智慧化企業整合

心律不整分類問題

108034553 尤子維

目錄

摘要

全世界有數百萬人有心律不整的問題。根據歐洲及北美於西元2014年統計，心房顫動患者占其人口約2%至3%。因「心房顫動」及「心房撲動」死亡人數從1990年的約29,000人攀升到2013年的約112,000人。全世界因心律不整而導致突發性心跳停止的案例占了所有心血管疾病相關患者的一半死亡原因、以及全球死亡原因的15%。因此藉由神經網路模型分析，由心電描記術(ECG)所紀錄心臟跳動的訊號，及早預測出病人的病狀並進行分類，就可以針對病狀及早治療。

關鍵字:心電描記術(ECG)、神經網路模型

1. 背景

1.1. 研究背景

心律不整是指心臟的跳動速率過慢、過快或是不規則跳動所引起的疾病，從心跳指令的出發點「節律點」到心肌之間只要出現任何異常狀況，都有可能導致心律不整。而節律點是一塊特化的心臟組織，它就像是心臟內的最高指揮中心，心臟跳動的訊號就是由此發佈，跳動訊號出發後會經過特定的路線到達心肌，期間只要出現任何問題，就有可能導致心律不整，這樣的異常可能是訊號減弱甚至消失，也可能是傳遞的時間延遲或加快，如果訊號傳遞到不該傳遞的地方，仍然算是異常的一種表現形式。

心電描記術(ECG) 是一種經胸腔的以時間為單位記錄心臟的電生理活動，並通過皮膚上的電極捕捉並記錄下來的診療技術。ECG是測量和診斷異常心臟節律最好的方法,其是診斷心電傳導組織受損時心臟的節律異常以及由於電解質平衡失調引起的心臟節律的改變。

1.2. 研究動機

隨著現代生活水平的提高，生活節奏加快，人們的壓力也越來越大，產生壓力的方面也越來越多了。房貸，車貸，結婚生子，工作壓力，教育問題等等。過度的壓力會對身心造成很大的負擔，而壓力正是導致心律不整的其中原因，其他還有像是飲食，新陳代謝等等。往往病人在求診時，都是透過醫生檢查心電圖，除了要記錄訊號外，還要進行分析，一來一往下就會花費很多時間，因此若是能利用神經網路建構模型，使其能夠快速分類出是哪一種病狀，讓病人獲得較完善的治療。

1.3. 研究目的

本研究利用 5W1H 的方法，先了解題及確認問題解決方式後，再使用類神經網路模型針對文本進行分類。因此本研究目的如下

(1) 提出一種深經網路模型，使其能快速且精準的預測病人具有那些病狀

(2) 降低病狀的診斷時間，及減輕醫生的壓力

1.4. 5W1H 分析法

本研究透過 5W1H 分析法，幫助我們了解問題並思考解決問題的方法，對選定的項目、工序或操作，都要從對象(何事 WHAT)、時間(何時 WHEN)、人員(何人 WHO)、地點(何地 WHERE)、原因(何因 WHY)、方法(何法 HOW)等六個方面提出問題進行思考，且對問題進行綜合分析研究，從而得到更具建設性設的決策。

利用此方法，我們主要想針對辨認心臟訊號的病徵來進行改善。

(1) What?: 要解決甚麼問題?
解決單獨由醫生判斷錯誤，可能準確率較低的問題。

(2) When?: 在甚麼時候要辨別病徵
當病人至醫院做心電圖檢測時

(3) Who?: 由誰來改善這些問題?
醫院的研究部門提出此模型，看診醫生同時採用此模型，增加判斷準確率。

(4) Where?: 在哪種地方可以進行問題改善?
醫院的心臟科

(5) Why?: 為什麼要進行改善此問題?
避免錯誤診斷導致病人看診時間延誤、提高醫院心臟科的名聲、給予病人更好的治療。

(6) How?: 如何解決此問題?
建構神經網路模型，協助醫生進行病人心臟訊號的病徵判斷。

2. 文獻回顧

2.1. 卷積神經網路(Convolutional Neural Network, CNN)

卷積神經網路是一種前饋式的神經網路，其發展最早由 Hubel 和 Wiesel 在研究貓的視覺皮層時發現了特殊的神經網路架構能夠降低神經網路的複雜性，後續進而提出了卷積神經網路。直到 1989 年 Yann LeCun 所提出了 LeNet-5，其為第一個卷積神經網路框架，此框架包括了卷積層、池化層以及全連接層，與現今的卷經神經網路結構相似。而卷積網路架構特別在辨認圖像及影像時，透過特徵擷取，可以得到非常好的成效。

2.2. XGboost(Extreme Gradient Boosting)

XGBoost 是 2014 年 2 月誕生的專注於梯度提升演算法的機器學習函式庫，此函式庫因其優良的學習效果以及高效的訓練速度而獲得廣泛的關注。僅在 2015 年，在 Kaggle 競賽中獲勝的 29 個演算法中，有 17 個使用了 XGBoost 庫，而作為對比，近年大熱的深度神經網路方法，這一資料則是 11 個。在 KDDCup 2015 競賽中，排名前十的隊伍全部使用了 XGBoost 庫，其不僅學習效果很好，而且速度也很快，相比梯度提升演算法在另一個常用機器學習庫 scikit-learn 中的實現，XGBoost 的效能經常有十倍以上的提升。

