MONASH University

Early Detection of Sepsis in Neonatal and Adult Patients Using Machine Learning Approaches - Aiding Diagnosis for Clinicians

Yifei Hu (*Bachelor of Engineering, Dalian University of Technology*; *Master of Computer Science, Southeast University*; *Master of Information Technology Systems, Monash University*)

A thesis submitted for the degree of *Doctor of Philosophy* at
Monash University in *2021*
Faculty of Information Technology

Copyright notice

Abstract

Sepsis is a life-threatening condition which causes millions of deaths every year. Since earlier detection and treatment can reduce mortality, developing an accurate and efficient computer-aided mechanism to predict the onset of sepsis demands urgent action. However, prediction of sepsis is not easy to implement since it is a highly variable disease process, making the work challenging. Previous researchers have mainly focused on investigating indicative biomarkers from a medical point of view or developing a scoring system based on logistic functions to help doctors with their diagnosis. As information technology researchers, we pay more attention to exploring the potential of machine learning (ML) methods, especially deep learning, in predicting impending sepsis before the clinicians' suspicion. This thesis makes three main contributions to the domain of sepsis detection. We first present a systematic literature review (SLR) that gives a comprehensive overview of the current progress of computer-aided sepsis detection. We proposed and implemented a unified sepsis detection framework to achieve the task. The framework defines a complete workflow of performing early detection of sepsis for both adults and infants, and provides a guidance for researchers from data collection to reflective evaluation. Finally, directed by action research methodology, we explore and verify the feasibility of multiple models and algorithms, including classic machine learning models, advanced deep learning models, and multi-instance learning approaches in three iterative research cycles.

Using the proposed unified sepsis detection framework, we have investigated several intelligent models and examined the feasibility of incorporating them into the sepsis diagnosis process. Despite the limitations of current data collection systems in the hospital, with proper preprocessing steps, collected data can still be used to develop machine learning models. Two additional public datasets are included to expand the training samples. Multiple models have been demonstrated to produce acceptable results, of which the Long Short-Term Memory (LSTM) model with a fully-connected encoder has improved detection time for sepsis by up to six hours and achieved a promising AUC of 0.95. With its superior performance compared to many of the methods reviewed in the SLR, it could be used as a decision support tool for clinicians.

Acknowledgements

First of all, I would like to express my gratitude to my supervisors: Assoc. Prof. Vincent C.S. Lee, Dr. Kenneth Tan for their support and guidance throughout my PhD career. I learned a lot from them, ranging from the knowledge about data science and sepsis diagnosis to academic skills such as writing and presentation, especially for their consistent support for the thesis writing. I enjoyed the wonderful time I worked with them, and their experience and advice are invaluable for my study and my future career.

I am deeply grateful for the help from my panel members Assoc. Prof. Guido Tack, Dr. Lan Du and Dr. Yuanfang Li. They have provided insightful comments on my research as well as much useful feedback to help me finish my research project and thesis.

Meanwhile, I would take this chance to acknowledge Faculty of IT (FIT), Monash Institute of Medical Engineering (MIME) for their generous financial and resource support in the past four years, as well as Monash Children's Hospital and Monash Health for their close collaboration, allowing me to have access to valuable real-time NICU data which is a really important part of my resarch.

Furthermore, I appreciate Ms Danette Deriane and all the other staff from FIT Graduate Team for their valuable time to help me arrange compulsory activities (commencement, course, milestones) and answer my enquiries before and after I started my study.

I also want to thank Ms Julie Holden and her GSAS classes for providing an excellent training opportunity for my academic communication skills.

Finally, I would like to express my deepest appreciation to my parents and my partner Jiaman, who give unconditional love, encouragement and support not only in my four-year PhD career but also in many other things I have done in my life.

Publications during enrolment

Publication(s) included in this thesis:

1. Y. Hu, V. C. S. Lee, and K. Tan, ’An Application of Convolutional Neural Networks for the Early Detection of Late-onset Neonatal Sepsis’, in 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, Jul. 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8851683. (Core Rank A)
2. Y. Hu, V. C. S. Lee, and K. Tan, ’Prediction of clinicians’ treatment in preterm infants with suspected late-onset sepsis An ML approach’, in 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), May 2018, pp. 1177-1182, doi: 10.1109/ICIEA.2018.8397888.

Publication(s) under review:

1. Y. Hu, V. C. S. Lee, and K. Tan, ’Systematic Review of Sepsis Detection Using AI Methods’, submitted to Artificial Intelligence in Medicine, Dec 2020.

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature:

Print Name: Yifei Hu

Date: 28/06/2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of the Clinical Problem

Sepsis is an extreme response in our body to an infection, which is likely to be fatal in some cases, according to the CDC of the United States, and it can affect patients of all ages [1]. We explore both conditions in this thesis, and examine how the proposed methods perform in each. Sepsis is a highly prevalent condition that accounts for 10% of adults admissions to an intensive care unit (ICU). Sepsis, along with severe sepsis and septic shock is associated with a 10% to 20% in-hospital mortality rate [2, 3]. The World Health Organization estimates that more than six million people die of sepsis annually, and many of these deaths are preventable. Various definitions of sepsis have been made to guide clinicians' diagnosis, for example, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) [4] created in 2016 replaced the systemic inflammatory response syndrome (SIRS) due to its poor sensitivity and specificity. Other tools like Sepsis-related Organ Failure Assessment (SOFA)[5] score and quick SOFA (qSOFA) [4] were designed to describe quantitatively and objectively the degree of dysfunction over time for sepsis patients.

Neonatal sepsis (NS) refers to bloodstream infections that occur during the neonatal period, i.e. the first 28 days for termed babies and up to 4 weeks after the expected delivery date for preterms. The prevalence of NS has a major impact on the mortality and morbidity of newborn babies [6]. Recently it was reported that in developing countries, NS causes about 1.6 million neonatal deaths every year [7]. Preterm infants, especially very low birth weight (VLBW; <1500g birthweight) infants can more easily infected because of their immature immune system and prolonged hospitalisation, contributing to the higher mortality [8]. The total incidence of neonatal sepsis is one to eight per 1000 live births, and up to 20% of neonatal intensive care unit (NICU) admission [9, 10]. Neonatal sepsis can be generally classified into early- or late-onset, depending on whether it occurs within 48 hours [9], some other reports define it as 24 or 72 [11]. Early-onset sepsis (EOS) has an incidence of 0.9/1000 live birth and 9/1000 NICU admission [6], and it is predominantly due to microorganisms from the birth canal, and over 80% of these related microorganisms are group B streptococcus (GBS) and Gram negative bacteria [11]. However, late-onset sepsis (LOS) is caused mainly by nosocomial bacteria, and approximately 50% of late-onset cases are associated with coagulase-negative staphylococci (CoNS). The incidence of late-onset sepsis is around 8/1000 live births and 70/1000 NICU admission [6], though it varies between 11% and 53% worldwide [12]. The risk is inversely correlated with birth weight and gestational age [13–16], preterm babies that are less than 2500g or delivered before 37 weeks have the greatest risk [6]. Mortality of late-onset sepsis ranges from 5 to 50% worldwide [17] and in Australia it is about 6% [18].

The absence of a genuine gold standard in the diagnosis of sepsis is one of the most important concerns and it leads to inconsistency among clinicians and researchers [19]. The most widely used gold standard at the moment is the blood culture test [8, 20, 21], and based on the result sepsis is defined as either culture-proven sepsis, in which case the result of blood culture test is positive, or clinical sepsis where other laboratory parameters make the clinicians believe the presence of sepsis but the blood culture test output is negative [8]. Apart from blood culture tests, there are other standards for diagnosing sepsis. It is reasonable to suspect sepsis when an adult patient meets at least two of the following clinical criteria: a respiratory rate of 22/min or higher, altered mentation, and a systolic blood pressure of 100 mm Hg [4]. According to the paediatric consensus definition for sepsis [22], established in 2005, evidence of SIRS must exist as a prerequisite to meeting sepsis criteria. The SIRS requires either an abnormal WBC count (an increased or decreased total WBC for an individual's age or a concentration of more than 10% immature neutrophils) or an abnormal core temperature ($> 38.5°C$ or $< 36°C$). Although blood culture is considered as the gold standard, it has several disadvantages. First, it is a tedious and cumbersome process which significantly influence the management of patients. Blood culture test requires 24-48 hours before results are generated, and it will delay the treatment if clinicians wait until the test results come out [21]. Second, the specificity of blood culture remains debatable, and it has too many false negative outcomes [23, 24]. For clinical sepsis diagnosis, it also mainly depends on the experience of the doctor. No specific signs will indicate the presence of neonatal sepsis and sometimes many non-specific symptoms could possibly occur due to sepsis [9]. For each hour treatment initiation is delayed after diagnosis, sepsis-related mortality increases by approximately 8% [25]. Third, for newborns, especially very-low-birth-weight preterm infants, blood collection is restricted to a single sample with a minimal volume (1 ml), hence the limitation of the blood volume could further hinder the pathogen capture which usually leads to false negative results [26, 27].

Currently, the commonly used approach to neonatal sepsis is the administration of empirical antibiotic therapy [8, 9]. To prevent deterioration, clinicians are encouraged to use antibiotics before the result of the blood culture test comes out. The excessive use of antibiotics can result in antibiotic resistance, predisposing to fungal infection, necrotising enterocolitis (NEC) and even death [28]. The dilemma is that if clinicians wait for the blood culture results, the patients may deteriorate during that time, and lose the best opportunity for treatment. Under this circumstance, an early detection system is urgently need. One of the motivations of our research is to solve this dilemma, make antibiotic administration targeted at the right patients, and avoid the overuse of antibiotics. Furthermore, staff fatigue due to current practice workload is one of the common situation that need to deal with, and monitoring all kinds of vital signs from tens of cots are demanding on staff time and makes the process vulnerable to human errors. To facilitate the diagnosis process of sepsis for clinicians, many groups have developed early detection methods. However, most of them were score-based or rule-based, focusing on some key parameters from a medical viewpoint or merely exploiting a simple logistic regression model to establish a correlation between certain measurements and potential future sepsis onset. Recent advances in computation technology and expansion of data capacity caused by an increasing number of computational devices, popularisation of Internet and better network conditions allowed Artificial Intelligence (AI) to be explosively developed since the beginning of the twenty first century. Multidisciplinary research drew more attention in academic society, where AI technology has been widely used in many cross-subject fields, providing novel and unprecedented solutions. As a rule of thumb, AI-based models are built on top of large datasets, in order to exploit the full potential of AI technology to uncover useful, though hidden, information that was previously unknown within the datasets. In this regard, another issue we have to properly address is to collect sufficient amounts of relevant data, and make sure it covers as many variables as possible. In this thesis, we aimed to explore the robust capability of more advanced machine learning and deep learning models and verify the feasibility of application in the medical field, and, to be more specific, the early detection of sepsis. Incorporating intelligent models into the current clinical workflow could potentially make the diagnosis of sepsis more efficient and effective for clinicians.

1.2 Research Objective and Questions

The overall research project's aim have been divided into sections, based on the following research objectives (RO) and questions (RQ).

1. **RO-1**: Design an efficient data acquisition scheme.

   Different data is stored in multiple sub-systems in the hospital, however, due to the initial design purpose, these systems often face a number of challenges in meeting research and confidentiality requirements, sustainable storage requirements and an inability to export data automatically.

   - **RQ-1**: How should the vital signs be collected from bedside monitors in NICU?

     As the most important type of clinical data, vital signs are ideally suited to machine learning algorithms since they are continuously accumulated and thus comprise a substantial proportion of the dataset. Normally, vital signs are surveilled by bedside monitors linked to patients by attached sensors, but the raw data collected are not processed or stored. Depending on the different standards in each hospital, vital signs are reserved only for a reasonably short time, e.g. 10 hours or 24 hours. Therefore, it is essential to effectively collect and store each patient's vital signs in the NICU for further use.

2. **RO-2**: Investigate the relations between physiological parameters and sepsis, and find critical ones that are most related to sepsis.

   - **RQ-2**: How many physiological parameters are available to researchers?

     Due to the need to protect the privacy of patients' medical data, we are restricted in our access to them all. Additionally, numerous types of laboratory test results are maintained in distinct subsystems, complicating our collection efforts when we attempt to extend the dimension of the data. The first issue we must address is coordinating with the hospital in order to collect as much data as possible while obtaining all relevant permits. It is also vital to investigate the available properties for publicly accessible online datasets in order to maintain synchronisation.

   - **RQ-3**: Which critical physiological parameters can predict sepsis before it occurs?

     Data is the fundamental basis of the early detection system. Considering previous studies and pathology knowledge, we need to analyse all the variables we can obtain and determine which ones are associated with sepsis. By selecting relevant parameters and excluding those that have little relation to sepsis, we can develop effective feature sets.

3. **RO-3**: Design, develop and implement suitable algorithms for the clinical data, and apply adjustments to improve the performance.

    - **RQ-4**: Can existing methods in literatures fulfil the current requirement of early detection?

      Many studies have proposed different methods for detecting and diagnosing neonatal sepsis, including traditional biomarkers, statistical techniques, and some machine learning methods. It is necessary to review and evaluate existing work to determine whether it meets the current demand for early detection.

    - **RQ-5**: Is it possible to adjust existing methods to improve their performance?

      As in some of the research, some preliminary machine learning models have already been applied to sepsis prediction, we should examine those results and determine whether appropriate adjustments could improve their performance before incorporating new methods and algorithms that have never been tried before. If they are promising, we will add new features, add new techniques, or merge multiple methods, to make the process more effective.

    - **RQ-6**: Is it possible for any new approach to have a better performance for this early detection task?

      The analysis of algorithms forms a major part of this research. Our approach to achieving this goal relies primarily on cutting-edge machine learning techniques, particularly deep learning. Different types of algorithms are examined, such as classic machine learning models, convolutional neural networks, and recurrent neural networks.

1.3 Research Contributions

This research explored a significant depth of knowledge by first systematically reviewing the existing state of the field, then designing and developing the data collection schema both from hospital and public medical databases. We worked closely with medical domain experts and users using an action research framework embedded in the research methodology. As a result, we were able to ensure that the work being done in this thesis was both valid and useful to the medical domain of research, and contributed to further development of the methodology. We conceptualised factors for sepsis in both infants and adults by reviewing and identifying data collected by existing studies and constructed multiple models for early detection of sepsis. These models were developed and refined in conjunction with medical domain experts, and represent the application of multi-discipline technologies. We began this thesis by narrowing down a specific area of research to focus on, and then drilling down into its depths. The results and findings that we have arrived at are rich and have allowed us to identify further issues and areas for future research. Specific contributions to this thesis are described in the following subsections.

Significant parts of this thesis have been published in:

- Y. Hu, V. C. S. Lee, and K. Tan, An Application of Convolutional Neural Networks for the Early Detection of Late-onset Neonatal Sepsis, in 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, Jul. 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8851683. (Core Rank A)
- Y. Hu, V. C. S. Lee, and K. Tan, Prediction of clinicians treatment in preterm infants with suspected late-onset sepsis An ML approach, in 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), May 2018, pp. 1177-1182, doi: 10.1109/ICIEA.2018.8397888.

1.3.1 Contribution to Knowledge

A systematic literature review (SLR) on existing status of sepsis detection. During the course of this thesis, a systematic review of existing research related to sepsis detection is presented, providing insight into the methods that have been utilised and are currently being utilised. There is a comprehensive description of the entire review process, and the results are analysed and discussed in depth. In addition to presenting a guide for anyone who aspires to enter this field, it summarises what predecessors have achieved, and provides directions for future research including the scope of this thesis.

A unified sepsis detection framework for both adults and infants. We proposed the architecture of a unified sepsis detection framework covering all the steps from pre-data collection practice to the analysis and evaluation of the results. Following the proposed framework, one can run through the procedure step by step until having an output of sepsis prediction, regardless of whether the target subjects are adults or infants. Meanwhile, this framework summarises many unique processes that other machine learning projects might not necessarily need to include.

1.3.2 Contribution to practice

Verifying the feasibility of incorporating raw data from ICU to facilitate sepsis prediction. Live data collected from bedside monitors is discarded after 24 hours when rule-based sepsis diagnosis methods are applied. Clinicians and nurses mainly focus on whether any value goes beyond the threshold. Yet with machine learning algorithms involved, information hidden among historical data is revealed by feeding them to a well-trained model after several steps of pre-processing.

Application of multiple types of machine learning based models in predicting sepsis onset. On different datasets, we have examined a variety of classic machine learning models and deep learning models in the detection of sepsis. Furthermore, we have improved the multi-instance learning algorithm by trading the amount of time a prediction can be made in advance, with higher accuracy, and obtained an acceptable outcome.

The novel method of converting multi-variate time series data to image in classification task. The transformation of univariate time series into images can be accomplished in a variety of ways. Two approaches were presented that can be used to produce colour images with multivariate time series, providing further inspiration on how to deal with time series data. In addition, it makes it possible to utilise the remarkable performance of convolutional neural networks on time series data rather than images.

1.4 Thesis Structure

This thesis has seven chapters in total, and is structured as follows: Chapter 1 is the introduction of our research, illustrating the background, motivation of the project, our research objectives and research questions derived from those objectives, and the contribution we have made throughout our research. The next chapter presents a systematic literature review of related works published in the past decade. We examine different sepsis detection approaches, assess the PhysioNet Challenge that is aiming to predict sepsis based on clinical data, summarise the current state of research in this area, and identify research gaps. A description of the research methods used in this thesis is provided in Chapter 3. A unified sepsis detection framework was designed to fill one of the gaps that we identified during the literature review in the context of the action research framework we follow. We analyse datasets from various sources in Chapter 4 before implementing the proposed framework in order to gain insight into the problem. Our chapter 5 and 6 discuss the use of the cyclic research methods of action research, examine the performance of multiple models and algorithms including classical machine learning, deep learning and multi-instance learning methods, respectively, to predict neonatal and general sepsis, and evaluate the results. As directed by the principles of action research, reflection has also been conducted in these chapters. Finally in Chapter 7, we summarise the thesis by concluding our contributions, answering the research questions posed in Chapter 1, and discussing the potential directions for future research.

Chapter 2

Systematic Literature Review

This chapter systematically summarises the literature relevant to this thesis on the topic of sepsis prediction, especially neonatal sepsis. The whole review process is fully described and the results are analysed and discussed in depth.

2.1 Review Definition

In this systematic literature review , our focus is on research into sepsis prediction, especially neonatal sepsis prediction achieved by using artificial intelligence (AI)-based methodologies, such as machine learning and deep learning in the past decade. The related techniques include conventional machine learning models, deep learning models, and other statistical models. Medical approaches like diagnostic models based on clinical rules or workflow on certain types of biomedical metrics are excluded, since they are not within the scope of our research. This systematic literature review has three aims: (1) to identify the current trends of AI-based sepsis prediction, (2) to highlight key challenges researchers have to deal with to increase the credibility of their work and (3) most importantly to introduce a classification framework of existing methods. To achieve these goals, we set the time scope of screening criteria from 2010 to 2020, and all papers are indexed in the following four well-known online databases: IEEE, ACM, ELSEVIER and PUBMED.

2.2 Review Methodology

We adopted Booth et al.'s [29] systematic approach to literature review and followed the three-phase methodology employed by Pourhabibi et al. [30], as depicted in Figure 2.1.

The first phase is ″research definition″ which is identifying the research area, formulating review goals, and defining the research scope, and these are already illustrated in section 2.1.

The second phase is ″research methodology″. The literature search process began with the creation of criteria to determine the articles to include in, or exclude from our analysis. We set up five rules that the article must (1) be published in a peer-reviewed academic journal or conference, (2) be written in English, (3) be published between 2010 and 2020, (4) have

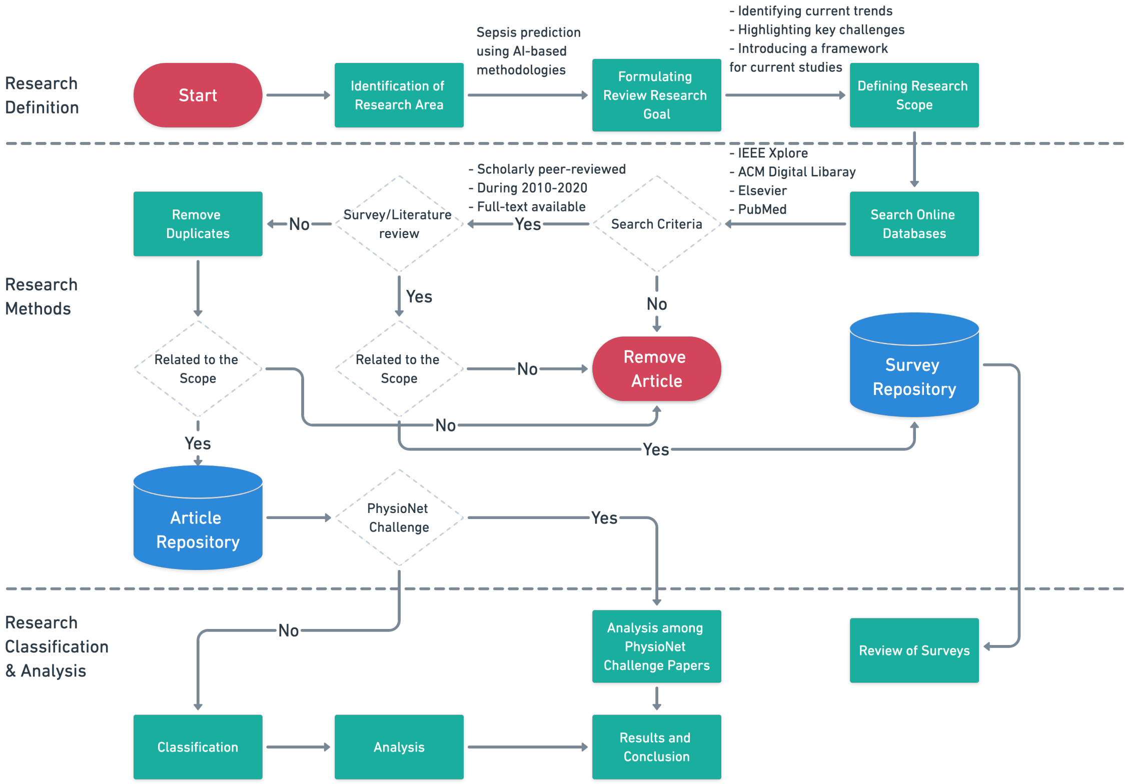


Figure 2.1: Systematic literature review process [29, 30].

its full text available and (5) use AI-related methods to solve the sepsis detection problem. The keywords we input to conduct the search are ”neonatal sepsis”, ”sepsis prediction” and ”sepsis detection” in plain text connected with boolean operator ”OR”, and after the filtering process and the ”Abstract Reading and Skimming”, 102 papers are left in the search results for the final analysis. Note that 41 out of them are from the PhysioNet 2019 challenge, and we will analyse them separately in another section later.

For the final ”classification and analysis” phase, we have posed six guided questions to facilitate the sorting and classification process, which are:

1. What were the study trends and focus?
2. What were the features used to predict sepsis?
3. What were the research datasets?
4. What were the models used to predict sepsis?
5. What were the measures used to evaluate the model performance?
6. What were the contributions and limitations that researchers faced, and possible future directions?

These six questions provide six distinct different levels based on which we can systematically categorise all the available papers under a classification framework. The next section explains the classification framework proposed. Because this is a PhD thesis, all the screening and data extraction process were conducted by me only, not by two or more authors as recommended, but the results have been audited by my supervisors.

2.2.1 Classification Framework

To align with the six guided questions listed in the previous section, the proposed classification framework begins with identifying study trends and focuses. We describe some of the components of the framework individually below.

Prediction Features

The commonly used features in sepsis prediction models can be categorised into three classes: (1) demographic data, (2) vital signs, and (3) laboratory test results. Table 2.1 demonstrates the meaning of three types of features. Actually, in each class, multiple different factors will be chosen to build the input feature vector depending on what kind of dataset and model are used. Besides the features used, we also considered the process the features underwent before being fed to the model, a step called feature preprocessing. Feature preprocessing is an effective way to improve model performance and a variety of measurements may be taken to build up more meaningful and powerful features, depending on the nature of the data being used.

Table 2.1: Comparison of features used in AI models for sepsis prediction tasks.

	demographic data	vital signs	lab test
definition	population based factors	measurements of the body's most basic functions	results of related biochemical test mostly based on blood
selected examples	age, birth weight	heart rate, respiratory rate	C-reactive protein, white blood cell

Research Datasets

Among all the research we have reviewed, most datasets come from either hospitals or public medical databases. Normally, hospital store clinical data of patients admitted in a period of ten years or longer (or longer in infants for 25 years) , depending on jurisdiction and different requirements in each country. The clinical data may be stored as paper documents, scanned electronic copies of original paper documents, or in recent years in electronic medical records (EMR) databases. These records can be the source of retrospective data for developing and validating a prediction system. Physiological data from bedside monitors may not always be stored by hospitals as they require large storage memory capacity. Nevertheless, these types of high-frequency data can be accessed in real-time if appropriate equipment is available or if there is a regulatory requirement to store the data. For public medical database, MIMIC [31] and PhysioNet [32] are two typical ones that have been referenced most.

Prediction Models

As the predictive models will be considered from a technical perspective, only three types of models will be focused on:

- **conventional machine learning models** To distinguish from deep learning models, we call models with less complex structures as conventional machine learning models, including classic ones like logistic regression, support vector machine, and more advanced ones like ensembles and hidden Markov chain.

- **deep learning models** Deep learning models, refer to those with highly complicated structures, normally neural network related. Different from normal shallow neural networks, deep learning models have more layers along with some specially designed structures to capture latent features and patterns. Typical models include convolutional neural networks and recurrent neural networks.

- **other statistical models** Some other models beyond ML and DL.

2.3 Review Classification and Analysis

2.3.1 Overview of Surveys on Sepsis detection

There are several surveys or systematic reviews on sepsis detection [33–36] over the past decade. Since our focus is to review existing work on predicting or early detecting the onset of sepsis with AI-based methods, a paper such as [37] which reviews sepsis mortality prediction with medical biomarkers and [38] which compares the performance of two medical criteria qSOFA and SIRS in the diagnosis of sepsis will not be in the scope of our review.

Laurel [33] has reviewed thirteen studies between January 2005 and January 2015, describing automated detection approaches with potential to detect sepsis or sepsis-related deterioration in real or near-real time, focused on emergency departments and hospitalized neonatal, paediatric, or adult patients. This review primarily reported the data used in the reviewed articles and the comparison on their performance. The most commonly used variables in the sepsis detection algorithm include vital signs and laboratory value criteria and they occurred in all thirteen papers. Another typical feature widely used in the thirteen reviewed papers is sepsis alert which is mentioned in seven out of thirteen papers, although some of them claimed no significant effect of sepsis alerts on patient outcomes [39, 40]. The usage of sepsis alerts was intuitively driven by clinical knowledge, but somehow did not increase the accuracy of the detection algorithm. The limitation of this review is that it did not introduce and compare the method each paper used, making it difficult to evaluate the detection approaches.

In another work [34] presented by Mehanas and Pushpalatha, ten different machine learning based techniques used to predict and detect sepsis were analysed. The authors described each paper in detail, but no comparison was provided. Five techniques used were listed, which included Bayesian Network, Conditional Independent Maps, Kernel Extreme Learning Machine, Chaotic Fruit Fly Optimisation, and Hierarchical Analysis, and most of them are not commonly used machine learning algorithms with certain limitations. Besides, the lack of in-depth analysis makes this review superficial.

Another systematic review was conducted by Alejandro et al [35],targeted at finding papers where computational intelligence is used to predict infections in patients using physiological data as features, which are exactly aligned with our interests. The only bias is that this paper considers the general concept of infection, but our focus is on sepsis only, which is one type of infectious disease. However, the underlying theory behind infection or sepsis prediction is similar, so this review is still worthy of referring to. Most works use well-known machine learning models such as logistic regression, SVM, random forest and naive Bayes, etc. The authors have posed one major review question along with nine specific sub-questions, and analysed existing works to answer them. The major and sub questions are listed below:

RQ1. Does the literature document methods to predict infections given physiological data?

RQ1.1. Which are the infections or types of infections that are susceptible of prediction according to the literature?

RQ1.2. Do some of these documented methods involve machine learning?

RQ1.2.1. According to the literature, which are the machine learning techniques suitable for infection prediction?

RQ1.2.2. According to the literature, which is the impact of few training samples in infection prediction performance?

RQ1.2.3. According to the literature, which is the impact of a largely imbalanced dataset in infection prediction performance?

RQ1.3. Do some of these documented methods involve expert systems?

RQ1.3.1. According to the literature, which are relevant reasoning rules for infection prediction?

RQ1.4. Which are the available data sources for infections prediction?

RQ1.5. Which are the most frequently reported performance metrics for infection prediction?

Some of the questions are interesting and valuable for our study as well, e.g., RQ1.2.1. which are the machine learning techniques suitable for infection prediction, and RQ1.2.2. which are the impact of few training samples in infection prediction performance. According to this review, for all infections problems, Logistic Regression (LR) is the most common algorithm and is followed by Support Vector Machine (SVM) and Random Forest (RF), but for sepsis, LR and SVM have approximately the same number of usage, and both are more than others. As to the small data problems, three particular impacts were explicitly revealed: (1) low accuracy, (2) limited generalisation, and (3) unfair assessment. Some approaches were proposed to deal with the small data problem, for example, Stanculescu et al. [41] used a symmetric Dirichlet prior with optimised parameters in their autoregressive Hidden Markov Model to prevent the bias caused by the small datasets they have, Wiens et al. [42] applied a novel feature extraction scheme that fits better for small datasets. Imbalanced data samples were another common issue reported in 26 out of the 101 papers under review. Generally, this issue brings two impacts: (1) unfair assessment and (2) low accuracy, which are quite similar to those brought by small datasets. Normally, to alleviate the impact, higher misclassification cost could be assigned to the minority class, like what Monsalve et al. did in their work [43] when they trained the SVM model to predict infection. The other approach is to apply an under- or over-sampling scheme during the preprocess phase, trying to rebalance the data of the majority and the minority. Most of the papers reviewed (61 out of 101) explicitly or implicitly considered the impact brought by imbalance data, even if no measure is taken explicitly, an AUROC-based methodology which is less sensitive to imbalance data is used to evaluate their model performance.

Islam et al. have also presented a meta-analysis [36] on the topic of predicting sepsis patients using machine learning approaches. The paper filtered 7 out of 135 studies, met all of their inclusion criteria, and led to the conclusion that machine learning based approaches can achieve better performance than existing sepsis scoring systems such as Systematic Inflammatory Response Syndrome (SIRS), Modified Early Warning Systems (MEWS), Sequential Organ Failure Assessment (SOFA) and quick Sequential Organ Failure Assessment (qSOFA) in the task of sepsis prediction. The comparison was quite thorough, but the limitation was that only seven papers were reviewed and analysed, and the small number may not be representative.

However, the previous four reviews only covered the applications of some basic machine learning techniques in sepsis detection, but did not include deep learning or some other more advanced machine learning approaches such as XGBoost. This is the gap we are working on to fill in this systematic literature review . In the rest of this chapter, we will review the state-of-the-art machine learning based methods in sepsis detection, especially for neonates. Our contribution will lie in the following three areas: (1) We propose a classification framework to categorise existing works based on different aspects, offering a systematic analysis for researchers and providing an in-depth understanding of how machine learning based methods can be used to predict the onset of sepsis. (2) This literature highlights the current trends in the related research field, suggesting an optimal direction of future research to avoid the negative impact of the highlighted issues. (3) We review the work in the sepsis prediction field in a more technical manner, focusing more on advanced AI-based techniques, including deep learning algorithms.

2.3.2 Review Findings and Discussions

In total, we reviewed 56 papers which satisfied our inclusion requirements, except for the 41 PhysioNet challenge papers. With the proposed classification framework, we cataloged these 56 papers into four areas: prediction features, preprocessing methods, datasets, prediction models (see Table 2.2).

Table 2.2: Cataloging of sepsis prediction research

Reference	Features	Preprocessing	Datasets	Models	Evaluation
Calvert 2016[44]	DC, VS, LT	SW, DIS	MIMIC2	LR, SCORE	AUC, SEN, SPE, ACC
Gur 2014[45]	DC, VS	DIS	HSPT	-	SEN, SPE, OR, TP, FP
Ford 2016[46]	DC, VS	-	HSPT	LR, SCORE	AUC
Nemati 2018[47]	VS	-	HSPT	Others	AUC, SEN, SPE
Tekin 2019[48]	DC, VS	DIS	HSPT	KNN, BN	ACC, TP, FP, PRE, REC, F1
Raben 2018[49]	-	-	HSPT	Others	-
Navarro 2015[50]	RRV	SW, TFM	HSPT	LR	SEN, SPE, TP, FP
Stanculescu 2014[41]	VS	SW	HSPT	HMM	AUC, TP, FP, EE, AP, F1
Li 2019[51]	-	TFM, FF	MIMIC3	LSTM	ACC, F1, AUC
Gómez 2019[52]	HRC	SW, TFM	HSPT	AB, BCT, RF, LR, SVM, BN, DT, KNN	SEN, SPE, TP, FP, AUC
Zhang 2017[53]	DC, VS, LT	TFM, RM	MIMIC3	LR, SCORE	AUC
Nachimuthu 2012[54]	DC, VS, LT	DIS	HSPT	DBN	SEN, SPE, PRE, F1, AUC
Lauritsen 2020[55]	VS	DR, OS	HSPT	CNN, LSTM	AUC, AP, SEN, SPE, PRE
Lin 2018[56]	DC, VS	US	HSPT	CNN, LSTM	AUC, ACC, F1, PRE, REC
Demirer 2019[57]	VS	-	PN	Others	-
Saqib 2018[58]	DC, VS, LT	SW, FF	MIMIC3	LSTM, RF, LR	AUC, F1, PRE, REC
Shashikumar 2017[59]	DC, VS	DIS SW	HSPT	Others	AUC, SPE, ACC

Reference	Features	Preprocessing	Datasets	Models	Evaluation
Fu 2019[60]	DC, VS, LT	RM	MIMIC3	Others	AUC, SEN, SPE, ACC
Shimabukuro 2017[61]	DC, VS, LT	-	HSPT	LR	AUC, SEN, SPE
Barton 2019[62]	VS	-	MIMIC3	LR	AUC, SEN, SPE
Fairchild 2010[63]	HRC	-	-	-	-
Honoré 2010[64]	VS	RM	HSPT	HMM	ACC
Wyk 2019[65]	VS, LT	SW	HSPT	RF	ACC, SEN, SPE, PRE, F2
Kam 2017[66]	DC, VS, LT	SW, DIS, US, NF	MIMIC2	DNN, LSTM	ACC, SEN, SPE, AUC
Schamoni 2019[67]	DC, VS, LT	DF, TFM	HSPT	LR	AUC
Zhang 2017[68]	DC, VS, LT	FF	HSPT	LSTM	SEN, SPE, PRE, AUC, F1
Darwiche 2018[69]	VS, LT	SW, FF	MIMIC3	RF	ACC, SEN, SPE
Bloch 2019[70]	VS	US, DR	HSPT	LR, SVM, ANN	AUC, TP, FP, SEN, SPE, ACC
Vieira 2012[71]	-	DR	HSPT	ANN, SVM, FM	ACC, SEN, SPE
Marshall 2012[72]	-	-	HSPT	DT	-
Lin 2019[73]	VS, LT	US, FF	HSPT	CNN, LSTM	F1, AUC, PRE, REC, ACC
Mao 2018[74]	VS	FF	HSPT, MIMIC3	GBDT, TL	SEN, SPE, AUC, ACC
Liu 2019[75]	VS, LT	OS, FF	MIMIC3	XGB	AUC, SEN, SPE, PRE
Thakur 2018[76]	VS, LT	SW, RM	MIMIC3	LR	SEN, SPE, AUC
Ribas 2011[77]	-	-	HSPT	DT	-

Reference	Features	Preprocessing	Datasets	Models	Evaluation
Garcia-Gallo 2019[78]	DC, VS, LT	RM	MIMIC3	LR	AUC
Raknim 2019[79]	HRC	TFM	HSPT	Others	TP, FP, REC, PRE, SPE, ACC
Joshi 2020[80]	DC, VS	-	HSPT	BN, LR	AUC
Stojkovic 2017[81]	LT	-	-	Others	RMSE
Schlapbach 2017[82]	-	DF	HSPT	LR	AUC, OR
Desautels 2016[83]	DC, VS	SW, FF	MIMIC3	LR	AUC, SEN, SPE, F1, ACC
Khoshnevisan 2018[84]	VS, LT	-	HSPT	SVM	F1, AUC, ACC, REC, PRE
Mccoy 2017[85]	VS, LT	FF	HSPT	LR	-
Mellhammar 2020[86]	DC, VS, LT	MF	HSPT	SCORE	AUC, SEN, SPE, OR
Shuker 2018[87]	DC, VS	-	HSPT	SCORE	AUC, SEN, SPE, PRE
Jiang 2016[88]	VS, LT	-	HSPT	BN	-
Baghaei 2019[89]	DC, VS, LT	DIS, FF	MIMIC3	GRU	AUC, ACC
Ho 2014[90]	DC, VS, LT	MF, NF	MIMIC2	LR, SVM, DT, KNN	AUC, F1, F2
Nizami 2011[91]	HRC	-	-	-	-
Ribas 2012[92]	DC, VS, LT	DR	HSPT	LR	AUC, SEN, SPE, ACC
Ribas 2011[93]	DC, VS, LT	-	HSPT	SVM	AUC, SEN, SPE, ACC
Sheetrit 2019[94]	VS, LT	DIS	MIMIC3	RF, RNN	TP, FP
Godoy 2014[95]	HRC	TFM	HSPT	-	-

Reference	Features	Preprocessing	Datasets	Models	Evaluation
Luo 2019[96]	VS, LT	-	HSPT	-	AUC, SEN, SPE, PRE
Haug 2016[97]	DC, VS, LT	DR	HSPT	RF, BN	AUC, SEN, PRE
Mcgregor 2012[98]	HRC, RRV	-	HSPT	Others	-

Legend:

AB, adaptive boosting; ACC, accuracy; ANN, artificial neural network;
AP, average precision; AUC, area under the curve; BCT, bagged classification trees;
BN, bayesian network; CNN, convolutional neural network;
DBN, dynamic bayesian network; DC, demographical characteristics;
DF, set missing value to default; DIS, discretisation; DNN, deep neural network;
DR, dimension reduction; DT, decision tree; EE, equal error rate; F1, F1-score; F2, F2-score;
FF, forward filling; FM, fuzzy modelling; FP, false positive rate;
GBDT, gradient boosting decision tree; GRU, gated recurrent unit;
HMM, hidden Markov model; HRC, heart rate characteristics;
HSPT, collected from hospital; KNN, k-nearest neighbour; LR, logistic regression;
LSTM, long short-term memory; LT, lab test; MF, mean value filling;
MIMIC2: Multiparameter Intelligent Monitoring in Intensive Care Version 2;
MIMIC3: Multiparameter Intelligent Monitoring in Intensive Care Version 3;
NF, nearest filling; OR, odds ratio; OS, over-sampling;
PN: PhysioNet Challenge Dataset 2019; PRE, precision, REC, recall; RF, random forest;
RM, remove data with missing values; RMSE, root mean squared error;
RNN: Recurrent Neural Network; RRV, respiratory rate variability;
SCORE, scoring system; SEN, sensitivity; SPE, specificity; SVM, support vector machine;
SW, sliding window; TFM, transformation; TL, transfer learning; TP, true positive rate;
US, under-sampling; VS, vital signs; XGB: XGBoost;

In table 2.3, we summarise the aim, main contributions, challenges, and future directions of the reviewed papers. This table offers a quick guide to relevant work for researchers using artificial intelligent methods to detect sepsis. In preparing table 2.3, we classified the main contribution in terms of three aspects: feature, model, and performance and further identified key limitations and possible future research directions which serve as a scaffold to overcome these challenges.

Trends and Focus

We analysed the number of works published in the recent decade on the topic of sepsis prediction, Fig 2.2 depicts the trend of number of research. From the figure, we can see that the topic has drawn increasing interest recently and reached its peak in 2019 with 30 articles. The reason the number dropped in 2020 is because only six months had passed at the time this thesis was composed and the numbers may continue to increase in the remaining half of the year. Note that the calculation excluded the papers published for the PhysioNet 2019 Challenge, which we will cover later in a separated section. Since sepsis has been a global health concern recently due to its mortality and morbidity, and artificial intelligence techniques have advanced rapidly, more and more researchers have recognised that machine learning can be applied to the medical field, especially time series analysis such as sepsis prediction. This, we believe, is the cause of the increasing amount of work.

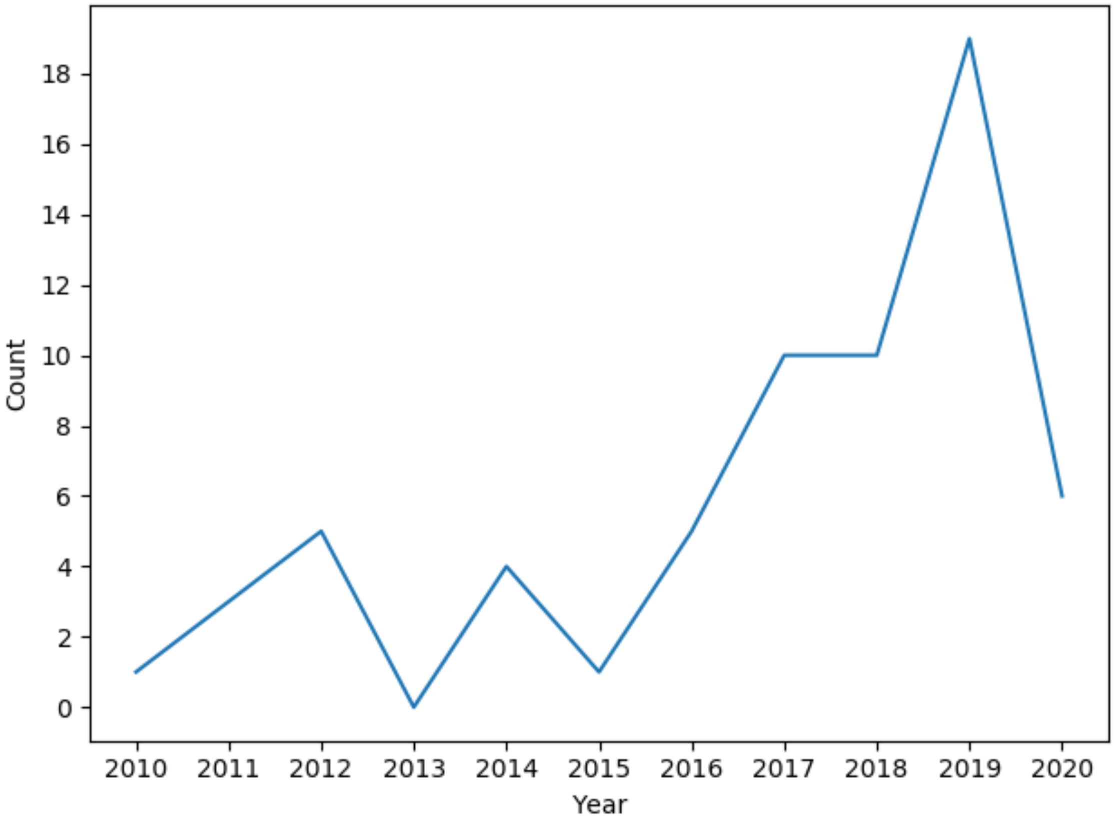


Figure 2.2: Number of publications in the recent decade.

Among all the research related to sepsis detection, except for 42 on normal sepsis, there are 14 targeted at neonatal sepsis specifically, since neonatal sepsis has threatened the lives

Table 2.3: Research aim, highlights of contribution, challenges and future directions

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Calvert 2016[44]	Sepsis onset	- Model: Introduced a machine learning based model *Insight* - Performance: Predicted sepsis up to 3h prior	L: - Limited clinical datasets D: - Implement *Insight* prospectively - Eliminate redundant factors
Gur 2014[45]	Neonatal sepsis	- Model: Evaluated *RALIS* model	L: - Limited clinical datasets - Selection of patients was performed within groups with or without culture proven sepsis
Ford 2016[46]	Mortality	- Feature: Using administrative data	L: - Some variables may have unclear clinical significance - Some administrative data may not be available in every hospital
Nemati 2018[47]	Sepsis onset	- Feature: Using widely available data - Feature: Combining data at different resolutions	L: - Suspicion of infection had to be inferred from systematic criteria - Data used was entered manually by nurses which may confer some information bias
Tekin 2019[48]	Neonatal sepsis	- Performance: Accuracy of 94.53%	- Not given
Raben 2018[49]	Sepsis onset	- Model: Applied functional resonance analysis method	L: - Only focused on how sepsis is detected immediately after admission and can not be translated to patients who develop sepsis during their stay
Navarro 2015[50]	Neonatal sepsis	- Feature: Respiratory patterns - Model: A complex model combines three consecutive models	- Not given
Stanculescu 2014[41]	Neonatal sepsis	- Feature: No need for lab test result - Model: Formulation of sepsis detection as inference and learning in an autoregressive hidden Markov model	L: - Time of the sepsis onset cannot be determined during labelling
Li 2019[51]	Sepsis severity	- Feature: Used MT-DNN model to extract information of organ systems - Model: Bidirectional LSTM network - Performance: 94.72% F1 score	L: - Model not suitable for other databases

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Gómez 2019[52]	Neonatal sepsis	- Feature: Heart Rate Characteristics - Model: Compared 8 different ML-based models - Performance: AdaBoost Classification Trees achieved 94%+ in sensitivity, specificity and AUC	- Not given
Zhang 2017[53]	Mortality	- Model: LASSO regression model for sepsis prediction score	L: - Used the dataset from a single center - The definition of ”routine” variable may vary widely
Nachimuthu 2012[54]	Sepsis onset	- Performance: Achieved an AUC of 94.4% with data of first 24 hours after admission	D: - Perform further experiments with datasets with real-world prior probability of sepsis - Include more clinical variables - Predict future probability of sepsis rather than estimating the current probability
Lauritsen 2020[55]	Sepsis onset	- Feature: Used health record events sequences	L: - Black box model - Oversampling worsened the test performance - Reproducibility is low D: - Add supporting explanation methods into the predictions to improve clinical acceptance
Lin 2018[56]	Septic shock	- Model: Demonstrated a novel neural network architecture specifically designed to handle static and dynamic information in EHRs. - Model: Explored the individual and combined efficacy of these two components i.e. CNN and LSTM - Model: Proposed a novel experimental design exploring two disease prediction tasks: visit-level and event-level	D: - Explore the model's behaviour for different patient subtypes - Test the model in different datasets
Demirer 2019[57]	Sepsis onset	- Can handle datasets with unevenly distributed samples and missing observations	- Not given
Saqib 2018[58]	Sepsis onset	- Feature: With only 24 and 36 hours of lab results and vital signs for a patient	D: - Integrate additional data from MIMIC3 database - Explore LSTM hyper-parameter space more thoroughly

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Fu 2019[60]	Sepsis onset	- Model: Applied improved cascade deep forest model	L: - Sample size is not large - Arguments of RF and XGBoost are too large to adjust
Shimabukuro 2017[61]	Mortality	- Feature: The algorithm only use six vital signs to provide higher sensitivity and specificity	L: - This is a single-centre study in the intensive care unit only D: - Validate the MLA's performance in non-critical care units
Barton 2019[62]	Sepsis onset	- Feature: Only six vital signs were used	L: - The algorithm may perform differently on real-time data - The sepsis gold standard used in this study is necessarily an imperfect characterization of sepsis
Fairchild 2010[63]	Neonatal sepsis	- Feature: Deeply analysed the HRC in the context of detection of neonatal sepsis	D: - Continuous heart rate characteristics monitoring
Honore 2020[64]	Neonatal sepsis	- Model: Tried classical GMM-HMM and recent Flow model based HMM - Model: Explored use of cross-entropy minimization based discriminative training	- Not given
Wyk 2019[65]	Sepsis onset	- Model: Two-layer RF was used to capture features in different granularities	L: - This was a retrospective research, limiting the ability to infer causality prospectively - Analysis was limited to patients who developed SIRS in the ICU D: - Apply this algorithm to prospective data streams
Kam 2017[66]	Sepsis onset	- Performance: Improved performance without feature extraction using domain knowledge - Performance: 92.9% AUC was achieved in the LSTM model	L: - Patients data was extremely imbalanced - Black box model D: - Apply the model in ED - Consider additional variables when applying to ED patients
Schamoni 2019[67]	Sepsis onset	-Not Clear	D: - Discovery new criteria responsible for the development of sepsis
Zhang 2017[68]	Septic shock	- Feature: Used two levels of imperfect labels	D: - Quantify the effectiveness of early prediction in terms of concrete time measure - Incorporate a larger dataset

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Darwiche 2018[69]	Septic shock	- Model: Cox Enhanced Random Forest model improved the lead time before the onset of the septic shock - Model: Offered a practical implementation for utilisation in medical care	D: - Use a different dataset - Create more prediction models based on other features and combine the result through a voting mechanism - Add an unsupervised machine learning technique to enhance the accuracy
Bloch 2019[70]	Sepsis onset	- Model: Presented a hypothesis that unstable patients are more prone to develop sepsis during ICU stay	L: - The dataset was rather small - Physicians determine sepsis onset as the moment in which antibiotics are administered.
Vieira 2012[71]	Mortality	- Model: Applied a new binary particle swarm optimisation method to feature selection	L: - Performance: Poor adjustment will lead to remarkable drop in model performance
Marshall 2012[72]	Neonatal sepsis	- Model: Incorporate classification trees into conditional component of the DC-S model	D: - Develop a statistical model that can benefit neonatal care management
Lin 2019[73]	Septic shock	- Feature: Facial representation - Model: Potentially improved the visualisation of high-dimensional multivariable time-series data like EHRs - Performance: Achieved an AUC of 96.05% with the combination of 2D-CNN and LSTM models	- Not given
Mao 2018[74]	Sepsis onset	- Model: The experiment was set up in three different scenarios: emergency department, general ward and intensive care unit - Model: The algorithm was robust to randomly missing data - Feature: Only six commonly used measured vital signs were used	L: - The retrospective nature of the study does not predict clinician reaction to information.
Liu 2019[75]	Septic shock	- Feature: NLP features extracted from clinical notes	L: - Measurements of many variables are required, but outside the ICU there might not be sufficient data - The criteria for diagnosing septic shock are treatment-based in Sepsis-3 - Data was collected from single hospital

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Thakur 2018[76]	Neonatal sepsis	- Model: Compared two models with invasive and non-invasive parameters	L: - The independent parameters values had to be identified and entered manually in the mobile application D: - Develop a system which can automatically enter the non-invasive parameters values
Ribas 2011[77]	Mortality	- Model: Regression Tree	- Not given
Garcia-Gallo 2019[78]	Mortality	- Feature: patient similarity	D: - Implement novel machine learning approaches on graph-structured data
Raknim 2019[79]	Neonatal sepsis	- Performance: Achieved a sensitivity of 92% and a specificity of 99%	- Not given
Joshi 2020[80]	Neonatal sepsis	- Feature: Estimated infant motion derived from ECG signal	L: - Information regarding supplemental oxygen was unavailable D: - Analysis for infants whose blood culture was negative
Stojkovic 2017[81]	Sepsis biomarker progression	- Model: Applied structured regression using Gaussian Conditional Random Fields	- Not given
Schlapbach 2017[82]	Mortality	- Feature: Cardiorespiratory and metabolic indicators within only one hour were needed	L: - Using mortality as outcome does not capture the impact of sepsis on patients surviving with major long-term morbidity
Desautels 2016[83]	Sepsis onset	- Validated *Insight* can perform better than other complex scoring systems even with less available patient data	L: - This study was performed exclusively on ICU data at a single center D: - Explore how *Insignt* can provide more explanation
Khoshnevisan 2018[84]	Septic shock	- Model: Evaluated RTP on both visit level and event level - Feature: Identified interpretable yet meaningful temporal patterns	D: - Apply pre-clustering of patients to discover more definitive patterns
Mccoy 2017[85]	Sepsis onset	- Not clear	L: - Questionable generalisability - Analyse the patient population based on SIRS criteria instead of ICD

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Mellhammar 2020[86]	Sepsis onset	- Model: Proved that NEWS2 score is better	L: - patients that are considered as falsely classified as positive still can suffer from other conditions - Used sepsis-2 instead of sepsis-3
Shukeri 2018[87]	Mortality	- Feature: The combined use of leukocytes count, PCT, IL-6 and OPN-1 activity	L: - Single centred dataset - This study used a convenience sample, selection bias may have led to a non-representative population
Jiang 2016[88]	Sepsis onset	- Model: An intensive sepsis patient screening and monitoring workflow based on Auto-BN	D: - Perform a cost benefit analysis from the management prospect - collect more real patient data during clinical practice
Baghaei 2019[89]	Sepsis onset	- Model: An attention-based interpretable approach	D: - Evaluate other methods to handle the missing data - A proper clustering of data would possibly yield more interpretability and much robust results
Ho 2014[90]	Septic shock	- Model: Can handle patients with partially missing observations - Feature: Utilised noisy and intermittently gathered non-invasive measurements as proxies for their invasive counterparts	D: - Incorporate time-series models into the framework
Nizami 2011[91]	Neonatal sepsis	- Model: Proposed a framework based on service oriented architecture to support real-time clinical artefact detection	D: - A common reference model shall be developed
Ribas 2012[92]	Mortality	- Model: The proposed method may be understood as a generalisation of the ROD formula	D: - In order to validate the generalisability of the method, future work should lead to a multicentric prospective study
Ribas 2011[93]	Mortality	- Same as last paper	- Same as last paper
Sheetrit 2019[94]	Sepsis onset	- Model: Introduced a new method for temporal-pattern-based classification exploiting similarity measure	L: - The original raw data are usually multivariate with various sampling frequencies and missing values
Godoy 2014[95]	Neonatal sepsis	- Feature: The transformation of features in time-domain and frequency domain	D: - Different mathematical approaches may be analysed in prospective sutdies

Reference	Research Aim	Highlights of Contribution	Limitations and future directions
Luo 2019[96]	Sepsis onset	- Performance: First prospective study to assess the diagnostic accuracy of qSOFA - Model: Sepsis was diagnosed in a real-world setting	L: - This was a single-centre study of selected wards with higher incidence of sepsis
Haug 2016[97]	Sepsis onset	- Performance: High AUC of 97%	D: - Add a subset of the more relevant algorithms - Accommodate R statistical packages
Mcgregor 2012[98]	Neonatal sepsis	- Performance: Distinguish between patients with low HRV due to sepsis and those due to other reasons such as surgery	D: - Work on the effect of drug existence and dosage on HRV

of newborn infants, especially preterm newborns, under care in neonatal intensive care units (NICU). Based on what kind of target is predicted, we can also divide the research into three categories, in which the majority is to predict the onset of sepsis (36 papers), and we have 10 papers for sepsis mortality prediction and 7 papers for septic shock prediction. The ratios are illustrated in Figures 2.3 and 2.4.

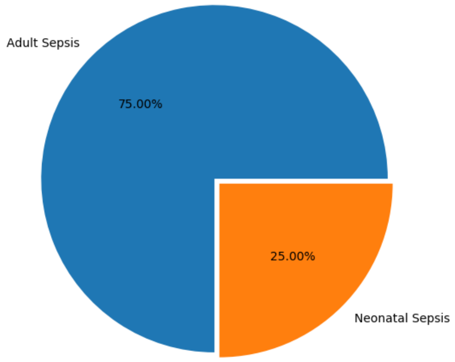


Figure 2.3: Ratio of publication on adult sepsis to neonatal sepsis.

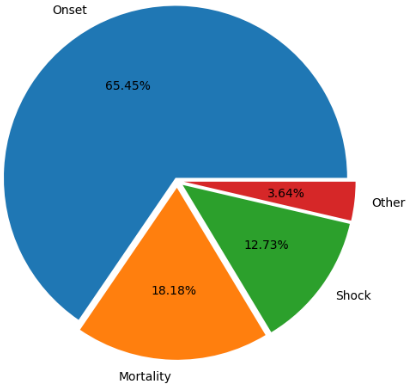


Figure 2.4: Ratio of publications on prediction of sepsis onset, mortality and shock.

Features Selection

For machine learning models, feature selection and organisation are two critical preprocessing steps before they are fed into the model. Features used in sepsis detection can be classified into two categories: raw features and derived features. Raw features are those variables we can directly use after we collect them from the monitors, systems or any other records, while derived features are calculated from raw features based on certain rules. Let us first check the usage of three types of basic raw features that are most commonly used.

