

元智大學資管系

學術類畢業專題頂石課程(二)

期末報告

基於深度學習之顱內出血分類模型研究與應用

1111646 梁凱晰、1111656 鄭宇涵、1111708 黃詩云

工作代碼：ZD2

校內專題

指導教授：郭文嘉 博士

中華民國 114 年 11 月

Nov, 2025

Abstract

隨著人工智慧技術在醫療領域的廣泛應用，結合深度學習與醫學影像分析已成為臨床輔助診斷的重要趨勢。顱內出血(Intracranial Hemorrhage, ICH)為一種需即時辨識與處理的急性腦部疾病，常見診斷工具為電腦斷層掃描(Computed Tomography, CT)，其影像判讀結果對病患後續處置具有決定性影響。然而，由於醫療影像中的病灶細節不易辨識，且臨床資源有限，醫師在高度疲勞或壓力下可能發生誤判，進而影響診療品質。因此，本研究希望結合深度學習與多窗口影像處理方法，開發出一套可協助顱內出血多標籤分類與診斷的輔助系統。

本研究以 RSNA 公開的顱內出血 DICOM 影像資料集作為訓練與測試基礎，共使用 16,293 張訓練影像與 933 張測試影像，涵蓋硬膜外出血(epidural)、硬膜下出血(subdural)、蛛網膜下腔出血(subarachnoid)、腦實質出血(intraparenchymal)、腦室內出血(intraventricular)及任一類型出血(any)等六種標籤[1]。針對 CT 影像的特性，我們採用多窗口增強策略，將單一 DICOM 影像轉換為三通道輸入：腦窗(Brain window, WL=40, WW=80)用於觀察腦組織結構、硬膜下窗(Subdural window, WL=50, WW=130)用於突顯硬膜下與蛛網膜下出血、以及骨窗(Bone window, WL=600, WW=2800)用於檢視顱骨與鈣化病灶，藉此在單一模型輸入中整合多層次的診斷資訊。

為了充分利用 CT 影像的三維空間資訊，本研究提出位置感知深度學習架構(Position-Aware Deep Learning)，將每張切片在掃描序列中的標準化 Z 軸位置($z_normalized$ ，範圍[0,1])透過位置嵌入多層感知機(Position Embedding MLP, 維度 64)進行編碼，並與影像特徵經由融合層(Fusion MLP, 維度 256)整合後輸出六個類別的預測機率。本研究評估了多種深度學習架構作為視覺特徵提取器，包括 ResNet50 與 EfficientNet-B4。實驗結果顯示，ResNet50 在精確率方面表現更優異($precision=0.9062$)，相較 EfficientNet-B4($precision=0.8746$)提升 3.6%，能有效降低臨床上的誤報風險。此外，ResNet50 在整體加權 F1 分數上達到 0.8849，優於 EfficientNet-B4 的 0.8715，提升 1.5%。雖然 EfficientNet-B4 透過其複合縮放策略(compound scaling)與高效的網路架構設計，能在相同參數量級下提取更豐富的多尺度特徵，但考量精確率在臨床應用中對於降低假陽性診斷的重要性，以及整體效能表現，本研究最終採用 ResNet50(特徵維度 2048)作為骨幹網路，接收經多窗口處理並調整至 320×320 像素的三通道影像[2][3]。

在模型訓練策略方面，我們採用 MultilabelStratifiedKFold 進行五折交叉驗證，以確保各類別出血在每折訓練與驗證集中的分布一致性。訓練過程使用 AdamW 優化器(學習率 $1e-4$, 權重衰減 $5e-4$)搭配 ReduceLROnPlateau 學習率調整策略，並引入 MixUp 資料增強($\alpha=0.2$)與標籤平滑(label smoothing=0.05)技術以提升模型泛化能力[4]。損失函數採用 RSNA 競賽定義的加權對數損失(weighted log loss)，針對「任一類型出血」賦予雙倍權重，以符合臨床優先級。此外，我們在驗證集上進行溫度校正(temperature scaling)與類別別閾值搜尋，以優化機率預測的校準度與分類決策邊界[5]。在測試階段，採用測試時間增強(Test Time Augmentation, TTA)技術，包含水平翻轉、垂直翻轉及 ± 10 度旋轉變換，取五次預測的平均值以提升模型穩定性[6]。

