

國立臺灣大學共同教育中心統計碩士學位學程



碩士論文

Master Program in Statistics

Center for General Education

National Taiwan University

Master Thesis

利用機器學習方法於 MIMIC 資料庫之早期敗血症預測

Early Prediction of Sepsis Using Machine Learning

Methods on MIMIC Database

吳承彥

Cheng-Yan Wu

指導教授: 周呈雲 博士

Advisor: Cheng-Ying Chou, Ph.D.

中華民國 111 年 7 月

July, 2022






摘要

敗血症是加護病房內嚴重疾病之一，此疾病發病後可能導致患者的高死亡率和多種併發症。由於不同的患者、生命特徵、敗血症標準和預測方法，敗血症的早期預測是具有挑戰性。本研究旨在通過機器學習算法和深度學習方法開發一種高準確率的早期敗血症預測模型，該模型可以提高敗血症的早期預測，藉此警示醫生那些未來可能發展成敗血症的病患，從而降低發病率和死亡率。

此研究所開發的模型分類結果顯示 XGBoost 和 CNN 預測模型在分類敗血症方面表現出很強的性能。在 MIMIC-III 資料庫中，使用 SIRS 標準和 XGBoost 模型在敗血症發病時的病患 AUROC 約為 0.876，發病前 8 小時的 AUROC 為 0.780。使用 qSOFA 標準在敗血症發病時的病患 AUROC 約為 0.942，發病前 8 小時的 AUROC 為 0.729。CNN 預測模型使用 SIRS 標準在敗血症發病時達到了 0.996 AUROC，在發病前 8 小時的 AUROC 值為 0.945。

在 MIMIC-IV 資料庫中，使用 SIRS 標準和 XGBoost 模型在敗血症發病時的病患 AUROC 約為 0.836，發病前 8 小時的 AUROC 為 0.902。使用 qSOFA 標準在敗血症發病時的病患 AUROC 約為 0.823，發病前 8 小時的 AUROC 為 0.737。CNN 預測模型使用 SIRS 標準在敗血症發病時達到了 0.992 的 AUROC，在發病前 8 小時的 AUROC 值為 0.917。

和前人做法不同的地方是我將一般的數值輸入轉換成圖像輸入，並且使用



圖像輸入比起數值輸入可以得到更好的分類效果。因此，相信 CNN 和 XGBoost 預測模型可以用於提前 8 小時預測敗血症發作。根據本研究的結果，CNN 和 XGBoost 預測模型可以使用八個特徵提前 8 小時準確預測敗血症發作。此外，僅使用八個特徵就獲得了這些高準確率的早期預測結果。總之，結果顯示 CNN 和 XGBoost 預測模型在敗血症的早期預測上可以得到很好的效果。

關鍵字：敗血症、早期預測、加護病房、機器學習、深度學習

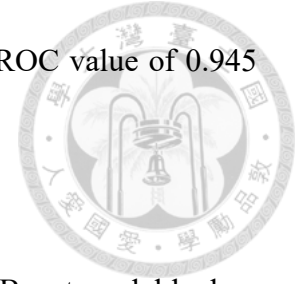


Abstract

Sepsis is one of the severe diseases which has high mortality, multiple complications, and cost overruns among patients treated in the intensive care unit (ICU). Because of variations in different patient cohorts, clinical variables, sepsis criterion, and prediction tasks, early clinical recognition of sepsis is challenging. This study aimed to develop a high-performance early sepsis prediction model by a machine learning algorithm and deep learning method that can improve the early detection of sepsis, thereby reducing morbidity and mortality.

The model classification results developed in this study show that the XGBoost and CNN prediction models exhibit strong performance in classifying sepsis. In the MIMIC-III dataset, subjects using the SIRS criterion and the XGBoost model had an AUROC of approximately 0.876 at the onset of sepsis and an AUROC of 0.780 eight hours before onset. Using the qSOFA criterion had an AUROC of 0.942 at the onset of sepsis and an AUROC of 0.729 eight hours before onset. The CNN prediction model achieved 0.996

AUROC at the onset of sepsis using the SIRS criterion and an AUROC value of 0.945 eight hours before the onset of sepsis.



In the MIMIC-IV dataset, using the SIRS criterion and the XGBoost model had an AUROC of 0.836 at the onset of sepsis and an AUROC of 0.902 eight hours before onset. Subjects using the qSOFA criterion had an AUROC of approximately 0.823 at the onset of sepsis and an AUROC of 0.737 eight hours before onset. Using the SIRS criterion, the CNN prediction model achieved an AUROC of 0.992 at the onset of sepsis and an AUROC value of 0.917 eight hours before the onset of sepsis.

According to the results, using eight features, the CNN and XGBoost prediction models could accurately predict sepsis onset up to eight hours in advance. Our model significantly outperformed previously existing ones. Furthermore, these high-accuracy early prediction results were obtained using only eight features. In summary, results demonstrated that the CNN and XGBoost prediction models could improve early sepsis detection.

Keywords: Sepsis, Early prediction, Intensive care unit, Machine learning, Deep learning



Contents

	Page
Verification Letter from the Oral Examination Committee	i
摘要	iii
Abstract	v
Contents	vii
List of Figures	ix
List of Tables	xi
Chapter 1 Introduction	1
Chapter 2 Literature Reviews	5
Chapter 3 Methods	9
3.1 Research Procedure	9
3.2 Sepsis Definition	10
3.3 Dataset	13
3.4 Patient Selection	14
3.5 Data Preprocessing	15
3.6 Heatmaps	20
3.7 Classification Models	24
3.7.1 Logistic regression	24

3.7.2	Long short-term memory	24
3.7.3	Extreme gradient boosting	25
3.7.4	Convolutional neural network	26
Chapter 4	Results	27
4.1	Patient Demographics	29
4.2	Numerical Model Results	30
4.2.1	MIMIC-III	30
4.2.2	MIMIC-IV	33
4.3	Heatmap Model Results	35
4.3.1	MIMIC-III	36
4.3.2	MIMIC-IV	38
4.4	Comparison	40
Chapter 5	Discussion	45
Chapter 6	Conclusion	47
	References	49





List of Figures

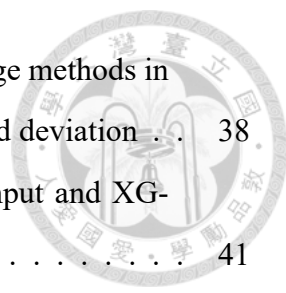
3.1	Flowchart of this study. The dashed box represented the machine learning model we used in this study.	11
3.2	Flowchart of patient selection in (a) MIMIC-III and (b) MIMIC-IV dataset.	15
3.3	Graphical explanation of setting nine time points.	18
3.4	An example of backfill method. The dot means the data exist in these time points.	19
3.5	A heatmap example from Bruno et al. [1].	20
3.6	Correlation coefficient matrix for eight features.	22
3.7	Hierarchical clustering for eight features.	23
3.8	A heatmaps example. The row represents the patient number. The column represents features.	23
4.1	Confusion matrix.	28
4.2	An example of ROC curve.	29
4.3	Bar plot of the heatmaps model testing result using CNN in MIMIC-III dataset.	37
4.4	Loss curves at three different time points.	38
4.5	Bar plot of the heatmaps model testing result using CNN in MIMIC-IV dataset.	39





List of Tables

2.1	Summary of previous studies on sepsis onset prediction.	7
3.1	SIRS and qSOFA scores are satisfied if more than two indicators are met [2, 3].	12
3.2	SOFA score is satisfied when one gets more than two points [4].	12
3.3	All feature IDs that need to be extracted from the database.	16
3.4	Filter criteria for features.	17
4.1	Demographics in MIMIC-III dataset and MIMIC-IV dataset. The value in the table represents the mean (standard deviation).	30
4.2	Patient numbers for whole clock prediction in MIMIC-III dataset.	31
4.3	Mean AUROC of three machine learning models at whole clock prediction in MIMIC-III dataset.	31
4.4	Patient numbers for different prediction intervals in MIMIC-III dataset.	32
4.5	Mean AUROC of three machine learning models at interval prediction in MIMIC-III dataset.	32
4.6	Patient numbers for whole clock prediction in MIMIC-IV dataset.	33
4.7	Mean AUROC of XGBoost model at whole clock prediction in MIMIC-IV dataset.	34
4.8	Patient numbers for different prediction intervals in MIMIC-IV dataset.	34
4.9	Mean AUROC of XGBoost model at interval prediction in MIMIC-IV dataset.	35
4.10	AUROC of CNN model results from different feature arrange methods in MIMIC-III dataset. Values in parentheses represent standard deviation	36



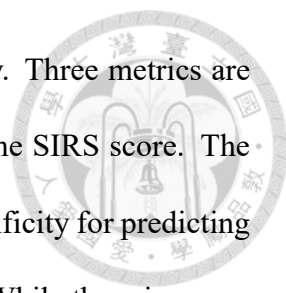
4.11 AUROC of CNN model results from different feature arrange methods in MIMIC-IV dataset. Values in parentheses represent standard deviation . . .	38
4.12 Comparison of the best training results using numerical input and XG-Boost model in the MIMIC-III and MIMIC-IV.	41
4.13 Comparison of the best training results using graphical input in the MIMIC-III and MIMIC-IV.	42
4.14 A comparison of my research with other researches.	43
5.1 The predicted features and the features used for the three scores. The bolded features overlap with the predicted features.	46



Chapter 1 Introduction

In both the intensive care unit and the general ward, sepsis is a disease with significant death and morbidity rates [5]. Sepsis typically results from an infection and can cause organ failure throughout the body. However, they might also be molds, viruses, or parasites [6]. Bacteria make up the majority of infectious pathogens. Lung, brain, urinary tract, skin, and abdominal organs are common primary infection sites in sepsis. Early detection and management of sepsis will lower patient mortality.

SIRS (systemic inflammatory response syndrome), SOFA (sequential organ failure assessment), and qSOFA (quick SOFA) are clinical criteria used to diagnose sepsis [2, 3]. Each of these judgment scores, however, has a disadvantage. The SIRS score looks for four vital signs to indicate organ failure. The SIRS score's high sensitivity and low specificity can identify critically ill patients in the ICU, but can easily misidentify those who do not have the disease [7, 8]. This could lead to a waste of medical resources. Aside from wasting medical resources, misdiagnosis of sepsis causes the doctor to use the incorrect treatment method, such as aggressive fluid administration, causing the patient to be treated inappropriately [9]. The SOFA score is calculated using six different markers, including laboratory test results and assessments of consciousness, as proof of organ failure. Although the SOFA score is more thorough than the SIRS score in determining organ failure, the process takes longer because laboratory test data are required. Addi-

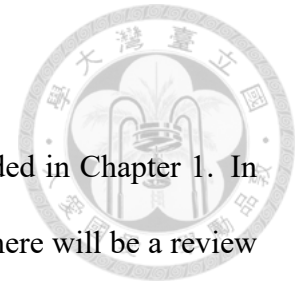


tionally, it can overlook the best opportunity to treat patients quickly. Three metrics are used in the qSOFA score, two from the SOFA score and one from the SIRS score. The time-consuming nature of assessing the SOFA score and its high specificity for predicting patient mortality are both issues resolved by the qSOFA score [10]. While there is no established, accepted standard, each of these three scores has advantages and disadvantages in diagnosing sepsis. The three criteria frequently employed in the clinical definition of sepsis were also utilized in this investigation as the sepsis diagnostic criteria.

Every minute and every second matter to the patient in the intensive care unit. A priceless life might be lost if the patient's prime time for treatment is missed. Numerous patient vital signs must be considered to identify sepsis because doctors cannot care for all patients simultaneously. So it is essential to create a high-performance machine learning model for early sepsis detection. The intended use of this machine learning model will be to build a warning system for monitoring patients' vital signs and laboratory data when admitted to the intensive care unit. This will be done with the help of the patients. Utilize this system to collect patients' data once they have been admitted to the ICU. Remind the doctor when sepsis may occur, and then ask the doctor to treat the patient as quickly as possible. This machine learning approach can lighten the load on ICU doctors. They are using it to foretell if a patient will experience sepsis and inform medical professionals so they can provide patients with the best care available as soon as feasible. In this study, the models were training using two datasets. This study also used numerical data, such as heartrate and blood pressure, and converted that data into heatmaps and deep learning techniques for early sepsis prediction. This study has also achieved high accuracy models utilizing deep learning and machine learning techniques. According to the model's test findings, if it performs well at a particular moment, it suggests that sepsis may be

accurately predicted.