There are 47 papers using at least one of the three raw features, among which 24 using demographical data, 40 using vital signs, and 24 using lab tests. Since vital signs are variables that can best represent the physiological status of patients, it is necessary for almost every work to incorporate them as the primary components of feature sets. Demographic data and laboratory tests are not used as much as vital signs because demographical data, like gestational age and gender, is usually static and will not form a time series such as vital signs, while in general, lab tests require quite a long time before the result can be acquired. Another reason for not using lab tests is that they often come with a blood test or other invasive operation which could bring potential side effects for the patient, especially for new born infants. Some researchers aiming at developing non-intrusive early detection of sepsis detection will choose not to include lab test data to overcome these issues [52, 83]

Except for the three raw types of features mentioned in chapter 2.2.1, which are used by almost every piece of work, we have reviewed some other types that are calculated by the raw features. **Measure of acute illness severity** are believed to be significantly associated with mortality in severe sepsis [46], for example, for respiratory failure, the authors constructed a variable termed 'early mechanical ventilation' which is constructed by using the date and time of admission combined with the date for a procedure code corresponding to mechanical ventilation. Any patients receiving mechanical ventilation within 0-2 days of the admission date and time will be identified as early mechanical ventilation cases, and for those that initiated mechanical ventilation after two days as late mechanical ventilation cases. Meanwhile, in the same paper, **pre-existing chronic disease** is also treated as one of the associating factors to the mortality in sepsis. Lauritsen et al. [55] proposed to detect sepsis by utilising deep learning on electronic health record **event sequences**. Various EHR events were monitored and calculated first before being passed to deep learning models, and then along with other vital signs, from which patterns were learned to make the prediction. Sometimes **statistical data** such as maximum values, mean values and standard deviations are more useful than the raw data itself. Multiscale blood pressure and heart rate (HR) time series dynamics are emphasised in another work [59]. Multiple types of blood pressure were extracted, like mean arterial pressure (MAP), systolic blood pressure (SBP) and diastolic blood pressure (DBP). The authors also calculated the following features from HR and MAP time series (2s resolution) derived from the bedside monitor's proprietary software from the ECG and BP waveforms: standard deviation of HR, standard deviation of MAP, multiscale entropy and conditional multiscale entropy of both HR and MAP within a 6-hour time window. Some other works that also used statistical data include [60, 65] were discovered. Continuous monitoring of neonatal **heart rate characteristics** (HRC), had been developed for earlier diagnosis and treatment of Late-onset neonatal sepsis in NICU patients [99–104] last decade by Griffin et al. They found that special heart rate characteristics will occur right before the onset of neonatal sepsis, including two parts: depressed heart rate variability and transient heart rate decelerations. This finding has a profound influence for the later researchers, and HRC was treated as features in some of their works [52, 63, 79, 85, 91]. An electronic questionnaire was created with attending physician's **daily judgements** of patients' status by Schamoni et al. [67], to exploit the implicit knowledge of practitioners. Except for data extracted from the EHR, it is interesting that comments from doctors were also included. Another work that utilised clinical notes was proposed by Liu et al. [75] in which a Natural Language Processing (NLP) technique was applied to generate NLP features as complementary to physiological data. Lin et al. [56] creatively researched how **facial expression** of patients can represent their status of sepsis. A general framework was presented for the extraction of temporal relationships and local patterns of evolving emotional expression in a unified and systematic way based on a patient's health condition. Furthermore, three types of **patient similarities** were calculated, which are Cosine Similarity (CS), Equally Contribution Similarity (ECS) and Weighted Contribution Similarity (WCS), and they were further extracted and input into a Linear Regression model to produce a predictive outcome [78]. In Joshi et al.'s work [80], along with demographic data and vital signs that are commonly used in diagnosis, another innovative feature, **infant motion**, was calculated. To estimate it, they introduced a new integrated measurement - Signal Instability Index (SII), which is derived from the ECG waveform by capturing both the extent and duration of movements. Briefly, the SII is a non-parametric measure based on the kernel density estimate that can be applied to a band-pass filtered ECG waveform to obtain an estimate of motion every second using the ECG-data for the past 10s. Lower values of SII indicate the absence of movement, while higher values are a quantitative estimate of body movement. In some cases, **features were extracted from existing models**. Kam and Kim [66] picked nine basic variables from 460 available ones in MIMIC II database, covering demographic data, vital signs and lab test result. Besides, features used in InSight [44] was also taken into calculation. SOFA and qSOFA are two commonly accepted assessment scores for patients with incipient sepsis, and they were incorporated as part of the feature set fed to the machine learning models [92, 96].

Data Preprocessing

In most cases, selected features can not be fed into machine learning models for training before they have been preprocessed, since outliers and missing values are inevitable, especially in medical records, due to human errors or unexpected activities after admission. In this phase, if appropriate measures are taken, issues like imbalanced data could be alleviated.

dimensionality reduction Sometimes not all of the available features are used in the model training, part of them might be not related to the disease as closely as the others, and we need to filter them out. The purpose of features filtering includes 1) Too many features will make the model complicated, which could possibly increase the chance of overfitting. 2) Incorporating non-related or less-related features might lead to the introduction of extra noise to the dataset, and consequently affect the final performance of the model. Haug and Ferraro [97] run a variable ranking according to the strength of their relationship to the target by chi-square test and then select the N strongest predictors, where N was set to two different values of 15 and 40. In the work of Ribas et al. [92], factor analysis (FA) was taken to select the features following a criterion based on the correlation between features of the observation vector. Bloch et al. [70] proposed to select the most important features in two phases. During the first phase, they have trained 5 different models and estimated the features' model-dependent importance. In the second phase, the top two most important features were selected for each model. Then a combined set of all model-specific features is used as a final feature set.

sliding window vs. stream Sliding window is a widely used technique that is able to convert streaming data into data blocks. Since many traditional models do not have the capability of handling data streams, data segmenting with certain length of time window has become one of the common steps in feature engineering. Desautels et al., Darwiche and Mukherjee [69, 83] segmented their extracted data within a one-hour window and calculated the average for further use. In this work [66] the authors summarised the continuous data in a 1h and 5h time window respectively by extracting the min, average, max values of measured variables. Three moving time windows were used to extract features from patients' physiological data collected at different granularities [65].

Sliding window has another function that can enrich the dataset of limited scale, which is used in [52]. A 15 minutes time window was used to segment the data records due to the low number of new-borns, especially with sepsis. Other publications that were involved in using sliding windows include [50] and [44] to which thirty-second and 5-hour sliding windows were applied.

under-sample vs. over-sample Data records of patients are always extremely imbalanced due to the morbidity of sepsis and modern medical administration. The number of patients who avoid the disease is 10 times or even greater than the number of patients who end up with at least one onset of sepsis. Normally, researchers have two ways to alleviate this problem, which are under- and over-sample. Under-sample means to randomly sample from the healthy patients, reducing the number of normal cases and consequently lead to a re-balance with sepsis cases. On the contrary, Over-sample is to create more data records for the class that is much less than others. In [66, 70, 73], it is an under-sample that is applied to the imbalanced data to equalise two types of data, while Liu et al. [75] chose the over-sample approach. Note that when undertaking an under-sample, we need to maintain the underlying distribution of age, gender or any other attributes as in the original datasets.

transformation To enrich the feature set, raw features are transformed to produce new ones. One typical application is to transform signals in the time domain and in the frequency domain. Navarro et al. [50] does this on respiratory signals while Godoy et al. [95] chose Heart Rate signals. Variables that could be achieved in the time domain include the absolute difference between the maximum and minimum value, standard deviation in the time domain, the root mean squared value, the kurtosis etc, and the kurtosis of the power spectral density (KPSD) and main frequency over a period T can be derived in the frequency domain, in which the main frequency could be expressed as:

$$MF = \frac{1}{T}\sum_{k=1}^{T} |resp(k) - resp(k-1)| \tag{2.1}$$

The Shannon entropy is another useful variable could be calculated[79] as:

$$E = -\sum_{k=1}^{T} resp^2 \log(resp^2) \tag{2.2}$$

For scale-sensitive models, standardisation is an effective way to improve the performance in preprocessing step. A typical way is z-transformation which transform all features x around zero with uniform variant $z = \frac{x-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation of the feature value [67].

discretisation Putting data into bins is a common way to aggregate continuous data into discrete values. After setting up a normal value as reference, data are discretised into bins of lower, normal and higher [89]. In Kam and Kim's work [66], an extra variable was calculated to indicate whether it is increasing, decreasing or roughly constant within a 5h window. [48] bin.

missing data handling Normally, the dataset used in the sepsis detection task is collected from a hospital during the normal course of care, even for those from a publicly accessible database, missing values are common. Multiple strategies could be applied to fill in the missing data, of which forward filling is an effective one. This imputation method applied the patient's last measured value to the following missing entries, in case that the first value was missing, the imputation followed a backward direction. It has been widely used in [51, 68, 69, 73–75, 83, 85, 89]. Mean filling as a complementary if all previous recored values are missing [51, 73]. Directly removing the data entries with missing value [53, 76, 78], setting the value to default defined by an experienced clinician [67] or imputed with the nearest measured value [66] are all acceptable measurements dealing with missing values.

Ho et al. [90] proposed a three-step missing data imputation approach in a specific ”global to local” order to estimate the missing observations in the medical records. Mean imputation is the most global approach available where a patient's missing observation is influenced by measurements from all of the other patients in the population. Neighbourhood-based imputation is on the opposite end of the spectrum, using local information (small subset of the patient population) to determine the missing value. Matrix factorisation methods can be viewed as a combination of the two approaches, imposing a global structure where the individual matrix values are then influenced by a smaller dimensional space.

Some research proposed the theory that missingness of specific data element, in itself, could be treated as a source of diagnostic information. Patterns of missing data was used as a useful proxy variable in the training data [97].

Others A peak detection algorithm was used to detect the R-peaks in the ECG-recordings, followed by calculating the R-R intervals, in order to get the Heart Rate Variability features [80]. For Natural Language Processing related approaches, clinical notes need to be processed, for instance, the notes are converted to lowercase and stripped of non-alphanumeric characters, to get ready for further steps [75].

Research Datasets

Research which studied the sepsis detection problem were usually collaborated with hospitals or other health institutes, so most of their dataset was collected from patients. Meanwhile, several public databases providing open access to clinical datasets of high quality are also popular among researchers, like MIMIC and PhysioNet. MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising of deidentified health data associated with approximately 60,000 intensive care unit admissions. It includes demographics, vital signs, laboratory tests, medications, and more [31]. Similarly, PhysioNet is another platform created for biomedical research and education by offering free access to large collections of physiological and clinical data and related open-source software [32]. There are 35 papers (62.5%) among all the reviewed papers using datasets from hospitals, institutes or their own projects, while 12 (21.4%) of them using MIMIC III, and 3 (5.4%) using MIMIC II, see Fig. 2.5. From the figure we can identify more than half of the research chose to collect their own data instead of using public datasets, due to the nature of MIMIC database that only focus on admissions in intensive care unit.

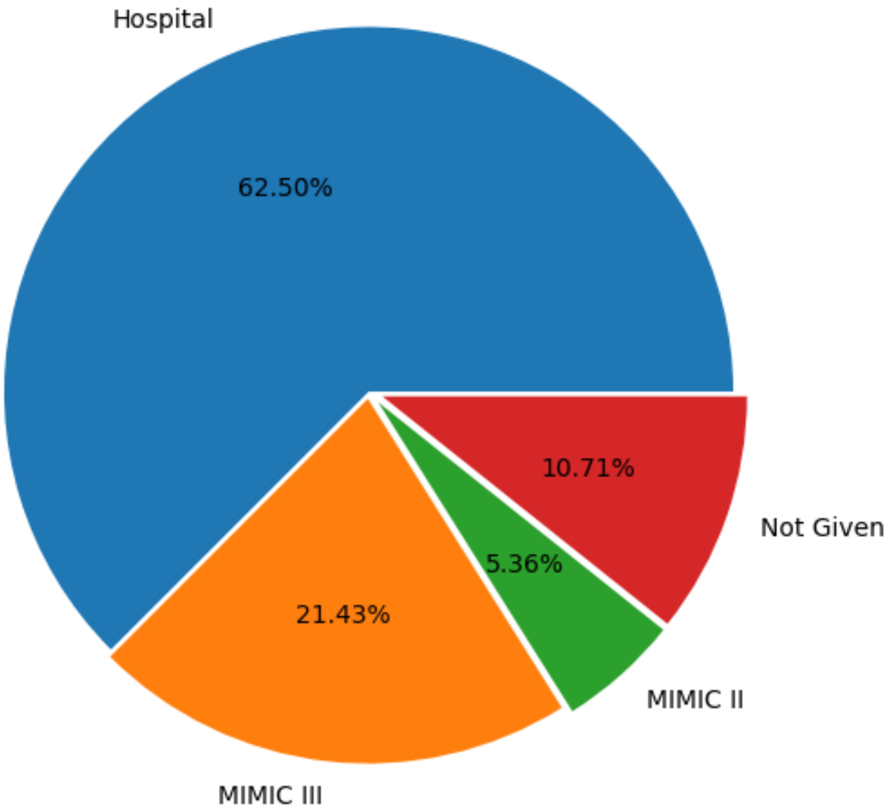


Figure 2.5: The distribution of datasets used in reviewed work.

Prediction Models

conventional machine learning models For the first type, conventional models, we have 28 papers (50%) which utilise normal machine learning algorithms. Logistic Regression (LR) was widely used due to its simplicity, 8 out of 56 (14.3%) papers among all the reviewed works have more or less utilised this method. Binary LR was used to develop and compare two prediction models using invasive and non-invasive parameters in [76]. This model calculated a probability of sepsis in an Android application, and predicted neonatal sepsis. An LR classifier could also be trained to identify body movement, as the first step of the sepsis detection method described in [50]. There are multiple publications proposing to facilitate sepsis prediction with scoring systems that calculated by LR models. Ford et al. [46] generated a sepsis severity score for use with administrative data with the multivariable logistic regression model, while in the work of Shukeri et al. [87], LR was used to derive a sepsis mortality score (SMS), a prediction equation describing the relationship between biomarkers and 30-day mortality. Other works incorporated with LR include [58, 82, 92].

Tekin et al.[48] tried two models, one was K-Nearest Neighbour (KNN) and the other, Naive Bayesian (NB) and compared the performance of the selected methods. Results showed that the basic ML model KNN slightly outperformed the probabilistic model NB.

We also found that the Support Vector Machine (SVM) algorithm was used to build models for early diagnosis of septic shock[84]. Six classic machine learning classifiers were compared and analysed, and SVM with Recent Temporal Patterns (RTPs) outperformed among all the candidates. SVM could be modified to be robust and interpretable in real-time decision making in the ICU. Ribas et al. [92] developed such a system based on SVM model and provided an automated ranking of relevance of the mortality predictors. The performance of SVM models very much depends on the kernel functions applied, a good example would be the work of Bloch et al. [70], in which SVM with radial basis function evidently has the highest AUS of 88.38% in the task of sepsis prediction.

The family of hidden Markov models (HMMs) is a flexible tool for generative probabilistic modelling of sequential data such as vital signs time series, commonly available in sepsis prediction, and it was applied in the work of Stanculescu et al. [41] with a slight enhancement by introducing a direct stochastic dependence between observations.

Ensemble learning is another popular technique that many researchers would like to use due to its outstanding performance. Random forest (RF) is one of the typical Ensemble learning algorithms, and van et al. [65] employed a two-layer stacked RF structure in their work because of its minimal hyper-parameterisation, faster training, and improved interpretability. Mao et al. [74] chose another subtype of ensemble learning which is Gradient Tree Boosting, and Liu et al. [75] used XGBoost which becomes increasingly popular recently.

deep learning models Deep learning has been explored extensively and applied in many research domains including healthcare for decades. In terms of deep learning models, we mainly discuss deep neural networks with special layers, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants. Long Short-

Term Memory (LSTM) is a specially designed neural unit for RNN, empowering RNN to record both recent and distant features, and to avoid gradient vanishing, a catastrophic consequence caused by the large number of layers. Kam and Kim [66] proposed the LSTM-based method of sepsis detection. The model consists of two components, a deep forward neural network that learns unknown high-level features without domain knowledge, and an LSTM network working on temporal features. When multiple layers of traditional LSTM are stacked, they are able to learn the discriminative patterns at different levels, hence increasing the comprehensive performance of prediction [68]. The other possible enhancement to the original LSTM is to train the model through both forward and backward directions so that the temporal dependency is captured bidirectionally [51]. However, we do not think the backward direction should be taken into consideration, because in reality, the status of the patient in the future is unknown and can not be input into the model.

Sometimes, CNN and RNN are combined to make the best use of each model. Lin et al. [73] presented a septic shock prediction framework based on the facial representation of patients, in which RNN extracted the temporal dependency among facial images at different times, whereas CNN summarised the local patterns, and finally generated the prediction result via another RNN. We can improve the performance of LSTM by adding extra components to it. In [55, 56], a CNN was concatenated to LSTM to obtain local characteristics of EHRs, and a fully connected (FC) neural network was followed, introducing static information to LSTM as a complementary. This type of conjunction between CNN and RNN is called Convolutional-LSTM or ConvLSTM, which has been widely applied to solve problems involving both spatial and temporal data simultaneously.

Despite their complex structure and advanced capability, deep learning models can not guarantee better results than others. The examples that LSTM underperform other models include [58, 84] in which LSTM didn't achieve a better performance than random forest and SVM respectively. We assume that the reason is probably the small size of the dataset, since deep learning models have more complicated structures and require a large amount of data during training to prevent overfitting.

other statistical models Besides the most focused conventional machine learning models and some more complex deep learning models, a variety of scoring systems were designed to be an indicator of the early signal of sepsis onset. *Insight* [44] was probably one of the most commonly used scoring systems proposed by Calvert et al back in 2016. The formula of the score is showed in Eq. 2.3, where observation of the i-th measurement with the 5-hour window is averaged and stored in M_i, the change of the observation will be classified as increasing, roughly constant or decreasing and indicated as $\hat{D}_i$. Similarly, trends of doublet and triplet measurements are denoted as $\hat{D}_{ij}$ and $\hat{D}_{ijk}$.

$$Insight\ Score = a\sum_{i\in A} p(M_i) + b\sum_{i\in B} p(\hat{D}_i) + c\sum_{(i,j)\in C} p(\hat{D}_{ij}) + d\sum_{(i,j,k)\in D} p(\hat{D}_ijk) \tag{2.3}$$

The least absolute shrinkage and selection operator (LASSO) was proposed to mitigate the limitation that covariates of high correlation can be included in conventional scoring methods [53], and LASSO method was able to retain important variables by penalising the coefficients of trivial variables to zero. Two LASSO regression were used to generate a new risk stratification score - Sepsis Early Warning Score (SEWS) proportion to the coefficients with the cut-off value of 7 points [86]. Logistic Regression was used to derive the Sepsis Mortality Score (SMS) which is a prediction equation describing the relationship between biomarkers and 30-day mortality of the sepsis patient, see Eq. 2.4 and 2.5.

$$SMS = [e^{logit(p)}/(1 + e^{logit(p)})] \times 100 \tag{2.4}$$

$$Logit(p) = 0.74 + (0.004 \times PCT) + (0.001 \times IL - 6) - (0.025 \times ARE) - (0.059 \times LC) \tag{2.5}$$

PCT stands for Procalcitonin, IL-6 is Interleukin-6, and LC indicates leukocyte count. There are more works developing the scoring system for sepsis prediction among our review but no specific formulas provided [46].

In addition to various scoring systems, some other methods are worth mentioning. RALIS, which is a computerised mathematical algorithm for continuous monitoring of patients was specifically developed to detect the potential onset of late-onset sepsis among premature infants [45]. Raben et al. proved that the systematic method of Functional Resonance Analysis Method (FRAM) offers a novel, alternative way of investigating the process of healthcare, in the case study of early detection of sepsis [49]. The method emphasised the six subsets of FRAM, through which it helps researchers to investigate the process in a systematic way and gain an understanding of how the process is adapted to everyday variability. Furthermore, fuzzy models were mentioned and compared, due to their better interpretability than traditional machine learning and deep learning models [71]. Fuzzy models are ”grey-box” rather than ”black-box” since its rule based nature allows for a linguistic description of knowledge. Finally, a probabilistic graphical framework - Gaussian Conditional Random Fields (GCRF) was examined in the task of vital biomarkers progression prediction, in particular cytokines which is important for sepsis[81]. This work was done upon the approach investigated in [105], extended the original work by replacing independent linear chains with a more general graph dense graph, and achieved better accuracy.

Network Structure

Specifically, we compare the network structures of models incorporating neural networks in our reviewed papers in this section, to highlight potential performance differences brought by changes in network structures.

In [66], a vanilla LSTM network was used as the model for prediction, comparing to a fully-connected feedforward network, although the authors applied LSTM network without any modification, just the vanilla structure of one hidden layer with 64 one cell memory blocks can outperform the feedforward model due to its ability to learn time dependent behaviour. There was another paper [68] incorporating vanilla LSTM network, in which the probability of getting septic shock was directly calculated by applying a sigmoid function to the hidden states h^t output from each LSTM cell, see figure 2.6, where y^t is the probability of getting a shock at time t.

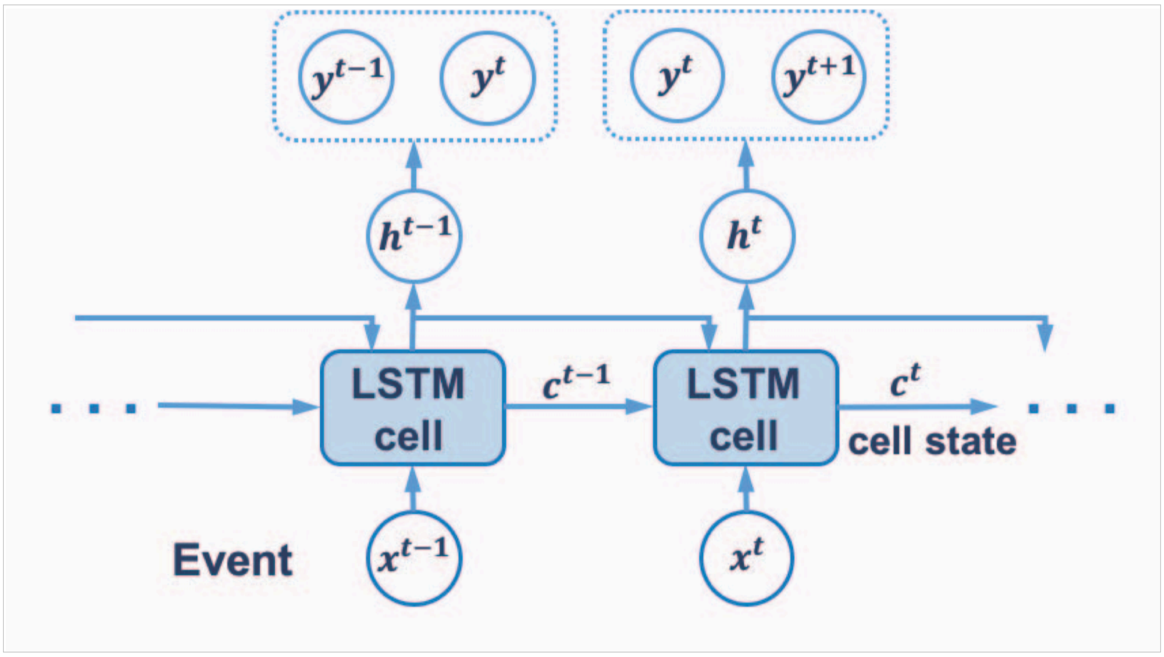


Figure 2.6: The vanilla LSTM structure used in [68].

Besides the basic form of LSTM network, li et al. [51] proposed a much more complicated and improved version of LSTM-based model. Figure 2.7 shows the layers and structure in detail. Basically, this complex model combines four major modules including Pretraining, Self Attention, Bidirectional LSTM network, and Attention. In pretraining, MT-DNN [106] was employed for encoding the data. These outputs were then passed to calculate a self-attention representation vectors to capture the meaning of time-series data, considering the human organ systems which may not be learned well due to the long-term dependency problems. Multi-head attention formulation was used in this module. The following LSTM module was not vanilla in this case, and it was modified to be a bidirectional LSTM network that consisted of forward and backward sub-LSTM networks. The LSTM from two different directions can capture not only relations to the past, but also relations from the future events. Finally, dot-product attention mechanism was applied to select important features and direct short-cuts were connected between the target and the source.

CNN is sometimes paired with LSTM and this usually gives a better result. An extremely complex model was presented to facilitate an innovative idea that was to predict septic shock from facial representations [73]. The network structure basically comprises of two parts: an image generator and a prediction module. For the image generator, static information was used to generate an identity which is a one hot encoding of 57 dimensions by the K-medoids clustering algorithm, grouping patients based on the attributes. As a result, patients with similar static information form a cluster and have the same identity vector. Another one hot encoding of 8 dimensions emotion vectors were learned by an LSTM model from patient data. More specifically, emotion vectors were calculated as the final hidden state of LSTM. The two one hot vectors will together generate a facial expression, which maps to the status of the patient, such as infection, organ failure or even septic shock. Both vectors are first independently passed to two fully connected layers (FC) with 512 neurons. The output is concatenated and passed to two FC layers with 1024 neurons and the third dense layer with 2560 neurons. Next, the output vector h with 2560 dimensions is reshaped to a 5×4 image of 128 channels. It is fed through 6 upsampling + deconvolution layers with 2×2 upsampling and each layer is followed by a convolutional layer. The upsampling + deconvolution operation can be seen as the opposite of the convolution + pooling steps performed in standard CNN. The upsampling step upsamples the feature maps to increase their space span, thus the width and height are both doubled after this operation is applied. The final convolutional layer produces an image of 640×512 with 3 channels. Note that each layer

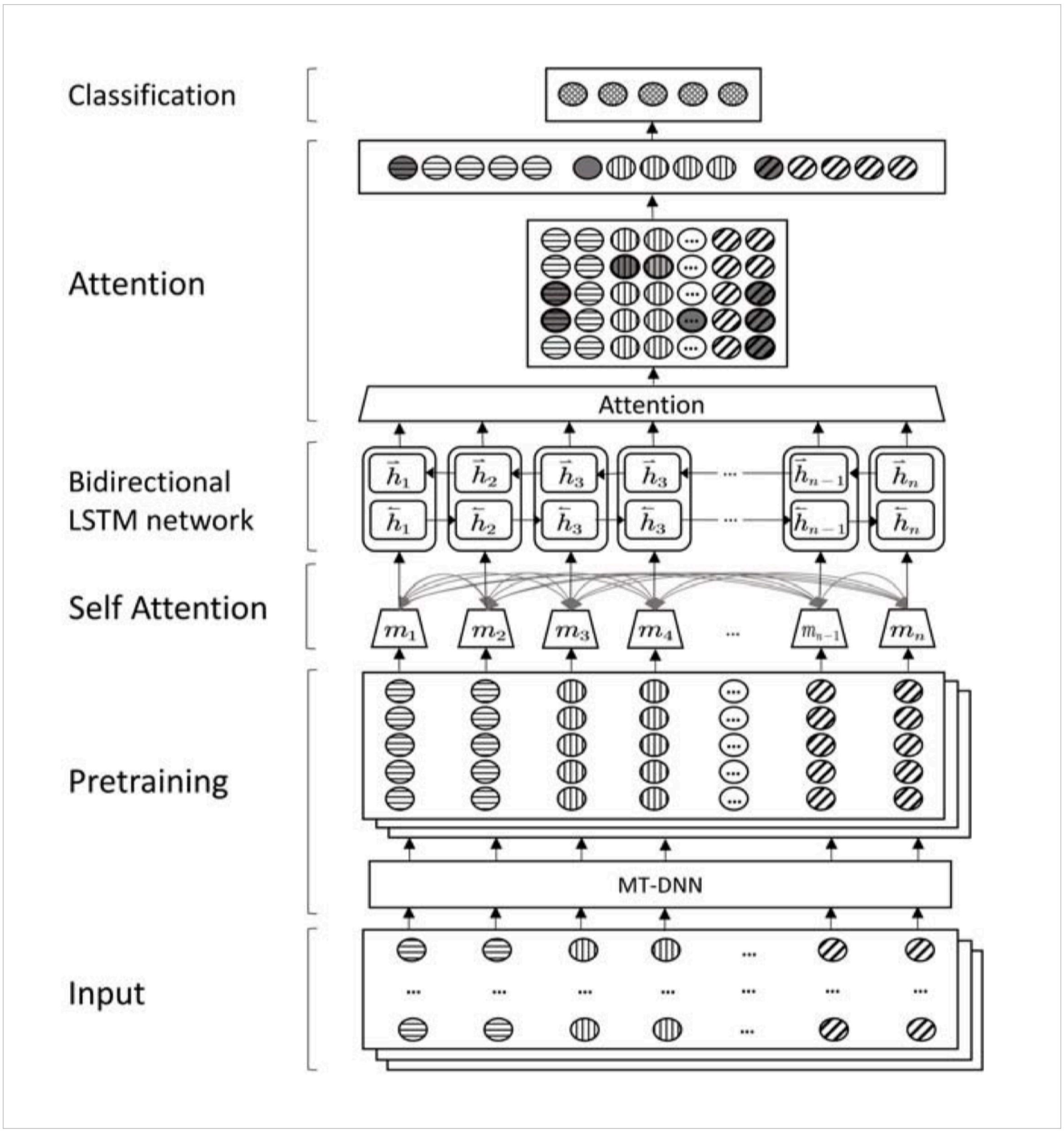


Figure 2.7: The complex LSTM-based network used in [51].

of the network is followed by a ReLU operation to introduce nonlinearity. The network is trained by minimising the reconstruction error, or Mean Squared Log Error (MSLE) between the original images and the generated images. For the prediction module, 2D CNNs were used to extract local features from the facial expression generated by the image generator in the previous step, and an LSTM based model will produce the final prediction. Static information was also considered to be involved in the classification in another work [56]. They proposed two different ways to incorporate static information, one is static-repeat mode and another is static-last mode, see figure 2.8 and 2.9. The only difference is how the model deals with the static information: it is concatenated to the output of each LSTM unit in static-repeat mode, while only in the last or final time step of the LSTM network the static information is involved in static-last mode. Experiments showed that the latter method has not only the simpler calculation but also the better performance.

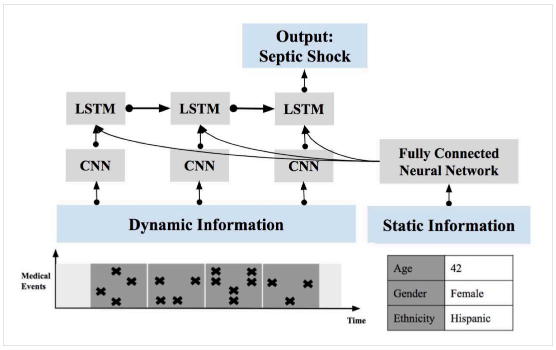


Figure 2.8: Static-repeat: the output of FC is concatenated with the hidden state of LSTM at every time step. [56]

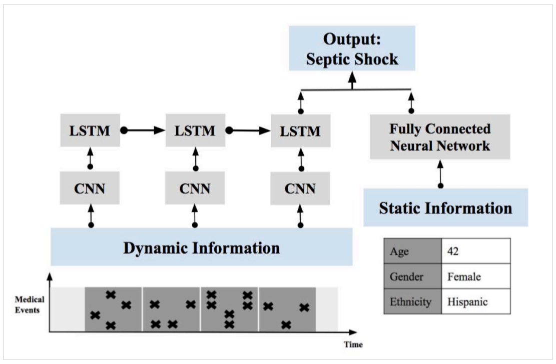


Figure 2.9: Static-last: the output of FC is concatenated with the output of LSTM at last step. [56]

There is a well-developed model which combines CNN and LSTM, also known as CNN-LSTM or Conv-LSTM which has shown to learn robust temporal feature representation in the convolutional layers, making it easier for LSTM layer to capture temporal dependencies compared to using the raw input [55]. The model first projected the sparse inputs into dense 1000-dimensional vectors, reducing the dimensionality for the following convolutional layer by a factor of five. Short-term temporal developments for a patient were then captured in the model by a stack of ”convolutional blocks” which consisted of two one-dimensional ReLU-activated convolutional layers followed by a max-pooling layer. All convolutional layers have kernels of size 3, a stride of 1, and zero-padding was used. All max-pooling layers have a kernel size of 2 and a stride of 2, halving the temporal width of the input. To ensure that information across the convolutional blocks obeys the ordering of the input information, without contaminating the output with information from the future, all kernels were causal in the sense that they only filtered input from the current time and the past. There were five convolutional blocks in the model. The initial block had a depth of 128 for both of the convolutional layers in the block, whereas the convolutional layers in the last four blocks all had a depth of 64. After the input filtered through the five convolutional blocks, the output vectors contained partly overlapping temporal information, where each vector spans 15 hours and 30 minutes of the original input, and the temporal distance between two succeeding vectors was 2 hours and 40 minutes. Finally, the model captured the long-term temporal development of a patient by allowing the output from the convolutional blocks to feed into an LSTM layer that incrementally builds up a representation of the temporal dependencies and continually predicts an output. The LSTM layer has 64 units and was initialised with a random initial state, see figure 2.10.

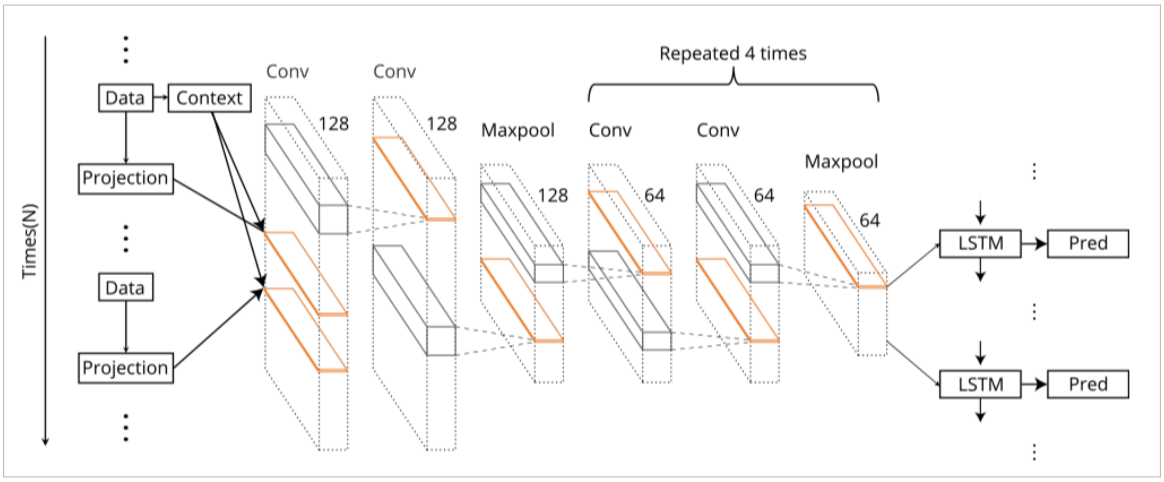


Figure 2.10: The CNN-LSTM model used in [55].

Performance Evaluation

As presented in Table 2.2, research studies have used different mathematical measures to evaluate the outcome of their proposed algorithms. For those with sufficiently available labeled data, the classical criteria based on receiver operating characteristic (ROC) e.g. AUC or precisionrecall (PR) curves e.g. APs have been used to analyse the performance of the proposed algorithms. ROC curves are commonly used to present the results for binary decision problems in machine learning, and among our reviewed publications, 36 out of 56 (64.3%) have used area under the ROC (AUC) as one of performance metrics. However, with highly skewed datasets, ROC does not provide much insight into the data, and PR curves tend to provide a more informative picture of an algorithm's performance [41]. In the task of sepsis detection, the number of negative samples considerably exceeds that of positive examples.

Consequently, a substantial change in the number of false positives (FP) can lead to a small change in the FP rate used in the ROC analysis [94].

Furthermore, sensitivity and specificity would be another pair of popular performance metrics which are widely applied in medical research with binary decision problems. From our review, they appeared in 27 out of 56 (48.2%) papers which is nearly a half. The definition of both of them is listed as follows.

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative} \tag{2.6}$$

$$specificity = \frac{true\ negative}{true\ negative + false\ positive} \tag{2.7}$$

Sensitivity represents the portion of correctly predicted patients out of all the positive patients, and it focuses more on the coverage of positive samples, while specificity is in the opposite direction, focusing on negatives. Due to the nature of the disease, we always try to cover as many positive cases as possible, so higher sensitivity stands for wider coverage of positive cases. Meanwhile, higher specificity means less mis-diagnosis of sepsis for healthy samples, see [59, 65, 71, 75, 83, 87].

In machine learning prediction problems, the performance of models is usually evaluated by a set of metrics comprising accuracy, precision, recall and F-measure. We observed a huge portion (53.6%) of reviewed works implementing this set of metrics for outcome evaluation. Precision and recall ofter comes in pair, see the follow definitions.

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{2.8}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{2.9}$$

According to above equations, precision captures the effect of many negative examples on the algorithm's performance by comparing false positives to true positives rather than true negatives, and recall is actually the same as sensitivity [56, 73, 79]. Comparing to the next measurement, i.e. accuracy, both of them focus on positive cases, but accuracy calculates all the correct predictions, combining positive and negative cases, which is inconsequential in sepsis detection tasks [48]. Considering the extremely imbalanced dataset of sepsis patients, the measurement weighs higher on true positives and false positives is more effective. In such case, F-measure is preferred as it balances precision and recall, resulting in a better evaluation of a sepsis prediction model.

$$F_\beta = (1 + \beta^2) \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{2.10}$$

Equation 2.10 denotes the general formula for F-score, and F1[41, 54, 58, 68, 73, 83] and F2[65, 90] are two most commonly used F-measurements as setting β to 1 and 2 respectively. Note that the intuition behind F2 score is that it weights recall higher than precision. The distribution of evaluation metrics is shown in Fig 2.11.

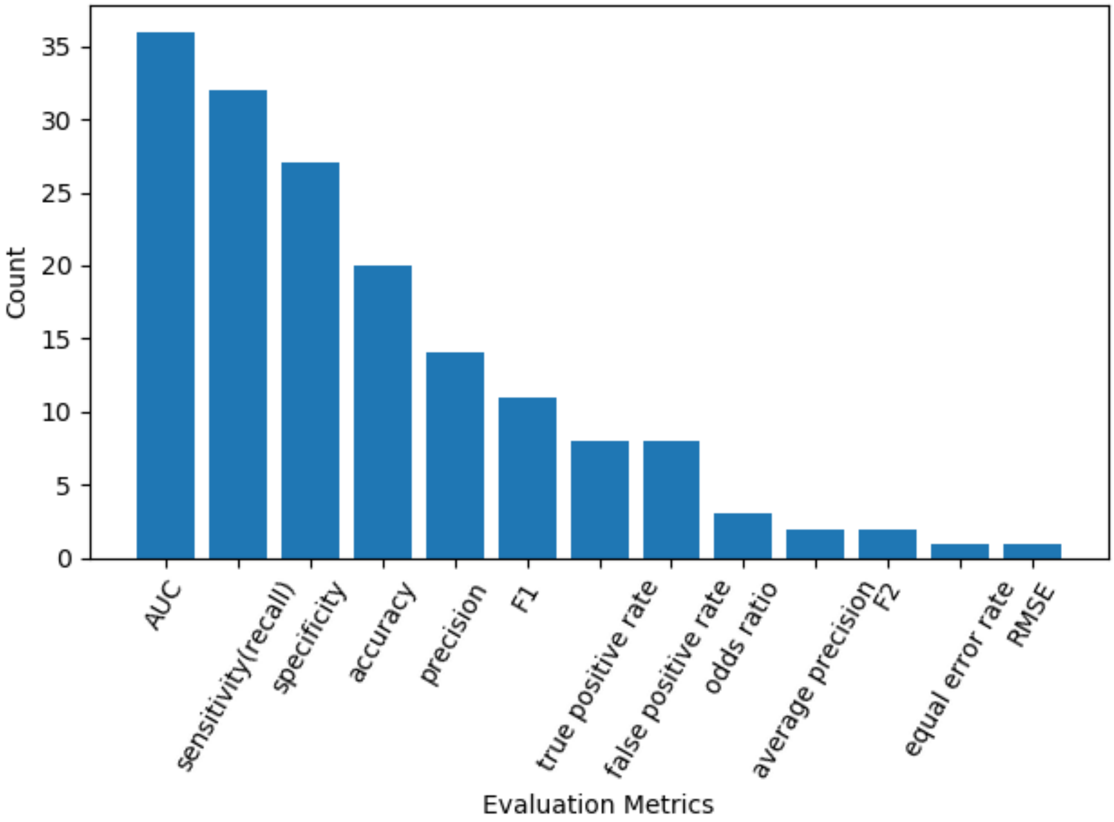


Figure 2.11: Distribution of evaluation metrics.

Limitations and Future Directions

As illustrated in Table 2.3, we have concluded limitations and possible future research directions that can address these limitations. In general, we identified three major limitations that commonly existed among related research. First, it is the limitation that comes from dataset. For those works involving data collected from local hospital wards, emergency department or other medical institutes, most of them will face the problem of that the size of dataset is small, and the models run on top of the small dataset may not be able to lead to a solid conclusion [44, 45, 60]. Another shortcoming of the dataset is that sepsis samples always go with a very small portion, and the ratio of negative and positive data is extremely imbalanced [54, 65]. Skewed data lead to models that are prone to predict the majority class as the result and this consequently reduces accuracy. Limited by the experiment conditions, researchers were not able to collect data from patients of different races at different locations around the world. Experiments were reported to use single-centred data samples which were a non-representative population, and the generated results might not be able to generalise to the rest of the world [52, 60, 75, 83, 87, 96]. There were some other generalisation related limitations reported [85]. Some research used administrative data that might not be available in every hospital [45], some designed the model specifically suitable for certain databases [51], and some used the original raw data sampled at a certain rate which may not be applicable in other places [94].

The nature of machine learning brings the second type of limitation related to the models which is the weak interpretability [55, 66]. Some of the machine learning models are relatively easy to understand, but they are usually simple models with fewer layers and parameters and less complex structures, like linear regression and decision tree. However, for more complicated models i.e. CNN or LSTM, even though they may have a better accuracy in the task of sepsis prediction, their training processes are more like a black box which is opaque to clinicians, so that they tend not to be confident about the result without appropriate explanation.

The gold standard of sepsis remains debatable among researchers and there are multiple versions like Sepsis-3 and SOFA, but none of them are perfect characterisations of sepsis [62]. Until now, there is no agreement on which version is the most accurate, which resulted in some issues during the diagnosis. Physicians determine sepsis onset as the moment at which antibiotics are administered. However, the onset might have happened several hours before that, this will bring negative influence to models and datasets that rely on the accurate onset time [70]. This is a limitation of the medical process which could not be improved easily.

During the data collection phase, it is inevitable to involve some manual labour, i.e., nurses manually entering the temperature, which will introduce information bias and human errors [47, 76]. Some necessary operations during the training process will also worsen the performance, for example, the over-sampling [55]. In some cases, poor adjustment of the parameters will lead to a remarkable drop in model performance [71].

Potential future research directions aiming at addressing the aforementioned challenges and expanding the research area are summarised. Multiple potential work extensions were proposed, but all of them were to solve the generalisation issue. The one that was mentioned the most is to upgrade the experiment to multiple-centre one with heterogeneous datasets [53, 56, 58, 68, 69]. Shimabukuro et al. [61], Kam and Kim [66] planned to validate the machine learning algorithm in non-critical care units and emergency department, respectively, both of which serve the purpose of widening the application scenario of proposed methods. To improve performance, new criteria responsible for the development of sepsis need to be discovered [67]. More prediction models based on more features would be created and the results would be combined through a voting mechanism to generate the final prediction [69, 95]. Improvement needs to be taken into consideration to increase the workflow efficiency and avoid manual labour along with the unexpected errors and bias [76]. In the work of Lauritsen et al. [55], they posed an interesting direction for future work which is to add supporting explanation methods into predictions to improve clinical acceptance.

2.4 PhysioNet Challenge 2019

In this section, we will summarise the work of a competition held by PhysioNet in 2019, targeted at predicting sepsis onset with clinical data. From all the 79 participating teams, we have retrieved 41 papers presenting their works. Through our analysis and comparison, current trends and patterns of AI-based approaches applied in sepsis prediction were identified.

2.4.1 Dataset

The datasets provided by the competition have data from 40,336 patients aged from 14 to 100 from two separate hospital systems. There was a third dataset for testing purposes, but unfortunately since the competition had finished, we were not able to assess it. Each data entry is a sequence of multivariable values consisting of 40 different features plus one label which indicates whether this patient will have an onset in the next six hours. Each row of the record contains a single hour's observation.

The feature set consists of three basic types as we mentioned in section 2.3.2: vital signs such as heart rate and oxygen saturation, 8 columns in total; laboratory values such as blood urea nitrogen and platelets, 26 columns in total; and demographics such as age and gender, 6 columns in total, see Table 2.4.

Table 2.4: Features of dataset provided by PhysioNet Challenge 2019

Type	Name	Description
Vital signs	HR	Heart rate (beats per minute)
	O2Sat	Pulse oximetry (%)
	Temp	Temperature (Deg C)
	SBP	Systolic BP (mm Hg)
	MAP	Mean arterial pressure (mm Hg)
	DBP	Diastolic BP (mm Hg)
	Resp	Respiration rate (breaths per minute)
	EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory values	BaseExcess	Measure of excess bicarbonate (mmol/L)
	HCO3	Bicarbonate (mmol/L)
	FiO2	Fraction of inspired oxygen (%)
	pH	N/A
	PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
	SaO2	Oxygen saturation from arterial blood (%)
	AST	Aspartate transaminase (IU/L)
	BUN	Blood urea nitrogen (mg/dL)
	Alkalinephos	Alkaline phosphatase (IU/L)
	Calcium	(mg/dL)
	Chloride	(mmol/L)
	Creatinine	(mg/dL)
	Bilirubin_direct	Bilirubin direct (mg/dL)
	Glucose	Serum glucose (mg/dL)
	Lactate	Lactic acid (mg/dL)
	Magnesium	(mmol/dL)
	Phosphate	(mg/dL)
	Potassium	(mmol/L)
	Bilirubin_total	Total bilirubin (mg/dL)
	TroponinI	Troponin I (ng/mL)
	Hct	Hematocrit (%)
	Hgb	Hemoglobin (g/dL)
	PTT	partial thromboplastin time (seconds)
	WBC	Leukocyte count (count*$10^3/\mu L$)
	Fibrinogen	(mg/dL)
	Platelets	(count*$10^3/\mu L$)
Demographics	Age	Years (100 for patients 90 or above)
	Gender	Female (0) or Male (1)
	Unit1	Administrative identifier for ICU unit (MICU)
	Unit2	Administrative identifier for ICU unit (SICU)
	HospAdmTime	Hours between hospital admit and ICU admit
	ICULOS	ICU length-of-stay (hours since ICU admit)

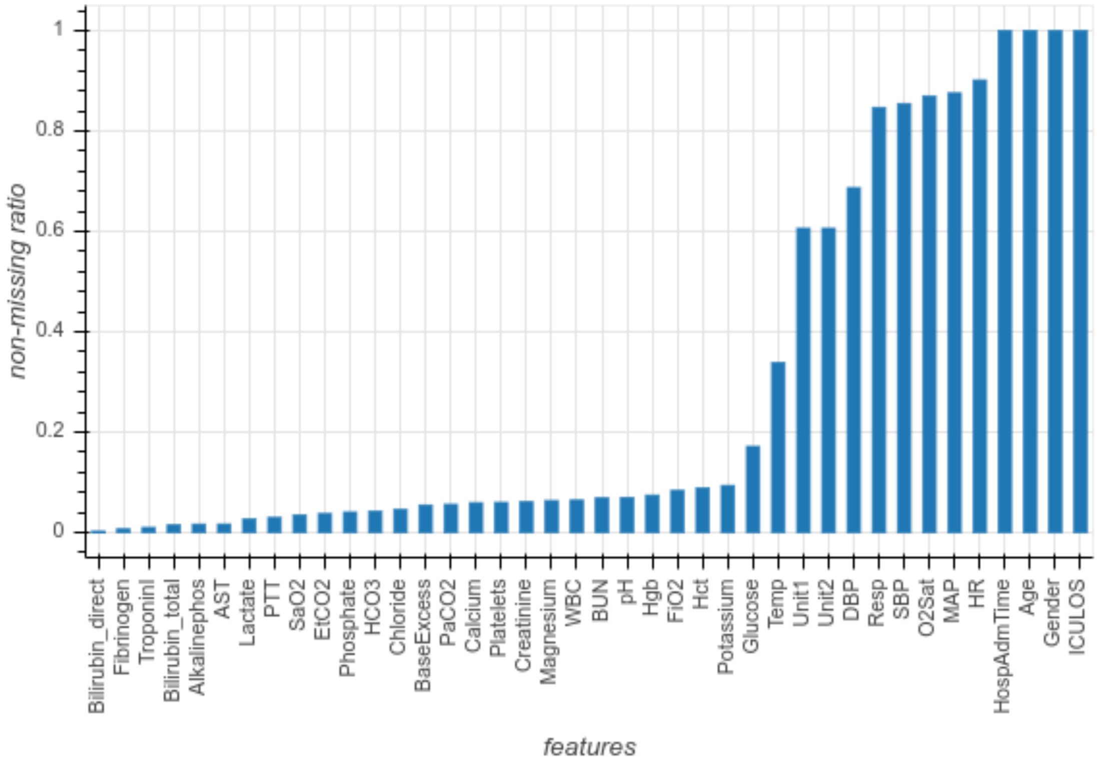


Figure 2.12: Available ratio of each feature in the dataset

However, not every feature is available in the raw dataset, due to the way this data is collected and due to human errors. In fact, we found that total non-missing ratio for all features is as high as 31.63%, which means more than half of data is not available. The non-missing ratio in terms of each feature can be seen in Figure 2.12, and it's easy to see that for 26 features, they have actual values less than 10%. There are only 9 features with missing values less than 20%, and only 3 features have full data.

On the other hand, the provided dataset is extremely imbalanced. From the point of view of patients, out of 40,336 patients, only 2,932 had sepsis which accounts for 7.27% of the total cohort. If we analyse the data row-wise, the imbalance is even worse. The entire dataset contains 1,552,210 rows of records, of which only 27,916 rows were marked as sepsis, taking a percentage of 1.8%, see Figure 2.13 and 2.14. The ratio of healthy against sepsis patient is 12.76:1 while that of sepsis records to healthy ones even reaches 54.6:1.

Another characteristic of the given dataset is that the length of the data sequence from each patient is different. They vary from 8 to 336 which requires the prediction model to have the capability to deal with variable-length sequence data. As shown in Figure 2.15, the distribution of data lengths is not balanced or uniformed, and most patients have a data length of less than 60 hours.

The above three characteristics are not only of the provided dataset of this competition, but also common to all the different patient datasets. They make the dataset sophisticated and increase the difficulty of the implementation of a high-accuracy prediction model on it.

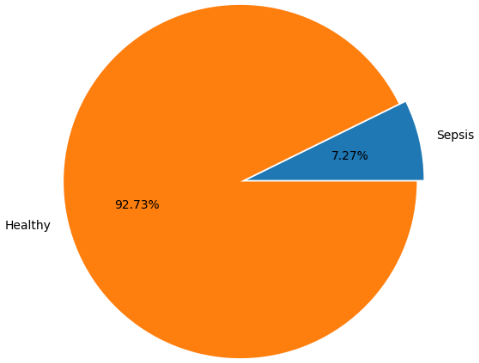


Figure 2.13: Ratio of sepsis patients to healthy ones.

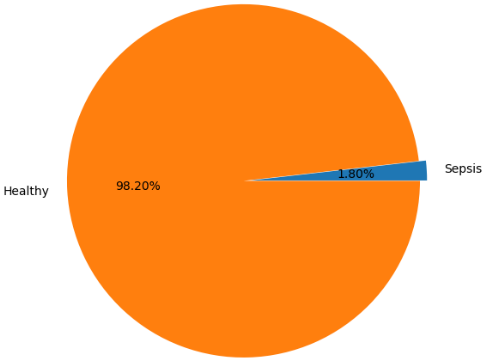


Figure 2.14: Ratio of sepsis records to healthy ones.

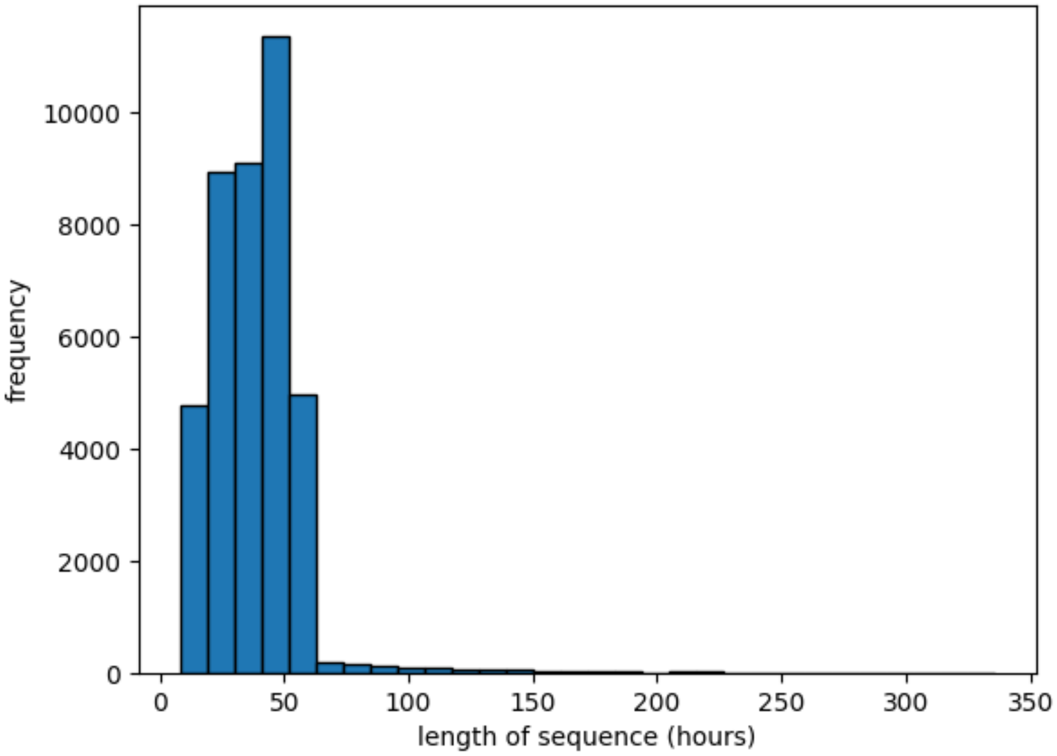


Figure 2.15: The distribution of data lengths for each patient.

2.4.2 Feature Extraction

Since all participants were using the same dataset, how they ”engineered” the features and how they chose and tuned their models would be the crucial factors of their final performance. In the following section, we will discuss several issues raised previously and how they were addressed.

Part of the feature engineering work is to build up an appropriate feature set by selecting and expanding the raw features. One of the feature selection ideas is to select features according to their importance calculated by certain methods. For every feature, t-test was conducted to compare the difference between sepsis and non-sepsis data, and finally 17 features with p-value less than 0.1 were selected [107]. Tree-based models were trained as feature selectors due to their inherent ability to evaluate feature importance, which is calculated by summing up the gain of the loss function when splitting on a particular feature. With a Random Forest model, 11 features of high importance were picked from the 40 variables [108, 108]; Similarly, LGBM and XGBoost models were built to generate the feature importance, in order to select the most valuable features [109, 110]. lyra et al. [111] created a new measure ”Normalised Observed Utility” (NOU) to calculate the feature importance as the performance loss when removing a feature during the training. Feature importance was represented by Joint Mutual Information (JMI) [112] which is the information between the target class and a random variable [113].

In addition to selecting valuable features from the given feature set, various types of derived features were created. The most common way to generate extra features is to incorporate statistical characteristics such as mean, variance, max, and min [111, 114, 115]. Except for these basic variables, a lot of other creative features were constructed, summarised in table 2.5.

Table 2.5: Novel Features Constructed in Reviewed Challenge Works

Reference	Novel Features Constructed
Noorzadeh 2019[107]	Sliding window based features i.e. skewness, kurtosis and quantile information were extracted to capture local and global information.

