國立陽明交通大學

生命科學系暨基因體科學研究所

碩士論文（初稿）

Department of Life Sciences and Institute of Genome Sciences

National Yang Ming Chiao Tung University

Master Thesis

建立一般病房預測心臟驟停之早期預警系統

Establish an early warning system for cardiac arrest in the general ward

研究生：李念恩（Li, Nien-En）

指導教授：巫坤品（Wu, Kun-Pin）

中華民國一一三年七月

July 2024

建立一般病房預測心臟驟停之早期預警系統

Establish an Early Warning System for Cardiac Arrest in the General Ward

研究生 ： 李念恩　　Student : Nien-En Li

指導教授 ： 巫坤品 博士　　Advisor : Dr. Kun-Pin Wu

國立陽明交通大學

生命科學系暨基因體科學研究所

碩士論文（初稿）

A Thesis Submitted to
Department of Life Sciences and Institute of Genome Sciences
College of Life Sciences
National Yang Ming Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master of Science
in Department of Life Sciences and Institute of Genome Sciences

July 2024

Taiwan, Republic of China

中華民國一一三年七月

摘要

心臟驟停（cardiac arrest）是指心臟停止機械活動，失去輸送血液到全身的功能。若不快速恢復心臟的跳動，患者有極高的機率會死亡。然而，高達八成的心臟驟停，在發生前皆出現可以被觀察到的徵兆。在一般病房中，醫療人員間隔 4–6 小時才會測量患者的生命徵象，並不會隨時注意每位患者。一旦有心臟驟停的事件發生，就不容易及時發現，進而錯過救援的時間。因此，本研究目的是利用公開的重症醫學資料，模擬一般病房量測患者生命徵象的項目和頻率，以建立一套可應用於一般病房的早期預警系統。透過此方式希望能克服難以取得一般病房資料的困境，並預測住院患者是否會在短時間內發生心臟驟停，提早讓醫療人員進行相關處置。

本研究的資料來源為急重症醫學資訊市集第四版（Medical Information Mart for Intensive Care, MIMIC IV），將發生院內心臟驟停及未發生心臟驟停的患者作為研究對象，利用體溫、心律、呼吸頻率、收縮壓、舒張壓及血氧飽和度等基本生命徵象作為資料特徵。為了掌握患者狀態的變化趨勢，我們考慮患者 48 小時內的多次量測資訊，透過 RNN、LSTM、GRU、CNN-LSTM 等深度學習方法來建立心臟驟停預測模型，並以多種指標評估模型的效能。本研究進一步比較不同患者組成的樣本是否對模型表現造成影響。結果發現，同時取自心臟驟停患者和未心臟驟停患者的各半資料，所建構之 GRU 模型有最佳的預測表現。本研究建立的模型表現皆優於傳統 NEWS（National Early Warning System），可作為辨識患者臨床惡化的有效篩檢工具。此外，我們嘗試使用馬偕醫院（台北院區）一般病房發生心臟驟停的患者資料進行外部驗證，初步獲得穩定的預測效果。未來期望該模型可發揮在臨床上，提升患者存活率並減輕醫療人員於照護上的負擔。

關鍵字：心臟驟停、生命徵象、MIMIC IV、早期預警系統、深度學習模型

Abstract

Cardiac arrest is a critical condition where the heart stops pumping effectively, often leading to death without prompt treatment. However, up to 80% of cardiac arrests are preceded by noticeable symptoms. In general wards, medical professionals measure the vital signs of patients in general wards every 4-6 hours and do not continuously monitor each patient. Once a cardiac arrest event occurs, it is not easily detected in a timely manner, leading to missing the rescue time. Therefore, the study aims to use publicly available critical care data to simulate the measurement items and frequencies of vital signs in general wards, in order to establish an early warning system applicable to general wards. This approach is intended to overcome the difficulty of obtaining data from general wards and to predict whether hospitalized patients will experience cardiac arrest in the short term, allowing medical professionals to take timely actions.

The data source for this study is the fourth edition of the Medical Information Mart for Intensive Care (MIMIC IV), which includes patients who experienced in-hospital cardiac arrest and those who did not. The data features used are basic vital signs such as body temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and oxygen saturation. To capture the trends in patient condition over time, we considered multiple measurements within a 48-hour period and used deep learning methods such as RNN, LSTM, GRU, and CNN-LSTM to develop models predicting cardiac arrest. The performance of these models was evaluated using various metrics. The results showed that the GRU model, built using data equally divided between patients who experienced cardiac arrest and those who did not, demonstrated the best predictive performance. The models developed in this study outperformed the traditional National Early Warning System (NEWS), providing an effective screening tool for identifying clinical deterioration. Additionally, we attempted external validation using data from patients who experienced cardiac arrest in the general wards of Mackay Memorial Hospital (Taipei branch), and initially achieved stable predictive results. In the future, it is hoped that the model can be applied clinically to improve patient survival rates and reduce the burden on medical professionals in medical care.

Keywords：cardiac arrest, vital signs, MIMIC IV, early warning system, deep learning model

目錄

摘要 i
Abstract ii
目錄 iii
圖目錄 v
表目錄 vi
第一章 緒論 1
1.1 研究背景 1
1.1.1 心臟驟停（Cardiac arrest） 1
1.1.2 快速反應系統（Rapid response systems） 2
1.1.3 早期預警分數（Early warning score） 2
1.1.4 早期預警系統於急診、加護病房與一般病房之情況 5
1.2 相關研究探討 5
1.2.1 早期預警系統應用於預測心臟驟停 5
1.2.2 機器學習應用於預測心臟驟停 5
1.2.3 預測模型應用場所 6
1.2.4 醫學資料庫應用於醫療研究 6
1.3 研究動機與目的 7
第二章 材料與方法 8
2.1 資料來源 8
2.1.1 急重症醫學資料庫（MIMIC） 8
2.2 研究對象 8
2.2.1 MIMIC IV 中發生心臟驟停之患者 8
2.2.2 從 MIMIC IV 中篩選未發生心臟驟停之患者 8
2.3 資料特徵 9
2.4 資料前處理 9
2.5 選取正負樣本 11
2.5.1 正樣本 12
2.5.2 負樣本 13
2.6 預測模型建立 14
2.6.1 深度模型架構 14
2.6.2 超參數設定與執行環境 18

2.6.3 自助抽樣法（bootstrap）.....18
2.7 模型評估.....19
2.8 統計分析.....21
2.9 NEWS 計分.....22
2.10 特徵重要性.....23
第三章 研究結果.....24
3.1 研究流程.....24
3.2 模型穩定性.....26
3.3 資料人口學及特徵分析.....27
3.4 不同負樣本與不同模型組成之比較.....28
3.4.1 負樣本取自心臟驟停患者資料所訓練之模型.....28
3.4.2 負樣本取自未心臟驟停患者資料所訓練之模型.....29
3.4.3 負樣本取自心臟驟停患者和未心臟驟停患者資料訓練之模型
30
3.5 NEWS 與 GRU 之比較.....33
3.6 特徵重要性.....34
第四章 討論.....36
4.1 資料人口學及特徵分析.....36
4.2 不同負樣本組成之比較.....36
4.3 NEWS 與 GRU 之比較.....37
4.4 特徵重要性.....37
4.5 外部驗證.....38
4.6 研究優點與限制.....39
第五章 結論.....41
參考文獻.....42
附錄.....47
A. 變項代碼對照表及數值範圍.....47
B. MIMIC IV 資料配置.....48
C. 馬偕紀念醫院（台北院區）一般病房之心臟驟停患者資料.....49

圖目錄

圖 1.1 MEWS 分數計算表.....2
圖 1.2 NEWS 分數計算表.....4
圖 1.3 NEWS 警示閾值對照表.....4
圖 2.1 模型預測目標.....11
圖 2.2 選取正樣本示意圖.....12
圖 2.3 選取心臟驟停患者之負樣本示意圖.....13
圖 2.4 選取未心臟驟停患者之負樣本示意圖.....14
圖 2.5 RNN 單元結構.....15
圖 2.6 LSTM 單元結構.....16
圖 2.7 GRU 單元結構.....17
圖 2.8 CNN-LSTM 模型結構.....17
圖 2.9 bootstrap 抽樣示意圖.....19
圖 2.10 混淆矩陣.....20
圖 2.11 AUROC 示意圖.....21
圖 2.12 正負樣本進行 NEWS 分數計算.....22
圖 2.13 SHAP 分析示意圖.....23
圖 3.1 研究流程。圖中以正負樣本皆取自心臟驟停患者資料為例.....24
圖 3.2 各模型於訓練過程所計算之損失值（loss）.....26
圖 3.3 負樣本取自心臟驟停患者資料所建構 GRU 之特徵重要性.....34
圖 3.4 負樣本取自未心臟驟停患者資料所建構 GRU 之特徵重要性.....35
圖 3.5 負樣本取自心臟驟停患者和未心臟驟停患者資料所建構 GRU 之特徵重要性.....35

表目錄

表 3.1 MIMIC IV 中，心臟驟停患者和未心臟驟停患者之人口學及特徵分析....27
表 3.2 負樣本取自心臟驟停患者資料所訓練模型之訓練和驗證結果.....28
表 3.3 負樣本取自心臟驟停患者資料所訓練模型之測試結果.....29
表 3.4 負樣本取自未心臟驟停患者資料所訓練模型之訓練和驗證結果.....29
表 3.5 負樣本取自未心臟驟停患者資料所訓練模型之測試結果.....30
表 3.6 負樣本取自心臟驟停患者和未心臟驟停患者資料所訓練模型之訓練和驗證結果.....31
表 3.7 負樣本取自心臟驟停患者和未心臟驟停患者資料訓練模型之測試結果...31
表 3.8 不同負樣本組成資料應用於 NEWS 之預測結果 33
表 3.9 不同負樣本組成資料應用於 GRU 之預測結果 33

第一章　緒論

1.1 研究背景

1.1.1 心臟驟停（Cardiac arrest）

心臟驟停是指心臟停止正常跳動，無法將血液輸送至全身，各器官和組織會因為缺乏氧氣，逐漸喪失功能、壞死[1]。引起心臟驟停的主要原因是心律不整（arrhythmia）、心肌梗塞（myocardial infarction），大約占 50%–60%，其次是呼吸功能不全，大約占 15%–40%[2]。而面對心臟驟停的關鍵干預措施（critical interventions）包括胸腔按壓（chest compressions）、輔助通氣（assisted ventilation）和早期除顫（early defibrillation）[3]。

根據美國心臟協會（American Heart Association）研究統計，美國每年有超過 29 萬名住院患者發生院內心臟驟停（in-hospital cardiac arrest, IHCA），出院存活率大約為 22.4%[4]；美國每年超過 35 萬名患者發生院外心臟驟停（out-of-hospital cardiac arrest, OHCA），出院存活率大約為 9.3%[5]。根據 Chen 等人在 2012 年研究統計，台灣該年每 1000 名住院患者中會發生 3.25 例 IHCA，出院存活率大約為 11.8%[6]； 根據 Wang 等人在 2000 年至 2012 年研究統計，台灣每年有超過 9 千名患者發生 OHCA，30 天存活率大約為 10.9%[7]。由上述的統計結果得知，只要有心臟驟停事件發生，若沒有立即採取適當的緊急處置，可能會提高患者的死亡率。

許多院內心臟驟停在回顧性研究中被認為是可以預防或避免的[8]，約 80% 心臟驟停的患者在事件前 8 小時內會出現生命徵象（vital signs）或意識水平改變等相關跡象[9, 10]，因此具備辨識病情惡化的相關系統可提供醫院一個更安全的環境，協助醫療人員觀察院內的患者，出現任何突發問題可以即時接受治療，有助於減少心臟驟停事件或更嚴重的後果發生。

1.1.2 快速反應系統（Rapid response systems）

為了因應患者突發的危急情況，部分醫院會配置快速反應系統，這套系統可對表現出惡化的患者，提供早期檢測和適當的醫療處理，進而降低患者不良預後和死亡的機率[11, 12]。快速反應系統中的追蹤和觸發系統（track-and-trigger systems）可用來監測患者狀態和發出警示，主要分為 single-parameter track-and-trigger system（SPTTS）以及 multiple-parameter track-and-trigger systems（MPTTS）。MPTTS 相對於 SPTTS，在預測院內心臟驟停和臨床惡化方面有更好的表現[13]，臨床上應用最常見的 MPTTS 是 modified early warning score 和 national early warning score[14]。

1.1.3 早期預警分數（Early warning score）

1.1.3.1 Modified Early Warning Score（MEWS）

MEWS 最初由 Morgan 等人所提出[15]，經過多項研究改進後在 2001 年得到驗證[16]，作為一種識別住院患者惡化和死亡風險的臨床工具。MEWS 的計分方式由五個生理參數組成：收縮壓、心率、呼吸頻率、體溫和意識狀態（AVPU score）。如圖 1.1 所示，當醫療人員測量患者的生命徵象後，各項數值會對應到表格上方不同的分數，有任何一項超出正常範圍的項目會得到 1 至 3 分，總分越高代表患者的情況越嚴重，需要由快速反應團隊檢查原因並決定後續治療和照護方式。

	Score						
Vital signs	3	2	1	0	1	2	3
Systolic blood pressure (mmHg)	<70	70–80	81–100	101–199		≥200	
Heart rate (bpm)		<40	40–50	51–100	101–110	111–129	≥130
Respiratory rate (bpm)		<9		9–14	15–20	21–29	≥30
Temperature (°C)		<35		35–38.4		≥38.5	
AVPU score				**A**lert	Reacting to **V**oice	Reacting to **P**ain	**U**nresponsive

Each component of MEWS has an associated score ranging from 0 to 3, based on the degree of derangement of the parameter. The total score is the sum of each component: the maximum possible score is 14.

