, there are limitations that remain in this thesis and room for improvement.

6.5.1 Number of phenotypes

In this study, we selected 4 as the number of clusters and shows the feasibility of using deep representation learning to partition sepsis patients; Future work can focus on discovering optimal number of clusters, it can be done using average silhouette method or Gap statistic method.

6.5.2 Introduction of temporal information

Temporal information is introduced with a purpose of obtaining how do patients recover in the ICU. MIMIC-III database is collected from one hospital, it is safe to assume that the treatment used for all sepsis patients are similar in most ways. If the study is extended to larger population base, use of temporal information may not be viable.

6.5.3 Architecture of Auto-encoder

In this study, fully connected autoencoder is used to obtain the abstract information of patients' EHR. With the introduction of temporal information, LSTM based autoencoder can be used instead to better capture temporal features.

6.5.4 Time domain information

We know that the deep clustering model learns some higher dimensional feature, which results in good separation of patients in terms of patient outcome. Future study can focus on unmasking the higher dimensional features, in other words, study the difference between patients from different clusters. We believe that it can provide us more insights regarding the true cause of mortality associated with sepsis.

Chapter 7
Conclusions

In this retrospective study, two novel phenotype discovery methods have been successfully developed and tested on MIMIC-III database; each of them showing promising results according to the designed validation process. In the first method, we took a traditional approach and developed a Principal Component Analysis (feature embedding) Incorporated with K means (unsupervised learning); the elbow method is used to find the best number of clusters. This model is run for two experiments. The first experiment utilizes data from all periods for each patient, three phenotypes are derived; the clusters show distinct characteristics in terms of patient outcome, with the highest group at 68.5% and lowest group at 5.8% in terms of mortality rate. The second experiment utilizes only first lab result for each patient, three phenotypes are discovered, with the first group of patients (population: 44.9%, mortality: 14.88%) having high possibility of respiration and renal failure; the second group of patients (population: 23.8%, mortality: 9.15%) having high possibility of liver and coagulation failure and third group of patients (population: 31.4%, mortality: 20.9%) having high possibility of cardiovascular and CNS failure.

In the second model, we adapted [1] to cluster sepsis patients into novel phenotypes. Temporal features are introduced to capture the effect of time on the patient's progress.The derived 4 phenotypes present a patient mortality standard deviation of 4.97%. SOFA means are also compared between clusters, which has shown some characteristic of each clusters. With bio-markers (in the current assessment and definition of sepsis and severe sepsis) not significantly different and outcome clearly separated, there are some hidden patterns that the clustering architecture captured. The results indicate that our deep clustering model can discover new features that are originally hidden in the high-dimensional space and cluster patients into meaningful groups, which we believe can lead to the discovery of the true cause of sepsis mortality.

Bibliography

[1] Xie, J., R. Girshick, and A. Farhadi (2016) "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, pp. 478–487.

[2] Vincent, J.-L., A. De Mendonça, F. Cantraine, R. Moreno, J. Takala, P. M. Suter, C. L. Sprung, F. Colardyn, and S. Blecher (1998) "Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study," *Critical care medicine*, **26**(11), pp. 1793–1800.

[3] Ferreira, F. L., D. P. Bota, A. Bross, C. Mélot, and J.-L. Vincent (2001) "Serial evaluation of the SOFA score to predict outcome in critically ill patients," *JAMA*, **286**(14), pp. 1754–1758.

[4] Bone, R. C., R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald (1992) "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, **101**(6), pp. 1644–1655.

[5] Singer, M., C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, et al. (2016) "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *JAMA*, **315**(8), pp. 801–810.

[6] Makaju, S., P. Prasad, A. Alsadoon, A. Singh, and A. Elchouemi (2018) "Lung cancer detection using CT scan images," *Procedia Computer Science*, **125**, pp. 107–114.

[7] Padmanabhan, R., N. Meskin, and W. M. Haddad (2017) "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment," *Mathematical biosciences*, **293**, pp. 11–20.

[8] Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark (2016) "MIMIC-III, a freely accessible critical care database," *Scientific data*, **3**, p. 160035.

[9] Seymour, C. W., J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez, et al. (2019) "Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis," *JAMA*, **321**(20), pp. 2003–2017.

[10] Zhou, F., T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. (2020) "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *The Lancet*.

[11] Fan, E., D. Brodie, and A. S. Slutsky (2018) "Acute respiratory distress syndrome: advances in diagnosis and treatment," *JAMA*, **319**(7), pp. 698–710.

[12] Goldman, L. and A. I. Schafer (2011) *Goldman's cecil medicine E-book*, Elsevier Health Sciences.

[13] Clark, A. L. and H. Dargie (2011) *Oxford textbook of heart failure*, Oxford University Press.