Reference	Novel Features Constructed
Pawar 2019[116]	Two simple change indices were created to investigate temporal relationships: first derivative and changes from baseline which the first observations after admission as the reference.
Schellenberger 2019[117]	1. Last reliable: Count how many hours have passed since last reliable value was observed, indicating the missing patterns, also seen in [114, 118, 119]. 2. Differences: changes compared to the previous measured value is included as a feature, to emphasise the difference over time, also used in [114, 120]. In [121], difference was slightly adjusted to 6 hours before rather than one hour before, focusing more on long term difference. 3. qSOFA and SOFA score which were used in [121, 122] as well.
Sarafrazi 2019[123]	To capture the clinician's suspicion, a new indicator was designed by checking the lab test time. Value 2 is for a newly reported test, 1 for a non-expired test, and 0 for either an expired or never-ordered test.
Sarafrazi 2019[123]	The authors employed the idea of anomaly detection, treating non-sepsis data as normal and identifying sepsis data as anomaly. Experiments showed that the reconstruction error is generally higher for sepsis data than non-sepsis data. The reconstruction error was incorporated as one feature for the further classification.
Patidar 2019[124]	By applying a genetic algorithm based optimisation algorithm, a set of clinically significant features are selected based on the normalised utility score. They also proposed a novel ratio and power-based feature up to the order of three, see following equation.

$$R = \frac{x^k}{y^m z^n} : x, y, z \in P; -3 \leq k, m, n \leq 3 \tag{2.11}$$

Reference	Novel Features Constructed
Biglarbeigi 2019[113]	Three screening tools - Prehospital Early Sepsis Detection (PRESEP), SOFA and Systemic Inflammatory Response Syndrome (SIRS) are all well accepted as decision support systems so variables examined in the three clinical screening tools are selected.
Wang 2019[121]	The author trained a LSTM network with a subset of raw data, then removed the output layer, used the hidden layer as features extractor.
Liu 2019[125]	Heterogeneous consecutive clinical events in a short period were aggregated, and temporal interactions among them were mined then.
Narayanaswamy 2019[126]	Defined a synthetic risk feature, where if the lab values for Lactate, pH and/or WBC exceeds a threshold for the selected time step.
Zabihi 2019[110]	Two types of features were extracted - first one is some common features from data whose availability ratio is higher than 70%, and another type focuses on the missing patterns, specifically mean and variance of the lengths of the missing value sequences and non-missing value sequences along each covariate.
Morrill 2019[127]	Two new features ”ShockIndex” which is defined by heart rate divided by the systolic blood pressure and ”BUN/CR” which is the ratio of levels of bilirubin to creatinine were introduced. In addition, there is another auxiliary index called ”PartialSOFA”, literally it is the score calculated according to SOFA rule, but only by the variable available in our dataset. To cooperate with ”PartialSOFA”, the author created the fourth new feature ”SOFA_Deterioration” indicating whether ”PartialSOFA” has increased in the past 24 hours.

The above table lists some of the extra features derived from the raw data, but there are other works which highlight the way these new features were discovered, and these are also worth discussion. As we presented in Section 2.3.2, using a sliding window of different sizes is an effective way to extract local patterns from time series data [109, 118, 125, 128–132]. Data binning was found useful in improving the result slightly by aggregating variables with a continuous value into ranges and intervals. It could reduce some variance in the signal, and the generated discrete values could be further converted to one-hot features [125, 129, 132]. Du et al. [120] examined the histogram of each feature, then applied log transformation to the data with exponential or long tail distribution, to make them more close to a Gaussian distribution. However, it is based on the underlying theory that the observed signals of patients obey the Gaussian distribution, which is not necessarily true. The autoencoder is a type is type of neural network that can learn the core features from the input and perform dimensionality reduction by reconstructing the feature set. It is employed to extract key factors from the raw dataset, whilist shrinking the size and complexity of the potential model [123, 130, 131]. Yao et al. [133] made an improvement upon the basic autoencoder, replacing the multi-layer neural network with LSTM units to make it a Temporal Autoencoder. While combining it with the vanilla version, they developed a hybrid spatio-temporal model that can learn from both spatial and temporal domains.

Before feature extraction, some researchers applied outlier detection techniques to filter out those abnormal values that could potentially affect the prediction results. The boxplot is a popular statistical graph that can reveal basic statistical characteristics of a dataset, and one of its functions is to identify the outliers, so it was used to clean up the outliers in the raw dataset before feature extraction [134]. Another way is called the ”plausibility filter”, invented by Firoozabadi and Babaeizadeh [135], and it identifies a range of valid values for each feature based on its actual distribution and the knowledge in the literature. Any value outside this range was treated as an outlier and marked missing for further imputation. Normalisation is another necessary operation for those scale sensitive models. Option one is the min-max operation which is defined as Eq. 2.12.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.12}$$

It is easy to implement but it has a significant downside that it cannot handle outliers very well [108, 126, 128]. Option two is the widely accepted z-score normalisation defined in Eq. 2.13, where μ is the sample mean and s is the sample standard deviation.

$$x_{norm} = \frac{x - \mu}{s} \tag{2.13}$$

This method is slightly more complex than the previous one, but outliers can barely impact the result, making it a more robust way to scale data into the same range [109, 110, 116, 120, 125, 135, 136]. There were other feasible normalisation approaches but not as common as the aforementioned two, such as a special normalisation method to fit all the features in the range of 1 to 5 [137], see Eq. 2.14

$$y = \begin{cases} 0, & \text{if } x \text{ is missing} \\ \frac{4(x - x_{min})}{x_{max} - x_{min}} + 1, & \text{otherwise} \end{cases} \tag{2.14}$$

2.4.3 Missing Value Handling

One of the biggest limitations is that there is a great portion of missing values in the raw dataset and a variety of missing value handling methods were applied to address this issue. First, let's examine some standard techniques. Forward-filling, which replaces the missing value with the last known value, was mostly used among all the works we have reviewed [109, 113, 115, 116, 118, 120–123, 127, 129, 135, 138–140] due to its easy implementation and explicit interpretability. Mean-filling is another straightforward method to fill the NaN values. This method simply calculated the average of each feature for every patient and substituted for non-available values [115, 116, 120, 129–131, 134, 135, 137, 141]. However, the main weakness of mean-filling is that if all items are not available, there will be no way to calculate the average. Except for filling with mean values, we also can consider constant values like -1 [108, 128, 138] and 0 [108, 121, 136, 142], since the specific constant values are unlikely to be part of the normal collected data of patients, thus indicating the missing status.

Among our reviewed papers, there were other complicated methods that have been explored to handle the missing value issue. Linear interpolation is to construct a linear model based on at least three available values, and then use this model to predict the missing ones [110, 124, 143, 144]. The linear model could be something even more complicated, for example a Markov Chain [113], which can simulate the columns that have not one single available value based on the assumption of multivariate normal distribution. Unlike existing imputation methods with certain assumptions of data distribution, Recurrent Imputation of Time series (RITS) treats the missing data as variables in a Recurrent Neural Network, and it is able to be updated through the back-propagation process[145, 146]. Another practical way to deal with the missing data is missing-mask, which contains 0 if a feature is missing and 1 if the value is monitored, see Eq 2.15. We believe that the presence and presence rate play a significant role in determining whether a patient is susceptible or not of developing a sepsis, since the measurement of a given lab value is highly informative of the patient's health deterioration.

$$m_t^d = \begin{cases} 1, & \text{if } x_t^d \text{ is observed} \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

The approach like missing-mask is more like a utilisation of missing values rather than an elimination, which brought a new angle to deal with missing values. However some researchers raised their concerns about the usage of certain standard interpolation methods, because future values of patients is unknown in the real case in the hospital. As a consequence, some of the standard interpolation becomes infeasible as unknown features in the future should not be involved in the missing value processing [137].

2.4.4 Countermeasures to Imbalanced Data

The second major limitation is that the data distribution is extremely imbalanced in terms of sepsis and non-sepsis data. To address this problem, generally we have three possible ways: 1. under-sampling, 2. over-sampling, 3. customised weighted loss function. The ordinary under-sampling method is to randomly select data samples from the majority class, so that the minority class could reach the same level [107, 113, 115, 117, 122, 131]. Mutual Information (MI) is a measurement of uncertainty of one random variable given another one, and it can be used to enhance the original random sampling approach. Assigning a score which is the sum of MI to all other patients, to each patient, grouping patients with similar scores together because they are considered highly dependent on each other, and finally, proportionally sampling from each group to generate the representative dataset are steps comprising MI based under-sampling [141].

An opposite approach for under-sampling from the majority class, is to over-sample from the minority class [108, 119]. A very classical over-sampling method - Synthetic Minority Over-sampling Technique (SMOTE) was used to rebalance the sepsis-normal ratio in the raw dataset [141]. A more complex over-sampling scheme, the Adaptive synthetic sampling algorithm, was employed in [142]. It was first proposed by Haibo et al. in 2008 [147]. Different from the SMOTE method, the key idea of ADASYN is to use a density distribution as a criterion to automatically determine the number of synthetic samples that need to be generated for each minority class sample, rather than equally generating the same number of synthetic data for every sample. This algorithm will also bring a benefit that forces the learning algorithm to be biased towards those data samples that are hard to learn. Unfortunately, despite experiments confirming this, they did not show better results compared to random under-sampling. Extending the prediction time from six hours up to twelve hours will result in the enrichment of sepsis data and the imbalanced situation could be slightly mitigated [111]. This tiny adjustment could be somehow treated as a special type of ”oversampling” whose purpose is to increase the number of data samples of minority class.

The third way to alleviate the impact of an imbalanced dataset focuses on the algorithm. When designing the loss function, higher weights are assigned to the minority class, so misclassification of this class will cause more penalty consequently, and we call this weighted loss function [117, 121, 137]. With a weighted loss function, the designed model biases towards the class with fewer data samples, which neutralises the impact caused by the imbalance in datasets. Adding a small number to the predicted probability of sepsis can lead to more positive cases, and the number was chosen as 0.0002 in [121].

2.4.5 Predictive Models

With the same given dataset, choosing an appropriate model will be another key to achieving a high performance in the competition. According to our review, participants mainly chose to use two types of models: gradient boosting machine (GBM) and deep neural networks (DNN), see Fig. 2.16 and 2.17.

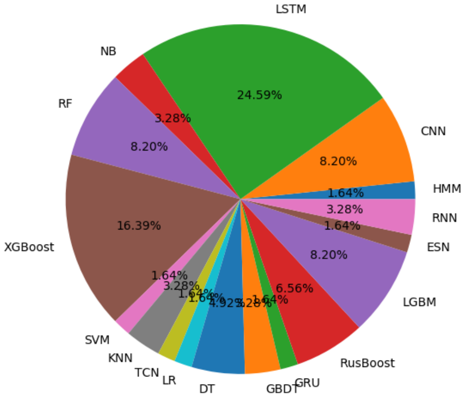


Figure 2.16: Distribution of models used in PhysioNet Challenge.

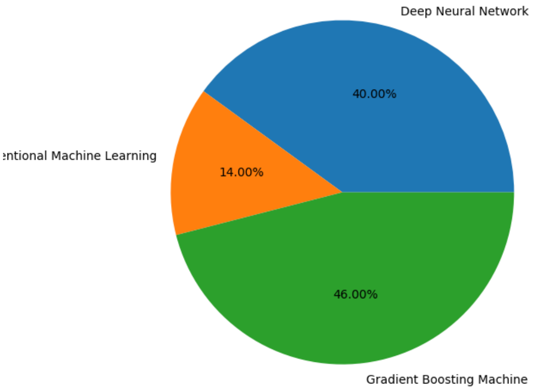


Figure 2.17: Distribution of models (categories) used in PhysioNet Challenge.

The idea of GBM is simple: it is basically an ensemble model that strategically integrates a set of weak classifiers to get an enhanced outcome. Each weak classifier learns the residual of the last one iteratively to minimise the final error. Gradient Boosting Decision Tree (GBDT) was one of the typical GBM models known for its great performance in most cases, four teams (9.8%) used it as the main classifier in the competition [118, 120, 122, 141]. Around 16.39% of the studies used XGBoost [114, 121, 122, 132, 148] which is getting more and more popular in recent years due to its effectiveness in a wide range of task especially prediction problems. Another 8.2% of the works chose Light-GBM, a more efficient and faster model, inspired with the novel histogram algorithm [109, 115, 127, 129, 140]. Combining undersampling with the boosting technique is the major difference between the Random Undersample Boost (RusBoost) and others. It is a modified version of Adaptive Boost (AdaBoost) with an extra step of under-sampling before the actual training, in which way, the disadvantage of under-sampling - loss of information could be overcome greatly [124, 139, 143].

Since the size of the dataset is relatively large, considering the total number of more than 40,000 records, complex neural networks were utilised, especially the RNN, which is good at extracting temporal relations. Among all the variations of RNN, LSTM stands out and was employed in many works in this competition [121, 122, 126, 130]. What's more, Liu et al. [125] used the LSTM bidirectionally to learn the latent patterns hidden in aggregated homogeneous clinical events; residual connection was added to the LSTM where the input of every block was a concatenation of the output of the previous block [119, 137]. During the training of the LSTM network, hyper-parameters were tuned through a differential evolution genetic algorithm, so that the potential of LSTM could be fully explored [138]. Although it has been developed for a long time, RNN with Gated Recurrent Unit (GRU) was still worth trying. Nonaka and Seita improved the original GRU in order to consider the static information available. The initial hidden states of GRU were learned from a fully connected neural network fed with the demographic information of patients [136]. Further examples of exploiting deep neural networks include a completely fully connected neural network, cooperating with a GBDT and a RF model, under the strategy that any classifier that predicts a positive outcome leads to an overall positive [131], and an echo-state network which is a subset of RNN that leverages Reservoir Computing properties [144].

Model ensemble was one of the many other interesting ideas we found during the review. Basically it is a strategy that integrates results from multiple models, like the very typical random forest model [111]. A CNN and a LSTM were combined under a bagging ensemble strategy [108] similar to the Conv-LSTM we discussed in the previous session, whereas a stacked ensemble of RNN and LGBM was explored in [109]. Two specific models were developed and pipelined in the work of Pimentel et al., the first one was to predict sepsis for each patient using common features like statistical variables, and the second one was to estimate the uncertainty of prediction generated from the first model with special features related to ”missingness”.[114]. TASP is a novel time-phased model, comprising two models according to the length of stay in the ICU. For the first 49 hours after admission into the ICU, LGBM models were applied, but for the time after the 50th hour, it is the RNN working to capture the long-term temporal relations [140].

2.4.6 Performance Evaluation

This competition has defined an official performance measurement instead of accuracy: the normalised utility score, which is a weighted sum of correct prediction plus penalty to the misclassification. The highest score is rewarded if the sepsis is detected at the optimal time which is six hours before the onset. Once the correct prediction is made but a bit earlier or later than optimal time, scores are also rewarded but not as much as the optimal one. On the other hand, if a non-sepsis patient was misclassified as sepsis at any time point, a small penalty will be inflicted. Please refer to Eq 2.16 - 2.21 for the exact rule of utility score calculation, where for each patient s at each time point t, an individual score is calculated according to the predicted outcome and its correctness, and finally summing them up. To improve the interpretability, the final score is normalised as Eq. 2.21 so that the optimal classifier (highest possible score) receives a normalised score of 1 and that a completely inactive classifier (no positive predictions) receives a normalised score of 0. In these equations, t_{sepsis} is the time sepsis onset occurs, $t_{optimal}$ is exact six hours before the onset, t_{early} is 12 hours before the onset which is the earliest time the prediction is expected at, and t_{late} is three hours after the onset after which the prediction will not be rewarded.

$$U_{TP}(s,t) = \begin{cases} -0.05, & \text{if } t < t_{early} \\ \frac{t-t_{early}}{6}, & \text{if } t_{early} \leq t \leq t_{optimal} \\ \frac{t_{late}-t}{9}, & \text{if } t_{optimal} < t < t_{late} \\ 0, & \text{if } t \geq t_{late} \end{cases} \tag{2.16}$$

$$U_{FN}(s,t) = \begin{cases} 0, & \text{if } t \leq t_{optimal} \\ -\frac{2(t-t_{optimal})}{9}, & \text{if } t_{optimal} < t < t_{late} \\ -2, & \text{if } t \geq t_{late} \end{cases} \tag{2.17}$$

$$U_{FP}(s,t) = 0.05 \tag{2.18}$$

$$U_{TN}(s,t) = 0 \tag{2.19}$$

$$U_{total} = \sum_{s \in S} \sum_{t \in T(s)} U(s,t) \tag{2.20}$$

$$U_{normalisation} = \frac{U_{total} - U_{no\ prediction}}{U_{optimal} - U_{no\ prediction}} \tag{2.21}$$

We have summarised the utility score for all the work we have reviewed in the following Table 2.6. The individual utility score is extracted from the official final results and is averaged with every work that uses the same model if necessary. We further calculate the overall utility score for every model category by simply using the mean value of all the models belonging to it. From the table we find that the GBM models have the overall best performance in the competition, followed by DNN and others, which aligns with the popularity of mod-

Table 2.6: Performance evaluation of models in PhysioNet Challenge.

Category	Overall Utility Score	Model	Utility Score
GBM	0.301	GBDT	0.339
		XGBoost	0.250
		LGBM	0.36
		RusBoost	0.256
DNN	0.275	LSTM	0.274
		TCN	0.342
		CNN	0.236
		RNN	0.279
		FCNN	0.284
		GRU	0.323
		ESN	0.188
Others	0.234	CRF	0.190
		LR	0.249
		RF	0.228
		KNN	0.270

Legend:
GBM: Gradient Boosting Machine; GBDT: Gradient Boosting Decision Tree;
XGBoost: Extreme Gradient Boosting; LGBM:Light Gradient Boosting Machine;
RusBoost: Random Undersampling Boost; DNN: Deep Neural Network;
LSTM: Long Short-Term Memory; TCN: Temporal Convolutional Network;
CNN: Convolutional Neural Network; RNN: Recurrent Neural Network;
FCNN: Fully-Connected Neural Network; GRU: Gated Recurrent Unit;
ESN: Echo State Network; CRF: Conditional Random Field;
LR: Logistic Regression; RF: Random Forest; KNN: K-Nearest Neighbour;

els we discussed in section 2.4.5. GBM and DNN models are mostly applied among all the reviewed work, since they have better than average performance. The reason why GBM can outperform models with a deep neural network architecture probably relates to the small number of sepsis data records and the complex nature of this disease.

We also verified the effect of ensemble strategy by comparing the utility score of individual models and that of them combined, see Table 2.7. By applying the ensemble strategy of multiple models, it is evident that the performance could be improved significantly. One of the ensemble model integrating XGBoost and GBDT [122] achieved the highest score of 0.406 and the only one above 0.4 among all reviewed works.

Table 2.7: Comparison of performance of individual models and ensembles.

Model	Utility Score	Improvement
CNN	0.236	22%
RNN	0.279	
CNN+RNN	0.288	
XGBoost	0.250	62.4%
GBDT	0.339	
XGBoost + GBDT	0.406	
XGBoost	0.250	35.6%
Ensemble of 5 XGBoost	0.339	
GBDT	0.339	38.7%
RNN	0.279	
GBDT + RNN	0.387	

Legend:
CNN: Convolutional Neural Network; RNN: Recurrent Neural Network;
GBDT: Gradient Boosting Decision Tree; XGBoost: Extreme Gradient Boosting;

2.5 Conclusion

Our task in this review chapter is to address the RQ-4 by systematically reviewing relevant works in the area of sepsis detection published in the recent decade and determining whether they can satisfy current requirements with regard to sepsis detection. We developed a classification framework that analysed 102 academic papers by answering ten review questions and sub-questions. Our study found that AI based techniques have been widely applied in healthcare, more specifically in sepsis related detection, and have promising performance.

The proposed classification framework provides an insightful analysis for current works of sepsis detection showing a clear roadmap to the researchers who are new in this research field. In-depth analysis and comparison of pre-processing techniques, learning models and performance evaluation were manifested to prove that predicting sepsis related symptoms with AI based technologies has an earlier warning, and significantly increased accuracy, reducing waste of valuable medical resources and mortality rate. Our review also summarised efforts put into the PhysioNet Challenge 2019 separately, elucidating various innovative ideas for solving the challenge problem, which is to predict sepsis with clinical data. The normalised utility score was defined as a customised performance metric to measure how the models performed on the given datasets. The proposed utility score is calculated based on prediction time and accuracy, so a correct prediction made six hours prior to the onset will have the highest score.

Meanwhile, our review identified several challenges and gaps researchers are facing in their work, such as imbalanced data and missing values issues that are not completely solved, weak interpretability of complex deep learning models, and the lack of a unified gold standard for sepsis diagnosis around the world. One salient gap existing in the reviewed works is that there is no universal framework to process and predict sepsis with heterogeneous data from different sources, hence a complete workflow of AI boosted sepsis detection methodology is needed as a decision support system for clinicians and researchers who have to deal with complex data to standardise their working procedures.

Chapter 3

Research Methods

In this chapter, we describe the Action Research methodology that we used in this thesis, as well as how we adapted the theory to our project. Further, we propose a unified framework for sepsis detection under which multiple algorithms and models can be developed to predict sepsis in both neonates and adults.

3.1 Action Research

3.1.1 The Origin of the Research Method

World War II brought massive damage to research society of social science, and at that time a new research methodology was needed. Kurt Lewin was developing the method in order to study social psychology within the framework of field theory [149], however another group at Tavistock Clinic developed a similar method which studied the psychological and social disorders caused by battlefields and prisoner-of-war-camps. At that time, scientists were not able to understand the complex causes of social illness and make universal treatment, because each case seemed different in some ways. Hence, scientists intervened in the experiments, being involved with the subjects and changing some aspects of the subjects' being or surroundings to get a deeper understanding.

Originally, the essence of Action Research was just a simple two-stage process [150], i.e. a diagnostic stage in which the problem is analysed and hypotheses are formulated, and a therapeutic stage in which hypotheses are tested by collaboratively changing conditions of experiments. However, as the method developed, more and more detailed steps were added to the original model. Lewin's original model [149] has six stages, i.e. (1) analysis, (2) fact-finding, (3) conceptualisation, (4) planning, (5) implementation of action, and (6) evaluation. Although it was not completely the same as the action research we are discussing nowadays, it was very similar though. The most prevalent description proposed by Susman and Evered [151] elucidated Action Research as a five-step cyclical process, required the client system to have an infrastructure to maintain and regulate some or all of these five phases together with researchers. The five steps include: (1) diagnosing, (2) action planning, (3) action taking, (4) evaluating and (5) specifying learning, and they are iterated during the experiment as a research cycle, see Fig. 3.1. It is noticed that, the five phases here have already been what Action Research is defined today.

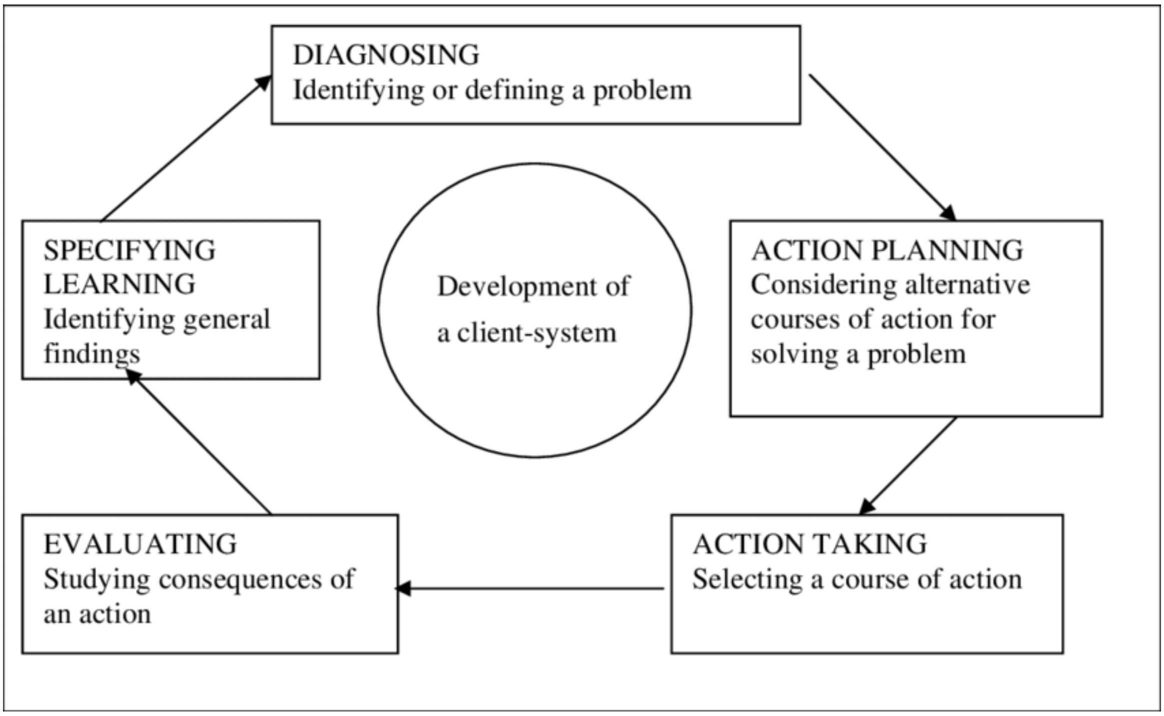

DIAGNOSING
Identifying or defining a problem
ACTION PLANNING
Considering alternative courses of action for solving a problem
ACTION TAKING
Selecting a course of action
EVALUATING
Studying consequences of an action
SPECIFYING LEARNING
Identifying general findings
Development of a client-system

Figure 3.1: The cyclical process of action research [151].

Action Research requires a client-system infrastructure, which in other words, is the norm and the protocol that will comprise the research environment. This infrastructure must define the responsibilities, boundaries, entries and exits of both researchers and clients. One of the key aspects of the infrastructure is the immersive role of researchers during the experiments. Research scientists work closely with the practitioners within the client system in order to get a better and deeper understanding, identifying the anomalies being studied.

Forms of Action Research

Action Research refers to not only a specific research method, but more of a set of research approaches sharing some similar characteristics which distinguish Action Research from other social science research methods. Four major characteristics of Action Research from Hult and Lennungs definition [152] are:

1. It aims at increasing understanding of an immediate problem within a certain social situation.

2. It facilitates practical problem solving and theoretical scientific knowledge expansion. Furthermore, two derived characteristics from this are:
    (a) Highly interpretive assumptions are made about observations.
    (b) There exists researchers' intervention in the problem setting.
3. Action Research is performed collaboratively and enhances the competencies of respective actors involved.
4. Action Research primarily applicable for the understanding of change process in social systems.

Another survey by Peters and Robinson [153] summarised four characteristics the methodologies within the class of Action Research have in common:

1. action and change oriented
2. problem focusing
3. systematic and iterative
4. collaborative

Lau [154] categorised them into four subtypes: action research, participatory action research, action science and action learning. Action Research has been described as a research method that is characterised by the intervention experiments on problems or questions extracted from the social context by practitioners. Participatory action research is distinguished by extra features which is the composite role as practitioners of both subjects and researchers. What makes action science different is the characteristic of a central emphasis on the spontaneous theories that participants bring to practice and research [155]. Action learning advocates group participation, programmed instructions, real actions and experiential learning within the social and organisational context.

Five phases of Action Research

The five cyclical phases in the research method are symbolic of Action Research, and they define a standard workflow when Action Research is applied in research.

The diagnosing phase, as the first step in the method, corresponds to the identification of problems and the proposal of hypotheses. In this phase, the context of research is collaboratively analysed to reveal the underlying cause, based on which, reasonable hypotheses are developed to be verified in the following phases. When AR is applied in social research, diagnosing often involves with systematic interpretation of the complex organisational problem in a holistic way [156].

The second phase is action planning, where plans for the actions that should be undertaken are made. Those actions should be able to answer the hypotheses brought up in the previous step. To make the plan, researchers and practitioners are supposed to collaborate like in the first phase. The action plan is guided by the theoretical framework, which indicates both some desired future states and the changes that would achieve such states. Besides, the plan set up the target of the change and the approach to change.

The plan is then implemented in the next stage, action-taking. Changes are realised in the client organisation by the intervention of researchers and practitioners. There are several different intervention strategies that are possible to apply. Directive intervention means it guides the changes directly, but on the contrary, changes will not be directed by the intervention in non-directive approach. Some other intervention tactics could be introduced to help the actions and changes be implemented correctly and smoothly. In other words, the action-taking phase is to select a course of action towards the target defined in the previous steps.

Evaluating phase assesses and analyses the outcome of changes, determine how the actions and changes are implemented and whether they are effective to solve the problem. If the changes were successful, it further evaluates whether the undertaking actions are the sole cause among all complex factors from both inside and outside the environment; otherwise, it figures out the reason of failure, reflects the initial hypotheses, adjust them if necessary, and designs an improved plan for the actions should be undertaken in the next iteration.

The focus of evaluating phase is on the outcome of changes, however, in the last phase - specifying learning, it is on the knowledge gained during the process. What gained knowledge could provide includes three aspects:

- It could be fed to ”double-loop learning” process [157] which reflects not only the action that has been done, but also the initial hypotheses. Unlike single-loop learning which is concerned with how to ”do things right”, double-loop aims at ”doing the right thing”. Through the process of double-loop learning, the organisation is able to reconstruct the norms with the help of gained knowledge.

- For the places where changes were not successful, the additional knowledge may provide sufficient and helpful information for diagnosing problems so that they could be fixed by improved intervention in the future iteration.

- The outcome itself, no matter success or failure, is important for the theoretical framework. It provides informative knowledge for the scientific community for future research.

The cycle of action research phases can continue regardless of whether the changes are successful or not, to investigate more knowledge of the organisation and verify the relevant theoretical framework. As a result, the organisation learns more about the nature of itself and is improved, while the research community can benefit from the continuous process and evolve.

Participatory Action Research

The traditional action research approach has been extended into a form known as ”Participatory Action Research”, and one of the changes is the realignment of the roles of researchers from observers to more of participants. It emphasises more on understanding and solving the problem rather than theory building. Researchers are not supposed to stay outside of the experiment observing, but to proactively engage in it with the subjects.

It is not necessary for the researchers to come out with the expected theoretical solution in the action plan, because solving the immediate problem requires in-depth and sufficient domain knowledge, of which the client professionals live in the context possess much more, than the researchers. Problem solving and action improvement could be achieved through full and extensive cooperation among researchers and client subjects in an organisation. Their mutual support is a significant benefit of the Action Research method, and could push the research toward the right direction.

3.1.2 Comparison with Design Science

Design Science is a research methodology conceptualised by Herbert Simon in 1996 [158]. It supports a pragmatic research paradigm that encourages the creation of IT artefacts to solve real-world problems. Back to the advent of Design Science, it is thought to be used in technical disciplines, but in early 1990s, the Information System community realised that Design Science could also be adopted to improve the effectiveness and utility.

There are many similarities when comparing the cyclical process of Action Research to the general workflow (see Fig. 3.2) of Design Science. For example, both of them have a five-step process with almost identical meanings, yet with different names. If we dig deeper, we can further compare their characteristics in a more detailed manner. Jarvinen [159] extracted six pairs of characteristics of both research methods from the literature review, and we will contrast them with each other.

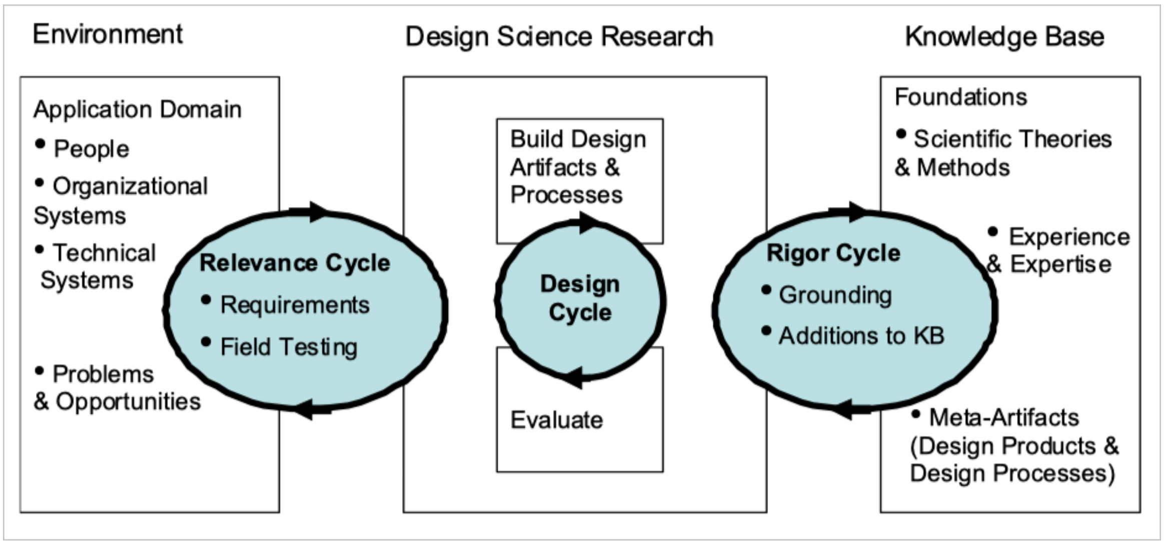


Figure 3.2: Design science research cycles [160].

AR1: Action research emphasises the utility aspect of the future system from peoples' point of view [151]

DS1: DS4: Design sciences products are assessed against criteria of value or utility [161].

One of Action Research's properties Susman and Evered [151] summarised is that it is future oriented, and should be designed to create a desirable future state. March and Smith referred to Simon's work [158], and gave a more specific research target: "It should be judged based on value or utility to a community of users"

AR2: Action research means both action taking and evaluating [162].

DS2: Building and evaluation are the two main activities of design science [161].

From the definition of the cyclical process, it is obvious that AR3 is correct. For Design Science, March and Smith thought evaluation is to develop criteria to assess the performance of artefacts (i.e. construction, model, method or instantiation) that have been made, so design science involves both building and evaluating, just like Action Research.

AR3: Action research modify a given reality or develops a new system [163].

DS3: Design science produces technical artefacts. [161].

Both Action Research and Design Science generate outputs, but the former produces a new system, while the latter produces four types of artefacts proposed by March and Smith [161].

AR4: Action research is carried out in collaboration between action researcher and the client system [151], and it requires action researchers to intervene [152].

DS4: Design science research is initiated by researchers interested in developing technological rules for a certain type of issue. Each individual case is primarily oriented at solving the local problem in close collaboration with the local people [164].

Literature has reached an agreement that a researcher's intervention is the symbolic feature of Action Research. Similarly, researchers and practitioners participate in experiments to solve problems in collaboration with local people. Both methods emphasise collaboration in research, and it has been reported that a lack of collaboration could be one of the reasons that leads to failure.

AR5: Knowledge is generated, used, tested and modied in the course of the action research project [165].

DS5: Knowledge is generated, used and evaluated through the building action [166].

Another output of both methods is knowledge, although in different forms.

AR6: Action research produces knowledge to guide practice in modication [163].

DS6: Design science produces design knowledge (constructs, models, methods and instantiations). [161].

Action Research produces knowledge concerning action guidance, while the reality is being modified by the action simultaneously. All the four types of products produced by Design Science could be treated as knowledge connected to design.

Why Action Research not Design Science

Design science is one of the major research paradigm that dominate Information System research. The target of Design Science is to create relevant artefacts that can extend the boundaries of human and organisational capabilities. According to Coughlan and Coghlan [167], Action Research is an appropriate approach when the research relates to understanding of the process of change or its improvement to learn about it. Although both Action Research and Design Science have been applied in Information System for decades, we choose use Action Research as our research method rather than Design Science. The main reason is that the symbolic five-step process fits our research workflow to a great extent, and it will be explicated in detail in section 3.1.3. Besides, there exist a series of challenges that Design Science is facing in IS community [168]. We summarise a few here:

1. The engineering discipline is a young field that Design Science has been applied in, and not very much cumulative theory basis has been built [169]. It is important to demonstrate the feasibility and utility of the technological oriented adoption of the social oriented research framework.

2. Design Science is not completely able to represent the technological environment due to insufficient sets of artefacts (constructs, models, methods and tools). It is criticised that the descriptive IS models do not have strong underlying theory base, which is to say finding an appropriate trade-off and balance between abstract theory and practical technology is difficult.

3. Artefacts created embody the understanding of the problem by researchers, but the artefacts themselves are sometimes not perfect and need upgrading and improving, because the existing knowledge base of Design Science is often insufficient, and not able to satisfy the increasing demand of the constantly changing situation of the environment as well as the technological discipline [170].

4. Design Science research is perishable due to the rapid development of technology. Emerging new tech could possibly invalidate existing artefacts before they are actually implemented or put into practice.

5. It is difficult to apply rigorous evaluation standards in Design Science research, for instance, a specially designed artefact may not generalise to different scenarios [171].

3.1.3 The Adoption of the Research Method

Application in Technical Research

Action Research was initially designed for social science research, but there has been a gradual transition since the 1970s to socio-technical disciplines or even purely technological projects. The area that adopts Action Research first and is most prevalent is Information Systems. According to [154], all the four types of Action Research, i.e. Action Research, Participatory Action Research, Action Science, and Action Learning, were applied in IS research. However, note that some of the papers did not clarify which one was used but only cited Action Research, and some other papers defined Action Research in ways which seem to deviate from our understanding. To illustrate, we take some cases as examples as follows. Fox [172] explored and discussed the development of the principles of designing a sociotechnical system, which blended two different aspects: social and technical. These two aspects must be considered independently due to the discrepancies in each system. Action Research was applied to conduct the system scan, technical analysis, social analysis, and finally summarised the reflection that staff seek more meaningful empowerment, greater productivity and viability in the organisation. In the work by Timpka et al. [173], five years of Participatory Action Research have been performed in the development of a medical hypermedia system. This is a typical case that discusses the collaboration between users and developers through Participatory Action Research. "Action Science" was one of the mainstream research methods that was able to learn from individual actions and practice into a model, helping the organisation to understand how and why individuals behave as they do during organisational IT implementation [174]. Finally, action learning has also been mentioned in [175], but action learning was only cited as the basis for the training and no clear definition was provided. The review [154] was conducted in 1997 which means that the Action Research methodology has been widely used in the area of Information System as early as the last century, however with the increasingly blurring border between social and technical research, Action Research has been adopted in more domains and reported lately. Staron published a book [176] to elucidate the application of Action Research in software engineering (SE). He claimed that the results of traditional SE experiment are difficult to be transferred to real industry, but its another story with Action Research due to its focus on the intervention, the context and learning. Ochodek et al. [177] conducted an Action Research project aiming to develop a machine learning method to detect lines of codes that violate coding guidelines which is another example of application in SE. This work follows a classic workflow of exploring a machine learning based solution with an Action Research methodology like ours: 1) understanding the problem, 2) experimenting with machine learning models, 3) evaluating the tools and outcomes, 4) reflecting. Academic research in applied disciplines such as construction engineering and management has the dual mission of simultaneously contributing to the solution of practical problems and establishing theoretical and conceptual knowledge [178], and Action Research happens to be the appropriate method to fulfil this particular demand. Also in the work by Coughlan [167], it was proven that Action Research is relevant and valid for the discipline of operation management (OM), and operation managers and researchers can learn from the applied activities that characterise the practice of OM by Action Research.

Application in Our Project

In this section, we will discuss how we adopt the Action Research method throughout our project in terms of the unique five-step pattern.

Diagnosing Diagnosing is the first step in which the problem is pinpointed and hypotheses are proposed. We identified the neonatal sepsis issue according to the demands from clinicians work in the frontline of newborn care. The commonly used approach to neonatal sepsis is the administration of empirical antibiotic therapy [9, 19]. To prevent deterioration, clinicians are encouraged to use antibiotics before the result of a blood culture comes out. The excessive use of antibiotics can result in antibiotic resistance, predispose to fungal infection, necrotising enterocolitis (NEC) and even death [179]. The dilemma is that if clinicians wait for blood culture result, the sepsis may deteriorate during the time, and miss the best opportunity of treatment. Under these circumstances, an early detection system is in urgent need. One of the motivations of our research is to solve this dilemma, make antibiotic administration targeted at the right patients, and avoid the overuse of antibiotics. Besides, staff fatigue due to current practice is one of the common situations that needs to be dealt with, and monitoring all kinds of vital signs from tens of cots are time consuming and vulnerable to human errors. Hence we want to introduce early detection systems into the current clinical workflow to make the diagnosis of late-onset sepsis more efficient and effective, and further extend it to general sepsis infections in adults.

We hypothesised that the pattern of clinical presentations in patients with neonatal sepsis can be summarised in data models using Machine Learning techniques and development of early detection algorithms is feasible. This theory should also work for adults with sepsis, since they are the similar cases in many ways. Clinical experience indicates that sepsis onset comes along with particular symptoms like many other diseases, so with continuously collected data with various types of sensors from patients, we could monitor the status and analyse the pattern in normal and infected cases.

Action Planning Based on the problems and hypotheses we had made, macro and micro plans were drawn up in detail. First, we proposed six research questions and three research objectives as guidance in the general direction, clarifying what we expected to get at the end of the research.

Research Objectives (RO):

RO-1: Design an efficient data collection scheme.

RO-2: Investigate the relations between physiological parameters and sepsis, and find the critical ones that are most related to sepsis.

RO-3: Design, develop and implement suitable algorithms for clinical data, and make some adjustments to improve the performance.

Research Questions (RQ):

RQ-1: How should the vital signs be collected from bedside monitors in NICU?

RQ-2: How many physiological parameters are available to researchers?

RQ-3: Which critical physiological parameters can predict sepsis before it occurs?

RQ-4: Can existing methods in literatures fulfil the the current requirement of early detection?

RQ-5: Is it possible to adjust existing methods to improve their performance?

RQ-6: Are there any new approaches more suitable for this early detection task?

Then a three-cycle experiment was delicately designed, by which we could verify and adjust our hypothesis, and help us finally achieve our pre-set goals.

Three cycles:

As the Action Research method is an iterative process, we designed a three-cycle iterative experiment to approach our goal step by step, aligning with our three research objectives. In cycle one, the primary task we need to handle is to set up an effective and efficient data collection scheme for patients in hospital, as multiple data sources and various formats of data are required in the experiment. There must be a unified and automated workflow to collect and process all the necessary data we need to conduct further analysis. The second cycle mainly focuses on figuring out the most related parameters to our task from all the available data. Analysis was made to distinguish important factors that are closely connected to the onset of sepsis, or the symptoms indicating the potential of infection. With filtered and refined datasets, models are supposed to have greater performance and efficiency. Finally, with data gathered in the previous two cycles, we explored the possibility of predicting sepsis with existing models in the last cycle. Cycle three is a self-iteration one which means it repeats, and the exploration upgrades during the iteration. We start by verifying existing models, then applyimprovements, try techniques that have never been used in sepsis detection.

Action Taking This phase corresponds to the implementation of the research plan. We chose the directive intervention tactic since clear goals have been set so we, as the researchers, can proactively make changes and guide the development of the experiment. To find the relevant parameters, we did some relevance analysis from the mathematical and Machine Learning algorithm aspects. A data collection workflow was implemented to gather data from multiple databases in the hospital. The most important part is the model training, which includes conventional machine learning algorithms, deep learning algorithms and multi-instance learning algorithms and necessary work like preprocessing data and parameter tuning were done to optimise the performance in this phase as well. The self-iteration nature of our three-cycle experiments requires us to make improvements, specify adjustments and repeat them in order to get an ideal outcome.

Evaluating What should be done in this phase basically is the outcome evaluation. Models have different results depending on the dataset, technique used, various combinations of parameters and model structures. Evaluation is essential to acquire the insights of different models, so we can make changes and improvements accordingly in the next iteration to progress. Multiple evaluation metrics were applied such as accuracy, training time, and false positive rate to get a full picture of performance as possible. The feedback from the evaluation will be analysed and then reflected in the following iteration.

Specifying Learning This is the step where summary and reflection happen. By analysing different methods, datasets used, and corresponding results, we could possibly draw some conclusions which are also part of our research outcome. Another important task is to sort out the relations between the raw data and sepsis onset, it is achieved by a complex analysis process, multi-dimensional data visualisation and other tools and methods. The output of specifying learning will be fed to a new iteration of experiments, as part of the guidance to direct further model-building.

3.2 Unified Sepsis Detection Framework

One of our contributions is developing a unified sepsis detection framework for sepsis in both adults and infants. The framework consists of three cycles and four tasks: data collection, data pre-processing, model training and performance evaluation, which covers the entire workflow of a sepsis detection task from collecting data at the beginning to evaluating and analysing in the end, see Fig. 3.3. Aligning with the Action Research framework, we designed the three-cycle experiment workflow to accomplish the pre-set goal, which is the early detection of sepsis. Each cycle contains one or more of the aforementioned tasks, and includes action taking, evaluating and specifying learning phases of the AR framework.

Following the proposed framework, one can run through the procedure step by step until having an output of sepsis prediction, no matter whether the target subjects are adults or infants. Meanwhile, this framework summarises many unique processes that other machine learning projects might not necessarily need to consider. With minor adjustments in one or more phases, this framework could be adopted to many other clinical problems that prediction time is crucial.

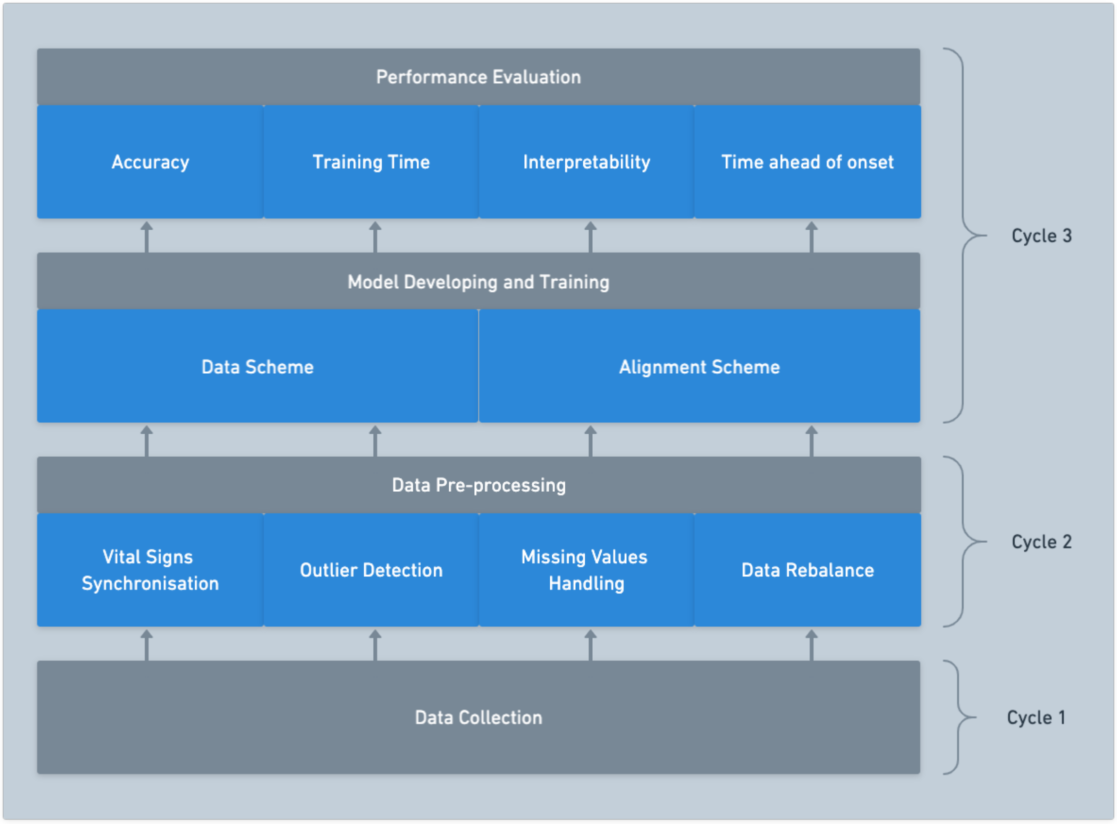


Figure 3.3: The unified sepsis detection framework.

3.2.1 Three Cycles

Cycle One

Objectives Because of the nature of clinical data, it comes from different sources in different formats, and the heterogeneity makes it difficult to efficiently collect large amounts of data for model training tasks. It is necessary to design and implement an effective workflow to facilitate the process of data collection as the quality and quantity of the data is the basic foundation of a satisfactory model.

Experimental Design In our research, we plan to use two parts of data. One is collected from the hospital, the other one is public medical dataset, such as MIMIC III. For the data from the hospital, it is relatively complex in terms of collection, since different data is distributed in different systems and databases. What we need to do includes 1) contact Dräger which is the manufacturer of the bedside monitors in NICU and get their help to extract real-time vital signs of babies under surveillance. 2) coordinate with hospital staff to get access to multiple databases such as the one that stores doctors notes, the one we can search for demographical data from, and the one that records biochemical laboratory test results. 3) Try to automate the procedure of exporting useful information from the databases in the format we want. For the public dataset, it is easier to retrieve because they have been already pre-processed for training in a standard format. Doing some research to select the most related datasets and applying them for research purposes are the only things we need to do.

Cycle Two

Objectives In the previous cycle, we focus on collecting data, but in this cycle, we shift the focus to data itself. The objectives of the second cycle lied on the relevance analysis of data, including the data that was already stored in the hospital and the data that were about to be collected. The purpose is to select the part of the data that most closely connected to sepsis, and get rid of those irrelevant ones. Although some of the advanced models could be modified to perform feature selection, but we believe that refining data in terms of features at an earlier stage is still beneficial to model training efficacy and efficiency.

Experimental Design In general, there are three types of feature selection methods i.e. filter, wrapper and embedded.

Filter method runs features selection based on certain metrics like importance or relevance before training any model, and the procedure of feature selection has nothing to do with the model itself [180]. In other words, it filters the raw feature set prior to the training. Typical filter methods include variance threshold test, Pearson correlation coefficient, distance correlation coefficient and Chi-Square correlation. The simplest way is Variance Threshold which could remove features that have variance lower than the specified threshold. The underlying theory is that low variance means less difference for data belonging to different classes, so it becomes less informative for the task of distinguishing multiple classes. Take an extreme case as an example, if one feature has the same value for all data, it is meaningless in terms of prediction. Pearson correlation coefficient is another metric that is simple but widely used to measure the relevance between features and corresponding variable. It is easy to calculate but it has a non-negligible disadvantage in that it measures only the linear correlation. If two variables are non-linear related, the Pearson correlation coefficient would also be around zero. To fix this issue, one could try Distance Correlation which is created to overcome the weakness of the Pearson method. Another technique to examine the feature relations is Chi-Square test. It tests the relevance of two variables, but only suitable for categorical variables is its weak point.

Unlike filter method, wrapper considers the performance of the prepared model as the evaluation metric, all selected features are specially desinged specifically for that model. Because of this, wrapper methods can select more suitable features than the filter methods, but the cost is much higher due to multiple training during the process. Recursive feature elimination is one of the wrapper methods, and it is implemented by recursively picking up the most/least significant features based on the coefficient or feature importance calculated by the candidate model.

The third category is embedded methods of which the representatives are regularisation-based method and tree-based method. The name embedded comes from the fact that in this method, feature selection is embedded in the process of training. L1 regularisation is able to suppress the weight of unimportant features inherently as well as L2 regularisation, but the L1 can only keep one of two highly related variables (make the weight 1 and 0 in extreme cases) while the L2 tends to average the weight of each variable and keep both making it a more stable method compared to L1 regularisation. Tree-based models have the nature of determining feature importance by counting samples falling into each leaf, for instance, random forest and XGBoost. Feature importance is available as long as they finish training, and the tree-based structure is robust to non-linear relations, these are two advantages of tree-based feature selection method.

Our plan is to utilise all the three types of methods we listed above, and then combine the generated results to achieve an appropriate feature selection outcome. For the filter method, variance threshold test and Chi-Square test were taken, distance correlation coefficient was calculated; For wrapper method, we did recursive feature elimination and for the embedded method, both L1 and L2 regularisation were tried, and xgboost was taken as the tree-based model to extract feature importance.

Cycle Three

Objectives The general objective of cycle three is to develop algorithms and models that are able to predict sepsis onset with limited data. As described above, this cycle is self-iterative which means it will repeat several times and upgrade some of the experiment settings each time. We aim to achieve multiple sub-objectives during the iteration, and the sub-objectives are: 1) Design and implement a unified sepsis detection framework for both adults and infants. 2) Embed classical machine learning models into the proposed framework to verify the performance of models and feasibility of the framework. 3) Bring in complex deep learning models with accumulated datasets. 4) Explore and adopt other techniques that could potentially improve prediction task performance.

Experimental Design Even though sepsis in infants and adults may have different clinical characteristics and symptoms, if we intend to detect sepsis by training a machine learning model, their workflows are similar. Therefore, we can unify the processes and develop a common framework for both cases. Many aspects need to be considered to design the framework, such as data collection, data pre-processing, model training, and finally the performance evaluation. Each part in the framework should be explored and researched to make sure it serves our goal and fits into the sepsis detection tasks for both adults and infants. Data collection has already been covered in previous cycle, and it should deal with gathering all kinds of available data including demographic characteristics, vital signs, and laboratory test results. Raw datasets can not be used directly unless they are being wrangled, so pre-processing is necessary before the actual model training and algorithm development. There are several issues that must be handled like outlier detection and elimination, missing value handling, and data re-balance, and we will develop an entire workflow to cover them. Sub-objectives two to four will be achieved in the model training section which involves model selection, algorithm design and reform, model tuning and evaluation. Performance evaluation is also important as other steps because it is where reflection and analysis happens. It corresponds to the evaluation and specifying learning sections of the Action Research methodology. Evaluation metrics will be delicately selected and thorough analysis will be performed in this cycle. A detailed description will be given in the next section.

3.2.2 Four Tasks

Data Collection

The first part of our framework is data collection. Normally, hospitals keep records of heterogeneous data from many dimensions, but we might not need all of them. The most useful variables are demographic characteristics, vital signs and laboratory test results. According to clinicians diagnosing routine and experience, these three types of data are the information they refer to determine antibiotic administration. Apart from the historical data that will be used for training collected from patients, we also need clinicians notes which record their diagnosis and corresponding treatments.

In the neonatal dataset collected from the NICU, it is difficult to set an exact time point after which the patient develops sepsis, but before which is perfectly healthy. Besides, even blood culture is not 100% accurate as we stated in chapter 1, i.e. quite a lot of the blood culture will be negative, hence we chose antibiotic administration as the sign of infection, as this is the level at which the medical staff is concerned about the possibility of sepsis to act, and make this the label for the data. In other words, if a patient accepts antibiotics on a particular day, he will be classified as a sepsis case from that day on. In the NICU, only systemic IV antibiotics is considered, and no other forms of antibiotic administration will be given to patients who are suspicious to neonatal sepsis. Patients that were treated with antibiotics for other reasons except for neonatal sepsis were excluded from ”sepsis” class. Note that the labels have been set for public datasets, so it is not necessary to define them once again.

Automating the process of data collection is another necessary component in this section, since it is impossible to collect the large amount of required patient data from multiple sources manually, not to mention other complex and time-consuming tasks to process this data. There has to be some kind of automated workflow, covering all the data-related work listed as follows:

1. Collecting vital signs in batches instead of patient by patient. Every time a new patient gets admitted, his data should be monitored and collected, and the system should be able to identify the data belonging to a new patient, but not the one who stayed in the same cot before. It could be implemented automatically by an algorithm and supported by the hardware manufacturer, or manually by the nurses records.

2. Collecting demographic data from databases. Unlike vital signs which are continuous time series data, demographic data such as gender and gestational age is static and it will not change, and they should be aligned to the vital signs of the same patient and left for further process.

3. Collecting lab tests result. Lab tests include many different types of biochemical tests, some of which are quite relevant to sepsis diagnosis, e.g. blood culture tests. Clinicians will request a lab test either because it is a routine examination or there is a suspicious symptom, therefore, only several test results might exist in the database during the stay. Similar to demographic data, it must be aligned with the vital signs of the same patient to form a full set of training data.

4. Decompressing data files if necessary. When some of the vital signs are collected by proprietary tools, they are possibly compressed or in a special format which can not be used directly without decryption. In this case, automatic decompression has become significant for generating useable content effectively and continuously.

5. Extracting clinicians notes. Clinicians notes are extremely meaningful because they contain large amount of information, for example, the treatment has been given to the patient, the judgement from clinicians and the one we care about most - the antibiotics usage.

6. Securely store the data. The data collection is a long lasting and accumulating process, and the patient's data itself is very sensitive, so how to securely persist the data is a serious question that needs to be handled properly.

The six points above are only for real-time data collection from patients in hospitals. For public available dataset, it will be much easier because in most cases public datasets have a wide variety of variables and are ready to use. Nevertheless, no matter in which way the data is collected, the output will be identical, which is in a tabular form of rows representing data records and columns indicating features respectively.