圖 1.1 MEWS 分數計算表

根據不同醫院或不同科別的要求，可以將 MEWS 的評估項目及範圍進行調整，例如在 Santiago González 等人的研究中，增加氧氣治療和血氧飽和度（SpO_2）的項目[17]；或是 Heitz 等人利用格拉斯科昏迷量表（GCS）取代 AVPU score 來確定患者的意識水平[18]。而 MEWS 發出預警的分數也可以依照各個單位的需求進行更改，根據研究統計，加總起來超過 5 分以上與進入加護病房、不良預後或死亡風險有較高的相關[19, 20]，因此 MEWS 大多會以 5 分作為系統的警示閾值，通知醫療人員查看患者是否需要執行相關處置。

1.1.3.2 National Early Warning Score（NEWS）

NEWS 由開發和實施小組代表皇家內科醫師學會（Royal College of Physicians）所制定，最初在 2012 年於英國發布，也是一種識別住院患者惡化和死亡風險的臨床工具，目前被英國的國民保健署（National Health Service, NHS）以及全球各地的醫療機構廣泛使用。NEWS 的計分方式由七個生理參數組成：呼吸頻率、血氧飽和度、有無供氧、收縮壓、脈搏、意識狀態和體溫。其中血氧飽和度分成 scale 1 和 scale 2。scale 1 適用於一般患者，scale 2 適用於具有慢性高碳酸血症呼吸衰竭（chronic hypercapnic respiratory failure, HCRF）的患者。HCRF 患者建議的血氧飽和度範圍為 88–92%，如果對患者提供過多的氧氣，可能使其面臨高碳酸血症（hypercapnia）迅速惡化和死亡的風險。因此，NEWS 新增 scale 2 的評分標準，讓這些患者的血氧飽和度不在一般範圍時，系統也不會頻繁地發出警示，從而在氧氣的使用上更加安全[21]。如圖 1.2 所示，NEWS 計算方式與 MEWS 相同，透過量測完患者的生命徵象，將各參數得到的分數進行加總，依據系統是否會發出預警得知每位患者的狀況。

Physiological parameter	Score						
	3	2	1	0	1	2	3
Respiration rate (per minute)	≤8		9–11	12–20		21–24	≥25
SpO_2 Scale 1 (%)	≤91	92–93	94–95	≥96			
SpO_2 Scale 2 (%)	≤83	84–85	86–87	88–92 ≥93 on air	93–94 on oxygen	95–96 on oxygen	≥97 on oxygen
Air or oxygen?		Oxygen		Air			
Systolic blood pressure (mmHg)	≤90	91–100	101–110	111–219			≥220
Pulse (per minute)	≤40		41–50	51–90	91–110	111–130	≥131
Consciousness				Alert			CVPU
Temperature (°C)	≤35.0		35.1–36.0	36.1–38.0	38.1–39.0	≥39.1	

圖 1.2 NEWS 分數計算表

NEWS 所評估的項目以及範圍是固定的，不可依照醫院或科別進行更改。如圖 1.3 所示，NEWS 總分 5–6 分為系統發出警示的門檻，代表患者處於中度臨床風險，需要立即通知醫療人員介入處理，並考慮將患者轉移至具備重症照護能力的醫療團隊。當總分達到 7 分或更高時，代表患者處於高度臨床風險，死亡率將快速上升，應該由具備重症照護能力的團隊進行緊急評估或處置，並考慮將患者轉移至等級更高的醫學中心。

NEW score	Clinical risk	Response
Aggregate score 0–4	Low	Ward-based response
Red score Score of 3 in any individual parameter	Low–medium	Urgent ward-based response*
Aggregate score 5–6	Medium	Key threshold for urgent response*
Aggregate score 7 or more	High	Urgent or emergency response**

圖 1.3 NEWS 警示閾值對照表

1.1.4 早期預警系統於急診、加護病房與一般病房之情況

早期預警系統常用於緊急服務和重症監護中。例如在急診（Emergency room）為了快速診斷患者的嚴重程度，許多警示系統被用來辨識已發生、即將發生危急的患者，或確定患者的住院需求[22]。在加護病房（Intensive Care Unit, ICU）的患者通常生命徵象較不穩定，病情可能會迅速變化，因此會設置警示系統持續觀察患者的狀態，讓醫療團隊可以即時協助處理。

在一般病房中，醫療人員間隔 4–6 小時才測量一次患者的生命徵象[23, 24]，相較於急診和加護病房的監測頻率是非常少；而一般病房所配置的醫療人力也不如急診和加護病房，因此無法隨時注意每位患者。如果患者突然發生惡化或心臟驟停，可能就無法在第一時間察覺，進而錯過救援的黃金時間。假設未來可以在一般病房設立預警系統，不但可以減輕醫療人員的照護負擔，也可以為住院患者提供優良的醫療品質和安全性。

1.2 相關研究探討

1.2.1 早期預警系統應用於預測心臟驟停

現今預測院內心臟驟停大多會選擇 MEWS 和 NEWS[25, 26]，但近期有多項研究顯示這兩種系統的表現並不穩定，例如出現靈敏度（sensitivity）低、誤報率（false-alarm rates）高的結果[27]，或是預測準確度（accuracy）存在很大的差異[28-30]，對於單獨使用 MEWS 或 NEWS 的醫療院所，這些問題可能會造成很大的困擾。理想的預警系統應該具備較高的靈敏度和特異度（specificity），確保能正確判斷出高風險患者，同時也要避免過度警報，才不會浪費醫療資源、降低護理品質。

1.2.2 機器學習應用於預測心臟驟停

近年來，許多學者開發人工智慧相關的預測模型，希望能改善使用 MEWS 和

NEWS 所存在的不確定性，目前多數研究採用機器學習演算法來建立預測院內心臟驟停的模型，如隨機森林（Random Forest, RF）、支持向量機（Support Vector Machine, SVM）、邏輯斯迴歸（Logistic Regression, LR）等[31]。然而，部分高性能的預測模型需要依賴大量的特徵輸入，包含人口統計資料、多種監測參數和實驗室檢驗結果，並經過複雜的運算[32, 33]。當模型處理高維數據時，過多的特徵容易導致模型出現過擬合（overfit），同時也會限制模型的使用區域和時機[34]。此外，臨床資料因各種不可預期的原因容易發生缺失，缺少的數據可能會降低預測表現，對於數據上的處理也是一大挑戰[35]，因此這些模型要應用在臨床實際場域仍有一定的困難度。

1.2.3 預測模型應用場所

根據 Perman 等人的研究，院內心臟驟停大多數發生在加護病房中[36]，代表目前預測心臟驟停的模型主要是依賴加護病房的患者資料，可能會限制模型在非加護病房環境中的適用性。雖然 Churpek 等人發表一般病房患者惡化的預測研究[37]，但相對於急診室或加護病房，一般病房的相關研究或文獻還是較為稀少。因為急診室和加護病房通常會處理更加急迫和重症的案例，針對密集監護和資源配置有較高的需求，這些場所會投入較多的相關研究。因此未來需要針對一般病房的醫療環境，進行相關的風險分析和模型建構。

1.2.4 醫學資料庫應用於醫療研究

公開的醫學資料庫是研究和開發醫療模型的重要資源，因為它們提供了患者的詳細醫療資訊。這些資料庫使研究人員和醫療專業人員能夠進行相關的臨床研究、疾病監測、評估醫療介入以及改善治療策略，從而提升醫療服務的品質和效率[38]。其中，常見的重症醫學資料庫包含 MIMIC（Medical Information Mart for Intensive Care）、eICU Collaborative Research Database 等；常見的急診醫學資料

庫包含 NHAMCS（National Hospital Ambulatory Medical Care Survey）、NEDS（Nationwide Emergency Department Sample）等。然而一般病房不像重症或急診有專門的醫學資料庫，因此不易取得一般病房的相關資料。

1.3 研究動機與目的

在現今的醫療領域，MEWS 和 NEWS 的評分系統無法有效處理大量的資料，也無法適應病情的快速變化，針對一般病房患者突然出現異常狀況，大多是依賴醫療人員和家屬的觀察與通報。隨著科技的進步，目前的資訊系統可透過自動化收集和分析醫療數據，根據患者生命徵象的變化趨勢，以即時且精確的警示預測未來可能出現的惡化風險。

而深度學習技術在事件預測方面的表現優於 MEWS、NEWS 及傳統的機器學習[39, 40]。深度學習模型在訓練階段能有效控制和學習大量特徵的資料，解決機器學習在處理複雜數據所面臨的困境[41]。例如，Hochreiter 和 Schmidhuber 在 1997 年的研究中提出了長短期記憶網路（Long Short-Term Memory, LSTM），成功克服長期時間序列數據在預測模型中的處理[42]。

因此綜合上述，本研究希望透過公開的重症醫學資料庫，模擬一般病房量測患者生命徵象的項目和頻率，利用深度學習技術建立一套放置在一般病房的早期預警系統。期望這套方法能克服不易取得一般病房資料的難處，並透過模型分析與監測患者的生命徵象，預測患者在近期內是否會出現心臟驟停，以便醫療人員能夠盡早採取適當的急救措施。

第二章　材料與方法

2.1　資料來源急重症醫學資料庫（MIMIC）

本研究使用急重症醫學資訊市集第四版（Medical Information Mart for Intensive Care IV, MIMIC IV）的資料進行模型建構。此大型且公開的資料庫收集 2008 年至 2019 年於美國貝斯以色列女執事醫療中心（Beth Israel Deaconess Medical Center, BIDMC），大約四萬多名加護病房患者的醫療數據。MIMIC IV 提供豐富的臨床資訊，包括人口統計、生理數值、實驗室相關檢驗、藥物使用和醫學影像等。為了保護患者隱私，BIDMC 會將每筆資料進行去辨識化處理[43, 44]。

2.2　研究對象

2.2.1　MIMIC IV 中發生心臟驟停之患者

本研究主要目的是預測心臟驟停。在 MIMIC IV 資料庫的 procedureevents 資料表記錄了患者在住院期間所發生的事件，根據 MIMIC IV 的代碼簿（codebook）所記錄，患者的 itemid 代碼為 225466 表示發生心臟驟停事件。我們使用 procedureevents 資料表和此代碼篩選出 575 起心臟驟停事件及患者資料。

2.2.2　從 MIMIC IV 中篩選未發生心臟驟停之患者

本研究建立預測模型時會使用未發生心臟驟停的患者資料。為了避免發生心臟驟停與未發生心臟驟停這兩族群的基本分佈差異過大，我們根據年齡與性別為每一位發生心臟驟停的患者進行配對，在 MIMIC IV 資料庫中隨機篩選未發生心臟驟停的患者。在 575 起心臟驟停事件中，經過資料處理後共包含 477 位患者。透過此方法收集 477 位未發生心臟驟停患者及其資料。

2.3 資料特徵

在 MIMIC IV 資料庫中，patient 資料表記錄患者的基本資料，chartevents 資料表記錄患者在住院期間所測量的項目及時間。本研究從中選取每位患者的入院號碼（stay_id）、年齡、性別、六種生命徵象（參照附錄 A：變項代碼對照表及數值範圍）、生命徵象測量時間和日期。

上述提及的六種生命徵象作為模型預測的資料特徵，分別為常見的收縮壓（Systolic Blood Pressure, SBP）、舒張壓（Diastolic Blood Pressure, DBP）、心率（Heart Rate, HR）、呼吸頻率（Respiratory Rate, RR）、體溫（Body Temperature, BT），以及血氧飽和度（oxygen saturation, SpO_2）。這些生命徵象不僅被 NEWS 所使用，也被專家認定為一般病房每天需要測量的數據，表示這些生命徵象對監測患者狀況和預測心臟驟停事件具有一定的關聯。

MIMIC IV 資料庫中所記錄的收縮壓和舒張壓，其數值來自侵入式量測（invasive）與非侵入式量測（non-invasive）。本研究欲發展應用在一般病房的系統，因此選擇使用非侵入式量測的血壓。而資料庫中所記錄的體溫，包含攝氏溫度和華氏溫度，在篩選資料同時會選擇這兩種溫度單位，考慮到國內慣用為攝氏溫度，我們一律將華氏溫度轉換為攝氏溫度。

此外，MIMIC IV 是麻省理工學院運算生理實驗室（Massachusetts Institute of Technology-Laboratory for Computational Physiology, MIT-LCP）和 BIDMC 共同開發的資料庫。因此，從 MIMIC IV 資料庫中篩選資料時，我們會使用 MIT-LCP 提供的 SQL 來檢查每個生命徵象的數據範圍是否正確[45]。

2.4 資料前處理

我們從 MIMIC IV 資料庫中，透過 chartevents 資料表挑選出所需的變項後，根據以下列規則進行資料前處理：

A. 排除年齡小於 20 歲的患者，超過 200 歲的患者則視為異常值並移除。

B. 同位患者在同次住院期間可能會經歷多次心臟驟停，而後續事件可能與第一次事件有關聯。為了避免研究結果受到多次事件的干擾，從 procedureevents 資料表篩選出來的心臟驟停事件，保留每位患者同次住院發生第一次心臟驟停的記錄，後續事件則不納入分析範圍。

C. 將保留後的心臟驟停時間與患者的生命徵象資料做合併，並移除每位患者發生心臟驟停後續的資料。

D. 移除記錄時間不明的資料，因無法確認測量和事件發生的時間差。

E. 移除六個生命徵象全是缺失值的資料，認定為嚴重缺失、無法進行補值。

F. 當生命徵象超過一定範圍時（參照附錄 A：變項代碼對照表及數值範圍），則視為異常值（outlier）並移除該數值。

G. 針對資料缺值採用線性內插法（linear interpolation）進行補值。假設兩個已知數值間的變化是線性，透過線性內插法估算這兩數值間的未知數值[46]。在醫學數據上，相鄰的資料點通常會是相似的數值，相較於常使用的平均數補值，線性內插法更能保留原始數據的趨勢[47]。而線性內插法需要至少兩個數值才能完成，若遇到無法補值的情況，考慮到加護病房量測時間相當密集，因此會直接採用最鄰近的資料點。

MIMIC IV 資料庫中的 575 起心臟驟停事件，共收集 148,334 筆生命徵象資料；資料清理後留下 479 起事件、68,973 筆生命徵象資料。而隨機配對 477 位未發生心臟驟停的患者，共收集 187,517 筆生命徵象資料；資料清理後留下 55,769 筆生命徵象資料。

2.5 選取正負樣本

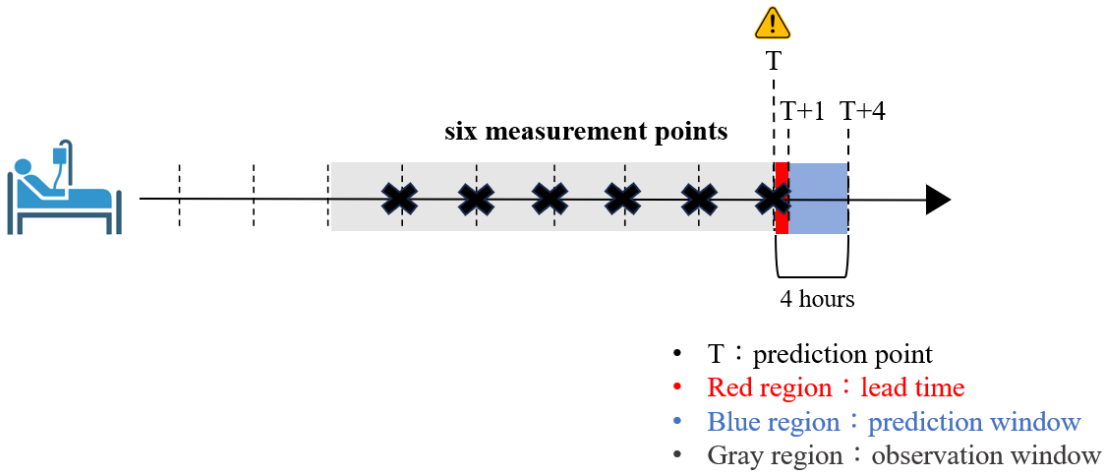

圖 2.1 模型預測目標

在患者資料選取正負樣本前，先介紹本研究模型的預測目標：模型將使用預警前 24 小時的數據，預測患者一小時後到下一次醫療人員測量前發生心臟驟停的可能性。如圖 2.1 所示，模型若在時間點 T 進行預測，則定義為預測點（prediction point）；時間點 T+1 到 T+4（藍色區間）定義為預測視窗（prediction window），表示模型預測可能會發生事件的時間範圍；時間點 T 到 T+1（紅色區間）定義為前置時間（lead time），表示模型發出警示到事件發生的時間範圍，預留一小時讓醫療人員有足夠時間為可能發生的心臟驟停做好準備，例如調整藥物、準備 CPR 和 AED 等臨床處置。此外，位於前置時間的資料並不會納入模型的訓練，因為模型在進行預測時，理論上是無法得知患者在此範圍的生命徵象。