首先 XGBoost 全名為 Extreme Gradient Boosting，主要是基於梯度提升決策樹 (Gradient Boosted Decision Tree，GBT)，被應用於解決監督式學習的問題，監督式學習主要是藉由多個特徵的訓練資料中學習建立一個模型，並且透過模型預測目標變數的結果。其中模型以數學函數表示，透過給定 X 針對 Y 進行預測的目標函數，其中模型的參數會從資料中學習調整，同時根據預測值的不同，我們可以將問題類型分為迴歸或分類。之所以 XGBoost 可以成為機器學習的大殺器，廣泛用於數據科學競賽和工業界，是因為它有許多優點：1.使用許多策略去防止過擬合，如：正則化項、Shrinkage and Column Subsampling 等。2.目標函數優化利用了損失函數關於待求函數的二階導數 3.支持並行化，這是 XGBoost 的閃光點，雖然樹與樹之間是串行關係，但是同層級節點可並行。具體的對於某個節點，節點內選擇最佳分裂點，候選分裂點計算增益用多線程並行。訓練速度快。4.添加了對稀疏數據的處理。5.交叉驗證，early stop，當預測結果已經很好的時候可以提前停止建樹，加快訓練速度。6.支持設置樣本權重，該權重體現在一階導數 g 和二階導數 h，通過調整權重可以去更加關注一些樣本。

3. 研究架構或方法

3.1. 研究架構

本研究之研究架構，大致可分為四個步驟:(1)資料輸入、(2)資料前處理、(3)模型建構、(4)輸出結果、(5)模型訓練，最後得出分類成效最佳的模型。

3.1.1. 資料輸入

本研究透過 kaggle 線上開放式資料庫所提供之 MIT-BIH 心律不整數據集，其中樣本數為 109446 筆(訓練集 87554 筆、測試集 21892 筆)，總共分成 5 個類別['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4](N: Non-ecotic beats (normal beat)，S: Supraventricular ectopic beats，V: Ventricular ectopic beats，F: Fusion Beats，Q: Unknown Beats)。其中訓練集各類別數量分別為('N': 72471, 'S': 2223, 'V': 5788, 'F': 641, 'Q': 6431)，測試集各類別數量分別為('N': 18118, 'S': 556, 'V': 1448, 'F': 162, 'Q': 1608)。

3.1.2. 資料前處理

本研究所做之資料前處理包含下列步驟:

(1)由於給定的訓練集資料不平衡，像是類別 N 的樣本高達 72471 筆，而類別 F 僅有 641 筆，會使得類別 F 的訓練不足，因此為了平衡此問題，我們採取過採樣及欠採樣的方式，將不足 35000 筆的類別，重複抽取至 35000 筆，反之將超過 35000 筆的類別，以隨機抽取 35000 筆的方式。最後訓練集共 5 個類別 175000 筆資料。

(2)將訓練集和驗證集的類別標籤 0~4 轉為 0,1 的表達方式

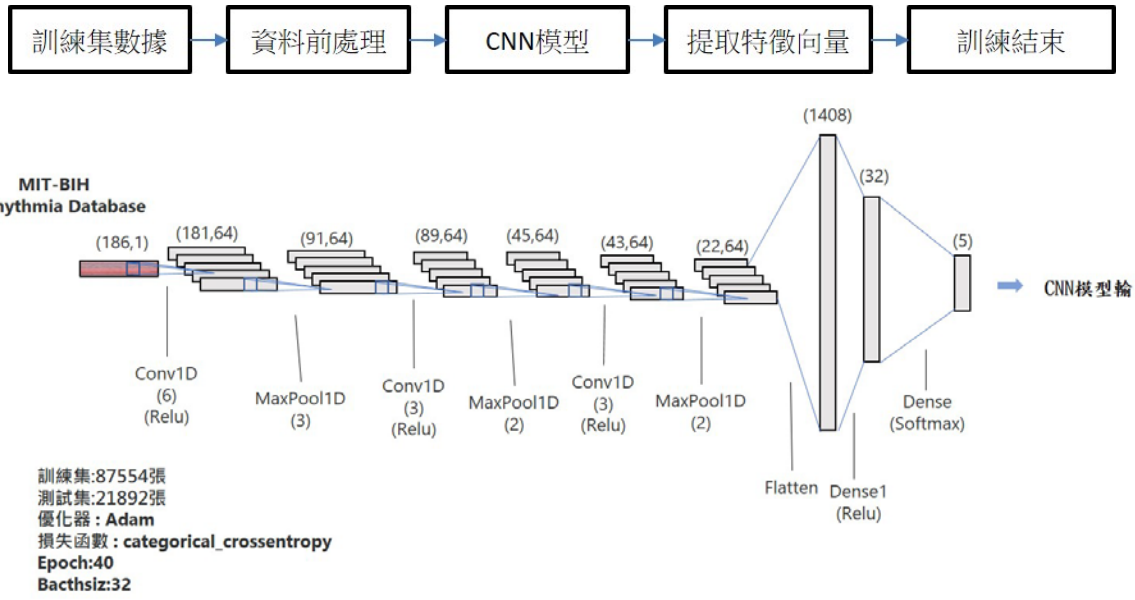
(3)將訓練集和驗證集以 reshape 的方式增加其維度，使其 fit CNN 模型的輸入

