Technische Universität Berlin

Master Thesis

Data-driven pattern discovery for large scale time series analytics in healthcare

Safiya Hashmi

Matriculation #: 399276

Supervisors: Prof. Dr. Markl, Volker
Advisors: Dr. Charfuelan, Marcela & Dr. Schwerk, Anne

30/11/2020

Erklärung (Declaration of Academic Honesty)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

I hereby declare to have written this thesis on my own and without forbidden help of others, using only the listed resources.

Datum	Safiya Hashmi

Contents

List of Figures

List of Listings

List of Tables

1 English Abstract

The growth of electronic data presents unprecedented opportunities for improvements in healthcare. However, it is often difficult to process and analyze such large-scale data efficiently and timely. Machine learning (ML) plays an important role in healthcare and is being applied increasingly to healthcare for computer-aided diagnosis, analyzing medical images, and predicting severe medical conditions. Thus, ML methods can be used to analyze such large-scale data and discover patterns to diagnose or predict a medical condition in advance and in real-time scenarios - thereby greatly decreasing time and human resources. This thesis proposes a data-driven ML approach to analyze and discover patterns in large-scale time series data continuously generated by the bedside monitors of ICU patients. Septic shock is one of the most critical medical conditions and a major cause of mortality in intensive care units (ICUs). Therefore, we present a data-driven tandem approach for predicting septic shock in sepsis patients by discovering patterns in commonly available large-scale vital signs' time series data. We perform multilinear regression to obtain a time series of the SOFA score, a score used to assess organ failure relevant to sepsis, from the commonly available vital signs time series data. We then predict septic shock occurrence by discovering patterns in large-scale vital signs and SOFA score time series data recorded between sepsis onset time and septic shock onset time. We use a data-driven feature engineering method of a sliding window, i.e. taking a window of fixed length of 1 hour initially and sliding it ahead every 15 mins to the next time step, to generate sequences from time series data of vital signs and the SOFA score and perform sequence classification to predict septic shock. To this end, we evaluated the proposed approach and observed that patients in the same group, i.e. patients that progressed into septic shock, showed high variability as expressed by high standard deviations in their vital signs and SOFA scores. High variability in data leads to an increase in heterogeneity, making it challenging to discover patterns in such highly heterogeneous data. As a result, the prediction of septic shock was not efficient in patients with highly variable data. Hence, we expect that the group of patients that progressed into septic shock was composed of subgroups that we did not account for, eventually due to a lack of extended data, such as comorbid conditions, drug usage etc, which could have overshadowed the differences we wanted to assess; yet this has to be verified by additional experiments. Another reason could be the improper assignment of sliding windows, which could have introduced

unwanted heterogeneity, which is not indicative of true differences between the shock and non-shock groups.

2 Deutscher Abstract

Das Anstieg elektronischer Daten bietet immense Möglichkeiten für Verbesserungen im Gesundheitswesen. Es ist jedoch oft schwierig, solche umfangreichen Daten effizient und zeitnah zu verarbeiten und zu analysieren. Maschinelles Lernen (ML) spielt eine wichtige Rolle im Gesundheitswesen und wird hier zunehmend für die medizinische Bildgebung, die computergestützte Diagnose und die Vorhersage schwerer Erkrankungen eingesetzt. Somit können ML-Verfahren verwendet werden, um solche großen Datenmengen zu analysieren und Muster zu entdecken, um einen medizinischen Zustand im Voraus zu diagnostizieren oder vorherzusagen. Diese Arbeit schlägt einen datengesteuerten ML-Ansatz vor, um Muster in großen Zeitreihendaten zu analysieren und zu entdecken, die kontinuierlich von den Monitoren am Krankenbett in der Notaufnahme erzeugt werden. Der septische Schock ist eine der kritischsten Erkrankungen und eine der Haupttodesursachen auf Intensivstationen. Daher präsentieren wir einen datengesteuerten Tandem-Ansatz zur Vorhersage des septischen Schocks bei Sepsis-Patienten, indem Muster in allgemein verfügbaren Zeitreihendaten von Vitalzeichen in großem Maßstab analysiert werden. Wir führen eine multilineare Regression durch, um aus den allgemein verfügbaren Zeitreihendaten der Vitalfunktionen eine Zeitreihe des SOFA-Scores zu erhalten. Wir sagen dann das Auftreten eines septischen Schocks voraus, indem wir Muster in den Vitalfunktionen und Zeitreihendaten des SOFA-Scores entdecken, die zwischen dem Beginn der Sepsis und dem Beginn des septischen Schocks aufgezeichnet wurden. Wir verwenden eine datengesteuerte Feature-Engineering-Methode eines Schiebefensters, d.h. zunächst ein Fenster mit fester Länge von 1 Stunde nehmen und alle 15 Minuten zum nächsten Zeitschritt vorschieben, um Sequenzen aus Zeitreihendaten von Vitalfunktionen und SOFA-Score zu generieren und eine Sequenzklassifizierung durchzuführen, um einen septischen Schock vorherzusagen. Zu diesem Zweck bewerteten wir den vorgeschlagenen Ansatz und beobachteten, dass Patienten derselben Gruppe, d. H. Patienten, bei denen ein septischer Schock auftrat, eine hohe Variabilität zeigten, die sich in hohen Standardabweichungen ihrer Vitalfunktionen und SOFA-Werte äußerte. Eine hohe Variabilität der Daten führt zu einer Zunahme der Heterogenität, was es schwierig macht, Muster in solch sehr heterogenen Daten zu entdecken. Infolgedessen war die Vorhersage eines septischen Schocks bei Patienten mit stark variablen Daten nicht effizient. Wir gehen daher davon aus, dass sich die Gruppe der Patienten, bei denen ein septischer Schock auftrat, aus Un-

tergruppen zusammensetzte, die wir nicht berücksichtigten, möglicherweise aufgrund des Mangels an erweiterten Daten wie komorbiden Zuständen, Drogenkonsum usw., die die Unterschiede, die wir hatten, überschatten könnten wollte beurteilen; Dies muss jedoch durch zusätzliche Experimente überprüft werden. Ein weiterer Grund könnte die falsche Zuordnung von Schiebefenstern sein, die zu unerwünschter Heterogenität führen könnte, was nicht auf echte Unterschiede zwischen der Schock- und der Nicht-Schock-Gruppe hinweist.

3 Introduction

With the progression of healthcare digitization, massive amounts of digital data is generated by modern hospitals every day at a tremendous rate. This data includes electronic health records (EHRs) of patients', which include results of medical examinations, lab data, medications used, insurance billing information (i.e. ICD-10 codes), and vital signs recordings,that are generated by different healthcare information and real-time health monitoring systems. Such data clearly satisfies the four characteristics of big data: volume, velocity, variety, and veracity. With the advent of machine learning (ML) methods and its applications, it is possible to analyze such big data to promptly predict or diagnose a medical condition, take preventive measures, and ensure timely and precise treatments [20].

For patients being monitored in intensive care units, the bedside monitors generate enormous amounts of continuous physiological time series data, reflecting minute or even seconds of intervals. A time series can be defined as a sequential set of data points that are measured typically over a continuous time period. It is mathematically defined as a set of vectors $x(t), t = 0, 1, 2, ...$ where t denotes the time passed [10]. The data points measured during an event in a time series are organized in a chronological order. In a medical setting, this data typically represents vital signs data such as heart rate, respiratory rate, blood pressure, temperature and oxygen saturation. The vital signs such as heart rate, respiratory rate, blood pressure are highly correlated and exhibit patterns related to clinical conditions [41]. In critical events those data can change rapidly, and even a small changes can signify an emergency and hence afford prompt interventions. They are also multi-dimensional and non-stationary. Because such a massive amount of time series data is generated rapidly, it becomes very difficult to interpret them, discover patterns, identify changes by solely human observation, as this can result in delayed treatment or even loss of life. Hence, there is an increasing need for automatic methods to recognize essential patterns in large-scale time series data. Abdur Rahim M. et al. [29] built a model, ViSiBiD to identify critical clinical events in a patient being monitored at home well in advance using knowledge discovered from the patterns of multiple vital signs from a large number of similar patients. Omar Bellorin et al. [14] performed a study to analyze a clinical correlation between abnormal vital signs and postoperative leaks and bleeding in patients undergoing laparoscopic gastric bypass (LRYGB). They conclude that patterns in time series of vital signs can be used to suspect postoperative leaks and bleeding in patients undergoing LRYGB. Jessie S. Davis et al. [21] discovered patterns in vital signs prior to the cardiopulmonary arrest due to shock that can be used to guide therapy to reverse the deteriorating condition and prevent the cardiopulmonary arrest.

To analyze such large-scale data and discover patterns, we use data-driven ML approaches in this thesis. Data-driven methods are used to analyze the data about a system to find

connections between the system state variables without the need for explicit knowledge of the physical behavior of the system and hence theory driven approaches. Using data-driven approaches, one can discover unseen correlations or patterns between the variables that were not known a priori [63].

In this thesis, we analyze the large scale time series data such as vital signs recorded by the bedside monitors for the sepsis patients to discover patterns in it and predict the occurrence of septic shock which is a critical condition in the ICU with a high mortality risk. We used the publicly available MIMIC-III Waveform Database Matched Subset for obtaining the time series of vital signs.

3.1 Motivation

3.1.1 Sepsis Definition and Prevalence

Sepsis is a life-threatening condition and a major cause of mortality in intensive care units (ICUs). More than 6 million people die annually worldwide due to sepsis and sepsis treatment is one of the most expensive treatments [28]. According to the world health organization, sepsis is a major public health problem and urges all the UN member states to improve the prevention, identification, and management of sepsis [28]. When sepsis is not identified and treated timely and appropriately, the condition can get worse and the patient may progress to septic shock. Septic shock is the most critical stage characterized by circulatory, cellular, and metabolic abnormalities and it has a higher risk of mortality than sepsis alone. Overall 37-56% of all inpatient deaths are caused by sepsis and septic shock [48]. In developed countries, 2% of the hospital admissions are diagnosed with sepsis and more than 50% are diagnosed with septic shock. Annually, for every 100,000 people in the world, the number of new cases is between 150 and 240 for sepsis, and approximately 11 for septic shock [27]. Studies have proved that when septic shock patients are treated within the first hour of diagnosis have a survival rate of 80% and delay in treatment for every hour increases the mortality rate by approximately 8%. This implies that the mortality rate increases with the delay in treatment [48].

3.1.2 Machine Learning and Septic Shock Detection

For patients being monitored intensively for sepsis, data is generated at a tremendously rapid rate by the bedside monitors, medication pumps, and ventilators, laboratory tests, etc. Such data clearly satisfies the four characteristics of big data: volume, velocity, variety, and veracity where volume refers to the huge size of generated data, velocity refers to

the speed at which data is generated, variety refers to the different types of data such as structured, semi-structured, and unstructured, and veracity refers to the quality of data. With increasing amounts of patients generated digitized data and considering the frequent need to act with high precision but under limited time frames, automatic methods to discover relevant patterns particularly in large and complex datasets such as recordings of vital signs can be significantly beneficial and critical. One way to automatically analyze such big data is to use machine learning (ML) approaches. ML, an application of artificial intelligence (AI), can be used to analyze the vast amounts of generated data, discover patterns, and predict events in the future based on the data available. In healthcare, ML can be critical in managing "golden hour" diseases such as septic shock where prompt and appropriate treatment saves lives. ML has a number of applications in healthcare because of: vast amounts of healthcare data are being captured and made available digitally, processing of large amounts of data has become cost-effective due to the increased computing power now available at affordable prices, need for better evidence-based care, and complexity of the critical illness. ML can also increase access to treatment in developing countries, it can improve the sensitivity of detection, add more value in treatment decisions, and help personalize treatment so that each patient gets the treatment that is best for them [13]. Because of the need and limited capability of the human brain to analyze vast amounts of generated data, sepsis, and septic shock is a good target for ML-based informed intervention. ML approaches can be used for analyzing the large amounts of patient's clinical and physiological data, building predictive and prognostic models, phenotyping, recognizing patterns, etc. that can result in not only improved and early diagnosis and treatment of sepsis but also to prevent the patients from progressing into a more severe stage i.e. septic shock by identifying the risk even before the onset of septic shock.

3.1.3 Heterogeneity in Sepsis

Sepsis can be caused due to a variety of infections such as viral, fungal, or parasitic and sometimes non-infectious etiologies (causes and origins) can also resemble sepsis [66]. It can be categorized according to clinical features, physiology, biology, and/or genetics. These categories are also known as sub-phenotypes, subclasses, subgroups, or endotypes [37]. Hence, we can say that sepsis is a heterogeneous disease rather than a uniform disease. Not only the early diagnosis of sepsis but also identifying the subgroups of sepsis patients that will progress into septic shock are significant to decide the clinical treatment path.

Because the infectious source is not immediately apparent, diagnosis of sepsis and septic shock is very difficult. Diagnostic tests such as blood culture tests commonly used in detection depend on the blood volume and time drawn, prior treatment with antibiotics,

presence of viable organisms which makes its sensitivity and specificity low. In addition, the delays in turnaround time of test results can expose patients to the risk of adverse events such as septic shock [66]. Techniques such as probing host response to infections can also be used as host response-based biomarkers can be measured in minutes and help in rapid identification, treatment of sepsis [66] and prevent the progression into septic shock. Biomarkers such as white blood cell count, C-reactive protein, procalcitonin, cellular function, mRNA, protein signatures, and metabolites can be used in probing the host response [66][45]. Information about the RNA that is transcribed by a genome in a specific tissue or cell type, at a specific development stage, and under a certain physiological or pathological condition is included a the transcriptome. Therefore, transcriptome profiling allows us to understand the human genome at a transcriptome level and provides comprehension of gene structure plus function, gene expression regulation, and genome plasticity. Most significantly, it can disclose the key alterations of biological processes triggering human diseases. However, transcriptomic profiling is time consuming too and not always available, thus causing further delays in diagnosis. Therefore, ability to identify the presence and severity of sepsis is very much limited. In such scenarios, commonly available physiological data such as heart rate, respiratory rate, blood pressure, etc. can be used for detecting Septic Shock as changes in the physiological variables take place gradually over a certain period of time in patients suffering from sepsis. Combining an ML approach with the commonly available data such as real-time vital signs recordings of heart rate, arterial blood pressure, respiratory rate, etc. can be helpful for early detection of Septic Shock. Using this approach of applying ML algorithms on continuously generated recordings of vital signs, we can predict whether or not the patient will progress into Septic Shock in a normal ICU setting without having to collect any special data such as blood culture and transcriptomic data. Hence, in this thesis, we use ML and large-scale time series data analysis of vital signs recorded by the bedside monitors in an ICU to learn and detect patterns related to sepsis patients that progress into septic shock.

3.1.4 Standard Risk Scoring Systems in Septic shock

Scoring systems such as systemic inflammatory response syndrome (SIRS), sequential organ failure assessment (SOFA), quick sequential organ failure assessment (qSOFA), and national early warning scores (NEWS) are some of the gold standard scoring systems that are used to diagnose sepsis and its severity in clinical practice. SIRS scoring system is used to diagnose an inflammatory condition caused by the body's response to an infectious or non-infectious insult [61]. SOFA is used to determine the extent of a patient's organ dysfunction or rate of failure [61]. qSOFA is a bedside prompt scoring system used to identify patients with suspected infection and high mortality risk [61]. NEWS is used to determine the degree of illness of a patient and further improves the detection and

response to clinical deterioration in adult patients [25]. However, these scoring systems suffer from the following limitations:

- SIRS criteria is an indicator of a systemic inflammation that can be caused due to an infectious or a non-infectious condition. Whereas, sepsis is a systemic response to infection. Hence, SIRS has low sensitivity and specificity for diagnosing septic shock [49].
- SOFA excludes lactate as an important marker and it requires calculation for subsequent days and laboratory tests to verify if the patients satisfy the diagnostic criteria for sepsis. Hence, using SOFA might delay the diagnosis of Septic Shock [49].
- qSOFA criteria has low sensitivity but high specificity to identify patients with a high risk of death. Hence, qSOFA score is used to predict mortality and morbidity and not to diagnose Septic Shock [67].
- NEWS has low sensitivity and specificity for the detection of Septic Shock. Although, it is highly sensitive to trigger an alert, it requires practitioner engagement for clinical judgment and decision making [25].

Because of the above limitations, Septic Shock diagnosis using these scoring systems does not provide accurate and reliable results which can further delay the detection of Septic Shock. In such scenarios, data-driven ML approaches might be used to automatically detect Septic Shock without having bias for any markers or high turnaround time for the diagnosis procedure.

3.2 Contributions

The contributions of this thesis are as follows:

1. This thesis is the first to propose the prediction of septic shock in sepsis patients applying the data-driven tandem ML method for pattern discovery in vital signs' time series data.
2. Stephen J. et al. [48] introduced a new state, i.e. 'pre-shock' state before the septic shock and stated that the sepsis patients that enter this new state are highly probable to progress into septic shock. To identify patients that enter this new state of pre-shock, they used the time series data of only two vitals, i.e. heart rate and respiratory rate. Additionally, they also used laboratory data such as platelet count,

creatinine, potassium, etc., in their methods. However, obtaining such laboratory data can be time consuming as the test results might not be available immediately and requires to have a clinician in loop. Hence, this makes timely prediction of septic shock impossible and increases the mortality risk.

We are the first to use an tandem method of regression and feature-based sequence classification on time series vital signs data only to predict the occurrence of septic shock in ICU patients. Using regression, we predict the SOFA scores from vital signs time series data and then use this predicted SOFA scores with vital signs to predict septic shock.

3. We provide the basis for analyzing large-scale time series of the commonly available physiological data and detecting patterns in them using ML to prevent critical clinical conditions in ICU such as septic shock and thus, providing appropriate treatment and medical care.

3.3 Thesis Outline

Section 4 In section 4, we present the definitions of sepsis and septic shock, including their identification process. This section concludes by reporting the recent studies done in the past to predict septic shock using ML approaches and their limitations.

Section 5 Section 5 provides an overview of the data sources used for this thesis. It contains information about the commonly used tables and the description of data stored in it. Further, in this section, we also explain the process followed in selecting cohort and identifying sepsis and septic shock patients based on the background information provided in section 4. We first start with selecting our initial cohort by applying basic filtering on the patients. We then identify patients with sepsis and their respective sepsis onset times. This is followed by the identification of sepsis patients that progressed into septic shock. We conclude this section by selecting our final cohort based on the availability of vital signs data in our used data source.

Section 6 Section 6 is split into three sub-sections for data extraction, pre-processing, and feature extraction. We start with extracting the time series data for patients in our final cohort selected in the previous section 5. We then describe the data pre-processing steps where we clean, transform, and perform missing value imputation. This section concludes with the feature extraction process to extract basic statistical features such as mean, maximum, minimum, sample entropy, coefficient of correlation, and standard deviation from the pre-processed time series data.

Section 7 Section 7 describes in detail the data-driven ML approach followed for pattern discovery in sepsis patients to predict the occurence of septic shock. This section consists of two main sub-sections, i.e. regression and classification. Each of these sub-sections includes the description of ML models used, experiments conducted, the evaluation metrics used, and each experiment's results.

Section 8 In section 8, we discuss and analyze the results of our best performing classification model. We compare the data of the best and worst predicted sepsis patients. We conclude this section with our observation and analysis results.

Section 9 Section 9 presents the conclusion along with the analysis made in this. It also mentions the results of the best performing classification model. Furthermore, it also provides practical limitations of this thesis and ideas to improve septic shock prediction in the future as a part of future work.

4 Background and Related Work

This chapter explains the backgrounds of this thesis. In section 4.1, we present different stages of sepsis and their respective definitions that have been revised over the past years. Section 4.2 provides the method to identify patients with suspicion of infection, which is the first step towards sepsis diagnosis. Section 4.3 presents the use of Sequential Organ Failure Score (SOFA) for diagnosis of sepsis-related organ dysfunction, and section 4.4 lists the vital signs that can be used for diagnosis and monitoring severity of sepsis.

4.1 Sepsis and Septic Shock

Clinicians agreed on three stages of sepsis i.e.sepsis, severe sepsis, and septic shock [57]. When not identified and treated timely and appropriately, the condition can worsen, and the patient may transition into severe sepsis and then even progress to septic shock, which is the most critical stage and has higher mortality rates in ICUs. The definition of sepsis, severe sepsis, and septic shock has been revised over the last two decades. Table 1 contains all the revised definitions of sepsis, severe sepsis, and septic shock, sepsis-3 definition by M. Singer et al. [61] being the latest.

In 1991, R. C. Bone et al. proposed broad definitions of SIRS, sepsis, severe sepsis, sepsis-induced hypotension, and septic shock. According to R. C. Bone et al., SIRS is defined as a systemic inflammatory response to diverse severe clinical insults. SIRS exhibits two or more of the following conditions [15]:

- Body temperature > 38°C or < 36°C
- Heart rate > 90 beats per minute
- Respiratory rate > 20 breaths per minute or PaC02 < 32 mm Hg
- White blood cell count $> 12{,}000$Icu mm, $< 4{,}000$/cu mm, or $> 10\%$ immature (band) forms.

The definitions introduced by R. C. Bone et al. [15] were widely used in practice for a large number of clinical trials and researches. However, according to the International Sepsis Definitions Conference held in 2001 [46], the former definitions introduced by R. C. Bone et al. do not allow accurate staging and prognosis of the host response to infection, and the diagnostic criteria for SIRS are excessively sensitive but non-specific. Hence, keeping the definition of severe sepsis unchanged, the definitions of sepsis and septic shock

Year	Sepsis	Severe Sepsis	Septic Shock
1991 [15]	The systemic response to infection, manifested by two or more of the SIRS criteria as a consequence of infection is known as sepsis.	Sepsis associated with organ dysfunction, hypoperfusion, or hypotension. Hypoperfusion and perfusion abnormalities may or may not include lactic acidosis, oliguria, or an acute alteration in mental status.	sepsis-induced with hypotension inspite of adequate fluid resuscitation along with the perfusion abnormalities that may or may not include lactic acidosis, oliguria, or an acute alteration in mental status. sepsis-induced hypotension is defined as a systolic blood pressure < 90 mm Hg or a reduction of 40 mm Hg from baseline in case of non-existence of other causes for hypotension.
2001 [46]	A clinical syndrome defined by the presence of both infection and systemic inflammatory response.	sepsis accompanied by organ dysfunction. Organ dysfunction can be identified by using the Sequential Organ Failure Assessment score.	Septic shock in adults refers to a state of acute circulatory failure distinguished by persistent arterial hypotension that is unexplained by other causes. Hypotension is defined by a Mean Arterial Pressure (MAP) lower than 60, systolic arterial pressure below 90 mmHg, or a reduction in systolic blood pressure of more than 40 mmHg from baseline, in spite of adequate volume resuscitation, in the absence of other cause of hypotension.
2016 [61]	A life-threatening organ dysfunction caused by a dysregulated host response to infection. organ dysfunction can be identified by an increase in the Sequential [sepsis-related] Organ Failure Assessment (SOFA) score of 2 points or more.	-	A subset of sepsis characterized by particularly profound circulatory, metabolic, and cellular abnormalities that are related to a higher risk of mortality than with sepsis alone. Septic shock patients can be clinically identified by a vasopressor requirement to maintain a MAP of 65 mm Hg or more and serum lactate level greater than 2 mmol/L ($>$ 18 mg/dL) in case of non-existence of hypovolemia.