Data Preprocessing

Most of unprocessed raw data is unable to be used directly due to many reasons. We list four main reasons specifically for the sepsis task and explain why we need pre-processing and how we perform it in order to improve the outcome. The input of pre-processing is the raw data collected in the previous section, and it produces a ready-to-use dataset which is exactly the material to train models.

Vital Signs Synchronisation If the training data is collected from patients in real-time, vital signs synchronisation is an issue that must be handled properly. This issue is caused by the different sample rates of each monitored variable. Usually, vital signs like heart rate have a higher sample rate up to 200Hz, blood pressure at 100Hz, and respiratory rate at 50Hz [181]. For those manually monitored like body temperature, they will be recorded only a couple of times in one day. Forcibly merging data at different frequencies will result in a large number of blank values in low-frequency features, which potentially reduces the performance of the upcoming model training. To mitigate this issue, a pivot sample rate should be set. In a specific sepsis prediction case, a super high sample rate does no good for the model training because of two reasons: 1) Higher sample rate means larger size of dataset which consumes more space in the computer. 2) Vital signs at two time points extremely close to each other barely change. 3) Higher sample rate brings unnecessary blank values for those of lower sample rate. In shorter words, we do not need a large dataset in which most records are almost identical, it is way less effective and informative given the size of the dataset. To capture sufficient information from the patients data and meanwhile keep it concise and practical, our suggestion for an appropriate pivot sampling rate is between once per second to once every ten minutes. In our experiment, we set the number to once per minute which is adequate to capture the trend of all types of vital signs without expanding the dataset size to a Gigabyte degree. For features like heart rate that have a higher sample rate than the pivot, under-sampling was applied to align it with others so that the tabular dataset could be created.

Outlier Detection Errors are inevitable during data monitoring, transmission and processing, and one of the problems it brings is outliers. Outliers are values that deviate far from the original distribution, and they have side effects on the final performance depending on which model is being trained. Therefore, they should be taken care of at the pre-processing stage, and make sure no outlier data is included or at least reduce its effect to the lowest level possible. There are a range of ways to detect and eliminate outliers, from the simplest z-score or boxplot methods, to more complex density based spatial clustering and isolation forests methods, one or more of which should be implemented in the framework.

Missing Value Handling Another issue caused by human and machine errors is missing values. The term ″missing value″ in our context generally refers to any irregular values in the dataset, such as blanks, ″***″, ″NA″ and so on. Missing values can be categorised into three types, which are Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). MAR means the missing values are not related to themselves but are related to part of those that have been observed, while in MNAR cases, they are related to the anticipated values of themselves or values of other features. MCAR indicates that missing values are completely irrelevant to any data so they do not affect the distribution.

Interpolation is the most widely used method to deal with missing values and there is a variety of different interpolation strategies available, but before applying interpolation, we should observe the ratio of missing values first. If the ratio is low, the most effective way is to directly delete those records with missing values; If the ratio of one feature is beyond the threshold, the entire set of this feature should be abandoned. Note that the deletion methods should be used with caution only in MAR and MCAR, for MNAR datasets, simply removing records with missing values will affect the original distribution and increase bias in the training results. To implement interpolation, there are multiple available options to choose, for example, for features that do not have explicit patterns, we can fill the missings with average values, mode values, or randomly select values in Gaussian distribution; If certain trends can be observed, linear interpolation would be a great choice; Treating missing values as a new category of data is another way when dealing with discrete categorical features; Even logistic regression algorithm can be applied to predict the missing values with other complete feature sets.

Data Rebalance Extreme imbalance is another typical property of clinical data for almost every type of disease, because positive cases are always way less than negative cases. However, imbalanced data hinders the model from making the right prediction, because in order to achieve higher accuracy, models tend to predict all results to the dominant negative cases. Data should be rebalanced before training, by either shrinking the size of the overwhelming class (under-sampling) or increasing the minor class (over-sampling). Under-sampling can not be done simply by randomly sampling from the original dataset, because vital signs are time series and they contain temporal information, unless sliding window technique is applied and datasets are divided into chunks. Under-sampling is easy to implement but its disadvantage is obvious though, which is that only small portion of the dataset is utilised and there could be latent information neglected by not using the entire dataset. Over-sampling is the approach to enrich the minor class with synthetic data. A simple implementation of over-sampling is to duplicate the minor class several times to produce more data, but the dummy repeating emphasise the minor training samples and raises the tendency of overfitting. Another typical and practical algorithm for over-sampling is called Synthetic Minority Oversampling Technique (SMOTE) proposed by Chawla et al. [182] which facilitates the generation of synthetic instances of the minor class by working on the feature space. The problem is that SMOTE was not designed for continuous time series data, so it can not be applied to our patient data directly.

Model Training

With a processed dataset, we can train machine learning models to predict sepsis patients now. There exists a wide range of models that can accomplish the prediction task, the target of model training section is to find out the best models among them with optimum parameters. Training procedure is quite similar to common machine learning projects except for two considerations: the data scheme and the alignment scheme.

Data Schemes To select a proper model for the sepsis prediction task, first we need to determine how each training data instance is formed, which we call data scheme. One way to feed the training data to the models is to divide it into chunks of a certain size by sliding windows. Sliding window is a popular technique used to convert continuous data stream to independent data chunks by filtering the original dataset with a fixed-sized window moving along the direction that the data extends. Since the training data has multiple dimensions, these data chunks are then fed into many of the classic models after a flattening operation which changes the shape from (n, m) to (n*m, 1) where n is the size of sliding window and m is the number of features (dimensions). Data chunks are suitable for classic models that deal with individual inputs separately, since the time info has been filtered out during the conversion. Another data scheme is to consider the data stream of every patient as one instance, in which way time info is kept, so more advanced models such as RNN can capture and utilise the time info to make predictions.

Alignment Schemes Another choice that has to be made before actual training is the selection of an alignment scheme. In left-aligned scheme, the first n hours of the patients data are made available to the models. The goal is to predict whether the patient will develop sepsis at any subsequent point. Data sequences are left aligned to carry out this task, i.e. only records within the first n hours are used for training and testing. This n-hour window is denoted as an observation window and is represented as a shaded area in Fig. 3.4.

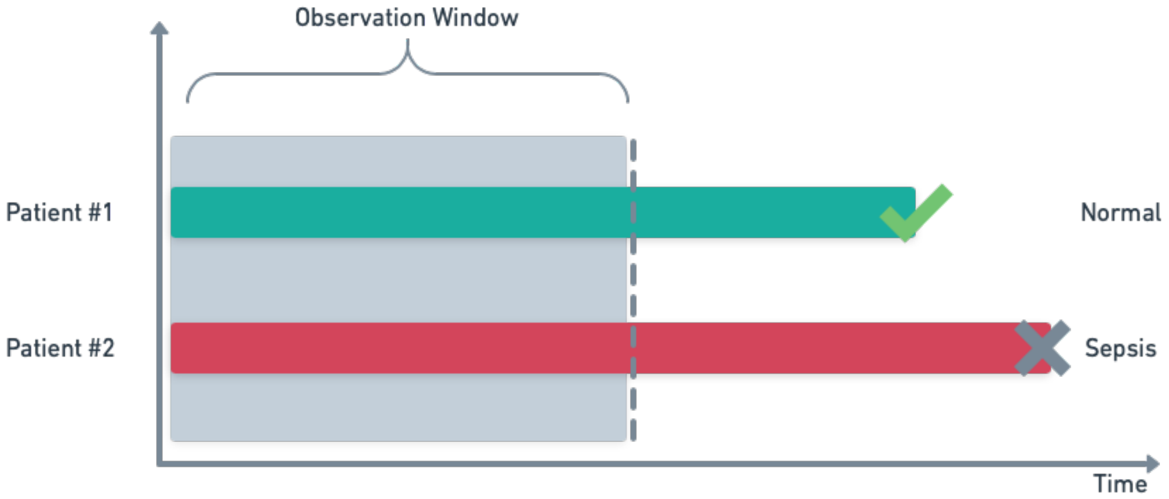


Figure 3.4: Left-aligned scheme.

For the right-aligned scheme, data sequences are aligned to the right which is the point where sepsis onset happens for positive cases, and the end of the sequence for normal cases. Our goal is to predict whether the patient will develop sepsis exactly m hours later, and the m-hour window is called the hold-off window. Only the data before the last m hours will be kept and all that within are omitted, see Fig. 3.5.

Performance Evaluation

Last but not least, evaluation is the final section in our framework. Only in the cycle ”train - evaluate - improve” can we push our models to the optimum. We are going to evaluate our models in the following four aspects: accuracy, training time, interpretability and time ahead of onset.

Accuracy ”Accuracy” is a general term here referring to a set of evaluation metrics that measure how well the models predict in terms of accuracy. AUC, F1 score, accuracy, re-

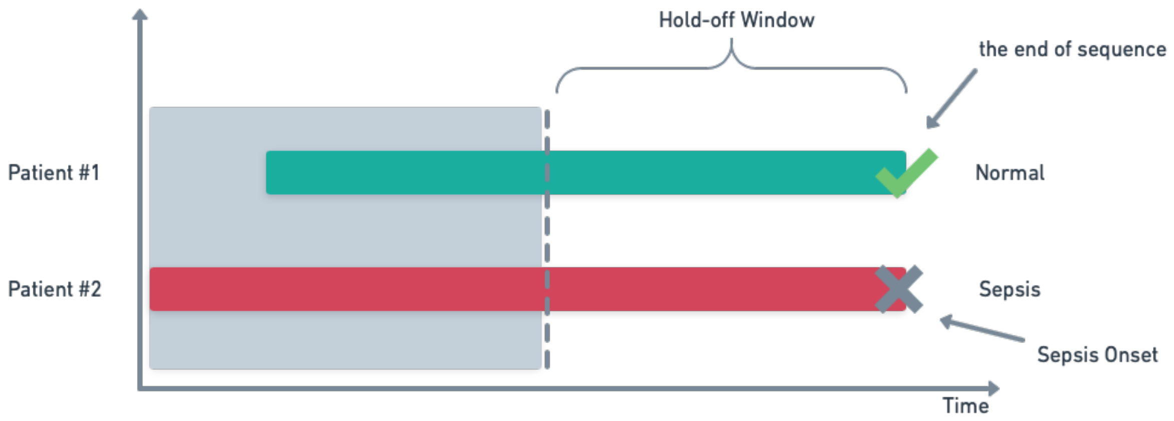


Figure 3.5: Right-aligned scheme.

call and precision are the most commonly used evaluation metrics. Accuracy describes the fraction of patients whose labels were correctly identified. Recall indicates the proportion of patients that actually had sepsis were correctly diagnosed. Precision equals the proportion of patients who were diagnosed as having sepsis actually had sepsis. Finally, AUC calculates the tradeoff between recall and specificity and F1 score is the tradeoff between precision and recall, so AUC and F1 score are two metrics that we prefer to use in terms of accuracy. Furthermore, we also focus on the false positive rate and the false negative rate, because the former increases the risk of antibiotic overdose and the latter misses the best curing windows.

Training Time Training time is closely connected to the model complexity. We do not want a complex model that tends to overfit while the performance does not have a significant increase. Shorter training time is preferred if performance remains at the same or similar level to save training time, especially when real-time learning is introduced in the future.

Interpretability Interpretability is also related to the model complexity, usually a model with more complex structure has lower interpretability. This has been a challenge for a long time since most of machine learning models are black boxes which can not earn clinicians trust even if they have considerably high performance. Visualisation will be a great tool for doctors to understand the machine learning models.

Time ahead of onset How long ahead the model can predict onset before it is happening is one of the key metrics if the scheme is right aligned. We surely want to predict the outcome as soon as possible, but if the size of the hold-off window is too big, it will certainly affect the accuracy of the model. We need to combine the performance evaluation of all four aspects and make the best tradeoff.

3.3 Ethical Considerations

Our research cooperation with Monash Childrens Hospital has been approved by the Monash Health Research Ethics Committee (#RES-17-0000-144Q, #RES-18-0000-24L). All the data used in our project has been de-identified, i.e. all sensitive personal information is omitted. Vital signs were collected within the hospital network and stored in a password protected Google drive under the Monash account.

3.4 Chapter Summary

Herein, the main purpose of the chapter is to describe the research methodology and approach used for the study and to explain how Action Research has been incorporated into our project. Fig. 3.6 summarised the proposed research activities and outputs.

We first explained and discussed what Action Research is and compared it with another popular research method in IS - Design Science, and illustrated its application with respect to IS research and to our project, followed by a description of our three-cycle experiment design. We then proposed a unified sepsis detection framework that works for both adult and infant cases of sepsis. Lastly, we discussed the ethical considerations in our research so that we were certain everything was legitimate.

The next chapter will analyse deeply both the dataset collected from NICU and publicly available, providing an insight into the data we used in our research.

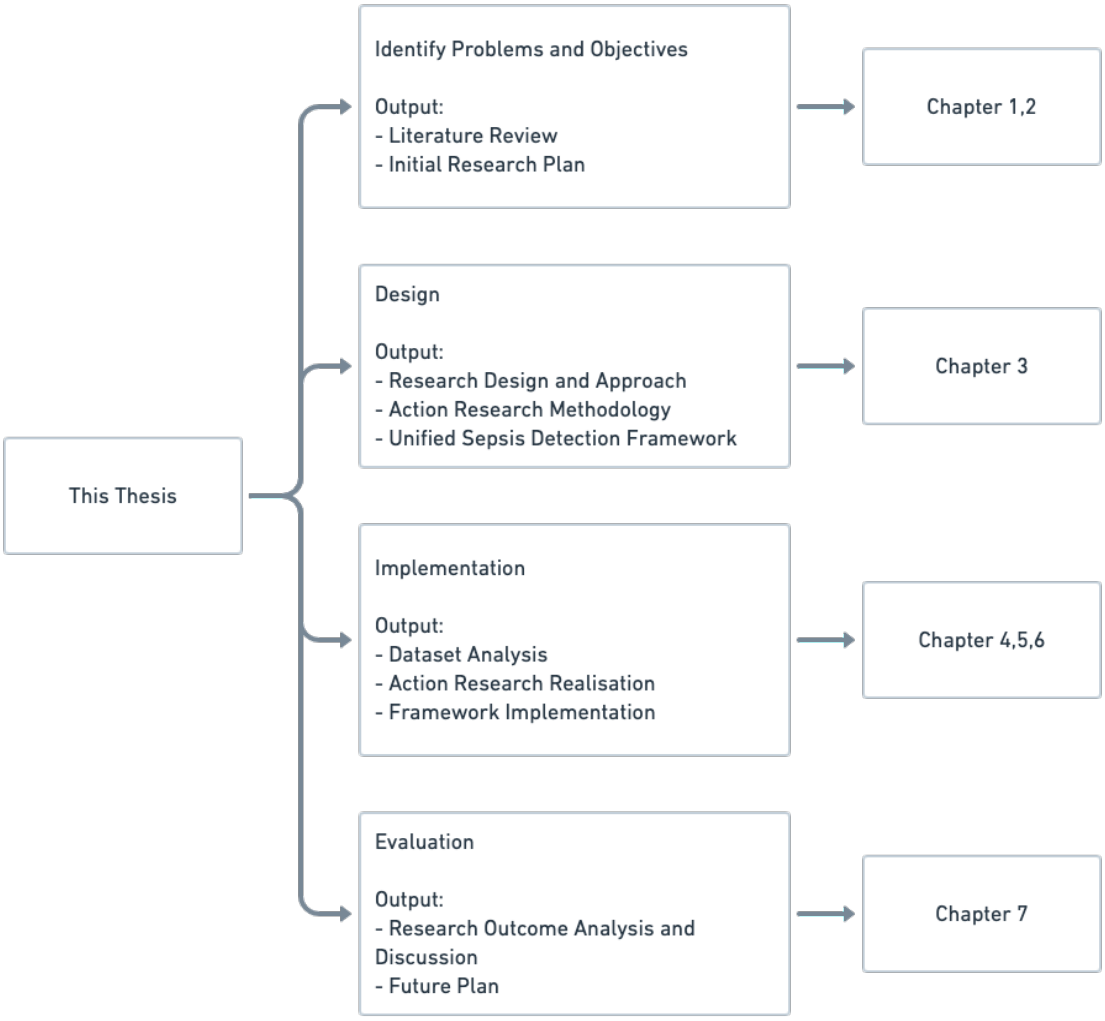

This Thesis
Identify Problems and Objectives
Output:
- Literature Review
- Initial Research Plan
Chapter 1,2
Design
Output:
- Research Design and Approach
- Action Research Methodology
- Unified Sepsis Detection Framework
Chapter 3
Implementation
Output:
- Dataset Analysis
- Action Research Realisation
- Framework Implementation
Chapter 4,5,6
Evaluation
Output:
- Research Outcome Analysis and Discussion
- Future Plan
Chapter 7

Figure 3.6: Summary of proposed research activities and outputs.

Chapter 4

Dataset Analysis

The quantity and quality of data are crucial to the success of a data science project. We examine all the data collected from three different sources in this chapter: data collected from the NICU at Monash Children's Hospital, MIMIC III public data, and the PhysioNet Challenge 2019 dataset. As these datasets consist of heterogeneous variables of different lengths, they must be analysed and processed prior to use, and only a small portion of them were selected based on clinical experience and the requirements of specific models we were using. With the incorporation of multiple datasets, we were able to extend our models and algorithms to both adults and infants, as well as being able to validate the intermediate models and generalise them. Samples extracted from all three datasets are provided in Appendix A.

4.1 Collection from NICU

The first dataset was collected in real-time from the NICU of Monash Children's Hospital and the subjects were all the 147 preterm infants admitted to the NICU between 23rd October 2017 and 26th February 2018. This dataset comprises of demographic data, vital signs, blood culture test results, and clinicians' notes. The gestation ages of the patients range from 166 days to 290 days with an average value of 221 days (Q1-Q3: 193-252), and the birth weights lie between 422 grams and 4,240 grams with an average value of 1,731 grams (Q1-Q3: 986-2549), 48.2% patients are female. All the vital signs used in our model training were from 32 cots in the NICU, and the demographic data, blood culture test outcome and clinician treatment details were extracted from the EMR system in the hospital. Available features are listed in Table 4.1 in detail. We could not capture all the patient data during that time due to a variety of issues such as unexpected laptop shutdowns, network issues, NICU duty nurses intervention and so on, and some features like NBP-S, NBP-D and NBP-M were recorded only once every several hours, presumably done by nurses manually. These features which are far less than normal will be discarded because the rare occurrence makes the dataset sparse.

Descriptive statistics were calculated before and after the preliminary cleaning of dataset,

Table 4.1: Raw features in collected dataset from NICU

Type	Name	Description
Vital signs	HR	Heart rate (beats per minute)
	ART-S	Arterial Blood Pressure (Systolic)
	ART-D	Arterial Blood Pressure (Diastolic)
	ART-M	Arterial Blood Pressure (Mean)
	NBP-S	Non-invasive Blood Pressure (Systolic)
	NBP-D	Non-invasive Blood Pressure (Diastolic)
	NBP-M	Non-invasive Blood Pressure (Mean)
	RESP	Respiration rate (breaths per minute)
	SpO2	Blood Oxygen Saturation
	PLS	Pulse
	25255	Readings from PS25255 sensors
Laboratory values	Blood Culture Test Result	Outcome of blood culture test
Demographics	Date of Birth	DAY/MONTH/YEAR
	Date of Admission	DAY/MONTH/YEAR
	Gender	Female (0) or Male (1)
	Gestation Age	Measured in weeks
	GA Days	Measured in days
	Birthweight	Measured in grams
Clinicians' notes	Antibiotic treatment	If antibiotic was used: yes(1), no(0)

see Table 4.2. The statistics between the raw dataset and the cleaned data are slightly different despite the cleaning being preliminary. Features like Arterial Blood Pressure are unlikely to be a negative number, so we simply remove all negative and zero values from the dataset. Major differences could be found in attributes ART-D, ART-M and ART-S, revealing the fact

Table 4.2: Descriptive statistics of NICU dataset

	Variable	Mean	Max	Min	Median	Std
Raw Data	HR	155.95	297.0	0.0	158.0	18.70
	SpO2	95.26	100.0	0.0	97.0	5.34
	RESP	55.88	154.0	0.0	54.0	20.91
	PLS	156.32	238.0	0.0	158.0	19.44
	25255	1.49	20.0	0.0	1.27	1.10
	ART-D	3.92	314.0	-98.0	0.0	12.13
	ART-M	5.21	314.0	-13.0	0.0	15.69
	ART-S	6.92	314.0	-12.0	0.0	20.53
Preliminarily Cleaned	HR	156.01	297.0	8.0	158.0	18.22
	SpO2	95.32	100.0	5.0	97.0	4.77
	RESP	55.95	154.0	1.0	54.0	20.83
	PLS	156.41	238.0	1.0	158.0	19.04
	25255	1.49	20.0	0.02	1.27	1.10
	ART-D	35.96	314.0	1.0	34.0	13.98
	ART-M	47.74	314.0	1.0	46.0	15.00
	ART-S	63.44	314.0	1.0	62.0	16.68

that there were large numbers of zeros and negative values which are treated as noise in these three. Mean values increased significantly from around 5 to around 50 when noise is removed, and medians surged from zero to 34.0, 46.0 and 62 respectively. Compared to the three ART attributes, others have no obvious changes between the statistics before and after the removal of zeros and negatives.

We define one patient-day as the data of one person within one day, and at the end of the data collection task, totally we had successfully retrieved 1204 patient-days of physiological data. As to the labels for training, antibiotic usage extracted from clinicians' daily notes was used as the reference that determines whether a patient should be classified as having potential sepsis or not. Clinicians record the problems list of the patients and their management plan in plain text according to their daily routing examination.

4.2 PhysioNet Challenge

The topic of the PhysioNet Challenge 2019 is Early Detection of Sepsis from Clinical Data which is exactly the same as our project. Data used in the competition was from three separate hospital systems, two of which are publicly available for training, but another one for testing is not, due to the fact that the competition has ended. The provided datasets have data from all 40336 patients aged from 14 to 100 (mean: 62, Q1-Q3: 51-74) in two separate hospital systems and 56% of them are male. Each data entry is a sequence of multi-variable values from one patient consisting of 40 different features plus one label which indicates whether this patient will have an onset in the next six hours. Each row of record contains a single hour's observation. Since every record has a corresponding label, this dataset is suitable for a right-aligned experiment, in which we predict the sepsis onset right before a fixed period of time, e.g. six hours. No overall collection time scope was specified, but the length of each data entry from individual patient varies from 8 hours minimum to 336 hours maximum, and the average duration is 38.48 hours which requires the prediction model to have the capability to deal with variable-length sequence data, see Fig. 4.1.

As shown in this figure, the distribution of data length is not balanced or uniformed, and most patients have a data with length less than 60 hours. The variation of the length could be caused by many reasons, and the variation itself contains information as well, indicating the how long the patient stayed in hospital. The 40 features provided by PhysioNet dataset can be categorised into three classes: vital signs such as heart rate (HR) and oxygen saturation

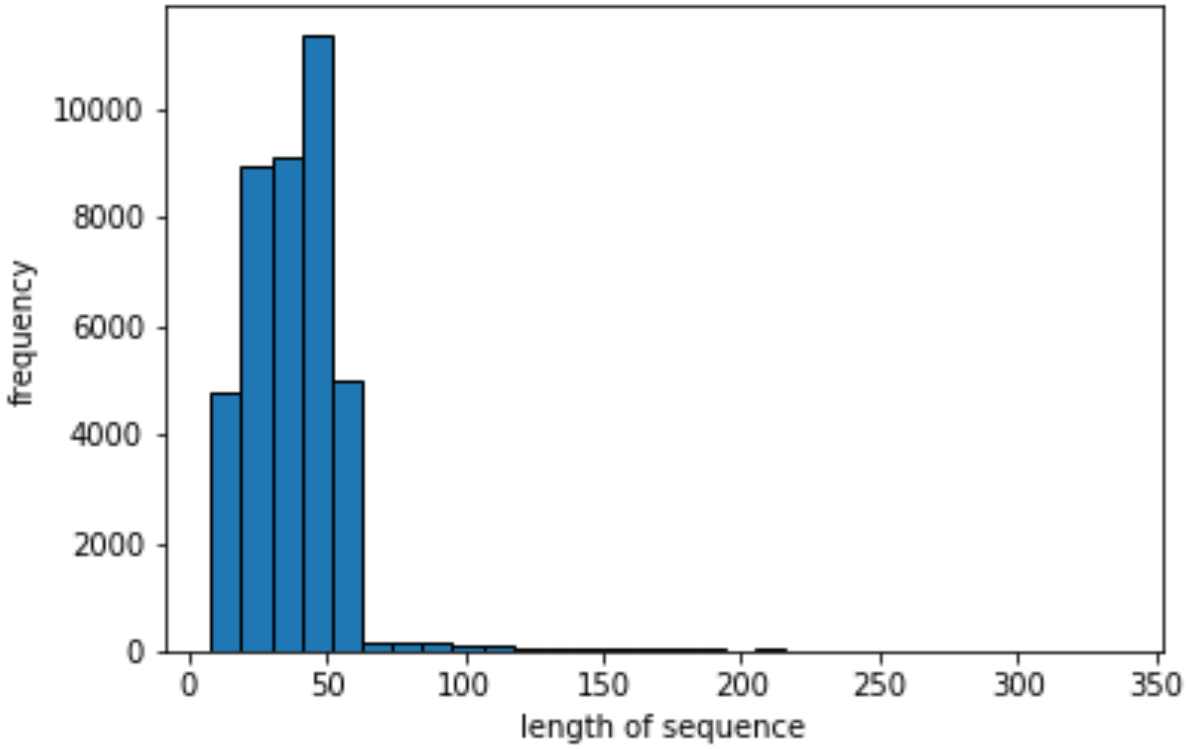


Figure 4.1: The distribution of data lengths for all patients in PhysioNet dataset.

(O2Sat), 8 columns in total; laboratory values such as blood urea nitrogen and platelets, 26 columns in total; and demographics such as age and gender, 6 columns in total, see details in Table 4.3.

Descriptive statistics were also examined for each variable in the PhysioNet Challenge datasets. The preliminary statistics showed that outliers were much less frequent than the NICU dataset, since most variables have a normal distribution, except for ”FiO2” and ”HospAdmTime” which have explicitly negative outliers, see Table 4.4.

According to the official document, the dataset defines the following time points.

$t_{suspicion}$: (1) Clinical suspicion of infection identified as the earlier timestamp of IV antibiotics and blood cultures within a specified duration. (2) If antibiotics were given first, then the cultures must have been obtained within 24 hours. If cultures were obtained first, then antibiotic must have been subsequently ordered within 72 hours.

t_{SOFA}: The occurrence of end organ damage as identified by a two-point deterioration in SOFA score within a 24-hour period.

t_{sepsis}: The onset time of sepsis is the earlier of $t_{suspicion}$ and t_{SOFA} as long as t_{SOFA} occurs no more than 24 hours before or 12 hours after $t_{suspicion}$; otherwise, the patient is not marked

Table 4.3: Raw features in PhysioNet challenge dataset

Type	Name	Description
Vital signs	HR	Heart rate (beats per minute)
	O2Sat	Pulse oximetry (%)
	Temp	Temperature (Deg C)
	SBP	Systolic BP (mm Hg)
	MAP	Mean arterial pressure (mm Hg)
	DBP	Diastolic BP (mm Hg)
	Resp	Respiration rate (breaths per minute)
	EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory values	BaseExcess	Measure of excess bicarbonate (mmol/L)
	HCO3	Bicarbonate (mmol/L)
	FiO2	Fraction of inspired oxygen (%)
	pH	N/A
	PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
	SaO2	Oxygen saturation from arterial blood (%)
	AST	Aspartate transaminase (IU/L)
	BUN	Blood urea nitrogen (mg/dL)
	Alkalinephos	Alkaline phosphatase (IU/L)
	Calcium	(mg/dL)
	Chloride	(mmol/L)
	Creatinine	(mg/dL)
	Bilirubin_direct	Bilirubin direct (mg/dL)
	Glucose	Serum glucose (mg/dL)
	Lactate	Lactic acid (mg/dL)
	Magnesium	(mmol/dL)
	Phosphate	(mg/dL)
	Potassium	(mmol/L)
	Bilirubin_total	Total bilirubin (mg/dL)
	TroponinI	Troponin I (ng/mL)
	Hct	Hematocrit (%)
	Hgb	Hemoglobin (g/dL)
	PTT	partial thromboplastin time (seconds)
	WBC	Leukocyte count (count*$10^3/\mu L$)
	Fibrinogen	(mg/dL)
	Platelets	(count*$10^3/\mu L$)
Demographics	Age	Years (100 for patients 90 or above)
	Gender	Female (0) or Male (1)
	Unit1	Administrative identifier for ICU unit (MICU)
	Unit2	Administrative identifier for ICU unit (SICU)
	HospAdmTime	Hours between hospital admit and ICU admit
	ICULOS	ICU length-of-stay (hours since ICU admit)

as a sepsis patient.

Table 4.4: Descriptive statistics of PhysioNet Challenge dataset

Variable	Mean	Max	Min	Median	Std
HR	84.58	280.0	20.0	83.5	17.33
O2Sat	97.19	100.0	20.0	98.0	2.94
Temp	36.98	50.0	20.9	37.0	0.77
SBP	123.75	300.0	20.0	121.0	23.23
MAP	82.4	300.0	20.0	80.0	16.34
DBP	63.83	300.0	20.0	62.0	13.96
Resp	18.73	100.0	1.0	18.0	5.1
EtCO2	32.96	100.0	10.0	33.0	7.95
BaseExcess	-0.69	100.0	-32.0	0.0	4.29
HCO3	24.08	55.0	0.0	24.0	4.38
FiO2	0.55	4000.0	-50.0	0.5	11.12
pH	7.38	7.93	6.62	7.38	0.07
PaCO2	41.02	100.0	10.0	40.0	9.27
SaO2	92.65	100.0	23.0	97.0	10.89
AST	260.22	9961.0	3.0	41.0	855.75
BUN	23.92	268.0	1.0	17.0	19.99
Alkalinephos	102.48	3833.0	7.0	74.0	120.12
Calcium	7.56	27.9	1.0	8.3	2.43
Chloride	105.83	145.0	26.0	106.0	5.88
Creatinine	1.51	46.6	0.1	0.94	1.81
Bilirubin_direct	1.84	37.5	0.01	0.44	3.69
Glucose	136.93	988.0	10.0	127.0	51.31
Lactate	2.65	31.0	0.2	1.8	2.53
Magnesium	2.05	9.8	0.2	2.0	0.4
Phosphate	3.54	18.8	0.2	3.3	1.42
Potassium	4.14	27.5	1.0	4.1	0.64
Bilirubin_total	2.11	49.6	0.1	0.9	4.31
TroponinI	8.29	440.0	0.01	0.3	24.81
Hct	30.79	71.7	5.5	30.3	5.49
Hgb	10.43	32.0	2.2	10.3	1.97
PTT	41.23	250.0	12.5	32.4	26.22
WBC	11.45	440.0	0.1	10.3	7.73
Fibrinogen	287.39	1760.0	34.0	250.0	153.0
Platelets	196.0	2322.0	1.0	181.0	103.62
Age	62.01	100.0	14.0	64.0	16.39
Gender	0.56	1.0	0.0	1.0	0.5
Unit1	0.5	1.0	0.0	0.0	0.5
Unit2	0.5	1.0	0.0	1.0	0.5
HospAdmTime	-56.09	23.99	-5366.86	-6.03	162.14
ICULOS	27.0	336.0	1.0	21.0	29.01

4.3 MIMIC III

MIMIC III (Medical Information Mart for Intensive Care III) [31] is a large and free medical database built by MIT Laboratory for Computational Physiology and collaborating research groups. It comprises de-identified clinical data from 38,657 distinct adult patients aged 16 years or above and associated 49,785 hospital admission, plus 7,863 neonates and their 9,191 admission. For patients older than 89 years, the age information was removed to protect the patient confidentiality, so we can not know the exact ages. Apart from those patients older than 89 years, the median age of the rest patients is 64, the mean value is 62, Q1 and Q3 are 51 and 76 respectively. Out of the entire cohort, male patients make up about 56%. These patients have accepted critical care in Beth Israel Deaconess Medical Centre between 2001 and 2012 [183]. Unlike the PhysioNet dataset, the diagnosis of a patient exists only once per admission, right at the time of discharge, which means the label indicating sepsis is admission-wise not record-wise. Left-aligned scheme is hence the right way to do the prediction.

MIMIC III could be seen as a complex database of multiple tables, each of which contains different information relating to patients. Not every single attribute in the tables is required by our research, so selection has to be made by joint query prior to any pre-processing steps. We select a subset of attributes from the provided tables based on clinicians' recommendation, trying to extract the same feature set as the PhysioNet datasets have. A list of 26 tables available in MIMIC III is presented below in Table 4.5, and another list of chosen attributes is depicted in Table 4.6. As a result, the selection of features is quite similar to the PhysioNet dataset, so we are able to compare the results retrieved from the two datasets and verify the generalisation of the proposed models. Only two attributes - ”AST” and ”BUN” were not included in the MIMIC III database, and since all the data in the database was collected from the ICU, there is no need to calculate the hours between hospital admission and ICU admission. Besides, ICU length-of-stay and age could be calculated from the date of birth and the subtraction of admission time and discharge time.The sepsis labels in MIMIC III dataset are retrieved from hospital database based on clinicians' diagnosis in ICD-9 code, however the criteria remains unexposed.

Table 4.5: Available tables provided MIMIC III database

Type	Name	Description
Track patients	ADMISSIONS	Every unique hospitalisation for each patient
	CALLOUT	Information regarding when a patient was cleared for ICU discharge
	ICUSTAYS	Every unique ICU stay in the database
	PATIENTS	Every unique patient in the database
	SERVICES	The clinical service under which a patient is registered
	TRANSFERS	Patient movement from bed to bed within the hospital, including ICU admission and discharge
Critical care unit	CAREGIVERS	Every caregiver who has recorded data in the database
	CHARTEVENTS	All charted observations for patients
	DATETIMEEVENTS	All recorded observations which are dates
	INPUTEVENTS_CV	Intake for patients monitored using the Philips CareVue system while in the ICU
	INPUTEVENTS_MV	Intake for patients monitored using the iMDSoft Metavision system while in the ICU
	NOTEEVENTS	Deidentified notes, including nursing and physician notes, ECG reports, imaging reports, and discharge summaries.
	OUTPUTEVENTS	Output information for patients while in the ICU
	PROCEDUREEVENTS_MV	Patient procedures for the subset of patients who were monitored in the ICU
Hospital records	CPTEVENTS	Procedures recorded as Current Procedural Terminology (CPT) codes
	DIAGNOSES_ICD	Hospital assigned diagnoses, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system
	DRGCODES	Diagnosis Related Groups (DRG), which are used by the hospital for billing purposes.
	LABEVENTS	Laboratory measurements for patients both within the hospital and in out patient clinics
	MICROBIOLOGYEVENTS	Microbiology measurements and sensitivities from the hospital database
	PRESCRIPTIONS	Medications ordered, and not necessarily administered, for a given patient
	PROCEDURES_ICD	Patient procedures, coded using ICD system
Dictionaries	D_CPT	High-level dictionary of CPT codes
	D_ICD_DIAGNOSES	Dictionary of ICD codes relating to diagnoses
	D_ICD_PROCEDURES	Dictionary of ICD codes relating to procedures
	D_ITEMS	Dictionary of ITEMIDs appearing in the MIMIC database, except those that relate to laboratory tests
	D_LABITEMS	Dictionary of ITEMIDs in the laboratory database that relate to laboratory tests

Table 4.6: Raw features in MIMIC III dataset

Type	Name	Description
Vital signs	HR	Heart rate (beats per minute)
	O2Sat	Pulse oximetry (%)
	Temp	Temperature (Deg C)
	SBP	Systolic BP (mm Hg)
	MAP	Mean arterial pressure (mm Hg)
	DBP	Diastolic BP (mm Hg)
	Resp	Respiration rate (breaths per minute)
	EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory values	BaseExcess	Measure of excess bicarbonate (mmol/L)
	HCO3	Bicarbonate (mmol/L)
	FiO2	Fraction of inspired oxygen (%)
	pH	N/A
	PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
	SaO2	Oxygen saturation from arterial blood (%)
	Alkalinephos	Alkaline phosphatase (IU/L)
	Calcium	(mg/dL)
	Chloride	(mmol/L)
	Creatinine	(mg/dL)
	Bilirubin_direct	Bilirubin direct (mg/dL)
	Glucose	Serum glucose (mg/dL)
	Lactate	Lactic acid (mg/dL)
	Magnesium	(mmol/dL)
	Phosphate	(mg/dL)
	Potassium	(mmol/L)
	Bilirubin_total	Total bilirubin (mg/dL)
	TroponinI	Troponin I (ng/mL)
	Hct	Hematocrit (%)
	Hgb	Hemoglobin (g/dL)
	PTT	partial thromboplastin time (seconds)
	WBC	Leukocyte count (count*$10^3/\mu L$)
	Fibrinogen	(mg/dL)
	Platelets	(count*$10^3/\mu L$)
Demographics	DOB	Date of Birth
	Gender	Female (0) or Male (1)
	Weight	Body weight at admission
	Admittime	Admission time
	Dischtime	Discharge time

4.4 Special Characteristics

4.4.1 Missing values

For many reasons such as human intervention or machine failure, and also asynchronized sampling rates, there could be missing data in the datasets from all three sources.

For NICU data, there are 1,419,731 lines with one or more missing values out of total 1,465,781 lines at the ratio of 96.86%, and the total missing ratio in terms of individual values is 4,350,597 out of 11,726,248 (37.10%). Presence ratio for each attributes is shown in Fig. 4.2. The figure showed the same conclusion as we drew from the preliminary data cleaning - the ART series has a large proportion of missing values, actually less than 10% as seen in the figure. With respect to blood pressure, there are another series of variables: NBP S, NBP M and NBP D which have a slightly larger number of occurrences than the ART series. According to our calculations, 90.22% of all records do not contain ART readings, 99.66% do not contain NBP readings, and at least one of the two blood pressure series is absent in 99.97% of total records.

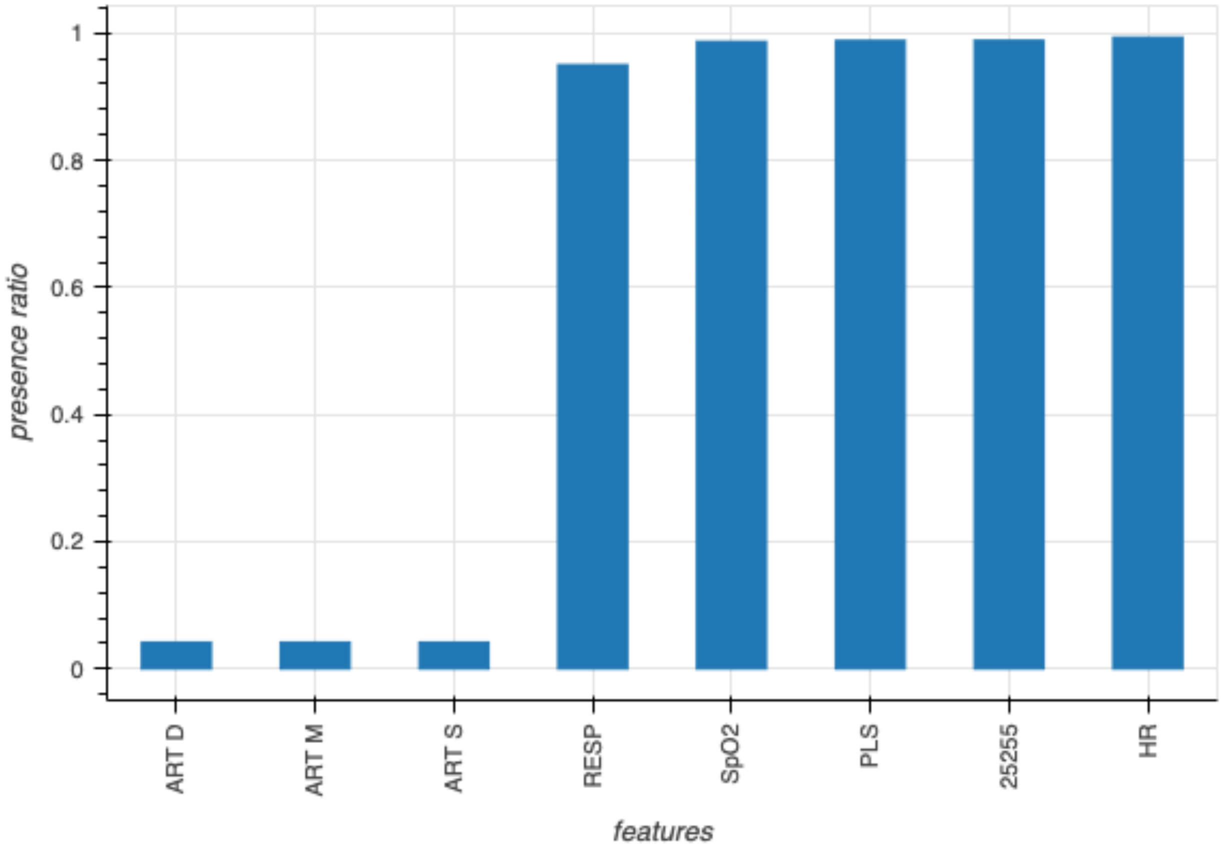


Figure 4.2: Presence ratio of each feature in the NICU dataset

For PhysioNet dataset, we found that total presence ratio for all features is as high as 31.63%, which means more than half of data is not available. Detailed presence ratio for each feature could be seen in Fig. 4.3, and it's easy to see that for 26 features, they have actual values less than 10%. There are only nine features with missing values less than 20%, and only three features have full data. Such characteristics also exist in MIMIC III dataset, see Fig. 4.4. Apart from the labels we manually attached, only four features have 100% presence rate, while almost 20 features have less than 10% data available. However, if we examine them closely, we can find that for both the PhysioNet and MIMIC III datasets, most features with large numbers of missing values are lab test results which are not typically taken for every patient every minute. By excluding these variables from the analysis, additional noise will be introduced into the dataset. What we did was to convert these sparse variables into categorical data, and we will elaborate on this in detail in Chapters 6.

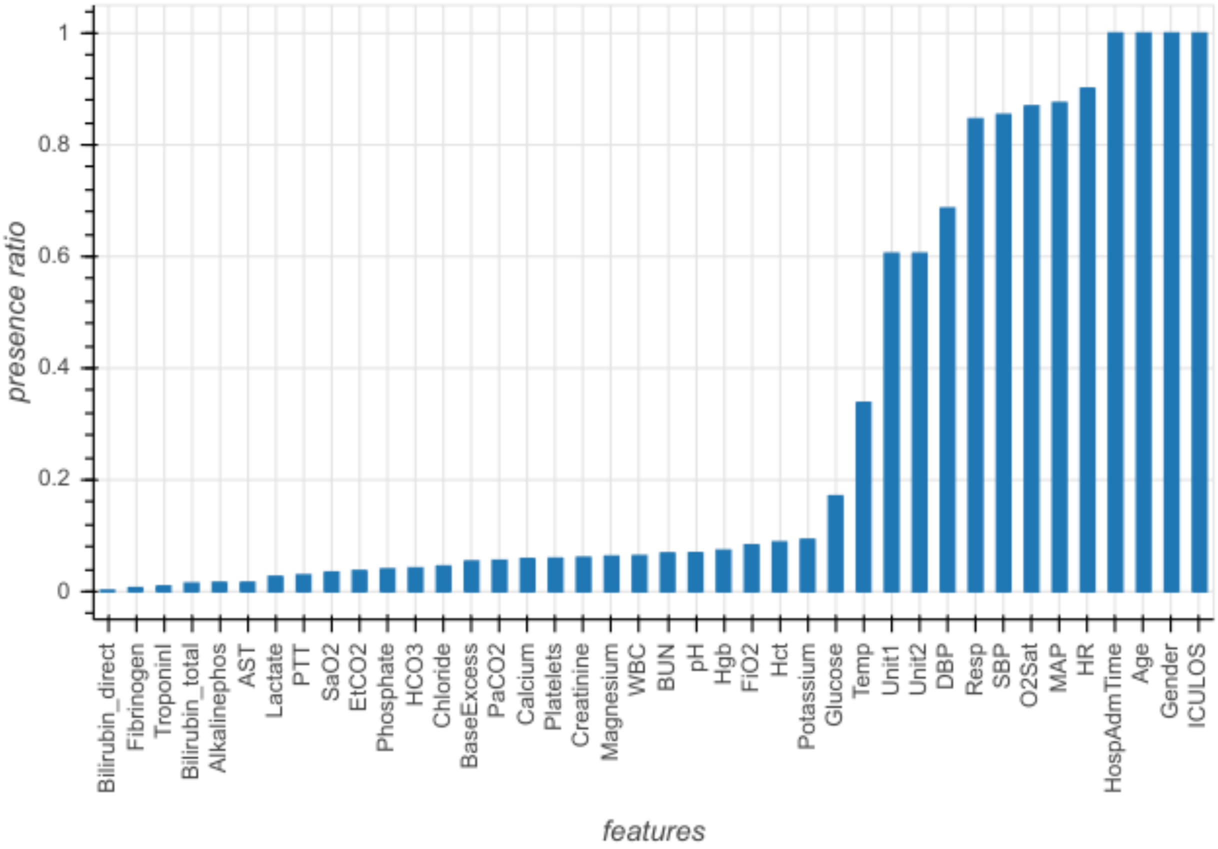


Figure 4.3: Presence ratio of each feature in the PhysioNet dataset.

4.4.2 Imbalance

Another major problem in our datasets was data imbalance, which is quite prevalent in most medical datasets. Taking the dataset from Monash Children's Hospital as an example,

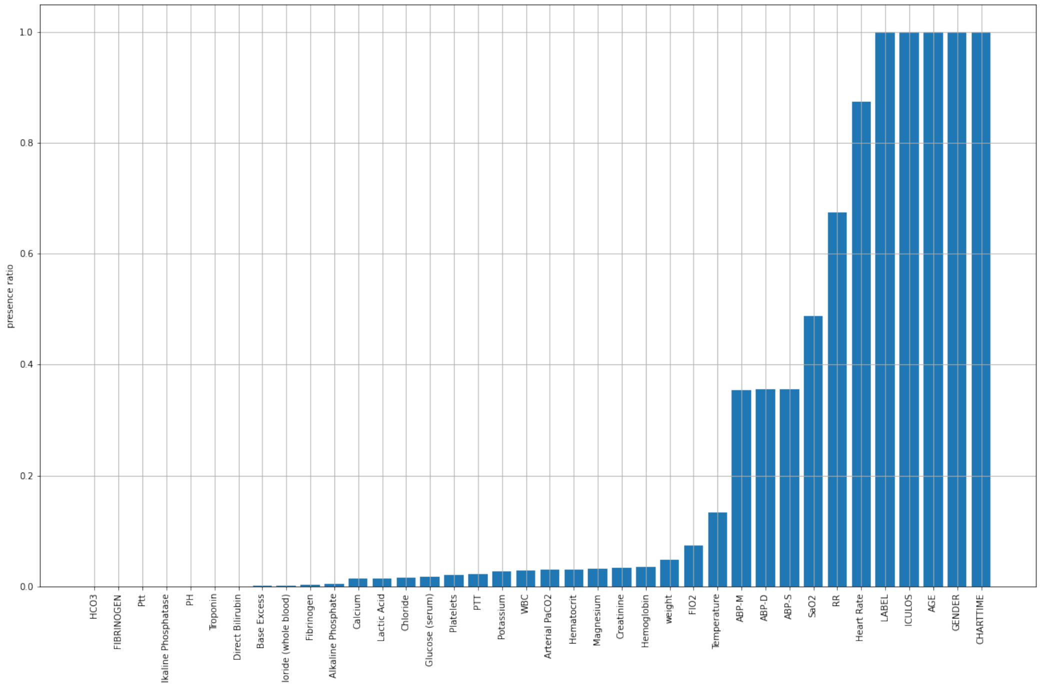


Figure 4.4: Presence ratio of each feature in the MIMIC III dataset.

the total number of sepsis patients is 49 while the number of normal cases is 98. If finer granularity is considered, total number of sepsis records is 163,074, but number of normal records is 1,008,969 which is 6.19 times as sepsis ones, see Fig. 4.5. Considering that the data had been preliminarily filtered prior to being analysed, the ratio would be even more polarised if all the raw data were considered. In addition, it should be clarified that, as we mentioned earlier, the sepsis label is based on the clinicians' treatment with antibiotics. Clinicians will consider not only the results of blood tests, but also a number of other factors, such as the clinical manifestations of the patient that are based on their professional experience.

Situation of data from PhysioNet was even worse. From the point of view of patients, out of 40336, only 2932 had sepsis which accounts for 7.27% of the total cohort. If we analyse the data row-wise, the imbalance is even worse. The entire dataset contains 1552210 rows of records, of which only 27916 rows were marked as sepsis, taking a percentage of 1.8%, see Fig. 4.6. The ratio of normal against sepsis patients is 12.76:1 while that of sepsis records to normal ones even reaches 54.6:1.

In 56884 hospital admission of MIMIC III dataset, only 4099 end up with the diagnosis of sepsis, accounting for 7.21%, while the proportion for non-sepsis cases is 92.79%, and the ratio between is 12.87:1 which is quite close to that of PhysioNet dataset. Since the label is

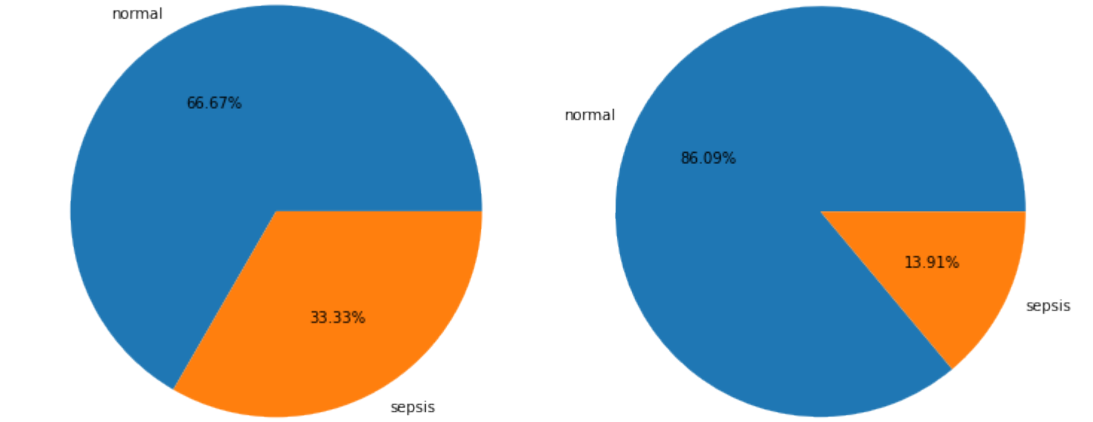


Figure 4.5: Ratio of sepsis patients to normal ones (left), and ratio of sepsis records (rows) to normal ones (right) in NICU dataset.

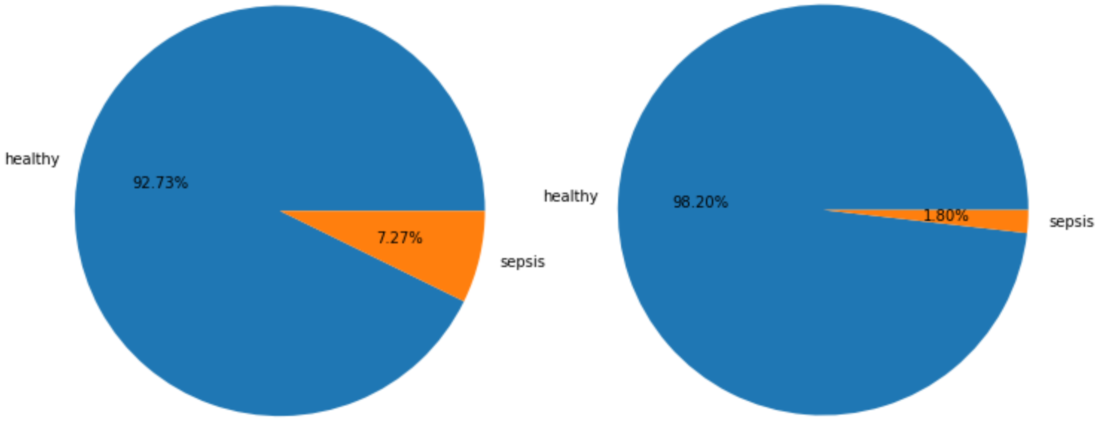


Figure 4.6: Ratio of sepsis patients to normal ones (left), and ratio of sepsis records (rows) to normal ones (right) in PhysioNet Challenge dataset.

assigned for each hospital admission, record-wise analysis is not applicable in MIMIC III datasets, see Fig. 4.7.

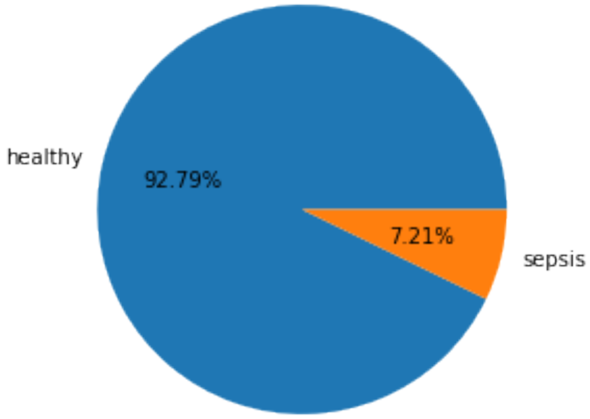


Figure 4.7: Ratio of sepsis admission to normal ones in MIMIC III dataset.

4.5 Chapter Summary

In this chapter, we examined datasets from a variety of sources, summarised their basic statistics, and determined some common characteristics, such as missing data and extreme imbalance. The way we deal with these issues will directly affect the model's subsequent performance, so getting an understanding of the available dataset is very helpful, particularly during the pre-processing process. Additionally, we used three datasets with heterogeneous structures to validate the generalisation of our unified sepsis detection framework.

By examining the datasets, The second research question, how many physiological parameters are available, has been answered. In the following chapters, we are going to discuss the implementation of proposed framework on both neonatal sepsis and adult sepsis prediction.

Chapter 5

Framework Implementation - Neonatal Sepsis Prediction

We presented a unified sepsis detection framework based on Action Research theory in Chapter 3. In this and the next chapters, we will discuss the specific implementation of the designed three-cycle process. Action research, as stated previously, is a cyclical process that repeats several times, identifying new problems and resolving them after reflection. Our research plan consists of three research cycles, aimed at different stages throughout the entire project. In cycle one, data is collected from various sources, while in cycle two, it is the task of identifying useful and informative variables from the collected information, making them accessible for the various models that will be learned in cycle three.The action research specification was followed to implement the last three phases of the process - action, evaluation, and reflection - in each cycle.

In this chapter, we focus on the prediction specifically of pre-term neonates. The dataset we used was collected from NICU of Monash Children's Hospital.

5.1 Experiment Environment

Our experiments were finished in a MacBook Pro with 32GB memory when the task is non-deep-learning related, while Google Colab was used when GPU accelerated computation was utilised in deep learning tasks. Python is the programming language to facilitate the building and training of all the models involved in our project, along with several useful packages designed for machine learning tasks like scikit-learn [184], and PyTorch [185] to handle classic machine learning models and deep learning models. Furthermore, We used PyTorch Lighting [186] as the framework to boost the development process, and wandb [187] as the experiments management tool to store and manage all the parameters, figures, metrics and other importance statistics of each experiemnt setups.

5.2 Cycle One

We began our research by collecting the necessary data from NICU of Monash Children's Hospital. Note that the NICU only provides data for newborn infants, so all the models were designed only for newborn infants and are not universally applicable for patients in the Emergency Department. However, the workflow was still completely under the unified sepsis detection framework we proposed.

The collaboration with Monash Children's Hospital makes it possible for us to collect real-time first-hand data from infants under intensive care. As no digital physiological data of infants was stored in the NICU, we have to collect real-time vital signs from beside monitors, extract demographic characteristics and lab test results from separate systems, and locate antibiotic usage and diagnosis among clinicians' notes. Lack of automatic design makes the workflow labour-intensive and time-consuming.