3.1.3. 模型建構

將資料做完前處理後，必須決定使用的神經網路類型以及其內部架構，由於我們投入神經網路模型的項目為數據資料，使用卷積神經網路模型能夠進行特徵的截取，幫助我們保留各種不同類別的特徵，從而進行病徵分類。

本研究為了建構適合分析心電圖訊號的模型，因此嘗試使用了卷積神經網路模型(CNN)，通常 CNN 用於處理圖像的神經網路，但也可以將其使用於訊號資料的分析，另外在建構模型的期間，發現到網路上有人嘗試使用 CNN+XGboost 的方式來改善模型的泛化能力，因此本研究將其納入分析與改善的嘗試範圍內。其初始 CNN 架構如下圖所示。

卷積神經網路(CNN)



原始模型當中，卷積神經網路的架構為:

(1)數據輸入(186,1)

(2)一層卷積核大小為 6，個數為 64 的卷積核

(3)一層池化大小為 3 的池化層

(4)一層卷積核大小為 3，個數為 64 的卷積核

(5)一層池化大小為 2 的池化層

(6)一層卷積核大小為 3，個數為 64 的卷積核

(7)一層池化大小為 2 的池化層

(8)一層平坦層將多維的輸入一維化

(9)各層激活函數:relu

(10)輸出層激活函數:softmax