模型使用預測點前 24 小時內的資料決定發出預警與否，表示時間點 T-24 到 T（灰色區間）定義為觀測視窗（observation window）。為了模擬一般病房測量的頻率，實驗設計每四小時為一個時間點，代表每次會選擇六個連續資料點視為一筆樣本（黑色叉叉）。然而，一般病房的醫療人員並不會每四小時準時量測患者的生命徵象，若該時間點沒有測量紀錄，則會在前後 30 分鐘內隨機選取一筆資料，作為該時間點的生命徵象紀錄。因此本研究模型會利用過去連續六次的測量結果，預測患者在下一次測量前是否會發生心臟驟停。

此外，考量到患者入院後的生命徵象為一筆一筆進行測量，一開始並不會一次就出現六次測量記錄，故利用零填充（zero-padding）將空白資料進行填補，使所有樣本具有相同的輸入格式，從而被模型有效處理。若患者的醫療數據紀錄未滿 24 小時，或是該時間點無測量記錄，一樣利用零填充的方式進行資料填補，確保所有具時間序列的資料能適當被保留下來並進行分析。

2.5.1 正樣本

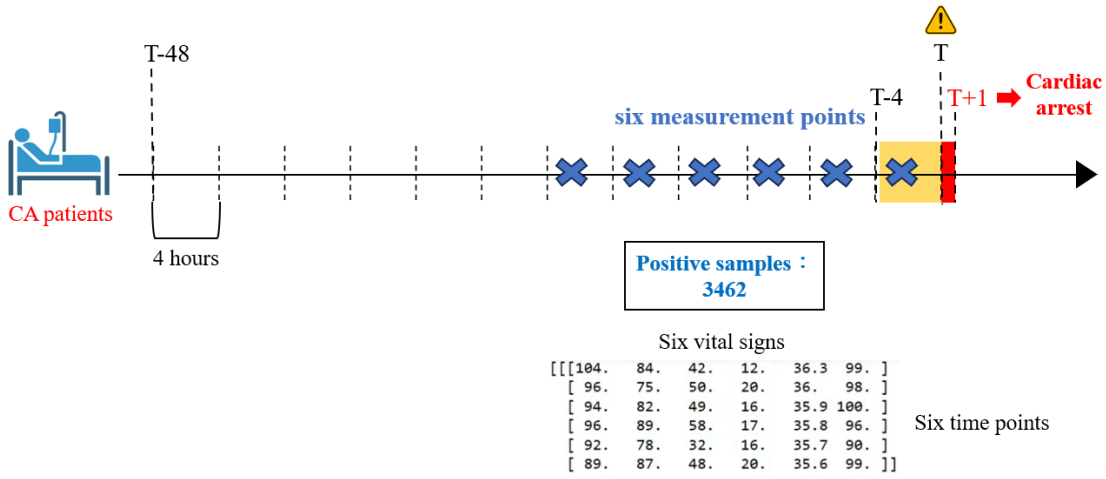

圖 2.2 選取正樣本示意圖

根據上述以連續六個測量點為一筆樣本的方式，套用在 MIMIC IV 資料庫進行正負樣本選取。而選取範圍為預測點 T 前 48 小時內的資料，因為此範圍的資料數量可以讓正負樣本的比例達到平衡。

一般而言，正樣本是從心臟驟停患者（CA patients）選取即將發生心臟驟停的資料。如圖 2.2 所示，假設某位患者在時間點 T+1 發生心臟驟停，若患者最後一次測量是落在時間點 T-4 到 T（黃色區間），則將這些資料視為正樣本，代表在這個時間內患者的生命徵象可能出現一些徵兆，因此在下一次測量前發生心臟驟停。這樣設計可防止正樣本的數量過少，也讓模型對於一個時間範圍（prediction window）進行預測才符合常理。最後共選取 3,462 筆正樣本，每筆樣本以矩陣形

式表示，欄（column）為六個生命徵象，列（row）為六個資料點。

2.5.2 負樣本

負樣本有兩種來源，一是從心臟驟停患者選取尚未發生心臟驟停的資料，二是直接選取未心臟驟停的患者資料，本研究將探討不同負樣本組成的模型預測效能。

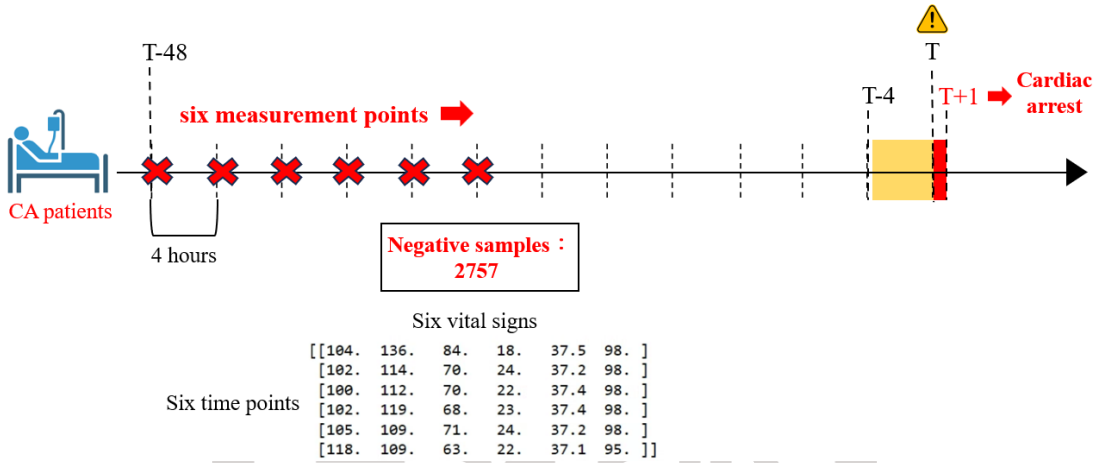

圖 2.3 選取心臟驟停患者之負樣本示意圖

首先是從心臟驟停患者選取尚未發生心臟驟停的資料。如圖 2.3 所示，假設某位患者在時間點 T+1 發生心臟驟停，根據每次選擇六個連續測量點的方式，將時間點 T-48 到 T-4 這個區間的資料視為負樣本，因為在下次醫療人員測量前，患者皆沒有發生心臟驟停事件，最後共選取 2,757 筆負樣本。

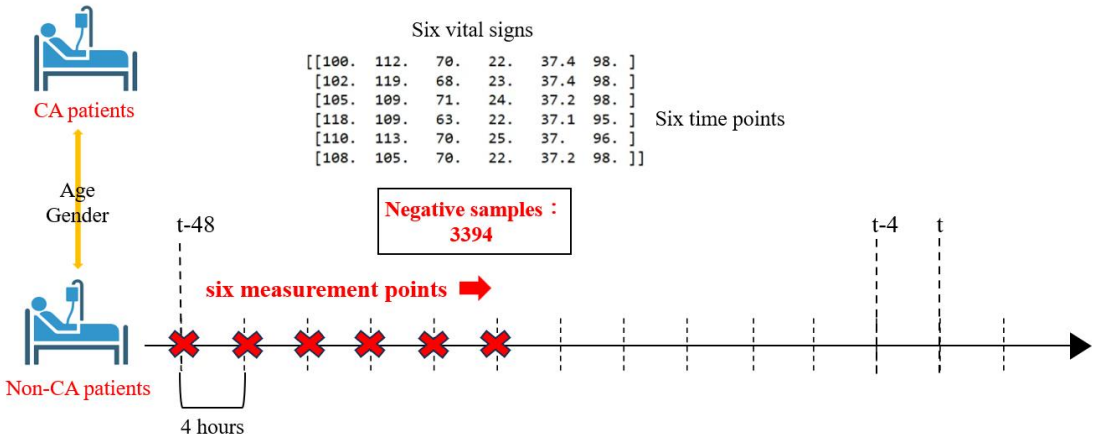

圖 2.4 選取未心臟驟停患者之負樣本示意圖

未心臟驟停患者（Non-CA patients）的選擇方式請參照章節 2.2.2，資料範圍是隨機選取每位患者 44 小時內的生命徵象。如圖 2.4 所示，假設選取到某位患者時間點 t-48 到 t-4 的範圍資料，則將此區間以六個連續測量資料視為一筆負樣本，最後共選取 3,394 筆負樣本。

2.6 預測模型建立

2.6.1 深度模型架構

本研究採用了四種不同的深度模型來處理時間序列資料，分別為 RNN、LSTM、GRU 和 CNN-LSTM。這些模型的不同網路層能夠在接收資料時，同時考慮先前網路層的輸出，可更精確地學習和預測時間序列數據，因此適用於章節 2.5.1 和 2.5.2 所選取的正負樣本。

A. 循環神經網路（Recurrent Neural Network, RNN）

RNN 是一種處理序列數據的神經網路架構，其特色在於神經單元存在循環連接，使 RNN 具有記憶的能力，利用過去的訊息來影響後續的輸出。然而，傳統的 RNN 容易受到梯度消失（gradient vanishing）或梯度爆炸（gradient explored）的影響，限制模型對於長序列數據的學習能力[48]。RNN 的單元

結構和運作原理如圖 2.5，在時間點 t 會傳送新輸入 X_t，隱藏層中的活化函數（activation function）負責結合 X_t 和前一個隱藏狀態 h_{t-1}，產生新的隱藏狀態 h_t。h_t 可直接作為該時間點的模型輸出 O_t，或是傳遞到下一個神經單元與 X_{t+1} 共同產生新的隱藏狀態 h_{t+1}。圖中使用雙曲正切函數（hyperbolic tan, tanh）作為活化函數，將值壓縮到-1 和 1 之間以控制數據的非線性特性。

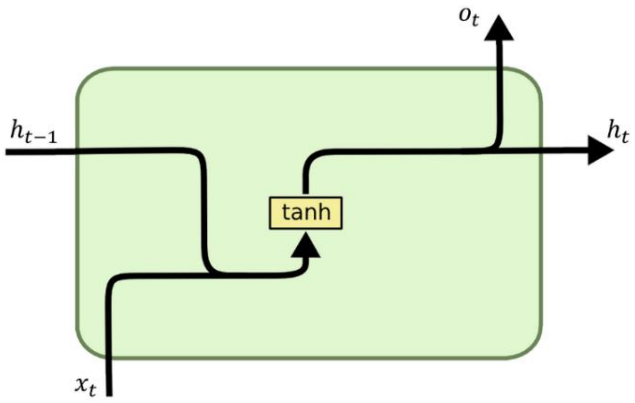

圖 2.5 RNN 單元結構

圖片來源：http://dprogrammer.org/rnn-lstm-gru

B. 長短期記憶網路（Long Short-Term Memory, LSTM）

LSTM 是一種常用於循環神經網路的架構，可改善 RNN 在處理長序列數據容易出現的梯度消失問題[42]。LSTM 透過遺忘門、輸入門和輸出門對模型內部進行訊號的控制，使模型能有效地學習和記憶序列數據中的長期依賴性。LSTM 的單元結構和運作原理如圖 2.6，遺忘門（Forget Gate, f_t）透過新向量 X_t 和前一個隱藏狀態 h_{t-1} 輸入 sigmoid 函數，產生一個 0 到 1 之間的值，此數值會與 C_{t-1} 相乘，因此遺忘門可決定前一個記憶單元 C_{t-1} 的哪些訊息需要被遺忘。輸入門（Input Gate, i_t）透過 X_t 和 h_{t-1} 輸入 sigmoid 函數，與經過 tanh 函數的結果相乘，所產生的候選記憶單元 $\tilde{C}_t$ 再與 C_{t-1} 進行相加，產生 C_t，因此輸入門可決定 X_t 和 h_{t-1} 的哪些訊息需要添加到記憶單元 C_t。輸出門（Output Gate, O_t）將 X_t 和 h_{t-1} 輸入 sigmoid 函數與 C_t 輸入 tanh 函數的結果相乘，產生隱藏狀態 h_t，因此輸出門可決定 C_t 的哪些訊息會被傳送到 h_t，或

是直接作為模型的輸出。

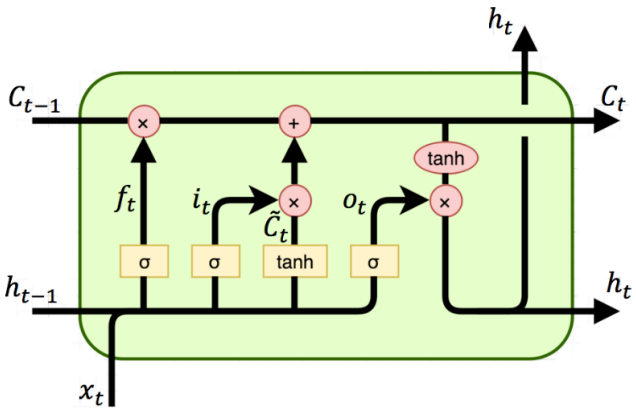

圖 2.6 LSTM 單元結構

圖片來源：http://dprogrammer.org/rnn-lstm-gru

C. 門控循環單元（Gated Recurrent Unit, GRU）

GRU 是一種常用於循環神經網路的架構，一樣可解決 RNN 容易出現的梯度消失問題[49]。GRU 透過重置門和更新門來控制訊號的傳輸。相較於 LSTM，GRU 少了一個閘門可提高執行速度與減少記憶體的用量。GRU 的單元結構和運作原理如圖 2.7，重置門（reset gate, r_t）透過新向量 X_t 和前一個隱藏狀態 h_{t-1} 輸入 sigmoid 函數，產生一個 0 到 1 之間的值，此數值會與 h_{t-1} 相乘，因此重置門可控制 h_{t-1} 對後續隱藏狀態的影響程度。更新門（update gate, z_t）將 X_t 和 h_{t-1} 輸入 sigmoid 函數，此數值會被 1 相減產生 1- z_t。而重置後的 h_{t-1} 與 X_t 輸入 tanh 函數形成候選隱藏狀態 $\tilde{h}_t$；$\tilde{h}_t$ 與 z_t 相乘，再加上 h_{t-1} 與 1- z_t 相乘，最終產生隱藏狀態 h_t。因此更新門可決定多少 h_{t-1} 需要被保留以及多少 $\tilde{h}_t$ 需要被添加到 h_t 中。

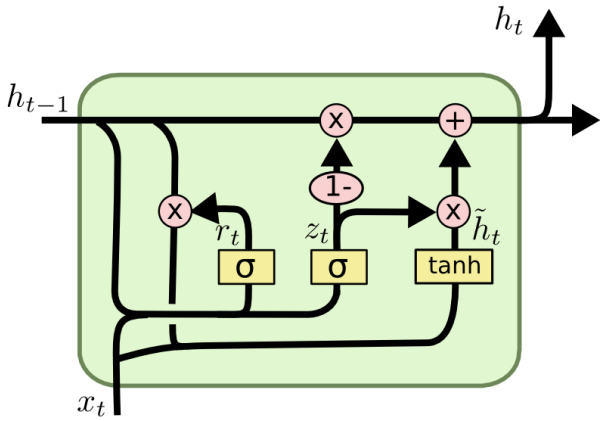

圖 2.7 GRU 單元結構

圖片來源：http://dprogrammer.org/rnn-lstm-gru

D. 卷積神經網路-長短期記憶網路（Convolutional Neural Network-Long Short-Term Memory, CNN-LSTM）

圖 2.8 呈現 CNN 結合 LSTM 的模型。首先，序列資料透過輸入層進入卷積層（convolutional layer），卷積層可從資料中提取重要的特徵，接著池化層（pooling layer）會去除資料雜質、減少資料維度，有助於降低計算複雜度和過擬合。這些資料經過 LSTM 單元處理後，會經過全連接層（fully connected layer）執行進一步計算，以統整學習到的特徵。最後，模型輸出預測可以是二元或多元分類的結果[50]。這種模型架構對於局部特徵（由 CNN 處理）及時間序列（由 LSTM 處理）的資料會產生不錯的預測結果。