Table 1: Definitions of sepsis, severe sepsis, and septic shock

were revised, and the list of diagnostic criteria was expanded in the SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference, 2001 [46].

Considering the advances made in the pathobiology and epidemiology of sepsis and considering the insufficient specificity of the SIRS criteria, definitions of sepsis and septic shock were re-defined in the Third International Consensus Definitions for Sepsis and Septic Shock, 2016 (sepsis-3) [61]. According to sepsis-3, sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated host response to an infection [61]. SOFA score is the most popular and commonly used scoring system to assess the severity of organ dysfunction and is based on clinical findings, laboratory data, or therapeutic interventions. According to the sepsis-3 definition, sepsis can be recognized by the presence of an infection and change in the SOFA score of two or more organs [61]. And septic shock is described as a subset of sepsis characterized by particularly profound circulatory, metabolic, and cellular abnormalities that are related to a higher risk of mortality than

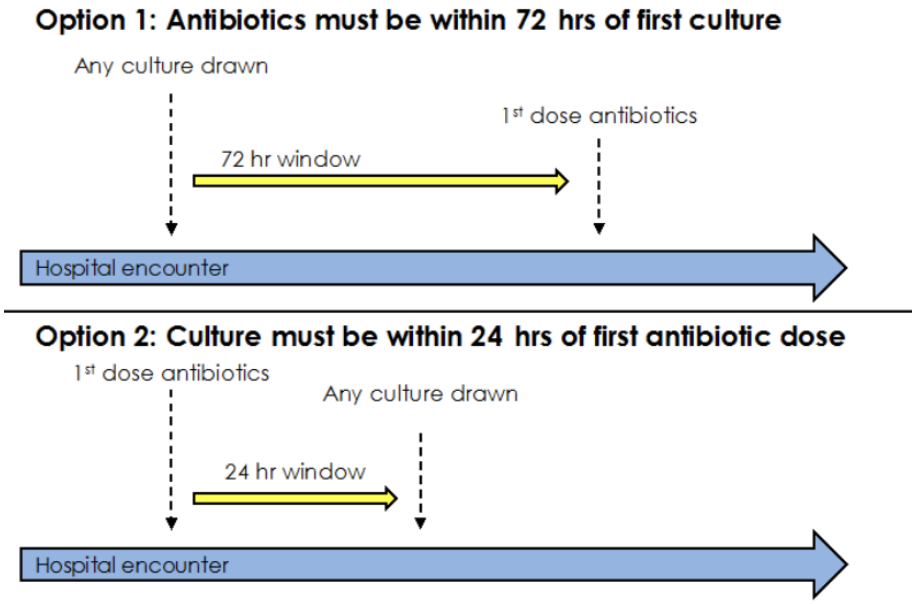


Figure 1: Suspicion of infection [58]

with sepsis alone. It can be identified by a vasopressor requirement to maintain a MAP of 65 mm Hg or greater and serum lactate level greater than 2 mmol/L (> 18 mg/dL) in the absence of hypovolemia. The sepsis-3 definition also introduced a new bedside clinical score termed quickSOFA (qSOFA). Its criteria are the respiratory rate of 22/min or more, altered mentation, systolic blood pressure of 100 mm Hg, or less. qSOFA can be used to instantly identify patients in out-of-hospital, emergency departments, or general hospital ward settings with suspected infection and are more likely to have sepsis if they meet at least 2 of the qSOFA criteria. Generally, qSOFA is not used to diagnose sepsis, but to predict mortality [67]. Figure 2 illustrates the criteria for SOFA scoring systems [61]. According to the sepsis-3 definitions, sepsis involves both infection and organ dysfunction, and this makes the definition of severe sepsis introduced earlier by R. C. Bone et al. [15] and the International Sepsis Definitions Conference (2001) [46] unnecessary. We consider the latest sepsis-3 definitions [61] for this thesis.

4.2 Suspicion of Infection

According to the sepsis-3 [61], sepsis is defined as a life-threatening organ dysfunction caused by the dysregulated host response to infection. This suggests that to diagnose patients that are at risk of sepsis, we must first identify patients with suspected infection.

According to Seymour et al. [58], the first episode of suspected infection can be known as the combination of fluid body cultures such as blood, urine, cerebrospinal fluid, etc. and antibiotics (oral or parenteral). If the patient was given the antibiotic first, then the culture sample must have been collected within 24 hours. If the culture sample was collected first, then the patient must have been given antibiotics within 72 hours. Figure 1 illustrates this process of identifying suspicion of infection. The suspected time of infection ($t_{suspicion}$) is defined as the time at which the first of these two events occurred.

4.3 Organ Dysfunction in Sepsis

According to sepsis-3 [61], sepsis-related organ dysfunction can be diagnosed by an increase in SOFA score of 2 or more consequent to the infection. The SOFA score is used to determine a patient's rate of organ failure in the ICUs, and according to the European Medicines Agency, a change in the SOFA score is an acceptable surrogate marker of effectiveness in exploratory trials of novel therapeutic agents in sepsis. The SOFA score was developed following a consensus meeting in 1994, the goal of which was to create a score to describe quantitatively and as objectively the degree of organ dysfunction/failure over a period of time in groups of patients or even individual patients. The score was developed to describe a series of complications of critical illness and not to predict the outcome. It was initially described as the sepsis-related organ failure assessment. However, the utility of the score for the assessment of acute morbidity in a range of critical illnesses was acknowledged early, resulting in a change of the title [44].

	Score				
System	0	1	2	3	4
Respiration					
Pao_2/Fio_2, mm Hg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, $\times 10^3/\mu L$	≥150	<150	<100	<50	<20
Liver					
Bilirubin, mg/dL (μmol/L)	<1.2 (20)	1.2-1.9 (20-32)	2.0-5.9 (33-101)	6.0-11.9 (102-204)	>12.0 (204)
Cardiovascular	MAP ≥70 mm Hg	MAP <70 mm Hg	Dopamine <5 or dobutamine (any dose)[b]	Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤0.1[b]	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b]
Central nervous system					
Glasgow Coma Scale score[c]	15	13-14	10-12	6-9	<6
Renal					
Creatinine, mg/dL (μmol/L)	<1.2 (110)	1.2-1.9 (110-170)	2.0-3.4 (171-299)	3.5-4.9 (300-440)	>5.0 (440)
Urine output, mL/d				<500	<200

Figure 2: The Sequential Organ Failure Assessment (SOFA) Scoring system [61]

Figure 2 illustrates the criteria for SOFA scoring system. SOFA score is evaluated on a scale of 0-4, where a SOFA score of 0 means there is no organ failure, and an increasing score reflects worsening organ dysfunction. It comprises following six different components:

- Respiratory
- Cardiovascular
- Renal
- Hepatic
- Coagulation
- Neurological system

The subscores for these six components are also evaluated on a scale of 0-4. These subscores represent the most severe or worst values during the observed period. The SOFA score is calculated as the arithmetic sum of each of the subscores.

4.4 Vital Signs and SOFA score

Vital signs play an essential role in determining the status of the body's critical functions and are used as a baseline reference for detecting or monitoring medical problems. These vital signs are commonly available and continuously recorded by the bedside monitors for the patients in ICU. For patients with sepsis, vital signs can help us understand the severity of sepsis and the rate at which the patient's health deteriorates. Also, vital signs and the SOFA score are significantly co-related in the case of sepsis [42]. Hence, we can predict the SOFA scores and prevent sepsis patients from progressing into Septic Shock using the vital signs. For the purpose of this thesis work, we consider the time series data of the following vital signs along with the time series data of SOFA scores:

- Heart Rate (HR)
- Respiratory Rate (RR)
- Oxygen Saturation (SPO2)
- Systolic Arterial Blood Pressure (ABPSYS)

- Diastolic Arterial Blood Pressure (ABPDIAS)
- Mean Arterial Blood Pressure (ABPMEAN)
- Temperature (TEMP)

4.5 Related Work

In the past recent years, some studies have been done for the early prediction of septic shock in sepsis patients using ML approaches.

Stephen J. et al. [48] introduced a new state, i.e., 'pre-shock' state before the septic shock. They hypothesize that sepsis patients that enter this new state are highly likely to develop septic shock. According to them, the physiology of sepsis patients who progress into septic shock changes gradually over time as their condition worsens, and hence, they move into the pre-shock state before transiting to septic shock. Therefore, to predict shock in sepsis patients who will develop shock, they identify the patients who transition into the pre-shock state. To demonstrate this, they used three different ML models, i.e. generalized linear models (GLM), XGBoost, recurrent neural networks (RNN). For this work, they consider the time series data of only two vitals, i.e., heart rate and respiratory rate, components of SOFA score such as cardio SOFA, kidney SOFA, respiration SOFA, coagulation SOFA in addition to the clinical values such as lactate, blood potassium level (PaO2), and fraction of inspired oxygen (FiO2) which are also parts of SOFA score components. They also used this data for different time window lengths, beginning at 12 hours preceding septic shock onset to 3 hours after septic shock onset. Christopher R. et al. [69] used Bayesian networks with laboratory data, International Classification of Diseases (ICD) diagnosis codes, and patient's other demographic data such as age, sex, etc., for 24 hours prior to septic shock onset to predict septic shock in sepsis patients. Josef F. et al. [27] used Long Short-Term Memory (LSTM) networks to study long-term dependencies in time series data and predict septic shock. They analyze clinical data such as creatinine, hemoglobin, potassium, etc., in addition to the time series data of vital signs 48 hours before the septic shock onset time. David et al. [34] developed a real-time warning score, a targeted early warning score (TREWScore), to predict which patients are at risk for septic shock. The TREWScore was calculated by fitting a cox proportional hazards model on features extracted from the physiological and laboratory data before septic shock onset. Nemati et al. [52] developed a Weibull-Cox proportional hazards model to predict sepsis before its onset using physiological and laboratory data. Alejandro B. et al. [53] predicted septic shock at the time of sepsis onset time applying ML on the vital signs time series data. They applied model-based and instance-based ML methods on the data acquired until sepsis onset time only. For model-based methods,

they use ML supervised classification models such as linear regression, support vector machine (SVM), LogitBoost, and random forest and perform for binary classification on the whole sequence of the vital sign time series data. In instance-based methods, they match the entire test sequence to all the training dataset sequences by calculating a pairwise similarity between them using dynamic time warping (DTW) and alignment of textures (ALoT) measures. The output class is then determined by the majority of labels owned by the sequence. Jacob et al. [50] developed an ML algorithm with gradient tree boosting and only the time series data of the vital signs to predict sepsis before the onset time.

The above studies suffer from some major limitations. In addition to the routine vital signs, they also use laboratory data such as platelet count, creatinine, potassium, etc., in their methods. Obtaining such laboratory data can be time consuming and requires to have a clinician in loop, thus making timely prediction of septic shock impossible. The methods proposed by Alejandro B. et al. [53] do not consider the golden hours' vital sign data, i.e., a few hours before the septic shock onset. In contrast, the vital signs of a patient progressing into septic shock recorded a few hours before the shock are highly indicative of the shock. Moreover, by performing classification on entire sequences, they do not consider the changes or patterns in the vitals that would occur over time as their condition worsens. And, Jacob et al. used to old definitions of sepsis for their analysis and work. Moreover, these studies do not consider the SOFA score data which plays an important role in identifying sepsis and it's severity in patients. The study done by Stephen J. et al.[48] overcomes this limitation of not using SOFA score for prediction of septic shock to some extent. They performed prediction based on components of SOFA scores and other clinical data in addition to the vital signs. However, the calculation of these SOFA score components is heavily dependent on the laboratory test results. Hence, such approaches based on laboratory data increase the turnaround time for predicting septic shock as the laboratory test results are not immediately available most of the time. Consequently, this further delays the treatment process and increases the mortality risk.

5 Cohort Selection

This chapter illustrates the data sources used and the different steps taken to build the cohort for this thesis. Section 5.1 describes the data sources used in detail. Section 5.2 explains the process of identifying the initial cohort, i.e. patients with infections. Sections 5.3 and 5.4 illustrate the approach used for identifying patients with sepsis and septic shock in our initial cohort, according to the sepsis-3 definitions. Sections 5.5 and 5.6 deal with checking for the availability of vital signs time series data for the time interval between sepsis onset time and septic shock time for all sepsis patients and discarding patients with missing data to build our final cohort.

5.1 Data Sources

The data from the Medical Information Mart for Intensive Care (MIMIC) III (version 1.4) was used for this thesis. MIMIC-III is an extensive, single-center database which contains deidentified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts [39].

Figure 3: MIMIC-III Database

MIMIC-III contains information for 53,423 unique hospital admissions for patients aged 16 years or above that were admitted to ICUs between 2001 and 2012. In addition to that, it also includes information about 7,870 neonates that were admitted between 2001 and 2008 [39]. The data in the MIMIC-III database is extracted from the following two different sources:

- MetaVision: It is provided by iMDSoft and records the data of patients that are admitted after 2008.

- CareVue: It is provided by Philips and records the data of patients that are admitted between 2001-2008.

Both MetaVision and CareVue are clinical information systems that store and display data at the bedside for patients in the ICU. MIMIC-III database is also widely accessible to researchers universally under a data use agreement and it is provided by PhysioNet [2]. Figure 3 shows that the MIMIC-III database consists of two parts: a clinical database [2] and a waveform database [3]. The MIMIC-III Waveform database comprises a subset, i.e., the MIMIC-III Waveform Database Matched Subset [4]. For this thesis, we used both the MIMIC-III clinical database (version 1.4) [2] and Waveform Database Matched Subset (version 1.0) [4]. We majorly used the MIMIC-III Clinical database for cohort selection and identification of the ground truth and the MIMIC-III Waveform Database Matched Subset for evaluating vital signs time series data and hence for our experiments and analysis.

5.1.1 MIMIC-III Clinical Database

MIMIC-III clinical database is a relational database containing a total of 26 tables with data such as vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more [2]. Table 2 lists some of the most commonly used MIMIC-III clinical database tables used in this thesis with their respective description. Access to the MIMIC-III Clinical database is restricted. It can only be obtained by registering on the Collaborative Institutional Training Initiative (CITI) program, which affords completing an online training course, and then applying for credentialed access via your PhysioNet account [9].

5.1.2 MIMIC-III Waveform Database Matched Subset

The MIMIC-III Waveform Database consists of physiologic signals (waveforms) and time series of vital signs (numerics) data recorded from the bedside monitors in the ICUs. For every ICU stay recording, two records exist, i.e., a waveform record and a numeric record, and are sampled once per second or once per minute [3].

The waveform records consist of one or more digitized signals such as electrocardiogram (ECG), arterial blood pressure (ABP), respiration, and photoplethysmogram (PPG) signal waveforms [3]. Each waveform record contains a *master header file* and a *layout header file*. The waveform records consist of multiple *segments*, each of which can be read as a separate record. Each *segment* consists of an uninterrupted recording of a set of simul-

taneously monitored signals. The *master header file* contains the list of all the *segments* and the *layout header file* for the waveform record. The *layout header file* specifies all of the observed signals in any of the *segments* belonging to the record. Also, each *segment* has its own *header file* and a matching binary signal (*.dat*) file. The monitor may be disconnected occasionally for a short period of time, and such intervals are reported as gaps in the *master header file*.

MIMIC-III Table	**Description of Table**
Admissions	All the details about every unique hospitalization for each patient such as demographic information, timing of admission and discharge, the source of the admission is stored in this table
Icustays	This table holds information about every unique ICU stay such as ICU in time and out time, length of stay, etc.
Patients	Information about each patient such as date of birth and death, whether the patient expired, etc. is saved in this table.
Chartevents	This tables consists of all the electronic chart data recorded during the ICU stay such as patients' routine vital signs, information related to mental status, ventilator settings, etc.
Inputevents_mv	Data related to enteral feeding and intravenous medication intake for patients is monitored using the iMDSoft MetaVision system during their ICU stays is stored in this table.
Outputevents	This table contains output information such as urine output for patients in ICU.
Labevents	This table stores details about laboratory measurements and results of laboratory tests for patients.
Microbiologyevents	Microbiology data such as cultures acquired, culture results, antibiotic sensitivities is stored in this table.
Procedureevents_mv	This table holds information about the procedures for the patients monitored in the ICU using the iMDSoft MetaVision system.
D_items	This table serves as a dimensional table or a dictionary of local codes (*ITEMIDs*) appearing in the MIMIC database, excluding the ones that are related to laboratory tests. e.g. For heart rate, the *ITEMID* is 220045.
D_labitems	This is a dimensional table that stores dictionary of local codes (*ITEMIDs*) appearing in the MIMIC database that are related to laboratory tests. e.g. For a Hemoglobin test, the *ITEMID* is 50811.
D_icd_diagnoses	This is a dimensional table and consists of dictionary of International Classification of Diseases Version 9 (ICD-9) codes for diagnoses that are assigned at the end of the patient's stay and is utilized by the hospital to bill for care provided.
D_icd_procedures	This is a dimensional table and contains a dictionary of International Classification of Diseases Version 9 (ICD-9) codes for the procedures that are assigned at the end of the patient's stay and is used by the hospital to bill for care provided.

Table 2: Commonly used MIMIC-III Clinical Database tables [2][39]

The numeric records usually consist of time series data of heart rates, oxygen saturation (SpO2), respiration rates, and systolic, diastolic, and mean blood pressure. The numeric records are designated by the letter *'n'* appended to the record name, and they are not divided into segments [3].

The MIMIC-III Waveform Database Matched Subset is a subset of the main MIMIC-III Waveform Database and contains 22,317 waveform records and 22,247 numeric records

that have been matched and time-aligned with the 10,282 patients of MIMIC-III clinical database records [4].

Both the MIMIC-III Waveform database and MIMIC-III Waveform Database Matched Subset have open access and can be directly downloaded from the physioNet server without any restriction. Which makes it one of the most widely used medical research databases and allows the required benchmarking of methods.

5.2 Initial Cohort Selection

The sepsis cohort selection was done based on sepsis-3 definition [61].

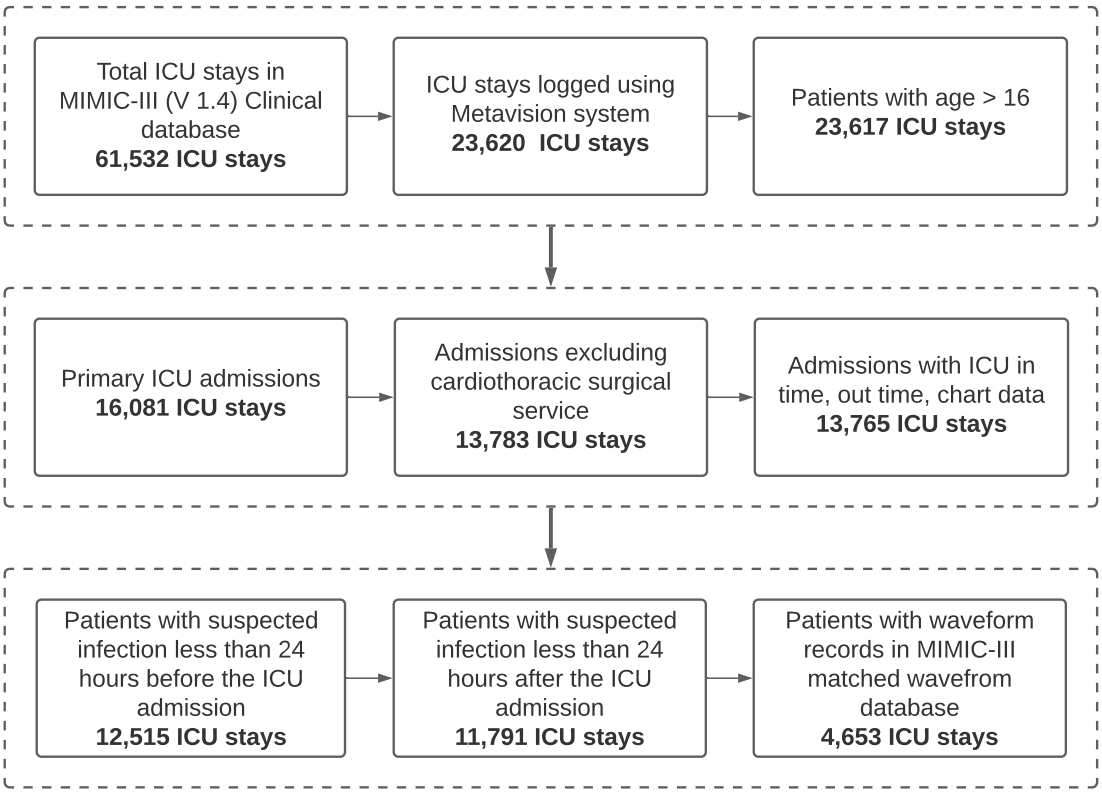


Figure 4: Initial Cohort Selection [38]

Out of the total 61,532 ICU admissions in MIMIC-III clinical database, we consider only the ones that were logged using MetaVision (23,620 ICU admissions) and not CareVue

system (37,912 ICU admissions) because the data logged using the CareVue system is not as detailed and comprehensive as for the data logged using MetaVision. We then apply the above method used by Seymour et al. [58] on 23,620 ICU admissions logged using MetaVision in MIMIC-III clinical database to identify patients with suspected infection and their respective time of suspicion of infection ($t_{suspicion}$). For identifying suspicion of infection, we used the MIMIC-III tables *microbiologyevents* to get information about the body cultures drawn and *inputevents_mv* for information about antibiotic administration. We then followed the same steps as in [38] to built our cohort.

As in [38], we started with analyzing 23,620 ICU admissions from the MIMIC-III clinical database. We applied the following exclusion criteria:

- Excluding non-adult patients (3 admissions).
- Avoiding repeated measures, i.e. excluding secondary (or greater) admissions for patients (7,536 admissions).
- Since the post-operative physiologic derangements of the admissions to the cardiothoracic surgical service does not translate to the same mortality risk as to the other ICU patients, and we excluded them (2,298 admissions).
- Excluding admissions with missing ICU stay data such as ICU in time and out time, routine vital signs, etc. (18 admissions).
- Excluding patients that had suspected infection more than 24 hours before ICU admission (1,250 admissions).
- Excluding patients that had suspected infection more than 24 hours after ICU admission as we chose to concentrate on the patients who are admitted to the ICU with sepsis (824 admissions).

Since we want to work with time series data for this thesis, we need to further verify that out of these 11,791 ICU stays, for how many ICU stays we have the respective time series data in the MIMIC-III matched waveform database. After excluding the ICU stays with no data in MIMIC-III matched waveform database, we get a total of 4,653 ICU stays. We then identify patients with sepsis and septic shock out of these 4,653 ICU stays. Figure 4 gives an overview of the process carried out for selecting the initial cohort.

SOFA Component	Materialized Views	MIMIC-III Tables
Respiration	*ventsettings,ventdurations, blood_gas, bloodgasarterial*	*procedureevents_mv,chartevents, labevents, icustays, admissions, patients*
Coagulation	*labs*	*labevents, icustays, admissions, patients*
Neurological System	*gcs*	*chartevents, icustays, admissions, patients*
Renal	*labs, urine_output*	*labevents, outputevents, icustays, admissions, patients*
Liver	*labs*	*labevents, icustays, admissions, patients*
Cardiovascular	*vitals*	*chartevents, inputevents_mv, icustays, admissions, patients*

Table 3: Materialized views created for calculation of SOFA score

5.3 Identifying Patients with Sepsis

Once we have identified patients with suspicion of infection and their respective time of infection suspicion ($t_{suspicion}$), we take a window of 48 hours before the suspicion time of infection ($t_{suspicion}$ - *48 hours*) and 24 hours after the suspicion time of infection ($t_{suspicion}$ + *24 hours*). For this 72 hours window, we calculate the SOFA score of every hour and compare it with the initial SOFA score at the start of the window. If the increase in SOFA score is greater than or equals to two for a particular hour, we define this hour as sepsis onset time ($t_{sepsis\ onset}$).