(11)損失函數:categorical_crossentropy

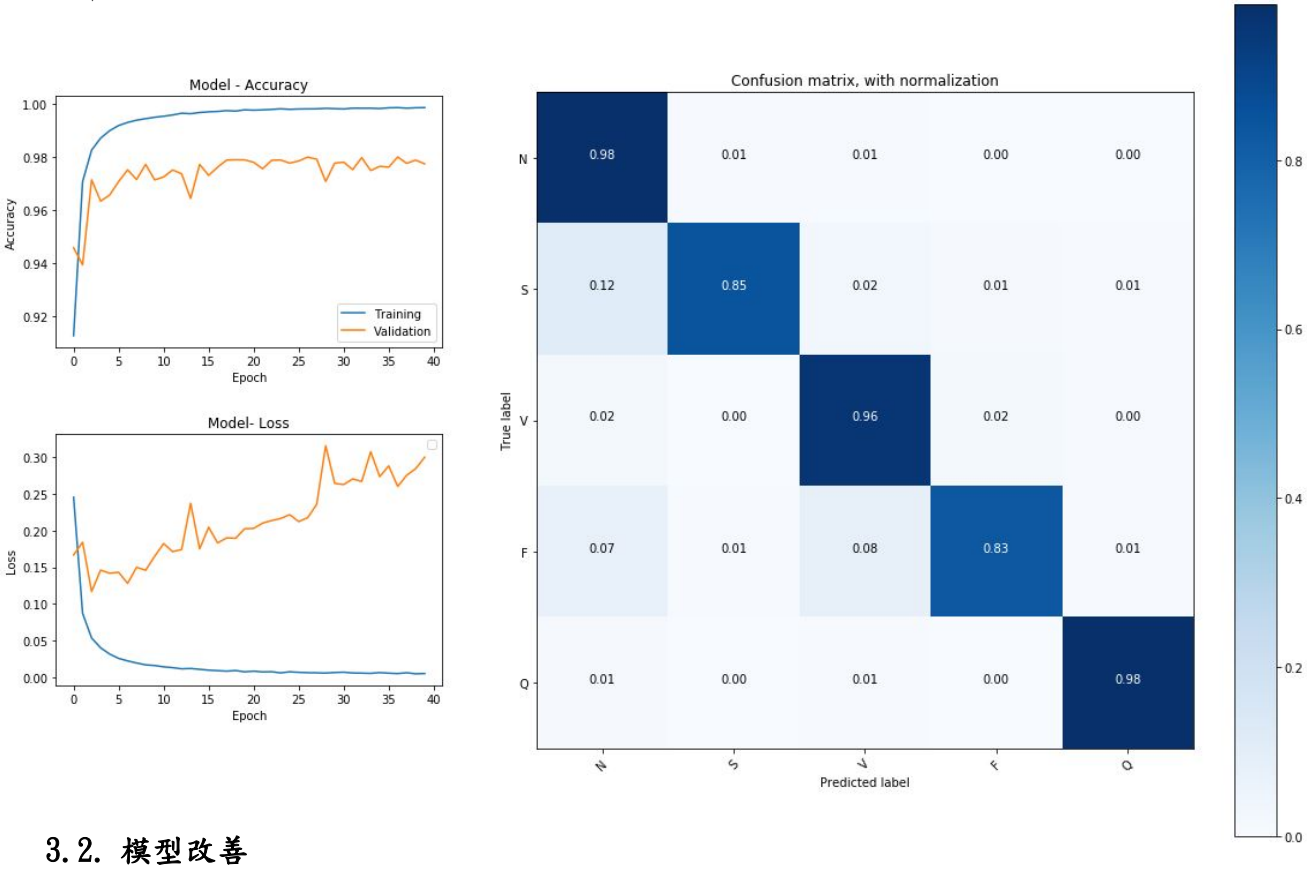
(12)Epoch:40

(13)Batchsize:50

(14)優化器為 Adam

3.1.4. 結果輸出

最後測試集準確率為 96.75%，以下圖形能更詳細的看出各類別的預測情形。可以看到左圖模型發生過擬合的情形，而右圖可以看出數量比較少的類別預測比較不準

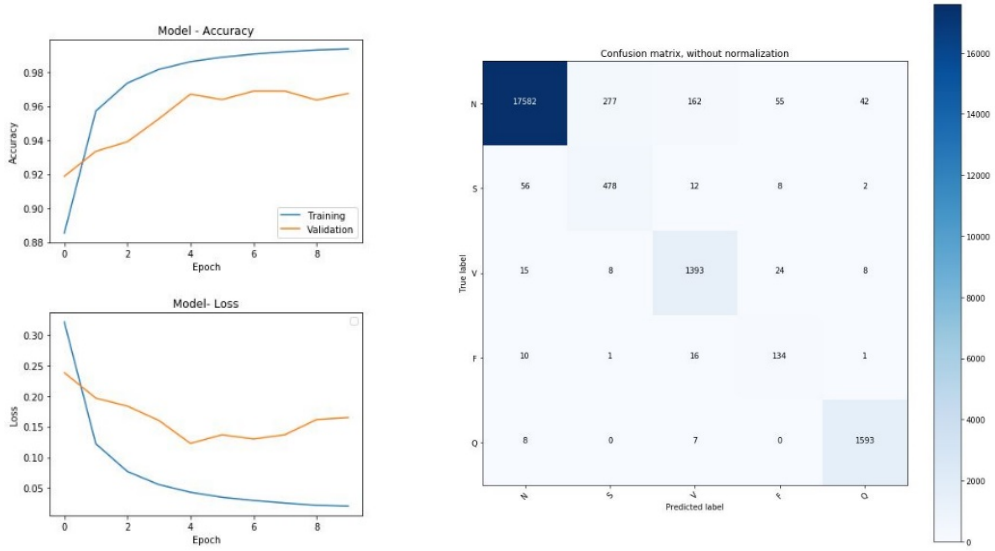


3.2. 模型改善

經由模型建構後跑出來的結果進行分析，針對覺得模型有不足的地方進改善，改善方法如下

3.2.1. 解決過擬合問題

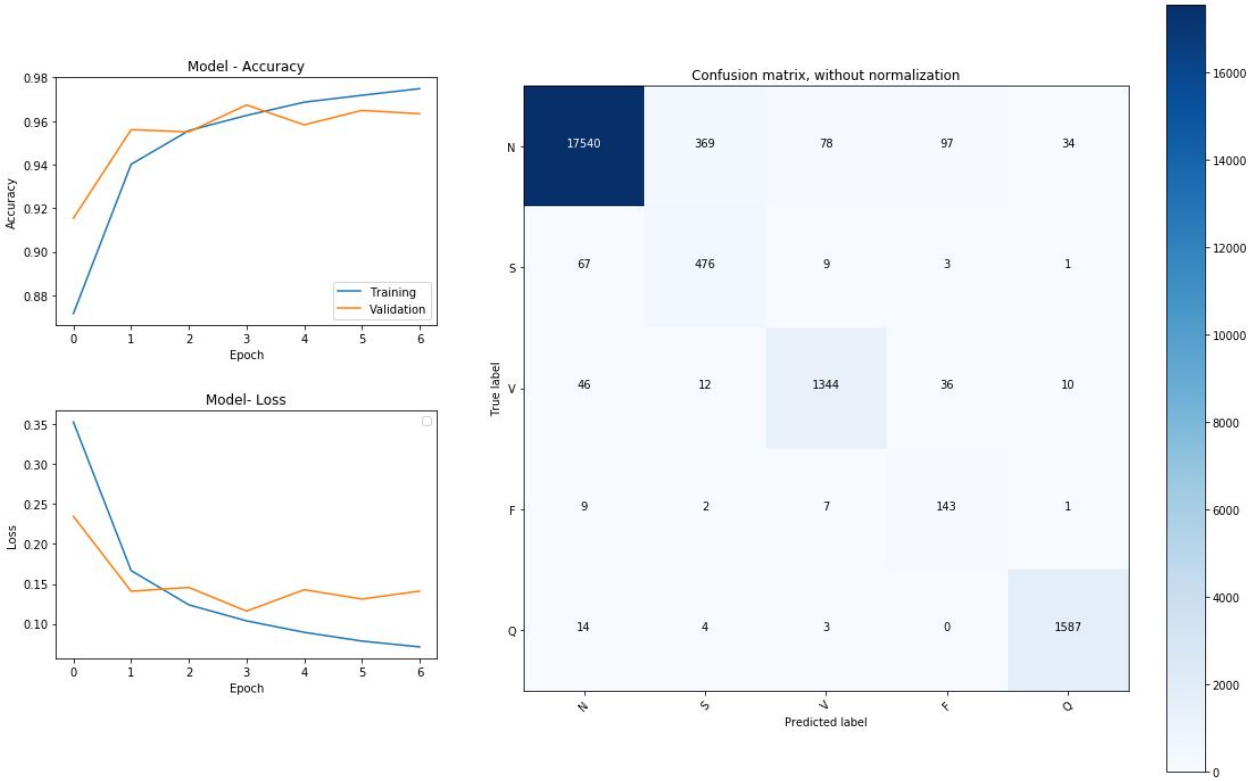
本研究認為有可能是因為數量少的類別重複抽取太多次，導致模型發生過擬合的原因，因此我們將每一個類別重複抽取從 35000 筆降至 25000 筆，另外在此階段也添加了提前終止(Early Stopping)的方法來防止模型發生過擬合的情況，其中 Patience 設定為 5，其結果如下