5.2.1 Vital Signs

The bedside monitoring machines attached to each cot are manufactured by Dräger Medical (Draëger Australia Pty. Ltd, Melbourne, AUS), and they broadcast all the parameters it is monitoring to the local area network within the same VLAN which makes it possible to capture some basic vital signs of newborns under intensive care. We use eDataGrabber (developed by Dräger), which is the client-side software working on the bedside monitors. It is able to capture real-time signals including multiple vital signs, alarm events and it also generates waveforms based on data it has retrieved from the network. In fact, different vital signs are not generated uniformly and every single one has its own sampling rate, for example, displayed ECG waveforms are sent every 200 ms and alarm events go on every second. There are two areas in the NICU of Monash Childrens Hospital with 16 NICU cots and 16 Special Care cots, and we set up two laptops as data collectors running eDataGrabber, each of which covered one area of 32 cots and kept them working 24/7 in order to get as much data as possible. A routine workflow (Fig. 5.1) is described as the following:

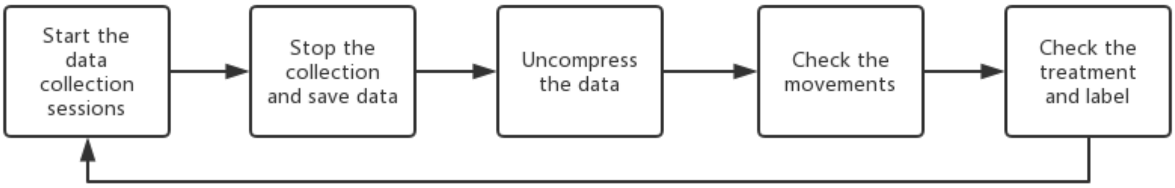

Start the data collection sessions
Stop the collection and save data
Uncompress the data
Check the movements
Check the treatment and label

Figure 5.1: Data collection workflow in the NICU.

1. Start eDataGrabber to capture the vital signs of patients of every cot in the area.

Due to the nature of this software, every single process can only connect to one cot, so 32 separated sessions are created.

2. Restart all the sessions every day and save the previous records in the hard drive.

   We name the data files after the patient's ID so that we can check the corresponding treatment records and clinicians comments in other systems. This data recording task restarts every day to prevent data loss caused by unexpected restarts or crashes of the collector laptops and cutting data into pieces for one day makes the following labelling work easier. Captured compressed data (raw data) for the previous day will be categorised based on date and cot number.

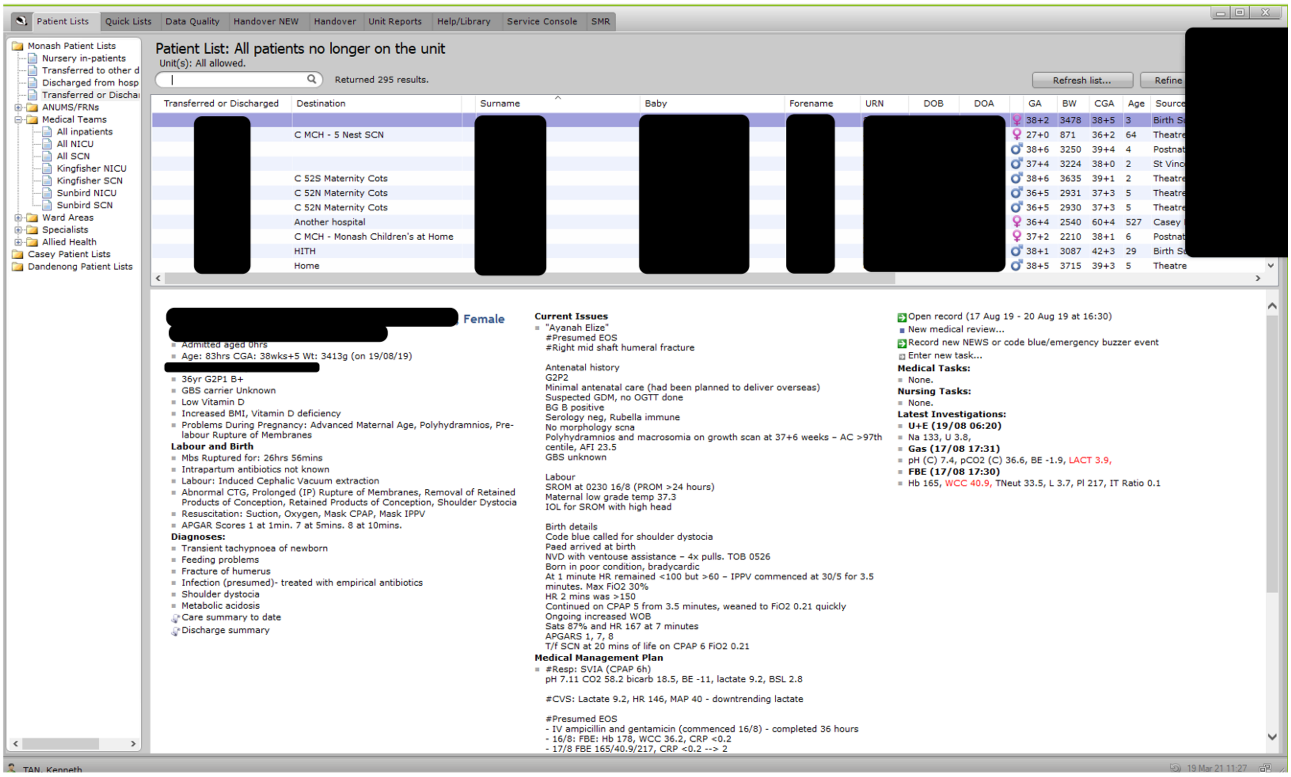
3. Uncompress the raw data, extract the vital signs trends and save them as csv files.

4. Check the patient movement status and the treatment records.

   The eDataGrabber software is associated with fixed cots (fixed IP address), so it does not recognise the patient in that cot or whether the patient in that cot has changed. The fact is that nurses often move babies around for staffing and patient acuity reasons. We have to check whether one baby was moved on the previous day to make sure the data we collected belongs to him. If he/she was moved, then we separate the data to several parts based on whom they belong to. Since the sensors are removed from the surface of the babies once they are moving to another cot, we infer the movement if there are big gap longer than ten minutes in time stamps, although it may not be 100% accurate. Some other actions such as feeding or treatment will cause transient interruptions in the data stream as well, but there is no precise record of the movement information and inference from gaps in time stamps is considered the best way at present.

5.2.2 Other Systems in the Hospital

Apart from vital signs, we also collect other auxiliary data to facilitate the sepsis detection task from other systems in the hospital. BadgeNet is the main source of EMR information from which we extract demographic characteristics and laboratory test results, see Fig. 5.2. Demographic information of each patient is displayed in the bottom window and also in the table upper side as long as we specify the person we are looking into by typing the name or ID in the search box at the top. A more detailed information page will be displayed when double clicking the entry in the table, showing clinicians' notes, care summary and medical management plan, shown in Fig. 5.3, where we collect the information about the diagnosis made by clinicians and set them as labels indicating the outcome of the patient - ”sepsis” or ”normal” respectively. The laboratory test results could be accessed via BadgeNet system as well under the charts tab. Here we can get biochemistry results like C-reactive protein Fig. 5.4 and haematology results like full blood count Fig. 5.5.



Patient List: All patients no longer on the unit
Unit(s): All allowed.

Returned 295 results.

[redacted]

Transferred or Discharged	Destination		Surname	Baby	Forename	URN	DOB	DOA		GA	BW	CGA	Age	Source
[redacted]			[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♀	38+2	3478	38+5	3	Birth S
[redacted]	C MCH - 5 Nest SCN		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♀	27+0	871	36+2	64	Theatre
[redacted]			[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	38+6	3250	39+4	4	Postnat
[redacted]			[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	37+4	3224	38+0	2	St Vinc
[redacted]	C 52S Maternity Cots		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	38+6	3635	39+1	2	Theatre
[redacted]	C 52N Maternity Cots		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	36+5	2931	37+3	5	Theatre
[redacted]	C 52N Maternity Cots		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	36+5	2930	37+3	5	Theatre
[redacted]	Another hospital		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♀	36+4	2540	60+4	527	Casey
[redacted]	C MCH - Monash Children's at Home		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♀	37+2	2210	38+1	6	Postnat
[redacted]	HITH		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	38+1	3087	42+3	29	Birth S
[redacted]	Home		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	♂	38+5	3715	39+3	5	Theatre

[redacted] Female

- Admitted aged 0hrs
- Age: 83hrs CGA: 38wks+5 Wt: 3413g (on 19/08/19)

[redacted]

- 36yr G2P1 B+
- GBS carrier Unknown
- Low Vitamin D
- Increased BMI, Vitamin D deficiency
- Problems During Pregnancy: Advanced Maternal Age, Polyhydramnios, Pre-labour Rupture of Membranes

Labour and Birth

- Mbs Ruptured for: 26hrs 56mins
- Intrapartum antibiotics not known
- Labour: Induced Cephalic Vacuum extraction
- Abnormal CTG, Prolonged (IP) Rupture of Membranes, Removal of Retained Products of Conception, Retained Products of Conception, Shoulder Dystocia
- Resuscitation: Suction, Oxygen, Mask CPAP, Mask IPPV
- APGAR Scores 1 at 1min. 7 at 5mins. 8 at 10mins.

Diagnoses:

- Transient tachypnoea of newborn
- Feeding problems
- Fracture of humerus
- Infection (presumed)- treated with empirical antibiotics
- Shoulder dystocia
- Metabolic acidosis

Care summary to date
Discharge summary

Current Issues

- "Ayanah Elize"
  #Presumed EOS
  #Right mid shaft humeral fracture

  Antenatal history
  G2P2
  Minimal antenatal care (had been planned to deliver overseas)
  Suspected GDM, no OGTT done
  BG B positive
  Serology neg, Rubella immune
  No morphology scna
  Polyhydramnios and macrosomia on growth scan at 37+6 weeks – AC >97th centile, AFI 23.5
  GBS unknown

  Labour
  SROM at 0230 16/8 (PROM >24 hours)
  Maternal low grade temp 37.3
  IOL for SROM with high head

  Birth details
  Code blue called for shoulder dystocia
  Paed arrived at birth
  NVD with ventouse assistance – 4x pulls. TOB 0526
  Born in poor condition, bradycardic
  At 1 minute HR remained <100 but >60 – IPPV commenced at 30/5 for 3.5 minutes. Max FiO2 30%
  HR 2 mins was >150
  Continued on CPAP 5 from 3.5 minutes, weaned to FiO2 0.21 quickly
  Ongoing increased WOB
  Sats 87% and HR 167 at 7 minutes
  APGARS 1, 7, 8
  T/f SCN at 20 mins of life on CPAP 6 FiO2 0.21

Medical Management Plan

- #Resp: SVIA (CPAP 6h)
  pH 7.11 CO2 58.2 bicarb 18.5, BE -11, lactate 9.2, BSL 2.8

  #CVS: Lactate 9.2, HR 146, MAP 40 - downtrending lactate

  #Presumed EOS
  - IV ampicillin and gentamicin (commenced 16/8) - completed 36 hours
  - 16/8: FBE: Hb 178, WCC 36.2, CRP <0.2
  - 17/8 FBE 165/40.9/217, CRP <0.2 --> 2

Open record (17 Aug 19 - 20 Aug 19 at 16:30)
New medical review...
Record new NEWS or code blue/emergency buzzer event
Enter new task...

Medical Tasks:

- None.

Nursing Tasks:

- None.

Latest Investigations:

- **U+E (19/08 06:20)**
- Na 133, U 3.8,
- **Gas (17/08 17:31)**
- pH (C) 7.4, pCO2 (C) 36.6, BE -1.9, LACT 3.9,
- **FBE (17/08 17:30)**
- Hb 165, WCC 40.9, TNeut 33.5, L 3.7, Pl 217, IT Ratio 0.1

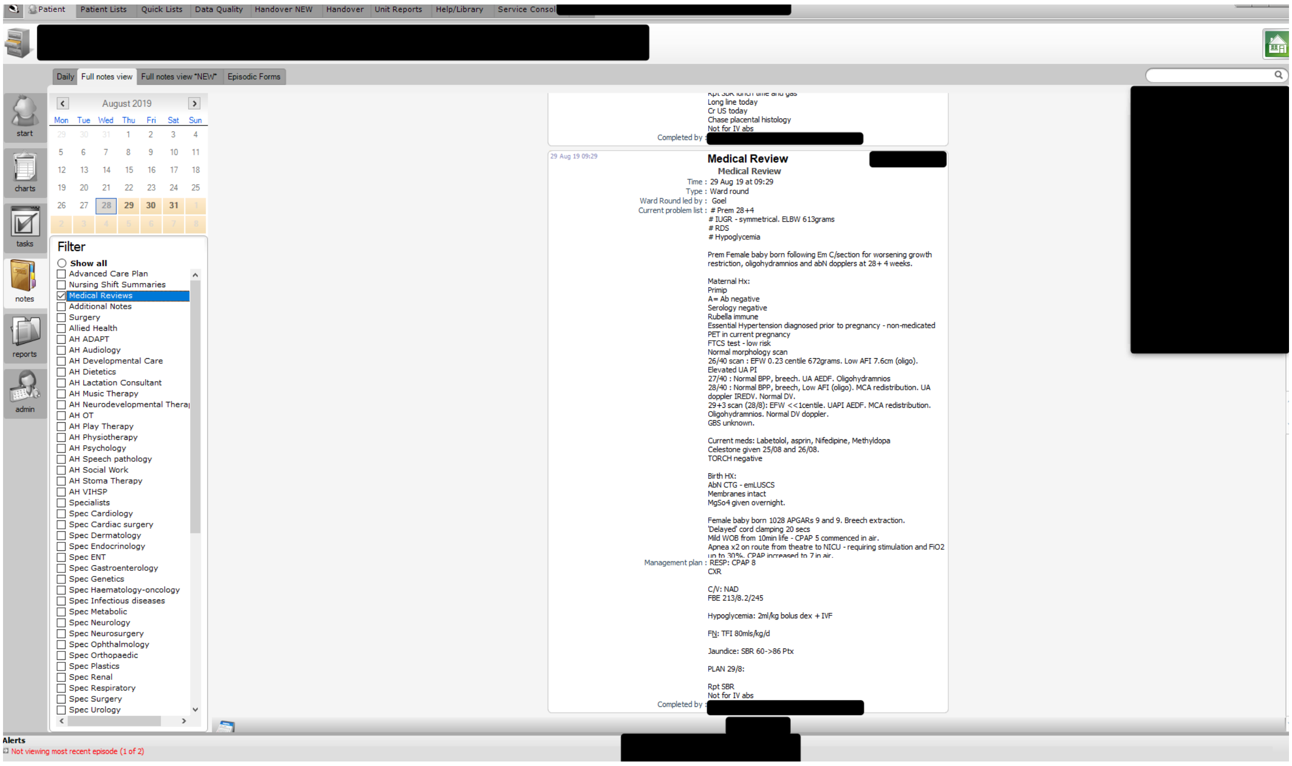
Figure 5.2: The interface of BadgeNet EMR system in Monash Children’s Hospital.

5.3 Cycle Two

The major objective of the second cycle is to investigate the collected datasets and select the best variables for our task. To achieve this, we tried to comprehend information from multiple viewpoints and methods including incorporating clinicians’ domain knowledge, statistical relation analysis and feature importance retrieved by pre-trained models.

5.3.1 Domain Knowledge

Limited by the functionality of bedside monitors, the number of physiological variables that could be recorded continuously was restrained to only eleven, out of which six blood-pressure-related metrics were manually recorded at a much lower frequency than that of machine records. Hence, clinicians do not have many choices in terms of real-time vital signs



[redacted]

[redacted]

Filter

○ Show all
□ Advanced Care Plan
□ Nursing Shift Summaries
☑ Medical Reviews
□ Additional Notes
□ Surgery
□ Allied Health
□ AH ADAPT
□ AH Audiology
□ AH Developmental Care
□ AH Dietetics
□ AH Lactation Consultant
□ AH Music Therapy
□ AH Neurodevelopmental Thera
□ AH OT
□ AH Play Therapy
□ AH Physiotherapy
□ AH Psychology
□ AH Speech pathology
□ AH Social Work
□ AH Stoma Therapy
□ AH VIHSP
□ Specialists
□ Spec Cardiology
□ Spec Cardiac surgery
□ Spec Dermatology
□ Spec Endocrinology
□ Spec ENT
□ Spec Gastroenterology
□ Spec Genetics
□ Spec Haematology-oncology
□ Spec Infectious diseases
□ Spec Metabolic
□ Spec Neurology
□ Spec Neurosurgery
□ Spec Ophthalmology
□ Spec Orthopaedic
□ Spec Plastics
□ Spec Renal
□ Spec Respiratory
□ Spec Surgery
□ Spec Urology

Rpt SBR lunch time and gas
Long line today
Cr US today
Chase placental histology
Not for IV abs
Completed by : [redacted]

29 Aug 19 09:29

Medical Review [redacted]

Medical Review

Time : 29 Aug 19 at 09:29
Type : Ward round
Ward Round led by : Goel
Current problem list : # Prem 28+4
IUGR - symmetrical. ELBW 613grams
RDS
Hypoglycemia

Prem Female baby born following Em C/section for worsening growth restriction, oligohydramnios and abN dopplers at 28+ 4 weeks.

Maternal Hx:
Primip
A= Ab negative
Serology negative
Rubella immune
Essential Hypertension diagnosed prior to pregnancy - non-medicated
PET in current pregnancy
FTCS test - low risk
Normal morphology scan
26/40 scan : EFW 0.23 centile 672grams. Low AFI 7.6cm (oligo).
Elevated UA PI
27/40 : Normal BPP, breech. UA AEDF. Oligohydramnios
28/40 : Normal BPP, breech, Low AFI (oligo). MCA redistribution. UA doppler IREDV. Normal DV.
29+3 scan (28/8): EFW <<1centile. UAPI AEDF. MCA redistribution. Oligohydramnios. Normal DV doppler.
GBS unknown.

Current meds: Labetolol, asprin, Nifedipine, Methyldopa
Celestone given 25/08 and 26/08.
TORCH negative

Birth HX:
AbN CTG - emLUSCS
Membranes intact
MgSo4 given overnight.

Female baby born 1028 APGARs 9 and 9. Breech extraction.
'Delayed' cord clamping 20 secs
Mild WOB from 10min life - CPAP 5 commenced in air.
Apnea x2 on route from theatre to NICU - requiring stimulation and FiO2 up to 30%. CPAP increased to 7 in air.

Management plan : RESP: CPAP 8
CXR

C/V: NAD
FBE 213/8.2/245

Hypoglycemia: 2ml/kg bolus dex + IVF

FN: TFI 80mls/kg/d

Jaundice: SBR 60->86 Ptx

PLAN 29/8:

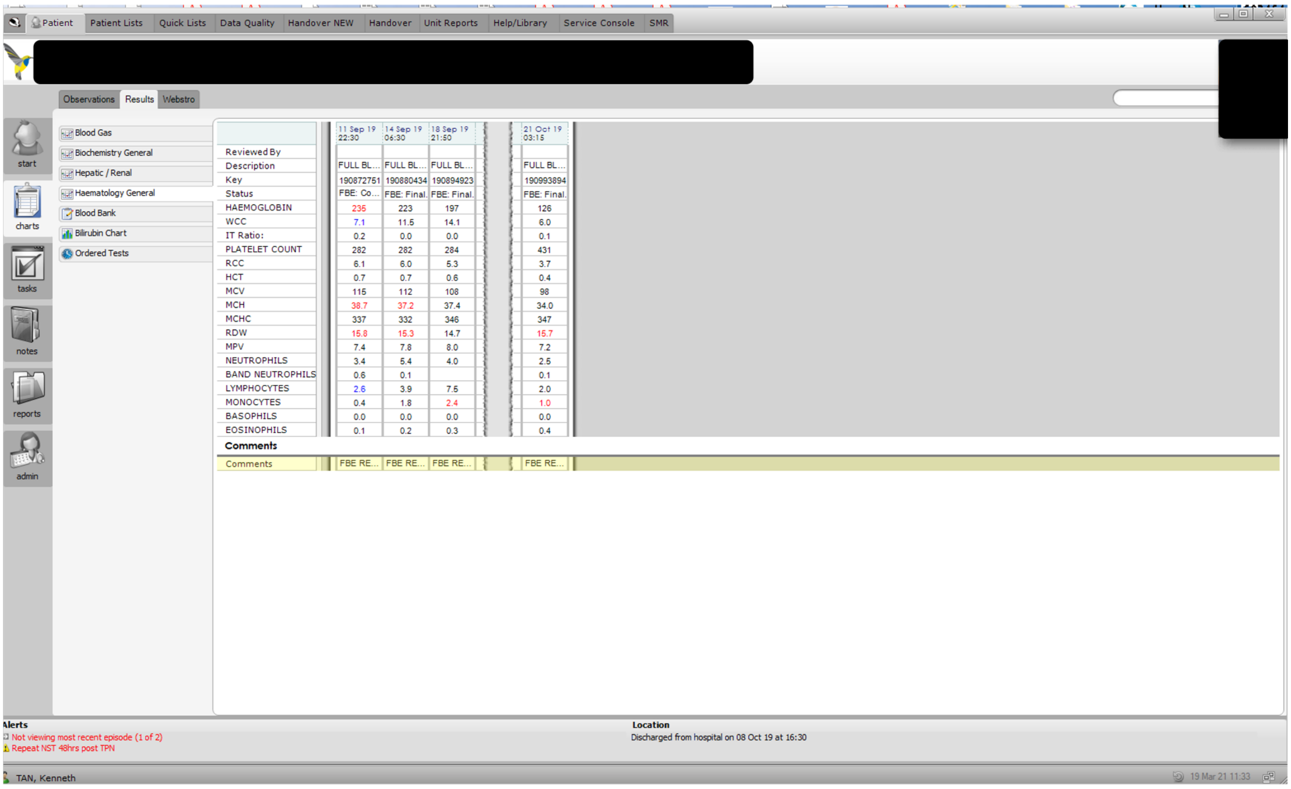
Rpt SBR
Not for IV abs
Completed by : [redacted]

[redacted]

[redacted]

Alerts
Not viewing most recent episode (1 of 2)

Figure 5.3: The interface of Clinicians' notes in BadgeNet EMR system.



[redacted]

[redacted]

	11 Sep 19 22:30	14 Sep 19 06:30	18 Sep 19 21:50	21 Oct 19 03:15
Reviewed By				
Description	FULL BL...	FULL BL...	FULL BL...	FULL BL...
Key	190872751	190880434	190894923	190993894
Status	FBE: Co...	FBE: Final.	FBE: Final.	FBE: Final.
HAEMOGLOBIN	235	223	197	126
WCC	7.1	11.5	14.1	6.0
IT Ratio:	0.2	0.0	0.0	0.1
PLATELET COUNT	282	282	284	431
RCC	6.1	6.0	5.3	3.7
HCT	0.7	0.7	0.6	0.4
MCV	115	112	108	98
MCH	38.7	37.2	37.4	34.0
MCHC	337	332	346	347
RDW	15.8	15.3	14.7	15.7
MPV	7.4	7.8	8.0	7.2
NEUTROPHILS	3.4	5.4	4.0	2.5
BAND NEUTROPHILS	0.6	0.1		0.1
LYMPHOCYTES	2.6	3.9	7.5	2.0
MONOCYTES	0.4	1.8	2.4	1.0
BASOPHILS	0.0	0.0	0.0	0.0
EOSINOPHILS	0.1	0.2	0.3	0.4
Comments				
Comments	FBE RE...	FBE RE...	FBE RE...	FBE RE...

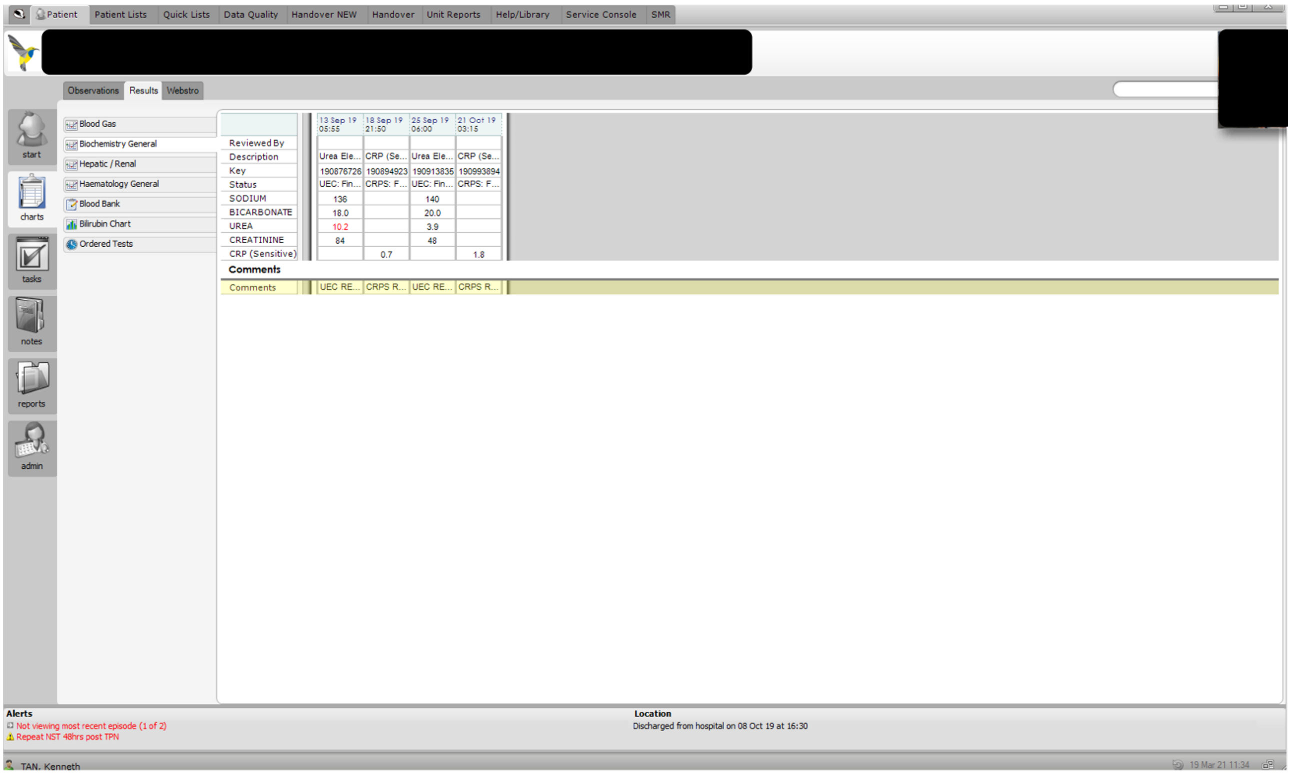
Alerts
Not viewing most recent episode (1 of 2)
Repeat NST 48hrs post TPN

Location
Discharged from hospital on 08 Oct 19 at 16:30

TAN, Kenneth

19 Mar 21 11:33

Figure 5.4: Biochemistry lab test result.



[redacted]

[redacted]

	13 Sep 19 05:55	18 Sep 19 21:50	25 Sep 19 06:00	21 Oct 19 03:15
Reviewed By				
Description	Urea Ele...	CRP (Se...	Urea Ele...	CRP (Se...
Key	190876726	190894923	190913835	190993894
Status	UEC: Fin...	CRPS: F...	UEC: Fin...	CRPS: F...
SODIUM	136		140	
BICARBONATE	18.0		20.0	
UREA	10.2		3.9	
CREATININE	84		48	
CRP (Sensitive)		0.7		1.8
Comments				
Comments	UEC RE...	CRPS R...	UEC RE...	CRPS R...

Alerts
Not viewing most recent episode (1 of 2)
⚠ Repeat NST 48hrs post TPN

Location
Discharged from hospital on 08 Oct 19 at 16:30

TAN, Kenneth

Figure 5.5: Haematology lab test result.

for diagnosis. As a consequence, they often tend to make diagnostic decisions based on not only available vital signs, but extra information, such as laboratory test results. In many cases, some crucial tests even have a higher priority than vital signs during the process. From a medical viewpoint, all the five vital signs are fundamental indexes that indicate patient status and could be used as references to support clinicians' decisions.

5.3.2 Feature Selection

In addition to seeking suggestions from professional medical staff, the selection of features could also be achieved by technical approaches. The purpose of feature selection is to find the optimal subset from the whole feature sets, and remove irrelevant and redundant ones to increase the accuracy ultimately reduce the training time. Feature selection methods can be roughly put into three categories, the filter, the wrapper and the embedded, but as preliminary processing we only used some simple methods to remove those features that are not significantly related to sepsis onset. Note that this feature selection process applies to only vital signs, but not to demographics since they are static. More complex pre-processing will be done in cycle three.

Variance-based selection

Variance-based method is one of the feature selection methods that do not rely on the feature vectors and the target labels. Basically variance of each feature is calculated, and compared to a certain threshold. Any feature that has a larger variance than the threshold will be kept, while those with a smaller variance will be abandoned. The underlying idea of this method is that features with small variance are stable but not discriminative with respect to the target classification. Min-Max scaling has been done on each feature before the calculation of variance to make sure they are at the same level. The variance of each feature in the NICU datasets is depicted in Fig. 5.6.

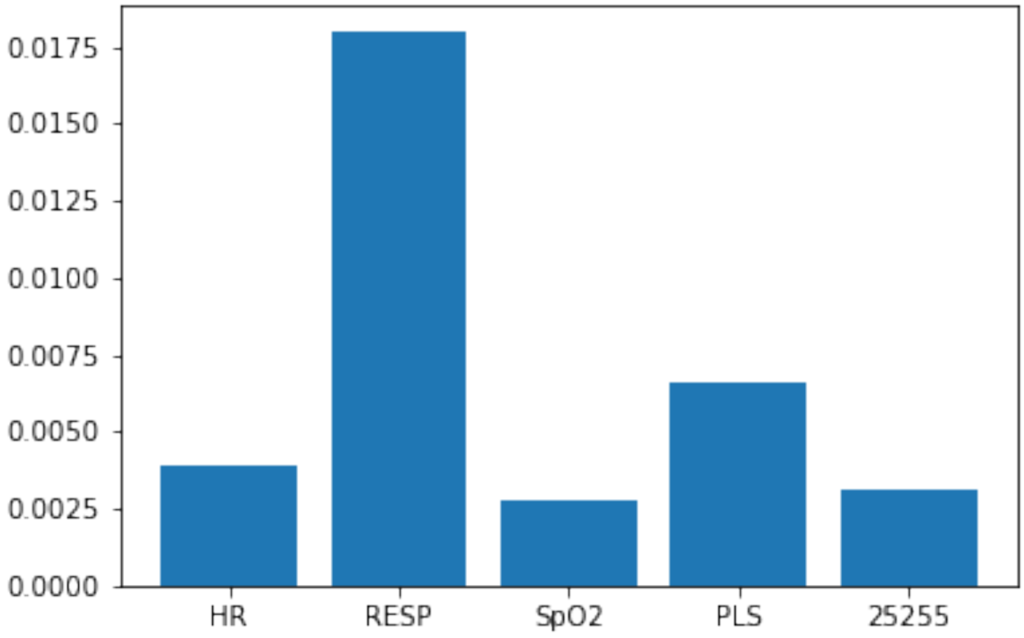


Figure 5.6: The variances of features in NICU datasets. Abbreviations: HR - heart rate, RESP - respiratory rate, SpO2 - blood oxygen saturation, PLS - pulse, 25255 - readings from PS25255 sensors

The scaled variances of five variables in NICU datasets lie between 0.0025 and 0.0175 and from the diagram we can see that respiratory rate has the highest values, aside from it the rest of the four variables are close to each other. Since nither one is significantly smaller than others, considering only five valid vital signs in NICU datasets, we keep all the five features. Meanwhile, the variance-based method has its underlying limitation as it suits the scenario when features have a similar distribution. However, if features are discrete and clustered around only a couple of values, this method has little value in terms of identifying informative features. Also, the selection of the threshold requires large amount of calculation and cross-validation to make sure the chosen features could produce the optimal performance.

Correlation Analysis

To achieve feature selection via correlation analysis has been very common. A well-known metric - Pearson Correlation Coefficient (PCC) was calculated for each pair of features in the datasets. The Pearson Correlation Coefficient is a measure of the linear correlation of two sets of continuous data in statistics. It is derived by dividing the covariance of two variables by the product of their standard deviations, see Formula 5.1.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{5.1}$$

The value of PCC falls between [-1,1], reflecting the degree that two sets of variables are linearly correlated. Closer to 1 means a stronger positive linear relation, while a trend toward the other direction indicates a negative linear relation. Since the target we aimed at is the future prediction of sepsis onset which is categorical data, the Pearson Correlation Coefficient could not be directly applied to calculate the relations between features and the target. In our feature selection task, we cross compared the PCC for each pair of features in the datasets, and listed the results including PCCs and their corresponding p-value below in Table. 5.1 and 5.2.

Table 5.1: Pearson correlation coefficient of each feature in NICU dataset. Abbreviations: HR - heart rate, RESP - respiratory rate, SpO2 - blood oxygen saturation, PLS - pulse, 25255 - readings from PS25255 sensors

	HR	RESP	SpO2	PLS	25255
HR	1.0	0.1103	-0.1436	0.9344	-0.0901
RESP	0.1103	1.0	0.0024	0.1049	0.0472
SpO2	-0.1436	0.0024	1.0	-0.1147	-0.0741
PLS	0.9344	0.1049	-0.1147	1.0	-0.144
25255	-0.0901	0.0472	-0.0741	-0.144	1.0

The values of PCC showed that all five features in the NICU dataset have little linear correlation to each other, supported by the extremely small p-value, except for the pair of 'heart rate' and 'pulse' which showed a very high linear correlation due to their biologically close connection. So as a result the attribute f pulse will be removed out of the final feature set.

The Pearson correlation coefficient can only be used to measure linear relations between

Table 5.2: p-value of PCC of each feature in NICU datasets. Abbreviations: HR - heart rate, RESP - respiratory rate, SpO2 - blood oxygen saturation, PLS - pulse, 25255 - readings from PS25255 sensors

	HR	RESP	SpO2	PLS	25255
HR	0.0	0.0	0.0	0.0	0.0
RESP	0.0	0.0	0.0044	0.0	0.0
SpO2	0.0	0.0044	0.0	0.0	0.0
PLS	0.0	0.0	0.0	0.0	0.0
25255	0.0	0.0	0.0	0.0	0.0

variables. To overcome this limitation, the distance correlation coefficient (DCC) was used to improve the quality of the feature analysis. DCC of two variable u and v is denoted as $dcor(u,v)$, and defined by the Formula 5.2

$$\hat{dcor}(u,v) = \frac{\hat{dcov}(u,v)}{\sqrt{\hat{dcov}(u,u)\hat{dcov}(v,v)}} \tag{5.2}$$

where

$$\hat{dcov}^2(u,v) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3 \tag{5.3}$$

and $\hat{S}_1, \hat{S}_2$ and $\hat{S}_3$ are calculated by the following formulas

$$\begin{cases} \hat{S}_1 & = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||u_i - u_j||_{d_u}||v_i - v_j||_{d_v} \\ \hat{S}_2 & = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||u_i - u_j||_{d_u}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||v_i - v_j||_{d_v} \\ \hat{S}_3 & = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}||u_i - u_l||_{d_u}||v_j - v_l||_{d_v} \end{cases} \tag{5.4}$$

In fact, the distance covariance is calculated by forming a double centred matrix from each variable vector, from which we know that the distance covariance (and correlation) is not the covariance (or correlation) between the distances themselves. It is the covariance (correlation) between the special scalar products. That is the reason that it can capture non-linear relation better than the PCC. Let us observe the DCC between feature pairs in our NICU dataset in Table 5.3 below. The DCC showed that most pairs of features are not very closely related, just like what the PCC values indicated, except for the pair of heart rate and pulse because they are biologically connected.

Tree-based methods

Using tree-based methods to filter useful features is actually one of the implementations of the embedded feature selection approaches. Some machine learning models have the

Table 5.3: Distance correlation coefficient of feature pairs in NICU dataset. Abbreviations: HR - heart rate, RESP - respiratory rate, SpO2 - blood oxygen saturation, PLS - pulse, 25255 - readings from PS25255 sensors

	HR	RESP	SpO2	PLS	25255
HR	1.0	0.1051	0.1602	0.9508	0.1401
RESP	0.1051	1.0	0.0371	0.1019	0.0577
SpO2	0.1602	0.0371	1.0	0.154	0.0627
PLS	0.9508	0.1019	0.154	1.0	0.1566
25255	0.1401	0.0577	0.0627	0.1566	1.0

ability to assess the importance of features inherently, such as Random Forest and Gradient Boosting Decision Tree. These models are friendly to non-linear relations between variables and target, and are easy to apply to them. We analysed the feature's importance by applying xgboost and random forest models and generated figures to illustrate the importance of each feature, see Fig. 5.7 and 5.8. As limited by the space, labels in x axis are impossible to be fully displayed. We noticed that the importance distribution derived from two ensemble methods is quite similar in that there is no significant difference between features except for a few of them on the left most side. Generally, the random forest model tends to lean on a small proportion of the entire feature set more than the xgboost. To compare the result from two methods, we also listed the top 15 feature from each side in Table. 5.4. The names of features are specified in the format of X_Y, where X is the real name of the feature, and Y is the suffix indicating the order within the 60-minute sliding window if Y is a numeric value, or the descriptive statistics if Y is a string like ”MAX”. Both methods indicate that the variance of respiratory rate is the most decisive feature. Besides, out of the 15 features, seven of them occurred twice, which showed the similarity as well. The scores from two methods were at difference scales since they are calculated in different approaches. For xgboost, the

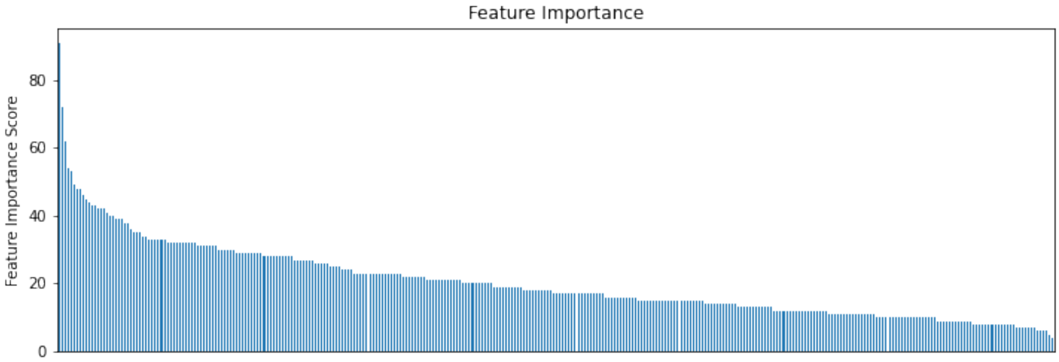


Figure 5.7: The importance of features in data chunks consisting 60-min data extracted from NICU dataset with xgboost.

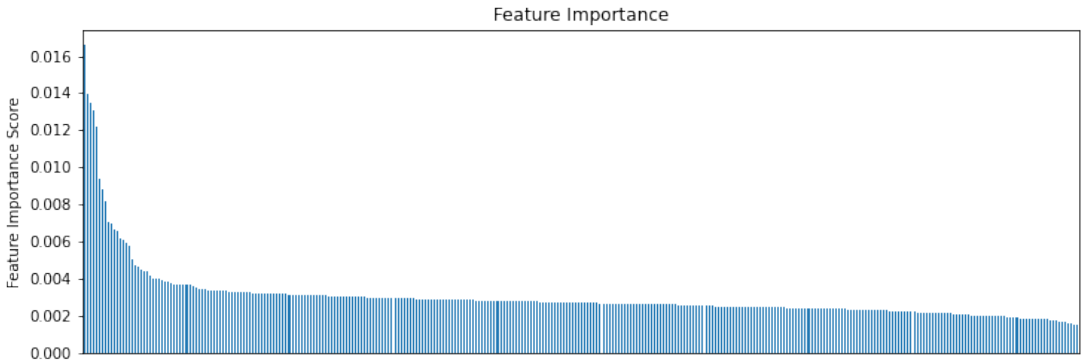


Figure 5.8: The importance of features in data chunks consisting 60-min data extracted from NICU dataset with RF.

Table 5.4: Top 15 important features derived from two methods. Abbreviations: HR - heart rate, RESP - respiratory rate, SpO2 - blood oxygen saturation, PLS - pulse, 25255 - readings from PS25255 sensors, VAR - variance, PTP - peak to peak, MAX - maximum

xgboost			random forest		
No.	Feature	Importance	No.	Feature	Importance
1	RESP_VAR	0.016613	1	RESP_VAR	91
2	RESP_PTP	0.013988	2	SpO2_VAR	72
3	PLS_VAR	0.013513	3	PLS_VAR	62
4	HR_VAR	0.013053	4	SpO2_MEAN	54
5	PLS_PTP	0.012171	5	HR_VAR	53
6	RESP_MAX	0.009408	6	RESP_MIN	49
7	PLS_MAX	0.008826	7	25255_29	48
8	HR_PTP	0.008185	8	25255_VAR	48
9	25255_VAR	0.007060	9	RESP_3	46
10	PLS_MEAN	0.007012	10	RESP_42	45
11	HR_MEAN	0.006626	11	PLS_MAX	44
12	HR_MAX	0.006593	12	RESP_2	43
13	SpO2_MEAN	0.006154	13	RESP_41	43
14	25255_PTP	0.006110	14	RESP_MAX	42
15	SpO2_MEDIAN	0.005963	15	RESP_14	42

value of importance denotes the number of the feature that was used to split data across all trees, while in random forest, it is called Gini importance measuring the average purity gain by splitting certain features.

5.4 Cycle Three

With the preliminary processed dataset produced in cycle two, we can explore how various machine learning models could learn from them and make predictions of sepsis to facilitate clinicians' diagnosis. In the third cycle we developed different types of models and algorithms, and trained them on the NICU dataset to achieve early detection of neonatal sepsis. After that, outcomes were evaluated iteratively to adjust and improve our models.

5.4.1 Classic Machine Learning

Many classic machine learning models have been shown to be effective in classification problems with less complex structures and parameters, like SVM, Random Forest. We designed a scheme to check how conventional ML models work with our sepsis data. Although some of these models were reported in previous research, with some changes in data processing and training, they still demonstrated a promising result. Fig. 5.9 shows the process of model building, training and evaluation.

Preprocessing

The bedside monitoring machine has a sampling rate of 200Hz and is able to output one record in CSV format every second. Data is accumulated at a very high speed and will bring excessive pressure on data storage, and such frequent data is not necessary for ML tasks. Thus, we scale down the data to one record per minute by sampling, which reduces the size of the dataset to 1/60. The variables used are chosen according to experienced paediatricians and our feature selection procedures, and given the limitations of the information the bedside monitoring machine broadcasts, we only use heart rate (HR), respiratory rate (RR) and blood oxygen saturation (SpO2) as input vital signs.

In most cases, clinicians would start using antibiotics while they found an infant showing suspicious symptoms and ordered blood culture test, so we define two types of patients which are ″sepsis″ and ″normal″ based on whether antibiotics treatment is taken. Any one conducted a blood culture test or accepted an antibiotic treatment should be labelled

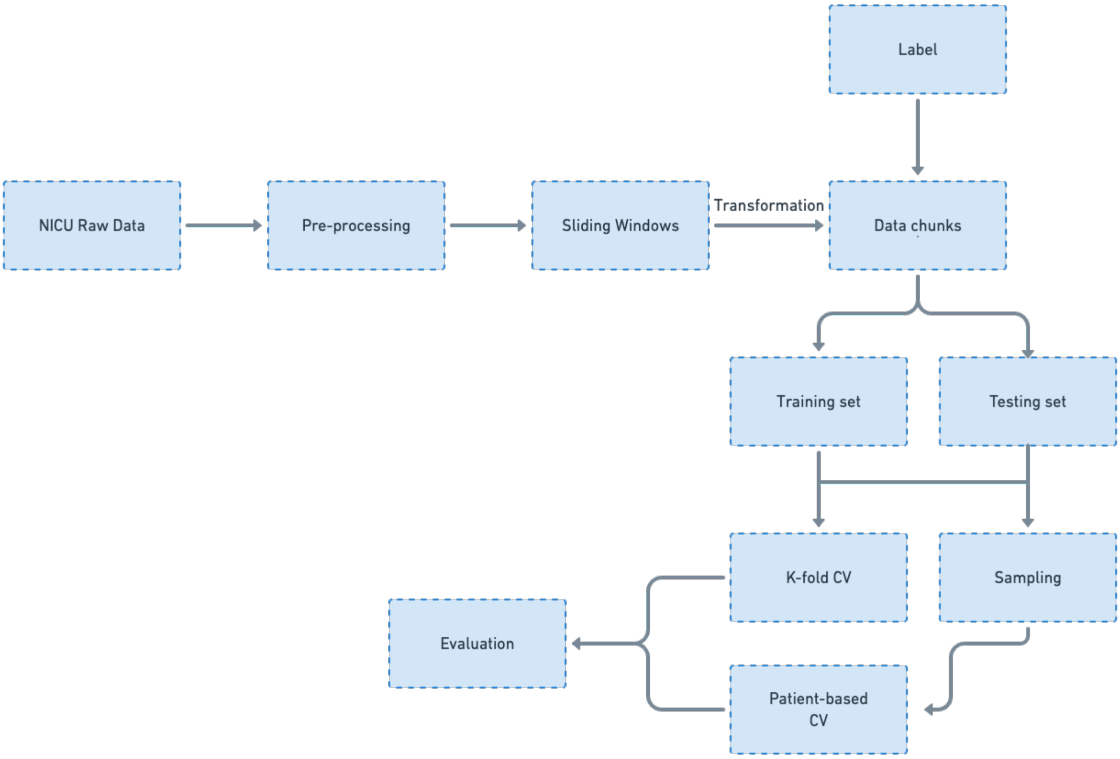


Figure 5.9: Global process of model building, training and evaluation with classic machine learning algorithms.

”sepsis” and the label ”normal” belongs to the rest. If one baby never did blood culture test or accepted antibiotics, he is absolutely classified into normal group, however if there existed suspected sepsis in this baby, the situation becomes subtle. Let the time of blood culture test or antibiotic therapy be denoted as t_0 , and the time of ceasing antibiotic as t'. Since we want to detect the sepsis 6 hours ahead, suspicious data should be expanded from t_0 to $t_0 - 6$. Note that usually the last 24 hours in the antibiotic treatment patients are almost fully recovered, so we set the end point of suspicious data at $t' - 24$, and all data fall into $t_0 - 6$ to $t' - 24$ is labelled ”sepsis” while the rest is labelled ”normal”.

Since the infants in the NICU may have moved to other locations and sensors may detach from the infants' bodies, there could be missing or error values in the dataset we have collected. Prior to importing data to the ML models, we have to filter out these invalid data. Two different strategies of data cleaning were tested, one is to delete the data block with missing or error data directly, the other is to replace them with the mean value. It turned out both strategies led to hardly any differences between the final results, so we decided to choose the simpler one which was to just delete the data block with invalid values. With all the pre-processing procedures done, finally we have got 4412 data samples in total, of which 3451 are with ”normal” label and 961 are with ”sepsis”.

Sliding Windows

The collected data stream is represented as a temporal sequence of a variety of parameters. To feed them into normal ML models, we performed some data transformation steps before training. We introduced a sliding window method to convert temporal data flow to data blocks. Specifically, raw data is cut into data blocks by sliding windows according to time stamps. The width of the sliding windows is set to 60 minutes, which means that in every data block we have 60 records of each variable since the sampling rate is once per minute. For each variable, not only do the 60 data points account for input features, we also calculate extra statistical parameters for 60 minutes. The statistical parameters are maximum and minimum value, mean and variance, medium value, and peak-to-peak value. Finally, each variable contributes 60 data points within the sample hour plus another 6 extra statistical parameters which totally 66 input features. Then the features from five variables (HR, RR, SpO2, Pulse and 25255) are concatenated and combined with static variable age and birthweight to form the final input feature vectors including 332 features, see Table 5.5.

Table 5.5: Structure of feature vectors. Abbreviations: HR - heart rate, RESP - respiratory rate, SpO2 - blood oxygen saturation, PLS - pulse, 25255 - readings from PS25255 sensors, BW - birthweight.

HR	Stats of HR	RR	Stats	SpO2	Stats	Pulse	Stats	25255	Stats	Age	BW
60	6	60	6	60	6	60	6	60	6	1	1

Data Rebalance

When sliding the window, different step sizes are applied based on labels. The step size is 60 minutes when the label is ”normal” and reduced to 10 minutes when the label is ”sepsis”. The reason we do this is that the number of normal cases and suspicious cases are extremely imbalanced - the former one is much more than the latter one. As we discussed in previous chapter, if we train ML models directly using these imbalanced data, they tend to classify all the data as ”normal” because in this way a better accuracy can be achieved. We set the step size smaller than window size in suspicious cases, so that there will be certain parts overlapping between two neighbouring windows, and in this case, with the same size of data stream we are able to create more data blocks, in other words training samples with ”sepsis” label. Note that the overlapping in our data samples will not increase the chance of overfitting because the relative positions of overlapping data in each data block are different. However, even expanding the suspicious data by overlapping the window, there is still a huge gap in terms of the amount between ″normal″ and ″sepsis″ data. We down-sampled ″normal″ data to make sure the size difference between the two classes was acceptable.

Training on Existing Models

A representative series of ML classification models were chosen for this suspicious patient prediction task. We chose Logistic Regression (LR), Support Vector Machine with Gaussian kernel (SVM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Multi-layer Perceptron Neural Network (NN) as target models, and trained them with preprocessed feature vectors with comparison.

1. K-fold cross validation

   In this scheme, all 4421 data samples were input to the algorithm without them being divided into a training set and a testing set. We set k to 4 which means the data was going to be divided into four segments after a random shuffle, and each segment would be taken as testing set for the model trained by the rest of the three segments.

   Given the random shuffle before the segmentation, data samples from different patients will be mixed. Our target is to test the prediction performance of chosen models under this scheme, because they could possibly learn from some patients part of whose data belongs to the training set while the other parts are included in the test set. In such a case we could examine whether our model could learn from what we have seen and predict the label correctly.

2. Patient-wise validation

   Unlike K-fold Cross Validation, we did not mix data from different patients. We shuffle patients instead of data samples, in order to keep the normal-sepsis ratios of training and testing data at the same level. ″Normal″ and ″Sepsis″ patients are shuffled separately but split into training set or testing set at the same ratio of 7:3 so we have 2416 ″normal″ data samples and 673 ″sepsis″ data samples in the training set, and for the testing set, the numbers are 1035 and 288 respectively. Once a patient is chosen to be in training or testing set, all data samples from him will be included in and only in that set, in other words, data from one patient will not be in training and testing set simultaneously.

   The purpose of designing this scheme is to verify the generalisability of our models. The models predict the condition of a group of patients by learning from another group of patients, which we believe is closer to the real circumstances in the NICU.

We trained the five models mentioned above with scikit-learn. For fine-tuning purposes, grid search method is employed. We have to cross check a variety of values in order to find the optimal one in each model. If a candidate value list of one parameter is provided, the GridSearchCV module is able to train the model with all the values in the list respectively and give a score defined by users, picking up the parameter makes the model have the best performance. For example, we tried different numbers of trees in RF in {10, 20, 50, 100, 150}, and the max depth of each tree is set in {1, 3, 5, 10, 15}. Another point worth mentioning is that although some measurements have been taken in the pre-processing stage, there are still many more normal data samples than suspicious ones, so we use the Cost-Sensitive Learning method during the training process, applying a higher weight to ″sepsis″ samples so that the error is enhanced when they are wrongly classified. In this way, the classifier care more about ″sepsis″ samples of small amount, and prevent the tendency of the bias to ″normal″ label which is the majority of the entire dataset.

Performance Evaluation

In this section, we report the result of the five models in two training schemes.

1. K-fold Cross Validation

   Table 5.6 shows the performance of five classification models trained by the dataset collected from 23 rd Oct to 23rd Nov with 5-fold cross validation. Note that for LR, SVM and NN, we normalised the data samples before training, scaling them down with a mean of 0 and a variance of 1, just to make sure the model would not be biased to the features with larger values. RF and GBDT do not need the normalisation though because they are both tree-based model.

Table 5.6: Performance of five models in 5-fold CV. Abbreviations: LR - logistic regression, SVM - support vector machine, RF - random forest, GBDT - gradient boosting decision tree, NN - neural network

Model Name	Accuracy	Precision	Recall	Weighted_F1	AUC
LR	0.66	0.66	0.65	0.65	0.71
SVM	0.72	0.74	0.67	0.71	0.79
RF	0.74	0.77	0.68	0.72	0.81
GBDT	0.73	0.76	0.68	0.72	0.81
NN	0.67	0.68	0.66	0.67	0.73

We can see that in general, the difference among five models is trivial, but LR, and NN did not perform very well compared to other models. Since the dataset has a complex structure with multiple variables, it must not be linearly separable. The linear model LR had a poor average performance of about 0.6+ on the existing data, while others could reach 0.8+. NN is capable of learning from a non-linear dataset, but it may need extra techniques or a more complex network structure to achieve better performance. The f1 value in the table is weighted according to the proportion of the data given the certain label, so it may be larger than precision and recall. The SVM and two tree-based ensemble learning models had a good performance with AUC of around 0.8, which we believe that it is possible to separate infants with suspected sepsis and healthy ones.

2. Patient-wise Validation

   We also conducted some tests in this more realistic situation in which models are trained with data from some patients and predictions are made on others. The parameters of each model are taken from the best model in a 5-fold cross validation scheme. To reduce possible errors, we run each test four times and calculate the mean value as the final result. Table 5.7 provides the final performance measurements. In this training scheme, the difference between the five models became larger, but still, RF achieved the best performance in terms of almost every evaluation metric and it performed way better than the other models, which we can tell from the ROC curve. Except precision, recall and f1 score, AUC was used as the general measurement because it provides an efficient measurement regardless of the size of data and is almost not sensitive to the imbalanced data samples[188]. Fig. 5.10, 5.11 shows the ROC curves of all the five models under two training schemes.

Table 5.7: Performance of five models in patient CV scheme. Abbreviations: LR - logistic regression, SVM - support vector machine, RF - random forest, GBDT - gradient boosting decision tree, NN - neural network

Model Name	Accuracy	Precision	Recall	Weighted_F1	AUC
LR	0.50	0.50	0.44	0.47	0.50
SVM	0.54	0.61	0.20	0.30	0.53
RF	0.77	0.75	0.81	0.78	0.77
GBDT	0.64	0.66	0.56	0.60	0.64
NN	0.50	0.50	0.85	0.63	0.50

In the 5-fold cross validation scenario, the ensemble models, especially the random forest performed much better than the other three models, and this should be attributed to their innovative strategies. RF is an enhanced version of decision tree with a bagging strategy. The two-step random sampling ensures the model will not easily overfit even without

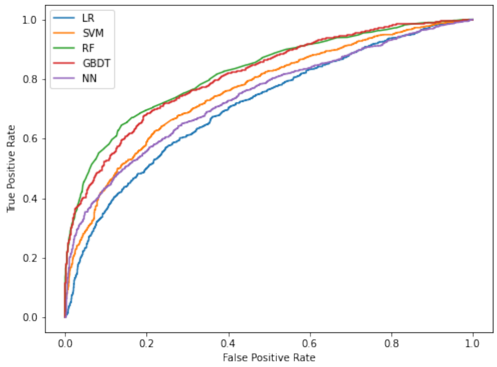


Figure 5.10: ROC curves in 5-fold CV schemes.

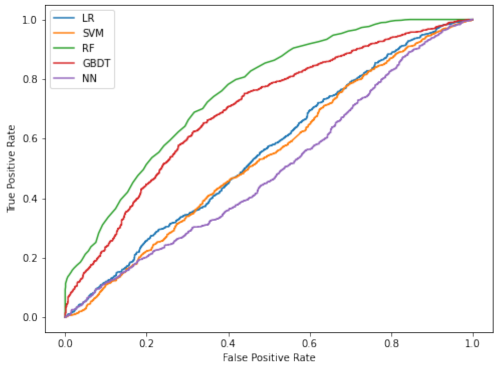


Figure 5.11: ROC curves in patient-wise validation schemes.

pruning. Every single tree in GBDT tries to fit the residual of its former ones result. When summing them up, the final result will be much better. Ensemble learning strategies make these two algorithms more capable of learning from large numbers of data with complex feature combinations.

In comparison to the first scheme, the AUC of RF and GBDT is lower in the patient-based validation scheme. We believe it is because of the nature of the data. Under this training scheme, training data and testing data are from different patients, the patterns that the models have learned from one patient may not be applicable to others. Every individual patient may have a unique reaction to the infection and produce a special pattern in physiological data streams due to different body conditions. Therefore, our models can not learn the pattern of a particular patient from others, unless we have a larger enough cohort of patients which can cover most infected cases. However, since the hospital had not previously kept backups of vital signs data of patients who were cared for in the NICU, and the data collection task was performed for just one month, all the data we have was from less than 180 individual infants, which obviously can not represent all the neonatal sepsis patients.