實驗結果顯示，本研究提出的位置感知 ResNet50 模型在五折交叉驗證中達到平均加權 F1 分數 0.8849、各類別平均精確率 0.9062、召回率 0.8674、RSNA 加權對數損失 0.360[7]。ResNet50 在精確率上的優勢能有效降低臨床上的假陽性診斷，同時維持良好的召回率以確保漏診風險可控。模型在各類型出血的辨識上均展現穩定且優異的效能，證明了整合多窗口視覺資訊、三維空間位置特徵與深度殘差網路架構的有效性。本研究開發的輔助診斷系統不僅提供高準確度的分類結果，更透過高召回率確保了臨床安全性，具備實際應用的潛力。未來，我們期望可將此系統拓展應用於其他類型醫學影像判讀任務，並與臨床工作流程整合，實現智慧醫療輔助診斷的落地化應用。

Contents

Abstract.....	2
Contents.....	3
Chapter 1 緒論.....	4
Chapter 2 相關技術與研究.....	5
2.1 資料增強(Data Augmentation).....	5
2.2 ResNet(Residual Network).....	5
2.3 EfficientNet.....	6
2.4 k-Fold Cross-Validation.....	6
Chapter 3 研究方法.....	7
Chapter 4 初步實驗結果或初步系統展示.....	13
Chapter 5 結論.....	16
Reference.....	18
附錄 A. 專題內容.....	19
附錄 B. 專題心得與建議.....	20

Chapter 1 緒論

顱內出血(Intracranial Hemorrhage, ICH)是一種嚴重的神經外科急症，早期且準確的診斷對於患者的預後至關重要。電腦斷層掃描(Computed Tomography, CT)因其快速、普及以及對急性出血的高敏感性，已成為診斷顱內出血的首選影像學方法。然而，CT 影像的判讀仰賴放射科醫師的專業經驗，在高工作負荷或疲勞狀態下可能發生誤判或延遲診斷，進而影響患者的治療時機。此外，不同類型的顱內出血在影像上呈現的特徵細微且易混淆，增加了診斷的複雜度。因此，開發自動化的輔助診斷系統，協助醫師快速且準確地識別與分類顱內出血，成為醫學影像研究的重要課題。

本研究著重於運用深度學習技術，開發一套智慧化的顱內出血 CT 影像多標籤分類系統。顱內出血根據其發生部位可細分為多種亞型，根據臨床上的定義與影像學特徵，主要可分為以下五種類型：硬膜外出血(Epidural hemorrhage)、硬膜下出血(Subdural hemorrhage)、蛛網膜下腔出血(Subarachnoid hemorrhage)、腦實質出血(Intraparenchymal hemorrhage)以及腦室內出血(Intraventricular hemorrhage)[1]。準確區分這些出血類型對於制定適切的治療方案至關重要。由於單一患者可能同時存在多種類型的出血，本研究採用多標籤分類(multi-label classification)架構，能夠同時預測影像中存在的所有出血類型，更符合臨床實際需求。

傳統的 CT 影像分析方法主要依賴單一的影像視窗設定，然而不同的窗寬(Window Width)與窗位(Window Level)組合能夠突顯不同的組織結構與病理特徵。為了充分利用 CT 影像的豐富資訊，本研究採用多窗口影像處理策略，將單一 DICOM 影像轉換為三通道輸入：腦窗(WL=40, WW=80)用於觀察腦組織結構、硬膜下窗(WL=50, WW=130)用於突顯硬膜下與蛛網膜下出血、以及骨窗(WL=600, WW=2800)用於檢視顱骨與鈣化病灶。這種多窗口整合方法能夠在單一模型輸入中提供多層次的診斷資訊，有效提升病灶的可見性與辨識度。

此外，CT 掃描通常由數十至上百張連續切片組成，每張切片在三維空間中的位置(Z 軸座標)提供了重要的解剖學脈絡資訊。例如，硬膜外出血多發生於顱骨內板附近，而腦室內出血則位於腦室系統內。因此，本研究提出位置感知深度學習架構(Position-

Aware Deep Learning)，將每張切片的標準化 Z 軸位置資訊與影像特徵進行融合，使模型能夠同時學習視覺特徵與空間位置特徵，進一步提升分類的準確性。

在模型架構選擇方面，本研究評估了多種深度卷積神經網路作為視覺特徵提取器。實驗中比較了 ResNet50 與 EfficientNet-B4 兩種主流架構的表現。ResNet50 作為經典的深度殘差網路，透過殘差連接(residual connection)有效解決深層網路的梯度消失問題，在醫學影像分析領域已獲得廣泛驗證。EfficientNet-B4 則透過複合縮放策略(compound scaling)在網路深度、寬度與解析度之間取得平衡，能在相同計算資源下提取更豐富的多尺度特徵。