由上圖可以看到準確率一樣落在 96.75%，但是解決了過擬合的問題，並且降低模型運作時間，提升效能。

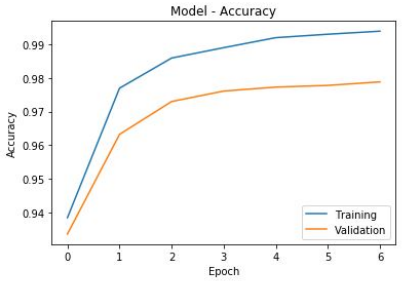
3.2.2. 增加高斯噪音

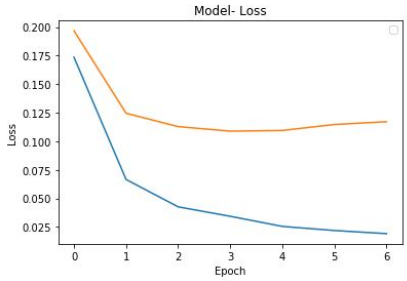
在進行資料蒐集的時候，發現增加噪音不僅可以讓數據更加真實，也可以讓模型有機會訓練得更好，故增加 noise 的嘗試，結果如下圖，可以發現到準確率下降至 96.34%外，各類別的準確率也下降了許多。

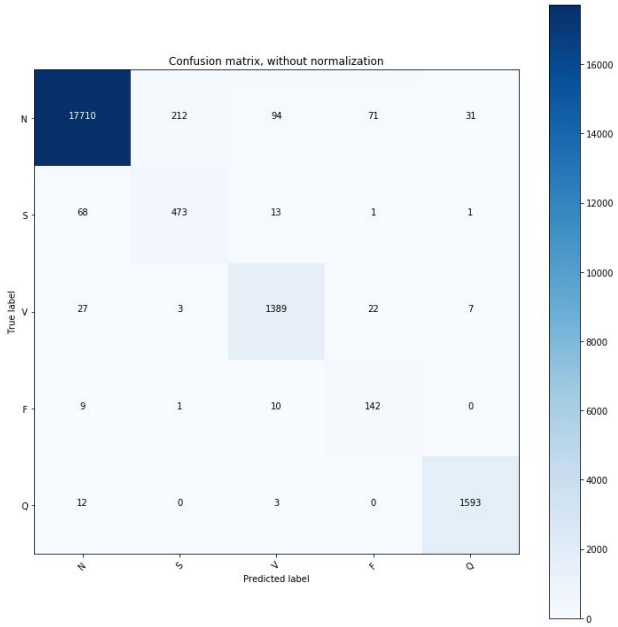


3.2.3. Batch Normalization

Batch Normalization，批標準化，和普通的數據標準化類似，是將分散的數據統一的一種做法，也是優化神經網絡的一種方法，具有統一規格的數據，能讓機器學習更容易學習到數據之中的規律。它的提出是為了克服深度神經網絡難以訓練的問題，使用 Batch Normalization 優點在於：快速學習（能增加學習率）、不會過度依賴預設值（不會對預設值產生過度反應）、控制過度學習（減少 Dropout 等必要性），我們將每一層卷積層後都加入一層的 batch normalization，其結果如下圖所示，可以看到準確率提升至 97.89%，各類別的準確率也有所提升