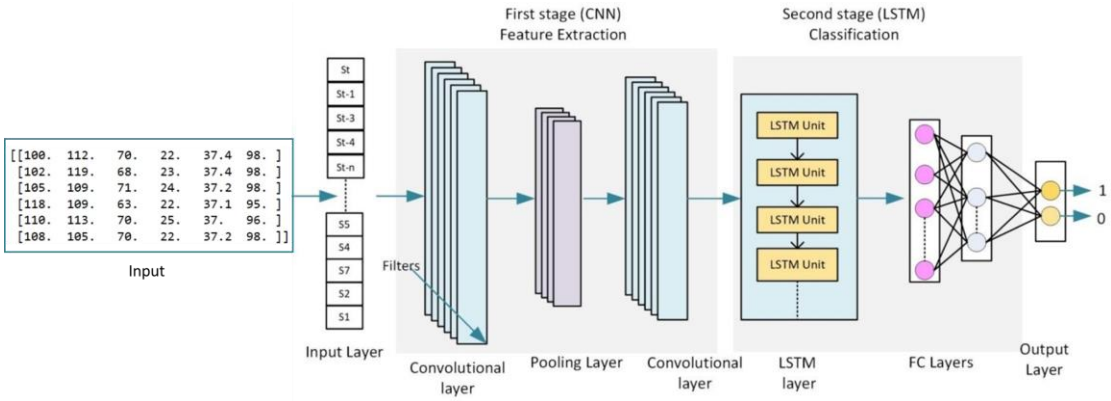

圖 2.8 CNN-LSTM 模型結構[50]

2.6.2 超參數設定與執行環境

模型的超參數經過多次測試後進行決定。批量大小（batch_size）為 64，優化器（optimizer）為 Adam（adaptive moment estimation），學習率（learning rate）為 0.001，損失函數（loss function）為二元交叉熵（binary_crossentropy），epoch 為 100。

本研究使用 Python 3.8.5 進行資料前處理，以及使用 Tensorflow 2.13.0 進行模型建構，並使用 matplotlib 3.3.2 進行視覺化分析。

2.6.3 自助抽樣法（bootstrap）

本研究採用 bootstrap 抽樣方法來測試模型的穩定性。Bootstrap 是一種利用有限樣本進行隨機且有放回的重複抽樣，能夠有效反映整體母體分布特性的方法。具體來說，透過 bootstrap 從訓練資料集（training dataset）中隨機抽取固定數量的訓練資料（training data），因為是有放回抽樣，所以某些樣本可能會被多次選中，而未被抽到的樣本則作為該次 bootstrap 的驗證資料（validation data）。因此，驗證資料的數量會依取後放回的抽樣結果而定。

如圖 2.9 所示，訓練資料集包含 10 個樣本，編號從 1 到 10，每個樣本被抽到的機率相同。第一次抽樣形成第一子樣本（bootstrap 1），抽到的編號 1、3、3、4、7 視為訓練資料 1（training data 1），其中編號 3 被抽到兩次；而未抽到的編號 2、5、6、8、9、10 則視為驗證資料 1（validation data 1）。第二次抽樣形成第二子樣本（bootstrap 2），抽到的編號 2、4、5、9、10 視為訓練資料 2（training data 2）；而未被抽到的編號 1、3、6、7、8 則視為驗證資料 2（validation data 2）。由此可知，bootstrap 抽樣方法可利用訓練資料集來生成多筆新的訓練資料和相對應的驗證資料。此外，bootstrap 透過反覆從同一資料集中抽樣，有助於減少數據變異性，因此它的優點是不需要對資料分布作出任何假設，適用於分析較多異常

值或者偏離常態分佈的數據集，進而提高推論統計的準確性[51]。

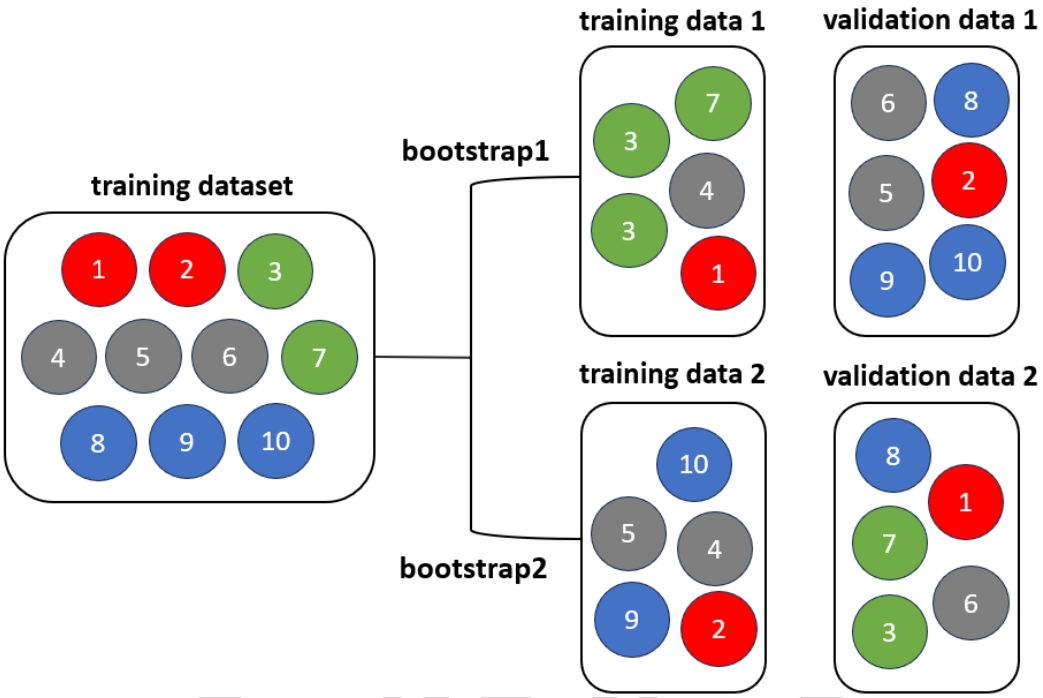

圖 2.9 bootstrap 抽樣示意圖

2.7 模型評估

為了評估模型的預測效能，我們根據預測結果與實際結果的比對來生成混淆矩陣（confusion matrix）。如圖 2.10 所示，混淆矩陣包含以下四個關鍵元素：真陽性（True Positive, TP）、真陰性（True Negative, TN）、偽陽性（False Positive, FP），亦稱為第一型錯誤（Type I error）、偽陰性（False Negative, FN），亦稱為第二型錯誤（Type II error）。這些元素是衡量各項性能指標的基礎，所計算出的指標數值介於 0 到 1 之間，數值越接近 1 表示模型在該項指標上的預測效能越高。本研究採用以下幾項指標來評估模型的表現：

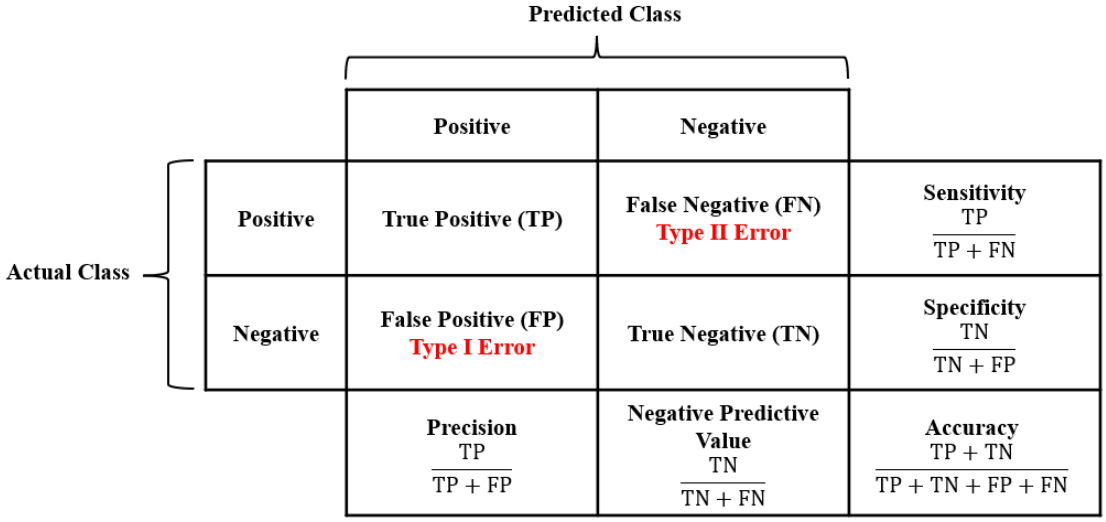

圖 2.10 混淆矩陣

A. 準確度（accuracy）：指在所有情況下，正負樣本被預測正確的比例。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

B. 精確度(precision)：指在預測為正樣本的情況下，正樣本被預測正確的比例。本研究中為模型預測會發生心臟驟停的資料，真正有發生的比例。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

C. 召回率（recall）：又稱為靈敏度（sensitivity）或是真陽率（True Positive Rate, TPR）。指所有正樣本被預測正確的比例，recall 越高，代表模型越能正確判斷正樣本。在本研究中為模型正確預測發生心臟驟停資料的比例。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

D. 特異度（specificity）：指所有負樣本被預測正確的比例，specificity 越高，代表模型越能正確判斷負樣本。在本研究中為模型正確預測沒發生心臟驟停資料的比例。

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

E. F1-score：為 precision 和 recall 的加權平均，適用於類別不平衡的情況。例如模型可能預測大多數為負樣本，容易產生較高 precision 或 recall，但並不代表模型的表現良好，大量的 true negative 可能會造成評估的偏差，因此透過 F1-score 平衡 precision 和 recall 來確認模型的整體性能。

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F. AUROC：ROC 曲線（Receiver Operating Characteristic curve）以模型在各閾值（threshold）產生的 false positive rate（FPR, 1-specificity）和 true positive rate（TPR）繪製而成。如圖 2.11 所示，FPR 為橫軸、TPR 為縱軸所形成的 ROC 曲線下面積為 AUROC，用來衡量模型對於正負樣本的判斷能力。AUROC 越高，代表模型越擅長區分兩個類別。

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

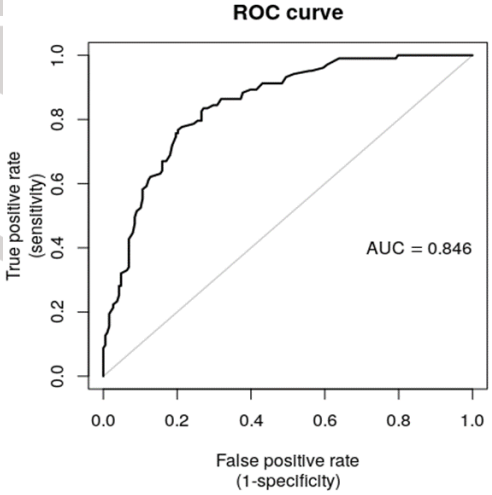

圖 2.11 AUROC 示意圖

2.8 統計分析

本研究使用 Shapiro-Wilk test 進行常態檢定。要檢測兩個獨立群體或樣本的平均是否存在顯著差異，如果資料呈常態分佈就使用獨立樣本 t 檢定（independent t test），否則就使用 Wilcoxon Rank Sum 檢定。

2.9 NEWS 計分

本研究使用章節 2.5 所選取的每一筆正負樣本，利用 NEWS 分數計算表（圖 1.2）進行計分。以 NEWS 警示閾值對照表為判斷標準（圖 1.3），若某樣本最後一筆生命徵象的總分大於五分以上，代表 NEWS 判定該樣本為正樣本，標記（label）為 1。因此透過正負樣本標記和 NEWS 預測標記，可將每筆樣本分類為 TP、TN、FP 和 FN（如圖 2.12 所示），並計算章節 2.7 所提及的各模型評估指標。

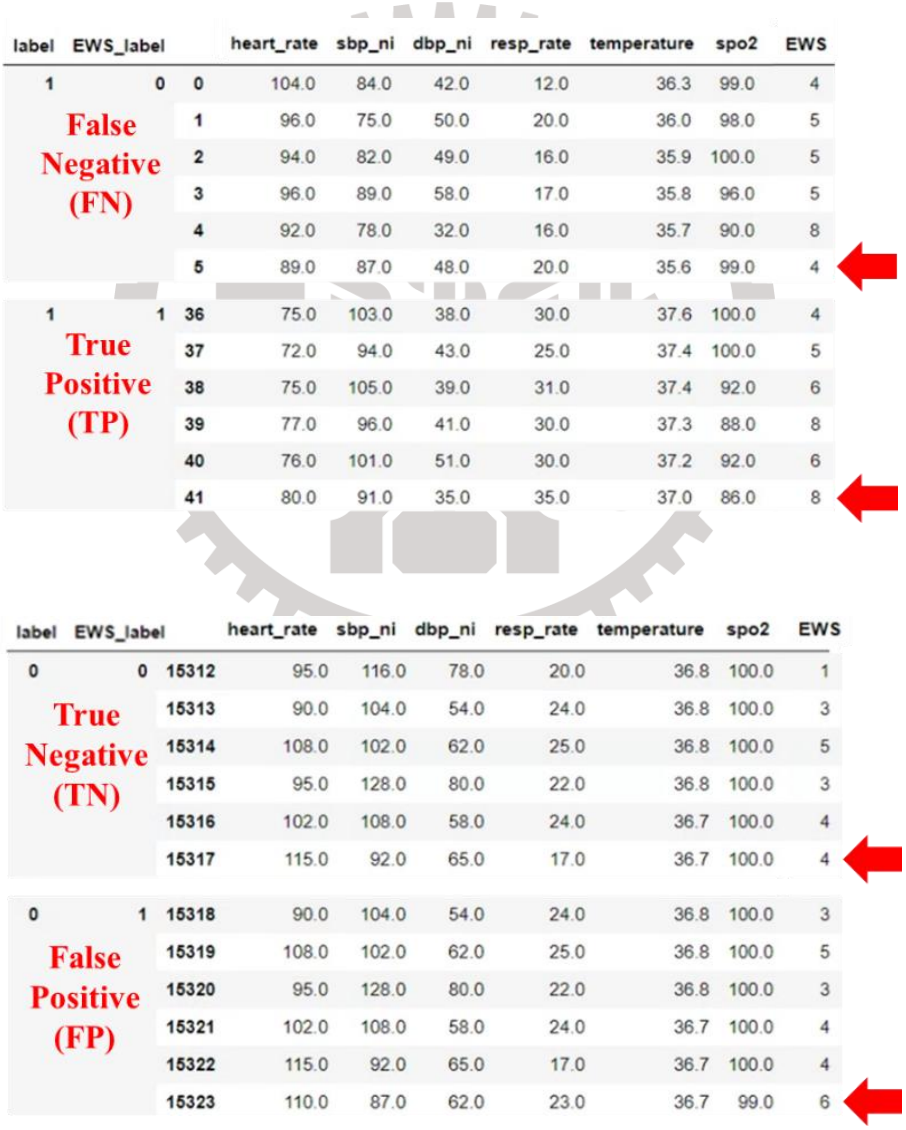

圖 2.12 正負樣本進行 NEWS 分數計算

2.10 特徵重要性

本研究使用 SHAP 進行各模型的特徵重要性分析。SHAP（SHapley Additive exPlanations）是一種以博弈論為基礎的機器學習解釋工具，它利用合作博弈中的 Shapley 值來解釋個別預測的結果[52]。Shapley 值原本被用於合作博弈中的公平分配收益，而在機器學習中，不論特徵的數量或模型的複雜性如何，可透過此方法衡量每個輸入特徵對模型預測的具體貢獻，使複雜的模型變得更透明和易於解釋。圖 2.13 為多種 SHAP 視覺化表示圖其中之一，橫軸的平均絕對值（SHAP value）表示各個特徵對於預測結果的重要性，其中 Relationship 的數值最大，代表此特徵對模型的預測影響最大，其餘特徵的影響依次遞減。透過分析特徵在所有可能組合中的貢獻，SHAP 不僅能夠提供對單個預測的詳細解釋，還能評估特徵對整個資料集的預測影響，提供局部和全局的解釋[53]。