A concise overview of the current condition of sepsis is provided in Chapter 1. In Chapter 2, where I will organize prior studies on sepsis prediction, there will be a review of the literature as well as a status report on the research that is currently being conducted. My study methodology is discussed in the third chapter, which covers topics such as the introduction of databases, patient screening, definition of sepsis, data preprocessing, machine learning algorithms employed, and the process of developing a model. In Chapter 4, the findings of the research are presented and compared. To explain my findings in their entirety, data visualization tools were used as well as tables. The results of my investigation will be discussed and concluded in Chapter 5. In Chapter 6, this study discussed potential developments in future research as well as the limits of this study.



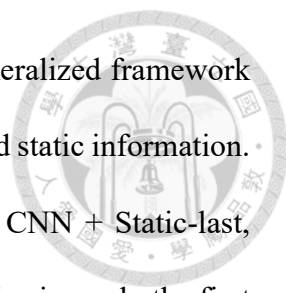




Chapter 2 Literature Reviews

Sepsis is one of the most deadly diseases in the intensive care unit because there is not a single biomarker that can be used to diagnose sepsis [4]. The relationship between different vital signs must be considered when diagnosing sepsis; therefore, sepsis is difficult to diagnose clinically and is both complicated and challenging. Vincent and Jean-Louis also pointed out the challenges of clinically judging sepsis in their study [11]. They pointed out that most studies have focused on changes in single indicators, and progress in diagnosing sepsis using biomarkers has been slow. This is partly because the sepsis response involves several players at different points during the disease process. However, it is essential to make a rapid diagnosis of sepsis to begin appropriate therapy promptly and provide patients with the best possible chance of survival.

In recent years, many machine learning approaches have been implemented in medicine due to advances in technology and software, as well as an increase in the use of big data. One of these ways is the early prediction of sepsis. Predictive machine learning approaches can be pretty helpful when looking at many data sets and identifying links between them. An LSTM-based model was suggested by Zhang et al. to diagnose sepsis utilizing the definition of Sepsis-2, which incorporates the SIRS score and probable infection [12]. Their research could accurately predict sepsis up to four hours before the condition manifested itself, and their model was able to attain AUROC of 0.84, which ensured both accurate



prediction and clinical applicability. Chen Lin et al. presented a generalized framework based on LSTM [13]. They used LSTM in combination with CNN and static information. Especially noteworthy is that their proposed architecture, LSTM + CNN + Static-last, achieved an AUC greater than 0.92 and an F1 Score greater than 0.85 using only the first three hours of the EHRs. Burdick et al. came up with a retrospective approach by combining the results of a Dascena analysis with the data from the Cabell Huntington Hospital dataset [14]. They determined the time to the onset of sepsis as the first hour in which two or more SIRS criteria were met along with at least one organ dysfunction criterion. This was the definition of the time to the onset of sepsis. The effectiveness of their strategy was assessed zero, four, six, twelve, twenty-four, and forty-eight hours before the start of sepsis. Their model exhibited a mean AUROC of 0.931 at the beginning of severe sepsis and 0.827 48 hours before the beginning. Table 2.1 provides a summary of the findings from the earlier studies, and compiled the results from 11 separate investigations. Most of the compiled information consists of retrospective studies that utilized medical databases and early sepsis diagnoses, much like this thesis. I found that these types of studies were very helpful. The findings of the final categorization can be affected by a wide range of circumstances, some of which include the databases consulted, the time of early prediction, and the definition of sepsis. Several of these studies, including mine, used the MIMIC database, which can be used for comparison. Although the procedures used are different and may not be suitable to be discussed together, the final results are compared to determine if there are significant differences between them.

Due to the difficulty and complexity of clinical diagnosis of sepsis, as well as the application of medical big data, the use of machine learning for early prediction of sepsis is emerging as a viable option, and the technology that is currently available for machine

Table 2.1: Summary of previous studies on sepsis onset prediction.

Authors	Dataset	Sepsis definition	Machine learning model	Hours before sepsis onset	AUROC
Zhang et al. [12]	Cerner Health Facts dataset	Sepsis-2	LSTM	4	0.84
Chen et al. [13]	Christiana Care Health System	Sepsis-3	LSTM based + CNN	3	0.92
Burdick et al. [14]	Dascena Analysis Dataset	Sepsis-3	MLA method	0	0.92
Futoma et al. [15]	Duke University Health System	Sepsis-2	MGP-RNN	0	0.91
Shashikumar et al. [16]	Emory healthcare system	Sepsis-3	ElasticNet	4	0.78
Guillen et al. [17]	MIMIC-II	Lactate concentration	SVM	2	0.87
Moor et al. [18]	MIMIC-III	Sepsis-3	MGP-TCN	0	0.91
Desautels et al. [19]	MIMIC-III	Sepsis-3	InSight	4	0.74
Nemati et al. [20]	MIMIC-III	Sepsis-3	Artificial Intelligence Sepsis Expert (AISE)	4	0.85
Barton et al. [21]	MIMIC-III	Sepsis-3	XGBoost	0	0.88
Scherpf et al. [22]	MIMIC-III	Sepsis-3	Recurrent neural network	3	0.81

learning can produce excellent early sepsis prediction results. These findings were found in several studies that were just discussed. On the other hand, most research does not validate their findings against other databases, which is necessary to generalize the proposed strategy. In my study, I used two datasets, MIMIC-III and MIMIC-IV, to validate my method and to make my method generalizable to different datasets. In addition, the majority of the study input is made up of numerical data. The features are first sorted according to different sorting methods and then normalized. After the feature arrangement and normalization, the data can be converted to different heatmaps and input into the models. In my research, this study provided a novel way to input graphified numerical data, which can also produce good results.





Chapter 3 Methods

3.1 Research Procedure

This research aims to develop a predictive model to predict sepsis early. The research process is shown in Figure 3.1, the dataset, machine learning method and data preprocessing process used in this study are explained in this figure. First, based on Burdick et al. [14] and collaborating physicians, this study selected eight predicted features age, systolic blood pressure, oxygen saturation measurement, diastolic blood pressure, heart rate, temperature, respiratory rate, and white blood cell count. Because diagnosing sepsis involves determining whether or not there is evidence of organ failure, the characteristics chosen for this study are indications of organ failure. Therefore, it is believed that using these features to predict sepsis can achieve good results. Some overlap between these eight traits and the three sepsis criteria may contribute to bias in my sepsis model. Because the onset time of sepsis is determined by different criteria, it is possible that the model using a score close to the predicted features could perform better. Fleuren et al. also classified the characteristics of sepsis prediction [23]. They analyzed 28 kinds of research and categorized each predictive feature's usage frequency. Heart rate, respiration rate, temperature, systolic blood pressure, oxygen saturation, white blood cell count, age, diastolic blood pressure, mean arterial blood pressure, and blood urea nitrogen are the top 10 features

ranking from the most frequent to least used. The eight features that have been chosen are within the top 10 features according to their feature usage frequency ranking. Thus it is appropriate to determine these features in predicting sepsis.



Then, these features were extracted according to preset nine-time points. I labeled the time as t_0 when the sepsis was presumed to have begun and then pushed it back by eight hours, giving a total of nine-time points ranging from t_0 to t_{-8} . After preprocessing the data, the numerical data were converted into heatmaps and were applied to deep learning methods. This was because this study aimed to use a different method to forecast sepsis and applied a graph input to produce good classification results instead of the initial numerical input. The machine learning methods are Logistic regression and XGBoost, and the deep learning method is a convolution neural network. The data were split into 80 % training set and 20% testing set. Because of utilizing the validation data to alter the model's parameters to get better training results, I split some of the data were intended for training into a separate set, which is referred to as the validation data. In the end, five-fold cross-validation method was performed to ensure that the model was reliable. I compared the testing results using two different datasets, three different sepsis-defining criteria, and three different machine learning approaches.

3.2 Sepsis Definition

Sepsis means bacteria invade the human body, and the immune capacity is insufficient, causing many bacteria to multiply in the blood and produce toxins [6]. This will damage the function of various human body organs and seriously lead to organ failure, hypotension, or shock state. Symptoms of sepsis include fever, chills, rapid heartbeat,

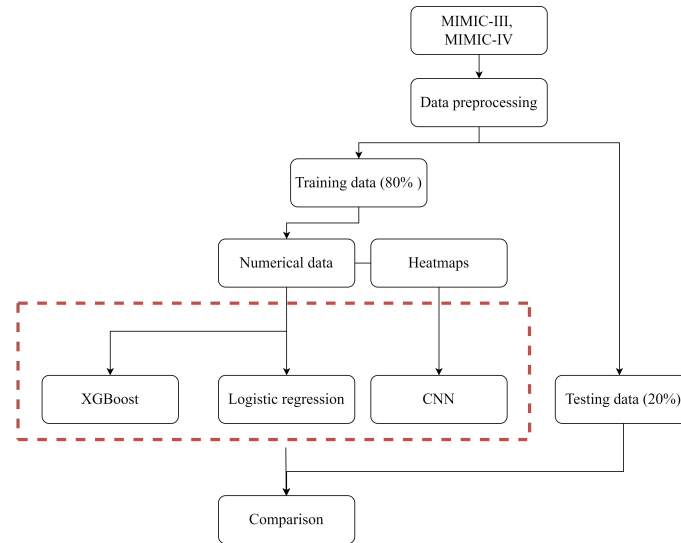


Figure 3.1: Flowchart of this study. The dashed box represented the machine learning model we used in this study.

shortness of breath, and confusion. The patient will have gastrointestinal, liver, and kidney dysfunction symptoms with systemic inflammatory response and thrombus formation everywhere. Without effective treatment, the patient will eventually begin to decrease urine output, lower blood pressure, and eventually lead to shock, or even death.

Sepsis is defined as an infection of the body plus organ failure. Three scores are commonly used to assess organ failure: SIRS, SOFA, and qSOFA. Table 3.1 and Table 3.2 show the detailed calculation process of these three scores. According to the Third International Consensus Conference on Definitions of Sepsis and Septic Shock in 2016, the Sepsis-3 criteria were proposed and the original 2001 definition of the Sepsis-2 criteria was abandoned [24, 25]. A patient with a simple infection was identified as having sepsis by Sepsis-2 in 2001. To determine whether or not the patient has organ failure, Sepsis-2 uses the SIRS score. Sepsis is identified in a patient when the SIRS score is greater than 2, and the patient is considered to have sepsis. The definition of sepsis according to Sepsis-3 in 2016 is the same as that according to Sepsis-2. This definition requires first determining whether the patient has a simple infection and then determining whether the patient has

organ failure. In contrast to Sepsis-2, Sepsis-3 uses both the qSOFA score and the SOFA score to determine whether or not organ failure has occurred.



Therefore, three different criteria for sepsis were used to determine the time of occurrence of sepsis. In the diagnostic criteria of Sepsis-3, the first thing is to confirm whether the patient is infected and then judge the evidence of organ failure according to the qSOFA score and SOFA score. In this research, I used the evidence of organ failure directly to evaluate sepsis without adding infection as a factor. When judging evidence of organ failure, I not only used the qSOFA score and SOFA score but also the SIRS score was used to believe the evidence of organ failure. I used these three scores to evaluate organ failure judgments and compare the results of the three scores applied to different models simultaneously.

Table 3.1: SIRS and qSOFA scores are satisfied if more than two indicators are met [2, 3].