Table 3 lists all the materialized views created for the calculation of scores for each of the six components of SOFA score and overall SOFA score. We refer to the GitHub MIMIC Code Repository as our base for creating these materialized views and calculating the SOFA score [1]. This repository is shared by the research community and is intended to be a primary hub for sharing, refining, and reusing code used for the analysis of the MIMIC critical care database. It contains a number of scripts, views to build the MIMIC-III Clinical database, to generate different organ failure scores and concepts related to it, and many other useful queries.

5.4 Identifying Patients with Septic shock

According to sepsis-3, patients with septic shock can be identified by sepsis persisting hypotension that requires vasopressors to keep the MAP >= 65 mm Hg and having a serum lactate level > 2 mmol/L (18 mg/dL) despite adequate fluid resuscitation [61]. The Surviving Sepsis Campaign guidelines for treatment [22] recommends that a patient

Shock Component	**Materialized Views**	**MIMIC-III Tables**
Weight	*mv_septic_shock_weight*	*chartevents, icustays, admissions, patients*
Fluids/crystalloids	*mv_septic_shock_fluids*	*inputevents_mv, icustays, admissions, patients*
Urine Output	*mv_septic_shock_urineoutput*	*outputevents, icustays, admissions, patients*
CVP	*mv_septic_shock_cvp*	*chartevents, icustays, admissions, patients*
MAP	*mv_septic_shock_map*	*chartevents, icustays, admissions, patients*
Lactate level	*mv_septic_shock_lactate*	*chartevents, icustays, admissions, patients*

Table 4: Materialized views created for identification of septic shock patients

is considered adequately fluid resuscitated if they have been administered at least 30 mL/kg of fluids/crystalloids, or if the urine output $>$ 0.5 mL/kg/hr or Central Venous Pressure (CVP) 8-12 mmHg have been met [48].

In order to identify septic shock patients in our cohort, we consider a window of one hour after the sepsis onset time ($t_{sepsis\ onset}$) and evaluate the following for each hour window.

- The total amount of fluids/crystalloids per kg is administered to a patient from the sepsis onset time until the current hour.
- Urine output per kg recorded for the current hour.
- CVP recorded for the current hour.

If a patient in a particular hour, satisfies any of the above-mentioned criteria for adequate fluid resuscitation, we consider that hour as the adequate fluid resuscitation time ($t_{adequate\ fluid\ resuscitation}$). After knowing the $t_{adequate\ fluid\ resuscitation}$, using the same hourly sliding window, we check for the value of MAP and serum lactate levels for every hour. If MAP $>=$ 65 mm Hg and serum lactate level $>$ 2 mmol/L, then the first such hour is considered as septic shock onset time ($t_{shock\ onset}$). Table 4 lists the materialized views created for the process of identifying patients with septic shock and calculation of the respective $t_{shock\ onset}$.

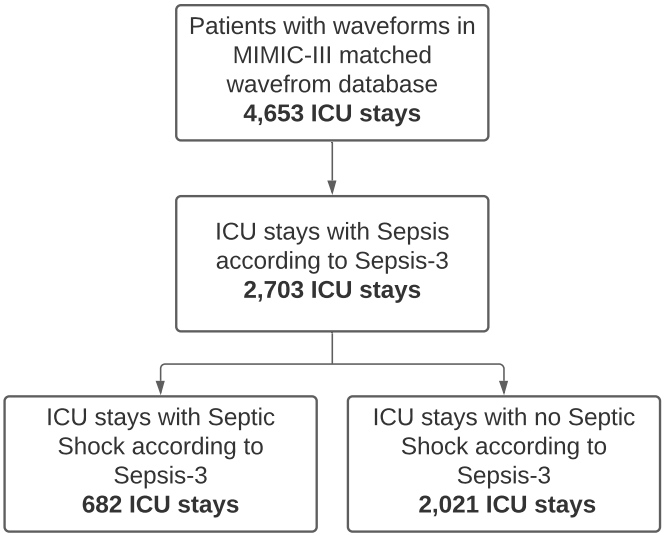


Figure 5: Number of ICU stays with sepsis and septic shock in initial cohort

Figure 5 shows the number of ICU stays with sepsis and septic shock out of the total 4,653 ICU stays.

5.5 Identifying Patients with Vital Signs Data

Since we want to focus on the time series data for this thesis, we consider the numeric records in the MIMIC-III Waveform Database Matched Subset to extract the time series of the vital signs mentioned in section 4.4.

Vital Sign	**MIMIC-III Waveform Database Matched Subset Representations**
Heart Rate	'HR'
Respiratory Rate	'RESP'
Oxygen Saturation	'%SpO2', 'SpO2'
Systolic Arterial Blood Pressure	'ABP SYS', 'ABP Sys'
Diastolic Arterial Blood Pressure	'ABP DIAS', 'ABP Dias'
Mean Arterial Blood Pressure	'ABP MEAN', 'ABP Mean'
Temperature	'TEMP'

Table 5: Vital signs representations in MIMIC-III Waveform Database Matched Subset

We first analyze all the numeric records to know the different representations of the vital signs recorded in the MIMIC-III Waveform Database Matched Subset. The table 5 shows the different representations of the the vital signs mentioned in section 4.4.

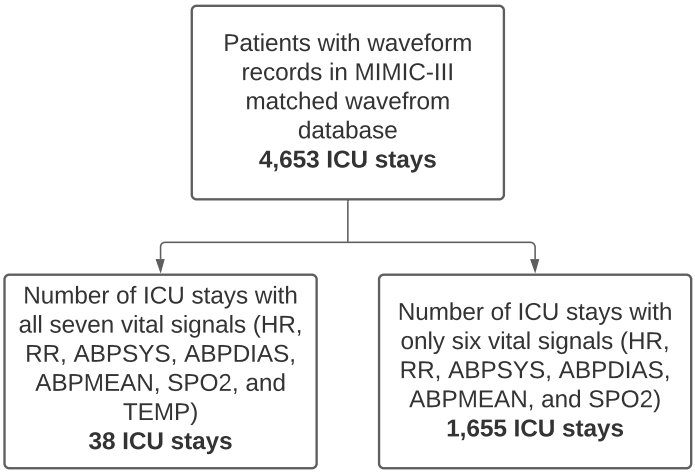


Figure 6: Number of ICU stays with and without Temperature data in MIMIC-III Waveform Database Matched Subset

We then check the number of ICU stays having data for all the seven vital signs, i.e., HR, RR, ABPSYS, ABPDIAS, ABPMEAN, SPO2, and TEMP in their respective numeric records. Figure 6 shows that out of 4,653 ICU stays, the number of ICU stays with data for all seven vital signs was 38 while the number of ICU stays with only six vitals (except TEMP) was 1,655. Since the number of ICU stays with data for all seven vital signs recorded in the matched waveform database is extremely low, we decided to consider and extract time series data of only six vital signs (HR, RR, ABPSYS, ABPDIAS, ABPMEAN, and SPO2) from the MIMIC-III Waveform Database Matched Subset [4] and extract TEMP data from the MIMIC-III clinical database [2] to construct the time series of TEMP.

5.6 Final Cohort Selection

As the primary goal of this thesis is to find the subgroup of sepsis patients that progress into septic shock, we consider the 2,703 ICU stays with sepsis, as shown in the figure 5 and look for their respective numeric records in MIMIC-III Waveform Database Matched Subset having data for all six vital signs (HR, RR, ABPSYS, ABPDIAS, ABPMEAN,

and SPO2) recorded between sepsis onset time ($t_{sepsis\ onset}$) and septic shock onset time ($t_{shock\ onset}$). Sepsis patients that did not progress into septic shock clearly will not have the shock onset time. Hence, we calculate the average number of hours between sepsis onset time and shock onset time ($AVG(\ t_{shock\ onset}$ - $t_{sepsis\ onset})$) for all sepsis patients that progressed into shock. We then consider the period between sepsis onset time and sepsis onset time + ($AVG(\ t_{shock\ onset}$ - $t_{sepsis\ onset})$), i.e., between $t_{sepsis\ onset}$ and $t_{sepsis\ onset}$ + 31 hours for the non-shock sepsis patients. Hence, for sepsis patients that developed septic shock and that did not develop septic shock, we consider the following time periods for extracting time series data:

- Sepsis patients with septic shock: Time series data between sepsis onset time ($t_{sepsis\ onset}$) and septic shock onset time ($t_{shock\ onset}$)
- Sepsis patients with no septic shock: Time series data between sepsis onset time ($t_{sepsis\ onset}$) and sepsis onset time plus 31 hours ($t_{sepsis\ onset}$ + 31 hrs)

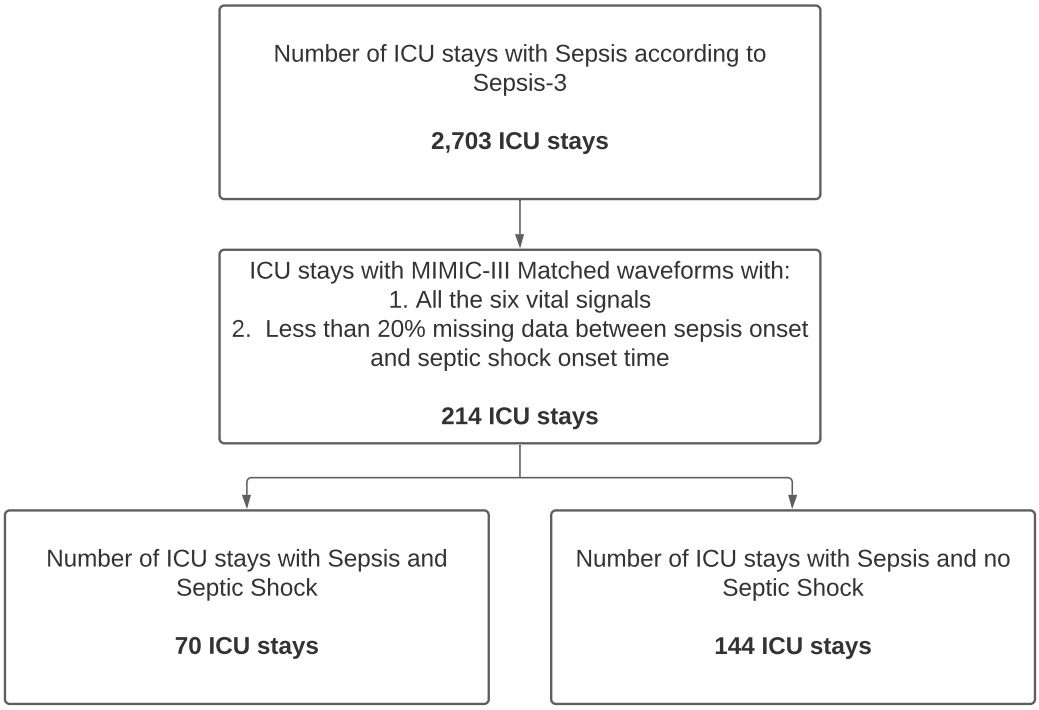


Figure 7: Final Cohort

We discard the ICU stays for which there exists no numeric record with all the six vital signals. In addition to that, we also drop the ICU stays that contain more than 20%

missing values. Figure 7 shows the number of ICU stays in our final cohort after applying all of the above filter criteria. We have a total of 214 sepsis patients in our final cohort that have a matching numeric record in the MIMIC-III Matched waveform subset with all six vital signs data, i.e. HR, RESP, SPO2, ABPSYS, ABPDIAS, and ABPMEAN and less than 20% missing data between $t_{sepsis\ onset}$ and $t_{shock\ onset}$ or between $t_{sepsis\ onset}$ and $t_{sepsis\ onset}$ + 31 hrs. Out of these 214 sepsis patients, 70 patients developed septic shock and 144 patients did not develop septic shock during their ICU stays.

6 Data Extraction and Pre-processing

In chapter 5, we performed cohort selection to select and identify sepsis patients that progressed into shock and that did not progress into shock. In this chapter, we explain the process followed for extracting and pre-processing the time series data of the vital signs mentioned in section 4.4 and for extracting features from them. In addition to vital signs, we also extract and pre-process the time series data of SOFA score. As stated and explained in section 5.5, we extracted the time series data for six vital signs, i.e., HR, RR, ABPSYS, ABPDIAS, ABPMEAN, and SPO2 from the MIMIC-III Waveform Database Matched Subset and time series data for TEMP from the MIMIC-III Clinical Database.

6.1 Extracting Data from MIMIC-III Waveform Database Matched Subset

For each sepsis patient in our cohort, we read their respective numeric records in the MIMIC-III Waveform Database Matched Subset using the *Python WFDB (WaveForm DataBase) Software Package version 3.1.1* .

The *WFDB (WaveForm DataBase) Software Package* is a General Public Licensed (GPL) successor to the Massachusetts Institute of Technology (MIT) DB Software Package. It consists of a *WFDB library* of tools that can be used for processing, automated analysis, viewing, annotating, and interactive analysis of the waveform signals from the web server directly without having to download the record first. It consists of many subpackages such as *rdrecord*, *rdheader*, *rdsamp*, *wrsamp*, *Record*, *MultiRecord*, *rdann*, *Annotation*. We used the *rdsamp* subpackage for the purpose of this thesis.

The *wfdb.rdsamp* package reads a WFDB record and returns the pysical signal as a 2D numpy array and the metadata with important descriptor fields as a a dictionary. The syntax for it is as follows:

$$signals, fields = wfdb.rdsamp(record\_name, sampfrom = 0, sampto = None, channels = [], \\ pn\_dir = None, channel\_names = [], warn\_empty = False, return\_res = 64)$$

Below is the description of all the *wfdb.rdsamp* input parameters:

- *record_name* : The name of the WFDB record to read.

Input Parameter	**Data Type**	**Default Value**	**Mandatory**
record_name	String	-	Y
sampfrom	Integer	0	N
sampto	Integer	length of entire signal	N
channels	List	All channels	N
pn_dir	String	None	N
channel_names	List	All channels	N
warn_empty	boolean	False	N
return_res	Integer	64	N

Table 6: Input Patameters for the *wfdb.rdsamp* package

- *sampfrom* : The starting sample number from which to start reading for all the requested channels.
- *sampto* : The starting sample number from which to start reading for all the requested channels.
- *channels* : List of integer indices specifying the channels to be read.
- *pn_dir* : This is used to stream data from Physionet. The Physionet database directory from which you want to read the WFDB record.
- *channel_names* : List of channel names to return from the WFDB record.
- *warn_empty* : Whether to display a warning if the specified channel indices or names do not exist in the WFDB record, and no signal is returned.
- *return_res* : Data type of the numpy array of the returned signals.

Table 6 lists the data type, default values of the input parameters and whether they are mandatory or optional.

Output of the *wfdb.rdsamp* consists of two parts: *signals* and *fields*.

- *signals* : A 2D numpy array containing the physical signals from the WFDB read record .
- *fields* : A dictionary containing the following key attributes describing the read record.
  - *fs*: The sampling frequency of the record. Value for this key is 0.0167 for sampling frequency of once per minute and 1 for sampling frequency of once per second.

- *sig_name*: The signal name for each of the extracted channel. For example, *'HR'*, *'ABP SYS'*, *'ABP DIAS'*, etc.
- *units*: The units for each of the extracted channel. For example, *'bpm'*, *'mmHg'*.
- *sig_len* : Length of the signal.
- *base_date* : Start date of the signal.
- *base_time*: Start time of the signal.
- *comments*: Any comments written in the header.

6.1.1 Example of wfdb.rdsamp package

Listing 1 is an example of a Python script for reading the numeric record for the patient ID '42930' from the MIMIC-III Waveform Database Matched Subset using the Python's *wfdb.rdsamp* Package. In the example below, we read the WFDB numeric record *p042930-2190-07-28-20-30n* directly from the Physionet directory *mimic3wdb/matched/p04/p042930/* without downloading it. Also, for the purpose of this example, we only extract specific channels, i.e. 'HR', 'ABP MEAN', 'ABP SYS', AND 'ABP DIAS' for 20 samples from 100 to 120.

```
!pip install wfdb
import pandas as pd
from IPython.display import display
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
import wfdb

signals,fields = wfdb.rdsamp( 'p042930-2190-07-28-20-30n',
             pn_dir='mimic3wdb/matched/p04/p042930/',
             channel_names=['HR','ABP MEAN', 'ABP SYS','ABP DIAS'],
             sampfrom=100,
             sampto=120 )

wfdb.plot_items(signal=signals, fs=fields['fs'])

print('Printing signals')
display(signals)
print('Printing fields')
display(fields)
```

Listing 1: A Python Program to extract time series data using *wfdb.rdsamp* Package

Listing 2 prints the *signals* 2d Numpy array and *fields* dictionary as an output of the Python program in listing 1. The *signals* array contains time series data of 'HR', 'ABP MEAN', 'ABP SYS', AND 'ABP DIAS' extracted from the numeric record *p042930-2190-07-28-20-30n*. As we see the values in *fields* dictionary, the extracted signal / record is sampled once per minute starting from *28.07.2190 22:10:08* and the signal length is *20*. It also gives us the unit of measurement for each of the extracted vital signs.

```
Printing signals:

array([[85.91608429, 68.08208835, 88.88301503, 57.20024612],
       [85.2326673 , 69.88438936, 91.40084976, 58.51664852],
       [84.7331388 , 70.43378967, 92.28381646, 58.99971279],
       [84.44990825, 71.68334038, 94.14977343, 59.78390802],
       [85.11641596, 72.6179909 , 95.23451182, 60.91628595],
       [84.93323202, 73.33489131, 96.10023307, 61.4829977 ],
       [85.21646257, 73.26789127, 96.05022129, 61.36693681],
       [85.2326673 , 72.86589104, 95.61736066, 60.99993344],
       [85.38273722, 72.0819906 , 94.18426432, 60.34957419],
       [85.3496232 , 73.59954146, 96.15024486, 61.68375168],
       [85.73290036, 73.23439125, 95.63288156, 61.26655982],
       [85.54971642, 70.98318998, 93.0012269 , 59.46709315],
       [86.16620082, 67.68343812, 88.85024869, 56.60007537],
       [85.76601438, 64.9833366 , 85.15110109, 54.45033482],
       [86.1330868 , 63.68353587, 82.93333707, 53.55007869],
       [86.24933815, 64.9833366 , 84.46645699, 54.70023171],
       [85.71669562, 64.88283655, 84.50094788, 54.70023171],
       [86.11617752, 65.8677371 , 85.83402065, 55.44992235],
       [85.54971642, 66.5846375 , 86.74975369, 55.94971612],
       [85.08330194, 68.14908839, 88.71745877, 57.34976601]])



Printing fields:

{'base_date': datetime.date(2190, 7, 28),
 'base_time': datetime.time(22, 10, 8),
 'comments': ['Location: tsicu'],
 'fs': 0.0166666666667,
 'n_sig': 4,
 'sig_len': 20,
 'sig_name': ['HR', 'ABP MEAN', 'ABP SYS', 'ABP DIAS'],
 'units': ['bpm', 'mmHg', 'mmHg', 'mmHg']}
```

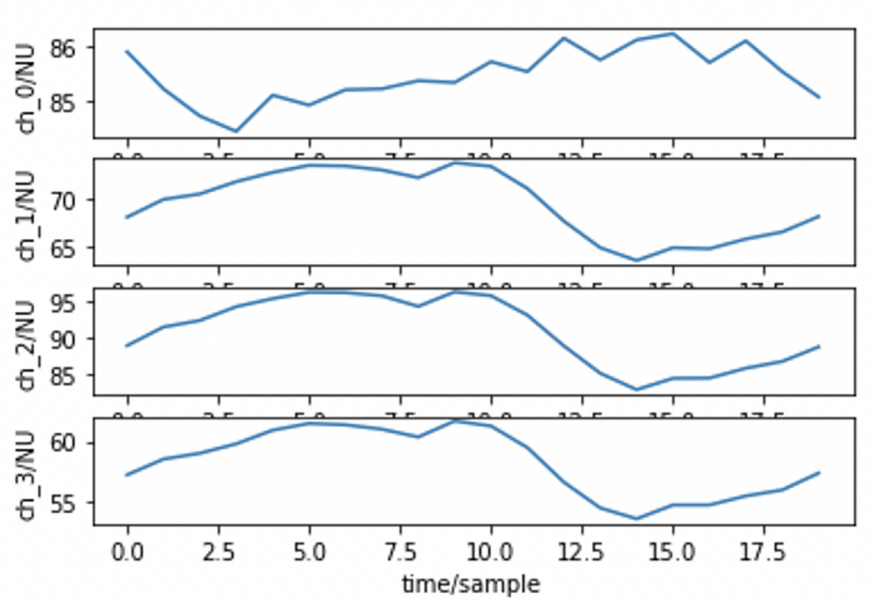


Figure 8: Output waveforms of the Python program in listing 1

Listing 2: Output *signals* and *fields* of the Python program in listing 1

Figure 8 displays the output waveforms of the Python program in listing 1. These waveforms are nothing but graphical representation of the output *signals* and *fields*. As shown in the figure, it displays four waveforms, one for each of the extracted vital signs. Since our *channel_names* list had values ['HR','ABP MEAN', 'ABP SYS','ABP DIAS'], it displays the waveforms in the same order and the labels on the Y-axis are representations of the extracted vital signs. label *ch_0/NU* represents 'HR', *ch_1/NU* represents 'ABP MEAN', *ch_2/NU* represents 'ABP SYS', and *ch_3/NU* represents 'ABP DIAS'. Since the sampling frequency for this record is once per minute, X-axis displays the time elapsed in minutes since the start time i.e. *28.07.2190 22:10:08*

Figure 9 displays the time series data of all the six vital signs, i.e., HR, RR, ABPSYS, ABPDIAS, ABPMEAN, and SPO2 for a sample patient extracted from the MIMIC-III Waveform Database Matched Subset.