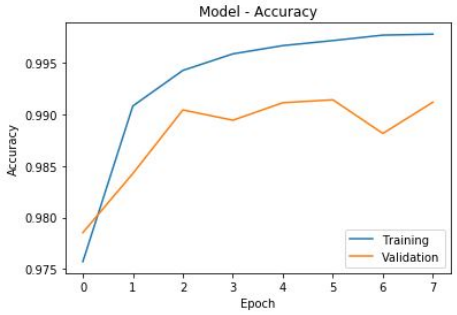


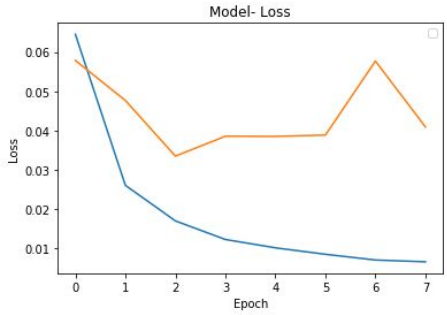


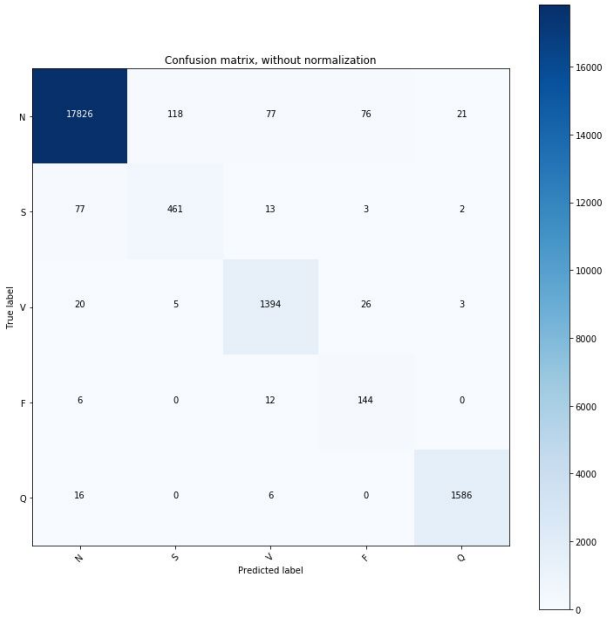


3.2.4. 其他調整

除了上述方法之外，本研究也在模型架構中增加 dense 層，以及調整 batch size 的參數，最終獲得最佳的模型，其準確度達到 99.2%，如下圖所示

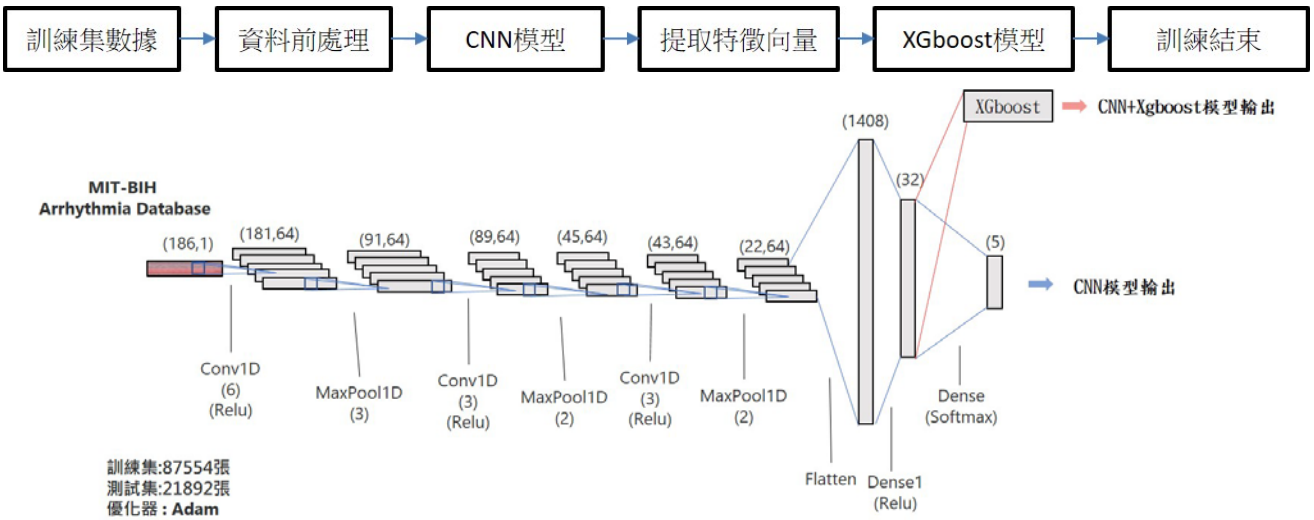


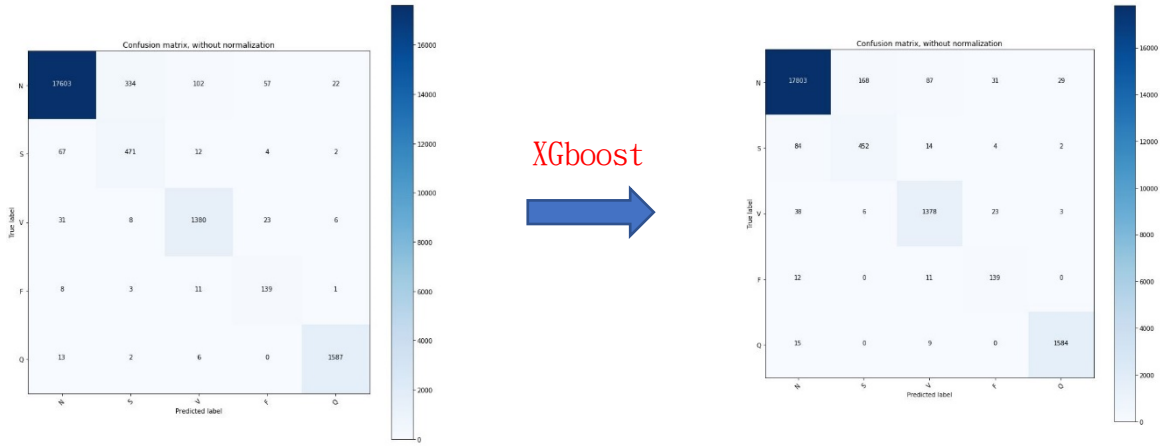