Criteria	SIRS	qSOFA
Temperature ($^{\circ}C$)	< 36 or > 38	–
Heart rate (beats/min)	> 90	–
White blood cell count ($10^3/uL$)	< 4 or > 12	–
Respiratory rate (breaths/min)	> 20	≥ 22
Glasgow coma scale	–	≤ 14
Systolic blood pressure (mmHg)	–	≤ 100

Table 3.2: SOFA score is satisfied when one gets more than two points [4].

Criteria	0	1	2	3	4
Cardiovascular	MAP ≥ 70 mmHg	MAP < 70 mmHg	dopamine ≤ 5 $\mu\text{g/kg/min}$ dobutamine (any dose)	dopamine > 5 $\mu\text{g/kg/min}$ epinephrine ≤ 0.1 $\mu\text{g/kg/min}$ norepinephrine ≤ 0.1 $\mu\text{g/kg/min}$	dopamine > 15 $\mu\text{g/kg/min}$ epinephrine > 0.1 $\mu\text{g/kg/min}$ norepinephrine > 0.1 $\mu\text{g/kg/min}$
PaO ₂ /FiO ₂ [mmHg] (kPa)	≥ 400 (53.3)	< 400 (53.3)	< 300 (40)	< 200 (26.7) with respiratory support	< 100 (13.3) with respiratory support
Platelets $\times 10^3/uL$	≥ 150	< 150	< 100	< 50	< 20
Bilirubin (mg/dl) [$\mu\text{mol/L}$]	< 1.2 [< 20]	1.2-1.9 [20-32]	2.0-5.9 [33-101]	6.0-11.9 [102-204]	> 12.0 [> 204]
Creatinine (mg/dl) [$\mu\text{mol/L}$] (or urine output)	< 1.2 [< 110]	1.2-1.9 [110-170]	2.0-3.4 [171-299]	3.5-4.9 [300-440] (or < 500 ml/day)	> 5.0 [> 440] (or < 200 ml/day)
Glasgow coma scale	15	13-14	10-12	6-9	< 6

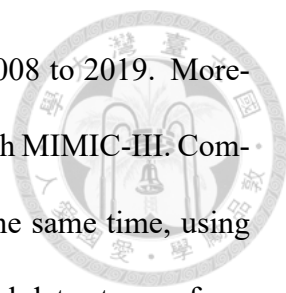
MAP, mean arterial pressure; PaO₂, partial pressure of oxygen.; FiO₂, fraction of inspired oxygen

3.3 Dataset



In this work, the Medical Information Mart for Intensive Care (MIMIC)-III and IV datasets were used [26, 27], which were built at the Beth Israel Deaconess Medical Center in Boston. The database had pre-de-identified patients to protect patient privacy. Before using this database, I must enter the CITI website and take the ethics exam. Before the exam, I needed to register an account and fill in my personal information. After completion, I needed to upload a certificate on physionet website to obtain permission to use the MIMIC database.

The MIMIC-III dataset collected more than 46,000 patients from 2001 to 2012. The data in MIMIC-III comes from two different ICU database systems: one is the CareVue database system [28], and the other is the MetaVision system. As the result, the same clinical data may correspond to multiple different IDs. Therefore, when extracting features, it must be noted that the features of both systems cannot be omitted. The data are stored in different CSV tables and there are 26 CSV tables. These tables detail the patients' data during ICU treatment, including vital signs, admission and discharge information, treatment process, nursing staff information, and more. Each table does not exist independently but was linked to other tables through different identifiers. The CHARTEVENTS table and the LABEVENTS table are the mainly used CSV tables. The CHARTEVENTS tables can extract important vital features such as heartrate, blood pressure, respiratory rate, and more. I needed to use these features when predicting sepsis and extract the features at nine time points. In the LABEVENTS table, laboratory test data were extracted, such as WBC count, which must be extracted from this table. The measurement time should also be extracted to help push back earlier features.



The MIMIC-IV set collected more than 60,000 patients from 2008 to 2019. Moreover, the MIMIC-IV set only uses the Metavison system compared with MIMIC-III. Compared with the confusion caused by using two database systems at the same time, using one database system can be more precise. It is newer, and the original data storage form has become more modular, which converts 26 original CSV tables into three modules: “hosp”, “core”, and “icu”. Module “hosp” contains laboratory measurement data, microbial cultures, medication administration, medication prescriptions, service-related data, supplier orders, and hospital billing information. Module “core” contains demographics, records of each admission, and records of ward admissions per admission. Module “icu” contains intravenous and fluid inputs, procedures, information documented such as a date or time, patient outputs, and other charted information. This makes data extraction easier. My methods were applied to these two datasets for result comparison.

3.4 Patient Selection

Patients in the MIMIC-III and MIMIC-IV datasets were screened using the same criteria. Adult patients 18 years of age or older with no missing data at t_0 to t_{-8} after imputation was the criterion employed in my research. Adults were selected since the average data of the vital signs of adults would be different from the data of kids. This was the reason behind why adults were chosen. Patients were screened to determine if they had sepsis based on the ICD-9 or ICD-10 codes provided by the database. Figure 3.2 illustrates my patient selection process for the MIMIC-III and MIMIC-IV datasets.

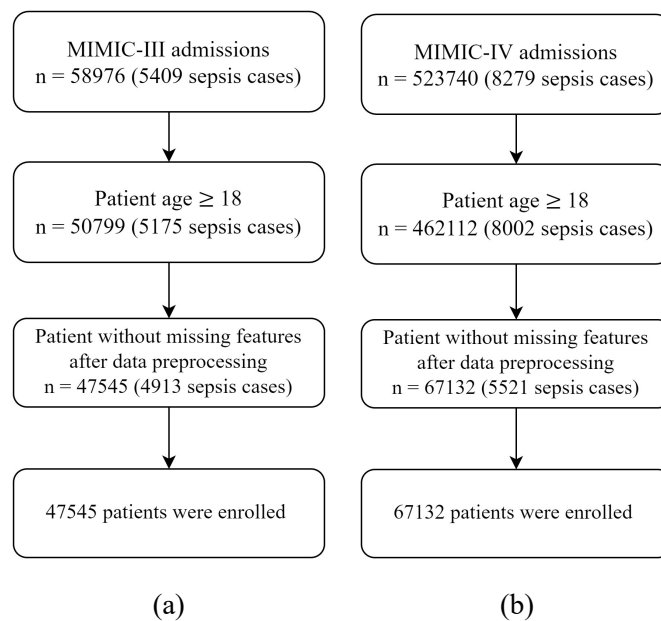


Figure 3.2: Flowchart of patient selection in (a) MIMIC-III and (b) MIMIC-IV dataset.

3.5 Data Preprocessing

This study selected age, systolic blood pressure, oxygen saturation measurement, diastolic blood pressure, heart rate, temperature, respiratory rate, and WBC as features to be included in the model. The features chosen from the datasets are based on Burdick et al. [14] and the physician’s professional clinical opinion. These vital signs are evidence that helps doctors in diagnosing sepsis. After selecting the features, they were from the database. In addition to filtering the predicted features, the features were used for calculating the SIRS score, SOFA score, and qSOFA score. The database medium was PostgreSQL. Different IDs were used to extract the features. All utilized characteristics are listed in Table 3.3. In addition, it is possible to deduce from the table that each measurement possesses more than one ID. This is because the MIMIC database uses not one but two different ICU database systems. The Carevue system and the Metavision system are the two that are utilized. It is important to remember that both systems’ IDs are included when extracting features from them.



Table 3.3: All feature IDs that need to be extracted from the database.

Measurement	CSV table	Item IDs
Heartrate	Chartevent	220045
Systolic blood pressure	Chartevent	220179, 220050
Diastolic blood pressure	Chartevent	220180, 220051
Respiratory rate	Chartevent	220210, 224609
Temperature	Chartevent	223761, 223762
SpO2	Chartevent	220277
White blood cell count	Chartevent	220546
Platelets	Labevent	51265
Bilirubin	Labevent	50885
Mean arterial pressure	Chartevent	456, 52, 6702, 443, 220052, 220181, 225312
Dopamine	Labevent	30043, 30307
Dobutamine	Labevent	30042, 30306
Epinephrine	Labevent	30044, 30119, 30309
Noepinephrine	Labevent	30047, 30120
GCS score	Chartevent	723, 454, 184 , 223900, 223901, 220739
Creatinine	Labevent	50912
Urine output	Ouputevent	40055, 43175

While extracting the data, some vital features are not reasonable, such as the heart rate of some patients being more significant than 300. Therefore, I had to set up some conditions to filter these variables which contained unreasonable values. Please refer to Table 3.4 for the criteria. The research conducted by Johnson et al. [26, 27, 29], who have made their code available on Github for use, served as the foundation for the establishment of these standards. The code's functionality covers the extraction of features, data preparation, handling inappropriate value ranges, and many other things. According to these criteria, more reasonable values could be left.

When trying to extract various features, I encountered numerous difficulties. The following paragraph described in further depth the solutions found to these issues and how to implement them because the database's de-identified actions for the patient and personal information were not provided in the database. When dealing with age, I have to calculate the age from the shifted dates. Because the database had the patient's processed

Table 3.4: Filter criteria for features.

Measurement	Filter criteria
Heartrate (bpm)	$0 \leq \text{value} \leq 300$
Systolic blood pressure (mmHg)	$0 \leq \text{value} \leq 400$
Diastolic blood pressure (mmHg)	$0 \leq \text{value} \leq 300$
Respiratory rate (breaths/min)	$0 \leq \text{value} \leq 70$
Temperature ($^{\circ}\text{F}$)	$70 \leq \text{value} \leq 120$
Temperature ($^{\circ}\text{C}$)	$10 \leq \text{value} \leq 50$
SpO2 (mmHg)	$0 \leq \text{value} \leq 100$
Mean arterial pressure (mmHg)	$0 \leq \text{value} \leq 300$
GCS eye	$0 \leq \text{value} \leq 5$
GCS motor	$0 \leq \text{value} \leq 6$
GCS verbal	$0 \leq \text{value} \leq 5$
FiO2 (mmHg)	$20 \leq \text{value} \leq 100$



date of birth and the processed date of the first admission, I could determine the patient's age during their first visit by subtracting the two dates and using the difference as the starting point. After doing the math, some of the patients aged more than 300 years old, which is quite peculiar. As described in MIMIC database's introduction, this group was older than the average and the median age of 91.4. As a result, the median age was used to replace the whole age group of patients who appeared to be more than 300 years old.

In the context of the GCS index, the MIMIC database does not directly supply this feature; hence, it needed to be derived. Eye-opening, verbal response, and motor reaction are the three components that make up the GCS index. The eye-opening component is the most important. A separate measured score is assigned to each of the indicators' scores. The patient's condition is considered more serious when the score is lower. The GCS index can be obtained by first obtaining these three indicators, then calculating the sum of those three indicators, and finally obtaining the GCS index. Because the database contains two different ICU versions, the body temperature records will contain Fahrenheit and Celsius units. I converted all body temperature readings into Fahrenheit to standardize the unit.

There was a problem when calculating the cumulative urine volume for 24 hours,

and it was not possible to obtain a whole day's worth of urine data. This was because the time interval between the urine output records was different, which caused a problem when calculating the cumulative urine volume for 24 hours. Following a conversation with the attending physician, the recommended solution was to determine the amount of urine output by taking the average of the recorded time intervals. For instance, if a patient's total urine output from 15:00 to 18:00 is 300 ml, that is, 300 ml/3 hr, I would average the patient's urine volume during these three hours, and the resultant value would be 100 ml/hr rather than the original 300 ml/3 hr. After completing the transformation, the whole 24-hour urine data was based on each hour's data.

After sorting out the unreasonable numerical data, nine-time points were preset to train different prediction models. The onset time was denoted as t_0 according to the definition of sepsis. After determining t_0 , eight features were extracted mentioned above at eight-time points, i.e., 1 to 8 hours before the onset of sepsis, and which was t_{-1} to t_{-8} . Different time points also corresponded to different numbers of patients, so the model training results at different times would not be the same. At nine-time points, the training results of different models were compared. The visualization of the preset nine-time points is shown in Figure 3.3, and the features are extracted at these nine-time points.

