

國立臺灣師範大學理學院

資訊工程研究所

碩士論文

Department of Computer Science and Information Engineering College  
of Science

National Taiwan Normal University

Master's Thesis

以異質網路圖學習病況事件表示法進行死亡風險預測

Data Representation Learning from Heterogeneous  
Network of Medical Data for Mortality Prediction

洪翊誠

Yi-Cheng Hung

指導教授：柯佳伶博士

Advisor：Jia-Ling Koh, Ph.D.

中華民國 一百一十一年 二 月

Feb. 2022

## 摘要

以異質網路圖學習病況事件表示法進行死亡風險預測

洪翊誠

近年來以機器自動學習數據的特徵表示法，已顯示有助於提升預測任務的準確率。本論文以電子病歷資料中相異類型的病況資料，依指定時間區間內病況事件同時發生的關聯，建立病況事件異質網路圖，並搭配不同的病況事件序列生成樣式，從取樣的事件序列中，學習儀器偵測數據特徵的病況事件表示法，用來從加護病房病患入病房後 48 小時的病況資料，以 LSTM 類神經網路架構進行死亡風險預測。本論文實驗比較使用同質特徵走訪路徑與異質特徵走訪路徑的擷取策略，所學習到的病況事件表示法對模型預測效果的差異。實驗在院內死亡預測及短期死亡預測的任務，初步顯示由異質特徵走訪路徑中學習的病況事件表示法，對兩個預測模型的預測效果皆有提昇。

**關鍵字:**異質網路圖、資料特徵表示法、死亡預測模型

# ABSTRACT

## Data Representation Learning from Heterogeneous Network of Medical Data for Mortality Prediction

by

Yi-Cheng Hung

In recent years, feature representation learning from data has been shown to be helpful for improving the accuracy of prediction tasks. In this thesis, the various types of attributes combined with the values in the electronic medical record, which implicitly describe patient's condition, are named clinical events. We constructed a heterogeneous network of clinical events according to their occurring on the same patient within a specified time interval. Then event sequences are sampled by visiting different meta-paths for learning the representations of chart events. The learned representations of chart events are used to input to a framework of LSTM neural network for predicting mortality of ICU patients according to their first 48 hours of in-ICU EMR data. In the experiments, we compared the prediction effectiveness of the learned event representations by changing the time interval of constructing the heterogeneous network and applying homogeneous or heterogeneous meta-path visiting. The preliminary results of experiments show that the representations of chart events learned from the heterogeneous meta-path effectively improve the recall and AUROC on both the tasks of in-hospital mortality prediction and short-term mortality prediction.

***Keywords : heterogeneous network, data representation, mortality prediction model***

## 致謝

首先，我最想感謝我的指導教授-柯佳伶教授，在我論文研究受挫的時候，不惜辛勞的幫助我，從老師身上我學習到做事的態度與方法，也學習到可貴的處事精神。接著，想感謝研究室的學長姐-家儀、修毅、雅雯，在研究所的這段期間每當我不知所措時，總是不吝給我建議。也感謝我的同袍-明潔和盈翔，與你們一起學習成長，讓我在你們身上獲益良多！感謝研究室的學弟妹們，在我準備口試期間與口試當天幫忙處理瑣碎事務，讓我可以專心準備論文口試，祝福你們未來順利完成論文。還要感謝我的高中好友們，天團的好夥伴，常常督促、激勵我，給我勇氣去面對論文。

感謝我的父母親，在精神上及經濟上不斷支持我，並全心相信我，讓我能無後顧之憂的完成論文。感謝我的女友庭馨，在我最辛苦的這一年，常常鼓勵我、照顧我，使我更有力量面對論文的挑戰。衷心感謝在這段路上幫助我的所有人，這段路走來不易，因為有各位，我才能完成，謝謝大家。

洪翊誠 謹識

於國立臺灣師範大學資訊工程所

2022 年 2 月

## 目錄

<b>第1章 緒論</b>	<b>1</b>
1.1 研究動機	1
1.2 研究目的	3
1.3 研究限制與範圍	5
1.4 論文方法	7
1.5 論文架構	11
<b>第2章 文獻探討</b>	<b>12</b>
2.1 表徵學習法	12
2.1.1 文字資料表徵學習	12
2.1.2 醫療代碼表徵學習法	13
2.1.3 網路圖表徵學習模型	18
2.1.4 醫療異質網路表徵學習法	21
2.2 死亡率預測模型	23
<b>第3章 問題定義與系統架構</b>	<b>28</b>
3.1 預測任務定義	28
3.2 系統流程架構	29
<b>第4章 研究方法</b>	<b>32</b>
4.1 資料前處理	32
4.1.1 填補遺失值	32
4.1.2 清除異常值	35
4.1.3 病況事件編碼	36
4.2 特徵表示法學習	40
4.2.1 建立病況事件異質網路圖	40
4.2.2 病況事件特徵表示法學習	42
4.3 死亡率預測模型	47
4.3.1 模型架構	47
4.3.2 擴展預測模型	51
<b>第5章 實驗結果與探討</b>	<b>53</b>
5.1 實驗資料集及模型參數說明	53
5.2 評估指標	56
5.3 病況事件表示法學習效果分析	59
5.3.1 分析表示法相似搜尋	59
5.3.2 儀器偵測數據類型病況事件分群	64

5.4	預測效能評估.....	70
5.4.1	資料前處理效果比較.....	70
5.4.2	病況事件表示法之訓練資料擷取策略效果評估.....	72
5.4.3	加入實驗室化驗結果病況事件資訊之模型預測效果.....	79
<b>第6章</b>	<b>結論與未來研究方向 .....</b>	<b>83</b>
<b>參考文獻</b> .....		<b>85</b>
<b>附錄</b> .....		<b>89</b>

## 附圖目錄

圖 1.1 患者 A 進入 ICU 後 48 小時內發生的病況事件.....	7
圖 1.2 病況事件網路圖.....	8
圖 2.1 Skip-gram 模型[7][8].....	13
圖 2.2 Med2vec 模型[1] .....	14
圖 2.3 Dipole 雙向遞歸式神經網路模型[3] .....	16
圖 2.4 GRAM 模型[2].....	17
圖 2.5 KAME 模型[5].....	18
圖 2.7 HeteroMed 的模型架構[6].....	21
圖 2.8 論文[11]模型架構.....	25
圖 2.9 論文[31]模型架構.....	26
圖 3.1 模型訓練流程圖.....	30
圖 3.2 模型測試流程圖.....	31
圖 4.1 以 Skip-gram 模型架構圖.....	44
圖 4.2 死亡率預測模型架構圖.....	47
圖 5.1 前處理對預測模型的效果比較.....	70
圖 5.2 加上 $W_{class}$ 比重後前處理對預測模型的效能比較.....	71
圖 5.3 表示法訓練資料擷取策略在院內死亡預測的預測準確率.....	72
圖 5.4 表示法訓練資料擷取策略在院內死亡預測的預測召回率.....	73
圖 5.5 表示法訓練資料擷取策略在院內死亡預測的預測 AUROC .....	73
圖 5.6 表示法訓練資料擷取策略在短期死亡預測的預測準確率.....	76
圖 5.7 表示法訓練資料擷取策略在短期死亡預測的預測召回率.....	76
圖 5.8 表示法訓練資料擷取策略在短期死亡預測的預測 AUROC .....	77
圖 5.9 加入 $V_{lab}$ 病況事件表示法在院內死亡預測之準確率 .....	79
圖 5.10 加入 $V_{lab}$ 病況事件表示法在院內死亡預測之召回率 .....	80
圖 5.11 加入 $V_{lab}$ 病況事件表示法在院內死亡預測之 AUROC .....	80
圖 5.12 加入 $V_{lab}$ 病況事件表示法在短期死亡預測之準確率 .....	81
圖 5.13 加入 $V_{lab}$ 病況事件表示法在短期死亡預測之召回率 .....	81
圖 5.14 加入 $V_{lab}$ 病況事件表示法在短期死亡預測之 AUROC .....	82

## 附表目錄

表 1.1 儀器偵測數據特徵.....	6
表 4.1 病患入住加護病房紀錄範例.....	32
表 4.2 儀器偵測數據病況紀錄.....	32
表 4.3 訓練集資料範例.....	33
表 4.4 儀器偵測數據特徵估算值.....	34
表 4.5 儀器偵測數據正常範圍值.....	36
表 4.6 連續型態病況事件劃分區間大小及其編號範圍.....	37
表 4.7 儀器偵測數據類別型態病況事件編碼範圍.....	38
表 4.8 已劃分區間的儀器偵測數據病況事件紀錄.....	40
表 4.9 以本論文資料集筆數計算比重值.....	50
表 5.1 資料筆數.....	53
表 5.2 正負樣本數.....	54
表 5.3 參數設定值.....	55
表 5.4 混淆矩陣範例.....	56
表 5.5 「GCS = 3」病況事件相似度前十名病況事件結果 1.....	60
表 5.6 「GCS = 3 分」病況事件相似度前十名病況事件結果 2 .....	62
表 5.7 病況事件分群結果.....	64



# 第 1 章 緒論

## 1.1 研究動機

醫療保健是全民日常經常需要的服務，民眾在醫療場所的每次看診，醫療單位皆會記錄病歷。由於紙本病歷保存不易且調閱困難，隨著電子化的發展，現今醫療單位大多採用電子化病歷。單在美國，每年有超過 3000 萬的人口就診，當中有百分之八十三使用電子病歷(Electronic medical record, EMR)記錄就診資料。改善患者的醫療照護是醫生和研究人員的共同目標，而從大量的電子醫療紀錄中進行探勘提供了實現這個目標的可能性[17, 18, 19, 20]。隨著近幾年數據分析領域的發展，這些臨床紀錄的醫療數據庫提供大量的資料，可用來分析醫療資料間的關聯，進行不同的醫療任務預測，例如檢傷分類(early triage)、風險評估(risk assessment)、生理性代償失調(physiological decompensation)等預測。

加護病房(Intensive care unit, ICU) 是醫院內為需要高度密集醫療照料的重病傷患所特設的病房，加護病房內有諸多設備與儀器隨時監測患者的身體變化，並由醫療人員 24 小時輪班照顧。患者在送至加護病房時，其身體狀況有一定的嚴重程度，因此病患在短時間內可能因病情突然惡化造成死亡。為了降低病患的死亡風險，若能由病患在加護病房的儀器偵測數據，預測病患在接下來一段時間內可能死亡的機率，可以幫助醫護評估病患病況的嚴重程度，協助醫

護更留意病患處置的後續演變。

近年來深度學習技術的發展，除了改善機器學習預測模型的準確率，也運用在數據特徵表示法的自動學習。表徵學習法的目標是以機器學習方法自動提取特徵，著重在如何從原始資料中有效提取特徵，以助於提升目標任務預測的準確度。由於患者在加護病房中監測所得的資料非常多樣，如何將其轉成有效的特徵表示法來反應病況，進而有效預測病患的病程發展，是重症監護研究中的重要問題。

## 1.2 研究目的

加護病房病患的 EMR 資料除了病患的基本資料，還紀錄了病患進入加護病房後的各種病況資料，包括病患的儀器偵測數據(Chart)、實驗室測量值(Laboratory test)、藥物處方(Prescription)、治療程序(Procedure)、微生物的測量值跟敏感性(Microbiology)等不同類型的資料。

以往也有不少醫療保健研究運用特徵學習法，讓機器從看診紀錄中學習醫療代碼表示法[1, 3, 4]。其中 Med2Vec[1]認為當次及前後次看診的醫療代碼具相關聯性，因此提出一個類似 Word2Vec 的神經網路模型，作為醫療代碼的特徵表示學習架構。不同於 Med2Vec[1]的架構，[3, 4]皆是基於遞歸式神經網路(RNN)為基礎設計出的模型，根據醫療代碼間出現序列的資訊，以 RNN 網路對醫療代碼的表示法進行學習。[3]與[4]兩者的差別在於[3]的醫療代碼是由最近到最久前的看診輸入 RNN 網路，而[4]則使用雙向的遞歸式神經網路(Bidirectional recurrent neural networks, BRNNs)[21]綜合醫療看診代碼排序由最久遠到最近及最近到最久遠皆輸入 RNN 網路，以預測患者下一次看診醫療代碼的任務學習出醫療代碼表示法。

前述方法[1, 3, 4]只考慮由醫療編碼的出現關聯學習其特徵表示法，HeteroMed[6]是首篇考慮 EMR 中不同類型病況資料之出現關聯並從中學習病況資料特徵表示法的研究。雖然 HeteroMed[6]採用不同類型病況資料，使病況資料的特徵表示學習資訊更完整，但[6]在將學習到的特徵進行後續預測任務時，

所建構的模型並未考慮病況資料出現的時序性。論文[11]以 ICU 病患院內死亡率為預測任務，將病患進入 ICU 後 48 小時內發生的多種儀器偵測數據直接做為輸入，以長短期記憶模型(LSTM)單元組合成的模型進行預測。

此方式在預測任務中雖然有考慮儀器偵測值出現的時序性，但並未將這些病況資料先進行表示法學習。因此，本論文的研究目的是探討如何運用不同類型病況資料的出現關聯，對儀器偵測數據進行病況資料的特徵表示法學習，再輸入[11]中提出的模型架構進行預測任務，觀察是否可有效提升預測正確率。

### 1.3 研究限制與範圍

MIMIC III(Medical Information Mart for Intensive Care III)資料集，是一個免費的大型醫療數據庫，記錄許多加護病房(Intensive Care Unit,ICU)病患的不同類型病況資料，其中包含臨床儀器偵測數據(Char)、實驗室化驗結果(Laboratory test)、微生物培養結果、醫護人員文字紀錄等類型。本研究在預測任務所輸入之病況資料類型與[11]相同，包括儀器偵測數據中共 17 種資料欄位，可顯示患者身體質量、生理監視、神經反應，及生化檢測四大特性的資料，各資料欄位說明如表 1.1 所列。

為了利用不同類型資料的出現關聯學習病況資料特徵表示法，本研究另納入實驗室化驗結果的病況資料，包括如紅細胞、白細胞、淋巴細胞數等實驗室檢測數據，用來輔助學習儀器偵測數據的病況資料表示法。

表 1.1 儀器偵測數據特徵

英文名稱	特性分類	中文名稱	數據型態	數據範圍
Weight	身體質量	體重	連續	0-250
Height	身體質量	身高	連續	0-240
Temperature	生理監視	體溫	連續	20-45
Systolic blood pressure	生理監視	收縮壓	連續	0-375
Diastolic blood pressure	生理監視	舒張壓	連續	0-375
Mean blood pressure	生理監視	平均血壓	連續	14-330
Heart Rate	生理監視	心率	連續	0-350
Capillary refill rate	生理監視	微血管回充率	列舉	{0, 1}
Respiratory rate	生理監視	呼吸頻率	連續	0-300
Oxygen saturation	生理監視	血氧飽和度	連續	0-100
Fraction inspired oxygen	生理監視	吸入氧氣分率	連續	0.21-1
GCS motor response	神經反應	運動反應	列舉	詳見附錄 A-1
Glasgow coma scale total	神經反應	昏迷量表	列舉	詳見附錄 A-1
GCS verbal response	神經反應	語言反應	列舉	詳見附錄 A-1
GCS eye opening	神經反應	睜眼反應	列舉	詳見附錄 A-1
Glucose	生化檢測	血糖	連續	33-2000
pH	生化檢測	血液酸鹼值	連續	6.3-8.4

## 1.4 論文方法

本研究認為 EMR 中這些不同類型的病況資料，由特徵及其出現的偵測值可用來反應某個病況事件，因此可根據同一病患在一定時間區間內發生的病況事件關聯進行事件的表示法學習。以圖 1.1 為例，其表示一個患者 A 被送入 ICU 後 48 小時內的儀器偵測數據特徵及實驗室化驗結果特徵，對應到發生的四個病況事件。而這些事件間發生的間隔若在一定時間區間內(小於等於  $k$  小時)，則可能是有關聯的。

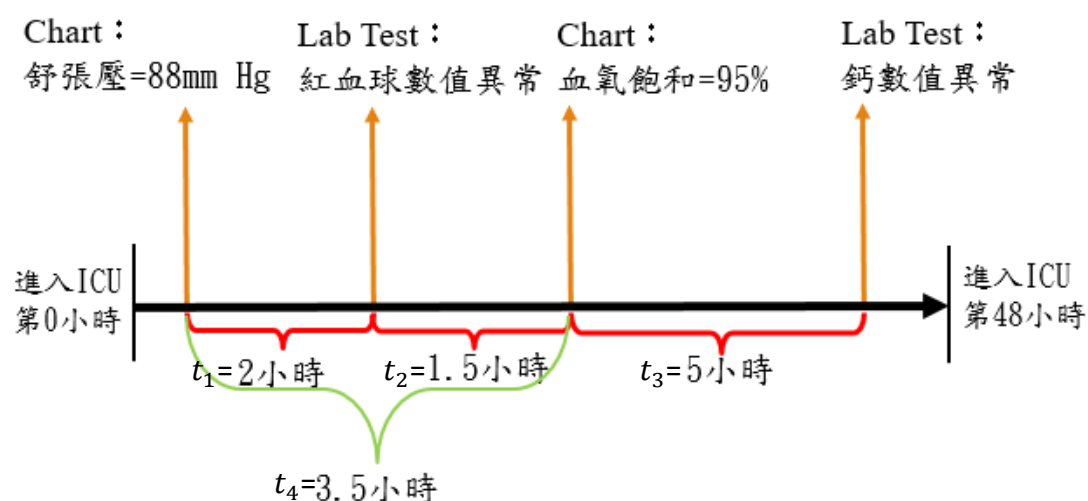


圖 1.1 患者 A 進入 ICU 後 48 小時內發生的病況事件

將特徵的資料值對應到病況事件，將病況事件以圖形節點表示，且假設在一定時間內同一病患發生的病況事件間有關聯，因此建立對應的邊，則可建構出一個病況事件網路圖，再從中進行病況事件表示法學習。圖 1.2 所示是由圖 1.1 所建立的病況事件網路圖，其中是設定在  $k$  小時區間內所發生的事件可能有關聯。當  $k$  設為 6，圖 1.1 中的  $t_1$ 、 $t_2$ 、 $t_3$  及  $t_4$  代表的時間間隔值皆小於等於  $k$ ，所以將「舒張壓=88mmHg」與「紅血球數值異常」兩個事件間視為可能有關聯；

「紅血球數值異常」與「血氧飽和=95%」有關聯；「血氧飽和=95%」與「鈣數值異常」有關聯；且「舒張壓=88mmHg」與「血氧飽和=95%」有關聯，因此在圖 1.2 中建立 $e_1$ 、 $e_2$ 、 $e_3$ 及 $e_4$ 四條邊。本研究將實驗改變發生有關聯事件間的時間間隔長短門檻值 $k$ ，比較依該設定建立的事件網路圖訓練出來的病況事件表示法效果差異。