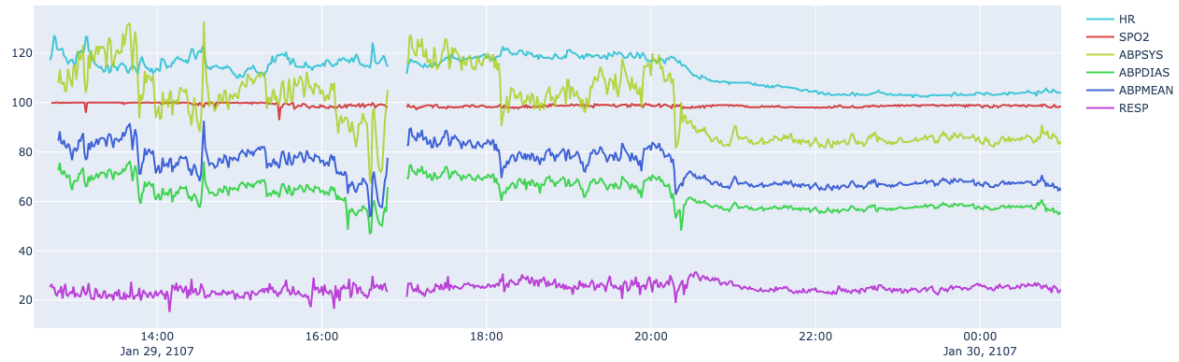


Figure 9: time series data extracted from MIMIC-III Waveform Database Matched Subset

6.2 Extracting Data from MIMIC-III Clinical Database

As mentioned in section 5.5, for each sepsis patient in our cohort, we extract the temperature (TEMP) time series data by querying the *chartevents* and *d_items* tables in MIMIC-III Clinical Database. For any given patient, the temperature data is recorded in the *chartevents* table 3-4 times a day. The TEMP data is measured in two units, i.e. either in *Fahrenheit* or in *Celsius*. For uniformity, we convert all the TEMP data measured in *Fahrenheit* to *Celsius*. Listing 3 shows the SQL query used to extract and transform the TEMP data.

```
select ch.charttime as recorded_time,
       (case
         when di.label = 'Temperature Fahrenheit' then
           (ch.valuenum - 32)/1.8
       else
          ch.valuenum
       end ) as recorded_temp_value
from chartevents ch, d_items di
where ch.itemid = di.itemid
and di.label in ('Temperature Fahrenheit', 'Temperature Celsius')
and ch.icustay_id = :icutay_id
order by ch.charttime;
```

Listing 3: Query to extract temperature (TEMP) time series data

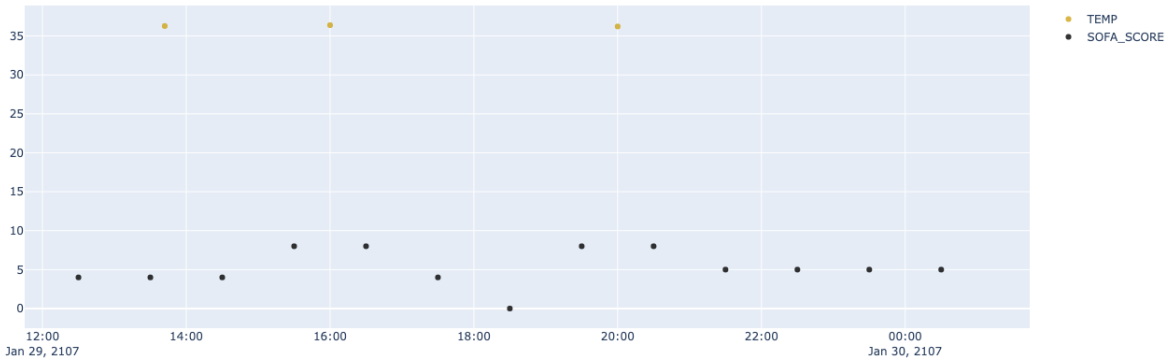


Figure 10: time series data extracted from MIMIC-III Clinical Database

As mentioned in 4.4, vital signs and SOFA scores are correlated and indicate the severity of sepsis. SOFA score is used to determine a patient's rate of organ failure in the ICUs, and according to the European Medicines Agency, a change in the SOFA score is an acceptable surrogate marker of effectiveness in exploratory trials of novel therapeutic agents in sepsis. The SOFA score was developed following a consensus meeting in 1994, the goal of which was to create a score to describe quantitively and as objectively the degree of organ dysfunction/failure over a period of time in groups of patients or even individual patients. The score was developed to describe a series of complications of critical illness and not to predict the outcome [44]. Hence, in addition to the time series data of vital signs, we also calculate and generate hourly time series of SOFA scores using the same Materialized views mentioned in table 3. Figure 10 displays the time series data of TEMP and the SOFA_SCORE for a sample patient extracted from the MIMIC-III Clinical Database.

6.3 Data Pre-processing

In this stage, we prepare the extracted time series data for feature extraction and implementation. This stage comprises two steps, i.e., transformation (aggregation and merge) and missing value imputation.

- Transformation: As mentioned above in section 5.1.2. Numeric records in MIMIC-III Waveform Database Matched Subset are either sampled once per second or once per minute. Hence, the time series data extracted for the six vital signs, i.e., HR, RR, ABPSYS, ABPDIAS, ABPMEAN, and SPO2 have a sampling frequency of

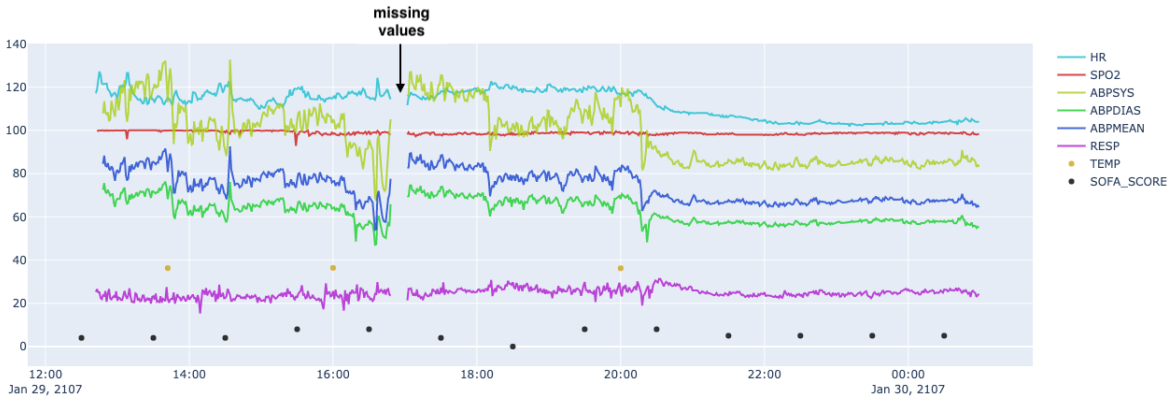


Figure 11: time series data of a sample patient before missing value imputation

once per second and were transformed into once per minute for uniformity. This is achieved by calculating the average values of these vital signs every 60 seconds.

In this step, we also merge the time series data extracted from different sources, i.e., time series data of six vital signs (HR, RR, ABPSYS, ABPDIAS, ABPMEAN, and SPO2) extracted from the MIMIC-III Waveform Database Matched Subset, time series data of TEMP and SOFA score extracted from the MIMIC-III Clinical Database.

- Missing value imputation: As mentioned in the 5.1.2, the bedside monitor used for recording the vital signs might be occasionally disconnected for some time. As a result, the numeric records consist of gaps having null values for the disconnected time intervals. In order to fill these gaps, we impute the missing values during these intervals using the Last Observation Carry Forward (LOCF) or forward-fill method [53] [27] [23]. According to LOCF, the missing values for a patient are replaced by or imputed with the patient's previously observed latest value, i.e., the last observation is carry forward. If the missing values precede the start of observed values, the first observed value is propagated backward to impute the missing values. This process of replacing missing values with the next observed value is known as backward-filling. We used Python *pandas.DataFrame.ffill* for LOCF or forward-fill and *pandas.DataFrame.bfill* for backward-fill.

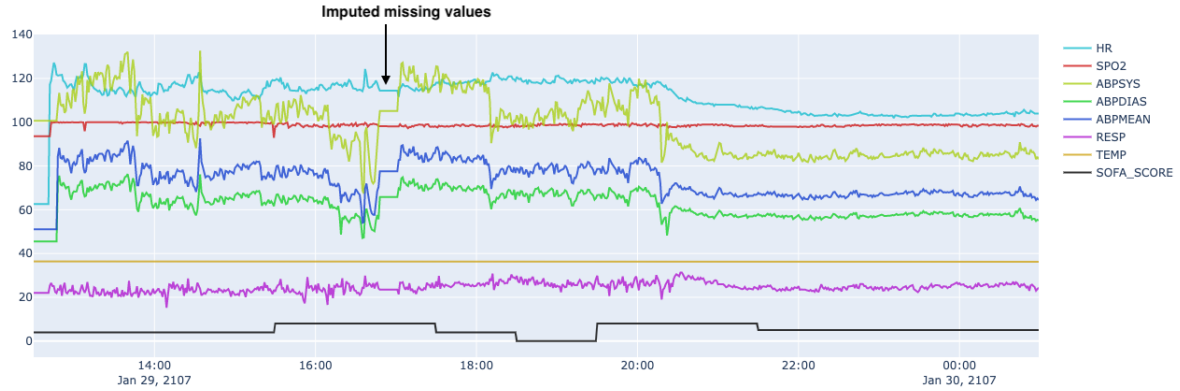


Figure 12: time series data of a sample patient after missing value imputation

Figures 11 and 12 show the time series data of a sample patient before and after missing value imputations respectively. As we see in figure 11, there exists missing values or gaps for the time interval between 16:00 and 18:00 before the imputation. Whereas, in figure 12, we can see that these gaps or missing values are imputed using the LOCF method.

6.4 Feature Extraction

This stage consists of obtaining statistical features from the pre-processed time series data of vital signs and SOFA score. For extracting features, we considered the pre-processed time series data between sepsis onset time and shock onset time for sepsis patients that progressed into shock, and sepsis onset time and shock onset time + 31 h for sepsis patients that did not progress into shock.

6.4.1 Correlation Coefficient

Correlation between two variables can be determined using the Pearson correlation coefficient (r). It is a statistical measure of the strength of the linear relationship between two variables. The Pearson correlation coefficient ($\rho_{x,y}$) for a given paired data (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) of two variables x and y is calculated using the formula 1 , where n is the sample size, x_i and y_i are the individual data points, and $\bar{x}$ and $\bar{y}$ are the sample means

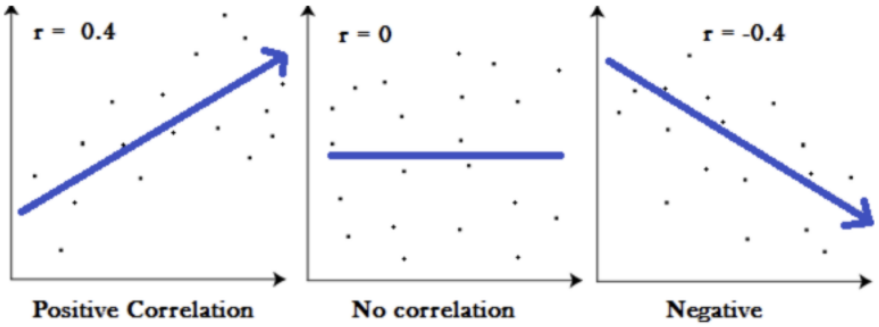


Figure 13: Correlation Coefficient [11]

for $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ respectively . It's values range between -1.0 and 1.0. A value of 1 indicates a strong positive correlation, value of 0 indicates no correlation and value of -1 indicates a strong negative correlation between the two variables. The closer the correlation coefficient to 1.0, the stronger the correlation between the two variables [40]. As seen in the figure 13 , the value of correlation coefficient greater than 0 indicates a positive relationship, equals 0 indicates no relationship, and less than 0 indicates an inverse relationship between the two variables.

$$\rho_{\mathrm{x,y}} = \frac{\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})(y_{\mathrm{i}} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_{\mathrm{i}} - \bar{y})^2}} \tag{1}$$

We analysed the correlation between all the extracted vital signs and SOFA score time series data using the Pearson correlation coefficient. We used Python's *pandas.DataFrame.corr* function [7]. It returns a correlation matrix by computing pairwise correlation of data in columns of a dataframe. Using *pandas.DataFrame.corr*, the correlation matrix can be computed using different correlation methods such as *'pearson', 'kendall', 'spearman'*. Figure 14 shows the correlation matrix for our extracted time series data of vital signs and SOFA score. The plot reveals that ABPDIAS - HR, RESP - HR, ABPDIAS - ABPSYS, ABPMEAN - ABPSYS, and ABPMEAN - ABPDIAS time series are positively and highly correlated. Hence, we calculate the value of correlation coefficient between these vital signs for every hour and use them as hourly correlation features. These features are labeled as *'ABPDIAS_HR_CORR', 'RESP_HR_CORR', 'ABPDIAS_ABPSYS_CORR', 'ABPMEAN_ABPSYS_CORR', 'ABPMEAN_ABPDIAS_CORR'*.

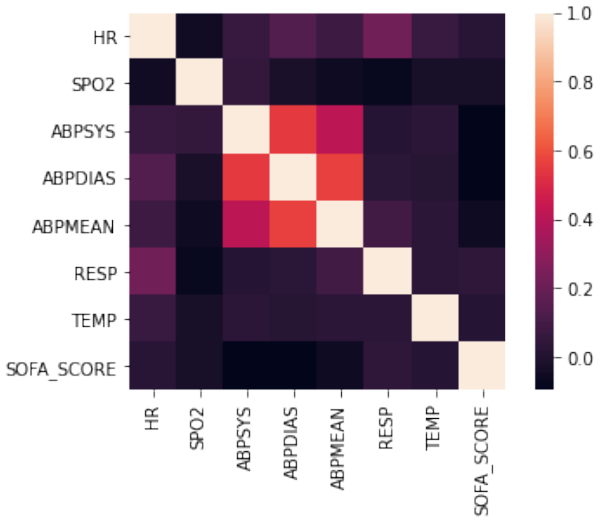


Figure 14: Vital signs and SOFA score correlation matrix

6.4.2 Entropy

According to Z. S. Spakovszky [64], entropy is a measure of the randomness of a system. When applied in ML, entropy is a measure of uncertainty in the data being processed. Approximate entropy (ApEn) and sample entropy (SampEn) are the two types of entropy used most commonly for analyzing the irregularity or unpredictability in time series data [51].

For a given time series of length N having subseries length m, and similarity tolerance of r, ApEn(m, r, N) is expressed as a negative average natural logarithm of the conditional probability that two similar subseries of length m remain similar for subseries of length (m + 1) with a tolerance no more than +/- r times the standard deviation of the time series. Most importantly, it considers each subseries as a match to itself, and hence, it develops a bias towards regularity. This shortcoming of ApEn can be avoided by using SampEn. The SampEn is similar to ApEn, except for the fact that it does not consider the self-matching of subseries. For a given time series of length N, SampEn(m, r, N) is calculated as a negative average natural logarithm of the conditional probability that two similar subseries of length m remain similar for subseries of length (m + 1) with tolerance no more than +/- r times the standard deviation of the time series where the subseries self-matches are not included. Additionally, unlike ApEn, SampEn is independent of data length [51] [55]. Hence, SampEn is more robust and consistent when compared with ApEn.

Feature Name	**Feature Representation**
Correlation coefficient	*'ABPDIAS_HR_CORR', 'RESP_HR_CORR', 'ABPDIAS_ABPSYS_CORR', 'ABPMEAN_ABPSYS_CORR', 'ABPMEAN_ABPDIAS_CORR'*
Sample entropy	*'HR_ENT', 'RESP_ENT', 'ABPSYS_ENT', 'ABPDIAS_ENT', 'ABPMEAN_ENT', 'SPO2_ENT', 'TEMP_ENT'*
Mean values	*'HR', 'RESP', 'ABPSYS', 'ABPDIAS', 'ABPMEAN', 'SPO2', 'TEMP', 'SOFA_SCORE'*
Minimum values	*'HR_MIN', 'RESP_MIN', 'ABPSYS_MIN', 'ABPDIAS_MIN', 'ABPMEAN_MIN', 'SPO2_MIN', 'TEMP_MIN'*
Maximum values	*'HR_MAX', 'RESP_MAX', 'ABPSYS_MAX', 'ABPDIAS_MAX', 'ABPMEAN_MAX', 'SPO2_MAX', 'TEMP_MAX'*
Standard deviation	*'HR_STD', 'RESP_STD', 'ABPSYS_STD', 'ABPDIAS_STD', 'ABPMEAN_STD', 'SPO2_STD', 'TEMP_STD'*

Table 7: List of extracted Features

Supreeth P. Shashikumar et al. [60] and Shamim Nemati et al. [59] demonstrated that entropy in physiological time series data such as time series of heart rate and blood pressure improves the prediction of sepsis and it's severity in adult patients. We calculated SampEn as a feature for all of the extracted vital signs' time series data using the Python's scipy.stats.entropy function [8]. These features are labeled as *'HR_ENT', 'RESP_ENT', 'ABPSYS_ENT', 'ABPDIAS_ENT', 'ABPMEAN_ENT', 'SPO2_ENT', 'TEMP_ENT'*.

6.4.3 Other Statistical Features

In addition to analyzing and calculating correlation coefficients and sample entropy, We aggregate the pre-processed time per minute, time series data of vital signs, and SOFA score into hourly data by averaging their respective values. These hourly mean values are labled as *'HR', 'RESP', 'ABPSYS', 'ABPDIAS', 'ABPMEAN', 'SPO2', 'TEMP', 'SOFA_SCORE'*. We also calculated the hourly minimum, maximum, and standard deviation values for the vital signs' time series data. Table 7 shows the complete list of extracted features and their respective representations. *'HR_MIN', 'RESP_MIN', 'ABPSYS_MIN', 'ABPDIAS_MIN', 'ABPMEAN_MIN', 'SPO2_MIN', 'TEMP_MIN'* represent hourly minimum values, *'HR_MAX', 'RESP_MAX', 'ABPSYS_MAX', 'ABPDIAS_MAX', 'ABPMEAN_MAX', 'SPO2_MAX', 'TEMP_MAX'* express the hourly maximum values, and the hourly standard deviations are presented as *'HR_STD', 'RESP_STD', 'ABPSYS_STD', 'ABPDIAS_STD', 'ABPMEAN_STD', 'SPO2_STD', 'TEMP_STD'*.

7 Time Series Analytic Methods

As we saw in the previous chapter, we extracted hourly features from the time series data of vital signs and SOFA score in data extraction and pre-processing step. This chapter explains the ML methods used to achieve the goal of this thesis, i.e. to predict whether a sepsis patient will develop septic shock or not. This chapter also consists of detailed description of the experiments followed with their respective evaluation methods and results.

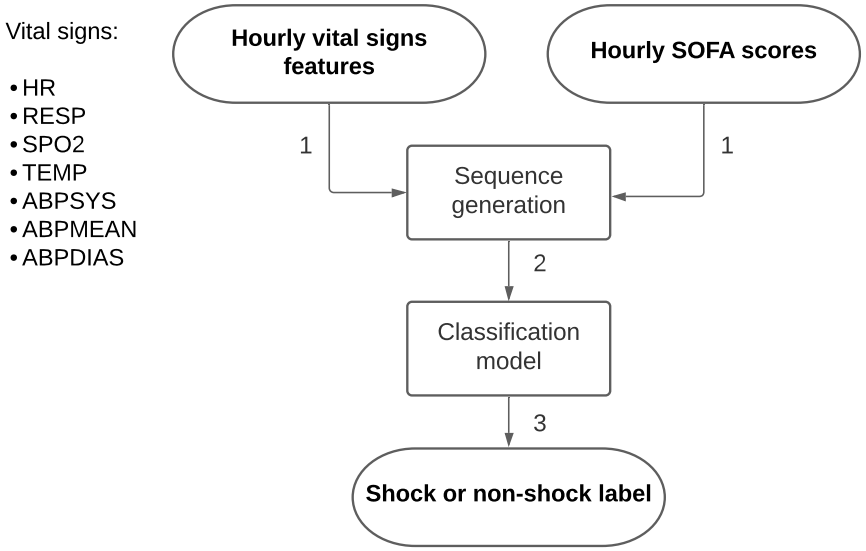


Figure 15: Prediction of septic shock using vital signs and SOFA score

Joseph L. et al. [48] used vital signs, components of SOFA score such as cardiovascular SOFA, respiratory SOFA, coagulation SOFA, renal SOFA scores and laboratory data such as blood potassium levels, lactate, and fraction of inspired oxygen values for prediction of septic shock in sepsis patients. Figure 15 shows the normal stand alone classification approach to predict septic shock. As illustrated in the figure 15, we can use the hourly extracted features of the vital signs along with the hourly SOFA scores for sequence generation and classification. However, obtaining the SOFA score components and it's associated laboratory values increases the turnaround time and results in delayed prediction or diagnosis of septic shock and it's treatment, thus, increasing the mortality risk. Hence, it is more feasible to use only the commonly available vital signs for prediction of septic shock.

In this thesis, we followed a tandem approach of combining multilinear regression with sequence classification using only the vital signs, as shown in the figure 16. We considered

the time series data between sepsis onset time and septic shock onset time for all shock patients, time series data between sepsis onset time and sepsis onset time + 31 hours for all non-shock patients. Figure 16 depicts the followed tandem approach and the central idea of the experiments performed in the subsequent sections. First, a multilinear regression was performed to predict hourly (mean) SOFA scores from the hourly features (mean, minimum, maximum, standard deviation, correlation, sample entropy) of vital signs extracted in section 6.4. Further, these predicted hourly SOFA scores, along with hourly features of vital signs, were used for the sequence generation and classification into *'shock'* or *'non-shock'* class.

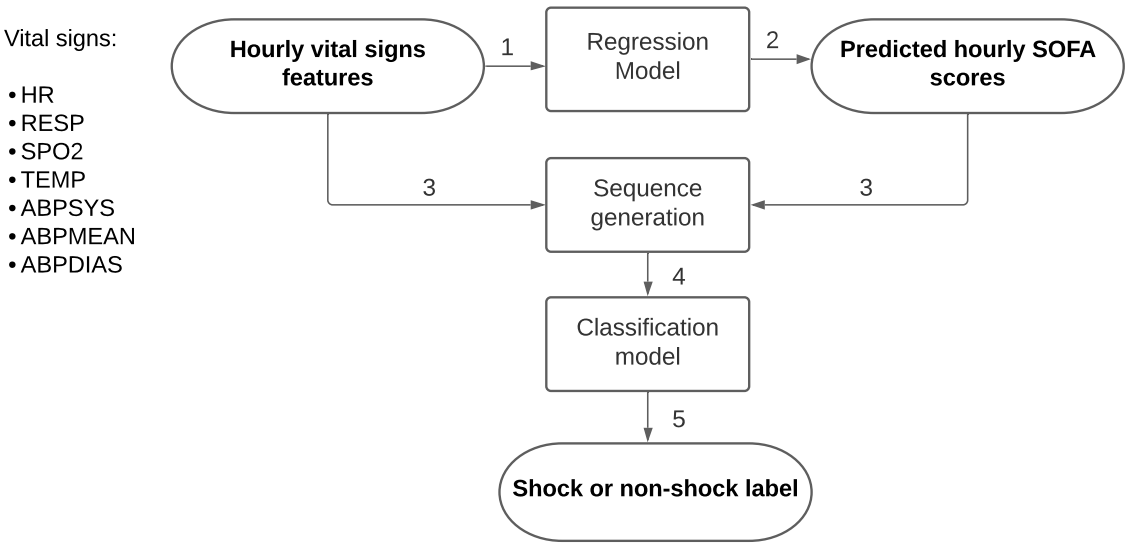


Figure 16: Tandem method for prediction of septic shock using only vital signs

In this chapter we also describe different experiments performed for regression and classification. To perform and evaluate our experiments, we apply 5-fold random cross validation method for 214 sepsis patients in our final cohort (as stated in section 5.6). Out of these 214 sepsis patients, 70 patients developed septic shock and remaining 144 patients did not develop septic shock. We repeat the experiments 5 times, each time shuffling and splitting the dataset into random 70% training and 30% testing datasets.

7.1 Regression

Regression is a supervised learning method that predicts continuous numerical values based on given input variables. The output variable is known as a dependent variable, while the input variables are known as independent variables.