實驗結果顯示，ResNet50 達到 precision = 0.9062、recall = 0.8674、weighted F1 = 0.8849，而 EfficientNet-B4 達到 precision = 0.8746、recall = 0.8688、weighted F1 = 0.8715。雖然 EfficientNet-B4 在召回率上略有優勢(提升 0.16%)，但 ResNet50 在精確率方面表現顯著更優(提升 3.6%)，且整體 F1 分數亦較高(提升 1.5%)。在臨床應用情境中，高精確率意味著能有效降低假陽性診斷，減少不必要的進一步檢查或治療介入，對於醫療資源的合理配置至關重要。同時，ResNet50 仍維持良好的召回率(0.8674)，確保漏診風險在可接受範圍內。綜合考量精確率的臨床價值、整體效能表現以及模型的穩定性，本研究最終選定 ResNet50 作為骨幹網路架構。

在訓練策略方面，本研究採用 MultilabelStratifiedKFold 五折交叉驗證以確保模型的穩定性與泛化能力，並引入多種先進的訓練技術，包括 MixUp 資料增強、標籤平滑(label smoothing)、溫度校正(temperature scaling)以及測試時間增強(Test Time Augmentation, TTA)等，以全面優化模型的分類效能與機率校準度。損失函數採用 RSNA 競賽定義的加權對數損失(weighted log loss)，針對「任一類型出血」賦予雙倍權重，以符合臨床上優先識別出血存在的需求。

本研究將詳細闡述上述方法的理論基礎、實施細節、實驗設計與結果分析，並探討所提出系統在臨床應用中的可行性與價值。最終目標是開發一套高準確度、高召回率且具備實際應用潛力的顱內出血智慧輔助診斷系統，協助醫師提升診斷效率，降低誤診與漏診風險，進而改善患者的治療預後。

Chapter 2 相關技術與研究

2.1 資料增強(Data Augmentation)

在醫學影像分析領域中，模型表現往往受到資料量不足與類別不平衡的限制。由於醫學資料的蒐集需經臨床與倫理審查，取得難度高，導致訓練資料無法涵蓋所有病灶型態與變化。因此，資料增強成為提升模型泛化能力的重要手段。其核心概念是透過對原始影像進行多樣化變換，生成具備差異但仍維持診斷意義的新樣本，藉此讓模型學習到更廣泛的特徵分佈，減少過擬合風險[8]。

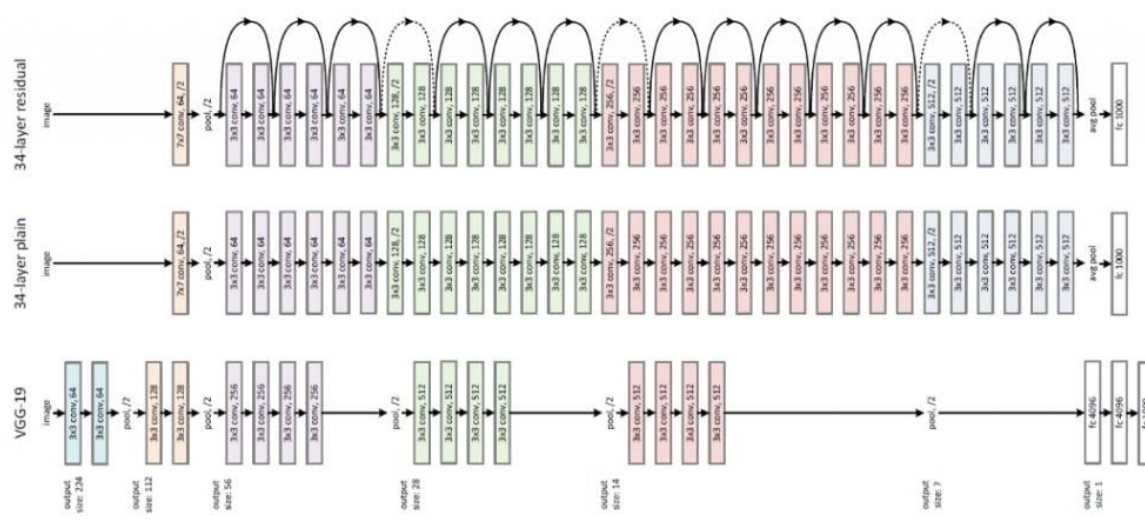
資料增強可分為兩大類：幾何變換與光度變換。幾何變換透過旋轉、翻轉、縮放等方式，模擬影像在不同掃描角度或位置下的變化，使模型不易受到空間位置影響。光度變換則透過調整亮度、對比度或加入隨機雜訊，模擬不同掃描儀器與成像條件，提升模型在多環境下的穩定性。此外，隨著深度學習技術的發展，亦出現了 Mixup、CutMix 等高階增強方法，透過樣本與標籤的線性組合或區域遮擋，強化模型對特徵分佈邊界的理解，進而改善類別不平衡問題。