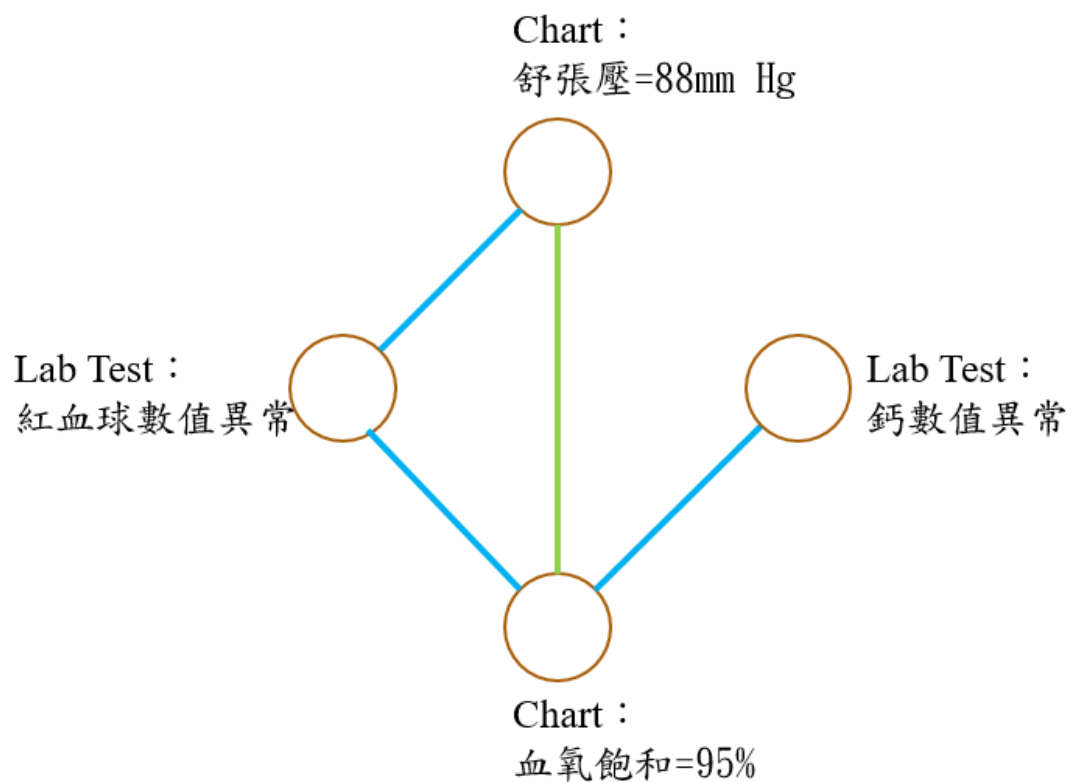


圖 1.2 病況事件網路圖



以往提出的網路圖特徵表示學習法[9, 10, 15]通常先採用一般隨機遊走的方式採樣產生節點序列，作為表示法學習的訓練資料。在針對同質節點所形成的網路圖，在進行隨機遊走時並不需要考慮網路圖路徑上所經節點的特徵類型。本研究所考慮的資料包括儀器偵測數據及實驗室化驗結果兩種不同類型特徵的病況事件，而如圖 1.2 的網路圖中包含兩種以上不同類型特徵的事件節點，即稱為一個異質網路圖(Heterogeneous Information Network)[13]。但異質網路圖中具有不同類型的節點，若未限制遊走路徑上的節點類型，可能產生無意義的節點遊走路徑。論文[12]提出了針對異質網路圖的節點特徵表示學習法 metapath2vec，不同於以往的方式，從網路圖採樣產生節點序列時會限制路徑所經節點需符合特定類型。本論文採用論文[12]的方式，預先訂定路徑型態，再隨機挑選符合類型的節點走訪路徑作為表示法學習的訓練資料。例如預定路徑型態是(儀器偵測數據-實驗室化驗結果)，隨機挑選出的節點走訪路徑都必須符合此路徑型態。上述圖 1.2 顯示的網路圖中包含儀器偵測數據及實驗室化驗結果兩種不同類型特徵的節點，本研究亦將探討採用同質節點序列及異質節點序列，對訓練特徵表示法的效果差異。

本研究會將取樣出的病況事件序列視為文字序列，採用 Word2Vec[16]的方法訓練出病況事件的表示法。原本 Word2Vec[16]採用文字序列中上下文出現的關聯學習文字的特徵表示法，而在本論文則運用事件路徑中事件節點出現的關聯進行事件的特徵表示法學習。在進行病況事件的表示法學習後，本研究最後

再採用長短期記憶網路(LSTM)單元的組合建構預測模型[11]，將病患在ICU儀器測量資料轉換成病況事件，以病況事件異質圖中各事件預學習到的表示向量代入，再以模型中學習到的參數比重加總，作為預測任務的輸入。

在實驗部分，本研究採用論文[11]所提出的標準資料集，用於以下兩個預測任務。第一個預測任務與論文[11]相同，以進入ICU的初期48小時內的儀器診斷數據作為輸入，預測病患會在院內發生死亡的機率，稱為院內死亡預測任務。此外，由於病患的住院時間長短差異很大，院內死亡發生時間未必在幾天內，因此造成病患死亡的病況事件因素未必只有發生在ICU的初期48小時，此情況下，這個預測任務的意義有限。所以本研究另實驗第二個預測任務，以進入ICU的初期48小時內的儀器診斷數據作為輸入，預測病患在接下來48小時內死亡的機率，以提供病患短時間內的死亡率預測，稱為短期死亡預測任務。由於[11]在預測任務的輸入採用儀器偵測數據原偵測值，未進行特徵表示法學習。本論文則以前處理將原偵測值先轉成病況事件，並進行病況事件表示法學習後作為輸入。因此在實驗中將進行以下評估：

- 1)將標準資料集做異常值處理與數值區間化等前處理後對模型預測效果的影響。
- 2)考慮以不同時間間隔關聯建立異質網路圖，並以不同的病況事件走訪樣式產生事件序列，比較所訓練出的病況事件表示法對模型預測效果的效果。
- 3)加入實驗室化驗結果病況事件資訊之模型預測效果。

## 1.5 論文架構

本論文以下章節架構說明如下：在第二章說明相關文獻探討，第三章說明本論文之問題定義及系統處理架構，第四章說明資料前處理、特徵產生方式及預測模型細節。第五章將呈現本論文方法的實驗結果評估與討論，最後在第六章提出總結並討論未來研究方向。

## 第 2 章 文獻探討

近十幾年來，許多研究學者針對如何從醫療資料庫的大量資料進行資料分析及預測的技術進行探討。本章節將介紹以往表徵學習法的技術，包含文字資料與醫療資料的表徵學習法，以及採用同質性或與異質性資料的表徵學習方法，最後介紹建立死亡風險預測任務模型的相關研究。

### 2.1 表徵學習法

表徵學習法是一種特徵提取的技術，讓機器可以自動從原始資料值提取出一個特徵向量，表達出在不同特徵維度上的數值，讓機器在後續的預測任務中獲得更好的表現。以下將依不同資料型態的表徵學習法進行介紹，分別為文字資料表徵學習、醫療代碼表徵學習法、同質醫療資料網路及異質醫療資料網路上的表徵學習。

#### 2.1.1 文字資料表徵學習

在自然語言中，文字資料的表徵學習，是為了將文字轉換成一個語意向量，用以有效計算文字間語意的相近程度，或用於後續句子及文章層級的語意理解。詞嵌入(word embedding)是自然語言中表徵學習過程的通稱，目標是將維度為所有詞數量的一個高維二元向量轉換到一個較低維度的連續向量空間當中，且能有效表達語意。近年來，運用最廣泛的詞嵌入模型是 2013 年 Google 的 Tomas Mikolov 團隊所提出的 Word2Vec[7][8]。Word2Vec 透過由大量文本資

料中字與字在前後文中一起出現的關係，對每個字學習一個向量表示法，使得語意較相近的字詞，其向量間具有較高的餘弦相似度。

Word2Vec 主要分成採用 CBOW 或 Skip-gram 兩種模型，Skip-gram 模型架構如圖 2.1，以輸入一個字，將換成字的嵌入式向量 $v(w_t)$ ，預測出其前後文一定範圍內，一起出現字有較高出現機率為目標，以學習出字的嵌入向量。CBOW 模型則是以前後文的字作為輸入，預測該字輸出機率值最高當作目標進行訓練。

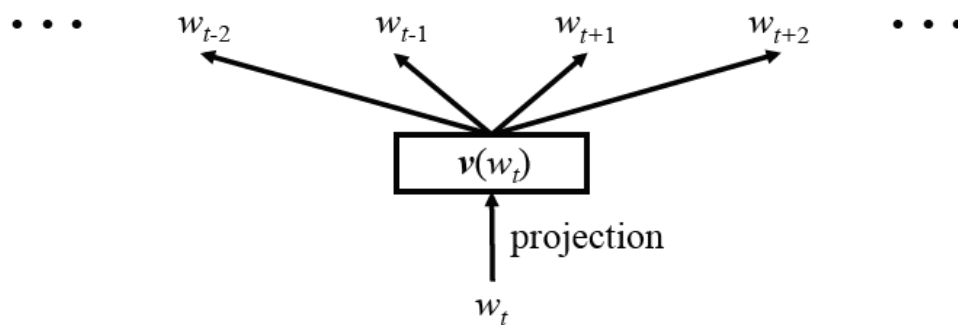


圖 2.1 Skip-gram 模型[7][8]

### 2.1.2 醫療代碼表徵學習法

Med2vec[1]擴展 Skip-gram 模型的概念，提出一種由病患看診資料中學習醫療代碼嵌入向量之表示學習模型。這個模型的學習基礎是根據同次及前後次看診中醫藥代碼的出現關聯，在網路架構中將同次看診醫療代碼表示法結合成一個向量後，以預測前後次出現病症為目標進行學習。圖 2.2 顯示 Med2vec 的模型架構，模型中同次看診醫療代碼集合以高維度的二元向量表示(multi-hot)

輸入，經過雙層的全連接層(Fully connected layer)及非線性激活函數 ReLU。第一層(code-level)的係數向量 $w_c$ 用來進行醫療代碼表示法的學習，此層的輸出會結合同次看診的醫療代碼表示法得到 $u_t$ ，再接合患者的個人基本資料(demographic information)  $d_t$ 作為第二層(visit-level)的輸入。第二層的輸出表示病患每次看診的表示法 $v_t$ ，經過輸出層預測病患的前後次看診的醫療代碼： $x_{t-2}, x_{t-1}, x_{t+1}, x_{t+2}$ 等，以此方式進行醫療代碼表示法的學習。

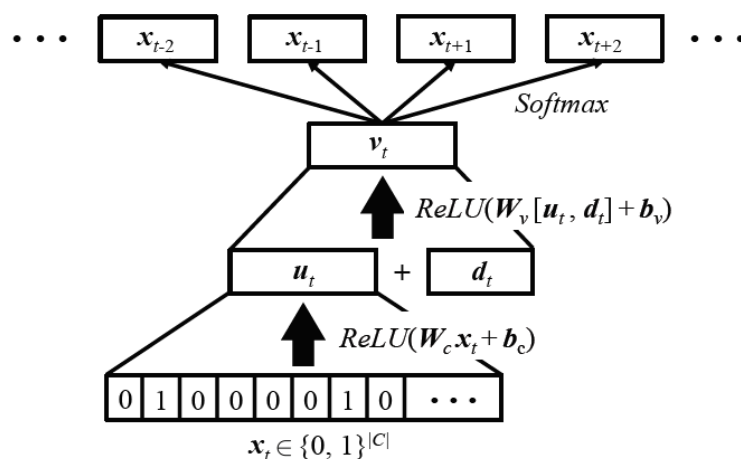


圖 2.2 Med2vec 模型[1]

Med2vec[1]在模型設計上未考慮看診資料中多次看診序列中醫療代碼出現的累積資訊。在深度學習的模型架構中，遞迴神經網路(RNN)可考慮資料出現的時序性，並可將先前資訊累積傳遞下去，因此醫療代碼的表示法學習研究陸續提出基於 RNN 架構設計的模型。RETAIN[3]模型使用兩層具注意力機制的 RNN 模型，模擬醫生看診時判斷病況的行為，依序從最近的病歷資料到最久遠的病歷資料輸入 RNN 模型，注意力機制用來學習先前診斷中出現的每個醫療代

碼對當前病症預測的影響程度，該論文認為上述的資料輸入順序可以使注意力機制權重生成更穩定。

由於單向遞歸式神經網路會因為序列過長遺忘較久遠的資料，導致模型預測效果下降，雙向遞歸式神經網路被認為可以用於改善此缺點而提升預測性能。所以 Dipole[3]採用雙向的遞歸式神經網路(BRNN)[21]，並加上注意力機制學習先前醫療紀錄對當前病症預測的影響比重。Dipole 的輸入資料為患者前  $t$  次看診(visit)的醫療代碼集合，以高維度的二元向量表示法輸入，經過一層嵌入向量層轉換每次看診的表示法，再輸入雙向遞歸式神經網路，得到每次看診綜合由前往後以及由後往前的隱藏狀態向量。Dipole 在遞歸式神經網路輸出的下一層加上注意力機制，由遞歸式神經網路輸出各時間點看診資料累積的隱藏狀態向量(Hidden state)中學習其相對重要性，用來比重加總前  $t$  次看診的隱藏狀態向量，預測第  $(t+1)$  次看診的醫療代碼。

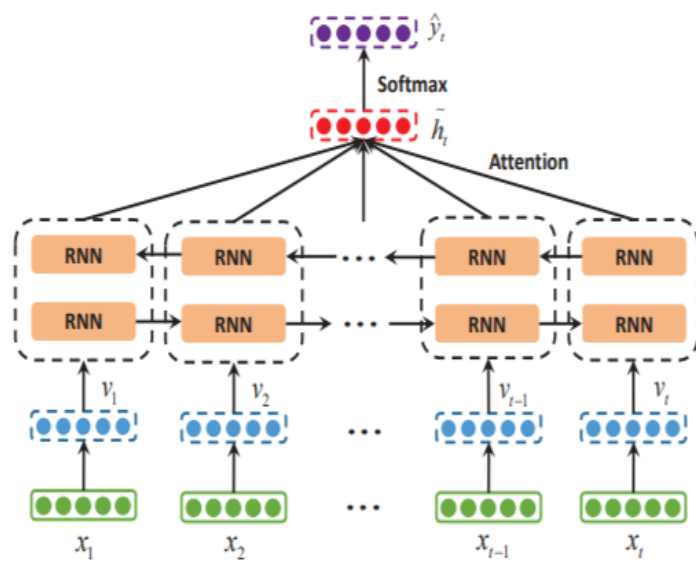


圖 2.3 Dipole 雙向遞歸式神經網路模型[3]



前述兩篇論文雖有考慮看診資料的時間性以 RNN 模型進行特徵擷取，但並未考慮醫療代碼間的語意架構關係。因此，GRAM[2]提出運用醫療資料的本體知識(medical ontology information)圖來組織醫療代碼，該組織圖描述醫療代碼間的語意包含關係。GRAM 的模型架構圖如圖 2.4 所示，圖中左側為 GRAM 的醫療訊息本體知識(medical ontology information)圖，該圖為一個有向無環圖(directed acyclic graph)，其中葉節點代表較為低階的醫療代碼，非葉節點(葉節點的祖先節點)代表較高階的醫療代碼， $e_i$ 代表節點對應的特徵表示法。注意力機制用在如何將醫療訊息本體(medical ontology information)組織圖中不同階層的嵌入向量，以比重加總組合出病患每次看診的醫療代碼表示向量，隨後再以 RNN 由看診序列為預測接下來會出現的醫療代碼。

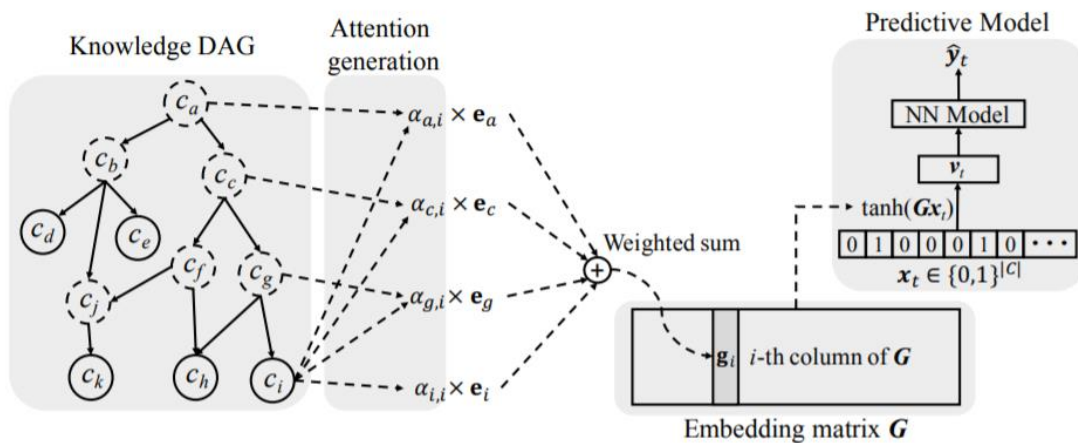


圖 2.4 GRAM 模型[2]

KAME[5]對 GRAM 的模型進行改良，其模型架構圖如圖 2.5 所示。KAME 採用 GRAM[2]中相同的方式學習看診資料的嵌入式向量，但在將看診資料的嵌入向量輸入 RNN 預測模型後，以隱藏向量 $h_t$ 作為查詢值(query)，學習本體組織圖中醫療代碼各祖先節點的特徵表示向量之注意力權重，比重加總組合出知識向量 $k_t$ ，最後接合 $h_t$ 與 $k_t$ 作為最後一層 Softmax 層的輸入，預測下一次看診的醫療代碼。

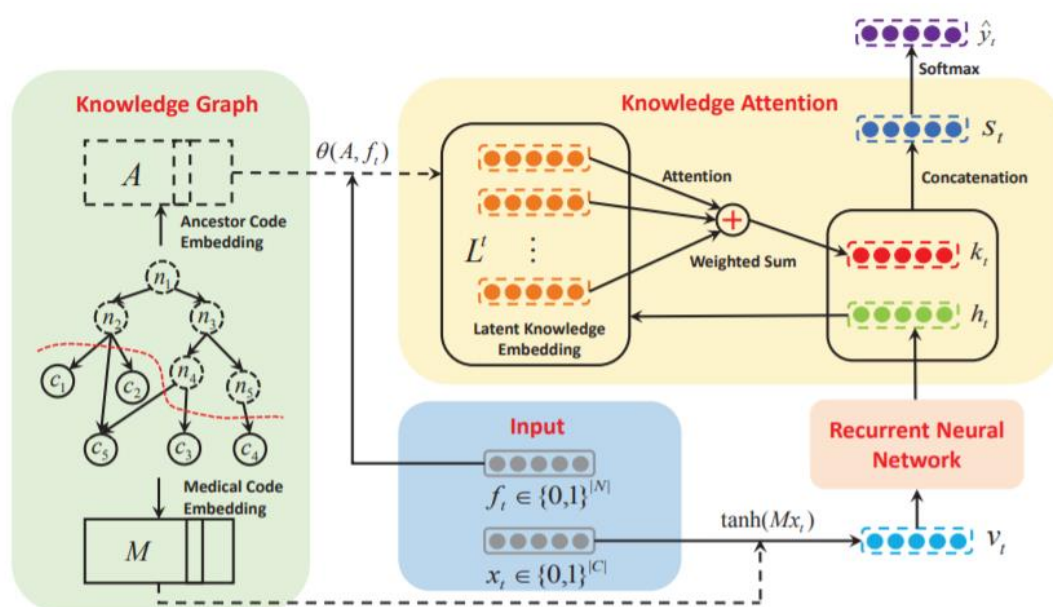


圖 2.5 KAME 模型[5]

### 2.1.3 網路圖表徵學習模型

上述 GRAM[2]及 KAME[5]在學習醫療代碼的表示法，是依據醫療代碼同質資料所形成的本體組織圖上的關係進行學習，且採用有向圖的概念，由下層節點表示法影響上層節點表示法。DeepWalk[9]和 Node2vec[10]則為目前常見的無向同質網路圖表示學習技術，通用於各種應用將資料關係表示成同質網路

圖後，對資料節點進行表示法學習。上述兩種方法皆擴展自 Word2Vec[7][8]，透過取樣網路圖中走訪節點的路徑得到節點序列，將節點序列視為包含字彙的句子，建構出訓練 Word2Vec[7][8]的資料集。

現實生活中有許多應用資料形成的網路圖都屬於異質網路，例如，社交媒體資料對應的節點類型可能包含商家帳號、用戶帳號、打卡地點、發文文章等，而電子商務系統中的資料節點類型可能包含用戶特徵:性別、年紀、瀏覽網頁紀錄，以及商品特徵:價格、類型(3C、家用、食品等)。有別於同質網路，異質網路中由於存在不同類型的節點，不同類型的節點間的走訪路徑具有特殊語意，例如醫生帳號節點連至醫院節點的邊，表示該醫生帳號在醫院看診的關係。若採用論文[9]及[10]同質網路圖的方法學習節點特徵表示，無法將異質網路中更豐富的連接邊資訊有效的學習到特徵表示向量中。