From the interpretability point of view, two tree-based models, RF and GBDT have an inherent advantage that conditions based on which branches were split help the interpretation of the model. Although ensemble strategies like bagging and boosting were employed, it did not hinder the models from being easily understood. LR is also a simple model that has high interpretability simply because of the linear relation between the input features. However, SVM and NN are less interpretable than the other three due to their non-linearity and mapping in different feature spaces.

5.4.2 Deep Learning 1 - Conversion to Images

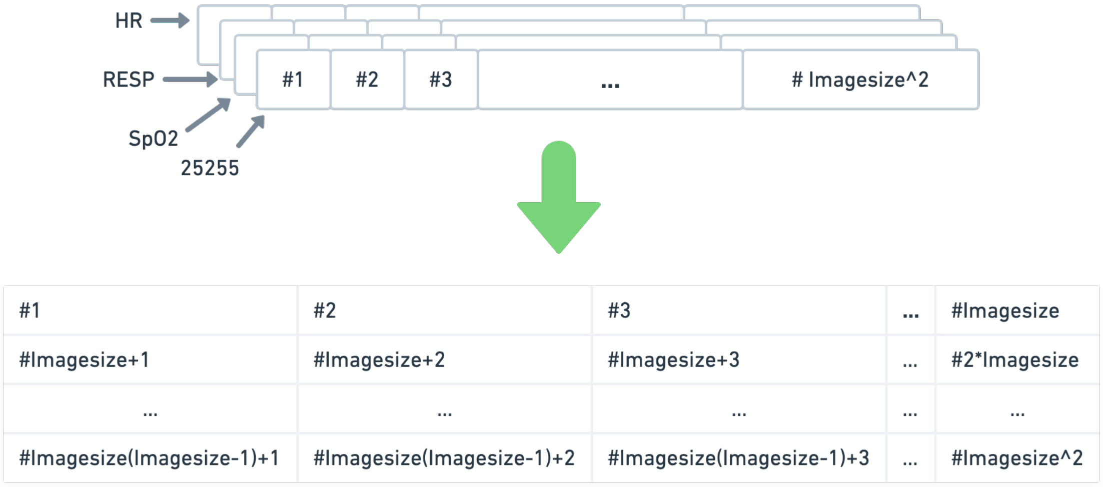
As more complex models can fit better into datasets, especially larger ones, we have also proposed two deep learning based approaches. The first one converts time series data into colour images, utilising the advantages of 2d Convolutional Neural Networks in terms of image processing.

Simple Transformation

The raw data captured by eDataGrabber software is a compressed file with a cpz suffix in the file name and a text file recording the vitals signs trend can be retrieved once the file is uncompressed. This trend file is stored in CSV format with timestamps on each line so that we can separate the file easily according to the date. The bedside monitor can sample HR at a rate of 200Hz and output one record per second, but considering comparison experiments, we also made another version of data by sampling it at one record per minute. Another step of preprocessing is data cleaning. For many reasons, such as human intervention or machine failure, there could be invalid data. To deal with them, we first removed all the data files in which invalid records ratio is more than 10%, then performed linear interpolation to replace the invalid record with a linear model of the particular features. If linear interpolation can not be done in cases where there are only two available samples, the nearest value will be filled in.

The novelty of this method lies in the transformation from vital signs streams to images. The continuously recorded vital signs from bedside monitors in the NICU could be considered as data sequences, but CNNs with 2-D filters do not work well with sequential data, so we cut the data sequence to a certain length, reshape it into a 2D matrix which is called a data chunk, and treat the stream data as flat images. The process is depicted in Fig. 5.12.

Since the maximum size of continuous data records from one patient is around 1,400, see Fig. 5.13, the number of data points within a single image have to be less than that. Considering the size of the data chunks we have we choose to map the clinical data to 16×16 and 32×32 images. If the image size is too large, as the data amount is fixed, we will get a smaller number of images, and for a deep learning model, less training samples usually leads to overfitting. On the contrary, if the image size is small, CNNs will not be able to extract enough features from the image which is also not good for training. As a coloured image contains red, green and blue (RGB) channels, and values in each channel are between 0 and 255. We scaled the time series data into the range of $[0, 255]$ by the Formula 5.5, making



#1	#2	#3	...	#Imagesize
#Imagesize+1	#Imagesize+2	#Imagesize+3	...	#2*Imagesize
...	...	...	...	...
#Imagesize(Imagesize-1)+1	#Imagesize(Imagesize-1)+2	#Imagesize(Imagesize-1)+3	...	#Imagesize^2

Figure 5.12: Process of simple transformation.

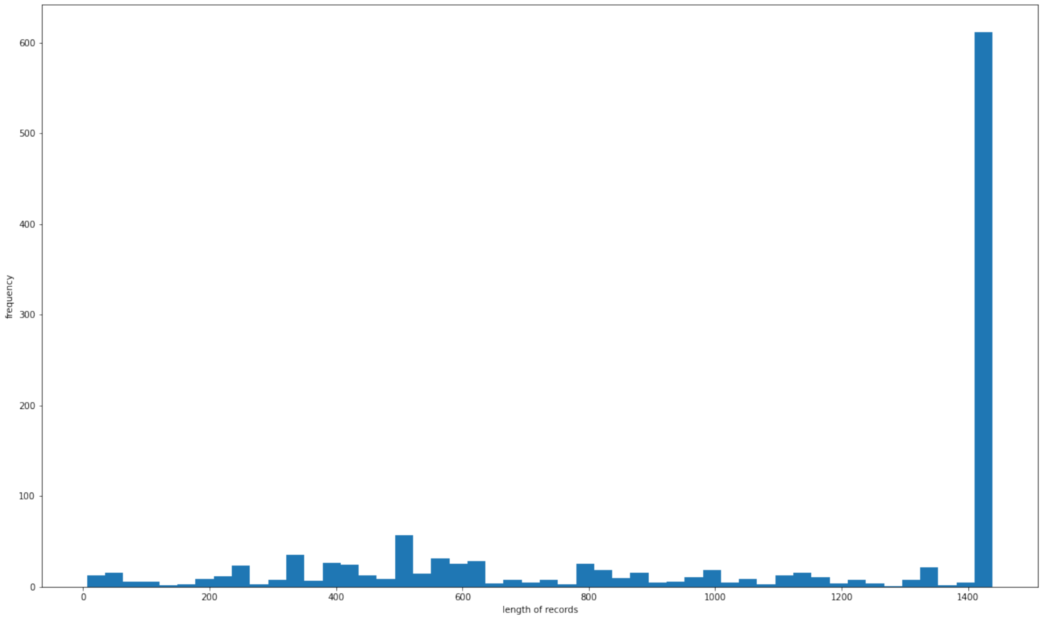


Figure 5.13: Length distribution of features in NICU dataset.

sure that every data point can be mapped to a valid colour. Sample images could be seen in Fig. 5.14. As the figure shows, different data segments can be mapped to different images with unique colours and textures. Since all the variables we handle stay in a certain range, the mapped images will not be very distinct especially in terms of the overall colour which is different shades of green. However, texture makes them unique from each other due to the random distribution of events and newborns different health conditions. Under this circumstance, our task becomes classifying normal images and sepsis images.

$$x_{new} = \frac{(x_{origin} - x_{min})}{x_{max}} \times 255 \tag{5.5}$$

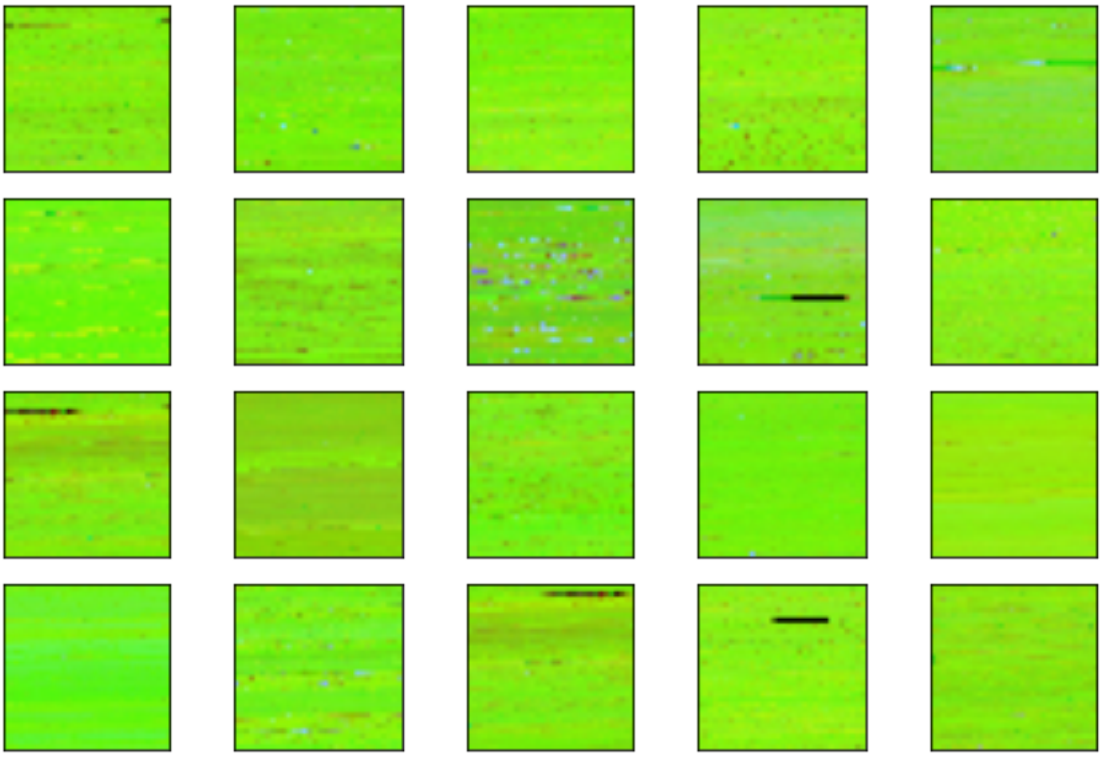


Figure 5.14: Colour image samples transformed from clinical data stream, HR, SpO2 and 25255 are mapped to RGB channels.

Images need to be labelled before fed into the CNN model in our task. As limited by the number of patients and infection ratio, we do not have enough number of proven sepsis data samples. Therefore, any episode suspected by clinicians will be considered as a sepsis case. We define two types of patients before we label them, normal and sepsis. For those who were suspected by clinicians at a certain period, no matter whether blood culture has a growth or not, they will be classified as sepsis. Mostly, clinicians will order antibiotics while they call a blood culture test just in case, and the time point when antibiotic treatment is ordered is noted as t_0 , all the data chunks from $t_0 - 6$ will be labelled as sepsis until antibiotic treatment stopped. Because some subtle changes may occur to babies before doctors or nurses notice that, data before t_0 may still contain sepsis information as well. On the other hand, it makes the prediction six hours ahead, which allows doctors more time to stop the disease.

Gramian Angular Field Based Transformation

Inspired by the work of Wang and Oates [189], we mapped univariate time series from traditional cartesian coordinates to polar coordinate, and generated two types of images: the Gramian Angular Summation Field (GASF) and Gramian Angular Difference Fields (GADF). Each element in these images is actually the summation or difference of the angles respectively. Rather than simply mapping time series values to RGB channels, the Gramian Angular Field (GAF) has the advantage that is able to keep the temporal information of the time series.

Gramian Matrix Gram matrix is a useful tool when calculating the relations among a set of vectors. It is basically a matrix consisting of the dot product of each pair of vectors, see Formula 5.6

$$G = \begin{Bmatrix} < u_1, v_1 > & < u_1, v_2 > & \cdots & < u_1, v_n > \\ < u_2, v_1 > & < u_2, v_2 > & \cdots & < u_2, v_n > \\ \vdots & \vdots & \ddots & \vdots \\ < u_n, v_1 > & < u_n, v_2 > & \cdots & < u_n, v_n > \end{Bmatrix} \tag{5.6}$$

Encoding To project time series data to polar coordinate, we need to consider two variables: the value and timestamp which are mapped to angle and radius. Since the angle falls in the range of [-1,1] in radian system, we must scale the values into the same range first. Given a time series $X = \{x_1, x_2, \ldots, x_n\}$ of n real-value observations, a min-max scaler was applied to achieve the rescale as follows:

$$\hat{x_i} = \frac{(x_i - max(x)) + (x_i - min(x))}{max(x) - min(x)} \tag{5.7}$$

Then we can represent the rescaled time series $\widetilde{X}$ in polar coordinates by encoding the value as the angular cosine and timestamp as the radius with the formula below:

$$\begin{cases} \phi & = \arccos(\widetilde{x_i}), -1 \leq \widetilde{x_i} \leq 1, \widetilde{x_i} \in \widetilde{X} \\ r & = \frac{t_i}{N}, t_i \in \mathbb{N} \end{cases} \tag{5.8}$$

when t_i is timestamps and $\mathbb{N}$ is a constant factor to regularise the span of the polar coordinate system. After the rescaling and transformation, we can easily exploit the angular perspective by defining GASF and GADF as:

$$GASF = [\cos(\phi_i + \phi_j)] \tag{5.9}$$

$$GADF = [\sin(\phi_i - \phi_j)] \tag{5.10}$$

where ϕ_i and ϕ_j are the angles of i-th and j-th vectors in the time series. The encoding process is also demonstrated in Fig. 5.15. We use pyts library [190] to implement the GASF image transformation, and sample images are shown in Fig. 5.16

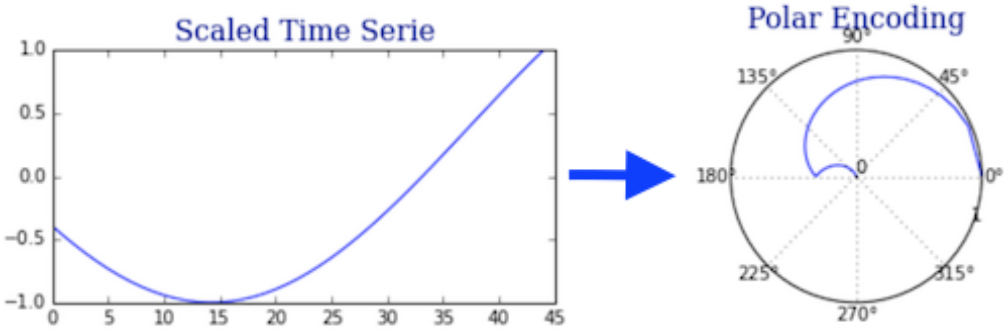


Figure 5.15: Encoding of GASF.

Convolutional Neural Network

Convolutional Neural Network is one branch of feedforward neural networks which performs well when dealing with data of grid-like structures, such as time series, voices and images. See Fig. 5.17 for a typical structure of CNN. It only depicts the key components but omits some other auxiliary layers such as pooling and dropout. What makes CNN special from other deep neural networks is the convolutional layers. To implement the convolution operations, a matrix of smaller size called the filter/kernel/weights is introduced, and by moving the filter throughout the input feature map, summing up all the products of the numbers from the input features and the filter at the corresponding positions, it produces the output feature map. Fig. 5.18 illustrates how the convolution is calculated. Because of the nature of the convolution operation, it is able to extract local features from the original input feature map, depending on the parameters in the filter.

The ordinary fully connected neural network requires a very high number of weights to

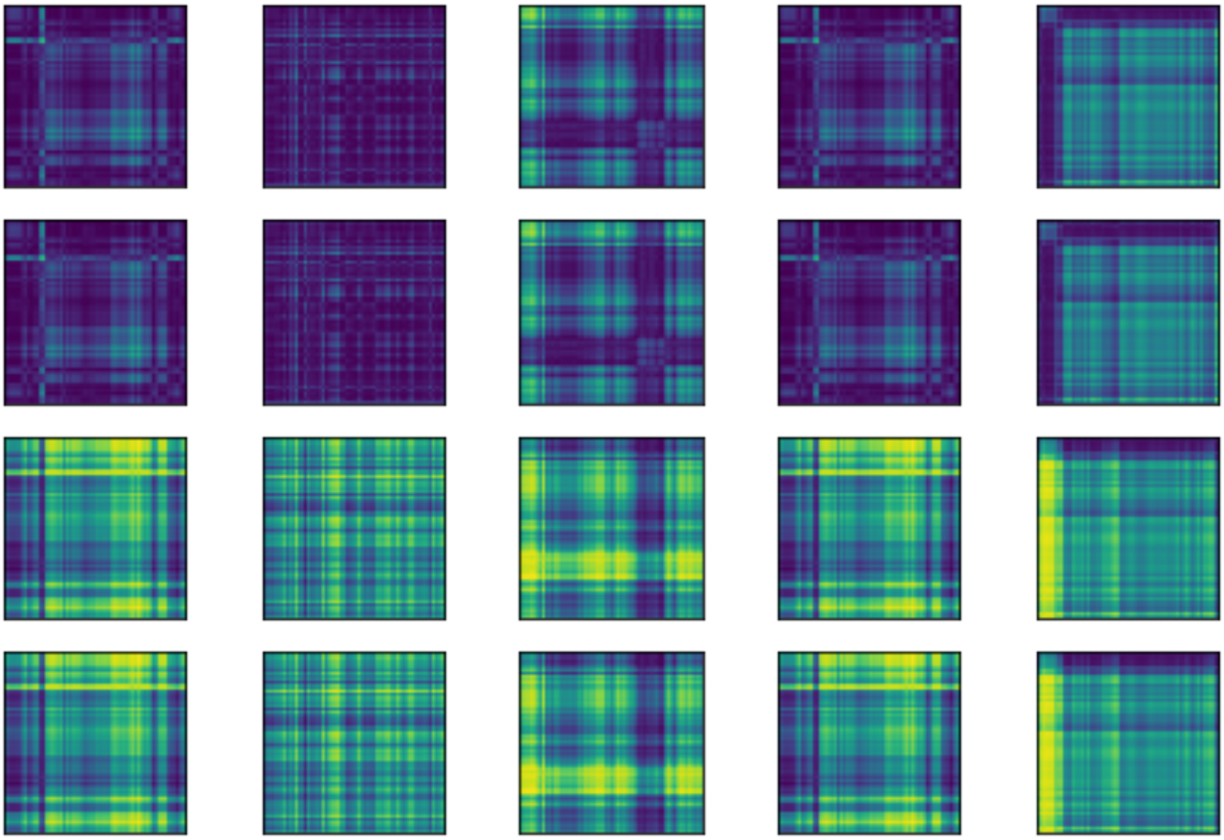


Figure 5.16: Sample images generated by GAF. (Top two lines are GASF, and bottom two lines are GADF.)

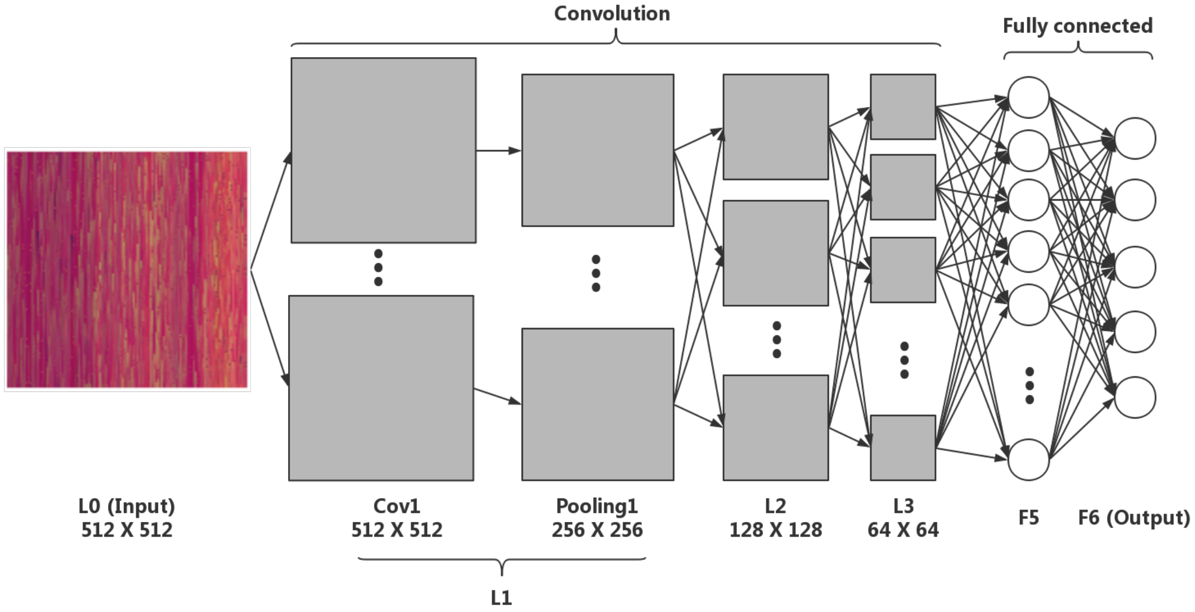


Figure 5.17: A simplified structure of convolutional neural network.

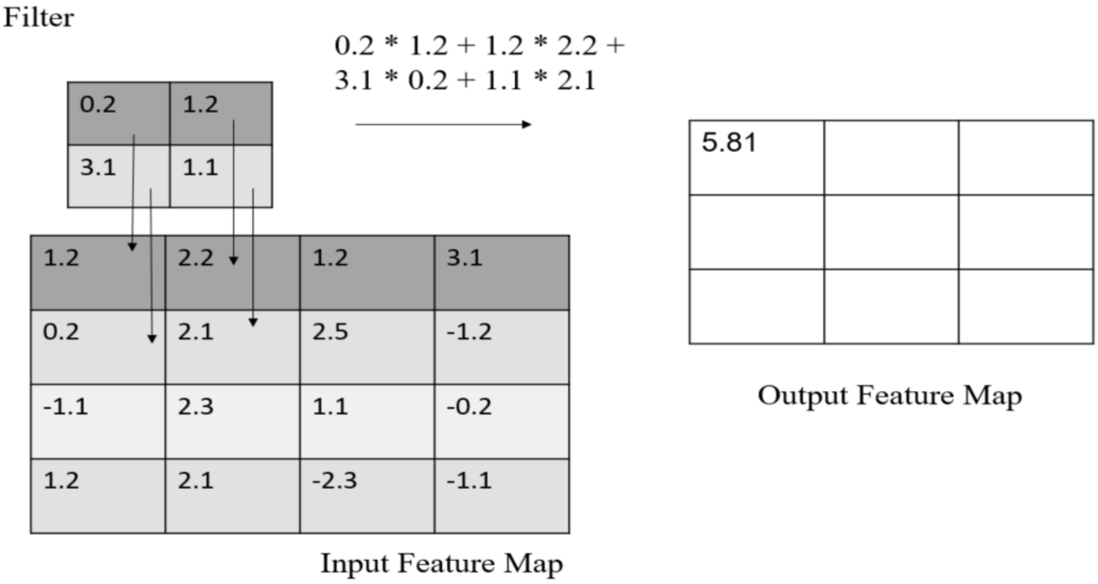


Figure 5.18: The convolution operation.

be tuned if the input size is large even in a shallow architecture. For instance, a 100×100 image has 10000 weights for each layer. The convolution operation mitigates this problem allowing the network to be deeper with fewer parameters. If the filter size is 5×5, only 25 learnable parameters are needed, regardless of the size of the input data size, because all the data points in this layer share the 25 parameters.

Network Structure

In this particular case of sepsis prediction, by folding the 1-D time series data into a 2-D image data, some data points which were not originally adjacent become neighbouring, so that it is possible for the moving filter to identify some latent relations between discrete data points. Besides, when sliding the same filter across the entire image, which is actually the data series, it means the feature that may determine the occurrence of sepsis can be learned no matter where it is. This property is called spatial invariance. In our thesis, we designed a 14-layer deep convolutional neural network for our sepsis image classification task. For all the training tasks, unless otherwise stated, this structure is used. Fig. 5.19 shows the structure in detail, and here is a brief explanation of the components in CNNs. The input layer is a r×r×3 matrix of three independent variables, which CNN treats like a coloured image. The convolutional layer uses a 3×3 filter to extract features from the so-called image which is actually transformed from clinical data chunks. Batch normalisation layer normalises data samples in every dimension of the input features, and ReLU layer adds non-linear parts in the model to make it different from the Multilayer Perceptron (MLP) model. Max pooling is a common approach used in deep learning models to reduce the dimension of the feature vector, and it helps avoid overfitting. The other way to avoid overfitting is to add a dropout layer before the fully connected (FC) layer. It randomly sets some neurons in the network to zero, applying a bagging-like strategy to the network, which has a great effect on the model optimisation. Finally, the fully connected layer and SoftMax layer, calculate the possibilities that the image belongs to each class. Considering the size of data samples, the CNN models should not be designed to be too complex, because a complex model with a small data size is prone to overfitting. Our model has only 14 layers, which is much shallower than the famous GoogleNet [191] and ResNet [192] which have 22 layers and 152 layers respectively.

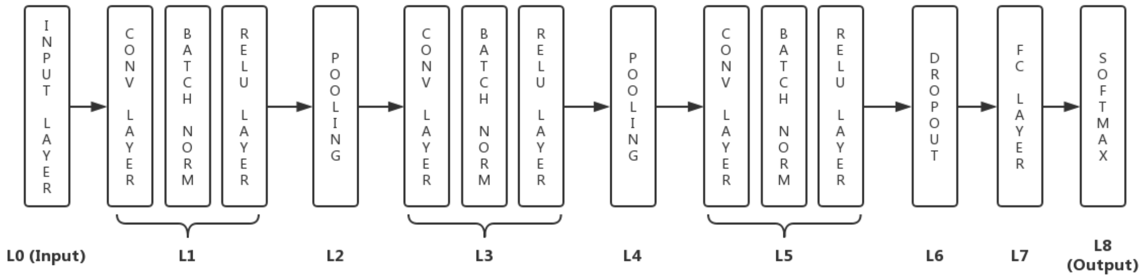


Figure 5.19: Detailed structure of CNN used in our method.

Performance Evaluation

To find a better solution in the simple transformation, we evaluated the model with several different setups.

1. Train a CNN model on 16×16 simple-transformed images.

2. Train a CNN model on 32×32 simple-transformed images.

3. Train a CNN model on 32×32 GASF-transformed images.

4. Train a CNN model on 128×128 GASF-transformed images.

5. Fine-tune a pre-trained AlexNet model on 227×227 GASF-transformed images.

In the first setting, we had transformed 27477 normal images and 5041 sepsis images while in the second setting, as the size of the image increased, the number of images dropped to 7670 and 1486 respectively. From the numbers, we can see that sepsis samples are much less than normal samples. To deal with this imbalance situation, we sample the normal images in the same number as the sepsis images and make sure the model performance will not be influenced by the significant difference in the numbers of both classes. Thirty percent of both normal and sepsis images were extracted from the validation set before training. Fig. 5.21 shows the learning curve under this setting.

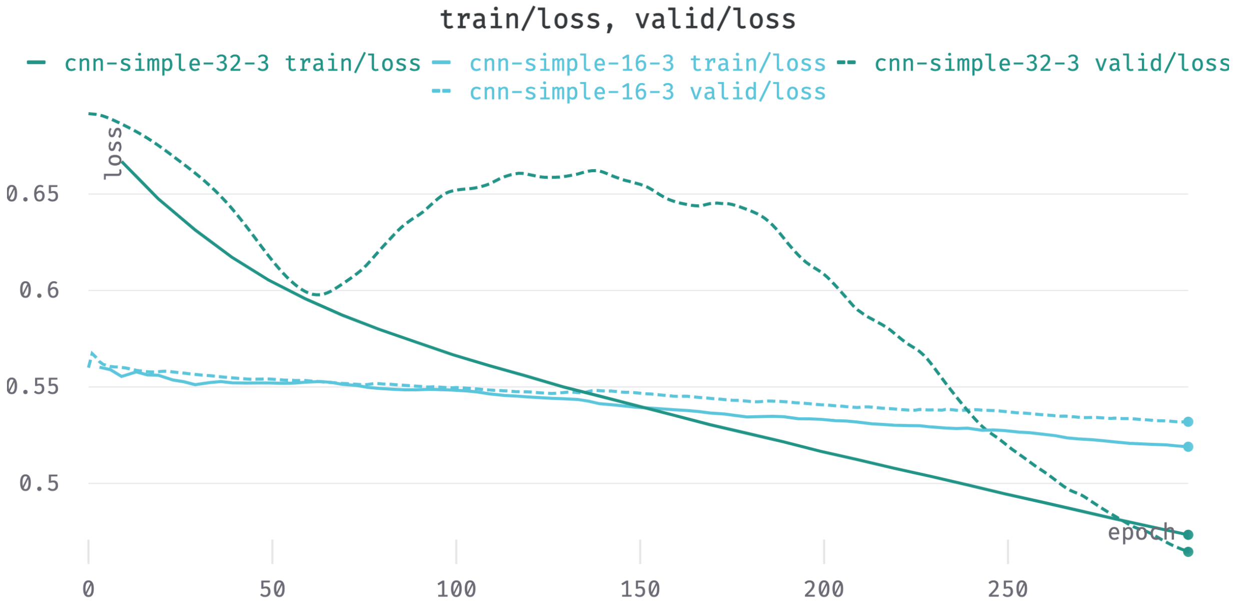


Figure 5.20: Training curve of 1st and 2nd settings on NICU dataset.

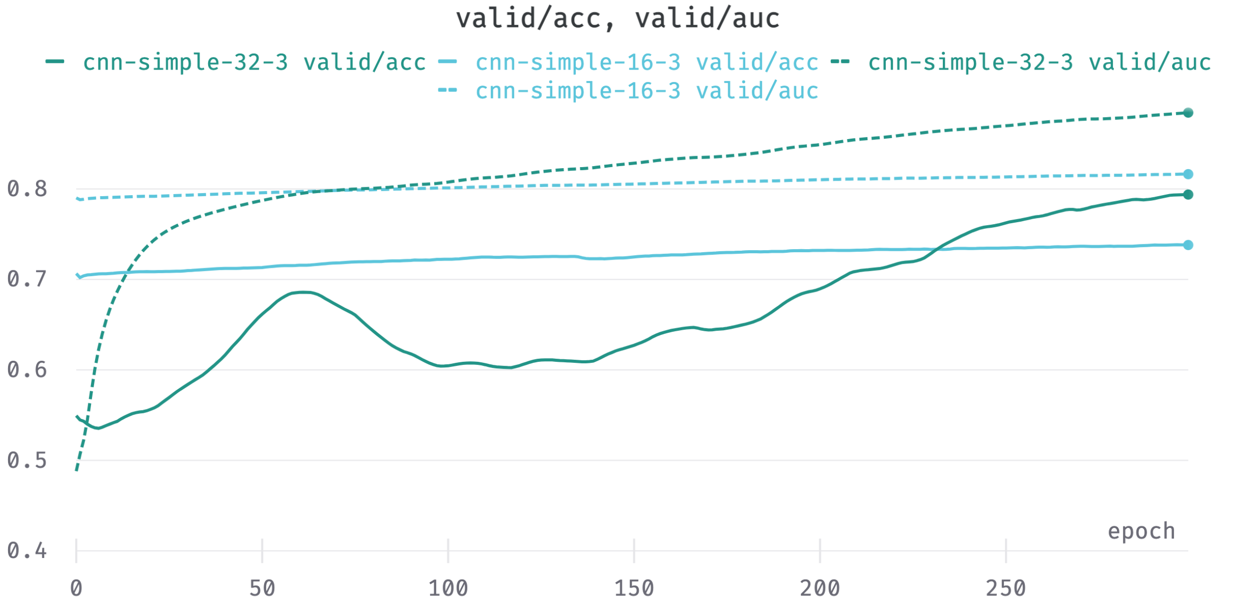


Figure 5.21: Accuracy and AUC of proposed model with setting 1 & 2.

As shown in the training curve, both experiments were converging and no overfitting was observed before epoch 300. Although the validation loss of setting 2 fluctuated during the course, it eventually reached the lowest point after all. To verify how image size would affect the model's performance, we found that larger size (32) images performed better than small ones. Other performance measures in Table. 5.8 could also prove that. When image size was scaled twice as the original, it actually contains four times more data points which means more information and features could be used to identify the imminent sepsis onset. Note that we adjusted sample rate of the raw data from one record per second to one record per minute before transforming to images because all the variables we monitored would not have a significant change within a few seconds. If we keep the data sample rate at once per second, all the data points within one image will be very likely the same.

As a comparison, GASF transformed images were tested as well in setting 3 and 4 with CNN models of the same structure. The only difference is the size of the transformed images. In GASF scheme, an image of size r×r could be generated from only r continuous time series data, so given the same dataset, it can generate more images than the simple transformation method, or images with larger size when keeping the amount at the same level. Besides, Since AlexNet has an outstanding performance in image classification tasks, we also tried to fine-tune a pre-trained AlexNet [193]. AlexNet is a deep convolutional neural network model proposed in the ILSVRC10 competition, and it won the first place that year. In our case, features learned by AlexNet were kept, but we replaced the last fully connected layer and retrained it to fit into our data. However, AlexNet requires the input images to be of size 227×227 and that is the reason we did not apply it on simple transformation images sets since we have only a few of patients who has more than 51529 (227 × 227) consecutive records. Also, AlexNet is designed to accept 3-channel input, thus Principle Component Analysis (PCA) was applied to reduce the dimension of input to three. There are many other successful models that have won the image classification competitions and have extraordinary performance, but compared to AlexNet, they have a much more complex structure. The training process of pre-trained AlexNet is shown in Fig. 5.22 and 5.23. From these figures we can clearly see that overfitting happens at around step 300 when validation loss started to increase while both acc and auc dropped from the peak.

To verify the feasibility of predicting sepsis with a machine learning approach, we applied conventional ML models like SVM and Random Forest (RF) in the previous section. Raw data is fed into five different models without any transformation in format. The result is shown in Table. 5.7. Two ensemble learning models, RF and Gradient Boosting Decision Tree (GBDT), have better results than others with an AUC of 0.77 and 0.64 respectively and there is no big difference regarding accuracy and AUC among the three models, both accuracy and AUC are about 0.50. Compared to previous models, our CNN-based model has a

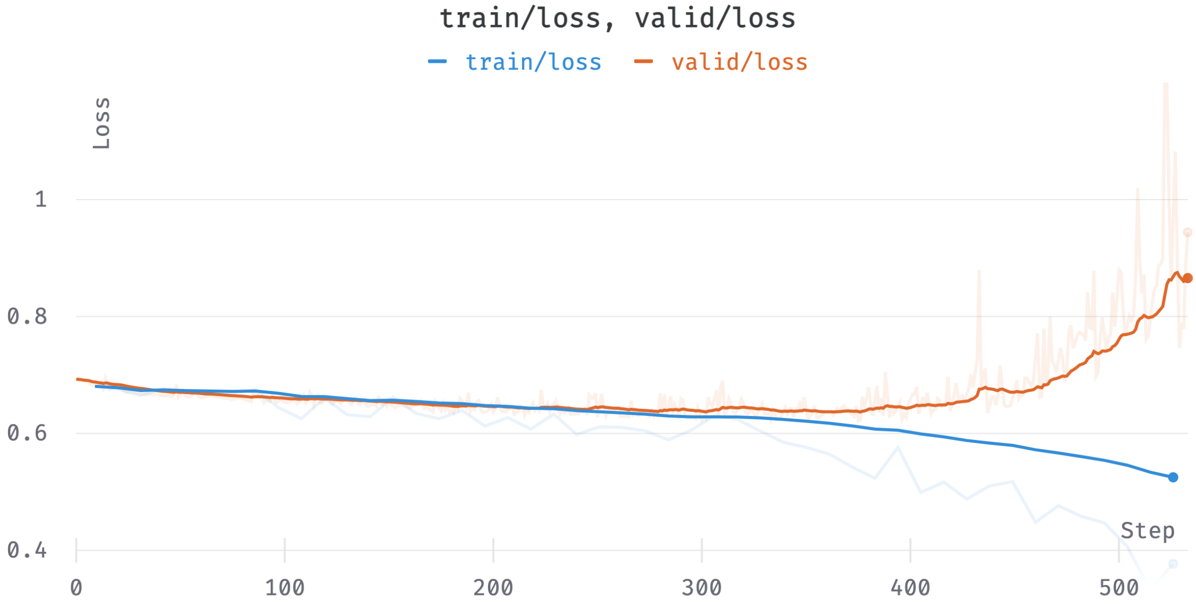


Figure 5.22: Training curve of the pre-trained AlexNet on NICU dataset.

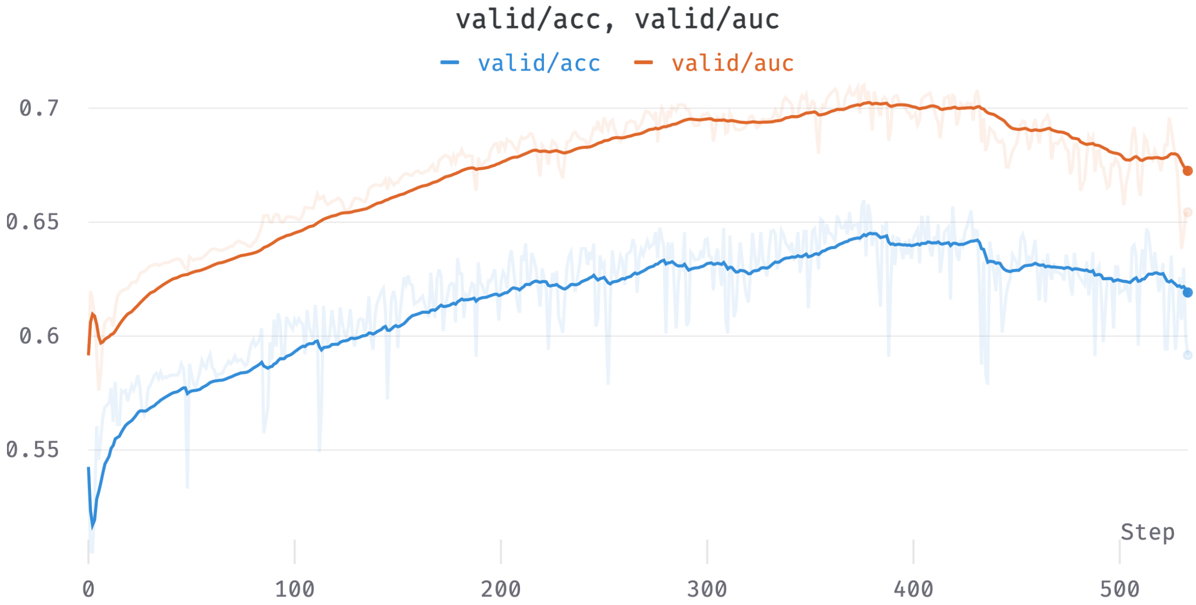


Figure 5.23: Accuracy and AUC of the pre-trained AlexNet on NICU dataset.

Table 5.8: Performance comparison between five settings. Abbreviations: CNN - Convolutional Neural Network, GASF - Gramian Angular Summation Field

No.	Setting	Accuracy	AUC
1	16×16 + CNN	0.74	0.82
2	32×32 + CNN	0.80	0.89
3	32×32 + GASF	0.54	0.64
4	128×128 + GASF	0.61	0.66
5	227×227 + AlexNet	0.59	0.65

better performance, and stands out in comparison to others' work as well. Griffin et al. [101] studied the relation between heart rate characteristics and neonatal sepsis, the AUC of their method was 0.77 originally, after two years another paper [104] from them improved this to 0.82. Two scoring systems [194, 195] managed to achieve an AUC of 0.85 and a sensitivity 0.83 (only 0.32 specificity) respectively, but due to the nature of rule based methods, rules of scoring varies from one cohort to another, and will be subjectively influenced by the experts who design the rules, which makes the evaluation metrics not consistent and unstable. Another machine learning method [196] proposed by Mani et al. has a 0.88 sensitivity and a 0.78 AUC, but a specificity of only 0.36. Some of previous models can outperform ours because our study population contains infants with only clinicians suspicion but not blood culture test ordered, and it is more difficult to predict definite and suspicious sepsis than definite sepsis alone. A comparison of the results of six models including ours is provided in Table. 5.9.

Table 5.9: Best performance of six models

Model author	Sensitivity	Specificity	AUC	Precision
Griffin 2003[101]	N/A	N/A	0.77	N/A
Griffin 2005[104]	N/A	N/A	0.82	N/A
Okascharoe 2005[194]	N/A	N/A	0.85	N/A
Singh 2003[195]	0.83	0.32	N/A	0.65
Mani 2014[196]	0.88	0.36	0.78	N/A
Proposed Model	0.88	0.84	0.89	0.74

Due to the nature of the CNN structure, the bigger the image size is, the more information it contains, but limited by the size of clinical data we collected, we cannot ensure an adequate number of images if we set the image size too big. Since the bedside monitors are not fully automated with intelligence, they are operated by nurses on duty most of the time, so it is inevitable to bring extra noise due to human errors. Some of these noise including missing, scrambled, or even incorrect data are impossible to eliminate because not every detailed record is maintained, making the prediction much more difficult. The other potential issue may be the transformation procedure from data chunks to images. The transformation proposed in this paper is a direct map from the HR, RR and SpO2 to RGB channels, because the differences among pixels are not as significant as in an actual image, CNN may not be able to extract the appropriate features.

The proposed GASF scheme did not achieve a promising result as expected, which is believed due to the limitation of data size incorporated by the algorithm. Even use the same size of 32×32, the GASF-based model was less effective because the images were generated from only 32 data points compared to 1024 data points in simple transformation case, hence were less informative.

Furthermore, the synthetic images do not contain common patterns that widely exist in normal images which may explain why pre-trained AlexNet had an excellent performance in the ImageNet competition but not as good in our scenario. However, our work has provided a new idea of transforming the time series data into 2-D images, and this idea of applying CNN in the task of predicting sepsis and the transformation of data into 2-D images are still a promising way to research on.

5.5 Reflection

In this section, we run through the entire workflow from data collection to developing the predictive model, targeting neonatal sepsis. During the process, some problems emerged nevertheless. The data collection program - eDataGrabber - was the first issue preventing us from accumulating large amounts of clinical data from the hospital. It was designed to capture data from a single patient within a short period of time, hence not very effective when dealing with all the patients from 32 cots in the entire NICU 24/7. Currently, there is no way to improve it unless Dräger updates the program with a more versatile mechanism. Almost every step in the workflow requires manual operations despite the eDataGrabber software. Its limited functions also add extra complexity to the workflow, e.g. every thread can only connect to one cot so multiple threads must be maintained simultaneously. The imperfection of the tools and workflow leads to an outcome that the quality and quantity of the collected NICU data is not as good as we expected, hence it became necessary to introduce datasets from other sources.

When pre-processing the raw data, we noticed that like most clinical data, it is extremely imbalanced with respect to the number of sepsis and normal records/patients. Without proper handling, the overwhelming normal cases would trick the models into making negative predictions in order to achieve the highest accuracy. We simply downsampled the normal records to shrink their size to the same level as sepsis records' to achieve the rebalance. Upsampling is another option to handle the imbalance problem, but not in our case since the most common SMOTE-based method is only suitable for time-independent data while our data is time series. However, downsampling brought the issue of significant loss of data, especially when the total amount is insufficient. In fact, after downsampling, only 2.5% of the entire dataset was used while leaving the rest 97.5% wasted. We believe that is one of the reasons that our models do not have outstanding performance.

We also found that the available features are much less than we expected, which makes all of them valuable, considering the numbers of monitored vital signs. It is necessary to not discard any of them when they were analysed during the feature selection step to keep our training samples with sufficient dimensions.

Finally, our proposed GASF-based transformation from time series to images did not achieve satisfactory performance in practice. Experiment results showed that it did not outperform the best baseline model, i.e. Random Forest. We tried to add dilation to the filters so they can capture patterns across longer intervals and concatenate multiple filters with different sizes of dilation, but it did not contribute much to the improvement. We assume that it was attributed to the data used to generate images. Even the generated images have the same size as those transformed by simple transformation methods, the GASF-based method used much fewer data from the raw dataset, hence less information in the images. Other CNN methodologies could be tried in the future to explore the potential, and hopefully exploiting the preserved time relations in the GASF generated image can lead to a more robust model and a better performance.

5.6 Chapter Summary

In this chapter, we implemented the action taking, evaluating and specifying learning steps in the Action Research framework to predict neonatal sepsis. Specifically, we divided the implementation into three cycles. During the first cycle, datasets were collected and prepared. In the second cycle, a more detailed analysis and feature selection were performed. To clarify the priority of each vital sign during their diagnosis procedures, we consulted clinicians to clarify the risk factors identified for sepsis, followed by three other methods. Fundamentally, we analysed the relationship between the features in order to retain the important ones and to eliminate the redundant ones. Lastly, in cycle 3, we developed a set of classical machine learning models as baselines and trained a CNN to perform the prediction task after we transformed the time series medical records into 2D images using two different methods. The outcomes were summarised and compared as the evaluation and reflection specified in the Action Research framework.

We have answered RQ-4 by checking a series of classic machine learning models, and answered RQ-5 by developing a model based on converting time series data to images. Additionally, features selection step identified the features that are most related to the neonatal sepsis prediction task which answered RQ-2. Similar experiments will be done in the next chapter on adult patients datasets.

Chapter 6

Framework Implementation - General Sepsis Prediction

In the absence of a comprehensive dataset from Monash Children's Hospital, we explored two public datasets available on the Internet, and widened the scope of the study from neonatal sepsis to general sepsis, including sepsis occurring in adults. In this chapter, we will elaborate on our work to predict sepsis using public datasets. We did not embed ourselves in the hospital since no data collection procedure was required, and Action Research was therefore not strictly applicable in this instance. The AR framework was, however, still followed in other areas such as the iterative phases of planning, evaluation and reflection. The experiment environment was the same as in Chapter 5.

6.1 Cycle One

6.1.1 Data Sources

To enrich our research datasets, we also collect sepsis-related data from two publicly available sources: one is the PhysioNet Challenge 2019 whose topic happened to be "Early Prediction of Sepsis from Clinical Data", another one is the very popular MIMIC III database among the research community. Compared to collecting real-time onsite from NICU in the hospital, it is much easier to download the public datasets. For the PhysioNet dataset, all the variables of each patient have been packed into a CSV file, we simply downloaded more than 40,000 files provided by the competition host. However, to obtain access to MIMIC III dataset, we have to complete a CITI Data or Specimens Only Research online course to make sure that we know how to deal with the patient data properly. As stated in previous chapters, the MIMIC III datasets consist of several tables, and we have to select the variables we were interested in by joint query.

6.1.2 Reflection

The purpose of introducing public datasets is to extend our research from neonatal sepsis to general sepsis which is broader by definition. Meanwhile, the extended work could supplement the work in the previous chapter, providing an opportunity to verify our models and frameworks in various use cases. The open-access datasets have sufficient sizes to train a more complex model, and they were well formatted. The issue of two public datasets we have possessed is the difference of available attributes in each of them, so feature synchronisation must be done.

6.2 Cycle Two

In this cycle the datasets were analysed just like Cycle 2 of neonatal sepsis case.

6.2.1 Domain Knowledge

Before the technical methods applied, we resorted to clinicians' expertise. Unlike NICU dataset, public datasets provide many variables. From clinicians' experience, some attributes have higher priorities to be examined when suspicious symptoms were found. Furthermore, the literature showed that some variables could be used as a key to indicate infection.

C-reactive protein (CRP) may be one of the most extensively studied and most frequently used parameters for the diagnosis of neonatal sepsis [197]. Despite known associations with other non-infectious complications, a CRP level of >10 mg/L does appear to be a highly accurate marker for infection with specificity values consistently >90% being reported. CRP is suitable for both EOS and LOS, and the sensitivity improves with serial measurements [198, 199]. The advantage in diagnosis accuracy is achieved when serial CRP are evaluated in conjunction with other parameters like haematological indices, cytokines, and cell surface markers [200, 201]. Note that preterm infants have lower CRP baseline values and a lower rise in response to infection, and some other non-infectious conditions may cause an elevation of CRP levels as well [21]. Because it takes 10-12 hours to change significantly after the onset of infection, the sensitivity of CRP is low during the early phase of sepsis. The specificity and positive predictive value of CRP ranges from 93%-100% [202]. Thus, CRP can be considered as a "specific" but "late" marker of neonatal infection. Due to non-infectious CRP elevations, the influence of gestational age and birth weight, and the lack of reliable age-specific reference values, the use of CRP requires further research to evolve as an ideal marker. Another test that almost every sepsis-like patient will do is the complete blood count (CBC) test, also called the Full Blood Examination (FBE). A large number of studies have been conducted to evaluate the use of CBC for the diagnosis of neonatal sepsis.

Low white blood counts (WBC) and absolute neutrophil counts (ANC), as well as a high immature-to-total neutrophil ratio (IT ratio) are associated with a high risk of getting infected [16]. In a large retrospective research, culture-proven EOS was proved to be related to low WBC and ANC, and the authors believed that it is worth postponing the antibiotic therapy until the result of CBC test come out [203]. In the context of LOS, another study showed the connection between WBC, ANC, IT ratio plus absolute band counts (ABC) and infection [16]. However, haematological parameters may perform well in the diagnosis of neonatal sepsis though, the clinical utility would probably lie in their combination with other biochemical markers like neutrophil CD64 (nCD64) [204]. More recently, advanced white blood cell indices such as mean neutrophil volume (MNV), mean monocyte volume (MMV), conductivity (MNC; MMC), scattering (MNS) (MMS), and distribution width (NDW; MDW) are emerging as possible additional markers of NS, and they may be useful in the differential diagnosis of neonatal sepsis [205], but both sensitivity and specificity were noted lower compared with other parameters such as CRP indicating these markers could only be the adjunctive measures. On the other side, research showed that haematological components could also be effective for identifying healthy newborns rather than distinguishing infected ones [24].

A lot of work has shown that Heart rate variability (HRV) could be a potential method of early detection of infants with neonatal sepsis and necrotising enterocolitis (NEC) prior to the onset of the systemic inflammatory response [99–102, 206–210]. Except for HRV, measurement of core-peripheral temperature difference showed promise as a predictor of LOS. Research demonstrated that a temperature difference 2.3 led to an overall accuracy of 90.9% in the identification of LOS, and a specificity of 100% if the difference reached $\geq$3.2°C [211]. Another recent study has replicated these findings using continuous monitoring of axillary and sole temperature differences and verified the feasibility of using temperature difference as another early warning signal of NS [212].

6.2.2 Feature Selection

Variance-based selection

Compared to the NICU dataset, it is another story when it comes to the PhysioNet dataset, see Fig. 6.1. We can see three features with extremely high variance close to 0.25, among which unit1 and unit2 are just administrative identifiers for MICU and SICU units,

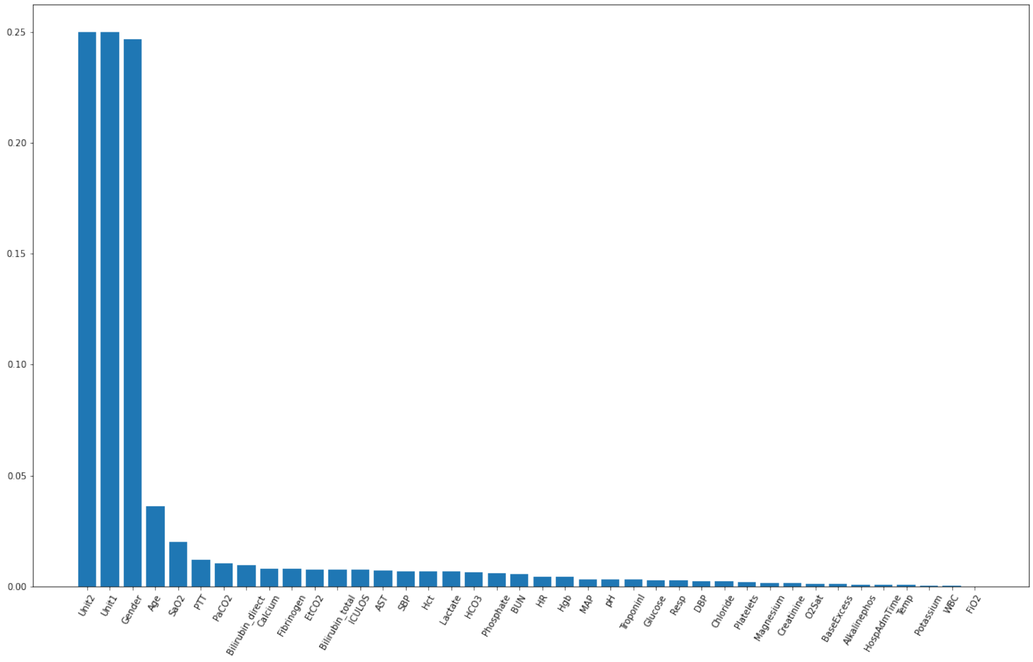


Figure 6.1: The variances of features in PhysioNet Challenge datasets.

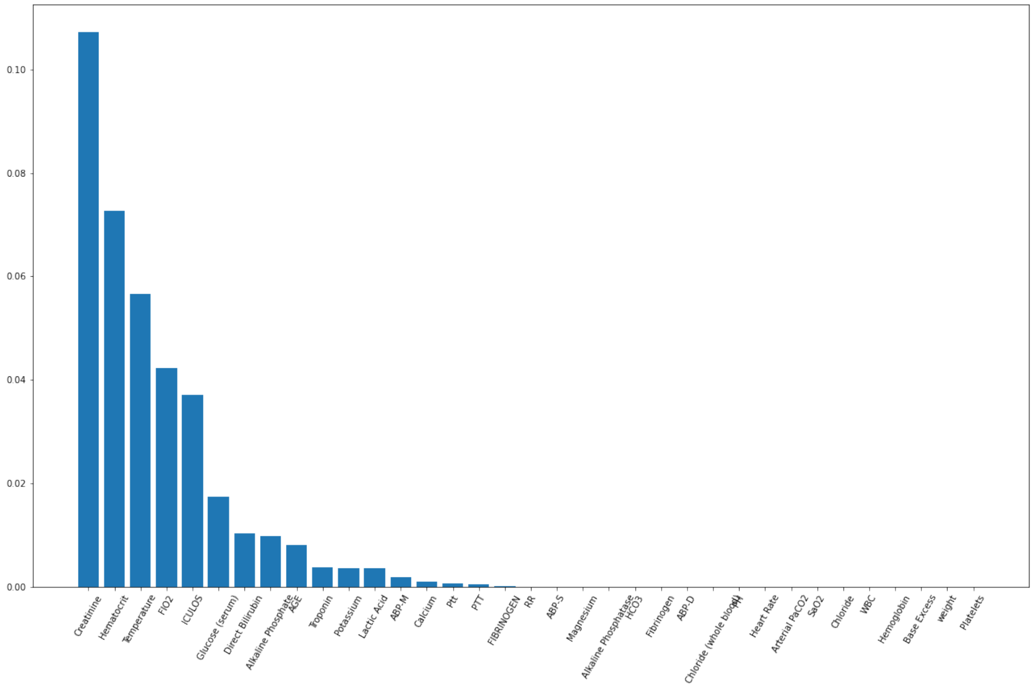


Figure 6.2: The variances of features in MIMIC III Challenge datasets.

they do not hold any information related to sepsis itself, and gender is roughly 50/50 distributed which we can infer from the value of 0.25. We listed the variances of the rest features in Table 6.1 for further analysis. Most values are between 0.002 and 0.01, so any feature below 0.002 was removed. In Fig. 6.2 we can clearly see that in MIMIC III dataset, approximate half of the feature set have a very low variance near zero. The same threshold as in PhysioNet Dataset was also applicable here to remove features with extremely small variance.

Table 6.1: Variances of features in PhysioNet datasets apart from the highest three. Abbreviations: refer to Table. 2.4

Feature	Variance	Feature	Variance
HR	0.0044	Creatinine	0.0015
O2Sat	0.0013	Bilirubin_direct	0.0097
Temp	0.0007	Glucose	0.0028
SBP	0.0069	Lactate	0.0067
MAP	0.0034	Magnesium	0.0017
DBP	0.0025	Phosphate	0.0059
Resp	0.0027	Potassium	0.0006
EtCO2	0.0078	Bilirubin_total	0.0076
BaseExcess	0.0011	TroponinI	0.0032
HCO3	0.0063	Hct	0.0069
FiO2	0.0	Hgb	0.0044
pH	0.0032	PTT	0.0122
PaCO2	0.0106	WBC	0.0003
SaO2	0.02	Fibrinogen	0.0079
AST	0.0074	Platelets	0.002
BUN	0.0056	Age	0.0363
Alkalinephos	0.001	HospAdmTime	0.0009
Calcium	0.0082	ICULOS	0.0075
Chloride	0.0024		

Correlation Analysis

We did Pearson correlation and distance correlation tests in two public datasets, the situation is quite similar to NICU dataset, see Table. 6.3 to 6.5. In the PhysioNet dataset, features are linearly independent according to the PCCs and p-values. Only one pair stands out, with a relatively high PCC of -0.2656, indicating a slight linear relation between diastolic arterial blood pressure and age. When examining the p-value we found that SaO2 has a 0.5155 p-value against ICU length of stay so the null hypothesis was supported. Large p-values also occurred in the features of respiratory rate and SaO2 in MIMIC III dataset.