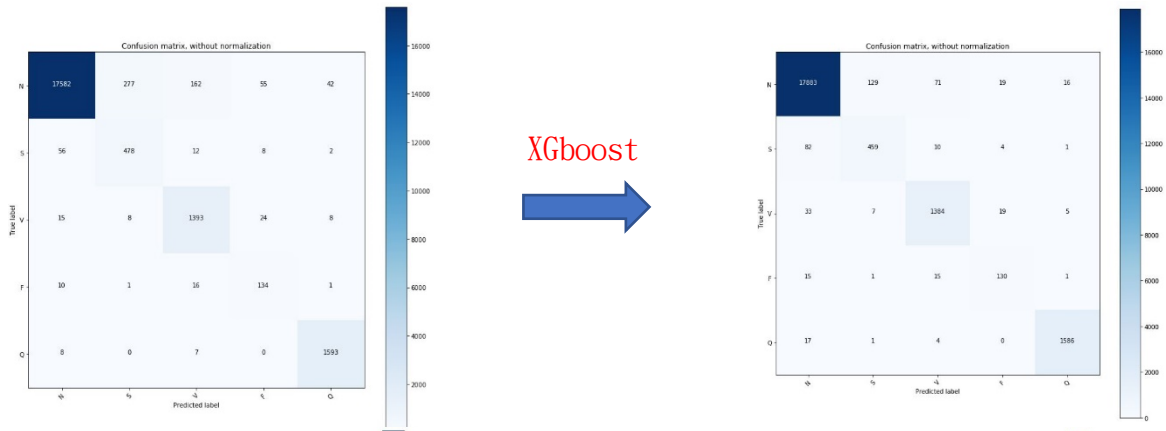


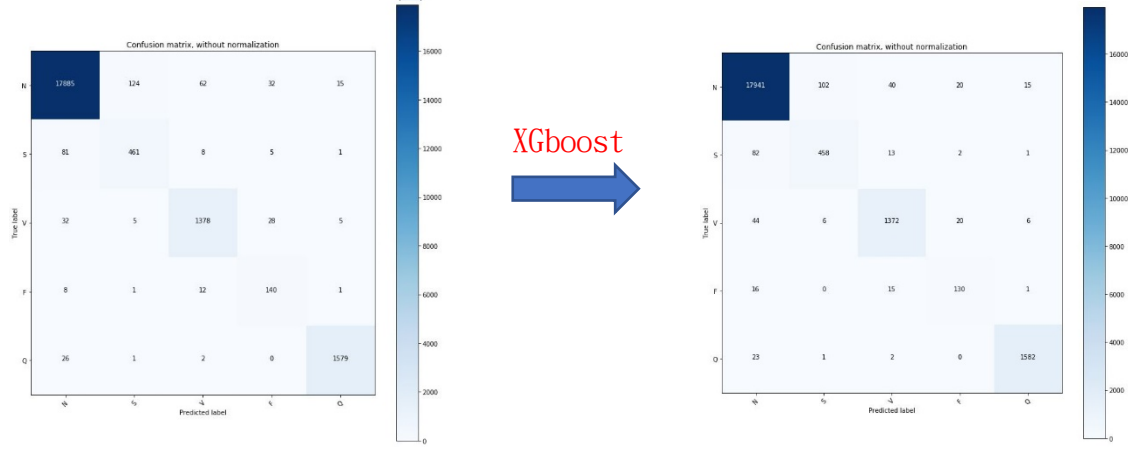
3.2.5. CNN+XGboost

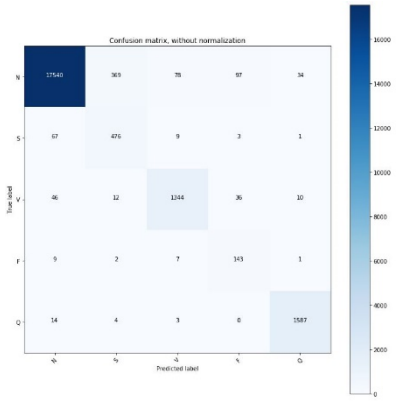
在蒐集資料的過程中，發現到網路上有研究結合 CNN 及 XGboost 的模型來提升泛化能力(其流程如下圖所示)，故本研究針對以往所做的測試，添加 XGboost 至模型中，來看 CNN+XGboost 是否能有效提升模型的泛化能力，其結果如下所示，可以發現幾乎所有模型添加 XGboost 後，類別 N 的訓練效果會變好，而其於類別的訓練效果則會變差