論文 metapath2vec[12]提出針對異質網路圖學習節點特徵表示法的方式，該論文所使用的資料集為「AMiner-Computer Science」資料集，該資料集中有作者、論文、研討會三種類型的節點。進行節點的表示法學習時經常採用的路徑型態(元路徑，metapath)有“作者-論文-作者”及“作者-論文-研討會-論文-作者”。第一種路徑可以採集到共同作者的資訊，第二種路徑則代表不同作者在同一個研討會發表論文的關係。論文[12]根據指定路徑上的節點類型進行採樣，以控制取出節點序列，做為表示法學習模型的訓練資料來源，再將節點序列視為文字序列，採用 word2vec[16]的方法學習得到各節點的表示法。論文[12]

在實驗中評估所學習節點表示法的分類及分群效果，結果顯示論文[12]的方法在異質網路圖中比論文[9][10]更能取到節點間的語意關係，嵌入更豐富的語意資訊。

## 2.1.4 醫療異質網路表徵學習法

EMR 資料庫中有許多醫療代碼以外的資料類型，但前述論文[1][2][3][4]皆未考慮到醫療代碼以外的資料類型。HeteroMed[6]將 EMR 中的病況資料關係以異質網路圖的方式表示，並採用異質節點表示學習法進行病況事件表示學習。HeteroMed[6]採用病況事件節點包含微生物培養(Microbiology test)、實驗室化驗數據(Laboratory test)、醫囑文字記錄(Symptoms)、基本資料(Biography)、用藥、醫療代碼、及醫療處置，儲存的資料值則分成列舉跟連續兩種型態。列舉型態可分別對應到節點，而數值型態則先切成區間，各區間再對應到節點。若一名病患的看診資料中若記錄某些病況事件，則以邊連接該名病患的病況事件對應的節點。根據上述方式建構出的異質網路圖，HeteroMed 採用論文[12]所提出的走訪序列取樣並經過模型訓練，學習出病況事件及醫療代碼的表示法。

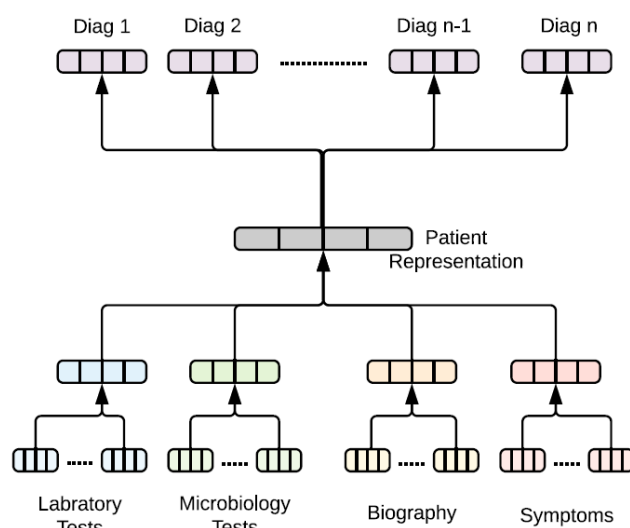


圖 2.6 HeteroMed 的模型架構[6]

圖 2.7 所示為 HeteroMed 的預測模型架構，經由計算病患的特徵表示法與醫療代碼表示法的相似度後進行排名，取相似度前幾名做為預測醫療代碼。在病患的表示法上，HeteroMed 模型會根據學習好的病況事件表示法，將同樣類型的病況事件節點表示法取平均為該類型的特徵向量，再以權重參數學習各個類型特徵向量的加權比重，以此加權比重加總得到該名病患的病況特徵表示法，最後會計算病患的病況特徵表示法與各醫療代碼表示法的相似度，取最相似的前幾名做為預測病人會被診斷的醫療編碼。

## 2.2 死亡率預測模型

以往加護病房臨床上預測病患死亡率，通常會採用嚴重程度評分指標，由專家訂好的規則評分；醫生經由經驗制定評分方式，用以評估加護病房的病患在臨床醫療上的危險程度。評估死亡率較常使用的指標有 APACHE(Acute Physiology and Chronic Health Evaluation)[23]、SAPS(Simplified Acute Physiology Score)[24]及 SOFA(Sequential Organ Failure Assessment)[25]。評分指標大致可以分成兩類：第一類指標是根據病患在入院時的生理數據、年齡、健康狀況，對其病況進行評分，其中生理數據包含體溫、心律、動脈壓、收縮壓以及血液中的物質(鈉、鉀)含量等，健康狀況則會評估人體之肝臟、心臟、呼吸系統、免疫系統及腎臟五個系統。該類指標中 APACHE 及 SAPS 的區別在於 APACHE 著重於血液的檢測，SAPS 多了肝功能的評估。第二類指標則著重在對病患器官的功能及衰竭程度進行評分，分別對呼吸系統、心血管系統、肝臟系統、凝結系統、腎臟系統、神經系統六大器官給予評分，此類指標如 SOFA 指標(Sequential Organ Failure Assessment)[25]。上述這些指標所計算出來的分數可經由回歸模型預測病患的死亡率。

1996 年 Dybowski 等人研究[22]從醫療資料預測患者經過醫生處置後未來的病症發展，發現以人工神經網路(artificial neural network, ANN)建模效果優於採用邏輯回歸分析。論文[26]以 ELM(Extreme Learning Machine)[27]人工神經網路的模型訓練加護病房死亡率模型，該論文將患者的心電圖臨床文本報告轉

換成相同維度的向量表示，作為 ELM 預測模型的輸入。ELM 屬於單隱藏層的前饋神經網路(single hidden layer Feedforward Neural Networks)，由兩層神經網路結構組成，該網路會隨機設定輸入層至隱藏層的參數，並訓練隱藏層至輸出層的網路參數。論文[28]則從手術紀錄中提取結構化的數據特徵作為模型輸入，例如血壓值、輸入的藥物及輸入的血液量等，以深度神經網路(deep neural network, DNN)為預測模型，模型共有四層隱藏層，皆以全連接的方式連接，最後以輸出層預測接受全身麻醉的手術患者於術後的院內死亡率。

醫療病歷中記錄的資料通常具有時序性，隨著病患在院內或加護病房的時間點不同，其病況資料也有會有所改變。ANN 與 DNN 模型無法準確的學習資料隨時間推移的資訊，而遞歸式神經網路更適合時序性的資料。2015 年 E Choi 等人提出了 DoctorAI[29]，開發出一種針對時序性電子病歷紀錄的通用模型，該模型採用遞歸式神經網路，以患者的病歷紀錄(包含醫師歷史診斷、用藥資訊、醫生處置方式)作為輸入資訊，預測患者後續的診斷編碼及用藥編碼。

2017 年 Harutyunyan 等人提出的論文[11]中將遞歸式神經網路的改良模型 LSTM 應用在預測加護病房患者院內死亡率任務。圖 2.8 所示為論文[11]提出的死亡率預測模型架構，每個時間點的輸入為  $x_t = (x_{t,1}, x_{t,2}, \dots, x_{t,17})$ ， $t$  表示入院後第  $t$  個小時，為介於 0 到 48 間的正整數。將  $x_t$  的 17 個特徵輸入雙向的長短期記憶模型，接著接合(concat)順向與逆向隱藏層(Hidden Layer)的輸出  $\vec{h}_t$  及  $\overleftarrow{h}_t$ ，再輸入一層單向的長短期記憶模型，以最後一個時間點的隱藏層  $h_{48}$  連接至輸出



層，並以 Sigmoid 函數做為激活函數，轉換輸出層的值介於 0 到 1 之間，預測死亡機率。此外，[11]認為 MIMIC III 資料集的提出，雖然提供醫療資料研究人員許多可使用資料，但此資料集太過複雜，導致研究者在處理資料上不一致，因此論文[11]提出一個具代表性且沿用至今的標準資料集(Benchmark)，該標準資料集提供了對 MIMIC III 資料集相同的資料前處理方式，制定了篩選病患資料及填入遺失值的準則，並以特定的 17 種儀器偵測數據特徵作為死亡率模型輸入。

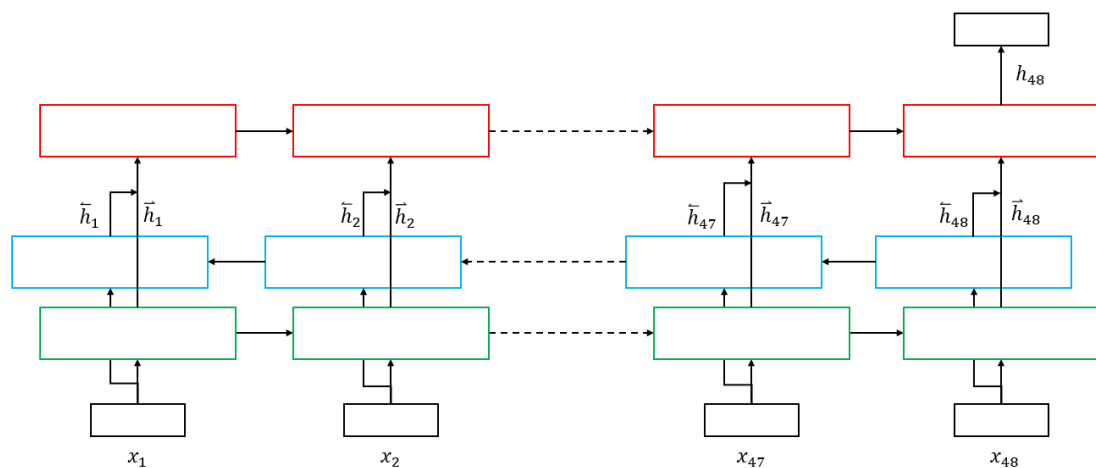


圖 2.7 論文[11]模型架構

此後，神經網路架構上新的技術也陸續被應用在院內死亡率預測任務。2018 年，SONG 等人提出 SAnD (Simply Attend and Diagnose)模型[30]，首次採用注意力機制在醫療序列資料上建模，但未使用遞歸式神經網路建模。該論文以病患的時序性醫療資料作為輸入，病患的醫療資料表示為 $(x_1, x_2, \dots, x_T)$ ，以向量 $x_t$ 表示時間點  $t$  的各式醫療偵測值，以注意力機制將 $x_t$ 作為查詢值，計算其他時間點向量的相對權重，進行比重加總，產生每個時間點的情境向量(context vector)。最後接合各時間點的時間向量，以線性連接輸出層預測患者

的院內死亡率。

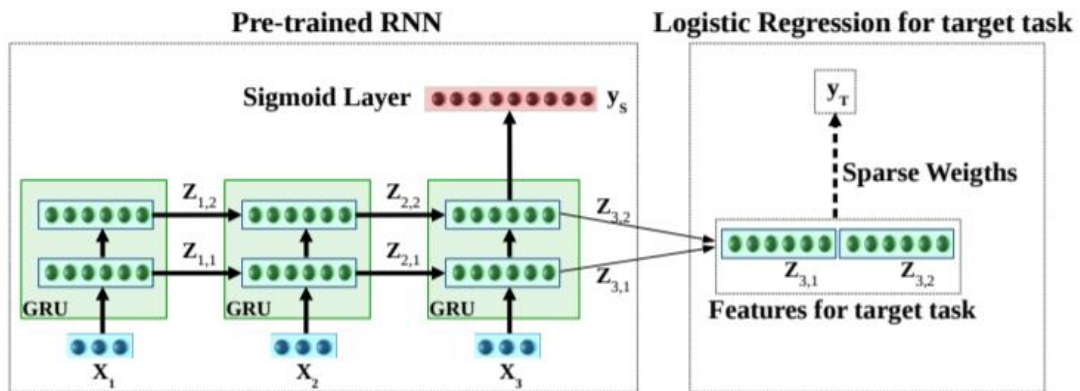


圖 2.8 論文[31]模型架構

2020 年，GUPTA 等人[31]則提出以遞歸式網路對醫療序列資料進行遷移式學習(Transfer Learning)建模，預測患者的死亡率。由於遞歸式神經網路模型需要大量訓練資料，遷移式學習提供了一種方式先在來源任務上預先訓練模型，以來源任務學習到的參數做為目標任務的初始輸入，對訓練資料數較少的目標任務提昇其模型預測效能。圖 2.9 顯示論文[31]的模型架構，以遞歸式神經網路作為來源任務訓練模型，採用論文[11]提出的標準資料集中的一部分作為來源任務訓練模型的輸入，並將來源任務設定為預測患者 25 種急性護理狀況分類，再取其他資料作為目標任務的訓練集。在進行目標任務訓練時，將資料輸入以來源資料集預先訓練好的遞歸式神經網路，擷取最後一層的隱藏層作為邏輯回歸模型的輸入，預測患者的死亡率。該論文亦以原本的儀器偵測值作為輸入，並未對特徵值進行特徵表示法學習。

前述論文[30][31]皆採用論文[11]提出的標準資料集上驗證其模型的效果，因此本論文亦採用此標準資料集驗證模型效果。除了採用論文[11]所制定的前處理準則，本論文另提出清理異常值的作法，以增進模型學習及預測效果。此外，不同於論文[11]以原偵測值作為輸入，本論文認為透過異質網路圖上特徵值的關係預先訓練特徵表示法，可以運用相異類型特徵值的出現，提供特徵表示法學習更多資訊，因此，本論文在前處理部分將連續型態數值區間化以建立異質網路圖，預先訓練病況事件的特徵表示法，並以病況事件特徵表示法做為院內死亡預測模型的輸入。

## 第 3 章 問題定義與系統架構

本章節將介紹在本論文中所提要預測的任務，並介紹系統處理架構。

### 3.1 預測任務定義

令  $\chi = (x_1, x_2, \dots, x_n)$ ，表示病患在加護病房的病歷紀錄，其中  $x_i$  為病患在加護病房第  $i$  個小時中的醫療病況資料， $x_t$  又表示為  $x_t = \{x_{t,1}, x_{t,2}, \dots, x_{t,17}\}$ ， $x_{t,j}$  表示病患第  $t$  個小時在第  $j$  種儀器偵測數據特徵的偵測值，17 種類型的儀器偵測數據特徵如第 1 章中的表 1.1 所列。

給定病患進入加護病房前 48 小時的病歷紀錄  $\chi = (x_1, x_2, \dots, x_{48})$  為輸入，預測病患發生院內死亡的機率值  $P_{mortality}(\chi)$ ，該機率值介於 0 至 1 之間，數值越大代表死亡風險越高。本論文第一個預測任務是院內死亡率，定義為病患進入加護病房第 48 小時後在出院前死亡的機率，第二個預測任務是短期死亡率，定義為病患進入加護病房後第 48 小時至第 96 小時間的死亡機率。

## 3.2 系統流程架構

本論文系統架構分為模型訓練流程與模型測試流程，其中模型訓練流程如圖 3.1 所示。

### (一) 資料前處理(Pre-processing)：

此階段將進行資料中的空值填補，處理資料中的異常值，並對儀器偵測數據連續型態特徵進行區間化，對應到不同的病況事件編碼。

### (二) 特徵表示法學習：

此處理階段將建立異質結構網路圖(Heterogeneous Network Construction)，並由異質網路圖學習病況資料中儀器偵測數據及實驗室化驗數據病況事件的特徵表示法(Representation Learning)。

### (三) 建立死亡率預測模型(Predict Model Construction)：

參照論文[11]所提模型，以階層式方式建立長短期記憶模型作為病況累積資訊的第一層編碼架構，再連接一層長短期記憶模型，最後以輸出層預測死亡率。此階段將載入預先訓練好的病況事件特徵表示法，並以病患訓練資料訓練死亡預測模型。

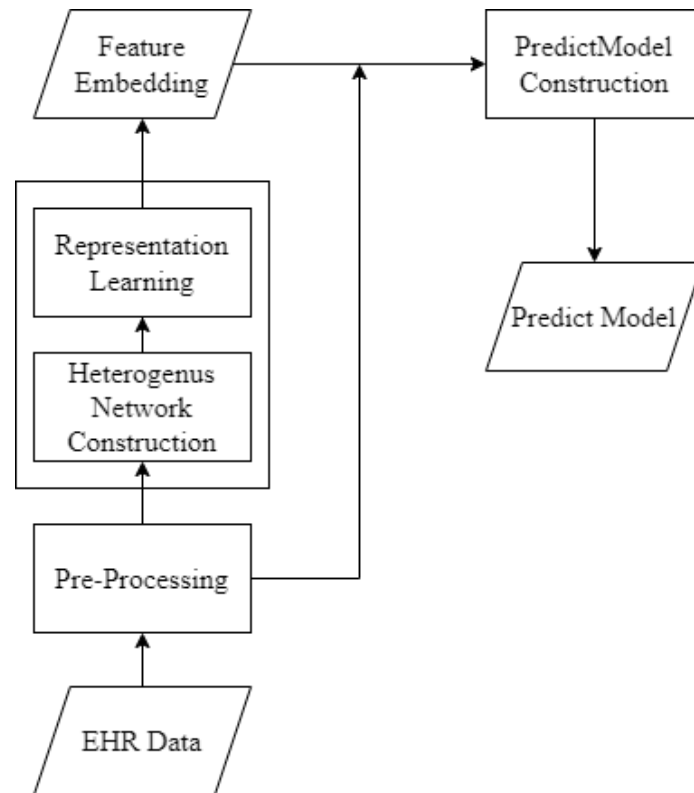


圖 3.1 模型訓練流程圖

模型測試流程如圖 3.2 所示，分為兩個處理階段。

(一)資料前處理(Pre-processing):

測試模型階段所輸入的測試資料集，若存在有空值與異常值，亦先以 4.1 所述方法進行空值填補與異常值處理。

(二)輸入測試模型:

載入預先訓練好的病況事件特徵表示法於預測模型中，將測試資料輸入死亡測試模型，得到預測機率值。

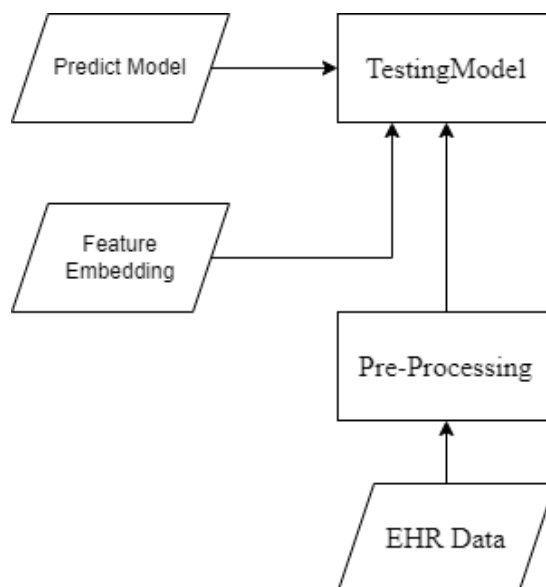


圖 3.2 模型測試流程圖

## 第 4 章 研究方法

### 4.1 資料前處理

#### 4.1.1 填補遺失值

本論文所採用的資料集是論文[11]對 MIMIC III 資料集提出的標準資料集，以表 4.1 為例，表中記錄病患入住加護病房的基本資料，包括病患編號 (SUBJECT\_ID)、入院編號 (HADM\_ID)、入住時間 (IN\_TIME) 與離開時間 (OUT\_TIME)。該資料集以入院編號 (HADM\_ID) 做為病患查找儀器偵測數據及實驗室化驗結果等病況紀錄的連結。

表 4.1 病患入住加護病房紀錄範例

SUBJECT_ID	HADM_ID	IN_TIME	OUT_TIME
82574	118464	2100/6/1 08:00:00 PM	2100/6/5 02:59:00 PM

表 4.2 顯示病患在加護病房的一次儀器偵測數據病況事件紀錄，每筆病況事件紀錄包括入院編號 (HADM\_ID)、發生時間 (TIME)、病況特徵 (FEATURE) 及對應偵測值 (VALUE)，原始資料中記錄每名病患在加護病房中發生的所有病況特徵與對應偵測值。

表 4.2 儀器偵測數據病況紀錄

HADM_ID	TIME	FEATURE	VALUE
118464	2100/6/7 08:00:00 PM	Heart Rate	107

本論文透過表 4.1 中病患的入院編號，查找病患在加護病房的前 48 小時的



病況事件紀錄，並以表 4.3 的格式紀錄。表中的資料以小時做為單位，依發生時間點先後排序，每筆資料中各欄位記錄該時間點發生的病況特徵對應的偵測值，其中病況特徵為表 1.1 所列舉的 17 種儀器偵測特徵。