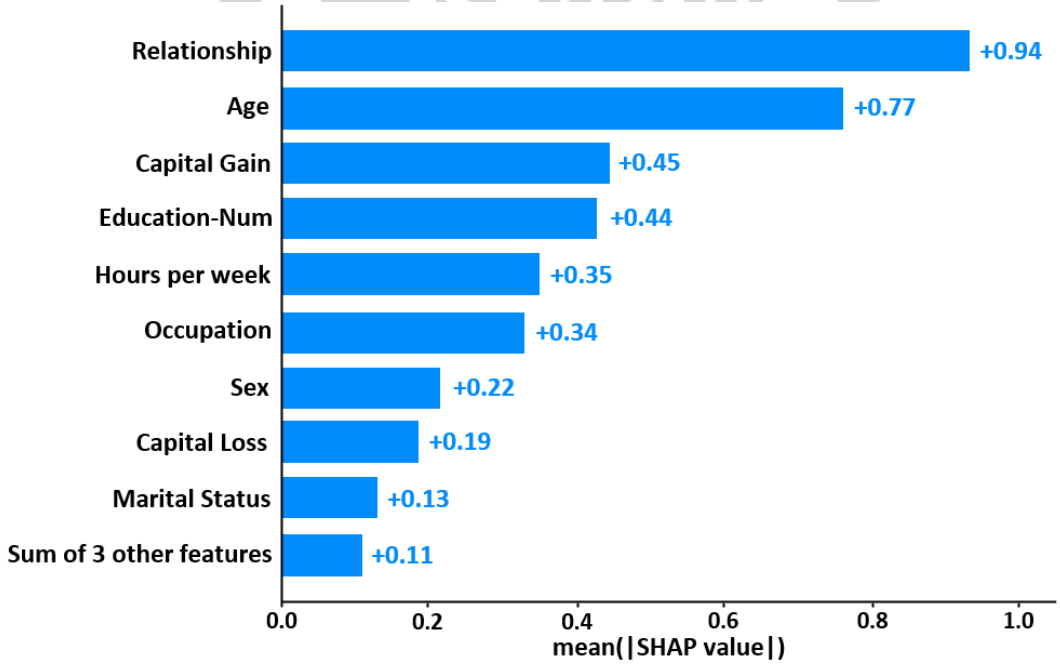

圖 2.13 SHAP 分析示意圖

圖片來源：

https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

第三章　研究結果

3.1　研究流程

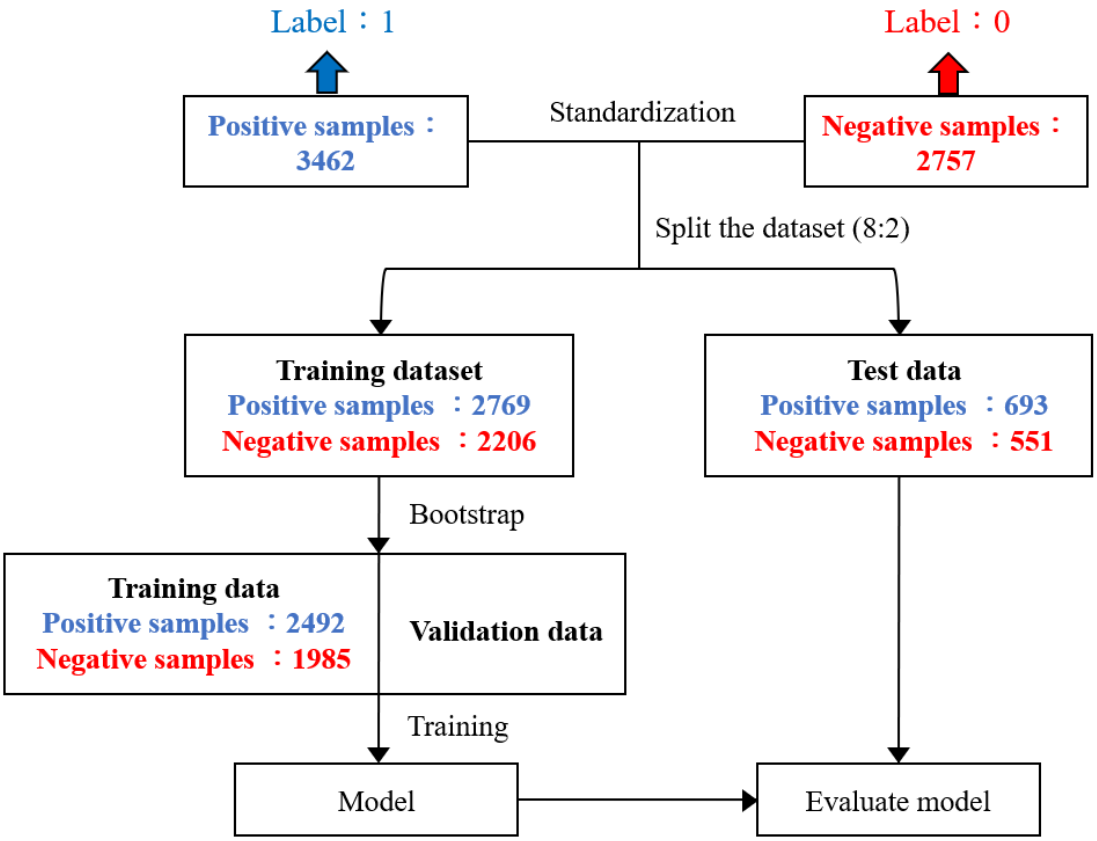

圖 3.1 研究流程。圖中以正負樣本皆取自心臟驟停患者資料為例

本研究流程如圖 3.1，模型進行二元分類（binary classification）的預測，因此將正樣本標記為 1，負樣本標記為 0。若模型預測結果為 1，代表模型認為這筆資料後續會發生心臟驟停，若預測結果為 0，代表模型認為這筆資料後續不會發生心臟驟停。而每筆正負樣本資料會進行標準化（Z-score standardization）的處理，標準化可將不同單位的數值轉換為具有相同單位的標準形式，讓數值具有平均值為 0 和標準差為 1 的分布[54]。

以正負樣本皆取自心臟驟停患者資料為例，首先將資料切分成 80% 訓練資料集（training dataset）和 20% 測試資料（test data）。為了測試模型的穩定度，針對 4975 筆訓練資料集進行 30 次 bootstrap 抽樣，每次 bootstrap 皆從訓練資料

集中隨機抽取 90% 訓練資料（training data），因為是進行取後放回的抽樣，抽到的 4477 筆訓練資料中可能會有重複資料，未被抽到的資料則作為驗證資料。因此，驗證資料數量會受抽樣結果的影響而有所變動，最後將每次 bootstrap 取樣出來的資料進行模型訓練。透過各模型在每次訓練所得到的結果，分別計算 30 次損失值（loss）的平均跟標準差，最後選擇驗證集損失值（validation loss）最低的權重（weight）作為最終模型，並利用測試資料進行模型評估。以上完整的訓練與驗證流程會套用於不同模型和資料，詳細的資料配置數量請參照附錄 B。

3.2 模型穩定性

在訓練模型時，隨著參數不斷被更新，模型預測結果和實際結果越接近，所計算出的損失值（loss）會逐漸降低，因此曲線呈現由左上逐漸往右下分佈。圖 3.2 為正負樣本皆取自心臟驟停患者資料所訓練的模型，在 30 次 bootstrap 的平均訓練損失和平均驗證損失隨著訓練週期（epochs）的變化結果。圖中的紅色虛線代表平均訓練損失（mean training loss），藍色虛線代表平均驗證損失（mean validation loss）。這兩條曲線都顯示隨著訓練週期的增加，損失逐漸下降且趨於平穩，代表模型在訓練過程中對數據的學習越來越良好。而圖中的陰影區域為損失的標準差，陰影越小代表模型在不同樣本上的表現越一致、穩定性越好。

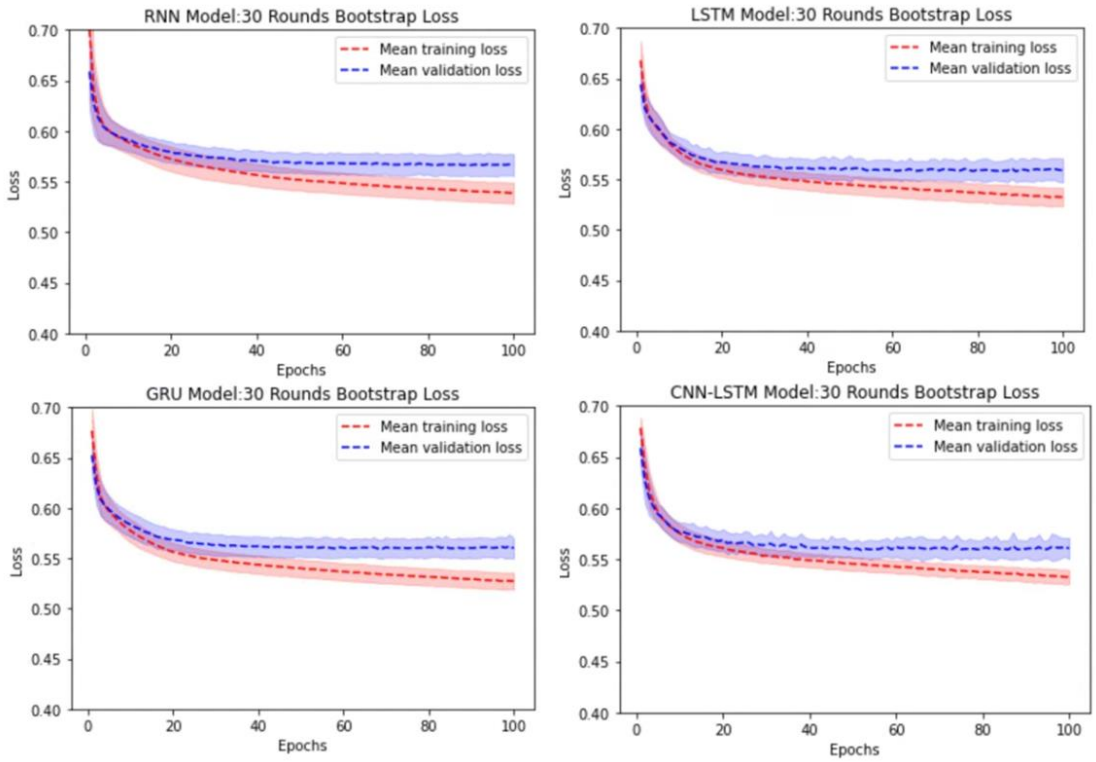

圖 3.2 各模型於訓練過程所計算之損失值（loss）

3.3 資料人口學及特徵分析

表 3.1 MIMIC IV 中，心臟驟停患者和未心臟驟停患者之人口學及特徵分析

Characteristics	**MIMIC IV (CA)**	**MIMIC IV (Non-CA)**	**p-value**
Number of total patients, n	477	477	
Number of total vital signs, n	68,973	55,769	
Age, y , mean ± SD	65.58 ± 15.91	65.58 ± 15.91	
Male ,sex , n (%)	289 (60.59%)	289 (60.59%)	
Features, mean ± SD			
SBP (mmHg)	110.69 ± 22.06	119.81 ± 22.60	$p < 0.05$
DBP (mmHg)	61.96 ± 15.74	69.80 ± 513.42	$p < 0.05$
HR (/min)	89.32 ± 20.88	84.74 ± 17.85	$p < 0.05$
RR (/min)	21.79 ± 6.45	19.80 ± 5.54	$p < 0.05$
BT (℃)	36.94 ± 1.04	37.00 ± 1.24	$p < 0.05$
SpO_2 (%)	96.50 ± 4.83	96.64 ± 5.09	$p < 0.05$

如表 3.1 所示，本研究從 MIMIC IV 資料庫收集 479 起心臟驟停事件，包含 477 位患者和 68,973 筆生命徵象記錄，該組別的平均年紀為 65.58 歲，男性的發生率較高(60.59% vs. 39.41%)。而未心臟驟停的患者共篩選出 477 位，其中包含 55,769 筆生命徵象記錄，該組別與心臟驟停患者的年紀和性別做篩選配對，因此平均年紀和性別比呈現相同結果。在生命徵象部分，相較於未心臟驟停的組別，發生心臟驟停的組別有較低的收縮壓（110.69 ± 22.06 mmHg vs. 119.81 ± 22.60 mmHg）、舒張壓（61.96 ± 15.74 mmHg vs. 69.80 ± 513.42 mmHg）、體溫（36.94 ± 1.04℃ vs. 37.00 ± 1.24℃）、血氧飽和度（96.50 ± 4.83% vs. 96.64 ± 5.09%），以及較高的心率（89.32 ± 20.88 次/min vs. 84.74 ± 17.85 次/ min）、呼吸頻率（21.79 ± 6.45 次 / min vs. 19.80 ± 5.54 次/ min）。

3.4 不同負樣本與不同模型組成之比較

3.4.1 負樣本取自心臟驟停患者資料所訓練之模型

本研究於 MIMIC IV 資料庫選取的正負樣本數量相近，因此會依 AUROC 的高低作為判斷不同模型預測的能力。表 3.2 呈現四種使用負樣本取自心臟驟停患者資料所訓練的模型，在訓練和驗證階段的各指標結果，包含 accuracy、precision、sensitivity、specificity、F1-score 和 AUROC。表 3.3 為各模型在測試階段的指標結果。由結果得知，所有模型在測試階段皆表現相當一致，AUROC 保持在 0.78、0.79，表示這些模型在區分正負樣本上具有良好的能力，其中 GRU、CNN-LSTM 的 AUROC 皆為 0.79，略高於 RNN 和 LSTM。