[14] Goyette, R. E., N. S. Key, and E. W. Ely (2004) "Hematologic changes in sepsis and their therapeutic implications," in *Seminars in respiratory and critical care medicine*, vol. 25, Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New ..., pp. 645–659.

[15] Pustavoitau, A. and R. D. Stevens (2008) "Mechanisms of neurologic failure in critical illness," *Critical care clinics*, **24**(1), pp. 1–24.

[16] Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986) "Learning representations by back-propagating errors," *nature*, **323**(6088), pp. 533–536.

[17] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996) "From data mining to knowledge discovery in databases," *AI magazine*, **17**(3), pp. 37–37.

[18] Pearson, K. (1901) "On lines of closes fit to system of points in space, London, E dinb," *Dublin Philos. Mag. J. Sci*, **2**, pp. 559–572.

[19] Gunter, T. D. and N. P. Terry (2005) "The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions," *Journal of medical Internet research*, **7**(1), p. e3.

[20] Huber, T. P., S. M. Shortell, and H. P. Rodriguez (2017) "Improving care transitions management: examining the role of accountable care organization participation and expanded electronic health record functionality," *Health services research*, **52**(4), pp. 1494–1510.

[21] Cebul, R. D., T. E. Love, A. K. Jain, and C. J. Hebert (2011) "Electronic health records and quality of diabetes care," *New England Journal of Medicine*, **365**(9), pp. 825–833.

[22] Churpek, M. M., T. C. Yuen, S. Y. Park, R. Gibbons, and D. P. Edelson (2014) "Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards," *Critical care medicine*, **42**(4), p. 841.

[23] Herrin, J., B. da Graca, D. Nicewander, C. Fullerton, P. Aponte, G. Stanek, T. Cowling, A. Collinsworth, N. S. Fleming, and D. J. Ballard (2012) "The effectiveness of implementing an electronic health record on diabetes care and outcomes," *Health services research*, **47**(4), pp. 1522–1540.

[24] Ramirez, A. H., Y. Shi, J. S. Schildcrout, J. T. Delaney, H. Xu, M. T. Oetjens, R. L. Zuvich, M. A. Basford, E. Bowton, M. Jiang, et al. (2012) "Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record," *Pharmacogenomics*, **13**(4), pp. 407–418.

[25] Knochel, J. P. (1977) "The pathophysiology and clinical characteristics of severe hypophosphatemia," *Archives of internal medicine*, **137**(2), pp. 203–220.

[26] Nemati, S., A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman (2018) "An interpretable machine learning model for accurate prediction of sepsis in the ICU," *Critical care medicine*, **46**(4), pp. 547–553.

[27] Che, Z., S. Purushotham, K. Cho, D. Sontag, and Y. Liu (2018) "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, **8**(1), pp. 1–12.

[28] Prasad, N., L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt (2017) "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*.
URL `http://auai.org/uai2017/proceedings/papers/209.pdf`

[29] Slee, V. N. (1978) "The International classification of diseases: ninth revision (ICD-9)," *Annals of internal medicine*, **88**(3), pp. 424–426.

[30] Ibrahim, I., I. G. Jacobs, S. A. Webb, J. Finn, et al. (2012) "Accuracy of International classification of diseases, 10th revision codes for identifying severe sepsis in patients admitted from the emergency department," *Critical Care and Resuscitation*, **14**(2), p. 112.

[31] Dellinger, R. P., M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke, et al. (2013) "Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock, 2012," *Intensive care medicine*, **39**(2), pp. 165–228.

[32] Schölkopf, B., A. Smola, and K.-R. Müller (1998) "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, **10**(5), pp. 1299–1319.

[33] Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013) "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119.

[34] Perozzi, B., R. Al-Rfou, and S. Skiena (2014) "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710.

[35] Kipf, T. N. and M. Welling (2017) "Semi-supervised classification with graph convolutional networks," *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[36] Veličković, P., G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio (2017) "Graph attention networks," *arXiv preprint arXiv:1710.10903*.

[37] Yang, B., X. Fu, N. D. Sidiropoulos, and M. Hong (2017) "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, pp. 3861–3870.

[38] Chen, G. (2015) "Deep learning with nonparametric clustering," *arXiv preprint arXiv:1501.03084*.

[39] Huang, P., Y. Huang, W. Wang, and L. Wang (2014) "Deep embedding network for clustering," in *2014 22nd International conference on pattern recognition*, IEEE, pp. 1532–1537.

[40] Johnson, A. E., J. Aboab, J. D. Raffa, T. J. Pollard, R. O. Deliberato, L. A. Celi, and D. J. Stone (2018) "A comparative analysis of sepsis identification methods in an electronic database," *Critical care medicine*, **46**(4), p. 494.

[41] Teasdale, G. and B. Jennett (1974) "Assessment of coma and impaired consciousness: a practical scale," *The Lancet*, **304**(7872), pp. 81–84.