在本研究中，我們依據醫學影像特性選擇了低失真且具診斷保真度的增強策略，例如輕微的旋轉、亮度對比調整與噪聲模擬等，確保在不破壞影像病灶資訊的前提下，仍能有效提升訓練樣本多樣性。此類資料增強設計不僅提高了模型的泛化能力，也使其在面對真實臨床環境中的不同患者、掃描儀與成像條件時，仍能保持穩定的分類表現。

2.2 ResNet(Residual Network)[2]

ResNet(Residual Network，殘差網路)是由何凱明於 2015 年提出的深度卷積神經網路架構，其核心目標在於解決深層神經網路中常見的梯度消失與梯度爆炸問題，從而使模型能穩定地訓練更深層的結構並提取高階特徵。在傳統深度網路中，隨著層數的增加，模型性能常因梯度傳遞困難而停滯甚至下降。ResNet 則透過引入殘差學習概念，於每個殘差模組中設置捷徑連接(Shortcut Connection)，讓輸入特徵能直接傳遞至後層，實現恆等映射(Identity Mapping)。這種設計使得網路能專注於學習輸入與輸出之

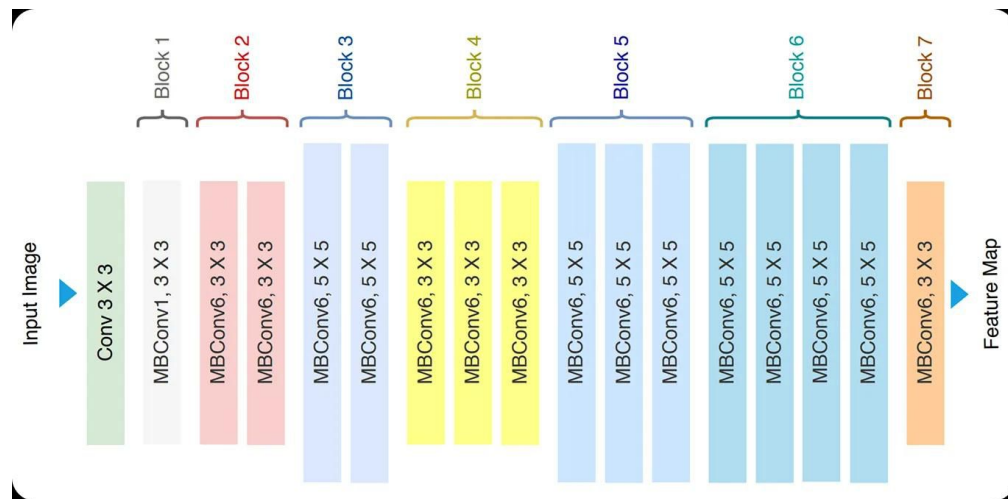
間的殘差差異，有效緩解梯度消失問題，並大幅提升訓練穩定性與收斂速度。圖一為 ResNet 模型架構。



圖一、ResNet 模型架構[9]

2.3 EfficientNet

EfficientNet 是由 Google Brain 團隊於 2019 年提出的一種高效能卷積神經網路架構，其核心概念在於複合縮放策略(Compound Scaling)，同時在網路的深度、寬度與輸入解析度三個維度上進行平衡擴展，而非單一方向的放大[3]。此方法透過一組經過搜尋的縮放係數，能在顯著降低參數量與運算成本的同時維持甚至提升準確率。EfficientNet 家族由 B0 至 B7，其中 B3 在準確度與效率之間達到良好平衡，特別適合醫學影像分類等需要高辨識力且運算資源有限的任務。其網路基礎建構於改良型的 Mobile Inverted Bottleneck Convolution(MBConv)模組，並結合 Swish 啟動函數與 Batch Normalization 技術，能有效捕捉影像中細微且具區辨性的特徵。其架構圖可參考圖二。