表 4.3 訓練集資料範例

Hour	Weight	Height	...	pH
1	空值	161.0	...	6.4
2	66.5	0.0	...	6.4
...	...	...	...	...
48	67.0	0.0	...	7.0

由於在加護病房中並非每個時間單位皆有紀錄上述 17 項特徵，本論文參考論文[11]中於前處理時填入資料空值的方式如下：填入該欄位比該時間點早且最接近的紀錄值。舉例來說，以 $x_{i,j}$ 表示第  $i$  個小時在特徵  $j$  的偵測值，若 $x_{10,j}$ 是空值，則取前一時間點的值 $x_{9,j}$ 填入，若 $x_{9,j}$ 亦為空值，則填入 $x_{8,j}$ ，以此類推；若比 $x_{10,j}$ 更早的時間點皆為空值，則會填入論文[11]提供的估算值(Impute Values)，各儀器偵測數據特徵的估算值如表 4.4 所列。

表 4.4 儀器偵測數據特徵估算值

Chart Feature	Impute Values
Weight	81.0
Height	170.0
Temperature	36.6
Systolic blood pressure	118.0
Diastolic blood pressure	59.0
Mean blood pressure	77.0
Heart Rate	86
Capillary refill rate	0.0
Respiratory rate	19
Oxygen saturation	98.0
Fraction inspired oxygen	0.21
GCS motor response	Obeys Commands
Glasgow coma scale total	15
GCS verbal response	Oriented
GCS eye opening	Spontaneously
Glucose	86
pH	7.4

## 4.1.2 清除異常值

原始資料中除了空值，另外會出現過高或過低，一些非常態下的數值，這些數值很可能是當初人工輸入資料庫時記錄錯誤，亦可能是該名病患的病況特殊而導致某些病況特徵出現極端的數值。論文[11]針對各特徵提供一個合理值的範圍，如表 4.5 所示。表 4.5 列出論文[11]針對各連續型態儀器偵測數據特徵訂出之低異常值(OUTLIER LOW)、低端值(VALID LOW)、高端值(VALID HIGH)及高異常值(OUTLIER HIGH)值，其中介於低端值與高端值之間的數值視為正常合理的數值範圍。但基於某些病患的病情特殊，再訂定超出正常範圍的極端值容許範圍，也就是在低異常值與高異常值間是可以容許的。

以 $F_{continuous}$ 表示連續型態儀器偵測數值特徵所成的集合，對於 $F_{continuous}$ 中的任一個特徵 $f$ ， $f_t$ 代表 $f$ 在第 $t$ 個小時的特徵值，若 $f_t$ 大於高端值，但小於等於高異常值，表示尚在容許範圍內，以高端值取代；若 $f_t$ 小於低端值，但大於等於低異常值，也表示尚在容許範圍內，以低端值取代。若 $f_t$ 小於低異常值或大於高異常值，則視為異常值。在處理異常值的策略上，本論文會以該時間點先往後尋找有正常值紀錄的最接近時間點之數值取代此異常值，若往後找不到，再往前找。若該筆資料的特徵 $f$ 在 48 小時中皆不存在正常值，將視此筆資料為異常資料，刪除該筆資料。舉例來說，若 $f_4$ 為異常值，則會查找下一小時的 $f_5$ 值，若 $f_5$ 為正常值則取代 $f_4$ 的異常值；否則再往下一小時查找 $f_6$ 。若 $f_5$ 到 $f_{48}$ 皆為異常值，則會往前一個小時查找 $f_3$ 。

表 4.5 儀器偵測數據正常範圍值

	<b>OUTLIER LOW</b>	<b>VALID LOW</b>	<b>VALID HIGH</b>	<b>OUTLIER HIGH</b>
<b>Weight</b>	0	0	250	250
<b>Height</b>	0	0	240	275
<b>Temperature</b>	14.2	26	45	47
<b>Systolic blood pressure</b>	0	0	375	375
<b>Diastolic blood pressure</b>	0	0	375	375
<b>Mean blood pressure</b>	0	14	330	375
<b>Heart rate</b>	0	0	350	390
<b>Respiratory rate</b>	0	0	300	330
<b>Oxygen saturation</b>	0	0	100	150
<b>Fraction inspired oxygen</b>	0.2	0.21	1	1.1
<b>Glucose</b>	0	33	2000	2200
<b>pH</b>	6.3	6.3	8.4	10

### 4.1.3 病況事件編碼

儀器偵測數據特徵中的值分為連續型態與類別型態，連續型態特徵必須先進行區間化處理，轉換成以不同的特徵值區間表示不同的病況事件，類別型態特徵中不同的列舉值則可各別對應到病況事件。以下將分別說明連續型態特徵及類別型態特徵的病況事件編碼方式。

### (一)連續型態特徵病況事件編碼

儀器偵測數據類型的病況特徵中有 12 個特徵屬於連續型態，將每個連續特徵依照表 4.5 所列出之高端值與低端值做為該特徵的最大值與最小值，以等距 (Equal Width) 切割成 10 個等分，每個區間對應到一個病況事件，並給予對應的病況事件編碼。經由上述劃分區間的方式，表 4.6 列出各連續型態特徵所劃分出的特徵值區間大小，以及所對應到的病況事件編號範圍。連續型態特徵對照病況事件的區間範圍與對應編碼請見附錄 A-2。

表 4.6 連續型態病況事件劃分區間大小及其編號範圍

特徵名稱	特徵值等分區間大小	病況事件編號範圍
Weight	25.0	131~140
Height	24.0	71~80
Temperature	1.9	121~130
Systolic blood pressure	37.5	111~120
Diastolic blood pressure	37.5	2~11
Mean blood pressure	31.6	81~90
Heart Rate	35.0	61~70
Respiratory rate	30.0	101~110
Oxygen saturation	10.0	91~100
Fraction inspired oxygen	0.08	12~21
Glucose	196.7	51~60
pH	0.21	141~150

## (二)類別型態特徵病況事件編碼

### (1)儀器偵測數據中的類別型態特徵

類別型態特徵的特徵值可被完整列舉出來，例如睜眼反應(Eye opening)為儀器偵測數據類型特徵中的一種類別型態特徵，其特徵值可列舉出有："Close"、"No Response"、"To Pain"、"To speech"、"Spontaneously"五種，本論文將該特徵結合其每一個特徵值視為一種病況事件。表 4.7 列出儀器偵測數據中各個類別型態特徵名稱，並列出對應的病況事件編碼範圍。舉例來說，若病患睜眼反應為" To Pain"，則該病況事件對應到編碼 24。類別型態特徵完整的病況事件對應編碼請見附錄 A-1。

表 4.7 儀器偵測數據類別型態病況事件編碼範圍

特徵名稱	特徵值個數	病況事件編碼範圍
Capillary refill rate	2	0~1
Motor response	6	27~32
GCS total	13	33~45
Verbal response	5	46~50
Eye opening	5	22~26

## (2)實驗室化驗結果特徵

在 MIMIC III 資料集中，病患若有進行某項化驗檢查，則會紀錄該項檢驗的實驗項目 (Laboratory test) 及數值是正常 (Normal) 或異常 (Abnormal)。本論文參考論文[6]提出對實驗室化驗病況事件編碼的方式，將實驗項目配上數值是否正常，各對應到一個病況事件。

經由上述病況事件編碼後，17 項儀器偵測數據共有 151 個儀器偵測數據病況事件，形成  $V_{chart}$ 。實驗室化驗項目中共有 753 項，所以共有 1506 個病況事件，形成  $V_{lab}$ 。

## 4.2 特徵表示法學習

本小節將介紹如何建立以病況事件為節點的異質網路圖，以及訓練病況事件特徵表示法細節。

### 4.2.1 建立病況事件異質網路圖

病患進入加護病房前 48 小時的病況資料經過比對病況事件的屬性值範圍，轉換成病況事件紀錄，紀錄格式如表 4.8 所示。表中每個橫列記錄該時間單位內是否發生儀器偵測數據中 151 項病況事件，若該小時內發生某個病況事件則標 1，未發生則標 0。令  $p_t$  表示病患第  $t$  個小時是否出現各病況事件紀錄的一個 151 維的二元向量(multi-hot)， $t \in [0,48]$ ， $p_t$  作為病患時間點  $t$  的病況輸入向量。

表 4.8 已劃分區間的儀器偵測數據病況事件紀錄

Time	$Event_0$	$Event_1$	...	$Event_{150}$
1	$p_{1,0} = 1$	$p_{1,1} = 0$		$p_{1,150} = 1$
2	$p_{2,0} = 0$	$p_{2,1} = 1$		$p_{2,150} = 2$
...				
48	$p_{48,0} = 0$	$p_{48,1} = 1$		$p_{48,150} = 1$

[定義] 異質網路圖(Heterogeneous Network):

異質網路圖的定義為一個圖型  $G = (V, E, A, g_v)$ ，其中  $V$  代表節點所成的集合， $E$  代表邊所成的集合， $A$  代表所有節點類型所成的集合，函式  $g_v$  將  $V$  對應到  $A$ ，其中  $|A| > 1$ 。



在本論文中， $V$  代表病況事件對應節點所成的集合， $A$  包括有儀器偵測數據及實驗室化驗結果兩種類型。因此將  $V$  中屬於儀器偵測數據類型的事件以  $V_{chart}$  表示，屬於實驗室化驗結果類型的事件以  $V_{lab}$  表示。

建立病況事件為節點的網路圖時，以病況事件對應到圖中的節點，節點間的邊決定方式則以同一個病患在固定時間區間內是否同時發生兩個病況事件作為判斷依據。例如表 4.8 中的病患在第一小時的病況事件中記錄  $p_{1,0} = 1$ ，且  $p_{1,150} = 1$ ，表示對同一病患這兩個病況事件  $Event_0$  和  $Event_{150}$  在同一時間點發生，可能具有關聯性，因此在網路圖上建立對應節點  $v_0$  與節點  $v_{151}$  相連的邊  $(v_0, v_{151})$ 。由於在臨床醫療上，除了在同一小時內發生的病況事件可能有關聯性，更長的時間間隔(Window Size)內共同發生的病況事件也可能有關聯。所以本論文方法可放寬時間範圍，根據在不同時間間隔內同一病患發生的病況事件來建立網路圖中的邊。本論文在實驗中將比較設定不同時間間隔建立網路圖，對訓練出來的節點表示法評估在預測模型上的效果。

此外，本論文受到論文[6]的啟發，以不同類型的病況事件建立異質網路圖，再透過異質網路圖上節點的連接關係，結合異質類型病況事件資訊學習病況事件表示法。因此除了建立  $V_{chart}$  與  $V_{chart}$  間的邊，也建立  $V_{chart}$  與  $V_{lab}$  間的邊。根據章節 4.1.3 對病況事件的編碼，編號 151 至編號 1656 是實驗室化驗結果類型的病況事件，因此將表 4.9 考慮的病況事件編號擴展到共 1656 種，涵蓋儀器偵

測數據與實驗室化驗結果的病況事件，同上所述，以相同的方式建立圖中同質及異質病況事件節點的邊，即可建立出病況事件異質網路圖。

## 4.2.2 病況事件特徵表示法學習

本論文採用 metapath2vec[12] 中的方式，以符合元路徑(meta-path)的隨機遊走取出走訪序列做為產生訓練資料的依據，接著以 Word2Vec 的方式學習各病況事件的特徵表示法，以下將說明元路徑序列的取樣方法與特徵表示法學習模型。

**[定義]**元路徑(Metapath)[14]:

元路徑用來指定網路圖中的節點路徑走訪型態，一個元路徑表示為  $A_1 \rightarrow A_2 \rightarrow \dots \rightarrow A_m$ ，其中  $A_i$  代表第  $i$  個節點的型態。

### (一)元路徑取樣

論文[12]提出基於元路徑的隨機遊走(Meta-Path-Based Random Walks)方式從異質網路圖取樣出節點走訪路徑。由給定的異質網路圖  $G=(V, E, A, g_v)$ ，以及元路徑  $\rho$  產生取樣路徑時，是對連接到  $v_t$  的各個相鄰節點中符合元路徑上限定型態的節點，計算走訪機率以取樣下個走訪節點產生走訪路徑。計算下一走訪節點的機率方式如公式 4.1 所示。

$$p(v_{t+1}|v_t, \rho) = \begin{cases} \frac{1}{|N(A_{t+1}, v_t)|} & (v_t, v_{t+1}) \in E, g_v(v_{t+1}) = A_{t+1} \\ 0 & (v_t, v_{t+1}) \in E, g_v(v_{t+1}) \neq A_{t+1} \\ 0 & (v_t, v_{t+1}) \notin E \end{cases} \quad (\text{公式 4.1})$$

其中 $v_t$ 表示走訪路徑上第  $t$  個節點， $A_{t+1}$ 表示元路徑上限定的下一節點型態， $N(A_{t+1}, v_t)$ 代表 $v_t$ 的鄰居節點型態屬於 $A_{t+1}$ 所成的集合。

由於儀器偵測數據病況特徵能直接反應病患的生理狀態，因此對應的病況事件做為預測模型的輸入。為了學習到這些病況事件的表示法，本研究設定的元路徑開頭皆為 $V_{chart}$ ，且為了避免在異質圖上路徑過長可能取出不相關的事件序列，因此採用以下三種元路徑的節點型態序列進行隨機遊走。

- 1)  $V_{chart} \rightarrow V_{chart}$
- 2)  $V_{chart} \rightarrow V_{lab}$
- 3)  $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$

## (二)特徵表示法學習模型

根據元路徑進行隨機遊走，取樣出事件序列為訓練資料集後，再採用 Word2vec[7][8] 中 Skip-gram 模型對事件節點的特徵表示法進行學習。

本論文方法取樣出的事件節點序列，可以視這些節點序列為文字形成的句子，以路徑中任一個事件當作輸入事件，當視窗大小(window size)設為 1，產生的訓練配對是以該事件在取樣路徑上前一個或後一個事件為輸出目標。舉例來說：以元路徑 $\rho = V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 為例，可能取樣出節點序列 $(v_0 \rightarrow v_{151} \rightarrow v_{10})$ ；以 $v_0$ 為輸入事件時，則訓練配對為 $(v_0, v_{151})$ ，以 $v_{151}$ 為輸入事件，訓練配對則有 $(v_{151}, v_0)$ 及 $(v_{151}, v_{10})$ 。同樣的，以 $v_{10}$ 為輸入事件，對應的訓練配對有 $(v_{10}, v_{151})$ 。當視窗大小(window size)設為 2，產生的訓練配對是以該事

件在取樣路徑上前 2 個或後 2 個範圍內的事件為輸出目標。因此對上述節點序列( $v_0 \rightarrow v_{151} \rightarrow v_{10}$ )，以 $v_0$ 為輸入事件時，則訓練配對為( $v_0, v_{151}$ )及( $v_0, v_{10}$ )，以 $v_{151}$ 為輸入事件，訓練配對則有( $v_{151}, v_0$ )及( $v_{151}, v_{10}$ )。同樣的，以 $v_{10}$ 為輸入事件，對應的訓練配對有( $v_{10}, v_{151}$ )及( $v_{10}, v_0$ )。對於元路徑 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ ，視窗大小(window size)設為 2 相當於較視窗大小(window size)設為 1 增加 $V_{chart}$ 與 $V_{chart}$ 的訓練配對。

圖 4.1 所示為本研究使用 Skip-gram 模型進行病況事件學習的架構圖，輸入層(Input layer)的維度(C)是儀器偵測數據的病況事件數量(151)加上實驗室化驗數據的病況事件數量(1506)，共 1657 維。中間層為嵌入層(Embedding layer)，嵌入層維度為 d，實驗中皆設定為 100。輸出層(Output layer)對應到每個病況事件出現的機率值，維度與輸入層維度一樣為 1657 維。

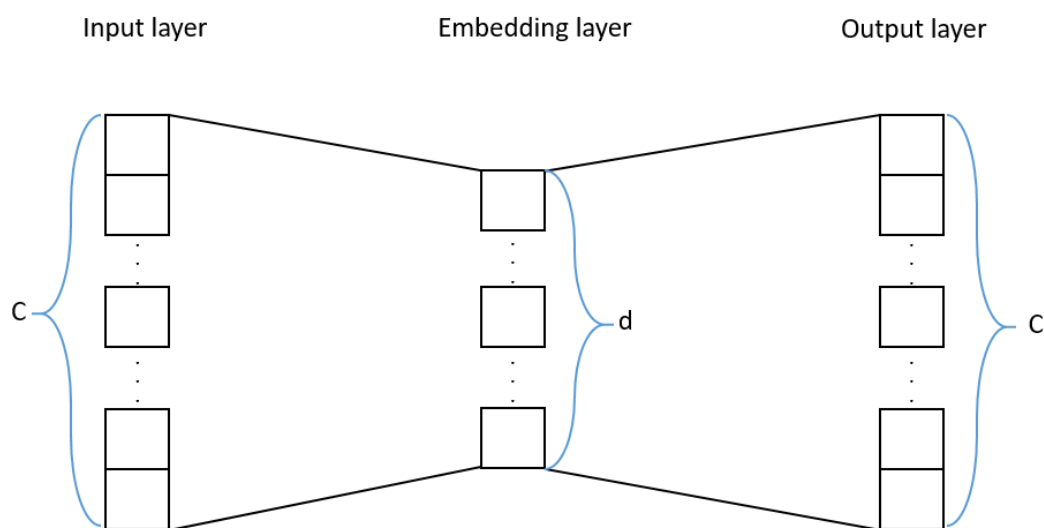


圖 4.1 以 Skip-gram 模型架構圖

對每一筆訓練資料配對 $(e_i, e_i^o)$ ，將輸入病況事件 $e_i$ 以一個獨熱向量 $v_e$ 表示， $v_e \in R^C$ ，經過嵌入層，轉成其嵌入向量 $emb_e$ ，計算方式如公式 4.2。

$$emb_e = W_{emb} * v_e \quad (\text{公式 4.2})$$

嵌入層的輸出為 $emb_e$ ，再輸入到輸出層。嵌入層與輸出層間以全連接的方式連接，並對各輸出結果代入 Sigmoid 函數，將數值轉成 0~1 的範圍，代表預測出現各事件的機率值。此步驟以公式 4.3 表示。

$$e_{predict} = Sigmoid(W_{out} * emb_e + b) \quad (\text{公式 4.3})$$

其中以 $W_{out}$ 表示全連接層訓練的模型參數， $W_{out} \in R^{d*C}$ ， $b$ 代表偏差(bias)向量， $e_{predict}$ 代表輸出層各事件的輸出機率值所組成的向量。

訓練模型的損失函數如公式 4.4[7]。

$$L = -\frac{1}{N} \sum_{i=1}^N [\log \widehat{y_{e_i^o}} + \log(1 - \widehat{y_{e_i^n}})] \quad (\text{公式 4.4})$$

其中  $N$  代表訓練資料筆數，對第  $i$  筆訓練資料配對 $(e_i, e_i^o)$ ，隨機取出一筆不是答案 $e_i^o$ 的事件 $e_i^n$ 與 $e_i$ 配對，形成 $e_i$ 的預測負樣本(negative sample) $(e_i, e_i^n)$ 。公式 4.4 中的 $\widehat{y_{e_i^o}}$ 代表輸入事件 $e_i$ 後模型預測事件 $e_i^o$ 的輸出機率； $\widehat{y_{e_i^n}}$ 代表輸入事件 $e_i$ 會預測到負樣本事件 $e_i^n$ 的輸出機率。

## 4.3 死亡率預測模型

### 4.3.1 模型架構

本章節將介紹由病患的一連串病況事件序列進行死亡率預測模型架構，如圖 4-3 所示，由下往上依序是輸入層(Input layer)、嵌入層(Embedding layer)、雙向長短期記憶層(Bi-LSTM layer)、長短期記憶層(LSTM layer)、及輸出層(Output layer)。