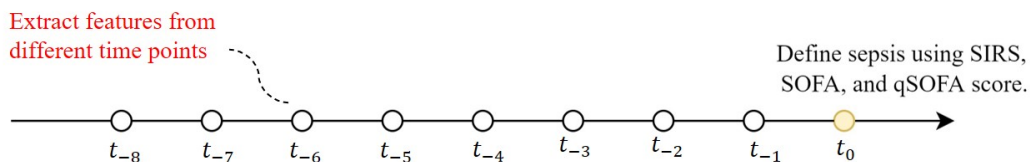


Figure 3.3: Graphical explanation of setting nine time points.

The data are stored initially in the form of long data, and there will be missing values when being converted to a wide data format. Because the time intervals of each variable record are different, there will be missing values of some time points when sorting vari-

ables from different time points to uniformly sampled time points. The backward filling method propagated the most recent value backwardly to a time point to facilitate the subsequent classification task. Using this method of filling missing values, each patient's vital characteristics can be filled by their previous data and would not be affected by outliers or other patient values. For example, if a patient's heart rate is null at a certain point, I would take the patient's most recent measurement to fill in, which is the most recent information the doctor can know about the patient at that time. Assuming that 15:00 is identified as t_0 , Figure 3.4 depicts the real procedure following this approach. It is clear that the values of blood pressure and body temperature are not available at this time points. Therefore, to fill in the missing number at 15:00, I would utilize the blood pressure reading taken at 13:00 and the temperature reading taken at 12:00. The filled information represented the latest information known at a specific time.

Time	Heartrate	Blood pressure	Temperature
2134/5/12 10:00	●	●	●
2134/5/12 11:00	●	●	●
2134/5/12 12:00	●		●
2134/5/12 13:00	●	●	↓
2134/5/12 14:00	●	↓	↓
2134/5/12 15:00	●	↓	↓

Figure 3.4: An example of backfill method. The dot means the data exist in these time points.

After the data pre-processing, clean data will be obtained. The content of the data is that there are eight features with no missing values at nine-time points, including age, systolic blood pressure, oxygen saturation measurement, diastolic blood pressure, heart rate, temperature, respiratory rate, and white blood cell count. I then applied different machine learning methods to these data to predict sepsis and compared the different models' results.



3.6 Heatmaps

Bruno and Calimeri provided evidence to show that applying heatmaps to the classification of tumors might produce valuable results [1]. They wanted to classify the disease by judging the information provided by DNA microarrays and gene expression profiles. They employed data on breast disease, mammography mass data, Parkinson’s disease, breast or kidney disease, and lymphoma data in their study. Principal component analysis was utilized in their research because determining which genes contribute to the categorization is an important task. As a result, dimension deduction was accomplished through the use of this technique. Because there are just eight predictive indicators included in my research, dimension deduction is not necessary. Figure 3.5 is an example of a heatmap that was used in their research. The purpose of using heatmaps was to depict the expression level of numerous genes or characteristics across several comparable samples. In addition to that, they classified the heatmaps using techniques such as deep neural networks and convolution neural networks.

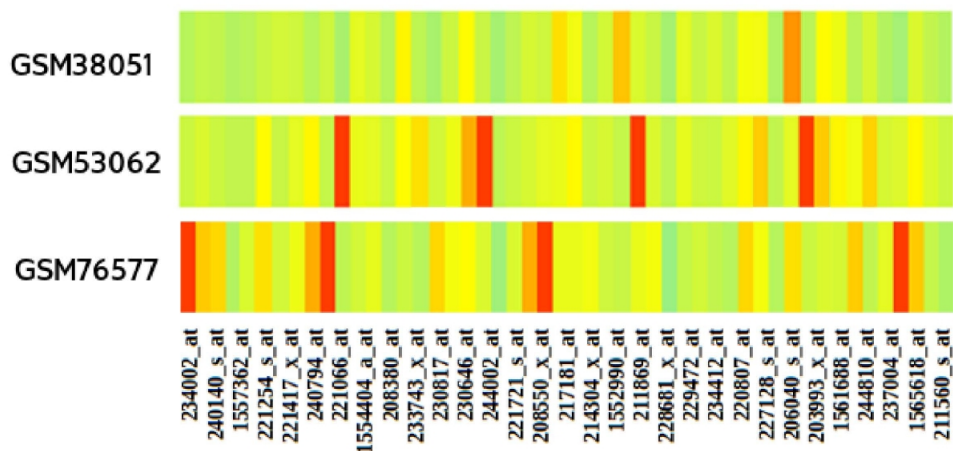
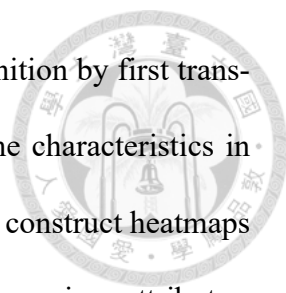


Figure 3.5: A heatmap example from Bruno et al. [1].

Because of the success of their study, this study endeavored to use the benefits of



convolutional neural networks in feature extraction and object recognition by first transforming the numerical data into visuals. Beginning by arranging the characteristics in random order to generate a heatmap from my data. After that, I could construct heatmaps at various time periods once the data had been normalized following the various attributes. For instance, if there were data from 100 patients at t_0 , the data were translated into one hundred individual heatmaps. The figures on my heatmaps and in the numerical data are comparable to one another. In addition, various techniques were utilized to organize the characteristics to investigate whether or not the order in which the features are listed impacted the categorization findings. They were organized in one of three ways: (1) in a random order, (2) according to the pairs of correlation coefficients, and (3) using hierarchical clustering.

After that, I will proceed to provide an in-depth explanation of the three different classification procedures. The first method was called random arrange, and it created a heat map by randomly arranging the features that make up the map, then combining and sorting those features in random order. The models were trained with 30 combinations of randomly sorted features and calculated the average AUROC using those results. The second approach was called paired correlation coefficients. I determined the correlation coefficient between each pair of characteristics, and then ranked the correlation coefficients from lowest to highest in magnitude. The correlation coefficient matrix can be found in Figure 3.6. Then, the feature sorting option was based on the correlation coefficients. The order of the sorted heatmap is as: DBP, SBP, age, heartrate, temperature, SpO2, respiratory rate, and WBC. Hierarchical clustering [30], which can also be referred to as hierarchical cluster analysis, is the final approach. This method is an algorithm that groups comparable items together that are referred to as clusters. The endpoint is a collec-

tion of clusters, each distinct from the others. The hierarchical clustering method was used to determine whether there is a clustering phenomenon between the features. I utilized the Euclidean distance and the complete clustering methods for my parameter setting. The hierarchical clustering method organized the related coefficients in the same cluster by looking at Figure 3.7. Because of this, this order was employed to build the heatmap. DBP comes first, followed by SBP, then respiratory rate, heartrate, temperature, SpO2, white blood cell count, and age.

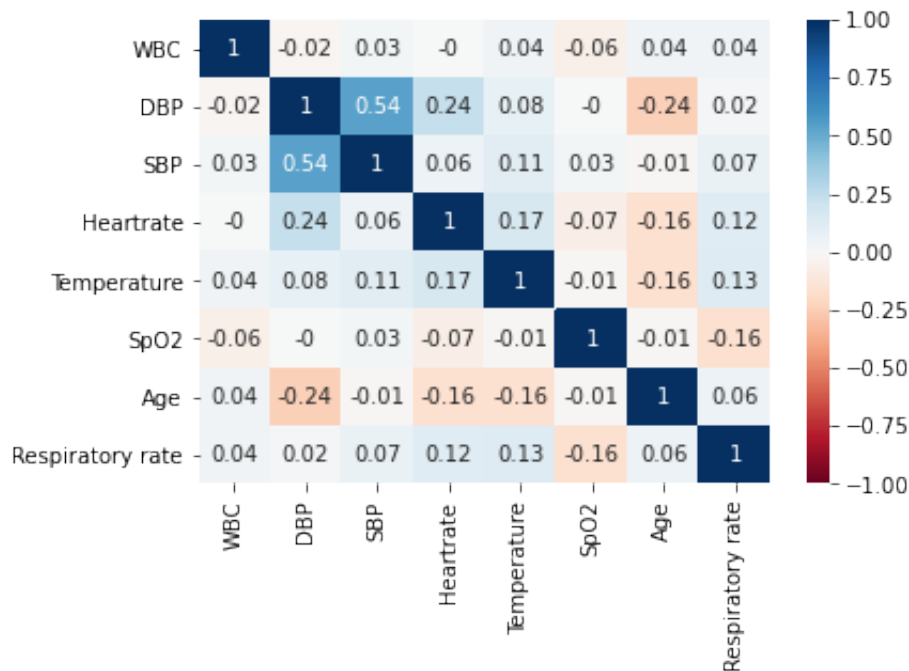


Figure 3.6: Correlation coefficient matrix for eight features.

Heatmaps were built using the normalized eight life qualities, and the results of establishing the heatmaps are depicted in Figure 3.8, respectively. The number serves as the graphic's vertical axis, while the various life characteristics serve as the graphic's horizontal axis. These life characteristics include age, systolic blood pressure, oxygen saturation measurement, diastolic blood pressure, heart rate, temperature, respiratory rate, and white blood cell count. Since it is difficult to tell with the naked eye whether the picture in this chapter represents the data of sepsis patients or ordinary individuals, a deep learning

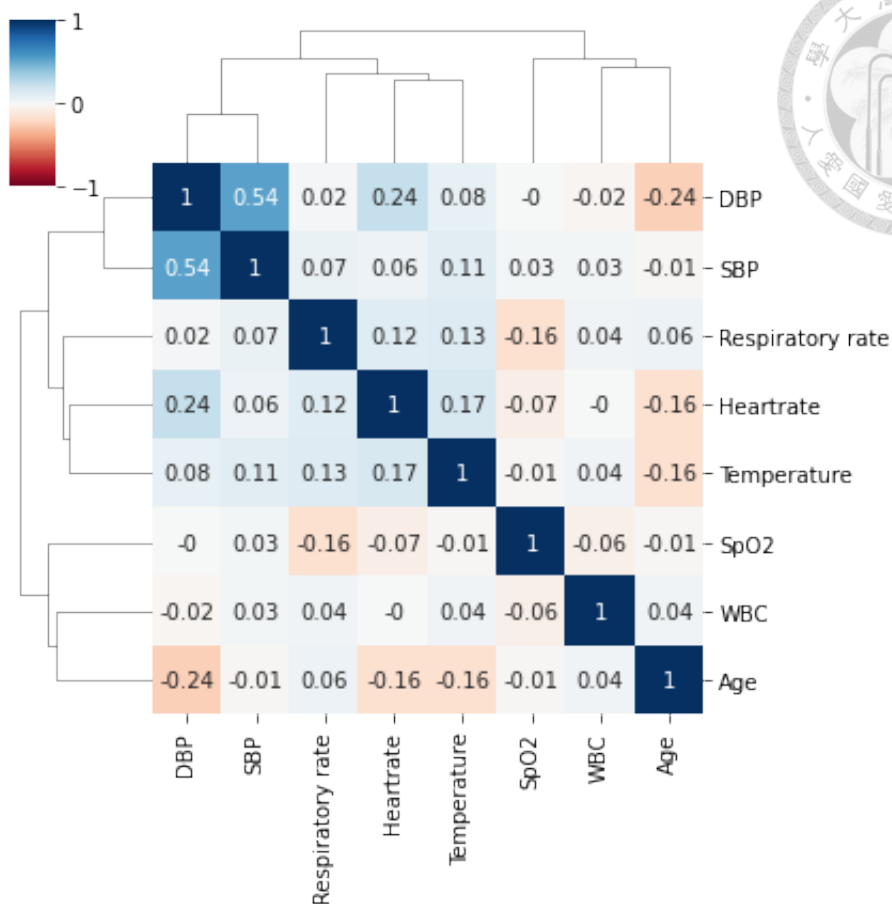


Figure 3.7: Hierarchical clustering for eight features.

method such as CNN was applied to categorize the heatmap.