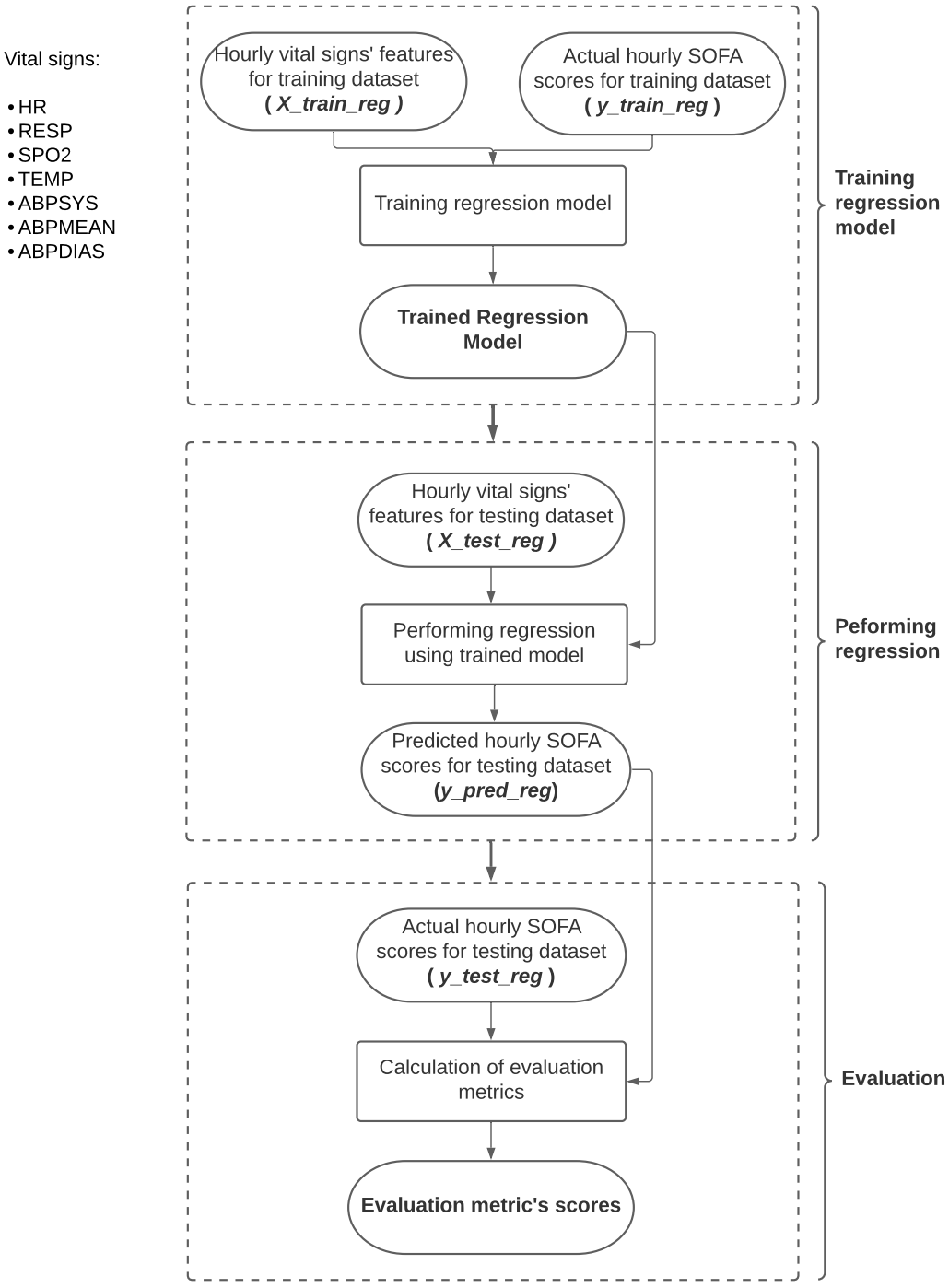


Figure 17: Multilinear regression for predicting hourly SOFA scores

Regression involving multiple independent variables is known as multivariate regression. The regression modeling involves approximating a mapping function from the independent variables to a continuous dependent variable. The basic form of regression is a linear regression. In linear regression, the model assumes a linear relationship between the dependent variable and independent variables. A linear regression with more then one independent input variables is known as multilinear regression. Joseph L. et al. [48] used multilinear regression to calculate a risk score based on vital signs and components of SOFA scores to identify patients entering a pre-shock state. As mentioned earlier, they hypothesized that patients entering this pre-shock state are highly likely to develop septic shock. Hence, inspired by this idea, we decided to use hourly vital signs features to predict hourly SOFA scores using multilinear regression. Figure 17 explains the process followed for predicting hourly SOFA score using regression. We trained the regression model with hourly features of time series data of vital signs (*X_train_reg*) and hourly SOFA score (*y_train_reg*) in the training dataset. Once we have the trained regression model, we performed regression by giving hourly features of time series data of vital signs (*X_test_reg*) in testing dataset as an input to the trained regression model. The model then gave us predicted hourly SOFA scores (*y_pred_reg*) for the *X_test_reg*. We evaluated our regression model by calculating scores for evaluation metrics using the *y_pred_reg* and *y_test_reg*.

7.1.1 Multilinear Regression Models

We implemented simple multivariate linear regression where the hourly SOFA score was the dependent variable, and the hourly features of vital signs were independent variables. We used two models for this: Generalized Linear Model (GLM) and eXtreme Gradient Boosting (XGBoost). The GLM models are a broad class of models, including linear regression, Poisson regression, logistic regression, etc. For this thesis, we used the classical and simple linear regression GLM model. Any GLM model consists of three components, i.e., random component, systematic component, and link function. The description of each of these components is as follows:

- Random component: Specifies the probability distribution of the response or dependent variable.
- Systematic component: Refers to the explanatory or independent variables and specifies their linear combination in creating a linear predictor in case of linear regression.

- Link function: indicates a link between the random and systematic components by determining how the expected values of the response or dependent variable is related to the linear predictor of the explanatory or independent variables.

It also uses maximum likelihood estimation (MLE) rather than least squares optimization, which results in the optimal parameters of a regression model. Least squares optimization is a method for estimating the model parameters by taking a collection of parameters that result in the smallest squared error between the actual values and predicted values of the dependent variable. Whereas the maximum likelihood estimation approach automatically finds the probability distribution and the model parameters that best describe the observed data [33][31].

XGBoost is an implementation of a gradient boosting decision tree algorithm and is commonly used for regression and classification tasks because of its advantages. It follows the ensemble modeling technique of boosting where the model is built using a series of poor models [16]. Every time it builds a model from the training data, it calculates errors and adds a second model for correcting the errors of the previous model. This process of adding models is repeated until the model predicts the complete training data correctly or until no improvement can be made. Additionally, XGBoost also pushes the boundary of computation resources for boosted tree algorithms [16]. Hence, XGBoost improves the speed and performance of difficult ML tasks.

7.1.2 Regression Experiments

We performed several regression experiments using both GLM and XGBoost models with 5-fold cross validation approach. In these experiments, we selected different hourly features of vital signs (independent variables) according to the F-value, mutual information between each of the hourly features of vital signs (independent variables) and the hourly SOFA scores (dependent variable) and XGBoost model based feature importance.

F-value is a result of an F-test, a test used for testing the overall significance of the regression model. In this test, there is a null hypothesis that the values of all of the regression coefficients are equal to zero, indicating that the regression model does not have any predictive capability. F-value is used to reject or accept this null hypothesis. The F-test distinguishes the regression model with no predictor variables (intercept-only model) and analyses whether the added coefficients improved the model [12]. It helps to test the individual effect of each of the many vital signs' features on the SOFA score. Higher the F-values, the higher the value of the feature.

Mutual information measures the decrease in uncertainty for one variable, given a known value of another variable. It refers to the amount of information about a random variable that can be obtained from another random variable. It is also referred to as a measure of dependency between the random variables. The higher the dependency between the two variables, the higher the value of mutual information [70].

In addition to these experiments, we also performed a few experiments on the XGBoost regression model by performing feature selection according to the feature importance determined by the model. The feature importance scores explain how important each feature was in the construction of boosted decision trees in the model. If a feature is used more to make important decisions with decision trees, its value of feature importance is high. To calculate the feature importance of a decision tree, we look at the improvement in performance measure caused by the amount of each feature split point. In XGBoost, we can calculate feature importance according to the following different importance types.

- *weight*: Refers to the number of times a feature is used for data split in all decision trees.
- *gain*: Specifies the average gain for all splits where the feature is used.
- *cover*: Specifies the average coverage for all splits where the feature is used.

For our experiments, we used *gain* as a feature importance type to discover the feature importance. When the importance type is *gain*, the feature importance is calculated as the information gain averaged over all trees. The model first computes the parent node's impurities using the Gini index or entropy and then compute the child node's impurities if you were to use a given feature for the split. Eventually, the information gain is calculated by subtracting the child node impurities from the parent node impurities.

Table 8 shows the hourly features of vital signs with numbering used for multilinear regression experiments. For selecting the best features according to the F-value and mutual information, we used Python's *sklearn.feature_selection.SelectKBest()* function with different score functions. This function selects the best k features according to the k highest scores. We used *sklearn.feature_selection.f_regression()* and *sklearn.feature_selection.mutual_info_regression()* as score functions to determine the F-value and mutual information between the vital signs' features and the SOFA scores respectively.

No.	Feature Label
0	*HR*
1	*RESP*
2	*ABPSYS*
3	*ABPDIAS*
4	*ABPMEAN*
5	*SPO2*
6	*TEMP*
7	*HR_STD*
8	*RESP_STD*
9	*ABPSYS_STD*
10	*ABPDIAS_STD*
11	*ABPMEAN_STD*
12	*SPO2_STD*
13	*TEMP_STD*
14	*HR_ENT*
15	*RESP_ENT*
16	*ABPSYS_ENT*
17	*ABPDIAS_ENT*
18	*ABPMEAN_ENT*
19	*SPO2_ENT*

No.	Feature Label
20	*TEMP_ENT*
21	*ABPDIAS_HR_CORR*
22	*RESP_HR_CORR*
23	*ABPDIAS_ABPSYS_CORR*
24	*ABPMEAN_ABPSYS_CORR*
25	*ABPMEAN_ABPDIAS_CORR*
26	*HR_MIN*
27	*RESP_MIN*
28	*SPO2_MIN*
29	*TEMP_MIN*
30	*ABPSYS_MIN*
31	*ABPDIAS_MIN*
32	*ABPMEAN_MIN*
33	*HR_MAX*
34	*RESP_MAX*
35	*SPO2_MAX*
36	*TEMP_MAX*
37	*ABPSYS_MAX*
38	*ABPDIAS_MAX*
39	*ABPMEAN_MAX*

Table 8: Hourly Vital Signs Features

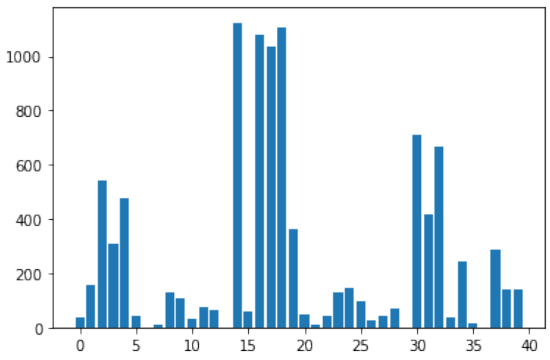


Figure 18: F-value between the features and SOFA score

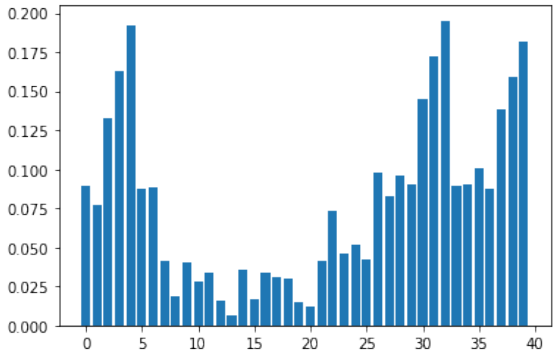


Figure 19: Mutual information between the features and SOFA score

Figures 18 and 19 show the F-values and mutual information between all the features and SOFA score. As illustrated in the figure 18, features *'ABPSYS'*, *'ABPMEAN'*, *'HR_ENT'*,*'ABPSYS_ENT'*, *'ABPDIAS_ENT'*, *'ABPMEAN_ENT'*, *'SPO2_ENT'*, *'ABPSYS_MIN'*, *'ABPDIAS_MIN'*, and *'APBMEAN_MIN'* have high F-values and thus, can be considered important in prediction of SOFA score. Figure 19 shows that features *'ABPSYS'* , *'ABPDIAS'*, *'ABPMEAN'*,*'ABPSYS_MIN'*, *'ABPDIAS_MIN'*, *'APBMEAN_MIN'*, *'ABPSYS_MAX'*, *'ABPDIAS_MAX'*, and *'APBMEAN_MAX'* have high mutual information scores and can be considered as best predictors of SOFA score.

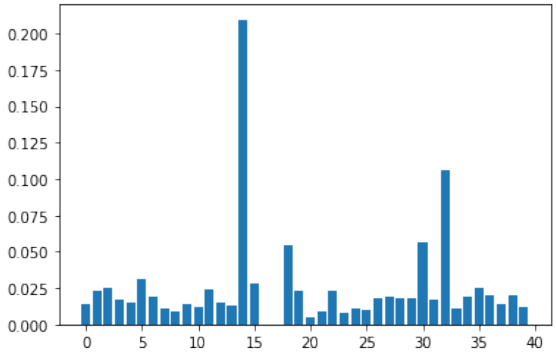


Figure 20: XGBoost model feature importance

Figure 20 shows the plot feature importance determined by the XGBoost regression model. In figure 20 we see that features *'RESP'*, *'ABPSYS'*, *'SPO2'*, *'HR_ENT'*, *'RESP_ENT'*, *'ABPMEAN_ENT'*,*'SPO2_ENT'*, *'ABPSYS_MIN'*, and *'ABPMEAN_MIN'* have high gain importance calculated by the XGBoost regression model.

7.1.3 Regression Evaluation and Results

For evaluating and comparing our experiments, we calculated the most commonly used metrics in regression problems: root mean square error (RMSE) and R-squared (R^2). RMSE is basically the square root of the mean square error (MSE). MSE provides an idea of the error's magnitude and represents the error or difference between the original and predicted values of the dependent variable by squaring the average difference over the data set. It is the standard deviation of the residuals (prediction errors) and that is commonly plotted along the regression line. RMSE is calculated by the formula 2 where N represents the number of samples,y_i and $\hat{y}$ denote the actual and predicted values respectively [19]. Its values can range from 0 to infinity; the lower, the better.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{\mathrm{i}} - \hat{y})^2} \tag{2}$$

R^2, also known as the coefficient of determination, is an statistical measure that indicates how much of the variance of the predicted values (dependent variables) are explained by the independent variable. It can be calculated by the formula 3, where y_i and $\hat{y}$ are the actual and predicted values respectively, and $\bar{y}$ denotes the mean of the actual values [19]. Its values normally range from 0 to 1. R^2 value of 0 indicates that the model does not explain additional variance, and values of 1 indicate that the model explains all of the variances; the higher the scores, the better. It is also possible to have negative values for R^2 if the predictions made by the model fit the data worse than the mean of the output values [19].

$$R^2 = 1 - \frac{\sum(y_{\mathrm{i}} - \hat{y})^2}{\sum(y_{\mathrm{i}} - \bar{y})^2} \tag{3}$$

Table 9 shows the values of RMSE andR^2 calculated for all our regression experiments with GLM and XGBoost. It is clear that the XGBoost model with the top 27 features performed the best with the RMSE value of 2.0457 and R^2 value of 0.0895. According to the figure 20 and table 8, these top 27 predcitors of SOFA score are *'HR'*, *'RESP'*, *'ABPSYS'*, *'ABPDIAS'*, *'ABPMEAN'*, *'SPO2'*, *'TEMP'*, *'ABPSYS_STD'*, *'ABPMEAN_STD'*, *'SPO2_STD'*, *'HR_ENT'*, *'RESP_ENT'*, *'ABPMEAN_ENT'*, *'SPO2_ENT'*, *'RESP_HR_CORR'*, *'HR_MIN'*, *'RESP_MIN'*, *'SPO2_MIN'*, *'TEMP_MIN'*, *'ABPSYS_MIN'*, *'ABPDIAS_MIN'*, *'ABPMEAN_MIN'*, *'RESP_MAX'*, *'SPO2_MAX'*, *'TEMP_MAX'*, *'ABPSYS_MAX'*, and *'ABPDIAS_MAX'* Hence, we chose the XGBoost regression model and the best 27 hourly features of vital signs for our regression task to predict the hourly SOFA scores.

Model	Feature Selection Method	Top K features	RMSE	R2
GLM	All features	-	2.0627	0.0742
	F-value	30	2.0672	0.0702
		25	2.0671	0.0703
		20	2.1093	0.0319
		15	2.1114	0.03
	Mutual Information	30	3.0427	-1.0144
		25	2.1112	0.0302
		20	2.1154	0.0264
		15	3.7558	-2.0692
XGBoost	All features	-	2.0513	0.0844
	F-value	30	2.0611	0.0757
		25	2.0625	0.0744
		20	2.0697	0.068
		15	2.0782	0.0603
	Mutual Information	30	2.0507	0.0849
		25	2.0554	0.0808
		20	2.0859	0.0533
		15	2.0896	0.0499
	Model Feature Importance	33	2.0513	0.0844
		27	**2.0457**	**0.0895**
		24	2.0528	0.0831
		21	2.0561	0.0801
		14	2.0541	0.0819
		12	2.0722	0.0656

Table 9: Regression experiments' results

7.2 Classification

Once we have the hourly predicted SOFA score as a result of our XGBoost regression model, we used them along with the extracted hourly features of the vital signs to perform supervised sequence classification.

Sequence classification has been an widely used for performing classification on video and speech signals but not much in healthcare domain. Zachary C. et al. [47] used LSTM networks to recognize patterns in multivariate time series data. They performed classification on the entire time series sequence by assigning a label to the entire time series sequence. However, they do not segment or split the entire time series sequence of a patient into smaller sequences using a sliding window. Shameek Gosh et al. [32] performed septic shock prediction by recognizing sequential patterns in time series data. They used a symbolic aggregate approximation (SAX) method to transform a time series signal into a discrete sequence. In their work, they also considered the order of the SAX symbols in the sequence to discover the most frequently occurring patterns. In order to find the patterns, they used sliding windows and identified the frequently occurring ordered sequences of items after applying the SAX transformation. However, while applying the SAX transformation, SAX maps a symbol from the average value of the data in a

particular window. It does not consider the actual values in the sequence and thus ignores important information such as the trend of the value change in the window, which might result in inaccurate classification [65].

We wanted to achieve the data-driven discovery of patterns in our time series data to predict whether a sepsis patient will progress into septic shock or not. Therefore, we performed a sequence classification with a sliding window on our time series data's hourly features. A window of fixed length, i.e. a window of 3 hours was moved continuously and step-wise over the time series data. This approach not only captures the data for a particular time window but also accounts for data in the recent past. Therefore, it is beneficial in detecting the changes and patterns in time series data that appear over time. The sliding window approach has been widely used for forecasting, predicting, detecting, and matching patterns in time series data [68] [18] [17] [43].

7.2.1 Sequence Generation and Labeling

In this thesis, the time series sequence classification consisted of two main steps, i.e. sequence and label generation for both the training dataset and the testing dataset. We slided a window of 3 hours every hour over the time series data of vital signs and of the SOFA score for sequence generation. This implies that we used a sliding window of 3 hours, with an overlap of 2 hours to generate sequences of 3 hours. Since we were using a supervised classification method, we also generated labels for each of the sequences generated according to the septic shock onset hour calculated in section 5.4. For every generated sequence of 3 hours, we checked if the patient developed septic shock in the immediate next hour, i.e., the fourth hour. If for a sequence, the patient did not develop a septic shock in the next forthcoming hour, then we labeled that sequence as a 'non-shock' sequence. For a sequence where the patient progressed into septic shock in the upcoming hour, we marked that sequence as a 'shock' sequence. Figures 21 and 22 illustrate the process of sequence and label generation for a sepsis patient that developed a septic shock using a sliding window of 3 hours. We started by taking a window of 3 hours from the hour of sepsis onset time and generated the first sequence, i.e. *sequence 1 (A, B, C)* . We labeled *sequence 1* as *'non-shock'* because the following hour after this sequence, i.e. *t+3 hr* was not a septic shock onset hour. We then slided the window one hour ahead to generate the second sequence, i.e., *sequence 2 (B, C, D)* and label it as *'non-shock'* since the next hour, i.e. *t+4 hr* was neither a septic shock onset hour. This process was repeated until we reached one hour before the septic shock onset hour. As we can see in the figure, for the last sequence, i.e. *sequence 7 (G, H, I)* , the label was *'shock'* as the upcoming hour after *t+8 hr* was a septic shock onset hour. For a sepsis patient that did not progress into septic shock, we generated the sequences in the same manner. However, the labels for all such sequences were always *'non-shock'*.

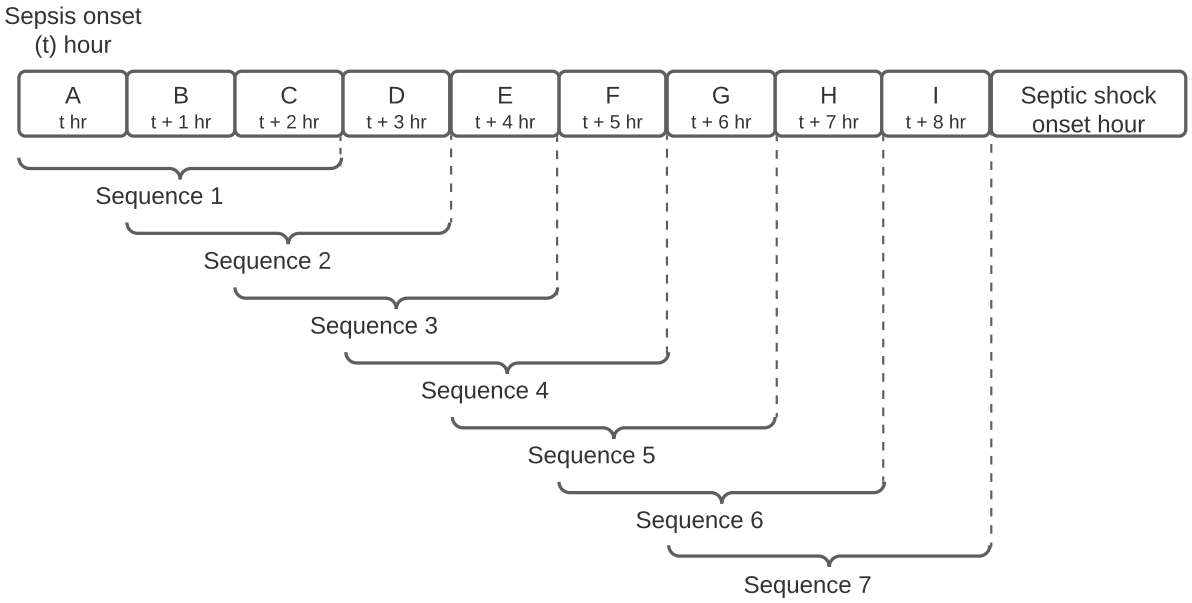


Figure 21: Sequence generation for a shock patient using a sliding window of 3 hours

Sequence 1 : [A, B, C] ; label : non-shock
Sequence 2 : [B, C, D] ; label : non-shock
Sequence 3 : [C, D, E] ; label : non-shock
Sequence 4 : [D, E, F] ; label : non-shock
Sequence 5 : [E, F, G] ; label : non-shock
Sequence 6 : [F, G, H] ; label : non-shock
Sequence 7 : [G, H, I] ; label : shock

Figure 22: Label generation for a shock patient

7.2.2 Classification Model

For classifying these sequences, we created a Long Short-Term Memory (LSTM) network. LSTM is a type of recurrent neural network (RNN) with the ability of learning long-term time dependencies. It was developed to overcome the shortcoming of RNNs. RNNs suffer from the problem of short-term memory. They have a hard time carrying information from earlier time steps to later ones if a sequence is too long enough. Hence, if we try to perform predictions on a long time series sequence, RNN might leave or forget important information from the initial time step. In LSTM, this problem is solved by a mechanism involving gates that regulate the flow of information [24]. These gates learn which data in the sequence is important, keeping only the important ones and throwing away the remaining ones. This makes it easier for these gates to only pass the relevant information for long sequences in order to make good predictions [24].