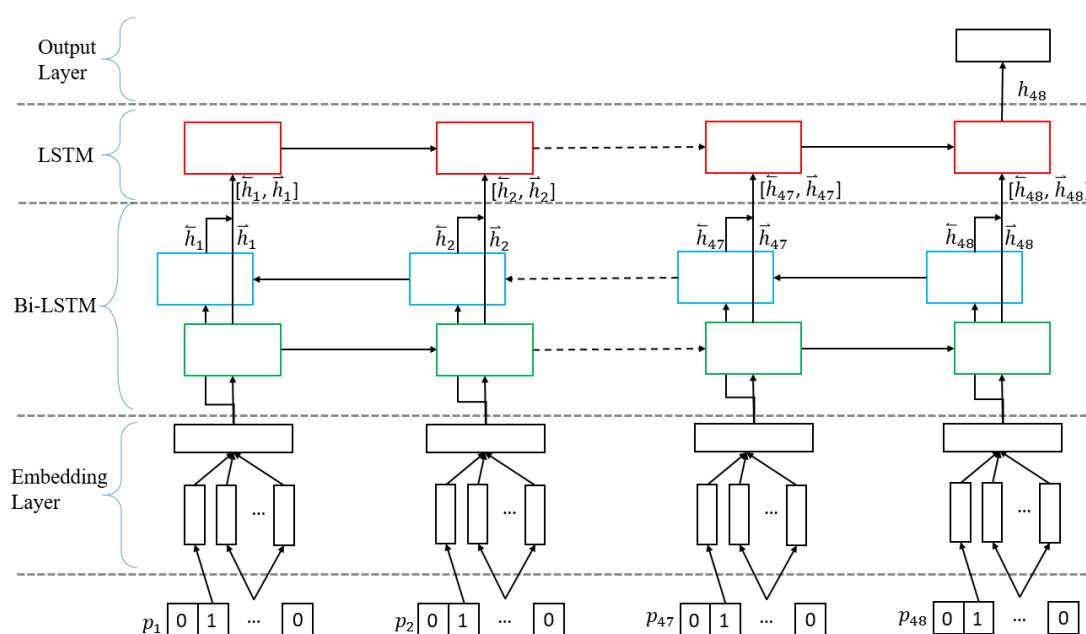


圖 4.2 死亡率預測模型架構圖

以下將會說明以這些網路層建立死亡預測模型的細節，包含每個網路層的計算方式及整體模型的損失函數計算方式。

### (1)輸入資料(Input data)

如圖 4.2 所示，死亡率預測模型輸入的資料，代表病患在時間點  $t$  發生的病況事件以 Multi-hot 二元編碼表示的向量， $p_t$  的維度為儀器偵測數據的病況事件數量，151。

### (2)嵌入層(Embedding layer)

由前一小節 4.2 介紹由異質網路圖預訓練好的病況事件的特徵表示矩陣  $W_{emb}$  會被載入此嵌入層，並在訓練時固定此嵌入向量層參數  $W_{emb}$ 。輸入的  $p_t$  經過這一層嵌入層得到  $emb_t$ 。

$$emb_t = emb_{chart} * \alpha \quad (\text{公式 4.5})$$

根據  $p_t$  中標為 1 的事件，從預先訓練好的儀器偵測病況事件表示法矩陣  $W_{emb}$  中取出對應的向量，對各特徵的病況事件取每個維度上的最大值，代表該名病患在時間點  $t$  的病況特徵矩陣，以  $emb_{chart}$  表示。公式中的  $\alpha$  是一個維度為 17 的向量，表示各個儀器偵測數據病況事件對應的重要性權重，是由模型學習出來的參數，因此  $emb_t$  表示將病患在時間點  $t$  發生的病況事件表示法進行比重加總的結果。



### (3)雙向長短期記憶層(Bi-LSTM Layer)

雙向長短期記憶模型，同時考慮到資料序列向前與向後累積的資訊，可以捕捉到較完整的語意關係。從圖 4.2 所示，每個時間點以嵌入層輸出 $emb_t$ 輸入 Bi-LSTM，由於 Bi-LSTM 包含順向及逆向的 LSTM 都有隱藏層狀態向量輸出，因此會接合順向隱藏狀態向量 $\vec{h}_t$ 與逆向隱藏狀態向量 $\overleftarrow{h}_t$ ，接合後的隱藏狀態向量為 $[\vec{h}_t, \overleftarrow{h}_t]$ 。

### (4)長短期記憶模型(LSTM Layer)

以上一層 Bi-LSTM 輸出的隱藏狀態向量 $[\vec{h}_t, \overleftarrow{h}_t]$ 做為本層 LSTM 每個時間點的輸入，每個神經單元會輸出一個隱藏狀態向量 $h_t$ ，取前後一個隱藏狀態向量 $h_{48}$ 做為病患的整體病患表示向量 $h_p$ 。

### (5)輸出層

以 $h_p$ 作為輸入，以全連接層輸出一個數值作為輸出值，並以 Sigmoid 做為激活函數，使輸出 $\hat{p}$ 的數值介於 0 到 1 之間，代表預測死亡率，如公式 4.6。

$$\hat{p} = \text{Sigmoid}(W'_{out} * h_p) + b_p \quad (\text{公式 4.6})$$

## (6)損失函數

由於預測任務為二分類，因此以二元值交叉熵(Binary cross entropy)做為損失函數，如公式 4.7 所示。

$$L = -\frac{1}{N} \sum_{i=1}^N [W_{class1} p_i \log \hat{p}_i + W_{class0} (1 - p_i) \log (1 - \hat{p}_i)] \quad (\text{公式 4.7})$$

其中，N 代表訓練資料筆數，第 i 筆資料的分類答案以  $p_i$  表示，模型輸出的預測機率值以  $\hat{p}_i$  表示。此外，本研究在模型訓練時考慮到資料正樣本與負樣本的資料數差異過大，所以設置正樣本分類的權重  $W_{class1}$  與負樣本分類的權重  $W_{class0}$ 。當  $W_{class1}$  設為 1 且  $W_{class0}$  設為 1，表示未考慮權重；若有考慮權重，則以公式 (4.8) 設定  $W_{class1}$  及  $W_{class0}$  比重值。

$$W_{class\ i} = \frac{|DB_{train}|}{2 * |\{x | x \in DB_{train} \wedge x \in C_i\}|} \quad (\text{公式 4.8})$$

其中  $DB_{train}$  表示訓練集， $C_1$  表示正樣本， $C_0$  表示負樣本， $i=0$  或  $1$ 。

本論文預測任務 1 中訓練集的正樣本數為 1984，負樣本數為 12632，在考慮權重的情況下以公式 4.8 計算  $W_{class1}$  及  $W_{class0}$  的結果如表 4.9 所示。

表 4.9 以本論文資料集筆數計算比重值

	$W_{class\ 1}$	$W_{class\ 0}$
未考慮權重	1	1
考慮權重	$\frac{1984+12632}{2*1984} = 3.68$	$\frac{1984+12632}{2*12632} = 0.57$

### 4.3.2 擴展預測模型

圖 4.2 所示之預測模型，因為要與[11]的輸入特徵一致，因此在輸入時只有考慮 $V_{chart}$ 中的病況事件。為了探討實驗室化驗結果的病況事件資訊對預測模型是否有幫助，本研究將 4.3.1 所介紹的預測模型，擴展成將實驗室化驗病況事件資訊也作為模型輸入。擴展模型和原模型的差別只在資料輸入時增加 $V_{lab}$ 中的病況事件，因此只有對輸入層與嵌入層如何融合實驗室化驗結果病況事件的輸入做說明。

#### (1)輸入資料

$V_{lab}$ 的病況事件有 1506 個，與原本 $p_t$ 相同採用一個二元編碼的 1506 維表示向量 $p_t'$ ，表示病患在時間點  $t$  發生的實驗室化驗結果病況事件。

#### (2)嵌入層

根據 $p_t'$ 中標為 1 的事件，從預先訓練好的實驗室化驗結果病況事件表示法矩陣 $W_{emb}'$ 中取出對應的向量，並取每個維度上的最大值，代表該名病患在時間點  $t$  的整合實驗室化驗病況特徵向量，以 $emb_{lab}$ 表示。

接著接合 17 個儀器偵測數據的事件表示向量跟 1 個實驗室化驗特徵向量，並由模型學習一個維度為 18 的比重向量 $\beta$ ，對 17 個儀器偵測數據病況事件表示向量與 1 個整合實驗室化驗病況特徵向量進行比重加總得到 $emb_t$ ，如公式 4.9 所示。

$$emb_t = [W_{emb} * D_{p_t} ; emb_{lab}] * \beta \quad (\text{公式 4.9})$$

其中  $D_{p_t}$  表示以  $p_t$  為對角的正方矩陣。

接下來依序輸入雙向長短期記憶層、長短期記憶層及輸出層，此三層的設計與 4.3.1 模型一樣。

## 第 5 章 實驗結果與探討

### 5.1 實驗資料集及模型參數說明

MIMIC III 資料集(Medical Information Mart for Intensive Care III)，是一個大型醫療數據庫，該資料集蒐集了 2001 年至 2012 年間以色列貝斯醫療中心(Beth Israel Deaconess Medical Center)入住重症監護室的病患醫療數據。論文[11]提出的標準資料集針對 MIMIC III 資料集，設計對病患資料的篩選準則以及填入遺失值的方式。該標準資料集中原先的資料總筆數為 21,139 筆，其中已分成訓練集、驗證集及測試集，各資料集筆數如表 5.1 所示。本論文採用這個標準資料集，經過章節 4.1.2 介紹的方法刪減異常資料後，資料筆數統計如表 5.1 所示，最後資料總數為 21,041 筆。

表 5.1 資料筆數

	訓練集	驗證集	測試集	總筆數
原始資料集	14,681	3,222	3,236	21,139
刪除異常值資料後	14,616	3,203	3,222	21,041

本論文的兩個預測任務皆為預測病患是否會發生死亡，此為一個二元分類預測任務，資料標記有兩種類別(Class)，分別為正樣本(Positive)為病患發生死亡，以及負樣本(Negative)為病患未發生死亡。兩個任務的差別是預測任務 1 的目標是預測病患在出院前死亡。預測任務 2 的目標是預測病患在接下來第 48 小時至第 49 小時內死亡。表 5-2 顯示兩個預測任務的訓練集、驗證集及測試集的正負樣本數。

表 5.2 正負樣本數

		正樣本(Positive)	負樣本(Negative)
預測任務 1	訓練集	1, 984	12, 632
	驗證集	434	2, 769
	測試集	374	2, 848
預測任務 2	訓練集	575	14, 041
	驗證集	139	3, 064
	測試集	127	3, 095

表 5.3 列出本論文在系統架構不同處理階段之參數設定值。

表 5.3 參數設定值

建立網路圖參數設定	
時間間隔	1/4/12/24/48 小時
元路徑類型	$V_{chart} \rightarrow V_{chart}$ $V_{chart} \rightarrow V_{lab}$ $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$
特徵表示法訓練模型參數設定	
特徵表示法維度 d	100
Skip-gram 前後文視窗範圍大小 (window size)	長度為 2 的元路徑： $V_{chart} \rightarrow V_{lab}$ 及 $V_{lab} \rightarrow V_{chart}$ window size 設為 1。 長度為 3 的元路徑： $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ window size 設為 1 或 2。
訓練批次大小(Batch Size)	1
訓練回合數(Epochs)	10
死亡率預測模型參數設定	
訓練批次大小(Batch Size)	8
訓練回合數(Epochs)	100

## 5.2 評估指標

本論文死亡率預測模型的評估指標採用二分類任務常用的評估指標，分別是準確率(Precision)、召回率(Recall)、F1 綜合評估指標(F1-Measure)及 AUROC。在分類時根據設定分類機率閾值為 0.5，機率大於等於 0.5 預測為發生死亡，小於 0.5 預測為存活。

根據比對測試資料預測結果與資料實際類別，可以統計出一個混淆矩陣，在混淆矩陣中共有四個象限，分別為 TP、FP、FN 及 TN，其中 TP(True Positive)表示預測為 1，且實際值為 1;FP(False Positive)表示預測為 1，但實際值為 0;FN(False Negative)表示預測為 0，但實際值為 1;TN(True Negative)表示預測為 0，且實際值為 0。在本研究的兩個預測任務來說，TP 代表預測死亡且實際死亡的病患數，FP 代表預測死亡但實際存活的病患數，FN 代表預測存活且實際死亡的病患數，TN 代表預測存活且實際存活的病患數，透過這四個數值可以計算出正樣本預測的準確率及召回率如下所述。

表 5.4 混淆矩陣範例

	實際為 1(實際死亡)	實際為 0(實際存活)
預測為 1(預測死亡)	TP =152	FP =123
預測為 0(預測存活)	FN =222	TN =2739



評估指標的計算公式如以下說明。

(一) 準確率(Precision)

$$Precision = \frac{TP}{TP + FP} \quad (\text{公式 5.1})$$

準確率代表模型預測患者發生死亡的準確比率值，計算預測為死亡的患者當中實際上發生死亡的比率有多少，如公式(5.1)。

如表 5.4 所示範例，經由上述公式 5.1，可以計算出正樣本的分類準確率為

$$\frac{152}{152+123} = 0.552。$$

(二) 召回率(Recall)

公式：

$$Recall = \frac{TP}{TP + FN} \quad (\text{公式 5.2})$$

召回率代表實際發生死亡的患者當中被預測為死亡的患者比例有多少，如公式(5.2)。

表 5.4 所示範例，經由上述公式(5.2)，可計算出正樣本的分類召回率為

$$\frac{152}{152+222} = 0.406。$$

(三) F1 綜合評估指標(F1-Measure)

$$F1\text{-Measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (\text{公式 5.3})$$

F1-Measure 是精確率與召回率的調和均值，計算方式如公式(5.3)所示。

#### (四) AUROC

此評估值代表動態改變分類門檻值，由模型分類預測的 FPR(False Positive Rate)與 TPR(True Positive Rate)為座標值，畫出 ROC 曲線，並計算曲線下的面積[32]。AUROC 值越接近 1 表示分類效果越好。

本論文會從模型訓練階段中儲存不同訓練回合數所學得的模型，以驗證集在模型上的平均損失值(Validation loss, val\_loss)做為指標，選出其中 val\_loss 最低的模型做為測試用的預測模型；採用此指標挑選模型可避免選到訓練過度而發生過度擬合(Overfitting)的模型。

## 5.3 病況事件表示法學習效果分析

為了初步評估以異質網路圖學習病況事件表示法的效果，本研究分別以特定病況事件之表示法搜尋相似的病況事件，並將各儀器測量數據對應之病況事件表示法進行階層式分群，從結果觀察其是否有病況關聯上的意義，將在以下兩個小節介紹。

### 5.3.1 分析表示法相似搜尋

本小節以特定病況事件學習到的表示法，查找與其相似度最高的前 10 名病況事件，觀察是否有一定程度上的病況關聯。學理上定義昏迷指數(GCS)在 3~8 間為嚴重的病況事件，因此本研究以“GCS=3”的病況事件查找 $V_{chart}$ 中和它最相似的前十名病況事件。由於 $V_{chart}$ 中事件的對應特徵可依特性分成：身體質量、生理監事、神經反應及生化檢查四種。其中身體質量特徵中單看體重或身高的病況事件意義不明確，且身高通常在住院一開始量過後不太會有改變，而神經反應相關特徵對應的其他病況事件也因為與“GCS=3”具相關特性而不特別被考慮，因此搜尋的病況事件範圍包括 $V_{chart}$ 中排除上述事件以外的 102 種病況事件。

表 5.5 「GCS = 3」病況事件相似度前十名病況事件結果 1

<p>病況事件：GCS = 3</p> <p>建構異質網路圖之時間間隔=24 小時</p> <p>元路徑 = <math>V_{chart} \rightarrow V_{Laboratory} \rightarrow V_{chart}</math> , window size = 1</p>				
	病況事件	數值範圍	正常範圍	解讀意義
1	Oxygen saturation	70.0 ~ 80.0	90~100	偏低
2	pH(血液)	7.14 ~ 7.35	7.4	
3	Heart Rate	140.0 ~ 175.0	60~100	過高
4	Respiratory rate	0.0 ~ 30.0	12~20	
5	Temperature	37.4 ~ 39.3	35~37	偏高
6	Heart Rate	0.0 ~ 35.0	60~100	過低
7	Fraction inspired oxygen	0.21 ~ 0.29	空氣中為 0.21	
8	Systolic blood pressure	37.5 ~ 75.0	90~140	過低
9	Fraction inspired oxygen	0.92 ~ 1.0	空氣中為 0.21	極高
10	Capillary refill rate	0	0	

表 5.5 所示是以時間間隔為 24 小時建構病況事件異質圖，且以元路徑取樣序列為  $V_{chart} \rightarrow V_{Laboratory} \rightarrow V_{chart}$ ，window size 設為 1，依學習出來的特徵表示法所找到與「GCS=3」相似度前十名的病況事件。本研究徵詢過醫療專業人

員，初步了解在臨床上當病患昏迷嚴重時，有許多數值出現偏低或偏高，甚至是過低或過高，而表 5.5 所顯示的相似病況事件符合此現象。例如一般人正常狀況下的血氧(Oxygen saturation)濃度為 97.5%以上，相似度排名第一的病況事件表示血氧濃度範圍偏低，屬於低氧症。此外，體溫(Temperature)偏高代表身體正在發炎。心率(Heart Rate)過高代表可能有嚴重感染，氧濃度(Fraction inspired oxygen)極高表示病患呼吸極度困難，需要提供集高的供氧濃度。血氧(Oxygen saturation)偏低跟血壓(Systolic blood pressure)過低亦代表具高度死亡風險。從表 5.5 的結果顯示，與‘GCS=3’病況事件相似度高的病況事件，有多項都是在臨床上高度死亡風險可能會發生的病況。

表 5.6 「GCS = 3 分」病況事件相似度前十名病況事件結果 2

<p>查找病況事件：GCS = 3 分</p> <p>建構異質網路圖之時間間隔=24 小時</p> <p>元路徑 = <math>V_{chart} \rightarrow V_{Laboratory}</math> , window size = 1</p>				
	病況事件	數值範圍	正常範圍	解讀意義
1	Systolic blood pressure	150.0 ~ 187.5	90~140	
2	<b>Temperature</b>	37.4 ~ 39.3	35~37	偏高
3	Fraction inspired oxygen	0.76 ~ 0.84	空氣中為 0.21	過高
4	<b>Oxygen saturation</b>	70.0 ~ 80.0	90~100	偏低
5	Mean blood pressure	14.0 ~ 45.6	70~105	過低
6	<b>Heart rate</b>	140.0 ~ 175.0	60~100	過高
7	<b>Respiratory rate</b>	150.0 ~ 180.0	12~20	極高
8	Glucose	33.0 ~ 229.7		
9	<b>Diastolic blood pressure</b>	225.0 ~ 262.	50~90	極高
10	Diastolic blood pressure	112.5 ~ 150.0	50~90	

表 5-6 所示是以時間間隔為 24 小時建構病況事件異質圖，且以元路徑取樣序列為  $V_{chart} \rightarrow V_{Laboratory}$ ，依學習出來的特徵表示法所找到相似度前十名的病況事件。將這些病況事件與表 5.5 的結果對照，發現兩者皆會找出血氧 (Oxygen saturation) 濃度偏低、體溫偏高 (Temperature) 以及心率 (Heart Rate) 過高的病況事件，皆為臨床上很嚴重的病況。此外，在病患血氧 (Oxygen

saturation)偏低的情況下，醫院會給予高濃度的氧氣(Fraction inspired oxygen)，希望提升其血氧，雖然氧氣濃度與表 5.5 找出的範圍不完全相同，但有類似的病況意義。

而表 5.6 所找出的呼吸率(Respiratory rate)=[150, 180]及舒張壓(Diastolic blood pressure)=[225, 262]病況事件皆屬於極端數值範圍，從原始資料集查找發生這兩個事件的次數都只有出現 1 次，所以可以視為離群病況事件。此結果初步發現以元路徑  $V_{chart} \rightarrow V_{Laboratory}$  取樣病況事件序列比起用  $V_{chart} \rightarrow V_{Laboratory} \rightarrow V_{chart}$  取樣，所學到的病況事件表示法較容易找到離群事件與其相似。

由上述兩個病況事件相似搜尋結果，可呈現以病況事件異質網路圖學習出的事件表示法，彼此間具有某種程度的病情狀況關聯資訊。此外，以元路徑  $V_{chart} \rightarrow V_{Laboratory} \rightarrow V_{chart}$  取樣的學習效果初步看來較穩定。