訓練集數據
資料前處理
CNN模型
提取特徵向量
XGboost模型
訓練結束
MIT-BIH
Arrhythmia Database
(186,1)
(181,64)
(91,64)
(89,64)
(45,64)
(43,64)
(22,64)
(1408)
(32)
(5)
XGboost
CNN+Xgboost模型輸出
CNN模型輸出
Conv1D (6) (Relu)
MaxPool1D (3)
Conv1D (3) (Relu)
MaxPool1D (2)
Conv1D (3) (Relu)
MaxPool1D (2)
Flatten
Dense1 (Relu)
Dense (Softmax)
訓練集:87554張
測試集:21892張
優化器：Adam
損失函數：categorical_crossentropy
Epoch:40
Bacthsiz:32

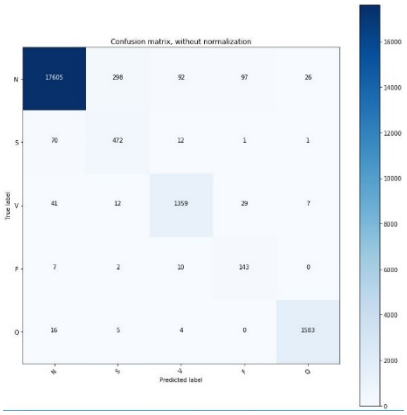

Confusion matrix, without normalization
XGboost
Confusion matrix, without normalization

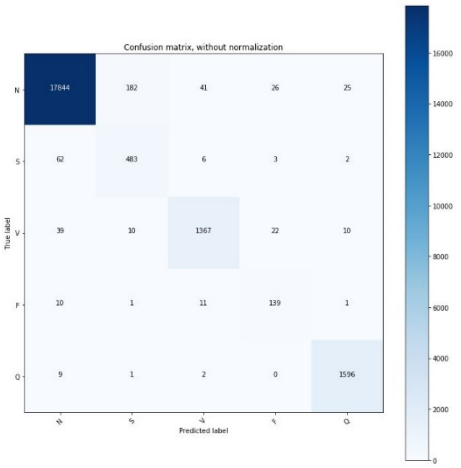

Confusion matrix, without normalization
XGboost
Confusion matrix, without normalization


Confusion matrix, without normalization
XGboost
Confusion matrix, without normalization

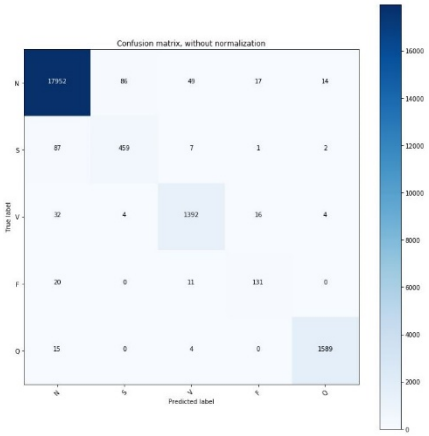


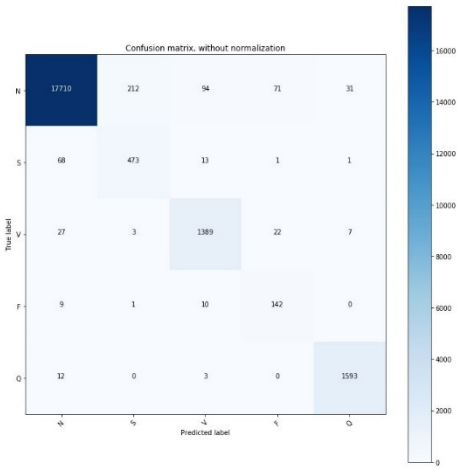
XGboost



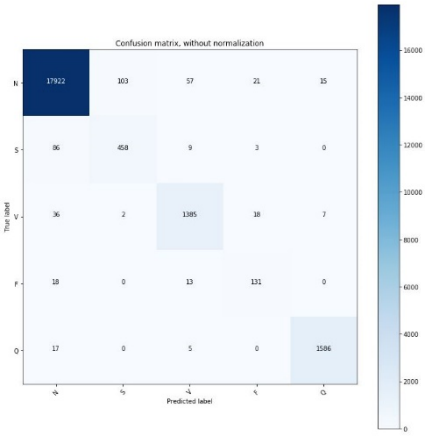


XGboost





XGboost



4. 結論

4.1. 分析與改善小結

根據本研究改善後所提出的模型，測試集準確率達 0.992，比起原先所提出的測試集 0.9665 有微幅的的提升，而看到 confusion matrix 上，比起原先模型，有小部分的改善，其中僅有類別 S 的病徵的預測能力下降。透過此模型可以協助醫生在評斷心律不整的病徵時，能夠多一個客觀的分析，讓醫生能夠快速判斷病徵，使病人及早治療。

4.2. 研究限制及後續優化

本研究透過實驗設計所訓練出的模型，雖然已經達到 99.2%的準確率，但是可以看到針對數量少的類別，即使有使用 resample 的方式減少訓練樣本不平衡，其訓練出來的模型對於數量少的類別，預測能力僅有 8 成左右，如下圖所示。

後續可以嘗試單獨使用 XGboost 建構模型，或許是因為承接 CNN 運行後的數據，導致效果沒有如預期的那麼好，其餘針對 XGboost 的參數還有很多可以嘗試調整的地方，可以使模型收斂的速度更快，預測更準。