Figure 23 gives an overview of the process followed for sequence classification. For training the classification model, we generated sequences and labels for the hourly features of vital signs and actual SOFA scores using the hourly sliding window of 3 hours for all the patients in the training dataset. The sequences and labels generated for the training dataset can be denoted by *X_train_sequences* and *y_train_labels* respectively. We then trained the classification model by providing *X_train_sequences* and *y_train_labels* as inputs to it. Once the classification model was trained, we used it for predicting whether the sepsis patient will progress into septic shock or not. We performed classification by taking hourly SOFA scores predicted by the regression model and hourly features of vital signs for all the patients in the testing dataset. We generated test sequences and labels using the same hourly sliding window of 3 hours. These sequences and labels can be represented as *X_test_sequences* and *y_test_labels*. The *X_test_sequences* were passed to the trained classification model, which predicted labels *y_pred_labels* for each of the *X_test_sequences*. Further, we evaluated the classification by calculating the scores for evaluation metrics, using the *y_test_labels* and *y_pred_labels* as our ground truth and predicted labels, respectively.

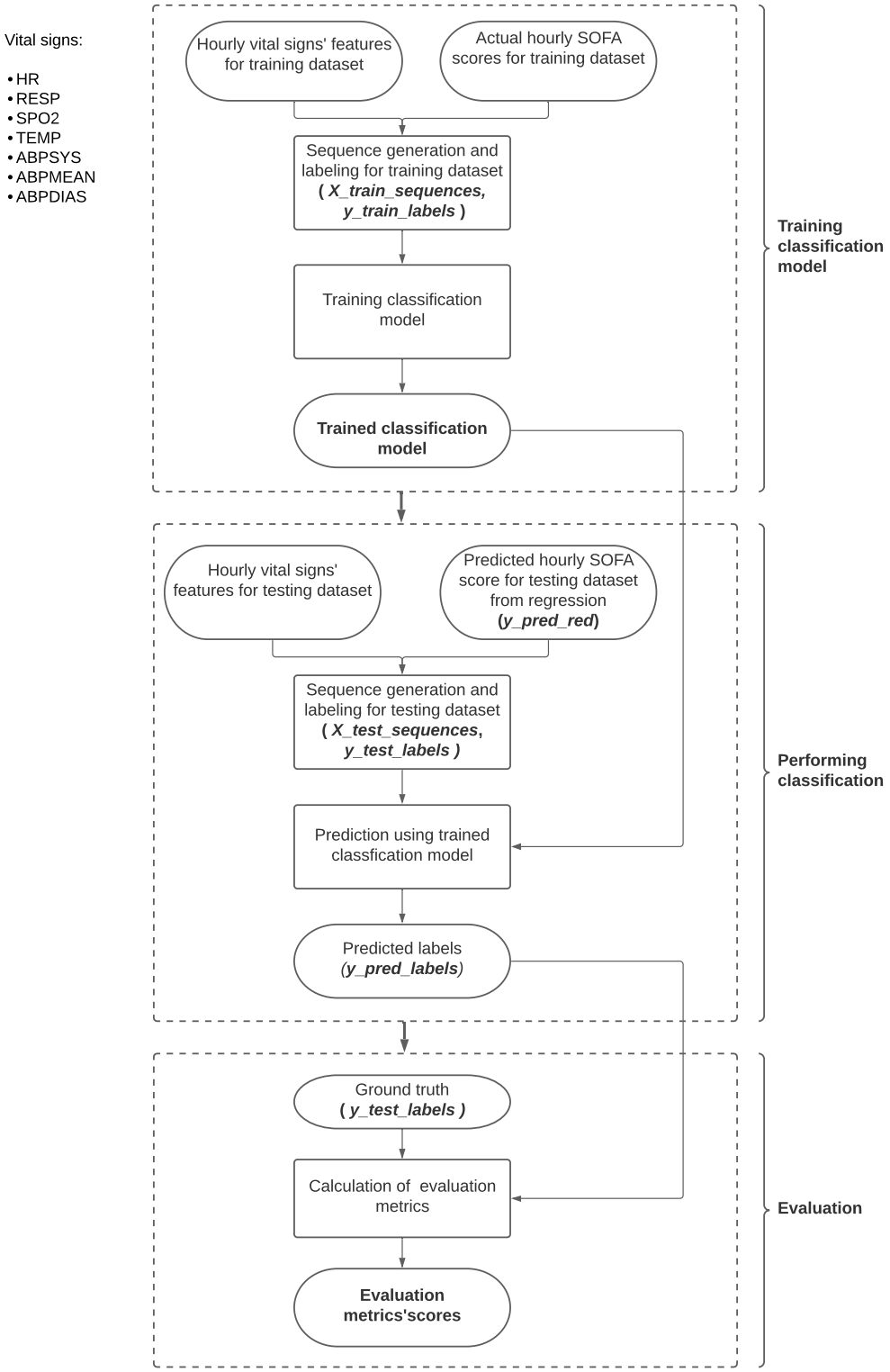


Figure 23: Sequence classification process for prediction of shock

7.2.3 Classification Experiments

We observed that our LSTM model performed poorly on sequence classification with the generated 3 hours sequences and their respective labels because of highly imbalanced data. We had more sequences with *'non-shock'* or *'0'* label than sequences with *'shock'* or *'1'* label.

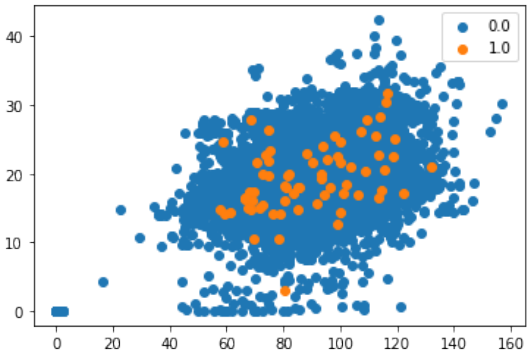


Figure 24: Imbalanced classification data

Figure 24 clearly illustrates how imbalanced our data was. To balance the data and improve our classification model performance, we performed several experiments by assigning different class weights to the LSTM model, using a different ML model, time-periods, sampling techniques, and sequence lengths with a 5-fold cross validation approach.

- Models:
  As mentioned above, we used an LTSM model for our initial classification. We also used an XGBoost model for our classification experiments.

- Time-periods:
  For our initial classification, we used time series data of vital signs and SOFA scores between sepsis onset time and septic shock onset time. For our experiments, we considered the time series data between different time-periods, i.e. from sepsis onset time to septic shock onset time + x hours, in addition to the time series data between the default time-period, i.e. between sepsis onset time and septic shock onset time.

- Sampling techniques:
  We considered two different sampling techniques to balance our dataset, i.e. undersampling the majority class and oversampling the minority class. We used the random undersampling method for undersampling and synthetic minority oversam-

pling technique (SMOTE) for oversampling. In random undersampling, the majority class is undersampled by randomly selecting samples from the majority class [54]. In this thesis, we used the random undersampling implementation of Python's imbalanced-learn API, i.e., *imblearn.under_sampling.RandomUnderSampler()* class [5].

SMOTE selects samples close in the feature space, draws a line between samples in the feature space, and draws a new sample at a point along that line. To create a synthetic instance, it randomly selects a minority class instance a and finds it k nearest minority class neighbors. It then chooses one of the k nearest neighbors b at random, connects a and b from a line segment in the feature space [54]. This approach is effective because the synthetic samples created are relatively close to the existing minority samples in the feature space. In this thesis, we used the SMOTE implementation provided by the imbalanced-learn Python, i.e., *imblearn.over_sampling.SMOTE()* class with default 5 nearest neighbours ($k = 5$) [6]

- Class weight:
  One way to handle imbalanced data is to consider the skewed distribution of the classes and assign different weights to both the majority and minority classes. These different weights affect the classification during the training of the classification model. The basic idea is to penalize the misclassification made by the minority class by setting a higher class weight and reducing weight for the majority class at the same time. We gave more importance or weight to the minority class in the cost function by assigning a higher weight to the minority class. The algorithm provided a higher penalty to the minority class, and the model could decrease errors for the minority class [62].

- Sequence lengths:
  We generated sequences of 3 hours for our initial classification. Additionally, we used sequences of different lengths for our experiments, such as sequences of 2 hours and 1 hour. For generating sequences of 2 hours, we slided a window of 2 hours every hour, and for sequences of 1 hour, we slided a window of 1 hour every 15 mins.

- Labeling: In our initial classification process, we labeled the sequences according to the septic shock onset time. We checked whether the upcoming hour after the current sequence was a septic shock onset hour or not. If the next upcoming hour was a septic shock onset hour, we assigned a *'shock'* or *'1'* label; else, we assigned a *'non-shock'* or *'0'* label to that particular sequence. We considered the following additional labeling techniques for our experiments.

  – Binary-labeling according to shock or non-shock condition.

– Binary-labeling according to an increase in SOFA score.

– Multi-labeling, according to an increase in the SOFA score.

Zachary C. et al. [47] assigned a label to the entire time series sequence according to the diagnosis. Hence, for binary-labeling according to shock or non-shock condition, we labeled all generated sequences of a shock patient as *'shock'* or *'1'* and all sequences of a non-shock patient as *'non-shock'* or *'0'*. For binary-labeling according to an increase in SOFA scores, we generated an additional hourly feature *hourlyclass*. We compared the SOFA score for each hour after the sepsis onset time with the SOFA score at sepsis onset time. For a particular hour after sepsis onset time, if the increase in the SOFA score was greater than 1, we assigned a *'shock'* or *'1'* value for the new hourly feature *hourlyclass*. If the increase in SOFA score was less than or equal to 1, we assigned a *'non-shock'* or *'0'* value for the new hourly feature *hourlyclass*. While labeling the generated sequences, we checked for the value of *hourlyclass* for the next forthcoming hour. If the value of *hourlyclass* for the next upcoming hour immediately after the sequence was *'shock'* or *'1'*, we assigned a *'shock'* or *'1'* label for that sequence; else, *'non-shock'* or the *'0'* label was assigned to that sequence.

In multi-labeling according to an increase in SOFA score, we considered 3 labels, i.e. *'non-shock'*, *'pre-shock'*, and *'shock'*. We generated a new hourly feature *hourlyclass*. We compared the SOFA score for each hour after the sepsis onset time with the SOFA score at sepsis onset time. For a particular hour after sepsis onset time, if the increase in SOFA score was less than or equals to 2, we assigned a *'pre-shock'* or *'1'* value for the new hourly feature *hourlyclass*. If the increase in SOFA score greater than 2, we assigned a *'shock'* or *'2'* value else, we assigned a *'non-shock'* or *'0'* value for the new hourly feature *hourlyclass*. While labeling the generated sequences, we checked for the value of *hourlyclass* for the next forthcoming hour. If the value of *hourlyclass* for the next upcoming hour immediately after the sequence was *'pre-shock'* or *'1'*, we assigned a *'pre-shock'* or *'1'* label for that sequence. If the value of *hourlyclass* for the next upcoming hour immediately after the sequence was *'shock'* or *'2'*, a *'shock'* or *'2'* label is assigned for that sequence; else, *'non-shock'* or *'0'* label is assigned to that sequence.

7.2.3.1 Experiments using different class weights

We performed this experiment to overcome the problem of imbalanced data in our initial classification using the LSTM model. This experiment considered hourly time series data of vital signs and SOFA scores between sepsis onset hour and septic shock onset hour. We generated sequences of 3 hours by a sliding window of 3 hours hourly, as shown in figure

21. We checked whether the upcoming hour after the current sequence is a septic shock onset hour or not for labeling the sequences, as shown in figure 22. If the next upcoming hour was a septic shock onset hour, we assigned a *'shock'* label; else, we assigned a *'non-shock'* label to that particular sequence. Table 10 shows the different values of different parameters used for this experiment.

	Model	**Time-period**	**Sampling**	**Class weight**	**Sequence length**	**Labeling method**
EXP 1	LSTM	Sepsis onset - shock onset	-	75-25	3 h	Binary labels according to shock onset time
EXP 2	LSTM	Sepsis onset - shock onset	-	80-20	3 h	Binary labels according to shock onset time
EXP 3	LSTM	Sepsis onset - shock onset	-	85-15	3 h	Binary labels according to shock onset time

Table 10: Parameters for experiments with different class weights

7.2.3.2 Experiments using data from different time-periods

As mentioned earlier, we considered the hourly features of time series data between sepsis onset time and shock onset time as our default time-period. For this experiment, we considered the hourly time series data of vital signs and the SOFA score between sepsis onset time and septic shock onset time + x hours. We assumed that the vital signs and SOFA score a few hours after the septic shock onset time are also indicative of septic shock. Therefore, we considered the time-period of sepsis onset time and septic shock onset time + 2 hours, sepsis onset time and septic shock onset time + 3 hours, sepsis onset time and septic shock onset time + 4 hours, and sepsis onset time and septic shock onset time + 5 hours. We also generated sequences of different lengths, e.g., 1 hour, 2 hours, and 3 hours sequences using a sliding window of different lengths. The sequence generation and labeling process applied was the same as shown in figures 21 and 22. We then used these sequences and labels with LSTM and XGBoost models for performing binary classification and followed the same testing and evaluation process as in our initial classification process (refer figure 23). Table 11 shows the different values of different parameters used for this experiment.

	Model	Time-period	Sampling	Class weight	Sequence length	Labeling method
EXP 4	LSTM	Sepsis onset - shock onset + 2h	-	-	3 h	binary labels according to shock onset time
EXP 5	LSTM	Sepsis onset - shock onset + 3h	-	-	3 h	binary labels according to shock onset time
EXP 6	LSTM	Sepsis onset - shock onset + 4h	-	-	3 h	binary labels according to shock onset time
EXP 7	LSTM	Sepsis onset - shock onset + 5h	-	-	3 h	binary labels according to shock onset time
EXP 8	XGBoost	Sepsis onset - shock onset + 2h	-	-	3 h	binary labels according to shock onset time
EXP 9	XGBoost	Sepsis onset - shock onset + 3h	-	-	3 h	binary labels according to shock onset time
EXP 10	XGBoost	Sepsis onset - shock onset + 4h	-	-	3 h	binary labels according to shock onset time
EXP 11	XGBoost	Sepsis onset - shock onset + 5h	-	-	3 h	binary labels according to shock onset time

Table 11: Parameters for experiments with different time-periods

7.2.3.3 Experiments using different sampling techniques

In this experiment, we converted the imbalanced data into balanced data by applying different sampling methods to improve the classification results. We considered the same time series data, sequence length, and process to generate and label the sequences as in our initial classification for this experiment. This experiment considered hourly time series data of vital signs and SOFA scores between sepsis onset hour and septic shock onset hour. We generated sequences of 3 hours by a sliding window of 3 hours hourly, as shown in figure 21. We checked whether the upcoming hour after the current sequence was a septic shock onset hour or not for labeling the sequences, as shown in figure 22. If the next upcoming hour was a septic shock onset hour, we assigned a *'shock'* label; else, we give a *'non-shock'* label to that particular sequence. Once we generated the training sequences and labels, we applied the sampling techniques to balance our training data. We applied SMOTE for oversampling and random undersampling for undersampling the training data. Using these sampling methods resulted in a balanced training dataset, which we then used to train our LSTM and XGBoost models. We then follow the same testing and evaluation process, as in our initial classification process (refer figure 23). The different values of different parameters used for this experiment are shown in the table 12.

	Model	Time-period	Sampling	Class weight	Sequence length	Labeling method
EXP 12	LSTM	Sepsis onset - shock onset	Undersampling	-	3 h	Binary labels according to shock onset time
EXP 13	LSTM	Sepsis onset - shock onset	Oversampling	-	3 h	Binary labels according to shock onset time
EXP 14	XGBoost	Sepsis onset - shock onset	Undersampling	-	3 h	Binary labels according to shock onset time
EXP 15	XGBoost	Sepsis onset - shock onset	Oversampling	-	3 h	Binary labels according to shock onset time

Table 12: Parameters for experiments with different sampling methods

After performing the above experiments, we noticed that the vital signs of the shock patient were highly unstable compared with that of a non-shock patient. Hence, we decided to experiment by changing how we labeled the generated sequences. We also used different sequences of 1 hour, 2 hours, and 3 hours and sampling methods, i.e., SMOTE and random undersampling used in section 7.2.3.3.

7.2.3.4 Experiments for labeling according to shock and non-shock condition

This experiment considered the hourly features of time series data of vital signs and SOFA scores in our default time-period, i.e., between sepsis onset hour and septic shock onset hour. We generated sequences of 1 hour, 2 hours, and 3 hours using the sliding window approach. All generated sequences of a shock patient were assigned a *'shock'* label or *'1'*. Whereas, we labeled all generated sequences of a non-shock patient as *'non-shock'* or *'0'*. We then performed and evaluated this classification using LSTM and XGBoost models following the same training, testing, and evaluation process as in our initial classification. Table 13 shows the different values of different parameters used for this experiment.

	Model	Time-period	Sampling	Class weight	Sequence length	Labeling method
EXP 16	XGBoost	Sepsis onset - shock onset	-	-	1 h	Binary labeling according to shock / non-shock
EXP 17	XGBoost	Sepsis onset - shock onset	-	-	2 h	Binary labeling according to shock / non-shock
EXP 18	XGBoost	Sepsis onset - shock onset	-	-	3 h	Binary labeling according to shock / non-shock
EXP 19	XGBoost	Sepsis onset - shock onset	Undersampling	-	1 h	Binary labeling according to shock / non-shock
EXP 20	XGBoost	Sepsis onset - shock onset	Undersampling	-	2 h	Binary labeling according to shock / non-shock
EXP 21	XGBoost	Sepsis onset - shock onset	Undersampling	-	3 h	Binary labeling according to shock / non-shock
EXP 22	XGBoost	Sepsis onset - shock onset	Oversampling	-	1 h	Binary labeling according to shock / non-shock
EXP 23	XGBoost	Sepsis onset - shock onset	Oversampling	-	2 h	Binary labeling according to shock / non-shock
EXP 24	XGBoost	Sepsis onset - shock onset	Oversampling	-	3 h	Binary labeling according to shock / non-shock

Table 13: Parameters for experiments with binary-labeling according to shock & non-shock condition

7.2.3.5 Experiments for binary-labeling according to an increase in SOFA score

Taking the hourly features extracted from the time series data of vital signs and SOFA scores between sepsis onset time and shock onset time, sequences of 1 hour, 2 hours, and 3 hours were created for the sliding window approach. Binary-labels (*'shock'* or *'non-shock'*) were assigned to these sequences according to an increase in SOFA score from sepsis onset time. We generated an additional hourly feature *hourlyclass*. We compared the SOFA score for each hour after the sepsis onset time with the SOFA score at sepsis onset time. For a particular hour after sepsis onset time, if an increase in the SOFA score was greater than 1, we assigned a *'shock'* or *'1'* value for the new hourly feature *hourlyclass*. If the increase in SOFA score was less than or equal to 1, we assigned a *'non-shock'* or *'0'* value for the new hourly feature *hourlyclass*. While labeling the generated sequences, we checked for the value of *hourlyclass* for the next forthcoming hour. If the value of *hourlyclass* for the next upcoming hour immediately after the sequence was *'shock'* or *'1'*, we assigned a *'shock'* or *'1'* label for that sequence; else, *'non-shock'* or *'0'* label was assigned to that sequence. The same training, testing, and evaluation process was carried

out as in our initial classification using both LSTM and XGBoost models. The different values of different parameters used for this experiment are shown in the table 14.

	Model	**Time-period**	**Sampling**	**Class weight**	**Sequence length**	**Labeling method**
EXP 25	XGBoost	Sepsis onset - shock onset	-	-	1 h	Binary labeling according to increase in SOFA
EXP 26	XGBoost	Sepsis onset - shock onset	-	-	2 h	Binary labeling according to increase in SOFA
EXP 27	XGBoost	Sepsis onset - shock onset	-	-	3 h	Binary labeling according to increase in SOFA
EXP 28	XGBoost	Sepsis onset - shock onset	Undersampling	-	1 h	Binary labeling according to increase in SOFA
EXP 29	XGBoost	Sepsis onset - shock onset	Undersampling	-	2 h	Binary labeling according to increase in SOFA
EXP 30	XGBoost	Sepsis onset - shock onset	Undersampling	-	3 h	Binary labeling according to increase in SOFA
EXP 31	XGBoost	Sepsis onset - shock onset	Oversampling	-	1 h	Binary labeling according to increase in SOFA
EXP 32	XGBoost	Sepsis onset - shock onset	Oversampling	-	2 h	Binary labeling according to increase in SOFA
EXP 33	XGBoost	Sepsis onset - shock onset	Oversampling	-	3 h	Binary labeling according to increase in SOFA

Table 14: Parameters for experiments with binary-labeling according to increase in SOFA score

7.2.3.6 Experiments for Multi-labeling according to an increase in SOFA score

Taking the hourly features extracted from the time series data of vital signs and SOFA score between sepsis onset time and shock onset time, sequences of 1 hour, 2 hours, and 3 hours were created the sliding window approach. Multi-labels (either *'non-shock'* or *'pre-shock'* or *'shock'*) are assigned to these sequences according to an increase in SOFA score from sepsis onset time. We generated a new hourly feature *hourlyclass*. We compared the SOFA score for each hour after the sepsis onset time with the SOFA score at sepsis onset time. For a particular hour after sepsis onset time, if an increase in the SOFA score was less than or equals to 2, we assigned a *'pre-shock'* or *'1'* value for the new hourly feature *hourlyclass*. If the increase in SOFA score was greater than 2, we assigned a *'shock'* or *'2'* value else, we assigned a *'non-shock'* or *'0'* value for the new hourly feature *hourlyclass*. While labeling the generated sequences, we checked for the value of *hourlyclass* for the next forthcoming hour. If the value of *hourlyclass* for the next upcoming hour immediately

after the sequence was *'pre-shock'* or *'1'*, we assigned a *'pre-shock'* or *'1'* label for that sequence. If the value of *hourlyclass* for the next upcoming hour immediately after the sequence was *'shock'* or *'2'*, a *'shock'* or *'2'* label was assigned for that sequence; else, *'non-shock'* or *'0'* label was assigned to that sequence. The same training, testing, and evaluation process, as in our initial classification, was followed. Also, for performing classification, we used both LSTM and XGBoost models. Table 15 shows the different values of different parameters used for this experiment.