表 3.2 負樣本取自心臟驟停患者資料所訓練模型之訓練和驗證結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN training	0.72 ± 0.01	0.79 ± 0.01	0.67 ± 0.01	0.89 ± 0.01	0.71 ± 0.00	0.79 ± 0.01
RNN validation	0.70 ± 0.01	0.78 ± 0.03	0.65 ± 0.03	0.88 ± 0.02	0.71 ± 0.01	0.77 ± 0.01
LSTM training	0.72 ± 0.01	0.79 ± 0.01	0.67 ± 0.01	0.89 ± 0.01	0.71 ± 0.00	0.79 ± 0.01
LSTM validation	0.71 ± 0.01	0.78 ± 0.03	0.66 ± 0.03	0.89 ± 0.02	0.71 ± 0.01	0.78 ± 0.02
GRU training	0.72 ± 0.00	0.79 ± 0.01	0.68 ± 0.01	0.90 ± 0.01	0.71 ± 0.00	0.79 ± 0.01
GRU validation	0.71 ± 0.01	0.77 ± 0.02	0.67 ± 0.03	0.89 ± 0.02	0.71 ± 0.01	0.78 ± 0.01
CNN-LSTM training	0.72 ± 0.01	0.78 ± 0.01	0.68 ± 0.02	0.89 ± 0.01	0.71 ± 0.00	0.79 ± 0.01
CNN-LSTM validation	0.71 ± 0.01	0.78 ± 0.03	0.66 ± 0.04	0.89 ± 0.01	0.71 ± 0.01	0.78 ± 0.01

表 3.3 負樣本取自心臟驟停患者資料所訓練模型之測試結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN	0.69	0.81	0.62	0.79	0.70	0.78
LSTM	0.70	0.78	0.68	0.73	0.72	0.78
GRU	0.70	0.81	0.63	0.80	0.71	0.79
CNN-LSTM	0.70	0.77	0.67	0.73	0.72	0.79

3.4.2 負樣本取自未心臟驟停患者資料所訓練之模型

表 3.4 呈現四種使用負樣本取自未心臟驟停患者資料所訓練的模型，在訓練和驗證階段的各指標結果。表 3.5 為各模型在測試階段的指標結果。由結果得知，所有模型在測試階段一樣皆表現相當一致，AUROC 為 0.88、0.89，表示這些模型在區分正負樣本上具有更好的能力，其中 LSTM、GRU 的數值為 0.89，略高於 RNN 和 CNN-LSTM。此外，該組別的模型相較於負樣本取自心臟驟停患者資料所訓練的模型（表 3.2、表 3.3），其整體表現還要來的好。

表 3.4 負樣本取自未心臟驟停患者資料所訓練模型之訓練和驗證結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN training	0.78 ± 0.01	0.80 ± 0.01	0.76 ± 0.01	0.97 ± 0.01	0.67 ± 0.00	0.87 ± 0.01
RNN validation	0.77 ± 0.01	0.80 ± 0.02	0.74 ± 0.04	0.96 ± 0.01	0.67 ± 0.01	0.86 ± 0.01
LSTM training	0.79 ± 0.01	0.81 ± 0.01	0.76 ± 0.01	0.97 ± 0.00	0.67 ± 0.00	0.88 ± 0.01
LSTM validation	0.78 ± 0.01	0.80 ± 0.02	0.75 ± 0.04	0.97 ± 0.01	0.67 ± 0.02	0.87± 0.01
GRU training	0.79 ± 0.01	0.82 ± 0.01	0.76 ± 0.01	0.97 ± 0.00	0.67 ± 0.00	0.88 ± 0.01
GRU validation	0.78 ± 0.01	0.81 ± 0.03	0.76 ± 0.05	0.97 ± 0.01	0.67 ± 0.01	0.88 ± 0.01
CNN-LSTM training	0.79 ± 0.01	0.81 ± 0.01	0.77 ± 0.01	0.97 ± 0.00	0.67 ± 0.00	0.88 ± 0.00
CNN-LSTM validation	0.79 ± 0.01	0.82 ± 0.03	0.75 ± 0.05	0.97 ± 0.01	0.67 ± 0.01	0.88 ± 0.01

表 3.5 負樣本取自未心臟驟停患者資料所訓練模型之測試結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN	0.78	0.81	0.77	0.81	0.79	0.88
LSTM	0.79	0.83	0.76	0.83	0.79	0.89
GRU	0.79	0.87	0.70	0.88	0.77	0.89
CNN-LSTM	0.79	0.83	0.75	0.83	0.79	0.88

3.4.3 負樣本取自心臟驟停患者和未心臟驟停患者資料訓練之模型

考量模型在進行預測時，無法得知患者後續是否會發生心臟驟停，因此在負樣本的選擇上，調整為各取一半發生心臟驟停和未發生心臟驟停的患者資料，進行模型的建構。表 3.6 呈現四種使用負樣本取自心臟驟停患者和未心臟驟停患者各半資料所訓練的模型，在訓練和驗證階段的各指標結果。表 3.7 為各模型在測試階段的指標結果。由結果得知所有模型的 AUROC 表現十分接近，代表它們在區分正負樣本的能力相差不大。此外，從這兩群患者中各取一半的資料作為負樣本，因此該組別的模型預測表現會介於上述兩組模型之間，AUROC 保持在 0.80 到 0.82，其中以 GRU 具有最好的模型性能。

表 3.6 負樣本取自心臟驟停患者和未心臟驟停患者資料所訓練模型之訓練和驗證結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN training	0.74 ± 0.01	0.77 ± 0.01	0.73 ± 0.01	0.92 ± 0.01	0.69 ± 0.00	0.82 ± 0.01
RNN validation	0.73 ± 0.01	0.77 ± 0.03	0.71 ± 0.03	0.91 ± 0.01	0.70 ± 0.01	0.81 ± 0.01
LSTM training	0.74 ± 0.00	0.77 ± 0.01	0.73 ± 0.01	0.93 ± 0.01	0.69 ± 0.00	0.83 ± 0.00
LSTM validation	0.73 ± 0.01	0.75 ± 0.03	0.73 ± 0.03	0.92 ± 0.01	0.69 ± 0.01	0.82 ± 0.01
GRU training	0.75 ± 0.00	0.77 ± 0.01	0.73 ± 0.01	0.93 ± 0.01	0.69 ± 0.00	0.83 ± 0.01
GRU validation	0.74 ± 0.01	0.78 ± 0.02	0.72 ± 0.03	0.93 ± 0.01	0.69 ± 0.01	0.82 ± 0.01
CNN-LSTM training	0.74 ± 0.00	0.77 ± 0.01	0.74 ± 0.01	0.93 ± 0.01	0.69 ± 0.00	0.83 ± 0.00
CNN-LSTM validation	0.73 ± 0.01	0.76 ± 0.02	0.73 ± 0.04	0.92 ± 0.01	0.69 ± 0.01	0.82 ± 0.01

表 3.7 負樣本取自心臟驟停患者和未心臟驟停患者資料訓練模型之測試結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN	0.72	0.76	0.71	0.74	0.73	0.80
LSTM	0.73	0.78	0.70	0.76	0.73	0.80
GRU	0.73	0.78	0.70	0.77	0.74	0.82
CNN-LSTM	0.73	0.78	0.69	0.76	0.73	0.81

綜合不同負樣本的組成應用在不同模型的實驗結果，這四種深度學習的模型（RNN, LSTM, GRU, CNN-LSTM）在各個指標上表現相似且沒有明顯差異，代表模型的架構變化對於本研究所使用的資料影響並不大。從上述的測試結果可看出，GRU 在多個指標以及不同資料均展現較高的準確度和穩定性，特別是區分正負樣本的能力，AUROC 幾乎達到 0.80 以上。考慮 GRU 是針對 RNN 的缺點做改良，以及 GRU 的單元結構少一個閘門，相較於 LSTM 和 CNN-LSTM 所設

定的參數少、執行速度快，因此後續實驗選用較高效的 GRU 更為合適。

負樣本以未心臟驟停患者資料所建立的模型，具有最好的預測整體表現。但考量到負樣本以未來可能發生以及不會發生心臟驟停的患者資料所組成，不僅讓樣本來源符合現實層面，也可使模型學習到更多元的生命徵象趨勢，對於心臟驟停的預測表現也沒有下降太多，因此本研究對於負樣本傾向選擇各取一半心臟驟停患者和未心臟驟停患者的資料。除此之外，後續會透過馬偕醫院一般病房的臨床資料進行部分外部驗證，確認模型在不同的資料集也具備相當的可解釋性（interpretability）和通用能力（generalizability）[55]。也確保此負樣本組成可以幫助模型達到一定的效能和實際應用。外部驗證的結果於附錄 C 和章節 4.5 進行更詳細的描述和討論。

3.5 NEWS 與 GRU 之比較

為了瞭解本研究建立的 GRU 模型，是否比傳統的早期預警系統有更好的預測效果，因此使用無法變更評估項目和範圍的 NEWS（National Early Warning System）進行比較。表 3.8 和表 3.9 分別呈現 NEWS 和 GRU 在不同負樣本組成下的預測表現，half 代表取自心臟驟停患者及未心臟驟停患者的各半資料。從結果來看，NEWS 針對負樣本取自未心臟驟停患者資料的預測效果一樣較好，AUROC 為 0.72。而面對不同負樣本組成的資料中，GRU 的 AUROC 皆達到 0.80 以上，而 NEWS 的 AUROC 僅介於 0.59 至 0.72 間，GRU 模型預測表現皆明顯優於 NEWS，代表 GRU 能夠有效地辨識各種情況下的心臟驟停風險，因此 GRU 可能成為比 NEWS 更有效的預測工具。

表 3.8 不同負樣本組成資料應用於 NEWS 之預測結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
NEWS_CA	0.59	0.60	0.60	0.59	0.60	0.59
NEWS_Non-CA	0.73	0.75	0.60	0.84	0.67	0.72
NEWS_half	0.67	0.67	0.60	0.74	0.63	0.67

表 3.9 不同負樣本組成資料應用於 GRU 之預測結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
GRU_CA	0.73	0.80	0.65	0.81	0.72	0.80
GRU_Non-CA	0.80	0.82	0.78	0.82	0.77	0.88
GRU_half	0.75	0.78	0.68	0.81	0.73	0.83

3.6 特徵重要性

本研究使用 SHAP 來分析各特徵的重要性，瞭解這些生命徵象對於模型預測的貢獻度，以提高模型在臨床應用中的接受度和價值。GRU 在不同負樣本所組成的資料當中，AUROC 的表現最為穩定，因此使用此模型進行特徵重要性的分析。圖 3.3 為負樣本取自心臟驟停患者資料所建構 GRU 之分析結果，重要的前三個特徵為體溫、心率、呼吸頻率；圖 3.4 為負樣本取自未心臟驟停的患者資料所建構 GRU 之分析結果，重要的前三個特徵為體溫、呼吸頻率、血氧飽和度。圖 3.5 負樣本取自心臟驟停患者和未心臟驟停患者資料所建構 GRU 之分析結果，重要的前三個特徵為體溫、血氧飽和度、呼吸頻率。因此綜合這些研究結果，體溫和呼吸頻率是對 GRU 模型預測最具影響力的特徵。

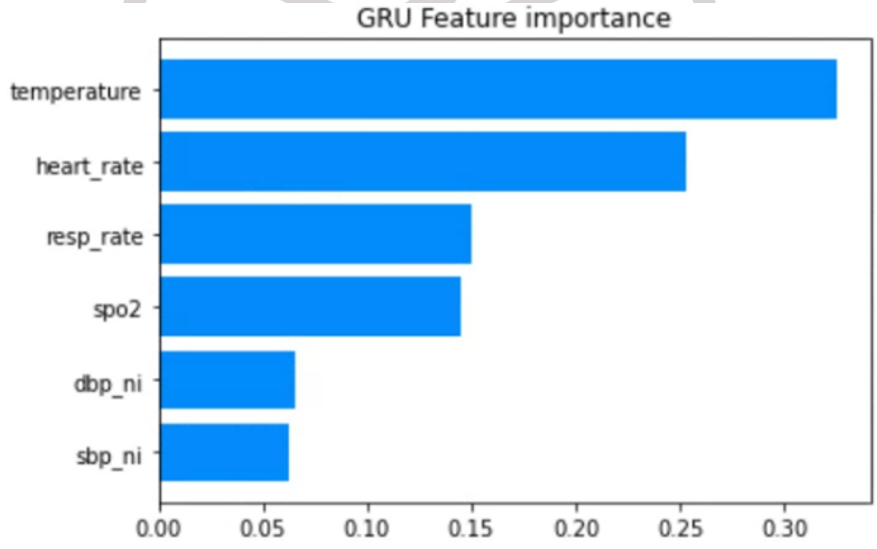

圖 3.3 負樣本取自心臟驟停患者資料所建構 GRU 之特徵重要性

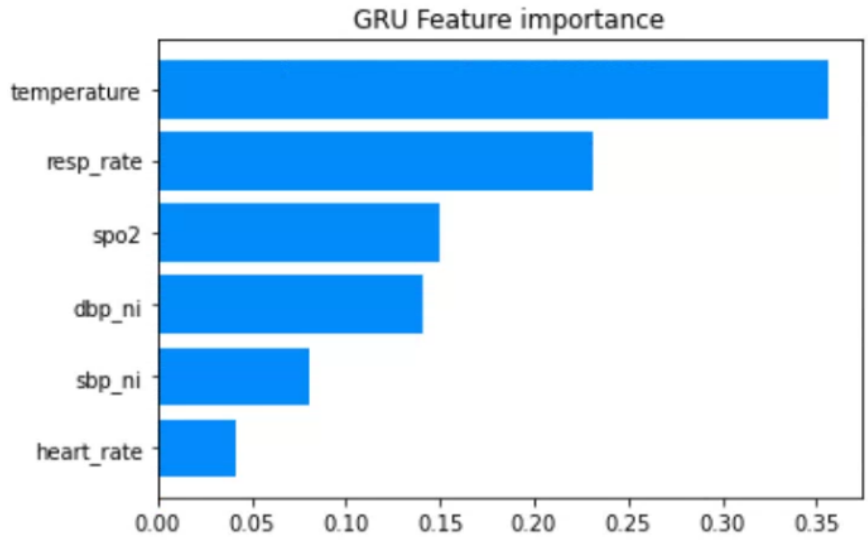

圖 3.4 負樣本取自未心臟驟停患者資料所建構 GRU 之特徵重要性

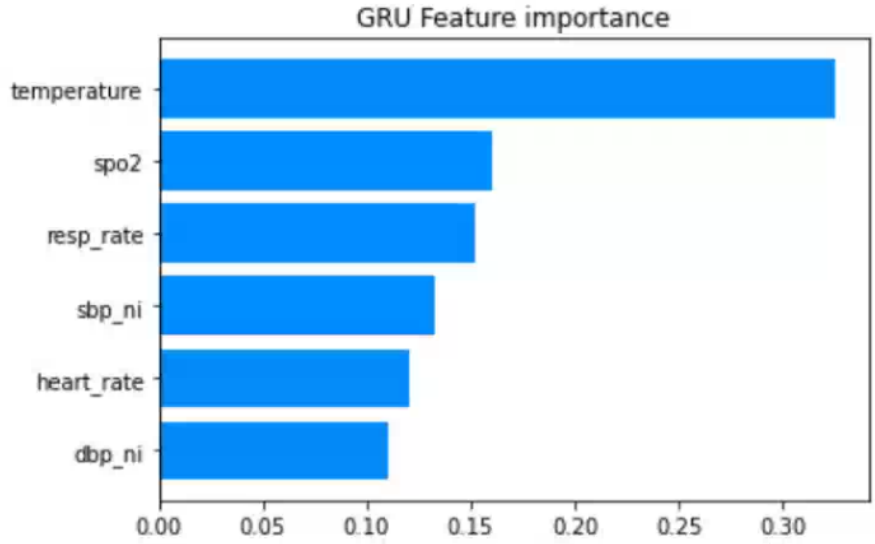

圖 3.5 負樣本取自心臟驟停患者和未心臟驟停患者資料所建構 GRU 之特徵重要性

第四章　討論

4.1　資料人口學及特徵分析

本研究利用 MIMIC IV 資料庫的統計顯示，心臟驟停的發生在男性中比例較高，整體平均年齡大約 65 歲。心臟驟停在性別上的顯著差異已被多項研究證實，Perman 等人的研究表示心臟驟停案例中男性佔六成，平均年齡為 66 歲[36]。Kim 等人的研究也發現，男性的心臟驟停發生率明顯高於女性，且男性發生年齡通常又比女性小，這可能與男性的生理差異和生活在高風險因素的環境中相關[56]。此外，隨著年齡的增加，人類的心血管系統會逐漸老化、脆弱，加上各種慢性病的累積，使得老年人更容易受到心臟驟停的威脅[57]，因此患者年齡越大，發生心臟驟停的機率越高。這些結果與本研究呈現相似的性別比例和年齡分布，並強調男性和年齡大於 60 歲是心臟驟停中的高風險條件。