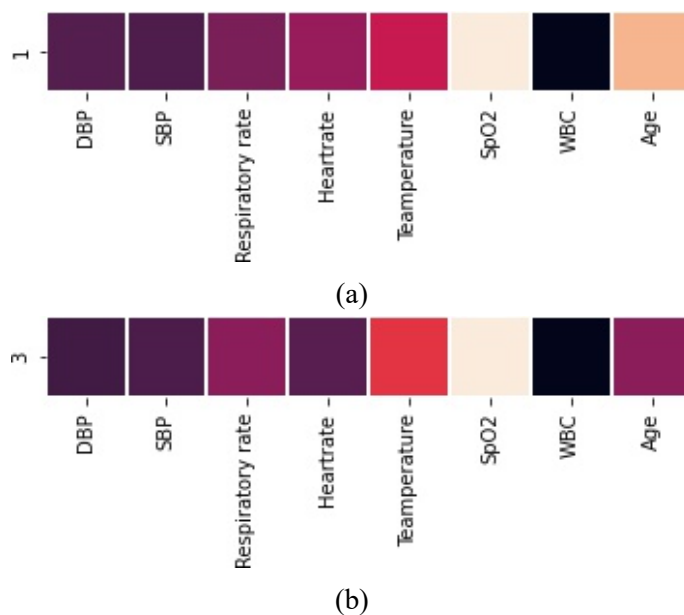


Figure 3.8: A heatmaps example. The row represents the patient number. The column represents features.



3.7 Classification Models

3.7.1 Logistic regression

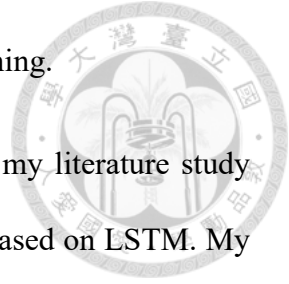
A classification model similar to linear regression is called logistic regression [31]. The purpose is to find a straight line referred to as a linear regression classifier that can separate and categorize all data. The primary focus of logistic regression is to investigate the connection between dependent and independent variables. In linear regression, the dependent variables are often continuous variables; however, in logistic regression, the dependent variables being examined are predominantly categorical variables, particularly ones that may be separated into two groups. In the context of this study, the objective was to use the electronic health record of a patient to make a binary categorization of a patient as having sepsis or not. The logistic regression approach can assist us in finishing the sepsis classification task while also allowing us to compare it to other machine learning methods.

3.7.2 Long short-term memory

Long Short-Term Memory [32], LSTM refers to a type of recurrent neural network (RNN) that was initially introduced in 1997 [33]. The one-of-a-kind architectural structure of the LSTM makes it well suited for the processing and forecasting of significant events in time series that are separated by very long intervals and delays. In general, LSTM solves various issues present in earlier RNN models, and LSTM is made up of four different units: an input gate, an output gate, a memory cell, and a forget gate. The amount of content that has been added has resulted in an increase in the number of parameters, as

well as a significant rise in the level of difficulty associated with training.

The usage of LSTM was addressed to generate predictions in my literature study [12, 13], and their researches were focused on making adjustments based on LSTM. My research was to review the literature. Since it is clear that they can also obtain good results in the early categorization of sepsis, I intend to use LSTM as one of the machine learning methods when comparing my results with those of others.



3.7.3 Extreme gradient boosting

The most popular method in Kaggle contests and the model utilized by most victors is Extreme Gradient Boosting, or XGBoost [34]. University of Washington Ph.D. candidate suggested this machine learning model Chen Tianqi [34]. Since XGBoost is a boosted tree model that combines many tree models to create a robust classifier, it is based on gradient boosting and incorporates a few novel approaches. One may argue that it combines the benefits of boosting and bagging. To ensure that the development of each decision tree is connected, XGBoost maintains gradient boosting. The objective is to anticipate that the later-generated tree will be able to fix the errors of the earlier trees. Additionally, XGboost employs the method of random feature sampling. It generates each tree by randomly extracting characteristics, similar to a random forest [35]. Therefore, not all traits are considered while making decisions during the development of each tree. Additionally, XGboost standardizes the objective function to further complicate the model. Because the model will produce several high-order functions to suit the training data, it is susceptible to noise and can overfit.

3.7.4 Convolutional neural network



Convolutional Neural Network, or CNN for short, is a highly effective tool for image identification [36–38]. Many image recognition models are also expanded based on the CNN architecture, and CNN is the foundation for many of these models. It is also important to point out that CNN is a deep learning model developed by referring to how the human brain processes visual information. To put it simply, the image pass after the convolution layer, pooling layer, and fully connected layer is the architecture of CNN.

The convolution layer primarily accomplishes the transformation from point comparison to local comparison by executing convolution operations on the input picture carried out by various kernels. It can get superior outcomes by researching the features of each block, making judgments about those features, and then progressively stacking the findings of extensive comparisons. The feature map is the outcome of the identification process, which is the picture obtained after the convolution. When extracting features from the image, the scale of the network should not affect the objective I am trying to achieve, which is the primary idea behind the pooling layer. By scaling the problem in this way, I may further minimize the number of parameters used by the neural network. To put it simply, the fully connected layer is a classifier that sorts the data into categories after they have been subjected to several convolutions and pooling operations.



Chapter 4 Results

During my investigation, the training procedure was validated by employing five-fold cross-validation and calculating to determine AUROC, sensitivity, and specificity. After drawing the ROC curve, AUROC can be calculated. The purpose of calculating an ROC curve is to analyze the degree to which the true positive rate and the false positive rate shift in response to alterations in the decision threshold. Figure 4.1 shows the definition of a confusion matrix. The confusion matrix was built by four indicators; true positive (TP), true negative (TN), false negative (FN), and false positive (FP). The calculation method of true positive rate (TPR) and false positive rate (FPR) is explained as

$$\text{TPR} = \frac{TP}{(TP + FN)} \quad (4.1)$$

and

$$\text{FPR} = \frac{FP}{(FP + TN)}. \quad (4.2)$$

Figure 4.2 offers an example of an ROC curve. The ROC curve has a certain area that it covers, and this area is referred to as the AUROC. The computation of the AUROC will be immediately followed by the determination of the sensitivity and specificity of the ideal cutoff point. Eq. (4.3) to Eq. (4.5) explain how to calculate accuracy, sensitivity and

specificity.

$$\text{Accuracy} = \frac{TN + TP}{(TN + TP + FN + FP)}, \quad (4.3)$$



$$\text{Sensitivity} = \frac{TP}{(TP + FN)}, \quad (4.4)$$

and

$$\text{Specificity} = \frac{TN}{(TN + FP)}. \quad (4.5)$$

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 4.1: Confusion matrix.

To facilitate comparison, it will be broken down into the following three subsections. The outcomes of applying the machine learning method to the MIMIC-III and MIMIC-IV datasets and three different sepsis judgment criteria will be shown in the two subsections in Sec. 4.2 and Sec. 4.3. In Sec. 4.4, there is a discussion of a detailed comparison of various datasets.

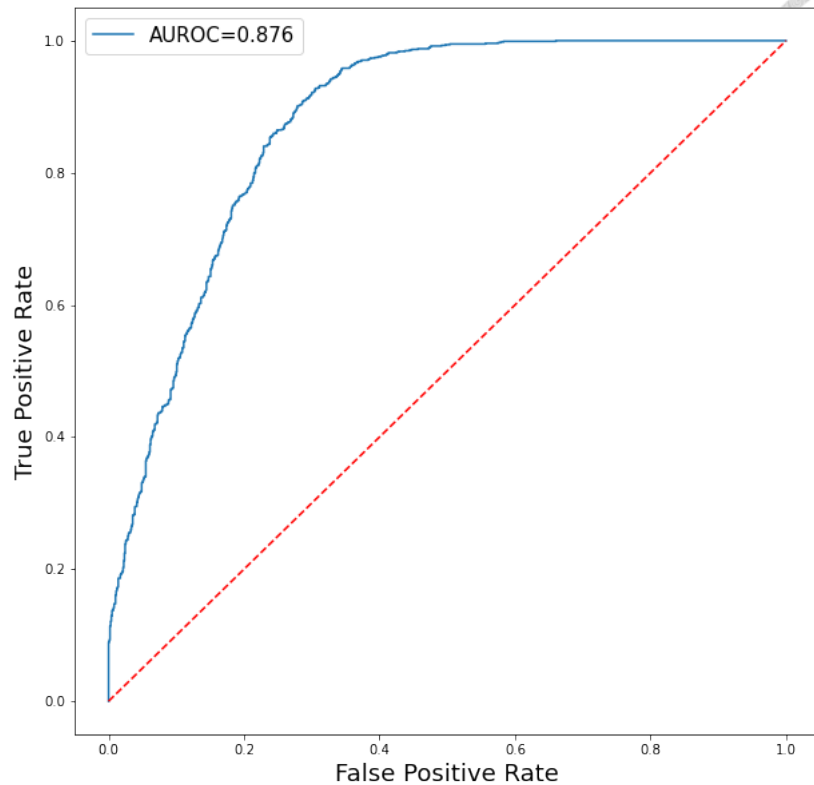


Figure 4.2: An example of ROC curve.

4.1 Patient Demographics

The results of my demographic calculations, which can be found in Table 4.1, were compiled from the MIMIC-III and MIMIC-IV repositories. The eight features used for model prediction are listed in Table 4.1. These features are as follows: age, systolic blood pressure, oxygen saturation measurement, diastolic blood pressure, heart rate, temperature, respiration rate, and white blood cell count. According to the ICD code, the patients were separated into those who had sepsis and those who did not have the condition to determine whether or not there was a significant difference between the two groups based on the data. From Table 4.1, there was not much difference between the two databases regarding the mean values of the patient attributes. The mean values of vital signs did not differ much between the septic and non-septic patients in the same dataset.

Table 4.1: Demographics in MIMIC-III dataset and MIMIC-IV dataset. The value in the table represents the mean (standard deviation).

Measurement	MIMIC-III		MIMIC-IV	
	Non-septic patient (n=42,632)	Septic patient (n=4,913)	Non-septic patient (n=61,612)	Septic patient (n=5,522)
Age	64.67 (19.30)	74.13 (20.41)	64.77 (16.25)	65.41 (16.02)
Diastolic blood pressure	62.08 (15.2)	60.11 (17.23)	64.25 (15.68)	60.02 (14.88)
Systolic blood pressure	122.63 (24.68)	116.11 (22.54)	120.43 (22.76)	114.68 (22.28)
Heartrate	87.72 (18.04)	92.08 (17.56)	85.91 (18.55)	92.12 (18.81)
Temperature	91.32 (19.82)	94.64 (20.81)	98.44 (1.08)	98.45 (1.39)
SpO2	99.43 (3.92)	96.91 (2.13)	96.69 (3.53)	96.85 (3.96)
Respiratory rate	21.37 (7.09)	20.45 (6.52)	19.87 (5.91)	21.43 (6.36)
White blood cell count	12.49 (8.21)	13.41 (7.79)	11.52 (8.29)	13.36 (9.51)

This may present a hurdle in constructing models because the vital signs of sepsis and non-septic sickness are very similar. There was no discernible difference between the patients regarding their essential characteristics.

4.2 Numerical Model Results

4.2.1 MIMIC-III

In the MIMIC-III dataset, this study included a total of 47,545 patients, and then retrieved varying numbers of patients based on the various sepsis diagnostic criteria and nine distinct time points (t_0 to t_{-8}). Table 4.2 displays the number of patients suffering from sepsis at various times. I chose an equal number of non-sepsis patients and patients with sepsis before including them in the model so that the total number of patients would be proportional. This means there would be an equal number of sepsis and non-sepsis patients. Table 4.3 shows the results of different models using three sepsis criteria, three machine learning models, and nine different time points. After doing five-fold cross-validation, the average AUROC findings are presented in Table 4.3.



Table 4.2: Patient numbers for whole clock prediction in MIMIC-III dataset.