	Model	**Time-period**	**Sampling**	**Class weight**	**Sequence length**	**Labeling method**
EXP 34	XGBoost	Sepsis onset - shock onset	-	-	1 h	Multi-labeling according to increase in SOFA
EXP 35	XGBoost	Sepsis onset - shock onset	-	-	2 h	Multi-labeling according to increase in SOFA
EXP 36	XGBoost	Sepsis onset - shock onset	-	-	3 h	Multi-labeling according to increase in SOFA
EXP 37	XGBoost	Sepsis onset - shock onset	Undersampling	-	1 h	Multi-labeling according to increase in SOFA
EXP 38	XGBoost	Sepsis onset - shock onset	Undersampling	-	2 h	Multi-labeling according to increase in SOFA
EXP 39	XGBoost	Sepsis onset - shock onset	Undersampling	-	3 h	Multi-labeling according to increase in SOFA
EXP 40	XGBoost	Sepsis onset - shock onset	Oversampling	-	1 h	Multi-labeling according to increase in SOFA
EXP 41	XGBoost	Sepsis onset - shock onset	Oversampling	-	2 h	Multi-labeling according to increase in SOFA
EXP 42	XGBoost	Sepsis onset - shock onset	Oversampling	-	3 h	Multi-labeling according to increase in SOFA

Table 15: Parameters for experiments with multi-labeling according to increase in SOFA score

7.2.4 Classification Evaluation and Results

We used sensitivity, specificity, precision, F-score, and balanced accuracy as evaluation metrics to assess the classification of the experiments. The values of these metrics range from 0 to 1, where 0 and 1 are considered as the worst and best values, respectively. In order to calculate these, we first identified the number of true positives (TP), true negative (TN), false positives (FP), and false negatives (FN).

- Sensitivity or recall: It determines how many observations out of all positive ones have been classified as positive. For our use case, It determines out of all sepsis patients the ones that progressed into septic shock patients, and how many were predicted correctly, i.e. that they would progress into septic shock [19] [30]. It is calculated by the below formula 4.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{4}$$

- Specificity: It determines how many observations out of all negative ones have been classified as negative. For our use case, It determines all sepsis patients that did not progress into septic shock patients, how many were predicted correctly, i.e. that they would not progress into septic shock [19] [30]. It is calculated by the below formula 5.

$$Specificity = \frac{TN}{(TN + FP)} \tag{5}$$

- Precision: It determines the number of observations that were predicted as positive and that are in reality positive. Considering our use case, this means for how many sepsis patients our model predicted that they would progress into septic shock out of all patients that actually progressed into septic shock [19] [30]. It is calculated by the below formula 6.

$$Precision = \frac{TP}{(TP + FP)} \tag{6}$$

- F-score: It is a harmonic mean of precision and recall. This metric balances both precision and recall. The value of the F-score is not high if the precision or recall is improved [19] [30]. It is calculated by the below formula 7.

$$F - score = \frac{2\ X\ Recall\ X\ Precision}{(Recall + Precision)} \tag{7}$$

- Balanced accuracy: Accuracy is the percentage of labels that the classification model predicts correctly. While balanced accuracy is the average accuracy of the model per class. Balanced accuracy is a good metric to use when the dataset is imbalanced. It indicates the accuracy of the classification model, considering the accuracy of

predictions made for each class [30]. For example, if you have 100 labels, 30 for class 1 and 70 for class 2. If your model predicted 80 out of 100 labels correctly, then the accuracy would be 80%. However, for calculating balanced accuracy, we need to consider the labels predicted correctly for each class, i.e., for class 1 with 30 labels, the model predicted 10 labels correctly, and for class 2 with 70 labels, the model predicted all 70 labels correctly. This implies that the accuracy for class 1 and class 2 is 10/30 = 33.33% and 70/70 = 100%. Balanced accuracy is then calculated as the average of accuracy for both the classes, i.e. (33.33 + 100)/2 = 66.66%. For our use case, balanced accuracy determines how accurate the classification model is, considering the patients were correctly classified into the 'shock' class and 'non-shock' class. It is calculated by the below formula 8.

$$Balanced\ Accuracy = \frac{(Sensitivity + Specificity)}{2} \tag{8}$$

Table 16 shows the scores for the above evaluation metrics calculated for all the classification experiments performed with hourly features of vital signs and hourly predicted SOFA score explained in section 7.2.3. On comparing the evaluation metrics' scores for all the experiments, we observed that the experiment with hourly sequence generation from the time series data of vital signs and SOFA score between sepsis onset time and shock onset time, applying SMOTE oversampling technique, and labeling according to shock and non-shock condition (EXP 22) mentioned in section 7.2.3.4 performed the best. For this best performing experiment, we achieved sensitivity of 87.69 %, specificity of 61.57 % , precision of 67.36% , F-score of 76.19 % , and balanced accuracy of 74.63 %.

After performing the experiments in section 7.2.3, we observed that the XGBoost classification model with hourly sequences and labeling according to the shock or non-shock condition performed the best. The hourly generated sequences were labeled as *'shock'* or *'1'* for all sepsis patients that progressed into septic shock and as *'non-shock'* or *'0'* for all sepsis patients that did not progress into septic shock. To further analyze this best-performing classification method's results, we analyze the septic shock patients who had the best and worst predictions made by the above classification model with hourly sequences.

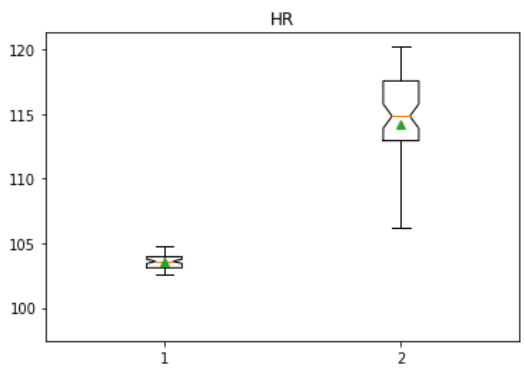
The best predicted patient had an F-score of 1 and 12 hours between sepsis onset time and septic shock onset time, while the worst predicted patient had an F-score of 0.2667 and 13 hours between sepsis onset time and septic shock onset time. Also, the balanced accuracy of the best and worst predictions made for the shock patients was 100% and 15.38%, respectively. This implies that all hourly sequences for the best predicted patient were correctly classified as *'shock'* with an F-score of 1, resulting in the prediction of septic shock immediately after sepsis onset time. However, for the worst predicted patient, only

15.38% of the hourly sequences were classified correctly as *'shock'* with an F-score of 0.2667, resulting in the worst and delayed prediction of septic shock.

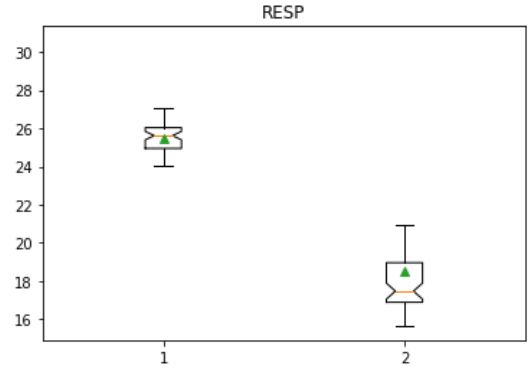
Exp No.	Sensitivity	Specificity	Precision	F-score	Balanced Accuracy
EXP 1	0.4664	0.663	0.4370	0.4512	0.5647
EXP 2	0.5731	0.6119	0.4531	0.5061	0.5925
EXP 3	0.5455	0.5143	0.3866	0.4525	0.5299
EXP 4	0.0761	0.9806	0.1346	0.0972	0.5183
EXP 5	0.0442	0.9878	0.1515	0.0685	0.5160
EXP 6	0.0588	0.9825	0.1633	0.0865	0.5207
EXP 7	0.0755	0.9681	0.1364	0.0972	0.5218
EXP 8	0.0333	1	1	0.0645	0.5167
EXP 9	0.0265	0.9987	0.5	0.0504	0.5126
EXP 10	0.0294	0.9996	0.8	0.0567	0.5145
EXP 11	0.0252	0.9996	0.8	0.0488	0.5124
EXP 12	0.3409	0.6831	0.0213	0.0401	0.5120
EXP 13	0	0.9954	0	0	0.4977
EXP 14	0.5454	0.6886	0.0342	0.0643	0.6170
EXP 15	0.0682	0.9853	0.0857	0.0759	0.5267
EXP 16	0.6958	0.7327	0.7019	0.6989	0.7143
EXP 17	0.3093	0.7595	0.5424	0.3940	0.5344
EXP 18	0.3290	0.7840	0.5874	0.4217	0.5565
EXP 19	0.6612	0.7407	0.6976	0.6789	0.7009
EXP 20	0.3476	0.7435	0.5554	0.4276	0.5456
EXP 21	0.3523	0.7709	0.5897	0.4411	0.5616
EXP 22	**0.8769**	**0.6157**	**0.6736**	**0.7619**	**0.7463**
EXP 23	0.6040	0.7216	0.6667	0.6338	0.6628
EXP 24	0.6570	0.7047	0.6753	0.6660	0.6809
EXP 25	0.6374	0.5039	0.6374	0.6374	0.6421
EXP 26	0.5816	0.5541	0.5816	0.5816	0.5847
EXP 27	0.5930	0.5501	0.5930	0.5930	0.5976
EXP 28	0.6341	0.4792	0.6341	0.6341	0.6396
EXP 29	0.5409	0.4403	0.5409	0.5409	0.5520
EXP 30	0.5799	0.4811	0.5799	0.5799	0.5905
EXP 31	0.6213	0.4476	0.6213	0.6213	0.6274
EXP 32	0.5488	0.4427	0.5488	0.5488	0.5605
EXP 33	0.5565	0.4672	0.5565	0.5565	0.5661
EXP 34	0.4254	1	0.4254	0.4254	0.4293
EXP 35	0.3746	0.5385	0.3747	0.3747	0.3759
EXP 36	0.3688	0.5	0.3688	0.3688	0.3743
EXP 37	0.4220	0.9524	0.4220	0.4220	0.4293
EXP 38	0.3637	0.4286	0.3637	0.3637	0.3859
EXP 39	0.3579	0.4667	0.3579	0.3579	0.3811
EXP 40	0.4154	0.9048	0.4154	0.4154	0.4237
EXP 41	0.3593	0.5333	0.3593	0.3593	0.3852
EXP 42	0.3363	0.2727	0.3363	0.3363	0.3633

Table 16: Results of classification experiments with predicted SOFA score

Since the vital signs and SOFA scores of the septic shock patients recorded a few minutes before the shock onset are highly indicative of shock, we analyze such data for the best and worst predicted patients using a box-plot. Box-plots can be useful to check the variability or dispersion of data based on the minimum, first quartile, median, third quartile, and maximum values.

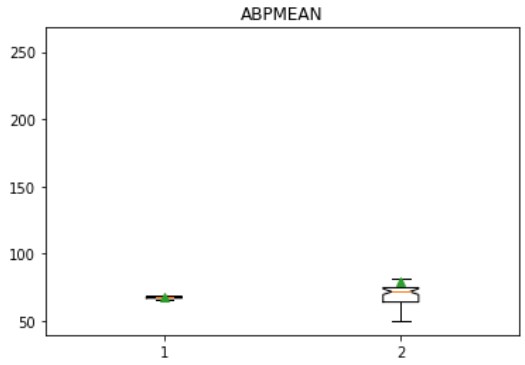


(a) HR of best and worst predicted patients

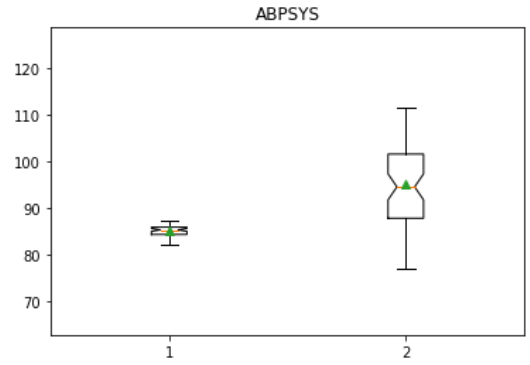


(b) RESP of best and worst predicted patients

Figure 25: Box plots for HR & RESP



(a) ABPMEAN of best and worst predicted patients

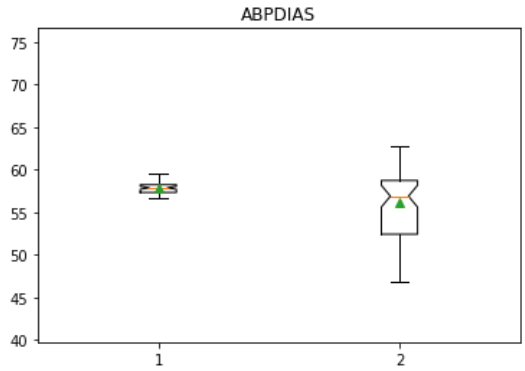


(b) ABPSYS of best and worst predicted patients

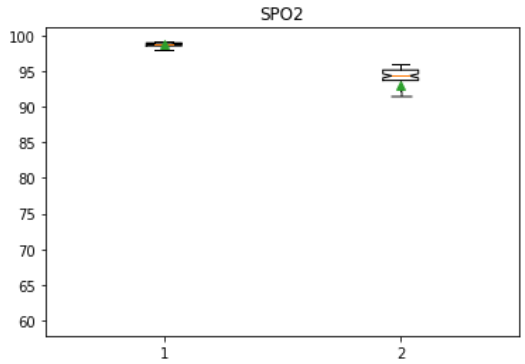
Figure 26: Box plots for ABPMEAN & ABPSYS

Figures 25, 26, 27, and 28 show the box-plots for the vital signs and SOFA scores included in a one hour sequence just before the septic shock onset, i.e. between 75 minutes before the shock onset (septic shock onset time - 75 minutes) and 15 minutes before the shock onset(septic shock onset time - 15 minutes), for the best and worst predicted patient. In these plots, *'1'* and *'2'* represents the best and worst predicted patients respectively. The orange line indicates the median, and the green triangle indicates the mean value. Using these plots, we made the following observations for each of the vital signs and the SOFA score.

- **HR :** figure 25(a) shows the box-plot of heart rate for the best and worst predicted patient. The plot is indicative of an abnormally high heart rate for both patients.



(a) ABPDIAS of best and worst predicted patients

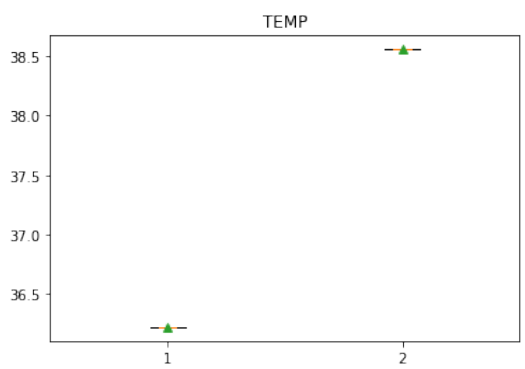


(b) SPO2 of best and worst predicted patients

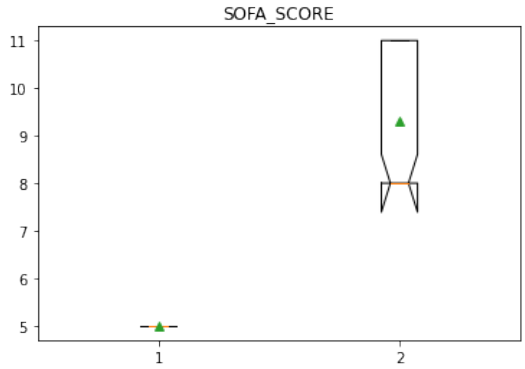
Figure 27: Box plots for ABPDIAS & SPO2

higher Lengths of the boxes and whiskers for the worst predicted patient indicate that the heart rate data for the worst predicted patient had high standard deviation as compared to that of the best predicted patient. Also, For the best predicted patient, the interquartile range was much smaller than that of the worst predicted patient. Hence, we can say that the data for worst predicted patient was more dispersed and variable than the data for best predicted patient. Moreover, the median and mean values for the best predicted patient were much lower than the median and mean heart rate of the worst predicted patient. The values for the best predicted patient were close to the median and mean values and the length of the whiskers and the boxes was also even on both the sides of the median value. Hence, we can say that for the best predicted patient, the data was symmetric and normally distributed. Whereas, for the worst predicted patient, the minimum value was far away from the median and mean values. Also, 75% of the values were much higher than the minimum value indicating that the data for the worst predicted patient was skewed with high standard deviation. This implies that the values of the heart rate for the best predicted patient were less variable with lower standard deviation while the values of heart rate for the worst predicted patient were highly variable with high standard devitation.

- **RESP :** Respiratory rate data for the best and worst predicted patient is presented in figure 25(b). As we can see, the values of respiratory rate for the best predicted patient were abnormally high, i.e. greater than 24 breaths per minute (bpm) , whereas the values for the worst predicted patient were lower than the minimum value of the best predicted patient. Lengths of the boxes and whiskers indicate the standard deviation in data. This implies that the values for the best predicted patient had less standard deviation as compared to the values of the worst predicted



(a) TEMP of best and worst predicted patients



(b) SOFA Score of best and worst predicted patients

Figure 28: Box plots for TEMP & SOFA score

patient because the lengths of the whiskers and boxes for the best predicted patient were smaller than that of the worst predicted patient. Also, the interquartile range value and the difference between the minimum and maximum values for the best predicted patient was much smaller than the that of the worst predicted patient. Moreover, for the best predicted patient, the length of whiskers on either side of the median value was even. However, length of the box between the median and first quartile (Q1) was slightly greater than the length of the box between median and third quartile (Q3) for the best predicted patient. Therefore, we can say that the values for the best predicted patient were slightly skewed but less variable and dispersed. In contrast, the maximum value of respiratory rate for the worst predicted patient was far from the median and the 50% of the values above the median were much more spread out and highly variable. Hence, the data for the worst predicted patient was highly skewed and variable as compared to the data for the best predicted patient. Consequently, the respiratory data for the best predicted patient was less variable with low standard deviations whereas the data data for the worst predicted patient was more variable and had high standard deviation.

- **ABPMEAN :** Figure 26(a) explains the variability and dispersion in mean arterial blood pressure values for the best and worst predicted patients. The mean arterial blood pressure values for the best predicted patient were concentrated around the median and mean values. Moreover, the difference between the minimum and maximum values was very small (almost negligible). This indicates that the data had no or a negligible standard devitation and consequently, had no variability in it. In contrast, for the worst predicted patient, length of the whiskers and boxes on either side of the median was different. The interquartile range value and the difference between the minimum and maximum value was also high. The lower whisker has

greater length indicating that the minimum value lies far from the Q1 and median. The 50% of the data below median was highly variable and dispersed. This implies that the the mean arterial blood pressure values for the worst predicted patient were more variable and had high standard deviation.

- **ABPSYS :** Dispersion and variability in systolic arterial blood pressure are depicted by figure 26(b). As we can see in the plot, for the best predicted patient, the length of the boxes and whiskers on either side of the median was even. Hence, the data was not skewed and is normally distributed. Moreover, the interquartile range value and the difference between the maximum and minimum values was very less. The values lie close to the median and mean values indicating less variability and low standard deviation. For the worst predicted patient, the length of the boxes and whiskers on either side of the median was even indicating a normal distibution and no skewness in the data. However, the lenghts of the boxes and whiskers was large. Moreover, the interqaurtile range value and the difference between the maximum and minimum value was large. This implies that the values for worst predicted patient were highly variable with larger standard deviation as compared to that of the best predicted patient.

- **ABPDIAS :** Figure 27(a) shows the plots for the diastolic arterial blood pressure values for the best and worst predicted patients. For the best predicted patient, the length of the upper whisker was greater than the length of the lower whisker indicating that the maximum value was a bit far form the median and mean values in the data. Hence, we can say that the data had some skewness in it. However, the values for the best predicted patient were very much close to the mean and median values. Moreover, the intequartile range value and difference between the maximum and minimum values was very less. This implies that the data had less variability and low standard deviation.

  For the worst predicted patient, the length of the whiskers and boxes were uneven on both sides of the median value indicating skewness in the data. The intequartile range value and difference between the maximum and minimum values was much higher than that of the best predicted patient. Moreover, the minimum value was too far from the median value and the data in lower quartile Q1 was highly variable. This implies that the systolic arterial blood pressure data for the worst predicted patient was more variable and had high standard deviation as compared to the systolic arterial blood pressure data for the best predicted patient.

- **SPO2 :** Figure 27(b) illustrates the blood oxygen saturation level data for the best and worst predicted patient. The blood oxygen saturation level values for the best predicted patient were concentrated around the median and mean values. Moreover, the difference between the minimum and maximum values was very small

(almost negligible). This indicates that the data had no or a negligible standard devitation and consequently, had no variability in it. In contrast, for the worst predicted patient, length of the whiskers and boxes on either side of the median was different. The interquartile range value and the difference between the minimum and maximum value was also high as compared to that of the best predicted patient. The lower whisker had greater length indicating that the minimum value lied far from the Q1 and median. The 50% of the data below median was highly variable and dispersed. This implies that the the blood oxygen saturation values for the worst predicted patient were more variable and had high standard deviation.

- **TEMP :** Figure 28(a) shows the temperature data for both best and worst predicted patients. For both best and worst predicted patients, there were no boxes and whiskers. Consequently, the values had no standard deviation and were not at all dispersed. Hence, we can say that the temperature data for both patients was uniform without any changes. However, as we can see in the plot, the best predicted patient had extremely low temperatures (less than 36.2 Degrees Celsius) while the worst predicted patient had an extremely high temperature (greater than 38.5 Degrees Celsius).

- **SOFA score :** SOFA score for the best and worst predicted patient is presented in figure 28(b). As we can see, for the best predicted patient, there were no boxes and whiskers, indicating no standard deviation and spread in the SOFA score data. Hence, we can say that the SOFA score was uniform and did not change for the best predicted patient. However, for the worst predicted patient, the length of the boxes on either side of the median vary a lot and there were no whiskers. This indicate that the minimum and maximum values lied in the interquartile range. However, the interquartile range value was too high. This implies that the SOFA score for the worst predicted patient had a high standard deviation and was very variable.

8 Discussion

In this thesis, we presented a data-driven tandem method for predicting septic shock in sepsis patients by discovering patterns in commonly available large-scale vital signs' time series data. In addition to the vital signs, the SOFA score is highly indicative of the severity of sepsis and highly related to vital signs. However, the calculation of SOFA score requires laboratory data such as lactate, blood potassium level (PaO2), and fraction of inspired oxygen (FiO2), etc. The laboratory data is not immediately available after the laboratory tests and can further delay the prediction of septic shock. Our strategy is to leverage big amounts of historical clinical time series data to train a ML model to predict SOFA score based on vital signs. Thus having only vital signs as an input, we predict an approximate value for SOFA score and use it together with other vital signs features to classify sepsis patients into shock and non-shock patients. Hence, we performed multilinear regression using an XGBoost model to obtain the SOFA score from the commonly available vital signs time series data. We predicted the occurrence of septic shock by discovering patterns in large-scale vital signs and SOFA score time series data recorded between sepsis onset time and septic shock onset time. We used a data-driven feature engineering method of a sliding window and performed sequence classification using an XGBoost model to discover time series data patterns and predict septic shock with a sensitivity of 83.56 %, specificity of 65.20 %, precision of 68.6 %, F-score of 75.34 %, and balanced accuracy of 74.38 %. On analyzing the vital signs and SOFA score data a few hours before shock onset time for the best and worst predicted patients, we observed that less variability in vital signs such as heart rate, respiratory rate, blood oxygen saturation, mean, systolic, and diastolic arterial blood pressure, temperature, and SOFA score results in a precise and accurate prediction of septic shock. High variability in data leads to increase in heterogeneity and thus, making it difficult to discover patterns in such a highly heterogeneous data. As a result, the prediction of septic shock is not efficient.