4.2　不同負樣本組成之比較

在進行心臟驟停模型預測的研究中，探討正負樣本的組成對模型表現有一定的重要性。根據本研究的測試結果，若正負樣本皆取自於心臟驟停的患者資料，模型表現略差一些。在這種實驗設計下，正負樣本間的差異可能相對較小，模型需要在較細微的變化中評估生命徵象，增加了預測的難度。另一方面，若正樣本取自心臟驟停患者資料，負樣本取自於未心臟驟停的患者資料，模型表現有顯著提升。因為這兩組患者的生理狀況可能存在較大的差異，使模型容易進行正確的分類預測。為了更接近臨床現實，若正樣本取自心臟驟停患者資料，負樣本取自心臟驟停患者和未心臟驟停患者的各半資料，模型表現可能介於上述兩種情況之間。這類型的負樣本組成不但能反應真實的臨床情境，又能在一定程度上提高預測的準確性。

目前預測心臟驟停模型所訓練的負樣本來源，多數取自未心臟驟停的患者資

料[58, 59]。然而，根據本研究的外部驗證（附錄 C 和章節 4.5）得知，馬偕醫院一般病房的心臟驟停患者資料應用在不同負樣本組成所訓練的模型中，驗證結果皆取得不錯的預測表現，彼此間亦沒有明顯差距。因此針對負樣本的組成，本研究仍然選擇心臟驟停患者和未心臟驟停患者的各半資料，期望能讓模型適應於不同資料集，並在預測效果和實際情形間取得平衡。

4.3 NEWS 與 GRU 之比較

本研究建立用於預測心臟驟停的 GRU 模型，並與 NEWS 進行比較。結果顯示，GRU 在多項性能指標上均優於 NEWS，因為 GRU 可透過具有時間序列的生命徵象數據，更瞭解患者狀態的變化趨勢。相比之下，NEWS 採用簡單的靜態評分，但面對複雜的生命徵象變化還是存在一定限制。Shamout 等人開發用於預測臨床惡化的模型 Deep Early Warning System（DEWS），在牛津大學醫院進行了訓練和驗證，測試後模型的 AUROC 為 0.88、靈敏度為 0.73，而傳統 NEWS 系統的 AUROC 為 0.86、靈敏度為 0.70，此模型表現出比 NEWS 更高的準確性[60]。Cho 等人研究發現：深度學習模型 DeepCARS™ 在預測院內心臟驟停明顯比 NEWS 好，DeepCARS™ 的 AUROC 優於 NEWS（0.869 vs. 0.767），且模型在召回率和特異度方面都有明顯提高[61]。這些文獻與本研究結果相呼應，進一步證實深度學習模型可處理複雜的醫療數據，以及比傳統 NEWS 具備更有效的預警能力。

4.4 特徵重要性

本研究在不同負樣本組成所建立的 GRU 模型進行特徵重要性的分析，以體溫、呼吸頻率這兩種生命徵象對模型預測的貢獻最大。在心臟驟停的患者中，生命徵象通常會出現一系列的衰退變化，表示體溫和呼吸頻率可能是患者惡化的關鍵指標，醫療人員可透過此結果有效地監控和評估患者的健康狀況。Andersen 等人指出，在心臟驟停前 1 至 4 小時內，大約會有六成的患者出現呼吸頻率（大於 20

次/min 或小於 10 次/min）、心率（大於 100 次/min 或小於 60 次/min）、收縮壓（大於 90 次/min）的異常，這些現象可能增加患者的院內死亡率[62]，代表這三項生命徵象可能對臨床結果具有重要的預測意義。根據本研究的特徵分析結果，體溫在某些情況下比呼吸頻率、心率和收縮壓更靈敏地反映心臟驟停的發生，Frei 等人表示患者體溫過低（低於 32℃）可能會導致心臟驟停，並且與較高的死亡率相關[63]，因此患者出現體溫降低，或許可讓模型預測到相關的心臟驟停風險。

4.5 外部驗證

本研究使用馬偕醫院一般病房發生心臟驟停的患者資料進行外部驗證，檢測模型應用在不同資料集是否也具備一定的預測表現，並確認在臨床環境中的實用性。在比較不同模型時，因馬偕醫院資料所選取的正負樣本數量不平衡（正樣本：282 筆，負樣本：560 筆），會依 F1-score 的高低作為判斷各模型預測的能力。附錄表 C.4.1 為馬偕醫院資料輸入負樣本取自心臟驟停患者資料所訓練模型的結果，各模型的 F1-score 表現相似，介於 0.65 至 0.71 間，其中 GRU 的整體表現最穩定，並沒有與 LSTM 和 CNN-LSTM 一樣傾向預測正樣本。附錄表 C.4.2 為馬偕醫院資料輸入負樣本取自未心臟驟停患者資料所訓練模型的結果，各模型的 F1-score 表現明顯有差異，範圍介於 0.58 至 0.70 間，其中 CNN-LSTM 的整體表現最為穩定，而 RNN 的表現最差。附錄表 C.4.3 為馬偕醫院資料輸入負樣本取自心臟驟停患者和未心臟驟停患者資料所訓練模型的結果，LSTM、GRU 和 CNN-LSTM 的外部驗證表現相近，F1-score 為 0.70、0.71，而 RNN 的 F1-score 為最低 0.51。

從上述結果中可以看出，利用加護病房環境下建立的模型應用於一般病房的臨床資料時，模型的特異度較低。表示模型可能過度評估患者的危險狀態，對正常的生命徵象發出不必要的警報，從而造成患者心理上的壓力及醫療資源的浪費。然而，本研究所訓練的模型還是能針對即將發生心臟驟停的患者成功進行預測，

其整體的預測效能和 F1-score 仍然維持在合理範圍。透過馬偕醫院一般病房資料所進行的外部驗證，我們可初步確認使用 MIMIC IV 資料庫來模擬和開發適用於一般病房的預警系統是可行的。這種方法在略微增加醫療成本的情況下，有助於挽救未來可能發生心臟驟停的患者。

本研究所建構的模型在馬偕醫院一般病房心臟驟停患者資料中進行外部驗證，結果顯示這些模型同樣具備一定的預測表現。Cho 等人使用 DeepCARS™的早期預警系統進行了多家醫院的外部驗證，預測其發生心臟驟停或非計畫性轉入 ICU 的可能性，其 AUROC 為 0.869 [61]。Kang 等人提出一個臨床實用且可解釋的深度學習模型，用於預測加護病房患者的死亡風險。模型使用 eICU 資料庫進行訓練，並使用同樣為公開資料庫的 MIMIC III 作為外部驗證，測試和驗證的 AUC 分別為 0.877、0.791，得知模型在過程中皆達到不錯的預測效果[64]。上述研究強調在醫療領域中，透過外部驗證可避免使用同一資料集進行訓練和驗證所帶來的偏差，能夠更準確地反映模型在實際應用中的表現，並確保模型在不同醫院、患者和臨床環境中的穩定性和可靠性，從而提高其臨床應用價值[65]。

4.6 研究優點與限制

本研究以公開資料庫 MIMIC IV 中，發生心臟驟停和未發生心臟驟停的加護病房患者資料做為模型開發來源，透過這些資料模擬一般病房測量的頻率和項目，克服目前未有一般病房公開資料庫，而不容易取得相關資料的困境。本研究也實際利用馬偕醫院一般病房發生心臟驟停的患者資料作為外部驗證，以台灣真實的醫療數據確保模型適合應用在臨床資料上。其中，觀察生命徵象趨勢對於早期辨識心臟驟停的患者非常重要，本研究建構的模型僅使用一般病房一天所記錄的六種生命徵象，這些項目不僅減少院方在資料收集上的負擔，還能使模型在現今醫療環境中更容易實施。而模型效能相較於使用許多特徵的研究[33]，具備一定的預測準確度。最後本研究探討正負樣本來源的組成，有助於確定資料的代表性，使

模型符合患者後續可能發生心臟驟停的風險。未來可針對類似模型制定更有效的數據收集，確保模型能夠廣泛應用在各種情況。

本研究存在一些未來需要改進的限制。在外部驗證方面，使用馬偕醫院台北院區一般病房發生心臟驟停患者的生命徵象資料，雖然模型具有不錯的預測效果，但院方目前僅提供一般病房中的心臟驟停患者資料，尚未取得一般病房占多數未發生心臟驟停患者的資料作為模型驗證，後續期望能補充相關資料使模型驗證可以更貼近臨床數據。此外，本研究使用加護病房的患者資料模擬一般病房的資料分布，但兩者的患者狀態實際上存在一定差別，加護病房的患者大多以呼吸器和藥物支撐，相較於一般病房的患者嚴重，因此未來需要透過其他醫院的一般病房資料，進行更多的外部驗證，確保重症醫學資料可用來預測一般病房患者的相關疾病。最後，不同醫療機構的差異，包含資源、設備和專業知識等，皆會影響患者的診療過程和病情發展。為了讓模型能夠更好判斷每位患者所產生的生理變化，未來研究預計納入來自不同地區和醫院層級的患者資料，藉此提高數據多樣性和模型在不同情境下的學習能力。

第五章　結論

本研究利用 MIMIC IV 資料庫進行深度學習模型的建立，欲盡早發現一般病房的心臟驟停患者。研究中使用常見的六種生命徵象（收縮壓、舒張壓、心率、呼吸頻率、體溫和血氧飽和度）作為模型特徵輸入，結果以負樣本取自心臟驟停患者和未心臟驟停患者各半資料所訓練的 GRU 模型，得到良好的預測效果，且符合臨床實際情況。此外，該模型優於傳統早期預警評分系統（NEWS），其中體溫和呼吸頻率作為預測心臟驟停風險的重要生理指標。研究最後，使用馬偕醫院（台北院區）一般病房發生心臟驟停的患者資料進行模型的部分外部驗證，初步推測公開重症醫學資料庫可模擬一般病房的臨床資料作為模型開發，期望未來能有效降低心臟驟停的發生率和死亡率，改善一般病房患者的住院安全。

參考文獻

1. Jacobs, I., et al., *Cardiac arrest and cardiopulmonary resuscitation outcome reports: update and simplification of the Utstein templates for resuscitation registries: a statement for healthcare professionals from a task force of the International Liaison Committee on Resuscitation (American Heart Association, European Resuscitation Council, Australian Resuscitation Council, New Zealand Resuscitation Council, Heart and Stroke Foundation of Canada, InterAmerican Heart Foundation, Resuscitation Councils of Southern Africa).* Circulation, 2004. **110**(21): p. 3385-97.
2. Andersen, L.W., et al., *In-Hospital Cardiac Arrest: A Review.* Jama, 2019. **321**(12): p. 1200-1210.
3. Neumar, R.W., et al., *Part 1: Executive Summary: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care.* Circulation, 2015. **132**(18 Suppl 2): p. S315-67.
4. Tsao, C.W., et al., *Heart Disease and Stroke Statistics-2022 Update: A Report From the American Heart Association.* Circulation, 2022. **145**(8): p. e153-e639.
5. Martin, S.S., et al., *2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association.* Circulation, 2024. **149**(8): p. e347-e913.
6. Chen, C.-T., et al., *Prognostic factors for survival outcome after in-hospital cardiac arrest: An observational study of the oriental population in Taiwan.* Journal of the Chinese Medical Association, 2016. **79**(1).
7. Wang, C.Y., et al., *The secular trends in the incidence rate and outcomes of out-of-hospital cardiac arrest in Taiwan--a nationwide population-based study.* PLoS One, 2015. **10**(4): p. e0122675.
8. Andersen, L.W., et al., *The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest.* Resuscitation, 2016. **98**: p. 112-117.
9. Schein, R.M.H., et al., *Clinical Antecedents to In-Hospital Cardiopulmonary Arrest.* Chest, 1990. **98**(6): p. 1388-1392.
10. Hodgetts, T.J., et al., *Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital.* Resuscitation, 2002. **54**(2): p. 115-123.
11. Devita, M.A., et al., *Findings of the first consensus conference on medical emergency teams.* Crit Care Med, 2006. **34**(9): p. 2463-78.
12. Jones, D.A., M.A. DeVita, and R. Bellomo, *Rapid-response teams.* N Engl J

Med, 2011. **365**(2): p. 139-46.
13. Smith, G.B., et al., *A review, and performance evaluation, of single-parameter "track and trigger" systems.* Resuscitation, 2008. **79**(1): p. 11-21.
14. Liu, V.X., et al., *Comparison of Early Warning Scoring Systems for Hospitalized Patients With and Without Infection at Risk for In-Hospital Mortality and Transfer to the Intensive Care Unit.* JAMA Netw Open, 2020. **3**(5): p. e205191.
15. Morgan, R., F. Williams, and M. Wright, *An early warning scoring system for detecting developing critical illness.* Clin Intensive Care, 1997. **8**(2): p. 100.
16. Subbe, C.P., et al., *Validation of a modified Early Warning Score in medical admissions.* QJM: An International Journal of Medicine, 2001. **94**(10): p. 521-526.
17. Santiago González, N., et al., *Modified Early Warning Score: Clinical Deterioration of Mexican Patients Hospitalized with COVID-19 and Chronic Disease.* Healthcare (Basel), 2023. **11**(19).
18. Heitz, C.R., et al., *Performance of the maximum modified early warning score to predict the need for higher care utilization among admitted emergency department patients.* J Hosp Med, 2010. **5**(1): p. E46-52.
19. Goldhill, D.R., et al., *A physiologically-based early warning score for ward patients: the association between score and outcome.* Anaesthesia, 2005. **60**(6): p. 547-53.
20. Reini, K., M. Fredrikson, and A. Oscarsson, *The prognostic value of the Modified Early Warning Score in critically ill patients: a prospective, observational study.* European Journal of Anaesthesiology | EJA, 2012. **29**(3).
21. Williams, B., *The National Early Warning Score 2 (NEWS2) in patients with hypercapnic respiratory failure.* Clin Med (Lond), 2019. **19**(1): p. 94-95.
22. Aygun, H. and S. Eraybar, *The role of emergency department triage early warning score (TREWS) and modified early warning score (MEWS) to predict in-hospital mortality in COVID-19 patients.* Ir J Med Sci, 2022. **191**(3): p. 997-1003.
23. Ghosh, E., et al., *Description of vital signs data measurement frequency in a medical/surgical unit at a community hospital in United States.* Data Brief, 2018. **16**: p. 612-616.
24. Smith, G.B., A. Recio-Saucedo, and P. Griffiths, *The measurement frequency and completeness of vital signs in general hospital wards: An evidence free zone?* International Journal of Nursing Studies, 2017. **74**: p. A1-A4.
25. Smith, G.B., et al., *The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death.* Resuscitation, 2013. **84**(4): p. 465-470.