Time points	SIRS	SOFA	qSOFA
t_0	4,915	1,028	4,064
t_{-1}	677	190	83
t_{-2}	555	155	261
t_{-3}	480	132	212
t_{-4}	432	121	1,396
t_{-5}	385	109	151
t_{-6}	341	104	211
t_{-7}	314	95	127
t_{-8}	288	89	903

Table 4.3: Mean AUROC of three machine learning models at whole clock prediction in MIMIC-III dataset.

Time points	SIRS			SOFA			qSOFA		
	XGBoost	Logistic Regression	LSTM	XGBoost	Logistic Regression	LSTM	XGBoost	Logistic Regression	LSTM
t_0	0.876	0.704	0.633	0.757	0.612	0.645	0.942	0.712	0.733
t_{-1}	0.798	0.617	0.611	0.670	0.683	0.612	0.652	0.687	0.641
t_{-2}	0.791	0.622	0.596	0.613	0.643	0.675	0.707	0.643	0.696
t_{-3}	0.809	0.674	0.661	0.647	0.672	0.616	0.673	0.662	0.561
t_{-4}	0.822	0.627	0.638	0.543	0.621	0.628	0.791	0.617	0.638
t_{-5}	0.822	0.682	0.556	0.591	0.543	0.568	0.754	0.643	0.556
t_{-6}	0.877	0.711	0.567	0.537	0.577	0.537	0.616	0.677	0.577
t_{-7}	0.835	0.645	0.641	0.684	0.589	0.541	0.793	0.579	0.541
t_{-8}	0.780	0.642	0.648	0.679	0.674	0.548	0.729	0.574	0.548

Furthermore, to see the results of the model training on a larger scale, this study attempted to incorporate additional patients by either adding or deducting ten minutes from the starting time of the hour. For instance, if the first forecast time was 120 minutes ago, then the interval prediction will become 110 minutes to 130 minutes. And the patient number was shown in Table 4.4. This will expand the number of samples and improve the model's capacity for learning new information. Table 4.5 shows the average AUROC results of the three machine learning models.

During the process of MIMIC-III XGBoost numerical model training, to prevent the issue of overfitting and ensure that the model is not improperly trained, an early stop was



Table 4.4: Patient numbers for different prediction intervals in MIMIC-III dataset.

Time points	SIRS	SOFA	qSOFA
t_0	9,295	1,892	4,066
t_{-1}	1,023	321	85
t_{-2}	785	222	266
t_{-3}	683	189	214
t_{-4}	607	186	1,400
t_{-5}	540	167	153
t_{-6}	464	152	213
t_{-7}	414	145	129
t_{-8}	379	133	906

Table 4.5: Mean AUROC of three machine learning models at interval prediction in MIMIC-III dataset.

Time points	SIRS			SOFA			qSOFA		
	XGBoost	Logist Regression	LSTM	XGBoost	Logist Regression	LSTM	XGBoost	Logist Regression	LSTM
t_0	0.856	0.721	0.623	0.747	0.672	0.615	0.912	0.752	0.753
t_{-1}	0.768	0.717	0.619	0.660	0.638	0.612	0.712	0.687	0.651
t_{-2}	0.751	0.634	0.696	0.615	0.631	0.676	0.717	0.613	0.712
t_{-3}	0.819	0.671	0.561	0.617	0.673	0.616	0.673	0.663	0.662
t_{-4}	0.833	0.637	0.538	0.613	0.631	0.638	0.791	0.617	0.654
t_{-5}	0.813	0.683	0.666	0.591	0.613	0.668	0.751	0.613	0.546
t_{-6}	0.817	0.611	0.568	0.637	0.677	0.537	0.616	0.677	0.577
t_{-7}	0.785	0.615	0.611	0.681	0.698	0.511	0.793	0.579	0.511
t_{-8}	0.780	0.613	0.618	0.671	0.676	0.518	0.739	0.671	0.518

set to interrupt the training process. To end training the model, the models were set to stop after 30 epochs in my parameter settings, then the model will cease training. The parameters of my model's training, as determined by the validation set, are the next thing to consider. The following are the six parameters obtained after applying the adjustment: learning rate = 0.30012, max delta step = 0, max depth = 6, nestimators = 100, njobs = 8, num parallel tree = 1; the remaining parameters are not listed here since there are too many values.



4.2.2 MIMIC-IV

In the MIMIC-IV dataset, this study included a total of 67,132 patients. I retrieved varying numbers of patients based on the two sepsis diagnostic criteria, which were SIRS and qSOFA, and nine distinct time points (t_0 to t_{-8}). On the other hand, from the extracted information in the MIMIC-IV dataset, I discovered that many patients with sepsis met the SOFA score on the first record. This was likely because these patients were already in serious condition at hospital admission. Because of this, no earlier vital signs before the onset can be extracted to forecast sepsis if the SOFA score was utilized to judge sepsis. Hence, two different sepsis judgment criteria were used to apply to the machine learning model. Table 4.6 displays the number of patients suffering from sepsis at various times among two scores. The machine learning strategies implemented on the MIMIC-IV dataset are outlined in Table 4.7. At nine different time points, the models were used. The findings of the 5-fold cross-validation are summarized in Table 4.7 as the average AUROC.

Table 4.6: Patient numbers for whole clock prediction in MIMIC-IV dataset.

Time points	SIRS	qSOFA
t_0	5,521	4,975
t_{-1}	809	1,191
t_{-2}	687	950
t_{-3}	590	794
t_{-4}	542	695
t_{-5}	495	614
t_{-6}	450	561
t_{-7}	413	494
t_{-8}	377	454

Furthermore, to see the results of the model training on a larger scale, I attempted to incorporate additional patients and shifted forward or backward ten minutes from the starting time on the hour. For instance, if the first forecast time was 120 minutes ago, then the interval prediction will become 110 minutes to 130 minutes. And the patient number



Table 4.7: Mean AUROC of XGBoost model at whole clock prediction in MIMIC-IV dataset.

Time points	SIRS	qSOFA
t_0	0.836	0.823
t_{-1}	0.831	0.777
t_{-2}	0.801	0.789
t_{-3}	0.826	0.743
t_{-4}	0.801	0.784
t_{-5}	0.863	0.833
t_{-6}	0.827	0.818
t_{-7}	0.841	0.818
t_{-8}	0.902	0.737

is shown in Table 4.8. This will allow us to expand the number of samples and improve the model's capacity for learning more information. Table 4.9 shows the average AUROC results of the three machine learning models.

Table 4.8: Patient numbers for different prediction intervals in MIMIC-IV dataset.

Time points	SIRS	qSOFA
t_0	11,582	10,718
t_{-1}	1,433	2,168
t_{-2}	1,111	1,557
t_{-3}	905	1,250
t_{-4}	879	1,097
t_{-5}	771	935
t_{-6}	671	801
t_{-7}	641	711
t_{-8}	567	694

During the process of training the MIMIC-IV numerical XGBoost model, an early stop was set to interrupt the training process to prevent the problem of overfitting and to ensure that the model is trained correctly. This was done to ensure that the model is not inappropriately trained. The parameter settings were set such that the training of the model will end after 30 iterations of the process. This indicates that the model will stop

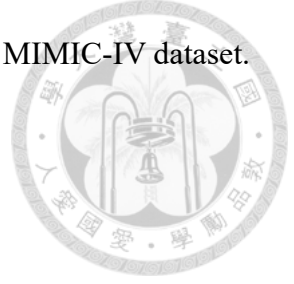


Table 4.9: Mean AUROC of XGBoost model at interval prediction in MIMIC-IV dataset.

Time points	SIRS	qSOFA
t_0	0.805	0.815
t_{-1}	0.787	0.731
t_{-2}	0.785	0.710
t_{-3}	0.733	0.679
t_{-4}	0.781	0.717
t_{-5}	0.777	0.665
t_{-6}	0.823	0.610
t_{-7}	0.761	0.655
t_{-8}	0.772	0.634

training after 30 epochs if the AUC does not show any signs of change during that time period. The second item that needs to be considered is the training parameters used in the model, as indicated by the validation set. After making the adjustment, the following results for the six parameters were obtained: learning rate = 0.30012, maximum depth = 6, maximum delta step = 0, nestimators = 100, number of jobs = 8, and number of parallel trees = 1. The other parameters are not included here since too many of them exist.

4.3 Heatmap Model Results

In this chapter, the EHR was converted into a heatmap, and then the images were classified images using the CNN model. Three distinct configurations of features were used to arrange the features, namely random order, pairs of correlation coefficients, and hierarchical clustering, to validate the model's reliability. The following findings were obtained by applying the procedures to the MIMIC-III and MIMIC-IV datasets. These configurations of features are shown in Section 4.3.1 and Section 4.3.2.



4.3.1 MIMIC-III

SIRS criteria were applied to the criterion for assessing sepsis to the CNN model and the model used the same number of patients in the prior numerical model for the MIMIC-III dataset. The mean AUROC values for image classification using the CNN model are presented in Table 4.10.

Table 4.10: AUROC of CNN model results from different feature arrange methods in MIMIC-III dataset. Values in parentheses represent standard deviation

Time points	Randomly	Pairs of correlation coefficients	Hierarchical clustering
t_0	0.996 (0.01)	0.994	0.932
t_{-1}	0.832 (0.03)	0.912	0.904
t_{-2}	0.915 (0.02)	0.923	0.921
t_{-3}	0.933 (0.01)	0.876	0.965
t_{-4}	0.923 (0.01)	0.932	0.917
t_{-5}	0.919 (0.02)	0.923	0.876
t_{-6}	0.819 (0.04)	0.952	0.877
t_{-7}	0.817 (0.08)	0.911	0.909
t_{-8}	0.821 (0.09)	0.945	0.884

The training results of t_0 to t_{-8} were presented in a bar plot, as shown in Figure 4.3, so that the three distinct approaches may be compared in a manner that is more comprehensible. This allowed me to thoroughly check the outcomes of my model's training. The chart makes it quite evident that all three of the feature arrangement approaches have the potential to produce good outcomes, the best of which can obtain an AUROC of approximately 0.99. However, after doing an in-depth comparison, it was discovered that the correlation pairs approach may have good results at 9 different time points, whereas the randomly arrange method can only have better classification results at t_0 . To make a con-

clusion, the above results show that using the correlation pairs method in the MIMIC-III database can obtained more consistent and better results at nine time points.

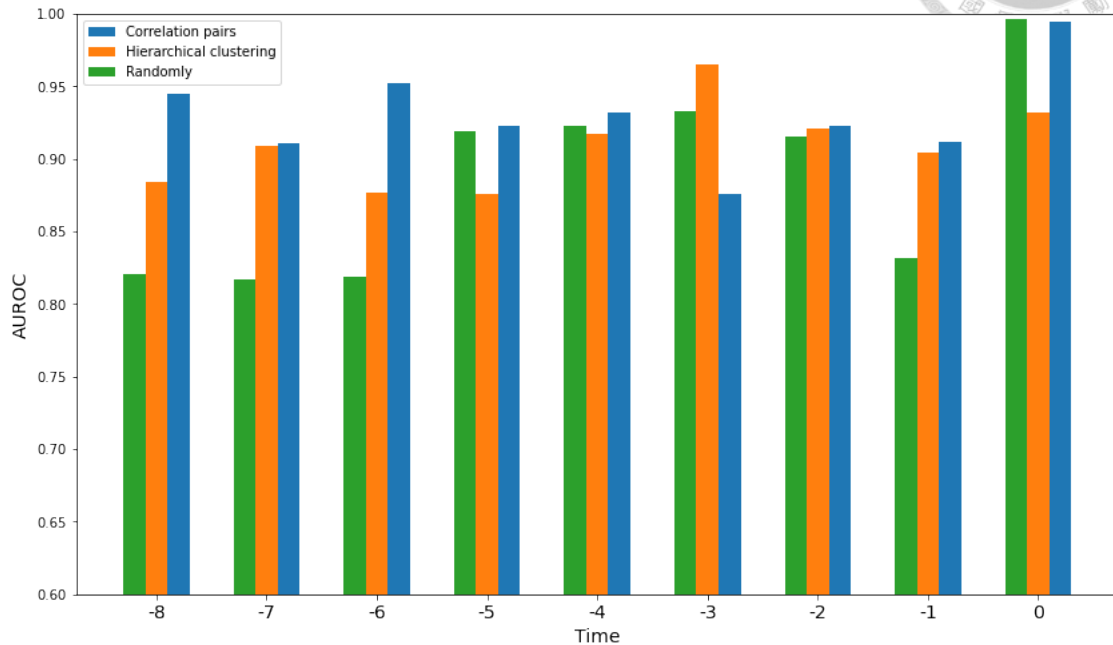
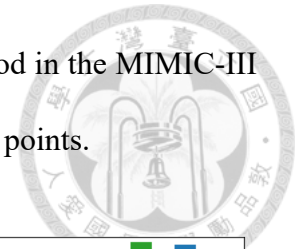


Figure 4.3: Bar plot of the heatmaps model testing result using CNN in MIMIC-III dataset.