8.1 Limitations

This thesis suffers from some practical limitations commonly encountered in computer-aided decision support in the healthcare domain [56]. These limitations are the unavailability of sufficient data, imbalanced classes, and missing data.

8.1.1 Insufficient Data

Matching the real-time data in MIMIC-III waveform dataset with the patients in MIMIC-III clinical database is an ongoing task [4]. Therefore, the MIMIC-III Waveform Database Matched Subset contains real-time data for only a few patients in the MIMIC-III clinical database, which resulted in a considerable reduction in the number of sepsis patients in our final cohort. Futhermore, excluding patients with more than 20% missing data also reduced the size of our cohort.

8.1.2 Unavailability of Sepsis Related Data

The sepsis and septic patients were not labeled in the MIMIC-III clinical database according to the sepsis-3 criteria [61]. Additionally, the MIMIC-III clinical database does not contain information about the sepsis onset time, septic shock onset time, and SOFA scores, which is highly relevant to our use case of predicting septic shock in sepsis patients. Therefore, we had to identify the sepsis and septic shock patients and calculate their respective SOFA scores, sepsis and shock onset times using the sepsis-3 criteria as mentioned in the sections 5.3 and 5.4. Moreover, we had no data to cross validate these patients and their sepsis and septic shock onset times.

8.1.3 Imbalanced Classes

We had more non-shock patients as compared to the number of shock patients, resulting in imbalanced classes and we used SMOTE to overcome this problem. However, SMOTE does not consider the neighboring samples from other classes which can lead to overlapping of classes and additional noise in the data. Also, SMOTE is not very effective for high-dimensional data like the multidimensional time series data used in this thesis [54]. To overcome the limitations of SMOTE and problem of imbalanced classes, we can combine SMOTE oversampling with the random undersampling or K-Medoids undersampling to reduce the bias or only apply K-Medoids undersampling technique. Nitesh Chawla et al. [54] stated that when SMOTE method is combined with random undersampling, results in improved performance. Rashmi Dubey et al. [26] studied random undersampling, random oversampling, K-Medoids based undersampling, and SMOTE oversampling techniques in Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. They also experimented with different rates of under and over sampling techniques and concluded that K-Medoids undersampling method performed the best.

8.1.4 Missing Data

The time series data in MIMIC-III Waveform Database Matched Subset also contains large gaps of 3-4 hours on average, implying a large amount of missing data. In medical domain, it is often difficult to understand and interpret the reason for unavailability of data without a domain expert. We can also consider modern imputation methods for handling missing values in future. Recent studies have shown that the performance of the prediction model can be improved by applying advanced imputation techniques [35][36]. J. Ghos et al. [35] [36] applied mean / median imputation , singular value based imputation (SVD), Probabilistic principal component analysis (PPCA), Bayesian principal component analysis (BPCA), and K-nearest neighbors (KNN) as missing value imputation methods on missing data of sepsis patients to predict septic shock.

8.1.5 Sequence Labeling

The process of labeling the data in supervised classification plays an important role in performance of the classification model. Accurately labeled data highlights the data features or characteristics that can be analyzed to discover patterns and predict the target class. While performing experiments, we also observed that the performance of sequence classification models is very much dependent on the way the sequences are labelled in the first place. The classification experiment described in the section 7.2.3.4, i.e. labeling all the generated sequences according to the shock or non-shock condition performed the best as compared to the other experiments with labeling according to septic shock onset time and changes in SOFA scores. Hence, considering various other factors such as initial fluid resuscitation time, patient's ICU discharge, or death time for generating time series sequences might further improve the prediction of septic shock in sepsis patients and their corresponding labels. We can also consider multilabel settings, for example, we can use the initial fluid resuscitation time and label only the sequences of shock patient after the initial fluid resuscitation time as *'shock'*. For shock patients that died in the ICU, we can also consider the sequences a few hours before the death and for the shock patients that recovered, we can consider sequences a few hours before the ICU discharge time because such a data will exhibit clear patterns of shock or non-shock condition.

9 Conclusion

This thesis demonstrates a method for predicting golden-hour diseases such as septic shock by applying data driven ML techniques for pattern discovery in large-scale data continuously generated by bedside monitors. We used a tandem method of multilinear regression and feature-based sequence classification on time series vital signs data to predict the occurrence of septic shock in ICU patients. Using multilinear regression, we predict the SOFA scores from vital signs time series data and then use this predicted SOFA scores with vital signs to predict septic shock. Thus, we provide the basis for analyzing large-scale time series of the commonly available physiological data and detecting patterns in them using ML to prevent critical clinical conditions in ICU such as septic shock. Such a method can help predict the occurrence of septic shock in sepsis patients before the shock onset without relying on lab data or having a clinician in the loop. It also ensures appropriate and timely treatment to prevent septic shock and, consequently, decrease sepsis patients' mortality risk.

9.1 Observations

After analyzing the results of our classification model, we observe that the vital signs and SOFA scores of patients in the same group, i.e. patients that progressed into septic shock, had high standard deviations leading to high variability in their vital signs and SOFA scores. High variability in data leads to an increase in heterogeneity, making it challenging to discover patterns in such highly heterogeneous data. As a result, the prediction of septic shock was not efficient in patients with highly variable vital signs and SOFA scores. Therefore, we expect that the group of sepsis patients that progressed into septic shock consisted of subgroups that we did not account for, eventually due to a lack of extended data, such as comorbid conditions, drug usage, etc., which could have concealed the differences we wanted to assess between the shock and non-shock groups; yet this can be verified by additional experiments. Additionally, the improper assignment of sliding windows could have introduced unwanted heterogeneity, which is not indicative of true differences between the shock and non-shock groups.

Moreover, this thesis suffers from limitations such as imbalanced classes, insufficient data, unavailability of labels for sepsis and septic shock, and missing data in real-time vital signs. We tried to overcome these by using methods such as last observation carry forward (LOCF) and SMOTE techniques for imputing missing values and balancing the imbalanced classes, respectively. During the prediction process, we also perceived that its performance was a lot reliant on the way sequences were labeled in the first place. Hence, we also performed various experiments for sequence labeling, such as binary labeling ac-

cording to septic shock onset time, binary labeling according to the shock or non-shock condition, Binary and multi labeling according to the changes in SOFA scores.

9.2 Future Work

We can improve the performance of our proposed method of pattern discovery and prediction of septic shock in sepsis patients in the future. We can apply other missing value imputation techniques such as mean/median imputation, singular value based imputation (SVD), Probabilistic principal component analysis (PPCA), Bayesian principal component analysis (BPCA), and K-nearest neighbors (KNN). We can implement methods such as SMOTE oversampling with the random undersampling or K-Medoids undersampling to reduce the bias or only apply K-Medoids undersampling techniques to balance the classes. Generating the time series sequences and their respective labels according to the factors such as initial fluid resuscitation time, patient's ICU discharge, or death time can further improve the performance of prediction of septic shock in the future.

References

[1] Github MIMIC code repository. https://github.com/MIT-LCP/mimic-code/. Online; Last accessed on 18 November 2020.

[2] MIMIC-III clinical dataset. https://mimic.physionet.org/. Online; Last accessed on 18 November 2020.

[3] MIMIC-III waveform database. https://archive.physionet.org/physiobank/database/mimic3wdb/. Online; Last accessed on 18 November 2020.

[4] MIMIC-III waveform database matched subset. https://archive.physionet.org/physiobank/database/mimic3wdb/matched/. Online; Last accessed on 18 November 2020.

[5] Python's imbalanced-learn api for random undersampling. https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.under_sampling.RandomUnderSampler.html. Online; Last accessed on 18 November 2020.

[6] Python's imbalanced-learn API for SMOTE. https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html. Online; Last accessed on 18 November 2020.

[7] Python's pandas.dataframe.corr function. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html. Online; Last accessed on 18 November 2020.

[8] Python's scipy.stats.entropy function. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html/. Online; Last accessed on 18 November 2020.

[9] Requesting access to MIMIC-III clinical database. https://mimic.physionet.org/gettingstarted/access. Online; Last accessed on 18 November 2020.

[10] Ratnadip Adhikari and Agrawal R, K. *An Introductory Study on Time Series Modeling and Forecasting*. John Wiley Sons, 2013.

[11] Sudharsan Asaithambi. Correlation coefficient: Simple definition, formula, easy steps. https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula. Online; Last accessed on 18 November 2020.

[12] Sudharsan Asaithambi. Why, how and when to apply feature selection. https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection. Online; Last accessed on 18 November 2020.

[13] Dr. Manjiri Bakre. 5 ways machine learning is redefining healthcare. http://entm.ag/zp71. Online; Last accessed on 18 November 2020.

[14] Omar Bellorin, Abraham Abdemur, Iswanto Sucandy, Samuel Szomstein, and Raul J. Rosenthal. Understanding the significance, reasons and patterns of abnormal vital signs after gastric bypass for morbid obesity. *Obesity Surgery*, 21(6):707–713, June 2010. https://doi.org/10.1007/s11695-010-0221-0.

[15] Roger C. Bone, Robert A. Balk, Frank B. Cerra, R. Phillip Dellinger, Alan M. Fein, William A. Knaus, Roland M.H. Schein, and William J. Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, June 1992. https://doi.org/10.1378/chest.101.6.1644.

[16] Tianqi Chen and Carlos Guestrin. XGBoost. August 2016. https://doi.org/10.1145/2939672.2939785.

[17] Yueguo Chen, Mario A. Nascimento, Beng Chin Ooi, and Anthony K. H. Tung. SpADe: On shape-based pattern detection in streaming time series. *abc*, April 2007. https://doi.org/10.1109/icde.2007.367924.

[18] Jui-Sheng Chou and Ngoc-Tri Ngo. Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Applied Energy*, 177:751–770, September 2016. https://doi.org/10.1016/j.apenergy.2016.05.074.

[19] Kwetishe Joro Danjuma. Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. *CoRR*, abs/1504.04646, 2015. http://arxiv.org/abs/1504.04646.

[20] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1), June 2019. https://doi.org/10.1186/s40537-019-0217-0.

[21] Jessie S. Davis, Jared A. Johns, David J. Olvera, Allen C. Wolfe, Alin Gragossian, Eliana M. Rees, Edward A. Pillar, and Daniel P. Davis. Vital sign patterns before shock-related cardiopulmonary arrest. *Resuscitation*, 139:337–342, June 2019. https://doi.org/10.1016/j.resuscitation.2019.03.028.

[22] R. P. Dellinger, , Mitchell M. Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M. Opal, Jonathan E. Sevransky, Charles L. Sprung, Ivor S. Douglas, Roman Jaeschke, Tiffany M. Osborn, Mark E. Nunnally, Sean R. Townsend, Konrad Reinhart, Ruth M. Kleinpell, Derek C. Angus, Clifford S. Deutschman, Flavia R. Machado, Gordon D. Rubenfeld, Steven Webb, Richard J. Beale, Jean-Louis Vincent, and Rui Moreno. Surviving sepsis campaign: International guidelines for management of severe sepsis and septic shock, 2012. *Intensive Care Medicine*, 39(2):165–228, 2013. https://doi.org/10.1007/s00134-012-2769-8.

[23] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D Feldman, Chris Barton, David J Wales, and Ritankar Das. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Medical Informatics*, 4(3):e28, September 2016. https://doi.org/10.2196/medinform.5909.

[24] Rian Dolphin. A comprehensive introduction to LSTM. https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9. Online; Last accessed on 18 November 2020.

[25] C.L. Downey, W. Tahir, R. Randell, J.M. Brown, and D.G. Jayne. Strengths and limitations of early warning scores: A systematic review and narrative synthesis. *International Journal of Nursing Studies*, 76:106–119, November 2017. https://doi.org/10.1016/j.ijnurstu.2017.09.003.

[26] Rashmi Dubey, Jiayu Zhou, Yalin Wang, Paul M. Thompson, and Jieping Ye. Analysis of sampling techniques for imbalanced data: An n=648 ADNI study. *NeuroImage*, 87:220–241, February 2014. https://doi.org/10.1016/j.neuroimage.2013.10.005.

[27] Josef Fagerström, Magnus Bång, Daniel Wilhelms, and Michelle S. Chew. LiSep LSTM: A machine learning algorithm for early detection of septic shock. *Scientific Reports*, 9(1), October 2019. https://doi.org/10.1038/s41598-019-51219-4.

[28] C. Fleischmann-Struzek, A. Mikolajetz, D. Schwarzkopf, J. Cohen, C. S. Hartog, M. Pletz, P. Gastmeier, and K. Reinhart. Challenges in assessing the burden of sepsis and understanding the inequalities of sepsis outcomes between national health systems: secular trends in sepsis and infection incidence and mortality in germany. *Intensive Care Medicine*, 44(11):1826–1835, October 2018. https://doi.org/10.1007/s00134-018-5377-4.

[29] Abdur Rahim Mohammad Forkan, Ibrahim Khalil, and Mohammed Atiquzzaman. ViSiBiD: A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. *Computer Networks*, 113:244–257, February 2017. https://doi.org/10.1016/j.comnet.2016.12.019.

[30] V. García, R. A. Mollineda, and J. S. Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. *abc*, pages 441–448, 2009. https://doi.org/10.1007/978-3-642-02172-5_57.

[31] Princeton university German Rodriguez. Generalized linear models. https://data.princeton.edu/wws509/notes/a2s2. Online; Last accessed on 18 November 2020.

[32] Shameek Ghosh, Jinyan Li, Longbing Cao, and Kotagiri Ramamohanarao. Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. *Journal of Biomedical Informatics*, 66:19–31, February 2017. https://doi.org/10.1016/j.jbi.2016.12.010.

[33] Hamidreza Hakimdavoodi and Maryam Amirmazlghani. Maximum likelihood estimation of generalized linear models with generalized gaussian residuals. December 2016. https://doi.org/10.1109/icspis.2016.7869893.

[34] Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, August 2015. https://doi.org/10.1126/scitranslmed.aab3719.

[35] Joyce C. Ho, Cheng H. Lee, and Joydeep Ghosh. Septic shock prediction for patients with missing data. *ACM Transactions on Management Information Systems*, 5(1):1–15, April 2014. https://doi.org/10.1145/2591676.

[36] Joyce C. Ho, Cheng H. Lee, and Joydeeph Ghosh. Imputation-enhanced prediction of septic shock in icu patients. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.469.8616. Online; Last accessed on 18 November 2020.

[37] Lauren Jacobs and Hector Wong. Sepsis subclasses. *Pediatric Critical Care Medicine*, 18(6):591–592, June 2017. https://doi.org/10.1097/pcc.0000000000001132.

[38] Alistair E. W. Johnson, Jerome Aboab, Jesse D. Raffa, Tom J. Pollard, Rodrigo O. Deliberato, Leo A. Celi, and David J. Stone. A comparative analysis of sepsis identification methods in an electronic database. *Critical Care Medicine*, 46(4):494–499, 2018. https://doi.org/10.1097/ccm.0000000000002965.

[39] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), May 2016. https://doi.org/10.1038/sdata.2016.35.

[40] Weaver Kathleen F, Vanessa Morales, Dunn Sarah L, and et al. *An Introduction to Statistical Analysis in Research: With Applications in the Biological and Life Sciences*. John Wiley Sons, 2017.

[41] Tsuneaki Kenzaka, Masanobu Okayama, Shigehiro Kuroki, Miho Fukui, Shinsuke Yahata, Hiroki Hayashi, Akihito Kitao, Daisuke Sugiyama, Eiji Kajii, and Masayoshi Hashimoto. Importance of vital signs to the early diagnosis and severity of sepsis: Association between vital signs and sequential organ failure assessment score in patients with sepsis. *Internal Medicine*, 51(8):871–876, 2012. https://doi.org/10.2169/internalmedicine.51.6951.

[42] Tsuneaki Kenzaka, Masanobu Okayama, Shigehiro Kuroki, Miho Fukui, Shinsuke Yahata, Hiroki Hayashi, Akihito Kitao, Daisuke Sugiyama, Eiji Kajii, and Masayoshi Hashimoto. Importance of vital signs to the early diagnosis and severity of sepsis: Association between vital signs and sequential organ failure assessment score in patients with sepsis. *Internal Medicine*, 51(8):871–876, 2012. https://doi.org/10.2169/internalmedicine.51.6951.

[43] Ku Ruhana Ku-Mahamud, Norharyani Zakaria, Norliza Katuk, and Mohamad Shbier. Flood pattern detection using sliding window technique. *abc*, 2009. https://doi.org/10.1109/ams.2009.15.

[44] Simon Lambden, Pierre Francois Laterre, Mitchell M. Levy, and Bruno Francois. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23(1), November 2019. https://doi.org/10.1186/s13054-019-2663-7.

[45] Aleksandra Leligdowicz and Michael A. Matthay. Heterogeneity in sepsis: new biological evidence with clinical applications. *Critical Care*, 23(1), March 2019. https://doi.org/10.1186/s13054-019-2372-2.

[46] Mitchell M. Levy, Mitchell P. Fink, John C. Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M. Opal, Jean-Louis Vincent, and Graham Ramsay. 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. *Critical Care Medicine*, 31(4):1250–1256, April 2003. https://doi.org/10.1097/01.ccm.0000050454.01978.3b.

[47] C. Zachary Lipton, C. David Kale, charles elkan, and randall wetzell. Learning to diagnose with lstm recurrent neural networks. *international conference on learning representations*, 2015. http://arxiv.org/abs/1511.03677.

[48] Ran Liu, Joseph L. Greenstein, Stephen J. Granite, James C. Fackler, Melania M. Bembea, Sridevi V. Sarma, and Raimond L. Winslow. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Scientific Reports*, 9(1), April 2019. https://doi.org/10.1038/s41598-019-42637-5.

[49] Flavia Ribeiro Machado, Murillo Santucci Cesar de Assunção, Alexandre Biasi Cavalcanti, André Miguel Japiassú, Luciano Cesar Pontes de Azevedo, and Mirella Cristine Oliveira. Getting a consensus: advantages and disadvantages of sepsis 3 in the context of middle-income settings. *Revista Brasileira de Terapia Intensiva*, 28(4), 2016. https://doi.org/10.5935/0103-507x.20160068.

[50] Qingqing Mao, Melissa Jay, Jana L Hoffman, Jacob Calvert, Christopher Barton, David Shimabukuro, Lisa Shieh, Uli Chettipally, Grant Fletcher, Yaniv Kerem, Yifan Zhou, and Ritankar Das. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*, 8(1):e017833, January 2018. https://doi.org/10.1136/bmjopen-2017-017833.

[51] Luis Montesinos, Rossana Castaldo, and Leandro Pecchia. On the use of approximate entropy and sample entropy with centre of pressure time-series. *Journal of NeuroEngineering and Rehabilitation*, 15(1), December 2018. https://doi.org/10.1186/s12984-018-0465-9.

[52] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine*, 46(4):547–553, April 2018. https://doi.org/10.1097/ccm.0000000000002936.

[53] Hasan Ogul, Alejandro Baldominos, Tunc Asuroglu, and Ricardo Colomo-Palacios. On computer-aided prognosis of septic shock from vital signs. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 87–92, 2019. https://doi.org/10.1109/cbms.2019.00028.

[54] M. Mostafizur Rahman and D. N. Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, pages 224–228, 2013. https://doi.org/10.7763/ijmlc.2013.v3.307.

[55] Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and*

Circulatory Physiology, 278(6):H2039–H2049, June 2000. https://doi.org/10.1152/ajpheart.2000.278.6.h2039.

[56] Alejandro Rodríguez-González, Jose Emilio Labra-Gayo, Ricardo Colomo-Palacios, Miguel A. Mayer, Juan Miguel Gómez-Berbís, and Angel García-Crespo. SeDeLo: Using semantics and description logics to support aided clinical diagnosis. *Journal of Medical Systems*, 36(4):2471–2481, May 2011. https://doi.org/10.1007/s10916-011-9714-1.

[57] Ryding Sara. The stages of sepsis. https://www.news-medical.net/health/The-Stages-of-Sepsis.aspx. Online; Last accessed on 18 November 2020.

[58] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of clinical criteria for sepsis. *JAMA*, 315(8):762, February 2016. https://doi.org/10.1001/jama.2016.0288.

[59] Supreeth P Shashikumar, Qiao Li, Gari D Clifford, and Shamim Nemati. Multiscale network representation of physiological time series for early prediction of sepsis. *Physiological Measurement*, 38(12):2235–2248, November 2017. https://doi.org/10.1088/1361-6579/aa9772.

[60] Supreeth P. Shashikumar, Matthew D. Stanley, Ismail Sadiq, Qiao Li, Andre Holder, Gari D. Clifford, and Shamim Nemati. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *Journal of Electrocardiology*, 50(6):739–743, November 2017. https://doi.org/10.1016/j.jelectrocard.2017.08.013.

[61] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801, February 2016. https://doi.org/10.1001/jama.2016.0287.

[62] Kamaldeep Singh. How to improve class imbalance using class weights in machine learning. https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/. Online; Last accessed on 18 November 2020.

[63] Dimitri P. Solomatine and Avi Ostfeld. Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 10(1):3–22, January 2008. https://doi.org/10.2166/hydro.2008.015.

[64] Z. S. Spakovszky. A statistical definition of entropy. https://web.mit.edu/16.unified/www/FALL/thermodynamics/notes/node56.html. Online; Last accessed on 18 November 2020.

[65] Youqiang Sun, Jiuyong Li, Jixue Liu, Bingyu Sun, and Christopher Chow. An improvement of symbolic aggregate approximation distance measure for time series. *Neurocomputing*, 138:189–198, August 2014. https://doi.org/10.1016/j.neucom.2014.01.045.

[66] Timothy E Sweeney, Oliver Liesenfeld, and Larissa May. Diagnosis of bacterial sepsis: why are tests for bacteremia not sufficient? *Expert Review of Molecular Diagnostics*, 19(11):959–962, August 2019. https://doi.org/10.1080/14737159.2019.1660644.

[67] Hongcheng Tian, Jianfang Zhou, Li Weng, Xiaoyun Hu, Jinmin Peng, Chunyao Wang, Wei Jiang, Xueping Du, Xiuming Xi, Youzhong An, Meili Duan, and Bin Du. Accuracy of qSOFA for the diagnosis of sepsis-3: a secondary analysis of a population-based cohort study. *Journal of Thoracic Disease*, 11(5):2034–2042, May 2019. https://doi.org/10.21037/jtd.2019.04.90.

[68] Majid Vafaeipour, Omid Rahbari, Marc A. Rosen, Farivar Fazelpour, and Pooyandeh Ansarirad. Application of sliding window technique for prediction of wind velocity time series. *International Journal of Energy and Environmental Engineering*, 5(2-3), May 2014. https://doi.org/10.1007/s40095-014-0105-5.

[69] Christopher R Yee, Niven R Narain, Viatcheslav R Akmaev, and Vijetha Vemulapalli. A data-driven approach to predicting septic shock in the intensive care unit. *Biomedical Informatics Insights*, 11:117822261988514, January 2019. https://doi.org/10.1177/1178222619885147.

[70] Zhihong Zhang and Edwin R. Hancock. Mutual information criteria for feature selection. pages 235–249, 2011. https://doi.org/10.1007/978-3-642-24471-1_17.