26. Wang, A.-Y., et al., *Periarrest Modified Early Warning Score (MEWS) predicts the outcome of in-hospital cardiac arrest.* Journal of the Formosan Medical Association, 2016. **115**(2): p. 76-82.
27. Kwon, J.M., et al., *An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest.* J Am Heart Assoc, 2018. **7**(13).
28. Mitsunaga, T., et al., *Comparison of the National Early Warning Score (NEWS) and the Modified Early Warning Score (MEWS) for predicting admission and in-hospital mortality in elderly patients in the pre-hospital setting and in the emergency department.* PeerJ, 2019. **7**: p. e6947.
29. Lee, Y.J., et al., *A multicentre validation study of the deep learning-based early warning score for predicting in-hospital cardiac arrest in patients admitted to general wards.* Resuscitation, 2021. **163**: p. 78-85.
30. Kim, W.Y., et al., *Modified Early Warning Score Changes Prior to Cardiac Arrest in General Wards.* PLoS One, 2015. **10**(6): p. e0130523.
31. Alamgir, A., O. Mousa, and Z. Shah, *Artificial Intelligence in Predicting Cardiac Arrest: Scoping Review.* JMIR Med Inform, 2021. **9**(12): p. e30798.
32. Wu, T.T., et al., *Machine learning for early prediction of in-hospital cardiac arrest in patients with acute coronary syndromes.* Clin Cardiol, 2021. **44**(3): p. 349-356.
33. Chae, M., et al., *Prediction of In-Hospital Cardiac Arrest Using Shallow and Deep Learning.* Diagnostics (Basel), 2021. **11**(7).
34. Cai, J., et al., *Feature selection in machine learning: A new perspective.* Neurocomputing, 2018. **300**: p. 70-79.
35. Pedersen, A.B., et al., *Missing data and multiple imputation in clinical epidemiological research.* Clin Epidemiol, 2017. **9**: p. 157-166.
36. Perman, S.M., et al., *Location of In-Hospital Cardiac Arrest in the United States-Variability in Event Rate and Outcomes.* J Am Heart Assoc, 2016. **5**(10).
37. Churpek, M.M., et al., *Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards.* Crit Care Med, 2016. **44**(2): p. 368-74.
38. Yang, J., et al., *Brief introduction of medical database and data mining technology in big data era.* Journal of Evidence-Based Medicine, 2020. **13**(1): p. 57-69.
39. Choi, E., et al. *Doctor ai: Predicting clinical events via recurrent neural networks*. in *Machine learning for healthcare conference*. 2016. PMLR.
40. Lipton, Z.C., et al., *Learning to diagnose with LSTM recurrent neural networks.* arXiv preprint arXiv:1511.03677, 2015.
41. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p.

436-444.
42. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.
43. Johnson, A., et al., *Mimic-iv.* PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021), 2020: p. 49-55.
44. Johnson, A.E., et al., *MIMIC-IV, a freely accessible electronic health record dataset.* Scientific data, 2023. **10**(1): p. 1.
45. Johnson, A.E., et al., *The MIMIC Code Repository: enabling reproducibility in critical care research.* J Am Med Inform Assoc, 2018. **25**(1): p. 32-39.
46. Huang, G. *Missing data filling method based on linear interpolation and lightgbm*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
47. Cesare, N. and L.P.O. Were, *A multi-step approach to managing missing data in time and patient variant electronic health records.* BMC Res Notes, 2022. **15**(1): p. 64.
48. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning representations by back-propagating errors.* nature, 1986. **323**(6088): p. 533-536.
49. Cho, K., et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078, 2014.
50. Ozcanli, A.K. and M. Baysal, *Islanding detection in microgrid using deep learning based on 1D CNN and CNN-LSTM networks.* Sustainable Energy, Grids and Networks, 2022. **32**: p. 100839.
51. Raschka, S., *Model evaluation, model selection, and algorithm selection in machine learning.* arXiv preprint arXiv:1811.12808, 2018.
52. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 2017. **30**.
53. Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis, *Explainable AI: A Review of Machine Learning Interpretability Methods.* Entropy (Basel), 2020. **23**(1).
54. Zheng, A. and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. 2018: " O'Reilly Media, Inc.".
55. Ho, S.Y., et al., *Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability.* Patterns (N Y), 2020. **1**(8): p. 100129.
56. Kim, C., et al., *Out-of-hospital cardiac arrest in men and women.* Circulation, 2001. **104**(22): p. 2699-703.
57. Penketh, J. and J.P. Nolan, *In-hospital cardiac arrest: the state of the art.* Critical Care, 2022. **26**(1): p. 376.
58. Soudan, B., F.F. Dandachi, and A.B. Nassif, *Attempting cardiac arrest*

prediction using artificial intelligence on vital signs from Electronic Health Records. Smart Health, 2022. **25**: p. 100294.

59. Yijing, L., et al., *Prediction of cardiac arrest in critically ill patients based on bedside vital signs monitoring.* Computer Methods and Programs in Biomedicine, 2022. **214**: p. 106568.
60. Shamout, F.E., et al., *Deep Interpretable Early Warning System for the Detection of Clinical Deterioration.* IEEE J Biomed Health Inform, 2020. **24**(2): p. 437-446.
61. Cho, K.J., et al., *Prospective, multicenter validation of the deep learning-based cardiac arrest risk management system for predicting in-hospital cardiac arrest or unplanned intensive care unit transfer in patients admitted to general wards.* Crit Care, 2023. **27**(1): p. 346.
62. Andersen, L.W., et al., *The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest.* Resuscitation, 2016. **98**: p. 112-117.
63. Frei, C., et al., *Clinical characteristics and outcomes of witnessed hypothermic cardiac arrest: A systematic review on rescue collapse.* Resuscitation, 2019. **137**: p. 41-48.
64. Kang, Y., et al., *A Clinically Practical and Interpretable Deep Model for ICU Mortality Prediction with External Validation.* AMIA Annu Symp Proc, 2020. **2020**: p. 629-637.
65. Ramspek, C.L., et al., *External validation of prognostic models: what, why, how, when and where?* Clin Kidney J, 2021. **14**(1): p. 49-58.

附錄

A. 變項代碼對照表及數值範圍

表 A.1　變項代碼對照表及數值範圍

變項	MIMIC 代碼	範圍
收縮壓 Non Invasive Systolic Blood Pressure, NISBP	220179	0 < NISBP < 400
舒張壓 Non Invasive Diastolic Blood Pressure, NIDBP	220180	0 < NIDBP < 300
心率 Heart Rate, HR	220045	0 < HR < 300
呼吸頻率 Respiratory Rate, RR	220210	0 < RR < 70
體溫（華氏） Body Temperature（Fahrenheit）, BT（℉）	223761	70 < BT（℉）< 120
體溫（攝氏） Body Temperature（Celsius）, BT（℃）	223762	10 < BT（℃）< 50
血氧飽和度 oxygen saturation, SpO_2	220277	0 < SpO_2 <= 100

B. MIMIC IV 資料配置

表 B.1 正負樣本皆取自心臟驟停患者資料

	Positive samples	Negative samples
Training dataset	2752	2223
Test data	710	534
Total	3462	2757

表 B.2 正樣本同上，負樣本取自未心臟驟停患者資料

	Positive samples	Negative samples
Training dataset	2769	2715
Test data	693	679
Total	3462	3394

表 B.3 正樣本同上，負樣本取自心臟驟停患者和未心臟驟停患者資料

	Positive samples	Negative samples
Training dataset	2736	2493
Test data	726	582
Total	3462	3075

C. 馬偕紀念醫院（台北院區）一般病房之心臟驟停患者資料

C.1 資料來源

本研究欲預測發生在一般病房的心臟驟停事件，因此使用馬偕紀念醫院（台北院區）一般病房的心臟驟停患者資料進行外部驗證。院方資訊室提供 2016 至 2021 年在加護病房、急診和一般病房等發生心臟驟停的所有相關資料，以及病歷系統中記錄為「心跳停止分析」且發生地點為一般病房的相關案例。本研究將這些案例與院方提供的完整資料進行比對，從 642 起心臟驟停事件中，篩選出 179 起於一般病房發生的事件。為了保護患者隱私，資訊室會將每筆資料進行去辨識化處理。本研究計畫已獲得馬偕紀念醫院人體研究倫理審查委員會之核准（編號：20MMHIS409e）。

C.2 資料特徵及前處理

馬偕醫院的資料在申請時已包含每位患者的病歷號（PNO）、入院日期、生日、性別、六種生命徵象（收縮壓、舒張壓、心率、呼吸頻率、體溫以及血氧飽和度）、測量時間和日期，以及心臟驟停時間。最終從院方資訊室得到每位患者心臟驟停前最多 48 小時的相關醫療數據。

我們從馬偕醫院得到相關病歷資料後，根據以下列規則進行資料前處理：

A. 使用生日及入院日期計算出年齡。排除年齡小於 20 歲的患者。超過 200 歲的患者則視為異常值並移除。

B. 只保留每位患者同次住院發生第一次心臟驟停的記錄，後續事件則不納入分析範圍。

C. 移除記錄時間不明的資料，因無法確認測量和事件發生的時間差。

D. 移除六個生命徵象全是缺失值的資料，認定為嚴重缺失、無法進行補值。

E. 由於一般病房患者的生命徵象通常需要醫療人員手動輸入，可能會出現標點符號錯誤（如：將 37.2 輸入為 37,2）或是誤按到英文字母（例如：將 118 輸入為 118A）等人為疏失，這可能導致程式解讀錯誤。為瞭解決這個問題使用正則表達式（Regular Expression, RE）移除不該出現的英文字母、中文字和特殊符號。若數值介於一段範圍（如：90–95），則更改成此範圍的平均。

F. 同位病患在相同時間點出現重複的資料，則隨機保留其中一筆。

G. 各生命徵象的數值範圍，與 MIMIC IV 資料庫使用相同標準（參照附錄 A：變項代碼對照表及數值範圍）。

H. 針對資料缺值採用線性內插法進行補值，若遇到無法補值的情況，則會利用每位患者該變數的平均值，針對有缺失值進行補植。

馬偕醫院（台北院區）一般病房中的 179 起心臟驟停事件，共收集 4,055 筆生命徵象資料；資料清理後留下 179 起事件、2,622 筆生命徵象資料。接著根據章節 2.5 選取正負樣本的方法，套用於資料清理後的心臟驟停患者資料，一共選取 282 筆正樣本、560 筆負樣本。後續以即將發生心臟驟停的正樣本，以及尚未發生心臟驟停的負樣本視為測試資料集，輸入訓練完成的各模型進行預測及評估。

C.3 MIMIC 資料庫及馬偕醫院中心臟驟停患者之人口學及特徵分析

表 C 3.1 MIMIC IV 資料庫及馬偕醫院中心臟驟停患者之人口學及特徵分析

Characteristics	**MIMIC IV (CA)**	**Mackay Hospital (Taipei) (CA)**	**p-value**
Number of total patients, n	477	178	
Number of total vital signs, n	68,973	2,622	
Age, y , mean ± SD	65.58 ± 15.91	70.70 ± 13.14	
Male ,sex , n (%)	289 (60.59%)	113 (63.48%)	
Features, mean（SD）			
SBP (mmHg)	110.69 ± 22.06	124.76 ± 27.05	$p < 0.05$
DBP (mmHg)	61.96 ± 15.74	67.15 ± 14.90	$p < 0.05$
HR (/min)	89.32 ± 20.88	93.53 ± 24.42	$p < 0.05$
RR (/min)	21.79 ± 6.45	19.66 ± 4.12	$p < 0.05$
BT (℃)	36.94 ± 1.04	37.00 ± 1.13	$p < 0.05$
SpO_2 (%)	96.50 ± 4.83	374.87 ± 3835.79	$p < 0.05$

本研究從馬偕醫院台北院區收集 179 起心臟驟停事件，包含 178 位患者和 2,622 筆生命徵象記錄。該組別的平均年紀為 70.70 歲，男性的發生率較高（63.48% vs. 36.52%）。在生命徵象部分，相較於 MIMIC IV 資料庫發生心臟驟停的組別，馬偕醫院發生心臟驟停的組別有較高的收縮壓（124.76 ± 27.05 mmHg vs. 110.69 ± 22.06 mmHg）、舒張壓（67.15 ± 14.90 mmHg vs. 61.96 ± 15.74 mmHg）、心率（93.53 ± 24.42 次/min vs. 89.32 ± 20.88 次/min）、體溫（37.00 ± 1.13℃ vs. 36.94 ± 1.04℃）和血氧飽和度（374.87 ± 3835.79% vs. 96.50 ± 4.83%），以及較低的呼吸頻率（19.66 ± 4.12 次/ min vs. 21.79 ± 6.45 次/ min）。其中，血氧飽和度的平均遠大於正常範圍，推測是有少數異常值（outlier）所造成。

C.4 外部驗證結果

表 C.4.1 負樣本取自心臟驟停患者資料所訓練模型之外部驗證結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN	0.78	0.71	0.60	0.88	0.65	0.87
LSTM	0.72	0.55	0.96	0.60	0.70	0.85
GRU	0.76	0.63	0.74	0.78	0.68	0.87
CNN-LSTM	0.58	0.44	0.99	0.38	0.71	0.85

表 C.4.2 負樣本取自未心臟驟停患者資料所訓練模型之外部驗證結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN	0.75	0.67	0.51	0.87	0.58	0.83
LSTM	0.72	0.55	0.96	0.61	0.70	0.86
GRU	0.76	0.64	0.65	0.82	0.65	0.84
CNN-LSTM	0.73	0.57	0.74	0.72	0.64	0.84

表 C.4.3 負樣本取自心臟驟停患者和未心臟驟停患者資料所訓練模型之外部驗證結果

	accuracy	precision	sensitivity	specificity	F1-score	AUROC
RNN	0.56	0.41	0.69	0.49	0.51	0.68
LSTM	0.76	0.60	0.87	0.71	0.71	0.88
GRU	0.75	0.58	0.93	0.66	0.71	0.86
CNN-LSTM	0.73	0.56	0.95	0.62	0.70	0.86