### 5.3.2 儀器偵測數據類型病況事件分群

此部分對 $V_{chart}$ 中的病況事件以時間間隔 24 小時建構病況異質圖，且以元路徑 $V_{chart} \rightarrow V_{Laboratory} \rightarrow V_{chart}$ 取樣事件序列，window size 設為 1。對 151 個病況事件學習出來的特徵表示法，本實驗對其進行階層式分群，再取 20 個分群的結果如表 5.7 所示，觀察是否群內的病況事件具有可解釋的病況關聯。

表 5.7 病況事件分群結果

群編號	該群中事件數量	包含病況事件
1	4	Heart rate=[280.0,315.0] Height=[96.0,120.0] Respiratory rate=[180.0,210.0] Temperature=[26.0,27.9]
2	9	Diastolic blood pressure=[150.0,187.5] Fraction inspired oxygen=[0.84,0.92] Heart rate=[0.0,35.0] Height=[192.0,216.0] Mean blood pressure=[172.0,203.6] Mean blood pressure=[235.2,266.8] Mean blood pressure=[266.8,298.4] Weight=[150.0,175.0] Weight=[200.0,225.0]
3	4	Fraction inspired oxygen=[0.6,0.68] Height=[0.0,24.0] pH=[6.51,6.72] pH=[7.77,7.98]
4	3	Respiratory rate=[150.0,180.0] Respiratory rate=[210.0,240.0] Temperature=[27.9,29.8]
5	6	Mean blood pressure=[298.4,330.0] Oxygen saturation=[10.0,20.0] Respiratory rate=[90.0,120.0] Weight=[0.0,25.0] Weight=[175.0,200.0] Weight=[225.0,250.0]



表 5.7 病況事件分群結果(續)

6	6	Glucose=[1016.5,1213.2] Heart rate=[175.0,210.0] Respiratory rate=[120.0,150.0] Temperature=[29.8,31.7] Temperature=[41.2,43.1] pH=[6.72,6.93]
7	13	Capillary refill rate=1.0 Fraction inspired oxygen=[0.68,0.76] Fraction inspired oxygen=[0.76,0.84] motor response=3 GCS total=5 verbal response=2 verbal response=3 Oxygen saturation=[40.0,50.0] Oxygen saturation=[50.0,60.0] Temperature=[31.7,33.6] Weight=[25.0,50.0] pH=[6.93,7.14] pH=[7.56,7.77]

表 5.7 病況事件分群結果(續)

8	30	<p>Capillary refill rate=0.0  Diastolic blood pressure=[237.5,75.0]  Diastolic blood pressure=[75.0,112.5]  Fraction inspired oxygen=[0.21,0.29]  <b>eye opening=2</b>  <b>eye opening=3</b>  <b>eye opening=4</b>  <b>motor response=5</b>  <b>motor response=6</b>  <b>GCS total=15</b>  verbal response=1  <b>verbal response=5</b>  Glucose=[33.0,229.7]  Heart rate=[35.0,70.0]  Heart rate=[70.0,105.0]  Heart rate=[105.0,140.0]  Height=[168.0,192.0]  Mean blood pressure=[45.6,77.2]  Mean blood pressure=[77.2,108.8]  Mean blood pressure=[108.8,140.4]  Oxygen saturation=[90.0,100.0]  Respiratory rate= [0.0,30.0]  Respiratory rate=[30.0,60.0]  Systolic blood pressure=[75.0,112.5]  Systolic blood pressure=[112.5,150.0]  Systolic blood pressure=[150.0,187.5]  Temperature=[35.5,37.4]  Temperature=[37.4,39.3]  Weight=[75.0,100.0]  pH=[7.35,7.56]</p>
9	2	<p>Height=[120.0,144.0]  <b>Respiratory rate=[240.0,270.0]</b></p>
10	2	<p><b>Glucose=[1213.2,1409.9]</b>  <b>Glucose=[1409.9,1606.6]</b></p>

表 5.7 病況事件分群結果(續)

11	34	Diastolic blood pressure=[0.0,37.5] Diastolic blood pressure=[112.5,150.0] eye opening=1 motor response=1 motor response=4 GCS total=3 GCS total=4 GCS total=6 GCS total=7 GCS total=8 GCS total=9 GCS total=10 GCS total=11 GCS total=12 GCS total=13 GCS total=14 verbal response=4 Glucose=[229.7,426.4] Heart rate=[140.0,175.0] Mean blood pressure=[14.0,45.6] Mean blood pressure=[140.4,172.0] Oxygen saturation=[60.0,70.0] Oxygen saturation=[70.0,80.0] Oxygen saturation=[80.0,90.0] Systolic blood pressure=[0.0,37.5] Systolic blood pressure=[37.5,75.0] Systolic blood pressure=[187.5,225.0] Temperature=[33.6,35.5] Temperature=[39.3,41.2] Weight=[50.0,75.0] Weight=[100.0,125.0] Weight=[125.0,150.0] pH=[6.3,6.51] pH=[7.14,7.35]
12	2	<b>Glucose=[1606.6,1803.3]</b> <b>Height=[72.0,96.0]</b>
13	2	<b>Diastolic blood pressure=[300.0,337.5]</b> <b>Systolic blood pressure=[300.0,337.5]</b>

表 5. 7 病況事件分群結果(續)

14	11	Diastolic blood pressure= [187.5,225.0] <b>motor response = 2</b> <b>Glucose= [426.4,623.1]</b> <b>Glucose= [623.1,819.8]</b> <b>Glucose= [819.8,1016.5]</b> <b>Mean blood pressure= [203.6,235.2]</b> <b>Oxygen saturation= [0.0,10.0]</b> <b>Oxygen saturation= [20.0,30.0]</b> <b>Oxygen saturation= [30.0,40.0]</b> <b>Respiratory rate= [60.0,90.0]</b> <b>Systolic blood pressure= [225.0,262.5]</b>
15	6	Fraction inspired oxygen= [0.37,0.45] Fraction inspired oxygen= [0.45,0.53] Fraction inspired oxygen= [0.53,0.6] Fraction inspired oxygen= [0.92,1.0] eye opening= 0(“Close”) Height= [144.0,168.0]
16	1	<b>Heart rate= [210.0,245.0]</b>
17	1	<b>Height= [48.0,72.0]</b>
18	1	<b>Respiratory rate= [270.0,300.0]</b>
19	3	Fraction inspired oxygen= [0.29,0.37] pH= [7.98,8.19] pH= [8.19,8.4]
20	3	Diastolic blood pressure= [225.0,262.5] Diastolic blood pressure= [262.5,300.0] Systolic blood pressure= [262.5,300.0]

由於分群效果受所採用分群演算法及分群數影響，所以並非每個分群中涵蓋的病況事件之關聯性皆可解釋，但表 5. 7 中仍顯示出一些有意義的病況事件聚集性。例如表 5. 7 中的第 8 群中「GCS\_Total=15」、「Eye opening=3、4 或 5」、「Motor response=5 或 6」及「Verbal response=5」等七個病況事件皆屬於神經反應的檢測，且這些病況事件在臨床上代表病患的昏迷狀況屬於仍有反應。同一群中還有病況事件「Oxygen saturation=[90. 0, 100. 0]」，該病況表示血氧飽和濃度正常，上述這些皆屬於症狀較不嚴重的病況事件。此外，平均血壓(Mean blood pressure)的正常範圍在 70~105，收縮壓(Systolic blood pressure)的

正常範圍在 90~140，此群中病況事件「Mean blood pressure=[45.6, 77.2] 及 [77.2, 108.8]」及「Systolic blood pressure=[75.0, 112.5] 或 [112.5, 150.0]」皆屬正常範圍，表示此群主要涵蓋病情較輕或較穩定的病況事件。第 14 群中的病況事件則大多顯示臨床上病情嚴重的狀況，例如血糖(Glucose)極高、血氧(Oxygen saturation)極低、收縮壓(Systolic blood pressure)極高及「Motor response=2」(代表病患對疼痛刺激的肢體反應僵硬)。此外，我們發現有幾個找出的群其包含的病況事件數不超過2，且這些病況事件皆對應到極端數據，在臨床上極少出現這些病況事件，包括第 10、12、13、16、17、18 群。若將這些離群事件(Outlier)先去除後再進行分群，有可能得到更有意義且病情嚴重程度更一致的病況事件分群結果。

## 5.4 預測效能評估

### 5.4.1 資料前處理效果比較

本研究針對[11]的標準資料集另進行異常值處理，並對連續數值特徵進行區間化處理，以便後續進行病況事件的表示法學習。此實驗將比較 1)以標準資料集的原始偵測值做為輸入(以 Baseline 表示)、2)以移除資料中異常值做為輸入(以 $Baseline_{w/o\ outlier}$ 表示)，以及 3)以 $Baseline_{w/o\ outlier}$ 資料再經過區間化處理轉成出現病況事件的二元化向量做為預測模型輸入(以 $Baseline_{discrete}$ 表示)，觀察進行這兩個前處理步驟後對訓練死亡率預測模型的影響效果。

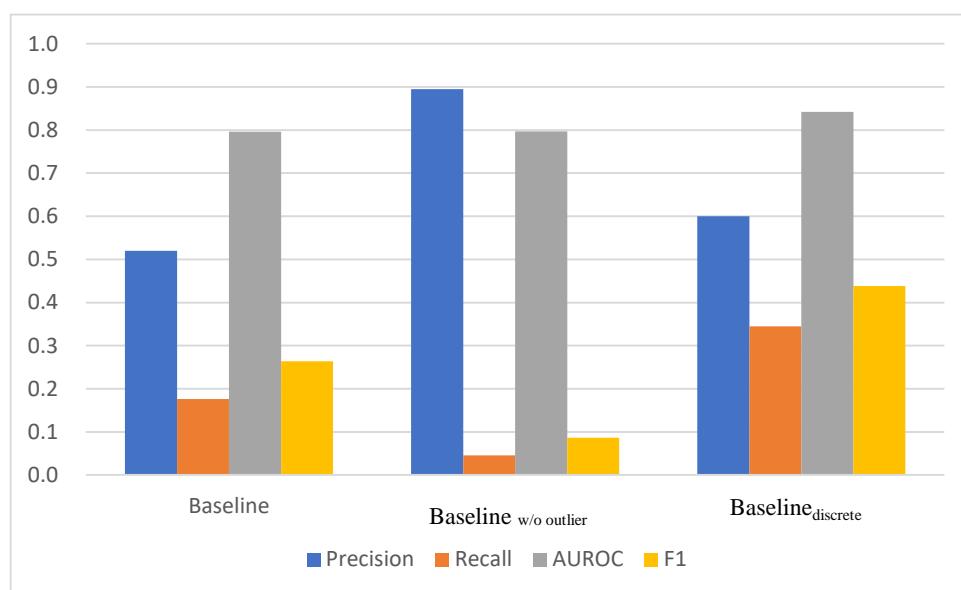


圖 5.1 前處理對預測模型的效果比較

圖 5.1 顯示資料前處理對模型預測效果影響的實驗結果，從圖 5.1 中可以看出 AUROC 平緩的上升，只移除異常值，在 Precision 有提升但 Recall 下降，F1 也跟著下降，表示去除異常值讓學習到的模型對死亡預測更嚴格，但也可能

學習到過於特定的資料。進行區間化處理後，能使模型學習到資料中較一般化的規則， $Baseline_{discrete}$ 的結果顯示較 Baseline 在 Precision、Recall、F1 及 AUROC 都有上升，上述結果可顯示本研究進行資料前處理及區間化處理的效果。

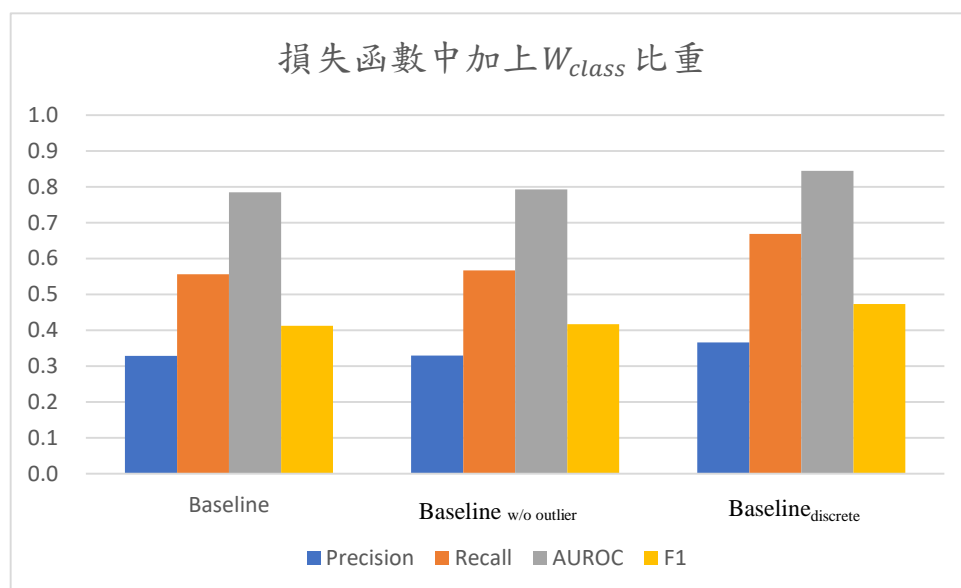


圖 5.2 加上 $W_{class}$ 比重後前處理對預測模型的效能比較

本研究在模型訓練時，考慮到資料集中正樣本與負樣本數差異很大，因此在模型中設定 $W_{class}$ 比重參數，使正負樣本在影響訓練損失的比例較平衡，以提高正樣本的學習。圖 5.2 顯示，若在損失函數加上 $W_{class}$ 比重參數，在處理異常值後，再進行區間化處理，與圖 5.1 相比，各項評估指標皆會平緩上升。雖然在 Precision 的表現較未加入比重參數前為低，但在 Recall 上有很顯著的提高，且整體的 F1 score 也提高。本研究認為以死亡預測任務來說，觀察 Recall 的效果應優先於 Precision 的效果，因此在後續實驗中，我們都採用在損失函數上加 $W_{class}$ 比重參數，使模型學習效果更穩定。

## 5.4.2 病況事件表示法之訓練資料擷取策略效果評估

本小節的實驗目的在比較採用不同時間間隔區間內的病況事件建立網路圖的邊，結合採用不同元路徑進行事件序列取樣，評估對死亡預測模型的預測效果。時間間隔分別設為 1、4、12、24、及 48 小時；三種元路徑分別為  $V_{chart} \rightarrow V_{chart}$  (以 C-C 表示)、 $V_{chart} \rightarrow V_{lab}$  (以 C-L 表示)、及  $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 。其中  $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$  元路徑又分成 window size 設為 1 (以 C-L-C 表示)，以及 window size 設為 2 (以 C-L-C+skip2 表示) 兩種訓練方式。

### (一) 院內死亡預測

在院內死亡預測的任務，採用這些不同的事件表示法訓練資料擷取策略，在任務預測上的準確率、召回率、及 AUROC 分別如圖 5.3、5.4、及 5.5 所示。由於預測發生死亡的門檻值會影響 Precision 及 Recall 的結果，因此以下討論以 AUROC 為主，Recall 的效果為輔。