A few time points were chosen throughout the process of model training to conduct observations to determine whether or not my model training suffers from overfitting issues. A good fit condition occurs when the plot of training loss lowers to the point of stability and the plot of validation loss declines to the end of stability and has a small gap with the training loss. This indicates that the training loss and validation loss are well matched. In an overfitting condition, the plot of the training loss curve continues to decline with the training process, while the plot of validation loss decreases to a point and then begins climbing again. This means that the model is already experiencing an overfitting state. The loss curve for the MIMIC-III model training is visualized in Figure 4.4. It can be deduced from these three images that the model has a decent fit and is convergent because both the training loss and the validation loss are getting closer and closer to a stable state. So from Figure 4.4, figures show that the training procedure has no problem with overfitting.

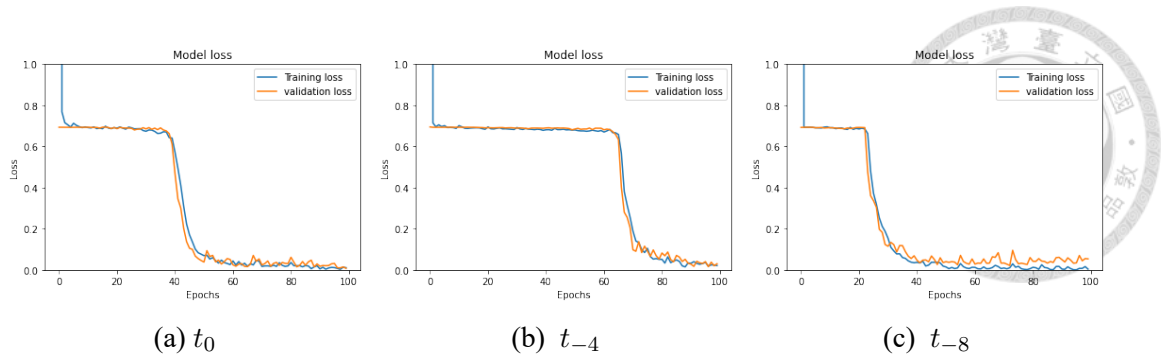


Figure 4.4: Loss curves at three different time points.

4.3.2 MIMIC-IV

SIRS scores were applied to the criterion for defining sepsis to the CNN model and the model used the same number of patients as the prior numerical model for the MIMIC-IV dataset. The mean AUROC values for image classification using the CNN model are presented in Table 4.11. The results presented in the table are the average AUROCs, and three sorting methods are used at the same time.

Table 4.11: AUROC of CNN model results from different feature arrange methods in MIMIC-IV dataset. Values in parentheses represent standard deviation

Time points	Randomly	Pairs of correlation coefficients	Hierarchical clustering
t_0	0.984 (0.02)	0.992	0.923
t_{-1}	0.953 (0.01)	0.964	0.874
t_{-2}	0.942 (0.06)	0.915	0.854
t_{-3}	0.914 (0.05)	0.832	0.931
t_{-4}	0.864 (0.02)	0.996	0.921
t_{-5}	0.885 (0.01)	0.913	0.902
t_{-6}	0.874 (0.07)	0.875	0.852
t_{-7}	0.874 (0.09)	0.913	0.941
t_{-8}	0.832 (0.07)	0.917	0.882

To check the results of my model's training more clearly, the training results of t_0 to t_{-8} were presented in a bar plot, as shown in Figure 4.5. This allows for more precise comparison of the three distinct approaches, which can be checked more easily. It can be

seen from the graph that the model will perform best at t_0 , which is exactly as anticipated, and that level will gradually decline with increasing time. However, after taking a closer look at the graph, it was discovered that the training outcomes of the correlation pairs approach performed well throughout all nine time points, and at some time points, they performed even more favorably than the other two methods. I assumed that the correlation pairs approach would be the most effective one to utilize because the features in this method group are comparable together, so the results would be better than the other two methods.

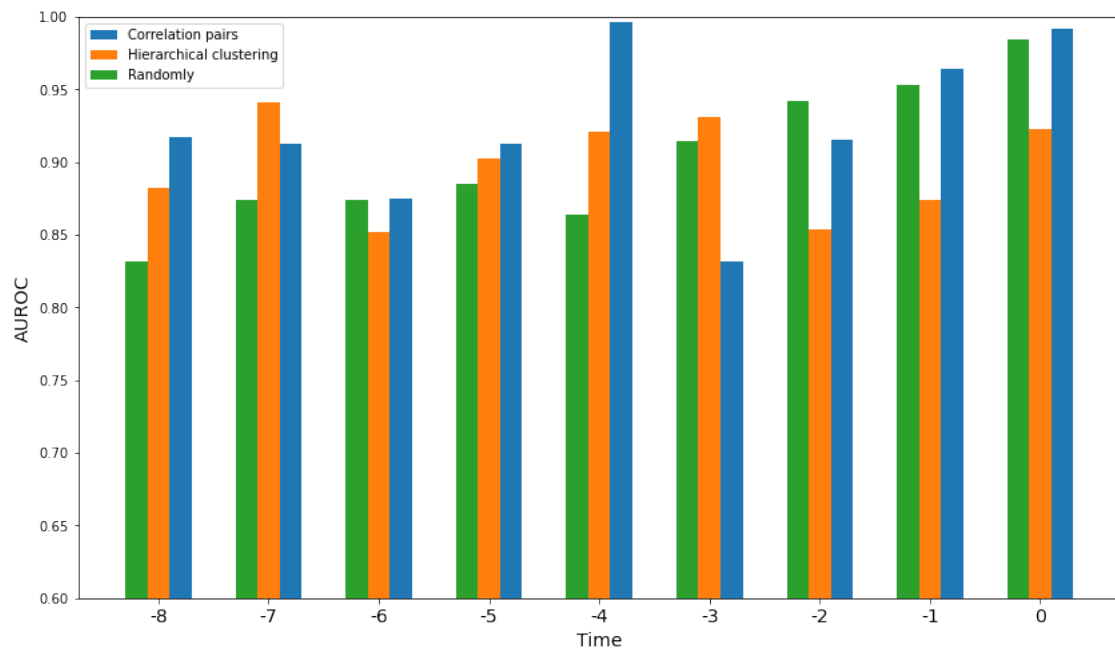
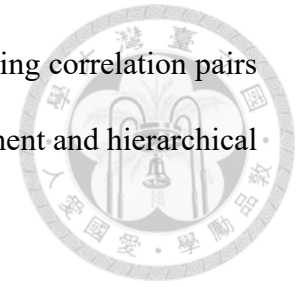


Figure 4.5: Bar plot of the heatmaps model testing result using CNN in MIMIC-IV dataset.

The outcomes of the three distinct approaches for rating features varied. According to the results, the best results can be obtained by employing the correlation pairs at t_0 , t_{-1} , t_{-4} , t_{-5} , t_{-6} and t_{-8} . At times t_{-7} and t_{-3} , the technique that made use of hierarchical clustering shows the best result at these time points. When using the random arrangement, the best results can only be obtained at time t_{-2} . According to the results shown above, the conclusion is that the sorting method of correlation pairs can produce more re-

liable and accurate classification results. This study provided that using correlation pairs arrangement methods could earn better results than random arrangement and hierarchical clustering methods.



4.4 Comparison

All the results are detailed in this subsection. It was possible to determine from Table 4.12 that the XGBoost model performed better in either the MIMIC-III or MIMIC-IV database by comparing the numerical results. My initial hypothesis was that this was because XGBoost was a superior model to the other two models. However, different patients will be determined according to different judgment scores, which may also affect the model's training. Therefore, the features defined for predicting sepsis will be discussed in detail in chapter 5. In particular, the prediction results for the complete clock are superior to the interval prediction results. The initial anticipation was that interval prediction might include more temporal information about the same patient. This would allow the model to be trained more realistically and bring it closer to the clinical situation. Our results show that using the whole clock prediction and the interval prediction do not differ by a significant number.

And the way to compare things is by looking at how different the training outcomes of a machine learning model are when it was applied to the two different datasets. In the MIMIC-III dataset, t_0 had the best performance among the nine time points, but in MIMIC-IV, the result at t_{-8} show the best. The problem may be caused by an insufficient quantity of samples. Therefore, more patients should be included in the related research in the future. Once more, it has been discovered that the SIRS standard is applied to achieve

superior results. However, in the model training results from t_{-2} to t_{-7} , there is not much difference between the two datasets. This could mean that in both training sets, good results can be obtained using XGBoost and using SIRS criteria. This also requires a larger sample size to validate the model generalization among two datasets.

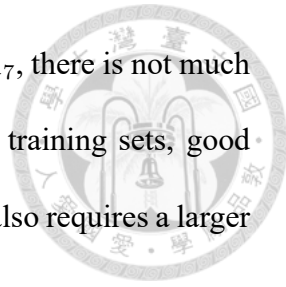
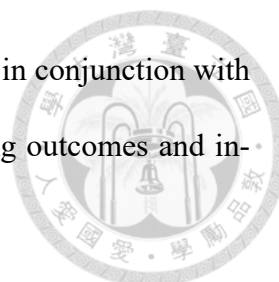


Table 4.12: Comparison of the best training results using numerical input and XGBoost model in the MIMIC-III and MIMIC-IV.

Dataset	Timepoints	Whole clock (A) or interval prediction (B)	Definition	AUROC
MIMIC-III	t_0	A	qSOFA	0.942
	t_{-1}	A	SIRS	0.798
	t_{-2}	A	SIRS	0.791
	t_{-3}	B	SIRS	0.819
	t_{-4}	B	SIRS	0.833
	t_{-5}	A	SIRS	0.822
	t_{-6}	A	SIRS	0.877
	t_{-7}	A	SIRS	0.835
	t_{-8}	B	SIRS	0.780
MIMIC-IV	t_0	A	SIRS	0.836
	t_{-1}	A	SIRS	0.831
	t_{-2}	A	SIRS	0.801
	t_{-3}	A	SIRS	0.826
	t_{-4}	A	SIRS	0.801
	t_{-5}	A	SIRS	0.863
	t_{-6}	A	SIRS	0.827
	t_{-7}	A	SIRS	0.841
	t_{-8}	A	SIRS	0.902

The training results of the heatmap are being discussed here. When contrasted with the numerical data presented in Table 4.13, the strategy of combining deep learning with image processing could produce superior learning outcomes at any given time point. This was found by examining Table 4.12. This is a noteworthy discovery. When the model was training, the training curve that the convergence speed of using correlation pairs will be faster, approximately 20 epochs, whereas the convergence speed of using randomly selected features is approximately 70 epochs. Three different methods of feature arrange-



ment do significantly affect the training results. Using a CNN model in conjunction with the correlation pairs approach can, in general, produce better training outcomes and increase training speed.

Table 4.13: Comparison of the best training results using graphical input in the MIMIC-III and MIMIC-IV.