The DCC showed that most pairs of features are not very closely related, just like what PCC values indicated. However, we discovered something not covered by PCC that those

Table 6.2: Pearson correlation coefficient of each feature in PhysioNet Challenge dataset. Abbreviations: refer to Table. 2.4

	Age	ICULOS	HR	Resp	SaO2	SBP	DBP
Age	1.0	0.0105	-0.1573	0.0326	-0.0949	0.0245	-0.2656
ICULOS	0.0105	1.0	0.0454	0.098	-0.0028	0.0563	0.0125
HR	-0.1573	0.0454	1.0	0.226	-0.0107	-0.0337	0.1294
Resp	0.0326	0.098	0.226	1.0	-0.0236	0.0459	0.0619
SaO2	-0.0949	-0.0028	-0.0107	-0.0236	1.0	0.106	0.0647
SBP	0.0245	0.0563	-0.0337	0.0459	0.106	1.0	0.5398
DBP	-0.2656	0.0125	0.1294	0.0619	0.0647	0.5398	1.0

Table 6.3: p-value of each feature in PhysioNet Challenge dataset. Abbreviations: refer to Table. 2.4

	Age	ICULOS	HR	Resp	SaO2	SBP	DBP
Age	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ICULOS	0.0	0.0	0.0	0.0	0.5155	0.0	0.0
HR	0.0	0.0	0.0	0.0	0.0156	0.0	0.0
Resp	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SaO2	0.0	0.5155	0.0156	0.0	0.0	0.0	0.0
SBP	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DBP	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6.4: Pearson correlation coefficient of each feature in MIMIC III dataset. Abbreviations: refer to Table. 2.4

	AGE	ICULOS	Heart Rate	RR	SaO2	ABP-S	ABP-D	ABP-M
AGE	1.0	-0.3184	-0.0073	-0.0003	0.0005	-0.0074	-0.0209	-0.0326
ICULOS	-0.3184	1.0	0.0031	0.0005	-0.0	-0.0024	-0.0013	-0.0021
Heart Rate	-0.0073	0.0031	1.0	0.1021	-0.0002	-0.0192	0.0148	0.0196
RR	-0.0003	0.0005	0.1021	1.0	0.0002	0.0167	0.0048	0.0121
SaO2	0.0005	-0.0	-0.0002	0.0002	1.0	0.0015	0.0002	0.0031
ABP-S	-0.0074	-0.0024	-0.0192	0.0167	0.0015	1.0	0.0136	0.0528
ABP-D	-0.0209	-0.0013	0.0148	0.0048	0.0002	0.0136	1.0	0.0135
ABP-M	-0.0326	-0.0021	0.0196	0.0121	0.0031	0.0528	0.0135	1.0

Table 6.5: p-value of PCC of each feature in MIMIC III datasets. Abbreviations: refer to Table. 2.4

	AGE	ICULOS	Heart Rate	RR	SaO2	ABP-S	ABP-D	ABP-M
AGE	0.0	0.0	0.0	0.5041	0.2988	0.0	0.0	0.0
ICULOS	0.0	0.0	0.0	0.226	0.9702	0.0	0.0245	0.0002
Heart Rate	0.0	0.0	0.0	0.0	0.6709	0.0	0.0	0.0
RR	0.5041	0.226	0.0	0.0	0.7206	0.0	0.0	0.0
SaO2	0.2988	0.9702	0.6709	0.7206	0.0	0.1171	0.8305	0.0014
ABP-S	0.0	0.0	0.0	0.0	0.1171	0.0	0.0	0.0
ABP-D	0.0	0.0245	0.0	0.0	0.8305	0.0	0.0	0.0
ABP-M	0.0	0.0002	0.0	0.0	0.0014	0.0	0.0	0.0

have biologically connected naturally, such as the heart rate and pulse, systolic, diastolic and mean blood pressure, have mild correlation measured by a DCC of about 0.5.

Table 6.6: Distance correlation coefficient of feature pairs in PhysioNet Challenge dataset. Abbreviations: refer to Table. 2.4

	Age	ICULOS	HR	Resp	SaO2	SBP	DBP
Age	1.0	0.0243	0.1472	0.0452	0.0957	0.0322	0.2706
ICULOS	0.0243	1.0	0.0455	0.0973	0.0655	0.0629	0.0256
HR	0.1472	0.0455	1.0	0.2096	0.0274	0.0343	0.1209
Resp	0.0452	0.0973	0.2096	1.0	0.0665	0.0493	0.0628
SaO2	0.0957	0.0655	0.0274	0.0665	1.0	0.1	0.065
SBP	0.0322	0.0629	0.0343	0.0493	0.1	1.0	0.5082
DBP	0.2706	0.0256	0.1209	0.0628	0.065	0.5082	1.0

Table 6.7: Distance correlation coefficient of feature pairs in MIMIC III dataset. Abbreviations: refer to Table. 2.4

	AGE	ICULOS	Heart Rate	RR	SaO2	ABP-S	ABP-D	ABP-M
AGE	1.0	0.3261	0.7924	0.0425	0.0167	0.0228	0.2959	0.1564
ICULOS	0.3261	1.0	0.3739	0.1513	0.0226	0.0489	0.0368	0.0351
Heart Rate	0.7924	0.3739	1.0	0.2425	0.0505	0.0668	0.1967	0.0887
RR	0.0425	0.1513	0.2425	1.0	0.06	0.0573	0.0556	0.059
SaO2	0.0167	0.0226	0.0505	0.06	1.0	0.0347	0.0201	0.0379
ABP-S	0.0228	0.0489	0.0668	0.0573	0.0347	1.0	0.5401	0.7856
ABP-D	0.2959	0.0368	0.1967	0.0556	0.0201	0.5401	1.0	0.8712
ABP-M	0.1564	0.0351	0.0887	0.059	0.0379	0.7856	0.8712	1.0

Tree-based methods

Fig. 6.3 and 6.4 demonstrated the feature importance in PhysioNet Challenge dataset and MIMIC III dataset. Both datasets have a similar distribution in the diagrams. Although the specific order might be slightly different, we could still identify that the most important ones are age, length of stay, heart rate and blood saturation in both datasets. Therefore during the training phase in the next cycle we should focus more on those features with higher importance and even ignore some of the features having extremely small importance.

6.2.3 Reflection

In cycle two, we performed some analysis on the features to figure out the effectiveness of each of them. Advice from professionals were referred, accompanied by three types of feature selection methods during the process. Unfortunately, measurements that clinicians are most concerned with are mostly lab test results like C-reactive protein, blood count, or

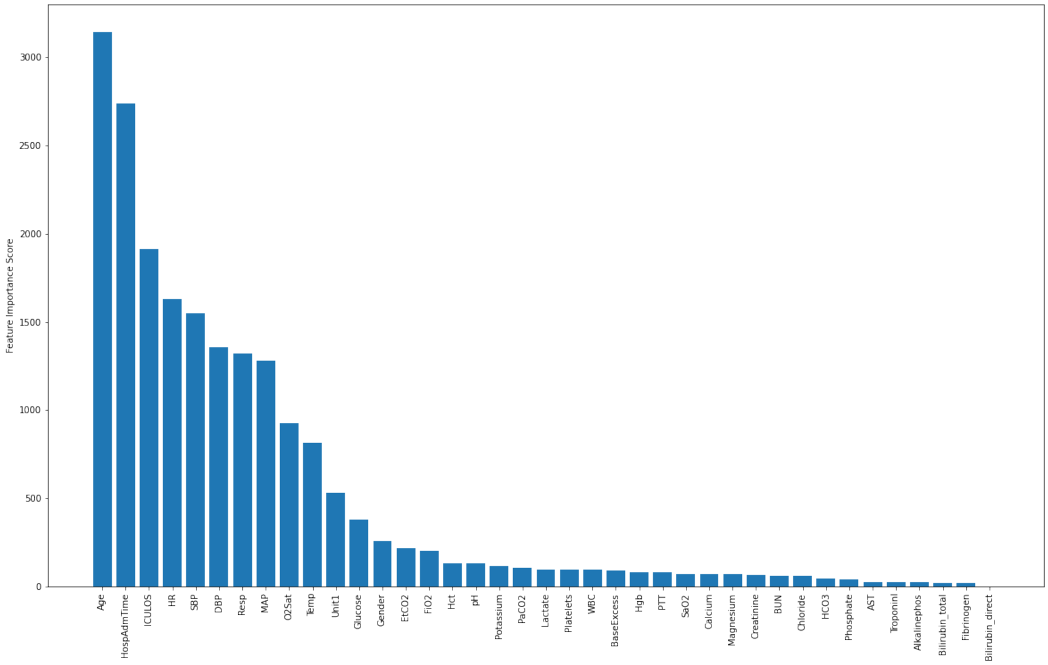


Figure 6.3: Features importance of features in PhysioNet Challenge dataset.

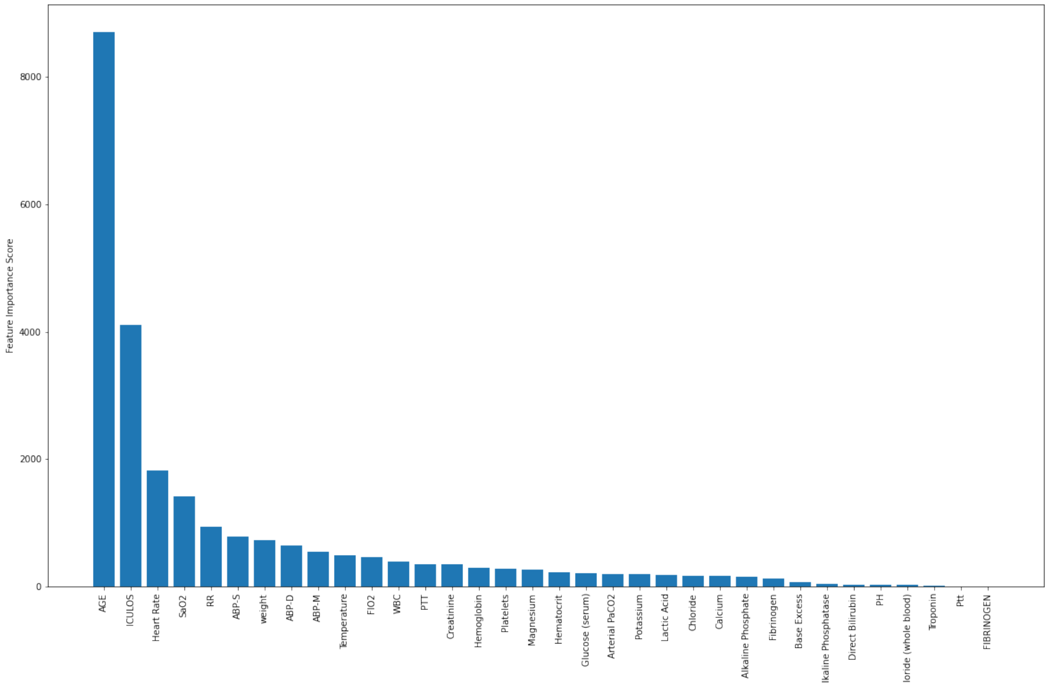


Figure 6.4: Features importance of features in MIMIC III dataset.

blood culture test, but they have to wait until the results come out which is not the preferable case. With the help of three technical approaches we can examine other measurements like demographic characteristics and vital signs which are easier to get but not paid as much attention as those lab results.

Setting a threshold of variance for each feature and removing those below it is a simple way to do a primitive check on the effectiveness. It is not the most accurate method to distinguish unrelated features, but at least we can have a big picture on them and remove those with zero variance or extremely small variance since they basically do not change for both sepsis and normal cases, which means not as discriminative as others. Then we checked the Pearson Correlation Coefficient - a metric that measures linear relations between variables. The result showed that there was no significant linear correlation among chosen features. Note that some of the PCC values like the one between SaO2 and ICULOS in the PhysioNet Challenge dataset or between AGE and RR in MIMIC III are approximately zero which indicates no linear correlation, but their corresponding calculated probabilities p are as high as 0.5, so their PCCs are not supposed to be accepted. Pearson Correlation Coefficient has such a limitation that in some cases even the PCC value is zero we can not determine that there is no linear correlation there. To supplement PCC, we also examined Distance Correlation Coefficient for corroboration. DCC works based on the distance covariance which is not limited to a linear relation. As a result, numbers proved that no significant relations exist among features except for the set of blood pressures (ABP-S, ABP-D and ABP-M) that inherently connected to each other. In addition to various correlation, a typical embedded feature selection method, the tree-based method, was employed. Figures have clearly depicted the importance of each feature, and we found that based on xgboost two datasets have a similar distribution in terms of features importance and same features of top importance.

The purpose of this cycle is to analyse the features in the datasets and select informative and discriminative ones while removing redundant features. By multiple analysis methods, we kept eight and seven features in MIMIMC dataset and PhysioNet dataset respectively due to the fact that mean arterial blood pressure is not available in the latter one, see Table 6.6 and 6.7.

6.3 Cycle Three

6.3.1 Deep Learning 2 - LSTM

We proposed the second deep learning method which is to combine CNN and LSTM together to build a hybrid model in this section. This model is established based on the encoder-decoder structure. Since our training samples (medical records from patients) could be treated as multi-variate time series data, LSTM is a suitable choice to try as the decoder due to its great power of dealing with temporal information and variant length input data. As to the encoder, 1d-Conv layer as well as fully-connected layer could be used as encoders to extract useful features from multi-variate records.

Preprocessing

Data filtering In order to improve the performance of our model, before training some preprocessing steps are necessary. First, to mitigate the impact caused by imbalanced data, we filtered the raw dataset. Since patient records are sequential, consecutive records from one patient should be treated as one data entry. In other words, the data filtering should be done patient-wise rather than row-wise. To re-balance data patient-wise, we down-sampled the patients with high sepsis-healthy records ratio, so after trial and error we set the threshold to 0 which means anyone who has even only one single data entry labelled as sepsis will be preserved, and for those totally healthy patients, their records are discarded. Setting a higher threshold will cause a dramatic drop in the number of data samples due to the imbalanced nature of the given dataset. To ensure data of both healthy and sepsis patients are trained, we add extra 500 healthy patients into the filtered dataset.

Imputation The next step of preprocessing is to impute the missing data. As mentioned, nearly 70% of the total data is not available. If normal routines are followed, most features will be dropped which will shrink the data notably, so we apply multiple strategies to fill the missing value. For continuous data like vital signs, we applied multiple interpolation approaches: linear interpolation was used for features that have at least two values, the nearest values were filled in the blanks if there is only one valid data among all records, and finally, if the feature has no valid value, the corresponding mean of all the records was calculated and placed. On the other hand, for sparse data that have a high absence ratio, lab test results, for instance, we converted them into categorical data based on the normal range.

Table. 6.8 shows the normal ranges of all the lab test indexes. Category 0 was assigned in the case of no valid data provided, and categories 1,2 and 3 were assigned when the value is lower, higher or just within the normal range, respectively.

Table 6.8: Normal ranges for lab test results.

Name	Range	Unit
Alkalinephos [213]	44 - 147	IU/L
BaseExcess [214]	-4 - +2	mmol/L
Bilirubin_direct [215]	0.0 - 0.3	mg/dL
Bilirubin_total [215]	0.3 - 1.2	mg/dL
Calcium [216]	2.2 - 2.6	mg/dL
Chloride [217]	96 - 106	mmol/L
Creatinine [216]	0.6 - 1.3	mg/dL
Fibrinogen [218]	200 - 400	mg/dL
Glucose [219]	72 - 135	mg/dL
HCO3 [220]	23 - 30	mmol/L
Hct [221]	35.5 - 44.9	%
Hgb [222]	12.0 - 17.5	g/dL
Lactate [223]	4.5 - 19.8	mg/dL
Magnesium [224]	0.85 - 1.10	mmol/L
p/H [214]	7.35 - 7.45	N/A
PaCO2 [214]	35 - 45	mmHg
Phosphate [225]	2.8 - 4.5	mg/dL
Platelets [226]	150 - 400	count*$10^3/\mu L$
Potassium [227]	3.6 - 5.2	mmol/L
PTT [228]	60 - 70	seconds
SaO2 [214]	95 - 100	%
Troponin [229]	0.00 - 0.04	ng/mL
WBC [230]	4.5 - 11	count*$10^3/\mu L$

1-d Convolutional Neural Network

Convolutional Neural Network was first created in 1989 in [231]. Due to its outstanding performance, it has drawn more and more attention and kept being used and improved in past decades. The key concept in a CNN is the convolution operation between a kernel (filter) and input data. One kernel will slide throughout the input data while doing convolution and this kind of operation has two advantages: first, unlike normal neural networks, weights are shared in CNN which makes it much easier to compute; second, the moving kernel facilitates the extraction of features in the local area.

Normally, when applied to image processing task, kernels of a CNN is 2d, and it slides from the top-left all the way down to the bottom-right. Since the patient record is just a row of 40 numerical values at a certain time point, we squeeze the kernel to 1d and slide it from the start to the end during convolution operation. Suppose the 1d kernel is $\boldsymbol{f} = \{f_1, f_2, ...f_m\}$

where m is the length of the kernel, input data $\boldsymbol{d} = \{d_1, d_2, ..., d_n\}$, the convolutional operation will produce the output as:

$$O_i = \varphi(\sum_{j=1}^{m} f_j d_{i+j-1} + b_i) \tag{6.1}$$

where $\varphi(.)$ is one of the nonlinear activation function like ReLu or Sigmoid function. Typically, a CNN has k kernels, each of which represents one type of feature extracted by the CNN. k kernels will produce k channels of outputs, and in this way raw features get expanded. Dropout is another technique often used in CNN. It simply drops some of the output randomly at a given ratio, and set the value to zero. By dropping some of the output values, it can effectively prevent over-fitting.

Recurrent Neural Network with LSTM units

A Recurrent Neural Network (RNN) is a neural network repeated over time. In particular, an RNN allows self-loop connections and shared parameters across different time steps. While a feedforward neural network maps an input vector into an output vector, an RNN maps a sequence into a sequence. The recurrent connections allow an RNN to memorise previous inputs and therefore capture longer dependencies. Since it was invented in the 1980s, it has become a powerful tool in time series data analysis, which makes it suitable in our scenario. In the traditional RNN model, given sequence data $\boldsymbol{x} = (x_1, x_2, ..., x_T)$, RNN updates its recurrent hidden state h_t by:

$$h_t = \varphi(\boldsymbol{W}_{ih} x_t + \boldsymbol{W}_{hh} h_{t-1} + b_h) \tag{6.2}$$

where x_t and h_t are data values and the recurrent hidden state at time step t, respectively, and $\varphi(.)$ represents the nonlinear activation function of a hidden layer, such as a sigmoid or hyperbolic tangent. t = 1 to T, W_{ih} denotes the input-hidden weight vector, W_{hh} represents the weight matrix of the hidden layer, and b_n is the hidden layer bias vector.

Long Short-Term Memory (LSTM) is short for RNN with LSTM hidden units, a modified version of RNN, to address the problem of longterm dependencies. Unlike traditional RNN which simply applies a transformation to a weighted linear sum of inputs, LSTM adds a linear self-loop memory cell which allows gradients to pass through longer sequences. The memory cell is gated to moderate the amount of information flow into or from the cell, and its activation is computed as:

$$h_t = o_t \varphi(c_t) \tag{6.3}$$

where o_t is the output gate that determines the portion of the memory cell content in time step t (c_t) to be exposed at the next time step. The recursive equation for updating o_t is:

$$o_t = \sigma(\boldsymbol{W}_{oi} x_t + \boldsymbol{W}_{oh} h_{t-1} + \boldsymbol{W}_{oc} c_{t-1} + b_o) \tag{6.4}$$

where $\sigma(.)$ is the logistic sigmoid function, $\boldsymbol{W}_{oh}$ is the hidden layer-output weight matrix, and $\boldsymbol{W}_{oc}$ is the memory-output weight matrix. The memory cell, c_t, is updated by adding new content, $\bar{c}_t$, and discarding part of the present memory:

$$c_t = \boldsymbol{i}_t \odot \bar{c}_t + \boldsymbol{f}_t \odot c_{t-1} \tag{6.5}$$

where $\odot$ is an element-wise multiplication and $\bar{c}_t$ is calculated as:

$$\bar{c}_t = \varphi(\boldsymbol{W}_{ci} x_t + \boldsymbol{W}_{ct} h_{t-1} + b_c) \tag{6.6}$$

In this equation, the **W** term represents weight matrices; e.g., $\boldsymbol{W}_{ci}$ is the input-memory weight matrix. Input gate **i**, and forget gate **f** determine the degree that new information is to be added and current information is to be removed, respectively, as follows:

$$\begin{aligned} \boldsymbol{i}_t &= \sigma(\boldsymbol{W}_{ix} x_t + \boldsymbol{W}_{ih} h_{t-1} + \boldsymbol{W}_{ic} c_{t-1} + b_i); \\ \boldsymbol{f}_t &= \sigma(\boldsymbol{W}_{fx} x_t + \boldsymbol{W}_{fh} h_{t-1} + \boldsymbol{W}_{fc} c_{t-1} + b_f) \end{aligned} \tag{6.7}$$

Proposed Structure

Our proposed model is based on the encoder-decoder architecture which has two parts: 1d-Convolutional layers as the feature extraction (encoder) and LSTM as the sequential prediction (decoder). For each row of data, two stacked 1d Convolutional layers are applied to extract features at different scales. It is implemented by adding two kernels of different sizes in a hierarchical structure - kernels in the lower layer find features in smaller granularity while kernels in the higher layer do the larger granularity features. Suppose the size of the kernel is l, the length of output the kernel could produce is $L_o = 40 - l + 1$ without padding, but to ensure the features lies at the beginning and end of a row are not neglected, we pad $(l-1)/2$ zeros at both ends so that the length of output remains the same as the input. Following each CNN layer is a dropout layer that randomly drops half of the data to prevent overfitting to some degree.

Two stacked LSTM layers are used to accept features passed from 1d-CNNs and calculate a label indicating if the sepsis onset will happen within the next six hours at each time step for any patients. Meanwhile, because of LSTM's self-loop structure, it can perfectly handle the various-length data by adjusting how many times it loops. As will be mentioned in the data analysis section, the length of records for each patient varies from one to another, making other models incapable of handling except for bringing in sliding windows. However, sliding windows will weaken the relations in the time dimension due to the isolation of time series data. The schema of the entire model is shown in Fig. 6.5.

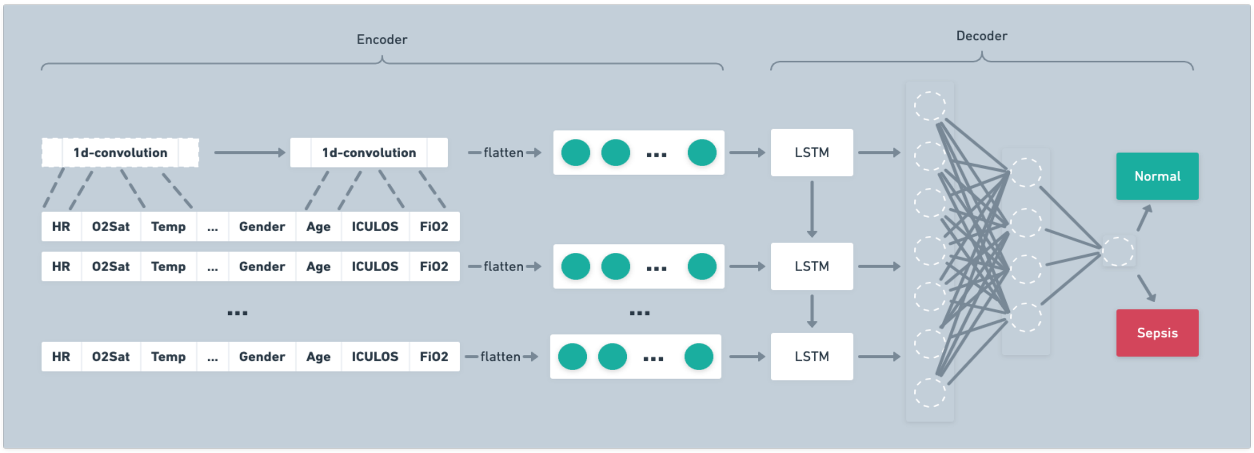


Figure 6.5: The scheme of proposed CNN+LSTM model.

Performance Evaluation

We designed a set of experiments to evaluate the performance of LSTM in our datasets, listed below:

1. A hybrid model including CNN encoder and one LSTM layer decoder.
2. A hybrid model including CNN encoder and two LSTM layers decoder.
3. A hybrid model including FC encoder and two LSTM layers decoder.
4. A CNN model on 32×32 GASF-transformed images.
5. A Simple Multilayer Perceptron Model.

For each of the above settings, we examined their accuracy and AUC and summarised them in Table. 6.9 and Figure. 6.6. From the comparison of setting 1 and 2, we found that adding an extra layer of LSTM did not significantly improve the predictive performance, as both accuracy and AUC were at the same level. When replacing the CNN encoder with a fully-connected one, both measurements got boosted to an optimal value. The better performance of FC encoder indicates it is able to extract more effective features from our data than the CNN encoder. Since the filter of CNN lies across feature dimensions, the 1d-Conv layer tends to capture patterns across neighbouring features, but in fact, the relation between features is not tight. Comparing to LSTM models, because the MLP is incapable of processing temporal information, it showed a poorer score, proving that our LSTM works well dealing with the underlying temporal information in the datasets. Also, we assume the reason GASF-based model achieved a much better performance than it did in NICU data is about the data size. The PhysioNet dataset has larger numbers of both samples and features, and with the abundant training dataset and the capability of preserving the original temporal relations in the synthetic images, GASF-based model also performed well in this experiment.

Table 6.9: Performance comparison between five settings. Abbreviations: LSTM - long short-term meomory, FC - fully connected, MLP - multi-layer perceptron, GASF - gramian angular summation field

No.	Setting	Accuracy	AUC
1	Single LSTM	0.79	0.85
2	Stacked LSTM	0.82	0.84
3	FC + LSTM	0.87	0.95
4	MLP baseline	0.52	0.53
5	GASF baseline	0.78	0.86

6.3.2 Multi-Instance Learning

Multiple instance learning (MIL) is a variation of supervised learning where data instances are aggregated randomly or in a certain way as bags, and a single class label is assigned to a bag instead of individual instances. A bag containing at least one positive instance will be labelled as sepsis, and a bag without any positive instance is labelled as normal. In our approach, we consider adjacent data samples from one patient as bags and every single data sample in it as an instance in multi-instance learning, and automatically learn a ranking model that predicts sepsis scores for sepsis patients. The idea behind the model is to sacrifice some time we can predict the onset in advance, in exchange for better classification accuracy.

To facilitate the ranking, we build a CNN to calculate a sepsis score for each instance, but for each bag, we only need to consider the instance which has the highest score. The data

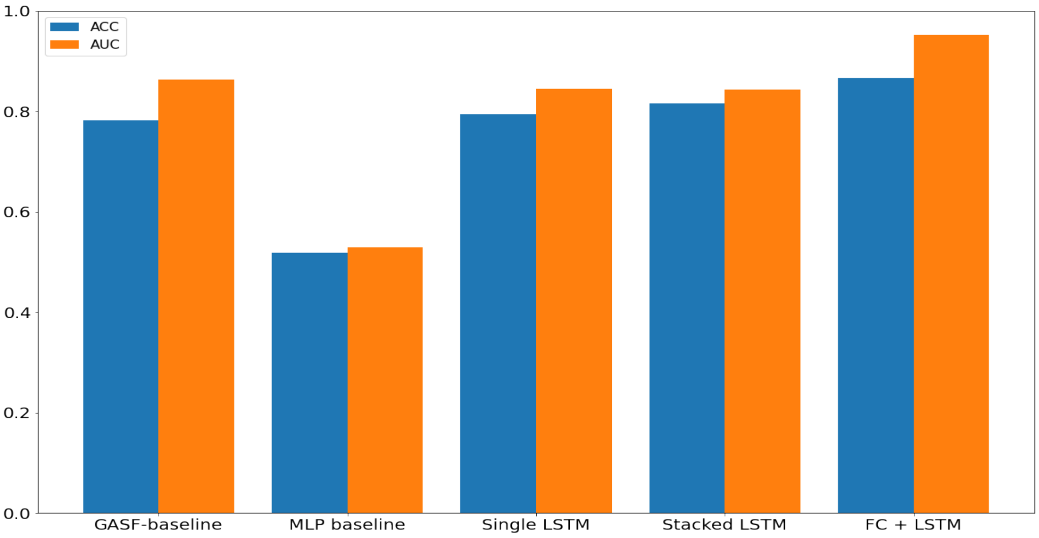


Figure 6.6: Performance comparison between five settings.

sample corresponding to the highest sepsis score in the positive bag is most likely to be the true positive, while the instance corresponding to the highest sepsis score in the negative bag is the one that looks most similar to a sepsis sample but actually is a normal instance. This negative instance is considered as a hard instance that may generate a false alarm in sepsis detection. The flow diagram of the proposed sepsis detection approach can be seen in Fig. 6.7.

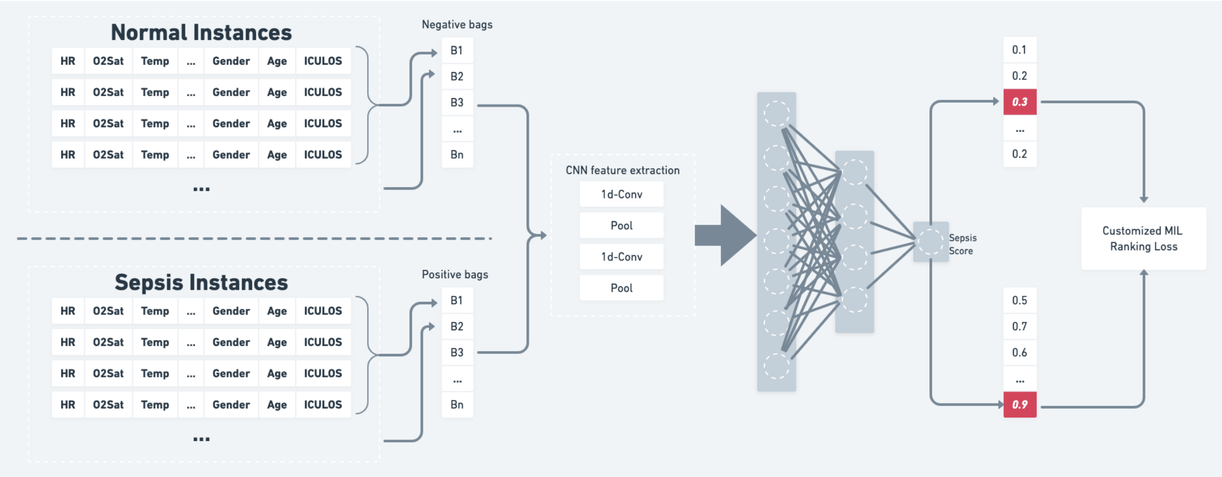


Figure 6.7: Flow diagram of proposed multi-instance learning approach.

Customised Loss Function

In order to implement the ranking model, we design a set of new ranking loss function. One can optimise the customised hinge loss function with respect to the maximum scored instance in each bag, like Eq. 6.8

$$\min_{w}[\frac{1}{z}\sum_{j=1}^{z}\max(0,1-Y_{B_j}(\max_{i\in B_j}(w\cdot\phi(x_i))-b))]+||w||^2, \tag{6.8}$$

where $\phi(x)$ denote the feature extracted by the Convolutional layer, b is bias, w is the classifier to be learned, Y_{B_j} denotes bag-level label, z is the total number of bags. In our proposed approach, we pose sepsis detection as a regression problem. We want the sepsis bags to have higher sepsis scores than the normal ones, so our ranking loss function is given as Eq. 6.9.

$$l(B_s,B_n)=\max(0,1-\max_{i\in B_s}f(D_s^i)+\max_{i\in B_n}f(D_n^i))+\lambda\sum_{i}^{n-1}(f(D_s^i)-f(D_s^{i+1}))^2+||w||^2 \tag{6.9}$$

where D_s and D_n means the sepsis and normal data sample respectively, $f()$ is the sepsis score, w is the set of model parameters and the squared sum of neighbouring score in the end is used to control the smoothness of the score, since we assume the fact that the neighbouring scores should have similar values.

Data Preparation

Data preprocessing procedures similar to our previous experiments were applied in the PhysioNet dataset, e.g. features-wise normalisation and data re-balance between sepsis and normal cases. Additionally, an extra step is required to bags aggregation for every patient. The bag size was carefully chosen which was set to 4, after multiple experiment comparisons have been made. In fact, we found that bag size has little influence as long as it stays within a reasonable range. A larger size leads to a smaller number of bag samples and sacrifices a longer time that we can make a prediction before the onset.

Table 6.10: Comparison of performance between multi-instance learning (MIL) and baseline (RF) models.

Model Name	Accuracy	Precision	Recall	Weighted_F1	AUC
RF	0.70	0.71	0.68	0.68	0.77
MIL	0.75	0.76	0.74	0.75	0.80

Performance Evaluation

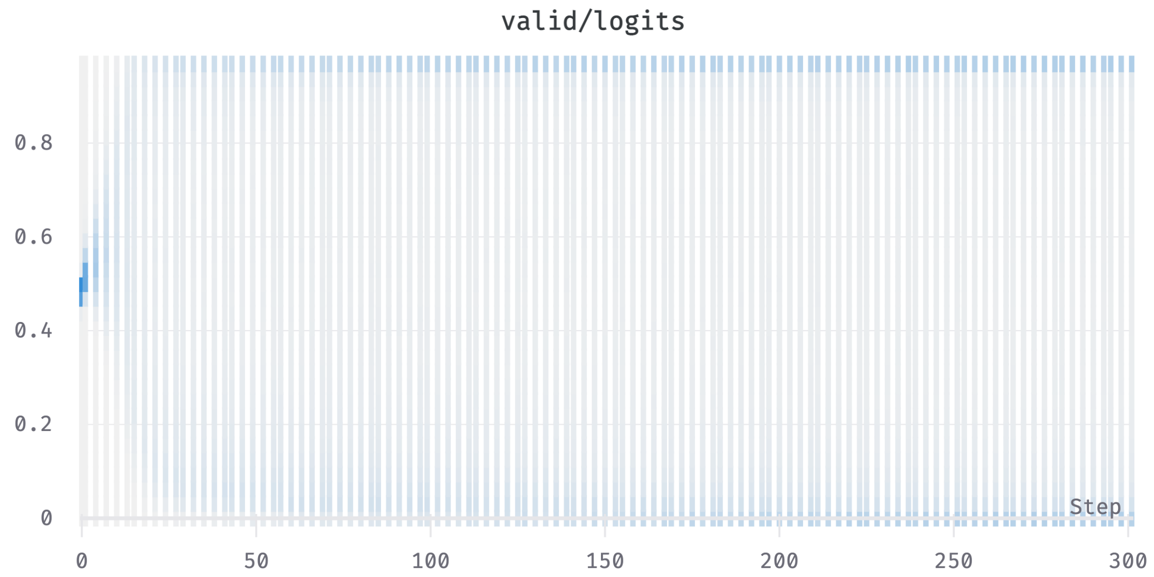


Figure 6.8: Predicted score distribution.

We set up a multi-layer neural network to calculate the proposed sepsis score as displayed in Fig. 6.9. Two 1d-Conv layers were used to extract features from a single patient record, followed by two fully-connected layers as the predictor. At last, we used the sigmoid function to scale the output score to the range of [0, 1]. So scores closer to 1 means a higher probability that this record belongs to a patient who is going to have a sepsis onset in the next six hours.

On the Physionet dataset, our multi-instance learning model achieved an accuracy of 0.75 and an AUC of 0.80, and compared to our baseline model RF which was the best model on the dataset collected from the NICU, they were 7.1% and 3.9% higher, respectively. Other comparison metrics are shown in Table. 6.10 below. We also noted down the predicted scores along the training course, and from the histogram (Fig. 6.8) it is clear to see that scores gathered around the value 0.5 at the beginning and polarised as the training course went on.

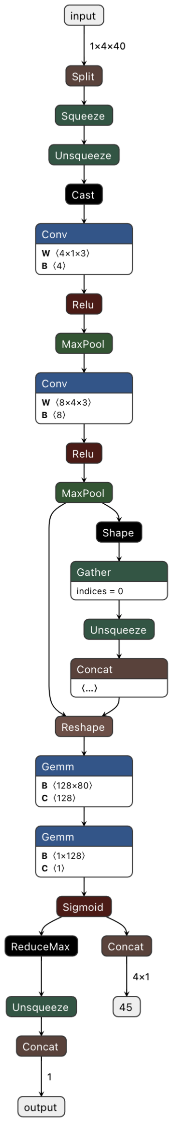


Figure 6.9: MIL network structure.

6.3.3 Reflection

We explored and evaluated two approaches to achieve sepsis prediction in this section. In the first method, we incorporated the popular LSTM to deal with the temporal information in the medical time series and had two findings. However, our proposed CNN-LSTM model did not achieve optimal performance. Instead, the FC encoder outperformed all other candidatures because cross-feature patterns are weak which is not the most suitable case for 1d-CNN. Also, we tested GASF-based model proposed in the last chapter and found it performed much better in PhysioNet dataset than it in NICU dataset due to the larger scale of training samples and its capability of preserving original temporal relation. Finally, we applied the multi-instance learning approach based on the trade-off between how long before the onset we can make predictions and how accurate our predictions are. When four continuous records are packed as a bag, if any record in it is classified as sepsis the label of the bag is set to sepsis. This could be used to improve the accuracy, but at the cost of four records every time a prediction is made. The performance of all the models we have explored (including those for neonatal sepsis detection in last chapter) are listed below in Table. 6.11. The optimal value in each measurement is highlighted. The model comprising FC encoder and LSTM decoder achieved the best performance among all the candidate models.

6.4 Chapter Summary

This chapter followed the workflow proposed in the unified sepsis detection framework to predict sepsis patient outcome based on two public datasets exploring the answers of RQ-3, RQ-5 and RQ-6 on adult patients. We implemented three cycles, as we did when predicting neonatal sepsis, which included collecting data, analysing data, and training data. We did not collect data in real-time, instead, used two public datasets from the PhysioNet Challenge 2019 and MIMIC III. Our feature analysis was performed to gain insights into the data, but we left it up to the model since complicated models such as encoder-decoder structures can automatically select features by training a series of weights. Finally, in cycle 3, we developed and refined two methods, an encoder-decoder structure and multi-instance learning, in order to achieve the predetermined goal of predicting sepsis. In accordance with the Action Research framework, the outcomes were summarised and compared. Analysing their performance, as well as identifying the underlying reasons for their variance, we were ultimately able to demonstrate the feasibility of predicting sepsis using AI-based models.

Table 6.11: Summary of all the models we have explored. Abbreviations: LR - logistic regression, SVM - support vector machine, RF - random forest, GBDT - gradient boosting decision tree, NN - neural network, CNN - convolutional neural network, GASF - gramian angular summation field, LSTM - long short-term memory, FC - fully connected, MLP - multi layer perceptron, MIL - multi instance learning

Type	Model Name	Accuracy	Precision	Recall	Weighted_F1	AUC
Classical Models	LR	0.50	0.51	0.44	0.47	0.50
	SVM	0.54	0.61	0.20	0.30	0.53
	RF	0.70	0.71	0.68	0.68	0.77
	GBDT	0.64	0.66	0.56	0.60	0.64
	NN	0.50	0.50	0.85	0.63	0.50
Image Transformation	16×16 + CNN	0.74	0.76	0.70	0.73	0.82
	32×32 + CNN	0.80	0.74	0.88	0.80	0.89
	16×16 + GASF	0.54	0.55	0.52	0.53	0.64
	128×128 + GASF	0.61	0.66	0.59	0.63	0.66
	227×227 + AlexNet	0.59	0.62	0.65	0.63	0.65
LSTM	Single LSTM	0.79	0.79	0.79	0.79	0.85
	Stacked LSTM	0.82	0.82	0.82	0.82	0.84
	FC + LSTM	0.87	0.87	0.87	0.87	0.95
	MLP baseline	0.52	0.52	0.52	0.52	0.53
	GASF baseline	0.78	0.77	0.79	0.78	0.86
MIL	MIL	0.75	0.76	0.74	0.75	0.80

Chapter 7

Conclusion and Discussion

Both adults and infants are susceptible to sepsis, particularly those in intensive care wards. Sepsis contributes to a substantial portion of mortality and morbidity in these patients. Currently, the diagnosis of sepsis is based upon a blood culture test, which can take up to two days. Clinicians tend to use antibiotics when they are waiting for test results in order to treat the patient in time and reduce the probability of disease progression that may result in patient deterioration. Overuse of antibiotics could lead to side effects for patients, thereby a mechanism for early detection is urgently needed.

First, we carried out a systematic review of research conducted in these fields throughout the last decade, revealing that most works including clinical trials were done from a medical point of view, rather than from a technical perspective. In spite of the fact that machine learning models were involved, only primitive ones were used. In addition, we noticed that no complete workflow had been defined for the detection of sepsis using machine learning methods. Our aim was to fill this gap by proposing a unified sepsis detection framework that described all aspects of the workflow from data collection to performance evaluation of machine learning based sepsis detection tasks for both adults and infants.

As part of this thesis, we used the action research methodology to guide and design a three-cycle process to implement and improve the previously mentioned sepsis detection framework. During the first cycle, data were collected from different sources, during the second cycle, the available features were analysed, and in the third cycle, multiple models, including classical machine learning as well as deep learning, were developed and evaluated as predictors for imminent sepsis. In this study, we converted the time series into 2D colour images, allowing us to take full advantage of the extraordinary capabilities of CNN and construct a deep neural network with higher accuracy than classic models. Our hybrid model combines CNN and RNN with LSTM units in order to exploit both the local feature extraction capability of CNN and the long-term memory capacity of LSTMs. Moreover, we utilised the multi-instance learning algorithm by combining a series of records into one bag so that any record classified as sepsis will lead to the sepsis label of the entire bag. The method was designed to balance time granularity and accuracy and, as long as the prediction is correct, it does not matter how many hours the onset occurs after the prediction.

7.1 Research Contributions

We have listed our contribution made by this thesis in Chapter 1, and this section will provide a more detailed description of the contributions of the research.

7.1.1 Contribution to Knowledge

A systematic literature review (SLR) on existing status of sepsis detection

We did a systematic review on the state-of-the-art approaches of sepsis-related detection, including neonatal sepsis, sepsis onset, septic shock and sepsis mortality in Chapter 2. The literature review was performed under the classification framework guided by six questions which gave a concise and clear concept of the current situation of research in this field. Specifically, we examined the works in PhysioNet Challenge 2019 which held the topic of ”early detection of sepsis from clinical data”. The methods adopted by each team and their outcomes were investigated and compared to provide information on the latest techniques and methods that could facilitate AI-aided sepsis detection. We believe our review revealed the trend of research in the related area, showed the latest cutting-edge techniques of sepsis detection with clinical data, and provided valuable insights for researchers just entering this domain.

A unified sepsis detection framework for both adults and infants

Another significant contribution of our research is the proposal of a unified sepsis detection framework, which defines the entire workflow of predicting sepsis with data-driven AI models. We discussed what types of data is needed, how medical records should be sampled and some other details in data collection. Then we specified procedures that have to be followed before the model training. Clinical data has some unique characteristics that might not exist in other forms of data, so it requires extra preprocessing steps like resampling at a different frequency compared to normal data projects. Unlike a traditional machine learning pipeline, this framework pre-defines the major variables fairly related to sepsis onset that researchers should focus on, and the model training step elicited two schemes that need to be considered when preparing data, one we call a data scheme, and another one is alignment scheme. Finally, we proposed the four evaluation metrics specifically for sepsis detection.

Our framework provides guidance for researchers new to the field of clinical data mining to start their research. Meanwhile, it preserves generalisability to some extent that one can easily implement a model from scratch just following our framework, and fine-tune the details according to individual cases which are also time sensitive medical events prediction problems..

7.1.2 Contribution to practice

Exploiting the feasibility of incorporating raw data from ICU to facilitate sepsis prediction

In chapter 5, raw data collected from NICU was used to predict the onset of neonatal sepsis. On these datasets, we have tested five classic machine learning models. In hospitals currently using rule-based sepsis diagnosis methods, live data would be temporarily stored and discarded after 24 hours. With our proposed methods, real-time monitored vital signs can be used to predict sepsis before it occurs, supporting clinicians' decision-making by incorporating them into their diagnosis workflow.

Application of multiple types of machine learning based models in predicting sepsis onset

On the basis of multiple datasets, we investigate the possibility and verify the feasibility of predicting sepsis with machine learning and deep learning methods. The results of the experiments presented in Chapters 5 and 6 demonstrated the effectiveness of AI-based approaches. Additionally, we proposed multiple candidature models, and evaluated their performance on different medical datasets, illustrating their effectiveness under different data schemes and alignment schemes. Our best AUC on the PhysioNet dataset was 0.95, and it corresponded to a prediction six hours before the onset of the disease. AUC of such a high level qualifies our model as an early warning system for clinicians, and could trigger an alarm in advance.

The novel method of converting multi-variate time series data to image in classification task

This thesis also suggested that the 1D time series of clinical data could be transformed into 2D images prior to feeding them to training models. We can benefit from the pre-trained models that may be applied directly to the transformed datasets since the simple transformation deals with data that has exactly three variables. Another way is based on the GASF transformation, which allows images of size $n \times n$ to be generated from only n data points while maintaining time relationships. Due to this characteristic, it is appropriate for datasets of small size. The methods are described in Chapter 5. Results from the experiment demonstrated the effectiveness of this technique, which can improve accuracy and AUC to 0.8 and 0.89, respectively. Further, the transformation of the raw data into images is only one of the many ways that raw data can be transformed into other forms prior to training, and our work is only a preliminary attempt that may lead to more refined ideas for future study.

Contribution to the application of multi-instance algorithm

In our thesis, we also developed a multi-instance learning model to make the prediction. Multi-instance learning is a technique that was previously used for the recognition of objects in images and videos, but we found that it could be used to improve the performance of our sepsis prediction model. Experiments and outcomes were reported in Chapter 6. Multiple instances are aggregated into bags with labels assigned to each bag instead of each instance, according to the algorithm. By doing so, we can benefit from two aspects. Any positive instance in a particular bag would indicate a positive label for that bag as opposed to the traditional method that requires an accurate prediction for every instance. Furthermore, according to their contents, labelling bags relieves the stress of labelling work during data collection.

7.2 Reviewing Research Objective 1

RO-1: Design an efficient data acquisition scheme. Since the bedside monitors broadcast all the data in LAN, we collect vital signs from bedside monitors in NICU by setting up two laptops in the ward, connecting to the local network. With the help of eDataGrabber software, provided by the same manufacturer of the monitor, data packets in the LAN are captured and stored in the local hard drive. However, other dimensions of data is much more difficult to access, lab test results and demographic data are distributed in multiple sub-systems within the hospital network, and they are not able to search or export. As a result, for these two types of data, a large amount of manual work is required, and no better way has been developed yet. PhysioNet Challenge and MIMIC III datasets are much easier to get, the only thing we need to do is to finish the test about human experiments ethics and then download the raw CSV files.

7.2.1 Answering Research Question 1

RQ-1: How should the vital signs be collected from bedside monitors in NICU? Our neonatal data is primarily obtained from the NICU at Monash Children's Hospital. In collaboration with Dräger, the manufacturer of the bedside monitors in the NICU, we have collected real-time vital signs data from patients under intensive care. The Dräger company provided us with the eDataGrabber software which enabled us to record vital signs broadcast across the local area network by their devices. These data were then securely stored in two laptop computers set up at nurse stations of the NICU.

7.3 Reviewing Research Objective 2

RO-2: Investigate the relations between physiological parameters and sepsis, and find critical ones that most related to sepsis. To fulfil this objective basically, what we need to do is feature selection. As stated in previous chapters, we have tried three different methods to filter valuable and informative features. This objective has two small research questions, and after the completion of our research, they can be answered as follows.

7.3.1 Answering Research Question 2

RQ-2: How many physiological parameters are available to researchers? With the help of the supervisors, we established very close cooperation with Monash Children's Hospital, so that we can have access to data from infants admitted in intensive care units. The accessible variables include heart rate, respiratory rate, oxygen saturation, pulse, readings from PS25255 sensors and multiple demographic data and laboratory test results, all of which contribute to our research. For public datasets, PhysioNet provides 40 features from two hospitals for the purpose of training a sepsis prediction model, as we have listed in Table. 4.3 and discussed in chapter 4.2, and MIMIC III databases contain even more medical measurements, see Table 4.6 but for the alignment of feature sets of others, we tried to select the same set of features as PhysioNet provides in our research.

7.3.2 Answering Research Question 3

RQ-3: Which critical physiological parameters can predict sepsis before it occurs? From our experiments, all physiological parameters contribute to the final prediction to certain degrees. We can answer this question by making feature selection in three ways. Specifically, for NICU datasets, out of 40 features in the public dataset from PhysioNet, ICULOS (ICU length-of-stay), Temperature and Calcium are the three most important variables with significance larger than 0.04 derived from xgboost model.

7.4 Reviewing Research Objective 3

RO-3: Design, develop and implement suitable algorithms for the clinical data, and make some adjustment to improve the performance. This objective is the most important among the three and includes our major contribution. To achieve this objective, three questions have to be answered.

7.4.1 Answering Research Question 4

RQ-4: Can existing methods fulfil the current demand of early detection? According to our review, current work in this field can not satisfy the demand for early detection in terms of the accuracy of the prediction and how long the prediction can be made before it actually happens, and that's the underlying motivation of our research. Another finding of our literature review is that most work that has been done is from the clinical angle, examining the relations between single vital sign or other measurements and sepsis onset, but little machine learning boosted methods were explored.

7.4.2 Answering Research Question 5

RQ-5: Is it possible to adjust existing methods to improve their performance? We have done some work to adapt existing machine learning models trying to improve their performance and achieve little progress. Most of them are basic machine learning models like SVM and Random Forest. Two ensemble models i.e. RF and GBDT achieved the best result out of five candidatures, with AUC of 0.77 and 0.64, but these are still not acceptable when dealing with serious problems like the diagnosis of diseases for humans. The conclusion we have is that limited by the relatively simple structure, and size of the NICU datasets, no significant improvement was made when experimenting with classic machine models.

7.4.3 Answering Research Question 6

RQ-6: Is it possible for any new approaches to have a better performance for this early detection task? This question is our major research area. We have examined both CNN and RNN and tried to combine them to fit them in our sepsis detection case and achieve a promising result. We also experimented with two different algorithms that convert 1-d time series data into images, in which way we were able to utilise the extraordinary capability of CNN in image classification. The multi-instance model is another way to improve the predictive accuracy by predicting the sepsis label of an aggregation of records rather than an individual one. The summary of results is listed in Table 6.11. Out of all the models we have examined, the one consisting of the fully-connected encoder and LSTM decoder achieved the highest accuracy and AUC, making it the optimal model, and its performance also showed the potential to be applied in the process of diagnosis to facilitate clinicians decision making.

7.5 Limitations and Future Research

We encountered a number of difficulties and limitations in the course of our research. Based on our findings, we were able to identify research directions that may be able to resolve the problems or address the limitations.

ICU patients are continuously monitored, and variables, particularly vital signs, could be considered a stream of dynamic data. The models we have tested are all offline, which means they cannot be incrementally trained with new inputs. One of the major weaknesses of offline models is that they cannot be automatically updated in real-time without manual intervention. To keep the offline model up-to-date, retraining the model with newer data periodically is frequently required. This incurs extra maintenance costs and results in less stable performance.

Our datasets are also extremely unbalanced in terms of sepsis and normal patients regardless of the data source. Based on our analysis, we can see that downsampling the majority class alone is an expedient. As a result, data size will be shrunk, important information will be lost, and the prediction performance will be negatively impacted. Alternatively, to rebalance data samples, one can upsample records from the minority class. However, currently there is no effective method that can synthesise a whole series of data while keeping its internal patterns.

Additionally, unexpected interruption is another issue during the data collection process.. For instance, in the NICU at Monash Children's Hospital, a nurse usually is responsible for approximately 3-4 infants, and they place the babies in cots within the same room, or as close as possible, but the sensors are attached to certain cots. Consequently, in the final collected datasets, it is almost impossible to determine what caused them when interruptions occur. A nurse might have moved the patient, and it may also have been caused by the sensors being detached by the unconscious movement of the patient. The interruptions lasting longer than ten minutes are interpreted as the babies transferring from one cot to another for now, but it is only a rough estimation and not 100% accurate.

Based on the current limitations, we also present some potential topics for future research. First, with access to the real-time data stream, it is perfect for the deployment of an online incremental learning algorithm, so that the model is always up-to-date, and can adjust to future data streams without human intervention. Second, develop a proprietary upsampling algorithm by synthesising time-series data to mitigate the data imbalance issue. Using a Generative Adversarial Network (GAN) seems to be a promising method to achieve this goal due to its known capability of data synthesis and generation. If data balance between two classes can be achieved without information loss, the performance of candidate models could be potentially increased. Third, understanding the reasons for interruptions and identifying when sensors are connected to another patient may improve the quality of data collection and preprocessing workflows. A careful examination of the patterns of data reading before and after transferring, as well as the interruption interval, will be helpful. By developing an algorithm that is capable of distinguishing baby transfer events, we are able to clearly and accurately label all the data segments, rather than discard those ambiguous records, preventing unnecessary waste.

Appendix A

Data Samples

This appendix contains data samples from NICU (Table. A.1), PhysioNet (Table. A.2 - A.4) and MIMIC III (Table. A.5 to A.10) datasets. Given the many features PhysioNet and MIMIC III datasets have, we split their samples into multiple tables for display convenience.

Table A.1: Data samples from NICU dataset.

	RELSEC	ABSSEC	RELTIME	JULIAN	DATE	TIME	RESP	%Pace	PLS	HR	25255	SpO2	NBP D	NBP M	NBP S
1936	116382	107830560	32:19:42	1510358459	11/11/2017	00:00:59	83		154	154	2.20	100			
1937	116442	107830620	32:20:42	1510358519	11/11/2017	00:01:59	65		151	152	2.41	100			
1938	116502	107830680	32:21:42	1510358579	11/11/2017	00:02:59	52		173	156	1.60	98			
1939	116562	107830740	32:22:42	1510358639	11/11/2017	00:03:59	92		147	147	2.06	96			
1940	116622	107830800	32:23:42	1510358699	11/11/2017	00:04:59	73		152	151	2.21	99			
1941	116682	107830860	32:24:42	1510358759	11/11/2017	00:05:59	71		151	151	2.25	99			
1942	116742	107830920	32:25:42	1510358819	11/11/2017	00:06:59	80		153	151	2.28	100			
1943	116802	107830980	32:26:42	1510358880	11/11/2017	00:08:00	71		147	146	2.01	99			
1944	116862	107831040	32:27:42	1510358940	11/11/2017	00:09:00	72		150	149	2.04	99			

Table A.2: Data samples from PhysioNet dataset - 1.

	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	FiO2	pH	PaCO2	SaO2	AST
0	97.0	98.5	35.89	145.0	100.33		19.0								
1	96.0	98.0		143.0	91.0		22.0								
2	97.0	98.0		137.0	93.67		27.0								
3	98.0	98.0		151.0	112.33		30.0		5.0			7.39	53.0		
4	104.0	98.0	37.11	135.0	79.67		21.0								
5	103.0	97.0		150.0	96.0		27.0								
6	98.0	96.0		140.0	95.33		30.0			31.0					
7	94.0	94.0		154.0	98.67		18.0								
8	93.0	97.0		146.0	90.0		26.0								
9															

Table A.3: Data samples from PhysioNet dataset - 2.

	BUN	Alkalinephos	Calcium	Chloride	Creatinine	Bilirubin_direct	Glucose	Lactate	Magnesium	Phosphate	Potassium
0											
1											
2											
3											
4											
5											
6	15.0		9.3	93.0	0.6		198.0		2.1	3.1	3.1
7											
8											
9											

Table A.4: Data samples from PhysioNet dataset - 3.