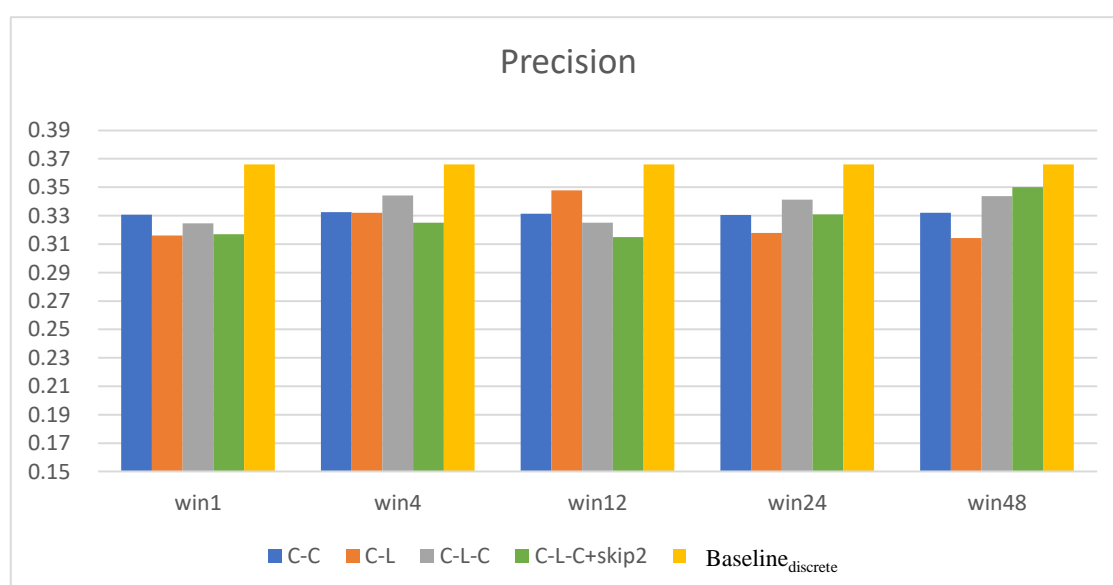


圖 5.3 表示法訓練資料擷取策略在院內死亡預測的預測準確率



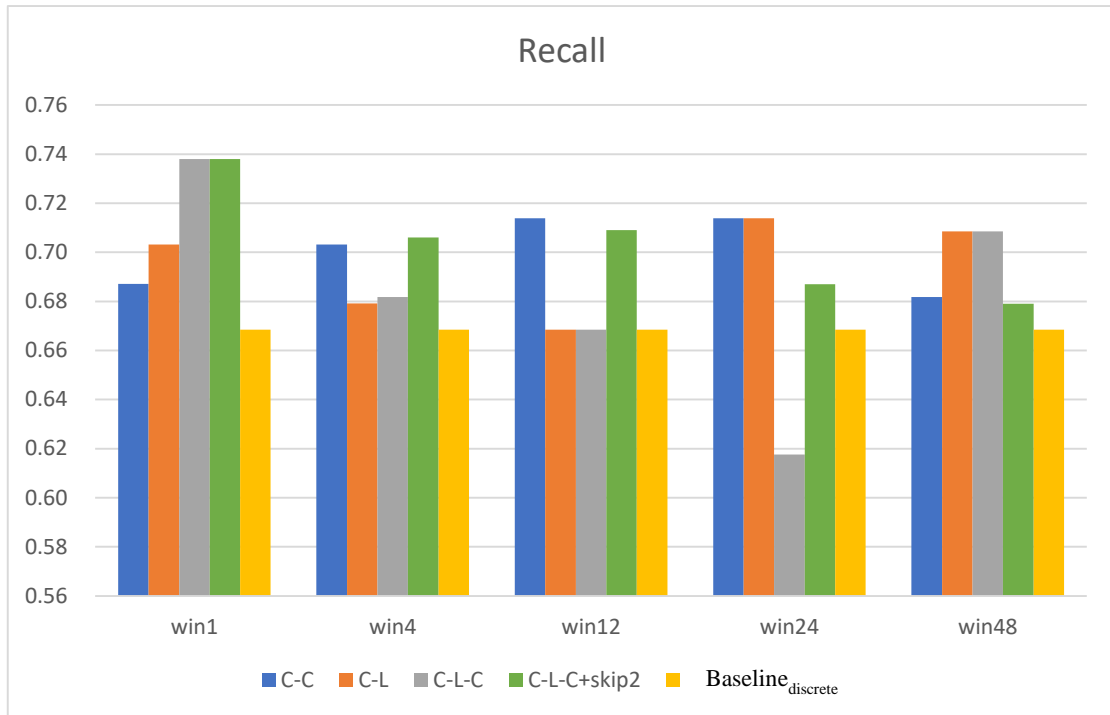


圖 5.4 表示法訓練資料擷取策略在院內死亡預測的預測召回率

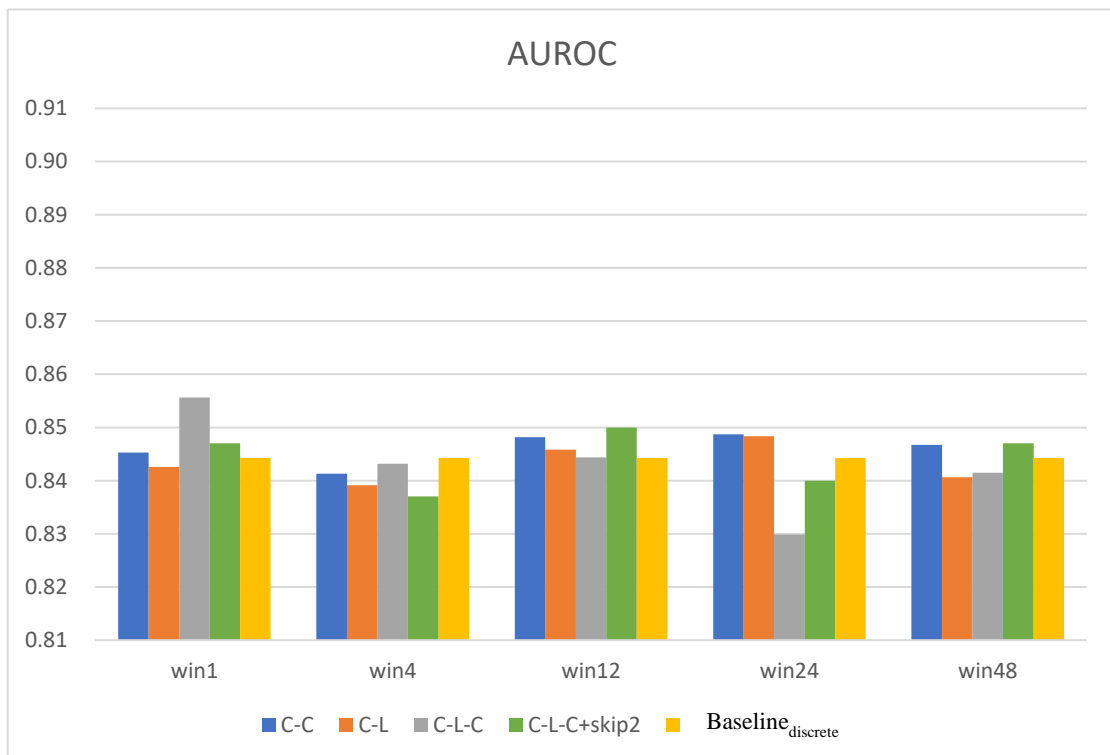


圖 5.5 表示法訓練資料擷取策略在院內死亡預測的預測 AUROC

從以上三個圖可以看出採用同質且長度為 2 的元路徑 $V_{chart} \rightarrow V_{chart}$ ，在召回率及 AUROC 的結果上，採用不同的時間間隔，結果差異分別在 0.03 及 0.01 以內，效果差距不明顯。採用異質且長度為 2 的元路徑 $V_{chart} \rightarrow V_{lab}$ ，不同時間間隔的結果差異亦分別在 0.04 及 0.01 以內。而採用元路徑 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 取樣事件序列的 C-L-C 及以 C-L-C+skip2，當配上不同時間間隔建構異質圖，所學得病況事件表示法的預測效果在召回率及 AUROC 的差異最大分別可達到 0.12 及 0.025，而 C-L-C 又比以 C-L-C+skip2 有較大差異。

以 AUROC 值來看，同質的 $V_{chart} \rightarrow V_{chart}$ 元路徑在 24 小時內，採用較長時間間隔有較好的預測效果，在 Recall 上也有一致的表現。上述結果顯示：放寬時間間隔區間到一天內，使儀器偵測數據之病況事件表示法，能利用更多事件關聯學習到更完整的病況資訊。採用異質的 $V_{chart} \rightarrow V_{lab}$ 元路徑，在 Recall 上的表現會隨時間間隔大小設定不同而有起伏，但同樣在設定成 24 小時可達到其最佳效果。而採用異質 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 元路徑的 C-L-C，因為時間間隔愈長愈可能混入不相關的事件序列，而讓預測效果不穩定，但在時間間隔設為 1 小時的情況下，則能達到最高的 AUROC(0.856)及 Recall(0.738)預測效果。C-L-C+skip2 在時間間隔設為 1 小時的情況下達到與 C-L-C 相同的最佳預測效果，但其 AUROC 略低；在時間間隔設為 12 小時的情況下達到其最佳的 AUROC(0.85)，但其對應的 Recall 降至 0.71。

上述結果初步顯示以短時間間隔(1 小時)， $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 取樣事件

序列，且設 window size 為 1 有最佳的表示法學習效果。此外，推測 C-L-C+skip2 因為可能取到跨不同病患的  $V_{chart}$  事件配對，因此學習效果不如 C-L-C。

## (二)短期(48 小時內)死亡預測

在短期(48 小時內)死亡預測的任務，這些不同的表示法訓練資料擷取策略，對任務預測上的準確率、召回率及 AUROC 分別如圖 5.6、圖 5.7 及圖 5.8 所示。

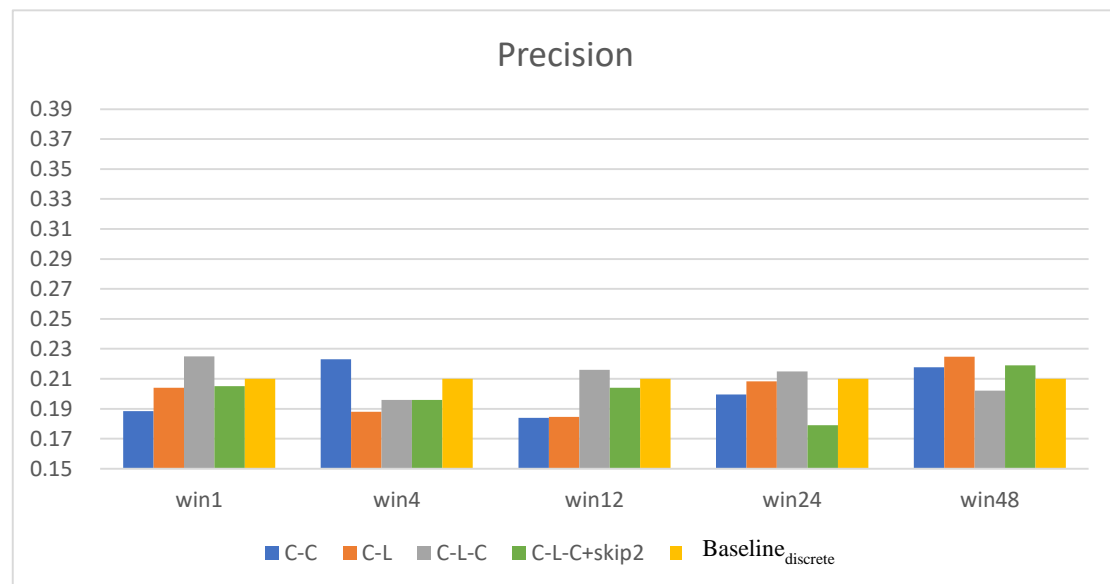


圖 5.6 表示法訓練資料擷取策略在短期死亡預測的預測準確率

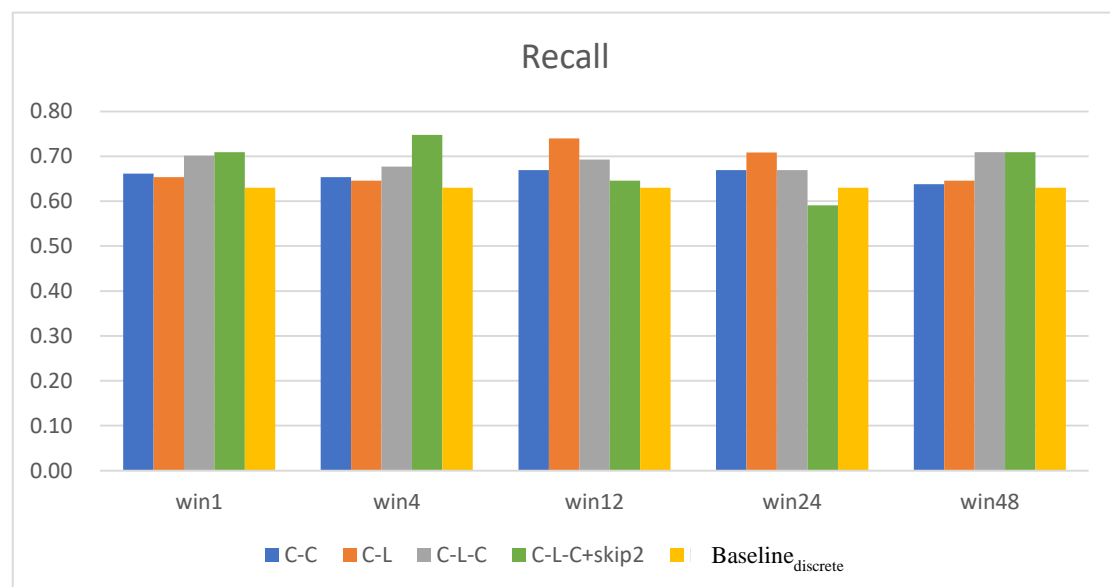


圖 5.7 表示法訓練資料擷取策略在短期死亡預測的預測召回率

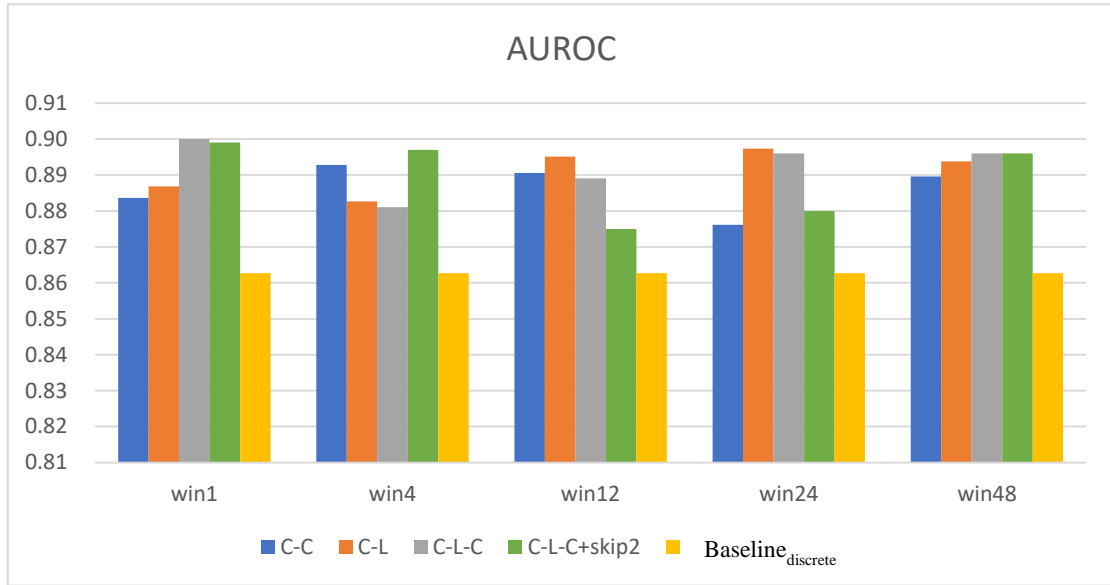


圖 5.8 表示法訓練資料擷取策略在短期死亡預測的預測 AUROC

由圖 5.7 及 5.8 顯示，在短期院內死亡預測任務，採用病況事件表示法學習皆較 $Baseline_{discrete}$ 有更佳的 AUROC 及 Recall 值。當採用元路徑 $V_{chart} \rightarrow V_{chart}$ 或 $V_{chart} \rightarrow V_{lab}$ ，結果顯示兩種元路徑結合時間間隔設為 12 小時是較好的訓練資料擷取策略。採用元路徑 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 時，則縮小時間間隔為 1 小時建構異質圖，不管在 AUROC、Recall、及 Precision 皆有最佳表現，且在 Precision 上也比 $Baseline_{discrete}$ 更高。當 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 的 window size 設為 2，其與 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 一樣在時間間隔 1 小時有最佳 AUROC 結果，在時間間隔 4 小時有最佳 recall 結果(0.748)但 AUROC 結果略低於時間間隔 1 小時的結果，此結果亦顯示 C-L-C+skip2 的學習效果較 C-L-C 不穩定。根據上述實驗結果，採用異質關係路徑 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 元路徑取樣，以時間間隔 1 小時建立異質網路圖，學習到的病況事件表示法，對短期死亡風險有最佳的預

測效果，且 recall 及 AUROC 分別可達到 0.7 及 0.9。

總結上述院內死亡預測任務及短期死亡預測任務的預測效果評估，以時間間隔 1 小時建立異質網路圖，並採用元路徑  $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$  取樣，並將 window size 設為 1，所學習到的病況事件表示法，能夠比 *Baseline<sub>discrete</sub>* 有效對院內死亡預測的召回率提高 7%，AUROC 值提高 0.01；並可對短期 48 小時內死亡預測的召回率提高 7%，AUROC 值提高約 0.04 達到 0.9，準確率亦提高 1.5%。

### 5.4.3 加入實驗室化驗結果病況事件資訊之模型預測效果

由於取樣元路徑 $V_{chart} \rightarrow V_{lab} \rightarrow V_{chart}$ 的序列中包含實驗室化驗結果事件，訓練特徵表示法時除了學習到 $V_{chart}$ 中病況事件的表示法，也學習到 $V_{lab}$ 中病況事件的表示法，於是本實驗嘗試加入 $V_{lab}$ 中的病況事件表示法到模型的輸入，評估對模型在院內死亡及短期死亡兩個預測任務的效果影響。

模型在院內死亡預測的實驗結果以 C-L-C+ $V_{lab}$  符號表示，並將 5.4.2 中院內死亡預測的 C-L-C 及  $Baseline_{discrete}$  的結果同時顯示做為比較。在院內死亡率預測的準確率、召回率、及 AUROC 分別如圖 5.9、圖 5.10、及圖 5.11 所示。

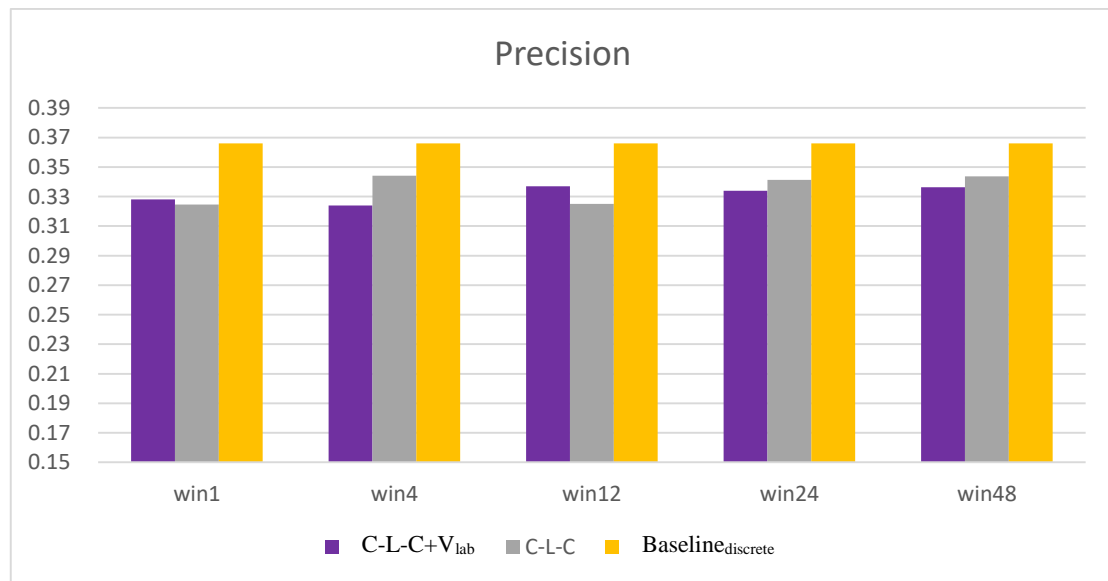


圖 5.9 加入 $V_{lab}$ 病況事件表示法在院內死亡預測之準確率

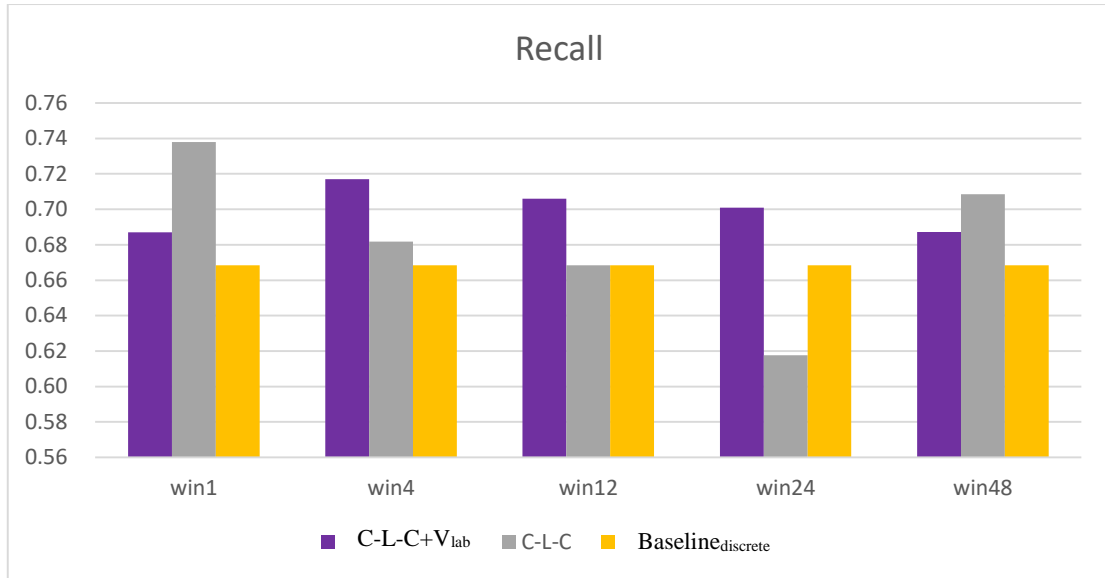


圖 5.10 加入 $V_{lab}$ 病況事件表示法在院內死亡預測之召回率

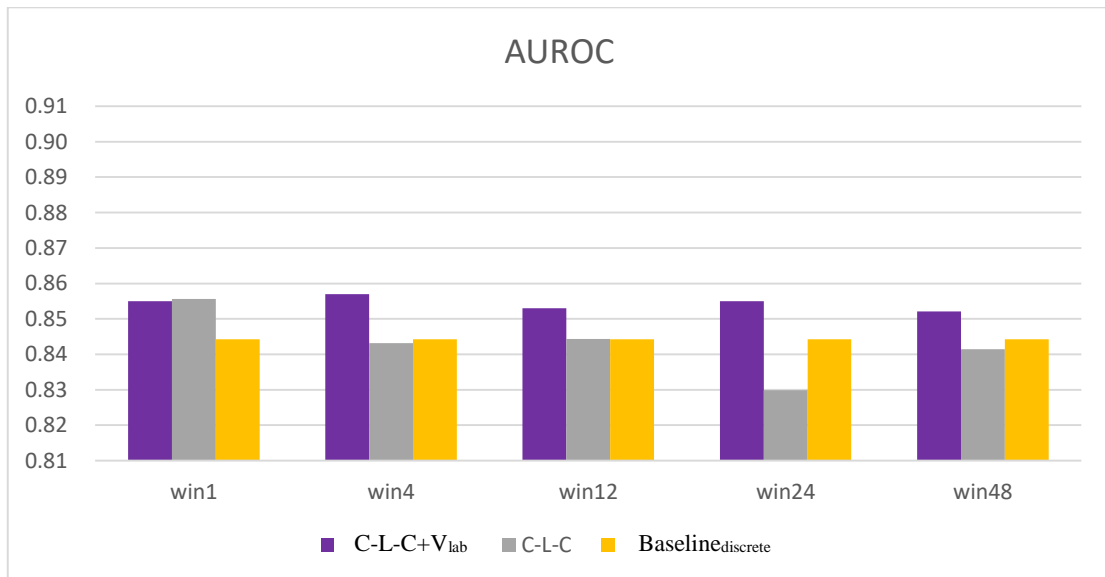


圖 5.11 加入 $V_{lab}$ 病況事件表示法在院內死亡預測之 AUROC

結果顯示加上 $V_{lab}$ 的病況事件表示法於模型輸入中，以時間間隔設為 4 小時建立網路圖，訓練出來的模型效果在召回率與 AUROC 可達到其最佳效果。和 C-L-C 結合時間間隔設為 1 小時的最佳效果相比，未加 $V_{lab}$ 事件表示法的 C-L-C 的最佳模型在 Recall 優於 C-L-C+ $V_{lab}$ 的最佳模型 2.1%，C-L-C+ $V_{lab}$ 只在 AUROC 值上較 C-L-C 略增，對於此結果，本研究認為是因為在學習病況事件表示法時已經將 $V_{lab}$ 的病況事件資訊貢獻到 $V_{chart}$ 的病況事件表示法，所以加入 $V_{lab}$ 的事