Dataset	Timepoints	Arrangement method	AUROC
MIMIC-III	t_0	Correlation pairs	0.996
	t_{-1}	Randomly	0.912
	t_{-2}	Correlation pairs	0.923
	t_{-3}	Hierarchical	0.965
	t_{-4}	Randomly	0.932
	t_{-5}	Correlation pairs	0.923
	t_{-6}	Correlation pairs	0.952
	t_{-7}	Correlation pairs	0.911
	t_{-8}	Correlation pairs	0.945
MIMIC-IV	t_0	Correlation pairs	0.992
	t_{-1}	Correlation pairs	0.964
	t_{-2}	Randomly	0.942
	t_{-3}	Hierarchical	0.931
	t_{-4}	Correlation pairs	0.996
	t_{-5}	Correlation pairs	0.913
	t_{-6}	Correlation pairs	0.875
	t_{-7}	Hierarchical	0.941
	t_{-8}	Correlation pairs	0.917

After comparing the results of this study with other studies, this study provided the following findings. Although the methods did not coincide, this can serve as a reference point. Three time points were chosen to compare with others: t_0 , t_{-4} , and t_{-8} , and used the best-performing results of my model to compare my results with others' results. The studies selected also used machine learning to predict sepsis. Because there are many different machine learning methods, I only list the research results and do not compare machine learning methods and databases in detail. The comparison results are shown in Table 4.14. Although the database or sepsis definition used may be different, which may result in bias, this comparison can still provide some reference. My results are better when compared with others. In contrast to the research conducted by other individuals, I utilized

not only a numerical model but also a heatmaps model. Overall, this study yields better results than others.

Table 4.14: A comparison of my research with other researches.

Study	Dataset	Input	0 hour before sepsis onset	4 hours before sepsis onset	8 hours before sepsis onset
Zhang et al. (2020)	Cerner Health Facts database	EHR	-	0.843	-
Nemati et al. (2018)	MIMIC-III	EHR	-	0.822	0.804
Burdick et al. (2020)	Dascena Analysis Dataset	EHR	0.924	0.851	-
My study	MIMIC-III	EHR	0.942	0.833	0.780
		Heatmaps	0.996	0.932	0.902
	MIMIC-IV	EHR	0.836	0.801	0.902
		Heatmaps	0.992	0.996	0.917





Chapter 5 Discussion

The outcome substantiates two of the research findings. First, the machine learning approach produced good classification results on both MIMIC-III and MIMIC-IV at specific time points. There is no significant difference between the results and those of other studies. Second, numerical data were converted into a heatmap before training the CNN model. The model achieved satisfactory results in terms of categorization, which was in line with expectations. This is a significant finding because no research has been done to turn numerical data into heatmaps for early sepsis prediction. Our contribution is the realization that good results may also be obtained when heatmaps are used in conjunction with image categorization.

Comparing the differences in the results of the SIRS, SOFA, and qSOFA scores to discuss the findings of the studies on the definition of sepsis, I used the SIRS score, the SOFA score, and the qSOFA score. Regardless of which machine learning model, the results at different time points showed promising results when using the SIRS criteria to classify sepsis. These findings are based on the results presented in Chapter 4. I arranged the predicted features, SIRS score, SOFA score, and qSOFA score to correspond to the information in Table 5.1. This table presents the features used in the model and the features for which the different scores were calculated, with the overlapping parts in bold. The markers used by the SIRS score and my predictive features all overlapped. This may have



Table 5.1: The predicted features and the features used for the three scores. The bolded features overlap with the predicted features.

	Features
Predicted features	DBP, SBP, age, heart rate, temperature, spo2, respiratory rate, WBC.
SIRS score	Temperature, respiratory rate, heartrate, WBC
SOFA score	Mean arterial pressure, PaO2/FiO2 , platelets, bilirubin, glasgow coma scale, creatinine, urine output, dopamine, dobutamine, norepinephrine
qSOFA score	SBP, respiratory rate , glasgow coma scale

resulted in an inaccurate initial definition of sepsis, which caused the model to use the SIRS score to perform better, and the findings are the same. My research shows that any of these three scores has its value.; but, when it comes to the classification of the model, the SIRS score is the one that is most similar to the machine learning model employed.

It lacks validation from other databases and comparisons with physicians; therefore, it has some limitations linked with those two factors. None of the studies have shown that the results of machine learning or deep learning models can be better than the diagnostic criteria used by clinicians. Therefore, further research should be carried out in the circumstances more representative of the actual world to generalize the model and compare it to clinicians.



Chapter 6 Conclusion

When applied to the datasets compiled for MIMIC-III and MIMIC-IV datasets, the model I have proposed can get good results in predicting sepsis up to eight hours before symptoms start. This research was not only the new trial that numerical data were transformed into heatmaps but also the new finding that CNNs were successfully employed to anticipate sepsis. Despite this, even if the model is well trained, there is still a significant amount of room for improvement in this specific area of research. This is because there is a lot of untapped potentials. On the other hand, because this study did not include any validation with any other databases or categorical comparison with any clinicians, this study may have specific limitations due to the variables above.

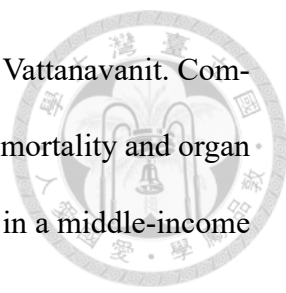
This research contributed to discovering a new prediction method distinct from the ones before. This new method would convert the general numerical data into image input and compare the practicability of using this method simultaneously with the use of two data sets. Future research should be conducted so that the model is generalized and can obtain good results in other databases. Alternatively, machine learning model methods should be used to compare with clinicians' diagnoses to demonstrate that machine learning is, in fact, helpful for intensive care units.

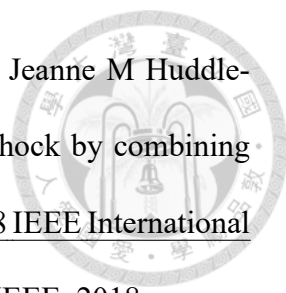




References

- [1] Pierangela Bruno and Francesco Calimeri. Using heatmaps for deep learning based disease classification. In 2019 IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pages 1–7. IEEE, 2019.
- [2] Paul E Marik and Abdalsamih M Taeb. SIRS, qSOFA and new sepsis definition. Journal of Thoracic Disease, 9(4):943, 2017.
- [3] Eamon P Raith, Andrew A Udy, Michael Bailey, Steven McGloughlin, Christopher MacIsaac, Rinaldo Bellomo, David V Pilcher, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. JAMA, 317(3):290–300, 2017.
- [4] Andrew Lever and Iain Mackenzie. Sepsis: definition, epidemiology, and diagnosis. BMJ, 335(7625):879–883, 2007.
- [5] Elizabeth K Stevenson, Amanda R Rubenstein, Gregory T Radin, Renda Soylemez Wiener, and Allan J Walkey. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis. Critical Care Medicine, 42(3):625, 2014.
- [6] Greg S Martin. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. Expert Review of Anti-Infective Therapy, 10(6):701–706, 2012.

- 
- [7] Bodin Khwannimit, Rungsun Bhurayanontachai, and Veerapong Vattanavanit. Comparison of the performance of sofa, qsofa and sirs for predicting mortality and organ failure among sepsis patients admitted to the intensive care unit in a middle-income country. Journal of Critical Care, 44:156–160, 2018.
- [8] Omar A Usman, Asad A Usman, and Michael A Ward. Comparison of sirs, qsofa, and news for the early identification of sepsis in the emergency department. The American Journal of Emergency Medicine, 37(8):1490–1497, 2019.
- [9] Jesus E Pino, Fergie J Ramos Tuarez, Jorge E Saona, Kai Chen, Endri Ceka, Julio Grajeda Chavez, Andres Chacon Martinez, Charles Bornmann, Pedro Torres, and Robert Chait. Misdiagnosis of sepsis in patients with acutely decompensated heart failure. real world outcomes. Journal of Cardiac Failure, 25(8):S150, 2019.
- [10] Åsa Askim, Florentin Moser, Lise T Gustad, Helga Stene, Maren Gundersen, Bjørn Olav Åsvold, Jostein Dale, Lars Petter Bjørnsen, Jan Kristian Damås, and Erik Solligård. Poor performance of quick-sofa (qsofa) score in predicting severe sepsis and mortality—a prospective study of patients admitted with infection to the emergency department. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 25(1):1–9, 2017.
- [11] Jean-Louis Vincent. The clinical challenge of sepsis identification and monitoring. PLoS Medicine, 13(5):e1002022, 2016.
- [12] Dongdong Zhang, Changchang Yin, Katherine M Hunold, Xiaoqian Jiang, Jeffrey M Caterino, and Ping Zhang. An interpretable deep-learning model for early prediction of sepsis in the emergency department. Patterns, 2(2):100196, 2021.

- 
- [13] 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. IEEE, 2018.
- [14] Hoyt Burdick, Eduardo Pino, Denise Gabel-Comeau, Carol Gu, Jonathan Roberts, Sidney Le, Joseph Slote, Nicholas Saber, Emily Pellegrini, Abigail Green-Saxena, et al. Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 us hospitals. BMC Medical Informatics and Decision Making, 20(1):1–10, 2020.
- [15] Joseph Futoma, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O’ brien. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In Machine Learning for Healthcare Conference, pages 243–254. PMLR, 2017.
- [16] Supreeth P Shashikumar, Qiao Li, Gari D Clifford, and Shamim Nemat. Multiscale network representation of physiological time series for early prediction of sepsis. Physiological Measurement, 38(12):2235, 2017.
- [17] Joseph Guillén, Jiankun Liu, Margaret Furr, Tianyao Wang, Stephen Strong, Christopher C Moore, Abigail Flower, and Laura E Barnes. Predictive models for severe sepsis in adult icu patients. In 2015 Systems and Information Engineering Design Symposium, pages 182–187. IEEE, 2015.
- [18] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. Early recognition of sepsis with gaussian process temporal convolutional networks

and dynamic time warping. In Machine Learning for Healthcare Conference, pages 2–26. PMLR, 2019.

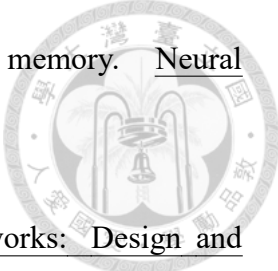


- [19] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D Feldman, Chris Barton, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Medical Informatics, 4(3):e5909, 2016.
- [20] 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, 2018.
- [21] 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, 2019.
- [22] Matthieu Scherpf, Felix Gräßer, Hagen Malberg, and Sebastian Zaunseder. Predicting sepsis with a recurrent neural network using the mimic iii database. Computers in Biology and Medicine, 113:103395, 2019.
- [23] Lucas M Fleuren, Thomas LT Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand RJ Girbes, Patrick Thoral, Ari Ercole, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Medicine, 46(3):383–400, 2020.
- [24] Mitchell M Levy, Mitchell P Fink, John C Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M Opal, Jean-Louis Vincent, and Graham

Ramsay. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. Intensive Care Medicine, 29(4):530–538, 2003.



- [25] Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA, 315(8):762–774, 2016.
- [26] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1):1–9, 2016.
- [27] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and R Mark IV. MIMIC-IV (version 0.4). PhysioNet, 2020.
- [28] M Michael Shabot. The hp carevue clinical information system. International Journal of Clinical Monitoring and Computing, 14(3):177–184, 1997.
- [29] Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The MIMIC code repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association, 25(1):32–39, 2018.
- [30] Stephen C Johnson. Hierarchical clustering schemes. Psychometrika, 32(3):241–254, 1967.
- [31] Scott Menard. Applied logistic regression analysis. Number 106. Sage, 2002.

- 
- [32] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
- [33] Larry Medsker and Lakhmi C Jain. Recurrent Neural Networks: Design and Applications. CRC press, 1999.
- [34] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015.
- [35] Gérard Biau and Erwan Scornet. A random forest guided tour. Test, 25(2):197–227, 2016.
- [36] Shiqi Yu, Sen Jia, and Chunyan Xu. Convolutional neural networks for hyperspectral image classification. Neurocomputing, 219:88–98, 2017.
- [37] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In 2014 13th international conference on control automation robotics & vision (ICARCV), pages 844–848. IEEE, 2014.
- [38] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 2285–2294, 2016.