	Bilirubin_total	TroponinI	Hct	Hgb	PTT	WBC	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	ICULOS	SepsisLabel
0									81.08	0	1	0	-0.03	2	0
1									81.08	0	1	0	-0.03	3	0
2									81.08	0	1	0	-0.03	4	0
3									81.08	0	1	0	-0.03	5	0
4									81.08	0	1	0	-0.03	6	0
5									81.08	0	1	0	-0.03	7	0
6			38.2	13.2		11.1		262.0	81.08	0	1	0	-0.03	8	0
7									81.08	0	1	0	-0.03	9	0
8									81.08	0	1	0	-0.03	10	0
9									81.08	0	1	0	-0.03	11	0

Table A.5: Samples from ADMISSION table of MIMIC III datasets - 1.

	ROW_ID	SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE
0	21	22	165315	2196-04-09 12:26:00	2196-04-10 15:54:00		EMERGENCY	EMERGENCY ROOM ADMIT	DISC-TRAN CANCER/CHLDRN H	Private
1	22	23	152223	2153-09-03 07:15:00	2153-09-08 19:10:00		ELECTIVE	PHYS REFERRAL/NORMAL DELI	HOME HEALTH CARE	Medicare
2	23	23	124321	2157-10-18 19:34:00	2157-10-25 14:00:00		EMERGENCY	TRANSFER FROM HOSP/EXTRAM	HOME HEALTH CARE	Medicare
3	24	24	161859	2139-06-06 16:14:00	2139-06-09 12:48:00		EMERGENCY	TRANSFER FROM HOSP/EXTRAM	HOME	Private
4	25	25	129635	2160-11-02 02:06:00	2160-11-05 14:55:00		EMERGENCY	EMERGENCY ROOM ADMIT	HOME	Private
5	26	26	197661	2126-05-06 15:16:00	2126-05-13 15:00:00		EMERGENCY	TRANSFER FROM HOSP/EXTRAM	HOME	Medicare
6	27	27	134931	2191-11-30 22:16:00	2191-12-03 14:45:00		NEWBORN	PHYS REFERRAL/NORMAL DELI	HOME	Private
7	28	28	162569	2177-09-01 07:15:00	2177-09-06 16:00:00		ELECTIVE	PHYS REFERRAL/NORMAL DELI	HOME HEALTH CARE	Medicare
8	29	30	104557	2172-10-14 14:17:00	2172-10-19 14:37:00		URGENT	TRANSFER FROM HOSP/EXTRAM	HOME HEALTH CARE	Medicare
9	30	31	128652	2108-08-22 23:27:00	2108-08-30 15:00:00	2108-08-30 15:00:00	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	DEAD/EXPIRED	Medicare

Table A.6: Samples from ADMISSION table of MIMIC III datasets - 2.

	LANGUAGE	RELIGION	MARITAL_STATUS	ETHNICITY	EDREGTIME	EDOUTTIME	DIAGNOSIS	HOSPITAL_EXPIRE_FLAG	HAS_CHARTEVENTS_DATA
0		UNOBTAINABLE	MARRIED	WHITE	2196-04-09 10:06:00	2196-04-09 13:24:00	BENZODIAZEPINE OVERDOSE	0	1
1		CATHOLIC	MARRIED	WHITE			CORONARY ARTERY DISEASE\CORONARY ARTERY BYPASS GRAFT/SDA	0	1
2	ENGL	CATHOLIC	MARRIED	WHITE			BRAIN MASS	0	1
3		PROTESTANT QUAKER	SINGLE	WHITE			INTERIOR MYOCARDIAL INFARCTION	0	1
4		UNOBTAINABLE	MARRIED	WHITE	2160-11-02 01:01:00	2160-11-02 04:27:00	ACUTE CORONARY SYNDROME	0	1
5		CATHOLIC	SINGLE	UNKNOWN/NOT SPECIFIED			V-TACH	0	1
6		CATHOLIC		WHITE			NEWBORN	0	1
7		CATHOLIC	MARRIED	WHITE			CORONARY ARTERY DISEASE\CORONARY ARTERY BYPASS GRAFT/SDA	0	1
8		CATHOLIC	MARRIED	UNKNOWN/NOT SPECIFIED			UNSTABLE ANGINA\CATH	0	1
9		CATHOLIC	MARRIED	WHITE			STATUS EPILEPTICUS	1	1

Table A.7: Samples from D_ICD_DIAGNOSES table of MIMIC III dataset.

	ROW_ID	ICD9_CODE	SHORT_TITLE	LONG_TITLE
0	174	01166	TB pneumonia-oth test	Tuberculous pneumonia [any form], tubercle bacilli not found by bacteriological or histological examination, but tuberculosis confirmed by other methods [inoculation of animals]
1	175	01170	TB pneumothorax-unspec	Tuberculous pneumothorax, unspecified
2	176	01171	TB pneumothorax-no exam	Tuberculous pneumothorax, bacteriological or histological examination not done
3	177	01172	TB pneumothorx-exam unkn	Tuberculous pneumothorax, bacteriological or histological examination unknown (at present)
4	178	01173	TB pneumothorax-micro dx	Tuberculous pneumothorax, tubercle bacilli found (in sputum) by microscopy
5	179	01174	TB pneumothorax-cult dx	Tuberculous pneumothorax, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture
6	180	01175	TB pneumothorax-histo dx	Tuberculous pneumothorax, tubercle bacilli not found by bacteriological examination, but tuberculosis confirmed histologically
7	181	01176	TB pneumothorax-oth test	Tuberculous pneumothorax, tubercle bacilli not found by bacteriological or histological examination, but tuberculosis confirmed by other methods [inoculation of animals]
8	182	01180	Pulmonary TB NEC-unspec	Other specified pulmonary tuberculosis, unspecified
9	183	01181	Pulmonary TB NEC-no exam	Other specified pulmonary tuberculosis, bacteriological or histological examination not done

Table A.8: Samples from D_ITMES table of MIMIC III dataset.

	ROW_ID	ITEMID	LABEL	ABBREVIATION	DBSOURCE	LINKSTO	CATEGORY	UNITNAME	PARAM_TYPE	CONCEPTID
0	457	497	Patient controlled analgesia (PCA) [Inject]		carevue	chartevents				
1	458	498	PCA Lockout (Min)		carevue	chartevents				
2	459	499	PCA Medication		carevue	chartevents				
3	460	500	PCA Total Dose		carevue	chartevents				
4	461	501	PCV Exh Vt (Obser)		carevue	chartevents				
5	1157	927	Allergy 2		carevue	chartevents				
6	1158	930	Ext		carevue	chartevents				
7	1159	935	Allergy 3		carevue	chartevents				
8	1160	938	blood cultures		carevue	chartevents				
9	1161	940	trach care		carevue	chartevents				

Table A.9: Samples from DIAGNOSES_ICD table of MIMIC III dataset.

	ROW_ID	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE
0	1297	109	172335	1.0	40301
1	1298	109	172335	2.0	486
2	1299	109	172335	3.0	58281
3	1300	109	172335	4.0	5855
4	1301	109	172335	5.0	4254
5	1302	109	172335	6.0	2762
6	1303	109	172335	7.0	7100
7	1304	109	172335	8.0	2767
8	1305	109	172335	9.0	7243
9	1306	109	172335	10.0	45829

Table A.10: Samples from PATIENTS table of MIMIC III dataset.

	ROW_ID	SUBJECT_ID	GENDER	DOB	DOD	DOD_HOSP	DOD_SSN	EXPIRE_FLAG
0	234	249	F	2075-03-13 00:00:00				0
1	235	250	F	2164-12-27 00:00:00	2188-11-22 00:00:00	2188-11-22 00:00:00		1
2	236	251	M	2090-03-15 00:00:00				0
3	237	252	M	2078-03-06 00:00:00				0
4	238	253	F	2089-11-26 00:00:00				0
5	239	255	M	2109-08-05 00:00:00				0
6	240	256	M	2086-07-31 00:00:00				0
7	241	257	F	2031-04-03 00:00:00	2121-07-08 00:00:00	2121-07-08 00:00:00	2121-07-08 00:00:00	1
8	242	258	F	2124-09-19 00:00:00				0
9	243	260	F	2105-03-23 00:00:00				0

Appendix B

Code Samples

B.1 Pseudo Code for GASF Algorithm

Algorithm 1 GASF

```
Normalise all the data to the range of [-1,1]: $\hat{x}_i \leftarrow ((x_i - max(x)) + (x_i - min(x)))/(max(x) - min(x))$
Convert to polar coordinate: $\phi \leftarrow arccos(\tilde{x}_i)$
Create the Gramian Matrix: $M_{ij} = cos(\phi_i + \phi_j)$
```

B.2 Pseudo Code for Multi-Instance Learning Procedure

Algorithm 2 Multi-Instance Learning

```
Require: $a \geq 0, b \leq 0$, and stop condition: $c$
repeat
    Randomly generate $n_i$, where $a \leq n_i \leq b$
    Pick $n_i$ consecutive instances
    Check the label of each instance
    if $\#positive \geq 1$ then
        Label the bag positive
    else
        Label the bag negative
    end if
until All instances have been assigned to one bag
repeat
    Calculate the loss:
    Take one positive bag $B_s$ and one negative bag $B_n$
    Calculate the hinge loss $l_h \leftarrow max(0, 1 - \max_{i \in B_s} f(D_s^i) + \max_{i \in B_n} f(D_n^i))$
    Calculate the smoothing factor $s \leftarrow \lambda \sum_i^{n-1} (f(D_s^i) - f(D_s^{i+1}))^2$
    Calculate the l2 normalisation penalty $p \leftarrow ||w||^2$
    Minimise the ranking loss of two bags $l(B_s, B_n) \leftarrow l_h + s + p$
until The stop condition $c$ is satisfied.
```

B.3 Github Repository

This is the Github repository for all the codes used in this thesis.

https://github.com/jsxhhyf/Thesis

References

[1] CDC. Sepsis is a medical emergency. ACT FAST. https://www.cdc.gov/sepsis/what-is-sepsis.html, August 2021.

[2] Yonathan Freund, Najla Lemachatti, Evguenia Krastinova, Marie Van Laer, Yann-Erick Claessens, Aurélie Avondo, Céline Occelli, Anne-Laure Feral-Pierssens, Jennifer Truchot, Mar Ortega, Bruno Carneiro, Julie Pernet, Pierre-Géraud Claret, Fabrice Dami, Ben Bloom, Bruno Riou, Sébastien Beaune, and for the French Society of Emergency Medicine Collaborators Group. Prognostic Accuracy of Sepsis-3 Criteria for In-Hospital Mortality Among Patients With Suspected Infection Presenting to the Emergency Department. *JAMA*, 317(3):301–308, January 2017.

[3] Medley O'Keefe Gatewood, Matthew Wemple, Sheryl Greco, Patricia A Kritek, and Raghu Durvasula. A quality improvement project to improve early sepsis care in the emergency department. *BMJ Quality & Safety*, 24(12):787–795, December 2015.

[4] 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):801–810, February 2016.

[5] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix). *Intensive Care Medicine*, 22(7):707–710, July 1996.

[6] Alison R. Bedford Russell. Neonatal sepsis. *Paediatrics and Child Health*, 21(6):265–269, June 2011.

[7] S Vergnano. Neonatal sepsis: An international perspective. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 90(3):F220–f224, 2005.

[8] Ilan Gur, Arieh Riskin, Gal Markel, David Bader, Yaron Nave, Bernard Barzilay, Fabien

Eyal, and Arik Eisenkraft. Pilot Study of a New Mathematical Algorithm for Early Detection of Late-Onset Sepsis in Very Low-Birth-Weight Infants. *American Journal of Perinatology*, 32(04):321–330, July 2014.

[9] Department of Health & Human Services. Sepsis in neonates. https://www2.health.vic.gov.au:443/hospitals-and-health-services/patient-care/perinatal-reproductive/neonatal-ehandbook/infections/sepsis.

[10] R. S. Baltimore, S. M. Huie, J. I. Meek, A. Schuchat, and K. L. O'Brien. Early-Onset Neonatal Sepsis in the Era of Group B Streptococcal Prevention. *PEDIATRICS*, 108(5):1094–1098, November 2001.

[11] Maria Pais, Elsa Sanatombi Devi, Muralidhar V Pai, Leslie Lewis, Anice Gorge, Sreemathi S Mayya, and Bairy KL. Neonatal Sepsis, Bacterial Isolates and Antibiotic Susceptibility Patterns among Neonates. *THE NURSING JOURNAL OF INDIA*, page 4.

[12] Onno K. Helder, Johannes Brug, Caspar W.N. Looman, Johannes B. van Goudoever, and René F. Kornelisse. The impact of an education program on hand hygiene compliance and nosocomial infection incidence in an urban Neonatal Intensive Care Unit: An intervention study with before and after comparison. *International Journal of Nursing Studies*, 47(10):1245–1252, October 2010.

[13] Barbara J Stoll, Nellie Hansen, Avroy a Fanaroff, Linda L Wright, Waldemar A Carlo, Richard a Ehrenkranz, James a Lemons, Edward F Donovan, Ann R Stark, Jon E Tyson, William Oh, Charles R Bauer, Sheldon B Korones, Seetha Shankaran, Abbot R Laptook, David K Stevenson, Lu-Ann Papile, W Kenneth Poole, and Abstract Objective. Late-Onset Sepsis in Very Low Birth Weight Neonates : The Experience of the NICHD Neonatal Research Network. *Pediatrics*, 110(2):285–291, 2015.

[14] I. R. Makhoul, P. Sujov, T. Smolkin, A. Lusky, and B. Reichman. Epidemiological, Clinical, and Microbiological Characteristics of Late-Onset Sepsis Among Very Low Birth Weight Infants in Israel: A National Survey. *PEDIATRICS*, 109(1):34–39, January 2002.

[15] Ichiro Morioka, Satoru Morikawa, Akihiro Miwa, Hirotaka Minami, Katsuhiko Yoshii, Masaaki Kugo, Yoshiki Kitsunezuka, Miki Enomoto, Takumi Jikimoto, Masakuni Nakamura, Naoki Yokoyama, Hisahide Nishio, Masafumi Matsuo, and Hideto Yamada. Culture-Proven Neonatal Sepsis in Japanese Neonatal Care Units in 2006–2008. *Neonatology*, 102(1):75–80, 2012.

[16] Christoph P. Hornik, Daniel K. Benjamin, Kristian C. Becker, Daniel K. Benjamin, Jen-

nifer Li, Reese H. Clark, Michael Cohen-Wolkowiez, and P. Brian Smith. Use of the Complete Blood Cell Count in Early-onset Neonatal Sepsis:. *The Pediatric Infectious Disease Journal*, 31(8):799–802, August 2012.

[17] J. H. Jiang, N. C. Chiu, F. Y. Huang, H. A. Kao, C. H. Hsu, H. Y. Hung, J. H. Chang, and C. C. Peng. Neonatal sepsis in the neonatal intensive care unit: Characteristics of early versus late onset. *Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi*, 37(5):301–306, October 2004.

[18] Harsha Gowda, Robert Norton, Andrew White, and Yogavijayan Kandasamy. Late-onset Neonatal Sepsis—A 10-year Review From North Queensland, Australia. *The Pediatric Infectious Disease Journal*, 36(9):883–888, September 2017.

[19] Margaret Gilfillan and Vineet Bhandari. Biomarkers for the diagnosis of neonatal sepsis and necrotizing enterocolitis: Clinical practice guidelines. *Early Human Development*, 105:25–33, February 2017.

[20] Andres Camacho-Gonzalez, Paul W. Spearman, and Barbara J. Stoll. Neonatal Infectious Diseases. *Pediatric Clinics of North America*, 60(2):367–389, April 2013.

[21] Birju A Shah and James F Padbury. Neonatal sepsis: An old problem with new insights. *Virulence*, 5(1):170–178, January 2014.

[22] James L. Wynn. Defining neonatal sepsis. *Current Opinion in Pediatrics*, 28(2):135–140, April 2016.

[23] Jeffrey S Gerdes. Diagnosis and management of bacterial infections in the neonate. *Pediatric Clinics of North America*, 51(4):939–959, August 2004.

[24] R. A. Polin and the COMMITTEE ON FETUS AND NEWBORN. Management of Neonates With Suspected or Proven Early-Onset Bacterial Sepsis. *PEDIATRICS*, 129(5):1006–1015, May 2012.

[25] Emanuele Rezoagli, Bairbre McNicholas, Peter Moran, and John G Laffey. Sepsis Therapies: Insights from Population Health to Cellular Therapies and Genomic Medicine. page 3.

[26] G Jawaheer, T J Neal, and N J Shaw. Blood culture volume and detection of coagulase negative staphylococcal septicaemia in neonates. page 3.

[27] PATRICIA R NEAL, MARTIN B KLEIMAN, JANET K REYNOLDS, STEPHEN D ALLEN, JAMES A LEMONS, and PAO-LO YU. Volume of Blood Submitted for Culture from Neonates. *J. CLIN. MICROBIOL.*, page 4, 2018.

[28] Joseph Y. Ting, Anne Synnes, Ashley Roberts, Akhil Deshpandey, Kimberly Dow, Eugene W. Yoon, Kyong-Soon Lee, Simon Dobson, Shoo K. Lee, Prakesh S. Shah, and

for the Canadian Neonatal Network Investigators. Association Between Antibiotic Use and Neonatal Mortality and Morbidities in Very Low-Birth-Weight Infants Without Culture-Proven Sepsis or Necrotizing Enterocolitis. *JAMA Pediatrics*, 170(12):1181, December 2016.

[29] Andrew Booth, Anthea Sutton, and Diana Papaioannou. *Systematic Approaches to a Successful Literature Review*. SAGE, May 2016.

[30] Tahereh Pourhabibi, Kok-Leong Ong, Booi H. Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, page 113303, April 2020.

[31] 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*, 3:160035, May 2016.

[32] Goldberger Ary L., Amaral Luis A. N., Glass Leon, Hausdorff Jeffrey M., Ivanov Plamen Ch., Mark Roger G., Mietus Joseph E., Moody George B., Peng Chung-Kang, and Stanley H. Eugene. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, June 2000.

[33] Laurel A. Despins. Automated Detection of Sepsis Using Electronic Medical Record Data: A Systematic Review. *Journal for Healthcare Quality: Official Publication of the National Association for Healthcare Quality*, 39(6):322–333, 2017 Nov/Dec.

[34] Mehanas Shahul and Mehanas Shahul. Machine Learning Based Analysis of Sepsis: Review. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*, pages 1–4, February 2020.

[35] Alejandro Baldominos, Adráan Puello, Hasan Oğul, Tunç Aşuroğlu, and Ricardo Colomo-Palacios. Predicting Infections Using Computational Intelligence – A Systematic Review. *IEEE Access*, 8:31083–31102, 2020.

[36] Md. Mohaimenul Islam, Tahmina Nasrin, Bruno Andreas Walther, Chieh-Chen Wu, Hsuan-Chia Yang, and Yu-Chuan Li. Prediction of sepsis patients using machine learning approach: A meta-analysis. *Computer Methods and Programs in Biomedicine*, 170:1–9, March 2019.

[37] Andreas Pregernig, Mattia Müller, Ulrike Held, and Beatrice Beck-Schimmer. Prediction of mortality in adult patients with sepsis using six biomarkers: A systematic review and meta-analysis. *Annals of Intensive Care*, 9(1):125, November 2019.

[38] Rodrigo Serafim, José Andrade Gomes, Jorge Salluh, and Pedro Póvoa. A Comparison

of the Quick-SOFA and Systemic Inflammatory Response Syndrome Criteria for the Diagnosis of Sepsis and Prediction of Mortality. *Chest*, 153(3):646–655, March 2018.

[39] Amber M. Sawyer, Eli N. Deal, Andrew J. Labelle, Chad Witt, Steven W. Thiel, Kevin Heard, Richard M. Reichley, Scott T. Micek, and Marin H. Kollef. Implementation of a real-time computerized sepsis alert in nonintensive care unit patients*:. *Critical Care Medicine*, 39(3):469–473, March 2011.

[40] Craig A. Umscheid, Joel Betesh, Christine VanZandbergen, Asaf Hanish, Gordon Tait, Mark E. Mikkelsen, Benjamin French, and Barry D. Fuchs. Development, implementation, and impact of an automated early warning and response system for sepsis: EWRS for Sepsis. *Journal of Hospital Medicine*, 10(1):26–31, January 2015.

[41] Ioan Stanculescu, Christopher K I Williams, and Yvonne Freer. Autoregressive Hidden Markov Models for the Early Detection of Neonatal Sepsis. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1560–1570, 2014.

[42] Jenna Wiens, Eric Horvitz, and John V Guttag. Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. page 9.

[43] Mauricio Monsalve, Sriram Pemmaraju, Sarah Johnson, and Philip M. Polgreen. Improving Risk Prediction of Clostridium Difficile Infection Using Temporal Event-Pairs. In *2015 International Conference on Healthcare Informatics*, pages 140–149, Dallas, TX, USA, October 2015. IEEE.

[44] Jacob S. Calvert, Daniel A. Price, Uli K. Chettipally, Christopher W. Barton, Mitchell D. Feldman, Jana L. Hoffman, Melissa Jay, and Ritankar Das. A computational approach to early sepsis detection. *Computers in Biology and Medicine*, 74:69–73, July 2016.

[45] Ilan Gur, Gal Markel, Yaron Nave, Igor Vainshtein, Arik Eisenkraft, and Arieh Riskin. A mathematical algorithm for detection of late-onset sepsis in very-low birth weight infants: A preliminary diagnostic test evaluation. *Indian pediatrics*, 51(8):647–650, 2014.

[46] Dee W. Ford, Andrew J. Goodwin, Annie N. Simpson, Emily Johnson, Nandita Nadig, and Kit N. Simpson. A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data. *Critical Care Medicine*, 44(2):319–327, February 2016.

[47] 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.

[48] Aytac Tekin, Mustafa Ulas, and Fatma Uzun. Analysis of the Neonatal Sepsis Data Set with Data Mining Methods. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pages 1–4, November 2019.

[49] Ditte Caroline Raben, Birgit Viskum, Kim L. Mikkelsen, Jeanette Hounsgaard, Søren Bie Bogh, and Erik Hollnagel. Application of a non-linear model to understand healthcare processes: Using the functional resonance analysis method on a case study of the early detection of sepsis. *Reliability Engineering & System Safety*, 177:1–11, September 2018.

[50] X. Navarro, F. Porée, A. Beuchée, and G. Carrault. Artifact rejection and cycle detection in immature breathing: Application to the early detection of neonatal sepsis. *Biomedical Signal Processing and Control*, 16:9–16, February 2015.

[51] Qing Li, L. Frank Huang, Jiang Zhong, Lili Li, Qi Li, and Junhao Hu. Data-driven Discovery of a Sepsis Patients Severity Prediction in the ICU via Pre-training BiLSTM Networks. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 668–673, November 2019.

[52] Rafael Gómez, Nuria García, Gonzalo Collantes, Francisco Ponce, and Pau Redon. Development of a Non-Invasive Procedure to Early Detect Neonatal Sepsis using HRV Monitoring and Machine Learning Algorithms. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 132–137, June 2019.

[53] Zhongheng Zhang and Yucai Hong. Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: The use of electronic healthcare records with LASSO regression. *Oncotarget*, 8(30):49637–49645, July 2017.

[54] Senthil K. Nachimuthu and Peter J. Haug. Early detection of sepsis in the emergency department using Dynamic Bayesian Networks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:653–662, 2012.

[55] Simon Meyer Lauritsen, Mads Ellersgaard Kalør, Emil Lund Kongsgaard, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artificial Intelligence in Medicine*, 104:101820, April 2020.

[56] Chen Lin, Yuan Zhang, Julie Ivy, Muge Capan, Ryan Arnold, Jeanne M. Huddleston, and Min Chi. Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 219–228, New York, NY, June 2018. IEEE.

[57] R Murat Demirer and Oya Demirer. Early Prediction of Sepsis from Clinical Data Using Artificial Intelligence. In *2019 Scientific Meeting on Electrical-Electronics Biomedical Engineering and Computer Science (EBBT)*, pages 1–4, April 2019.

[58] Mohammed Saqib, Ying Sha, and May D. Wang. Early Prediction of Sepsis in EMR

Records Using Traditional ML Techniques and Deep Learning LSTM Networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4038–4041, July 2018.

[59] 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.

[60] Mengsha Fu, Jiabin Yuan, and Chen Bei. Early Sepsis Prediction in ICU Trauma Patients with Using An Improved Cascade Deep Forest Model. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pages 634–637, October 2019.

[61] David W. Shimabukuro, Christopher W. Barton, Mitchell D. Feldman, Samson J. Mataraso, and Ritankar Das. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. *BMJ open respiratory research*, 4(1):e000234, 2017.

[62] Christopher Barton, Uli Chettipally, Yifan Zhou, Zirui Jiang, Anna Lynn-Palevsky, Sidney Le, Jacob Calvert, and Ritankar Das. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in Biology and Medicine*, 109:79–84, June 2019.

[63] Karen D. Fairchild and T. Michael O'Shea. Heart rate characteristics: Physiomarkers for detection of late-onset neonatal sepsis. *Clinics in Perinatology*, 37(3):581–598, September 2010.

[64] Antoine Honoré, Dong Liu, David Forsberg, Karen Coste, Eric Herlenius, Saikat Chatterjee, and Mikael Skoglund. Hidden Markov Models for Sepsis Detection in Preterm Infants. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1130–1134, May 2020.

[65] Franco van Wyk, Anahita Khojandi, and Rishikesan Kamaleswaran. Improving Prediction Performance Using Hierarchical Analysis of Real-Time Data: A Sepsis Case Study. *IEEE Journal of Biomedical and Health Informatics*, 23(3):978–986, May 2019.

[66] Hye Jin Kam and Ha Young Kim. Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine*, 89:248–255, October 2017.

[67] Shigehiko Schamoni, Holger A. Lindner, Verena Schneider-Lindner, Manfred Thiel, and Stefan Riezler. Leveraging implicit expert knowledge for non-circular machine

learning in sepsis prediction. *Artificial Intelligence in Medicine*, 100:101725, September 2019.

[68] Yuan Zhang, Chen Lin, Min Chi, Julie Ivy, Muge Capan, and Jeanne M. Huddleston. LSTM for septic shock: Adding unreliable labels to reliable predictions. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1233–1242, December 2017.

[69] Aiman Darwiche and Sumitra Mukherjee. Machine Learning Methods for Septic Shock Prediction. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality*, AIVR 2018, pages 104–110, Nagoya, Japan, November 2018. Association for Computing Machinery.

[70] Eli Bloch, Tammy Rotem, Jonathan Cohen, Pierre Singer, and Yehudit Aperstein. Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction. *Journal of Healthcare Engineering*, 2019:5930379, 2019.

[71] Susana M. Vieira, Luis F. Mendonça, Gonçalo J. Farinha, and João M.C. Sousa. Metaheuristics for feature selection: Application to sepsis outcome prediction. In *2012 IEEE Congress on Evolutionary Computation*, pages 1–8, June 2012.

[72] Adele H. Marshall, Kieran Payne, Karen J. Cairns, Stan Craig, and Emma McCall. Modelling the development of late onset sepsis and length of stay using discrete conditional survival models with a classification tree component. In *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6, June 2012.

[73] Chen Lin, Julie Ivy, and Min Chi. Multi-layer Facial Representation Learning for Early Prediction of Septic Shock. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 840–849, December 2019.

[74] 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.

[75] Ran Liu, Joseph L. Greenstein, Sridevi V. Sarma, and Raimond L. Winslow. Natural Language Processing of Clinical Notes for Improved Early Prediction of Septic Shock in the ICU. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6103–6108, July 2019.

[76] Jyoti Thakur, Sharvan Kumar Pahuja, and Roop Pahuja. Neonatal Sepsis Prediction Model for Resource-Poor Developing Countries. In *2018 2nd International Conference on Electronics, Materials Engineering Nano-Technology (IEMENTech)*, pages 1–5, May 2018.

[77] Vicent J. Ribas, Jesús Caballero López, Juan Carlos Ruiz-Rodríguez, Adolf Ruiz-Sanmartín, Jordi Rello, and Alfredo Vellido. On the use of decision trees for ICU outcome prediction in sepsis patients treated with statins. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 37–43, April 2011.

[78] J.E Garcia-Gallo, N.J Fonseca-Ruiz, L.A Celi, and J.F Duitama-Muñoz. One-Year Mortality Prediction in ICU Patients with Diagnosis of Sepsis Driven by Population Similarities. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 480–484, October 2019.

[79] Paweeya Raknim, Kun-chan Lan, Yung-Chieh Linker, and Yen-Tzu Lu. Position: On the Use of Low-cost Sensors for Non-intrustive Newborn Sepsis Monitoring. In *The 5th ACM Workshop on Wearable Systems and Applications*, WearSys '19, pages 39–40, Seoul, Republic of Korea, June 2019. Association for Computing Machinery.

[80] Rohan Joshi, Deedee Kommers, Laurien Oosterwijk, Loe Feijs, Carola van Pul, and Peter Andriessen. Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion. *IEEE Journal of Biomedical and Health Informatics*, 24(3):681–692, March 2020.

[81] Ivan Stojkovic and Zoran Obradovic. Predicting Sepsis Biomarker Progression under Therapy. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 19–24, June 2017.

[82] Luregn J. Schlapbach, Graeme MacLaren, Marino Festa, Janet Alexander, Simon Erickson, John Beca, Anthony Slater, Andreas Schibler, David Pilcher, Johnny Millar, Lahn Straney, and Australian & New Zealand Intensive Care Society (ANZICS) Centre for Outcomes & Resource Evaluation (CORE) and Australian & New Zealand Intensive Care Society (ANZICS) Paediatric Study Group. Prediction of pediatric sepsis mortality within 1 h of intensive care admission. *Intensive Care Medicine*, 43(8):1085–1096, August 2017.

[83] 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), September 2016.

[84] Farzaneh Khoshnevisan, Julie Ivy, Muge Capan, Ryan Arnold, Jeanne Huddleston, and Min Chi. Recent Temporal Pattern Mining for Septic Shock Early Prediction. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 229–240, June

2018.

[85] Andrea McCoy and Ritankar Das. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Quality*, 6(2):e000158, October 2017.

[86] Lisa Mellhammar, Adam Linder, Jonas Tverring, Bertil Christensson, John H. Boyd, Per Åkesson, and Fredrik Kahn. Scores for sepsis detection and risk stratification - construction of a novel score using a statistical approach and validation of RETTS. *PloS One*, 15(2):e0229210, 2020.

[87] Wan Fadzlina Wan Muhd Shukeri, Azrina Md Ralib, Nor Zamzila Abdulah, and Mohd Basri Mat-Nor. Sepsis mortality score for the prediction of mortality in septic patients. *Journal of Critical Care*, 43:163–168, February 2018.

[88] Yu Jiang, Lui Sha, Maryam Rahmaniheris, Binhua Wan, Mohammad Hosseini, Pengliu Tan, and Richard B. Berlin. Sepsis Patient Detection and Monitor Based on Auto-BN. *Journal of Medical Systems*, 40(4):111, April 2016.

[89] Kourosh T. Baghaei and Shahram Rahimi. Sepsis Prediction: An Attention-Based Interpretable Approach. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, June 2019.

[90] 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):1:1–1:15, April 2014.

[91] Shermeen Nizami, James Robert Green, and Carolyn McGregor. Service oriented architecture to support real-time implementation of artifact detection in critical care monitoring. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4925–4928, August 2011.

[92] Vicent J. Ribas, Alfredo Vellido, Juan Carlos Ruiz-Rodríguez, and Jordi Rello. Severe sepsis mortality prediction with logistic regression over latent factors. *Expert Systems with Applications*, 39(2):1937–1943, February 2012.

[93] Vicent J. Ribas, Jesús Caballero López, Adolf Ruiz-Sanmartín, Juan Carlos Ruiz-Rodríguez, Jordi Rello, Anna Wojdel, and Alfredo Vellido. Severe sepsis mortality prediction with relevance vector machines. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 100–103, August 2011.

[94] Eitam Sheetrit, Nir Nissim, Denis Klimov, and Yuval Shahar. Temporal Probabilistic Profiles for Sepsis Prediction in the ICU. In *Proceedings of the 25th ACM SIGKDD Inter-*

national Conference on Knowledge Discovery & Data Mining, KDD '19, pages 2961–2969, Anchorage, AK, USA, July 2019. Association for Computing Machinery.

[95] E Godoy, J López, L Bermúdez, A Ferrer, N García, C García Vicent, EF Lurbe, and J Saiz. Time-Domain, Frequency Domain and non-linear measurements in neonates' Heart Rate Variability with clinical sepsis. In *Computing in Cardiology 2014*, pages 429–432, September 2014.

[96] Jingchao Luo, Wei Jiang, Li Weng, Jinmin Peng, Xiaoyun Hu, Chunyao Wang, Guangyun Liu, Huibin Huang, and Bin Du. Usefulness of qSOFA and SIRS scores for detection of incipient sepsis in general ward patients: A prospective cohort study. *Journal of Critical Care*, 51:13–18, June 2019.

[97] Peter Haug and Jeffrey Ferraro. Using a Semi-Automated Modeling Environment to Construct a Bayesian, Sepsis Diagnostic System. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '16, pages 571–578, Seattle, WA, USA, October 2016. Association for Computing Machinery.

[98] Carolyn McGregor, Christina Catley, and Andrew James. Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. In *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–5, June 2012.

[99] M. Pamela Griffin, Douglas E Lake, Eric A Bissonette, Frank E Harrell, T. Michael O'Shea, and J. Randall Moorman. Heart Rate Characteristics: Novel Physiomarkers to Predict Neonatal Infection and Death. *Pediatrics*, 116(5):1070–1074, 2005.

[100] M. Pamela Griffin and J R Moorman. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics*, 107(1):97–104, 2001.

[101] M. Pamela Griffin, T. Michael O'Shea, Eric A. Bissonette, Frank E. Harrell, Douglas E. Lake, and J. Randall Moorman. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatric Research*, 53(6):920–926, 2003.

[102] J. Randall Moorman, Douglas E. Lake, and M. Pamela Griffin. Heart rate characteristics monitoring for neonatal sepsis. *IEEE Transactions on Biomedical Engineering*, 53(1):126–132, January 2006.

[103] M Pamela Griffin, Douglas E Lake, T Michael O'Shea, and J Randall Moorman. Heart Rate Characteristics and Clinical Signs in Neonatal Sepsis. *Pediatric Research*, 61(2):222–227, February 2007.

[104] M. P. Griffin, D. E. Lake, and J. R. Moorman. Heart Rate Characteristics and Laboratory

Tests in Neonatal Sepsis. *PEDIATRICS*, 115(4):937–941, April 2005.

[105] Vladan Radosavljevica, Kosta Ristovskia, and Zoran Obradovica. Gaussian Conditional Random Fields for Modeling Patients' Response to Acute Inflammation Treatment. page 8.

[106] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-Task Deep Neural Networks for Natural Language Understanding. *arXiv:1901.11504 [cs]*, May 2019.

[107] Saman Noorzadeh, Shahrooz Faghihroohi, and Mojtaba Zarei. A Comparative Analysis of HMM and CRF for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[108] Xin Li, G André Ng, and Fernando S Schlindwein. Convolutional and Recurrent Neural Networks for Early Detection of Sepsis Using Hourly Physiological Data from Patients in Intensive Care Unit. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[109] Matthieu Scherpf, Miriam Goldammer, Hagen Malberg, and Felix Gräßer. Sepsis Onset Prediction Applying a Stacked Combination of a Recurrent Neural Network and a Gradient Boosted Machine. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[110] Morteza Zabihi, Serkan Kiranyaz, and Moncef Gabbouj. Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[111] Simon Lyra, Steffen Leonhardt, and Christoph Hoog Antink. Early Prediction of Sepsis Using Random Forest Classification for Imbalanced Clinical Data. In *2019 Computing in Cardiology (CinC)*, pages 1–4, September 2019.

[112] Howard Hua Yang and John Moody. Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 687–693. MIT Press, 2000.

[113] P Biglarbeigi, D McLaughlin, K Rjoob, Abdullah Abdullah, N McCallan, A Jasinska-Piadlo, R Bond, D Finlay, KY Ng, A Kennedy, and J McLaughlin. Early Prediction of Sepsis Considering Early Warning Scoring Systems. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[114] Marco AF Pimentel, Adam Mahdi, Oliver Redfern, Mauro D Santos, and Lionel Tarassenko. Uncertainty-Aware Model for Reliable Prediction of Sepsis in the ICU. In *2019 Computing in Cardiology (CinC)*, pages 1–4, September 2019.

[115] Qiang Yu, Xiaolin Huang, Weifeng Li, Cheng Wang, Ying Chen, and Yun Ge. Us-

ing Features Extracted From Vital Time Series for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[116] Roshan Pawar, Jeffrey Bone, J Mark Ansermino, and Matthias Görges. An Algorithm for Early Detection of Sepsis Using Traditional Statistical Regression Modeling. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[117] Sven Schellenberger, Kilin Shi, Jan P Wiedemann, Fabian Lurz, Robert Weigel, and Alexander Koelpin. An Ensemble LSTM Architecture for Clinical Sepsis Detection. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[118] Ibrahim Hammoud, IV Ramakrishnan, and Mark Henry. Early Prediction of Sepsis Using Gradient Boosting Decision Trees with Optimal Sample Weighting. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[119] ByeongTak Lee, KyungJae Cho, Oyeon Kwon, and Yeha Lee. Improving the Performance of a Neural Network for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[120] John Anda Du, Nadi Sadr, and Philip de Chazal. Automated Prediction of Sepsis Onset Using Gradient Boosted Decision Trees. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[121] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, Junhui Zhang, and Xu Ma. Prediction of Sepsis from Clinical Data Using Long Short-Term Memory and eXtreme Gradient Boosting. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[122] Zhengling He, Xianxiang Chen, Zhen Fang, Weidong Yi, Chenshuo Wang, Li Jiang, Zhongkai Tong, Zhongrui Bai, Yueqi Li, and Yichen Pan. Early Sepsis Prediction Using Ensemble Learning with Features Extracted from LSTM Recurrent Neural Network. In *2019 Computing in Cardiology Conference*, December 2019.

[123] Soodabeh Sarafrazi, Rohini S Choudhari, Chiral Mehta, Himanshi K Mehta, Omid K Japalaghi, Jie Han, Kinjal A Mehta, Hyunyoung Han, and Patricia A Francis-Lyon. Cracking the "Sepsis" Code: Assessing Time Series Nature of EHR Data, and Using Deep Learning for Early Sepsis Prediction. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[124] Shivnarayan Patidar. Diagnosis of Sepsis Using Ratio Based Features. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[125] Shiyu Liu, Ming Lun Ong, Kar Kin Mun, Jia Yao, and Mehul Motani. Early Prediction of Sepsis via SMOTE Upsampling and Mutual Information Based Downsampling. In

2019 Computing in Cardiology (CinC), pages Page 1–Page 4, September 2019.

[126] Lakshman Narayanaswamy, Devendra Garg, Bhargavi Narra, and Ramkumar Narayanswamy. Machine Learning Algorithmic and System Level Considerations for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[127] James Morrill, Andrey Kormilitzin, Alejo Nevado-Holgado, Sumanth Swaminathan, Sam Howison, and Terry Lyons. The Signature-Based Model for Early Detection of Sepsis From Electronic Health Records in the Intensive Care Unit. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[128] Benjamin Roussel, Joachim Behar, and Julien Oster. A Recurrent Neural Network for the Prediction of Vital Sign Evolution and Sepsis in ICU. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[129] Soufiane Chami and Kouhyar Tavakolian. Early Prediction of Sepsis From Clinical Data Using Single Light-GBM Model. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[130] E Macias, G Boquet, J Serrano, JL Vicario, J Ibeas, and A Morel. Novel Imputing Method and Deep Learning Techniques for Early Prediction of Sepsis in Intensive Care Units. In *2019 Computing in Cardiology (CinC)*, pages 1–4, September 2019.

[131] Luan Tran, Manh Nguyen, and Cyrus Shahabi. Representation Learning for Early Sepsis Prediction. In *2019 Computing in Cardiology (CinC)*, pages 1–4, September 2019.

[132] Janmajay Singh, Kentaro Oshiro, Raghava Krishnan, Masahiro Sato, Tomoko Ohkuma, and Noriji Kato. Utilizing Informative Missingness for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages 1–4, September 2019.

[133] Jia Yao, Ming Lun Ong, Kar Kin Mun, Shiyu Liu, and Mehul Motani. Hybrid Feature Learning Using Autoencoders for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[134] Manmay Nakhashi, Anoop Toffy, P V Achuth, Lingaselvan Palanichamy, and C M Vikas. Early Prediction of Sepsis: Using State-of-the-art Machine Learning Techniques on Vital Sign Inputs. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[135] Reza Firoozabadi and Saeed Babaeizadeh. An Ensemble of Bagged Decision Trees for Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[136] Naoki Nonaka and Jun Seita. Demographic Information Initialized Stacked Gated Re-

current Unit for an Early Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages 1–4, September 2019.

[137] Tomas Vicar, Petra Novotna, Jakub Hejc, Marina Ronzhina, and Radovan Smisek. Sepsis Detection in Sparse Clinical Data Using Long Short-Term Memory Network with Dice Loss. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[138] Petr Nejedly, Filip Plesinger, Ivo Viscor, Josef Halamek, and Pavel Jurak. Prediction of Sepsis Using LSTM Neural Network With Hyperparameter Optimization With a Genetic Algorithm. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[139] Peter Doggart and Megan Rutherford. Randomly Under Sampled Boosted Tree for Predicting Sepsis From Intensive Care Unit Databases. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[140] Xiang Li, Yanni Kang, Xiaoyu Jia, Junmei Wang, and Guotong Xie. TASP: A Time-Phased Model for Sepsis Prediction. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[141] Luchen Liu, Haoxian Wu, Zichang Wang, Zequn Liu, and Ming Zhang. Early Prediction of Sepsis From Clinical Data via Heterogeneous Event Aggregation. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[142] Vytautas Abromavičius and Artūras Serackis. Sepsis Prediction Model Based on Vital Signs Related Features. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[143] Shailesh Nirgudkar and Tianyu Ding. Early Detection of Sepsis Using Ensemblers. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[144] Miquel Alfaras, Rui Varandas, and Hugo Gamboa. Ring-Topology Echo State Networks for ICU Sepsis Classification. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[145] Yale Chang, Jonathan Rubin, Gregory Boverman, Shruti Vij, Asif Rahman, Annamalai Natarajan, and Saman Parvaneh. A Multi-Task Imputation and Classification Neural Architecture for Early Prediction of Sepsis from Multivariate Clinical Time Series. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[146] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional Recurrent Imputation for Time Series. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Process-*

ing Systems 31, pages 6775–6785. Curran Associates, Inc., 2018.

[147] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, Hong Kong, China, June 2008. IEEE.

[148] Induparkavi Murugesan, Karthikeyan Murugesan, Lingeshwaran Balasubramanian, and Malathi Arumugam. Interpretation of Artificial Intelligence Algorithms in the Prediction of Sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, September 2019.

[149] Kurt Lewin. *Field Theory in Social Science: Selected Theoretical Papers*. Harper Torchbooks. Harper & Row, New York, 1st harper torchbook ed. edition, 1964.

[150] Fred H. Blum. Action Research–A Scientific Approach? *Philosophy of Science*, 22(1):1–7, January 1955.

[151] Gerald I. Susman and Roger D. Evered. An Assessment of the Scientific Merits of Action Research. *Administrative Science Quarterly*, 23(4):582, December 1978.

[152] Margareta Hult and Sven-Åke Lennung. Towards a Definition of Action Research: A Note And Bibliography. *Journal of Management Studies*, 17(2):241–250, May 1980.

[153] Michael Peters and Viviane Robinson. The Origins and Status of Action Research. *The Journal of Applied Behavioral Science*, 20(2):113–124, April 1984.

[154] F. Lau. A Review on the Use of Action Research in Information Systems Studies. In Allen S. Lee, Jonathan Liebenau, and Janice I. DeGross, editors, *Information Systems and Qualitative Research*, pages 31–68. Springer US, Boston, MA, 1997.

[155] CHRIS ARGYRIS and DONALD A. SCHÖN. Participatory Action Research and Action Science Compared: A Commentary. *American Behavioral Scientist*, 32(5):612–623, May 1989.

[156] Richard L Baskerville. Investigating Information Systems with Action Research. 2:33, 1999.

[157] Chris Argyris. Double loop learning in organizations. *Harvard business review*, 55(5):115–125, 1977.

[158] Herbert A. Simon. *The Sciences of the Artificial.* MIT press, 1996.

[159] Pertti Järvinen. Action Research is Similar to Design Science. *Quality & Quantity*, 41(1):37–54, February 2007.

[160] Alan R. Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.

[161] Salvatore T. March and Gerald F. Smith. Design and natural science research on information technology. *Decision Support Systems*, 15(4):251–266, December 1995.

[162] R Baskerville and AT Wood-Harper. Diversity in information systems action research methods. page 18, 1998.

[163] Paul Oquist. The Epistemology of Action Research. *Acta Sociologica*, 21(2):143–163, April 1978.

[164] Joan van Aken. Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological Rules. *Journal of Management Studies*, 41, February 2001.

[165] Evert Gummesson. *Qualitative Methods in Management Research*. Sage, 2000.

[166] Charles L Owen. Design research: Building the knowledge base. *Design Studies*, 19(1):9–20, January 1998.

[167] Paul Coughlan and David Coghlan. Action research for operations management. *Action research*, page 21, 2002.

[168] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design Science in Information Systems Research. page 33, 2004.

[169] Victor R Basili. The Role of Experimentation in Software Engineering: Past, Current, and Future. page 8.

[170] Allen Newel and Herbert A Simon. Completer Science asEmp rical Inquiry: Symbols and Search. 19(3):14, 1976.

[171] M. Lynne Markus, Ann Majchrzak, and Les Gasser. A Design Theory for Systems That Support Emergent Knowledge Processes. *MIS Quarterly*, 26(3):179–212, 2002.

[172] William M. Fox. Sociotechnical System Principles and Guidelines: Past and Present. *The Journal of Applied Behavioral Science*, 31(1):91–105, March 1995.

[173] Toomas Timpka, Cecilia Sjöberg, and Birgitta Svensson. The pragmatics of clinical hypermedia: Experiences from 5 years of participatory design in the MEDEA project. *Computer Methods and Programs in Biomedicine*, 46(2):175–186, February 1995.

[174] Harold G. Levine and Don Rossmoore. Diagnosing the Human Threats to Information Technology Implementation: A Missing Factor in Systems Analysis Illustrated in a Case Study. *Journal of Management Information Systems*, 10(2):55–73, September 1993.

[175] Evelyn Blennerhassett. RESEARCH REPORT: MANAGEMENT LEARNING GROUPS — A LESSON IN ACTION. *Journal of European Industrial Training*, 12(8):5–12, January 1988.

[176] Miroslaw Staron. *Action Research in Software Engineering: Theory and Applications*.

Springer International Publishing, Cham, 2020.

[177] Miroslaw Ochodek, Regina Hebig, Wilhelm Meding, Gert Frost, and Miroslaw Staron. Recognizing lines of code violating company-specific coding guidelines using machine learning: A Method and Its Evaluation. *Empirical Software Engineering*, 25(1):220–265, January 2020.

[178] Salman Azhar, Irtishad Ahmad, and Maung K. Sein. Action Research as a Proactive Research Method for Construction Engineering and Management. *Journal of Construction Engineering and Management*, 136(1):87–98, January 2010.

[179] Eirin Esaiassen, Jon Widding Fjalstad, Lene Kristine Juvet, John N. van den Anker, and Claus Klingenberg. Antibiotic exposure in neonates and early adverse outcomes: A systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, 72(7):1858–1870, July 2017.

[180] . . , 2016.

[181] Wolfgang Scholz. eDATA- Grabber : A Suite of Infinity Research Tools.

[182] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[183] Zheng Dai, Siru Liu, Jinfa Wu, Mengdie Li, Jialin Liu, and Ke Li. Analysis of adult disease characteristics and mortality on MIMIC-III. *PLOS ONE*, 15(4):e0232176, April 2020.

[184] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, page 6.

[185] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703 [cs, stat]*, December 2019.

[186] William Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

[187] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[188] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[189] Zhiguang Wang and Tim Oates. Imaging Time-Series to Improve Classification and Imputation. page 7.

[190] Johann Faouzi and Hicham Janati. Pyts: A Python Package for Time Series Classification. *Journal of Machine Learning Research*, 21(46):1–6, 2020.

[191] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA, June 2015. IEEE.

[192] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[193] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[194] Chusak Okascharoen, Sayomporn Sirinavin, Ammarin Thakkinstian, Dwip Kitayaporn, and Sarayut Supapanachart. A Bedside Prediction-Scoring Model for Late-Onset Neonatal Sepsis. *Journal of Perinatology*, 25(12):778–783, 2005.

[195] S. Amuchou Singh, Sourabh Dutta, and Anil Narang. Predictive clinical scores for diagnosis of late onset neonatal septicemia. *Journal of tropical pediatrics*, 49(4):235–239, 2003.

[196] Subramani Mani, Asli Ozdas, Constantin Aliferis, Huseyin Atakan Varol, Qingxia Chen, Randy Carnevale, Yukun Chen, Joann Romano-Keeler, Hui Nian, and Jörn-Hendrik Weitkamp. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *Journal of the American Medical Informatics Association*, 21(2):326–336, 2014.

[197] Nora Hofer, Eva Zacharias, Wilhelm Müller, and Bernhard Resch. An Update on the Use of C-Reactive Protein in Early-Onset Neonatal Sepsis: Current Insights and New Tasks. *Neonatology*, 102(1):25–36, 2012.

[198] Pak C Ng, Karen Li, Raymond P O Wong, Kit M Chui, Eric Wong, and Tai F Fok. Neutrophil CD64 Expression: A Sensitive Diagnostic Marker for Late-Onset Nosocomial Infection in Very Low Birthweight Infants. *Pediatric Research*, 51(3):296–303, March 2002.

[199] Efthalia Hotoura, Vasileios Giapros, Ageliki Kostoula, Polixeni Spyrou, and Styliani Andronikou. Pre-inflammatory Mediators and Lymphocyte Subpopulations in Preterm Neonates with Sepsis. *Inflammation*, 35(3):1094–1101, June 2012.

[200] Pak C Ng, Geng Li, Kit M Chui, Winnie C W Chu, Karen Li, Raymond P O Wong, Kai W Chik, Eric Wong, and Tai F Fok. Neutrophil CD64 Is a Sensitive Diagnostic Marker for Early-Onset Neonatal Infection. *Pediatric Research*, 56(5):796–803, November 2004.

[201] Axel R Franz, Karl Bauer, Andreas Schalk, Suzanne M Garland, Ellen D Bowman, Kerstin Rex, Calle Nyholm, Mikael Norman, Adel Bougatef, Martina Kron, Walter Andreas Mihatsch, and Frank Pohlandt. Measurement of Interleukin 8 in Combination With C-Reactive Protein Reduced Unnecessary Antibiotic Therapy in Newborn Infants: A Multicenter, Randomized, Controlled Trial. page 11, 2018.

[202] P C Ng, S H Cheng, K M Chui, T F Fok, M Y Wong, W Wong, R P O Wong, and K L Cheung. Diagnosis of late onset neonatal sepsis with cytokines, adhesion molecule, and C-reactive protein in preterm very low birthweight infants. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 77(3):F221–F227, November 1997.

[203] T. B. Newman, K. M. Puopolo, S. Wi, D. Draper, and G. J. Escobar. Interpreting Complete Blood Counts Soon After Birth in Newborns at Risk for Sepsis. *PEDIATRICS*, 126(5):903–909, November 2010.

[204] Iris Streimish, Matthew Bizzarro, Veronika Northrup, Chao Wang, Sara Renna, Nancy Koval, Fang-Yong Li, Richard Ehrenkranz, Henry Rinder, and Vineet Bhandari. Neutrophil CD64 with Hematologic Criteria for Diagnosis of Neonatal Sepsis. *American Journal of Perinatology*, 31(01):021–030, March 2013.

[205] H. Tolga Çelik, Oytun Portakal, Şule Yiğit, Gülşen Hasçelik, Ayşe Korkmaz, and Murat Yurdakök. Efficacy of new leukocyte parameters versus serum C-reactive protein, procalcitonin, and interleukin-6 in the diagnosis of neonatal sepsis: Leukocyte parameters in newborn sepsis. *Pediatrics International*, 58(2):119–125, February 2016.

[206] Brynne A. Sullivan and Karen D. Fairchild. Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock. *Seminars in Fetal and Neonatal Medicine*, 20(4):255–261, August 2015.

[207] Alan Jovic and Nikola Bogunovic. Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. *Artificial Intelligence in Medicine*, 51(3):175–186, March 2011.

[208] Karen D. Fairchild, Robert L. Schelonka, David A. Kaufman, Waldemar A. Carlo, John

Kattwinkel, Peter J. Porcelli, Cristina T. Navarrete, Eduardo Bancalari, Judy L. Aschner, M. Whit Walker, Jose A. Perez, Charles Palmer, Douglas E. Lake, T. Michael O'Shea, and J. Randall Moorman. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatric Research*, 74(5):570–575, November 2013.

[209] Douglas E. Lake, Karen D. Fairchild, and J. Randall Moorman. Complex signals bioinformatics: Evaluation of heart rate characteristics monitoring as a novel risk marker for neonatal sepsis. *Journal of Clinical Monitoring and Computing*, 28(4):329–339, 2014.

[210] Fredrick J. Bohanon, Amy A. Mrazek, Mohamed T. Shabana, Sarah Mims, Geetha L. Radhakrishnan, George C. Kramer, and Ravi S. Radhakrishnan. Heart rate variability analysis is more sensitive at identifying neonatal sepsis than conventional vital signs. *American Journal of Surgery*, 210(4):661–667, 2015.

[211] V. Bhandari and A. Narang. Thermoregulatory alterations as a marker for sepsis in normothermic premature neonates. *Indian pediatrics*, 29(5):571–575, May 1992.

[212] José Luis Leante-Castellanos, José M. Lloreda-García, Ana García-González, Caridad Llopis-Baño, Carmen Fuentes-Gutiérrez, José Ángel Alonso-Gallego, and Antonio Martínez-Gimeno. Central-peripheral temperature gradient: An early diagnostic sign of late-onset neonatal sepsis in very low birth weight infants. *Journal of Perinatal Medicine*, 40(5), January 2012.

[213] ALP - blood test: MedlinePlus Medical Encyclopedia. https://medlineplus.gov/ency/article/003470.htm.

[214] Danny Castro, Sachin M. Patil, and Michael Keenaghan. Arterial Blood Gas. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2021.

[215] What Is a Bilirubin Test? https://www.webmd.com/a-to-z-guides/bilirubin-test.

[216] Normal Calcium Levels: What is a high calcium level? Normal and High Calcium Level Symptoms, Treatment, Diagnosis - UCLA. https://www.uclahealth.org/endocrine-center/normal-calcium-levels.

[217] Serum Chloride. https://www.ucsfhealth.org/Medical Tests/003485.

[218] Fibrinogen. https://www.ucsfhealth.org/Medical Tests/003650.

[219] Blood glucose monitoring – Diabetes Australia. https://www.diabetesaustralia.com.au/living-with-diabetes/managing-your-diabetes/blood-glucose-monitoring/.

[220] Bicarbonate - Health Encyclopedia - University of Rochester Medical Center. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=b

[221] Hematocrit test - Mayo Clinic. https://www.mayoclinic.org/tests-

procedures/hematocrit/about/pac-20384728.

[222] Hemoglobin test - Mayo Clinic. https://www.mayoclinic.org/tests-procedures/hemoglobin-test/about/pac-20385075.

[223] Lactic acid test Information — Mount Sinai - New York. https://www.mountsinai.org/health-library/tests/lactic-acid-test.

[224] Serum Magnesium Test. https://www.ucsfhealth.org/Medical Tests/003487.

[225] Serum Phosphorus. https://www.ucsfhealth.org/Medical Tests/003478.

[226] Platelet Count. https://www.ucsfhealth.org/Medical Tests/003647.

[227] High potassium (hyperkalemia). https://www.mayoclinic.org/symptoms/hyperkalemia/basics-20050776.

[228] Activated Partial Thromboplastin Time (APTT) Test. https://www.webmd.com/a-to-z-guides/partial-thromboplastin-time-test.

[229] Troponin - Health Encyclopedia - University of Rochester Medical Center. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=tr

[230] WBC Count. https://www.ucsfhealth.org/Medical Tests/003643.

[231] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989.