件表示法作為模型輸入特徵，對模型的增進效果有限。

短期死亡風險預測模型中加入實驗室化驗結果為輸入的實驗結果在圖中亦以  $C-L-C+V_{lab}$  符號表示，並與 5.4.2 中短期死亡預測的  $C-L-C$  及  $Baseline_{discrete}$  的結果同時顯示做為比較。短期死亡率預測的準確率、召回率、及 AUROC 分別如圖 5.12、圖 5.13、及圖 5.14 所示。

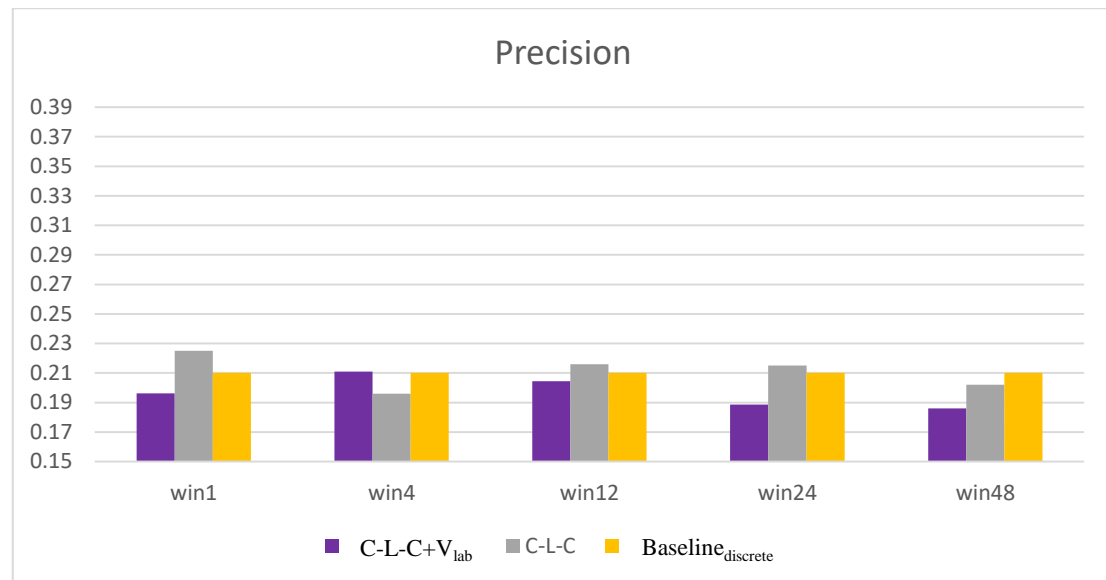


圖 5.12 加入  $V_{lab}$  病況事件表示法在短期死亡預測之準確率

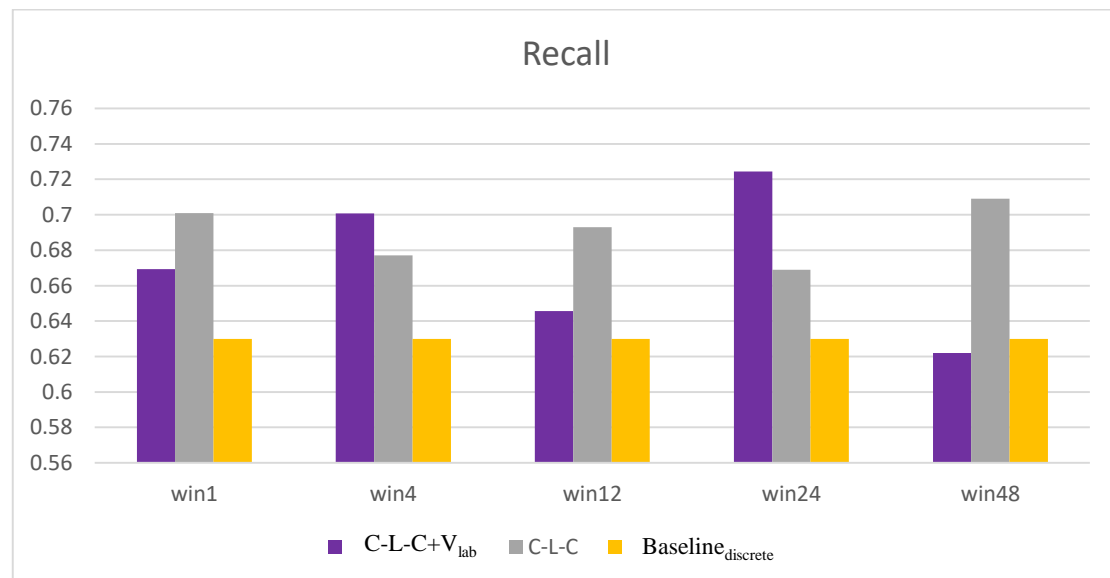


圖 5.13 加入  $V_{lab}$  病況事件表示法在短期死亡預測之召回率

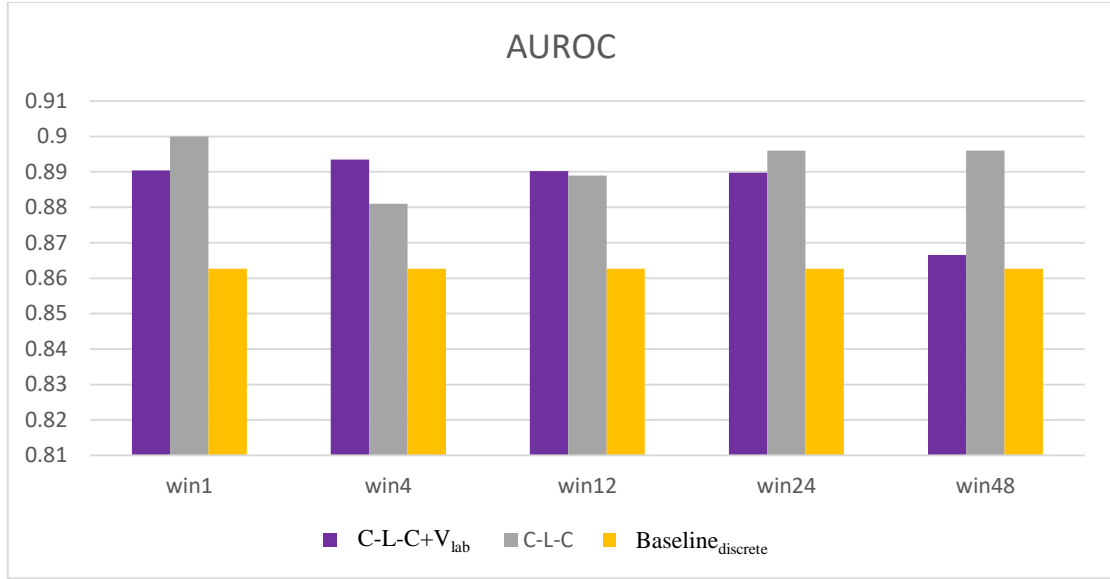


圖 5.14 加入 $V_{lab}$ 病況事件表示法在短期死亡預測之 AUROC

結果顯示加入 $V_{lab}$ 的病況事件表示法於模型輸入中，在短期死亡預測任務上，以時間間隔設為 4 小時建立網路圖，訓練出來的模型 AUROC 可達到其最佳效果，此結果與院內死亡預測任務相同，但得到次高的 Recall 值。若觀察 Recall 的結果，則在時間間隔設為 24 小時為最佳，但得到次高的 AUROC 值。與 C-L-C 結合時間間隔設為 1 小時的最佳效果相比，在 Recall 的表現上 C-L-C 與 C-L-C+ $V_{lab}$ 一樣好或略低，但 C-L-C 的最佳模型在 AUROC 及 Precision 皆高於 C-L-C+ $V_{lab}$ 的最佳模型。此結果與院內死亡預測任務的結果一致，在模型輸入時增加 $V_{lab}$ 的病況事件表示法，並無法再提高模型的預測效果，表示採用 $V_{chart}$ 的病況事件表示法輸入建構預測模型已足夠。

## 第 6 章 結論與未來研究方向

本論文探討如何運用不同類型病況資料的出現關聯，對儀器偵測數據進行病況資料的特徵表示法學習，再輸入以 LSTM 單元組合成的預測模型進行病患的死亡預測任務。本研究提出對原始資料進行異常值去除及連續型數值區間化之前處理，將原始資料值轉換成病況事件；此外，本研究在建立網路圖時改變不同病況事件間發生關聯的時間間隔門檻值設定，並實驗以異質網路圖上不同的元路徑進行序列採樣，評估以何種方式產生學習病況事件表示法的訓練資料能增進後續死亡預測任務的預測效果。

實驗結果顯示，以病況事件異質網路圖所學習到的病況事件表示法，其分群結果可顯示出某些反應病情嚴重或輕微狀況下之病況事件有出現聚集性。在用於院內死亡預測及短期死亡預測模型時，採用學習過的病況事件表示法為輸入，皆比採用原資料特徵值輸入，或只轉換成所發生病況事件作為輸入，得到更佳的預測效果。此外，本研究發現以短時間間隔(1 小時)下所建立出來的異質網路圖，採用兩段異質路徑(儀器偵測數據-實驗室化驗資料-儀器偵測數據)的元路徑取樣出的事件序列，所訓練出的儀器偵測數據病況事件表示法有最佳的模型預測效果。本論文的研究結果初步顯示以電子病歷資料中異質病況資料的出現關聯，對資料特徵表示法學習的確有所幫助，可以擴充以原本病況偵測值能表達的醫療語意資訊。

在醫療病歷資料中，醫療代碼表示醫生對病患症狀及檢查的臨床診斷，具有可明確對當次病況總結的資訊，因此未來可以考慮擴展預先訓練的異質病況事件表示法中包括醫療代碼，實驗是否能進一步提昇醫療預測任務模型的準確率，幫助患者獲得更好的醫療照護。

## 參考文獻

- [1] Choi, Edward, et al. "Multi-layer representation learning for medical concepts." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
- [2] Choi, Edward, et al. "GRAM: graph-based attention model for healthcare representation learning." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.
- [3] Choi, Edward, et al. "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism." *Advances in Neural Information Processing Systems*. 2016.
- [4] Ma, Fenglong, et al. "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks." *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2017.
- [5] Ma, Fenglong, et al. "Kame: Knowledge-based attention model for diagnosis prediction in healthcare." *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018.
- [6] Hosseini, Anahita, et al. "HeteroMed: Heterogeneous Information Network for Medical Diagnosis." *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018.
- [7] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*(2013).
- [8] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.
- [9] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

- [10] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016.
- [11] Harutyunyan, Hrayr, et al. "Multitask learning and benchmarking with clinical time series data." *arXiv preprint arXiv:1703.07771* (2017).
- [12] Dong, Yuxiao, Nitesh V. Chawla, and Ananthram Swami. "metapath2vec: Scalable representation learning for heterogeneous networks." *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017.
- [13] Sun, Yizhou, and Jiawei Han. "Mining heterogeneous information networks: a structural analysis approach." *Acm Sigkdd Explorations Newsletter* 14.2 (2013): 20-28.
- [14] Sun, Yizhou, and Jiawei Han. "Mining heterogeneous information networks: principles and methodologies." *Synthesis Lectures on Data Mining and Knowledge Discovery* 3.2 (2012): 1-159.
- [15] Wang, Jizhe, et al. "Billion-scale commodity embedding for e-commerce recommendation in alibaba." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.
- [16] Rong, Xin. "word2vec parameter learning explained." *arXiv preprint arXiv:1411.2738* (2014).
- [17] Chaudhry, Basit, et al. "Systematic review: impact of health information technology on quality, efficiency, and costs of medical care." *Annals of internal medicine* 144.10 (2006): 742-752.
- [18] Black, Ashly D., et al. "The impact of eHealth on the quality and safety of health care: a systematic overview." *PLoS medicine* 8.1 (2011): e1000387.
- [19] Goldzweig, Caroline Lubick, et al. "Costs And Benefits Of Health Information Technology: New Trends From The Literature: Since 2005, patient-focused applications have proliferated, but data on their costs and benefits remain sparse." *Health affairs* 28.Suppl2 (2009): w282-w293.

- [20] Jha, Ashish K., et al. "Use of electronic health records in US hospitals." *New England Journal of Medicine* 360.16 (2009): 1628-1638.
- [21] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11 (1997): 2673-2681.
- [22] Dybowski, Richard, et al. "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm." *The Lancet* 347.9009 (1996): 1146-1150.
- [23] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Critical care medicine*. 1985;13(10):818–829.
- [24] Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Jama*. 1993;270(24):2957–2963.
- [25] Vincent JL, De Mendonça A, Cantraine F, Moreno R, Takala J, Suter PM, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Critical care medicine*. 1998;26(11):1793–1800.
- [26] KRISHNAN, Gokul S.; KAMATH, S. Sowmya. A Supervised learning approach for ICU mortality prediction based on unstructured electrocardiogram text reports. In: *International Conference on Applications of Natural Language to Information Systems*. Springer, Cham, 2018. p. 126-134.
- [27] HUANG, Gao, et al. Trends in extreme learning machines: A review. *Neural Networks*, 2015, 61: 32-48.
- [28] LEE, Christine K., et al. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*, 2018, 129.4: 649-662.
- [29] CHOI, Edward, et al. Doctor ai: Predicting clinical events via recurrent neural networks. In: *Machine learning for healthcare conference*. PMLR, 2016. p. 301-318.

- [30] SONG, Huan, et al. Attend and diagnose: Clinical time series analysis using attention models. In: *Thirty-second AAAI conference on artificial intelligence*. 2018.
- [31] GUPTA, Priyanka, et al. Transfer learning for clinical time series analysis using deep neural networks. *Journal of Healthcare Informatics Research*, 2020, 4.2: 112-137.
- [32] <https://zh.wikipedia.org/wiki/ROC%E6%9B%B2%E7%BA%BF>



# 附錄

A-1

特徵名稱	特徵值	節點編碼
Capillary refill rate	0.0	0
Capillary refill rate	1.0	1
Eye opening	"Close"	22
Eye opening	"No Response"	23
Eye opening	"To Pain"	24
Eye opening	"To speech"	25
Eye opening	"Spontaneously"	26
Motor response	"No response"	27
Motor response	"Abnormal extensn"	28
Motor response	"Abnormal flexion"	29
Motor response	"Flex-withdraws"	30
Motor response	"Localizes Pain"	31
Motor response	"Obeyes Commands"	32
GCS total	3	33
GCS total	4	34
GCS total	5	35
GCS total	6	36
GCS total	7	37
GCS total	8	38

GCS total	9	39
GCS total	10	40
GCS total	11	41
GCS total	12	42
GCS total	13	43
GCS total	14	44
GCS total	15	45
verbal response	"No Response"	46
verbal response	"Incomprehensible sounds"	47
verbal response	"Inappropriate Words"	48
verbal response	"Confused"	49
verbal response	"Oriented"	50

中文名稱	英文名稱	區間	對應到的節點編號
體重	Weight	0.0 ~ 25.0	131~140
		25.0 ~ 50.0	
		50.0 ~ 75.0	
		75.0 ~ 100.0	
		100.0 ~ 125.0	
		125.0 ~ 150.0	
		150.0 ~ 175.0	
		175.0 ~ 200.0	
		200.0 ~ 225.0	
		225.0 ~ 250.0	
身高	Height	0.0 ~ 24.0	71~80
		24.0 ~ 48.0	
		48.0 ~ 72.0	
		72.0 ~ 96.0	
		96.0 ~ 120.0	
		120.0 ~ 144.0	
		144.0 ~ 168.0	
		168.0 ~ 192.0	
		192.0 ~ 216.0	
		216.0 ~ 240.0	
體溫	Temperature	26.0 ~ 27.9	121~130
		27.9 ~ 29.8	
		29.8 ~ 31.7	
		31.7 ~ 33.6	
		33.6 ~ 35.5	
		35.5 ~ 37.4	
		37.4 ~ 39.3	
		39.3 ~ 41.2	
		41.2 ~ 43.1	
		43.1 ~ 45.0	
收縮壓	Systolic blood pressure	0.0 ~ 37.5	111~120
		37.5 ~ 75.0	
		75.0 ~ 112.5	
		112.5 ~ 150.0	
		150.0 ~ 187.5	
		187.5 ~ 225.0	
		225.0 ~ 262.5	
		262.5 ~ 300.0	
		300.0 ~ 337.5	
		337.5 ~ 375.0	
舒張壓	Diastolic blood pressure	0.0 ~ 37.5	2~11
		37.5 ~ 75.0	

		75.0 ~ 112.5	
		112.5 ~ 150.0	
		150.0 ~ 187.5	
		187.5 ~ 225.0	
		225.0 ~ 262.5	
		262.5 ~ 300.0	
		300.0 ~ 337.5	
		337.5 ~ 375.0	
平均血壓	Mean blood pressure	14.0 ~ 45.6	81~90
		45.6 ~ 77.2	
		77.2 ~ 108.8	
		108.8 ~ 140.4	
		140.4 ~ 172.0	
		172.0 ~ 203.6	
		203.6 ~ 235.2	
		235.2 ~ 266.8	
		266.8 ~ 298.4	
		298.4 ~ 330.0	
心率	Heart Rate	0.0 ~ 35.0	61~70
		35.0 ~ 70.0	
		70.0 ~ 105.0	
		105.0 ~ 140.0	
		140.0 ~ 175.0	
		175.0 ~ 210.0	
		210.0 ~ 245.0	
		245.0 ~ 280.0	
		280.0 ~ 315.0	
		315.0 ~ 350.0	
呼吸頻率	Respiratory rate	0.0 ~ 30.0	101~110
		30.0 ~ 60.0	
		60.0 ~ 90.0	
		90.0 ~ 120.0	
		120.0 ~ 150.0	
		150.0 ~ 180.0	
		180.0 ~ 210.0	
		210.0 ~ 240.0	
		240.0 ~ 270.0	
		270.0 ~ 300.0	
血氧飽和度	Oxygen saturation	0.0 ~ 10.0	91~100
		10.0 ~ 20.0	
		20.0 ~ 30.0	
		30.0 ~ 40.0	
		40.0 ~ 50.0	
		50.0 ~ 60.0	
		60.0 ~ 70.0	

		70.0 ~ 80.0	
		80.0 ~ 90.0	
		90.0 ~ 100.0	
吸入氧氣分率	Fraction inspired oxygen	0.21 ~ 0.29	12~21
		0.29 ~ 0.37	
		0.37 ~ 0.45	
		0.45 ~ 0.53	
		0.53 ~ 0.6	
		0.6 ~ 0.68	
		0.68 ~ 0.76	
		0.76 ~ 0.84	
		0.84 ~ 0.92	
		0.92 ~ 1.0	
血糖	Glucose	33.0 ~ 229.7	51~60
		229.7 ~ 426.4	
		426.4 ~ 623.1	
		623.1 ~ 819.8	
		819.8 ~ 1016.5	
		1016.5 ~ 1213.2	
		1213.2 ~ 1409.9	
		1409.9 ~ 1606.6	
		1606.6 ~ 1803.3	
		1803.3 ~ 2000.0	
血液酸鹼值	pH	6.3 ~ 6.51	141~150
		6.51 ~ 6.72	
		6.72 ~ 6.93	
		6.93 ~ 7.14	
		7.14 ~ 7.35	
		7.35 ~ 7.56	
		7.56 ~ 7.77	
		7.77 ~ 7.98	
		7.98 ~ 8.19	
		8.19 ~ 8.4	