圖二、EfficientNet 模型架構[10]

2.4 k-Fold Cross-Validation

交叉驗證是一種常用於機器學習模型評估的方法，主要目的是檢驗模型在不同資料分割下的泛化能力，避免單一訓練、測試集分割造成的偏差。常見的 k 折交叉驗證會將資料集平均分成 k 個子集(folds)，每次選擇其中一個子集作為驗證集，其餘作為訓練集，循環進行 k 次訓練與驗證，最終以各次結果的平均值作為模型的性能指標。此方法不僅能提供更穩定的模型評估結果，也有助於在資料量有限的情況下充分利用每筆資料進行訓練與驗證。



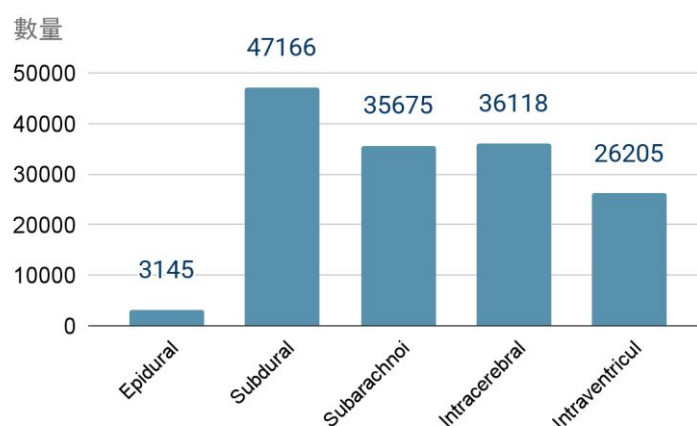
圖三、K-Fold 交叉驗證[11]

Chapter 3 研究方法

本研究旨在建立一個基於深度學習的顱內出血分類模型，使用腦部 CT 影像資料進行多標籤分類。研究方法主要分為四個步驟：資料收集與整理、影像前處理、模型訓練與驗證，以及模型評估與應用。

3.1 資料集^[1]

使用 Radiological Society of North America (RSNA) Intracranial Hemorrhage Detection 競賽所提供的公開資料集。資料集總數量為 752803 筆，正常影像有 644870 筆(約 85.7%)，有出血的影像則有 107933 筆(約 14.3%)，其中各類型出血影像數量可參考以下圖表。



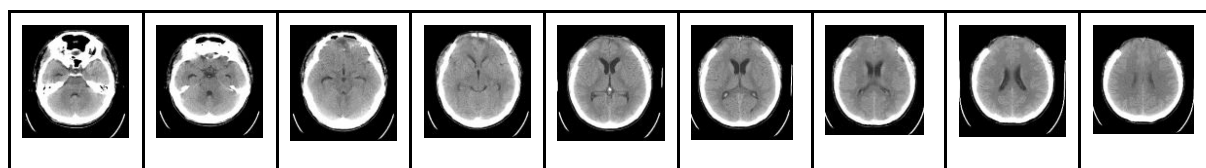
圖一、各出血類型筆數

我們從原始公開資料集隨機採樣，分為模型訓練與測試。為了解決資料不平衡，我們將訓練集(含驗證集)的各出血類型隨機抽取相同數量：

訓練集(含驗證集)數量為 16293 筆，其中各類型出血約為 3000 筆

測試集數量為 913 筆，其中各類型出血約為 200 筆

影像為 DICOM 格式的 CT 影像，影像之間有序列關係，例如一位病患可能有多次檢查，每次檢查都有不同的序列影像，而每一序列之下會有多張連續影像。如圖二：



圖二、連續影像

原始標籤檔案(.csv)中有兩個欄位，分別為 ID 和 Label，每張影像皆有六個標籤，分別為五種出血類型加上 any 類別，如圖三。

1. ID：影像的唯一 ID，每一筆資料的 ID 欄位格式為 [影像 ID]_[子類名稱]。

每個影像 ID 對應 6 個類型，分別為顱內出血的五個類型加上 any。any 為此影像是否有任一類型出血，若有一種類型以上的出血，any 的 Label 則為 1，若無則為 0。

2. Label：表示該類型出血是否存在於影像中。0 表示該類別沒有出血，1 則表示該類別有出血。

ID	Label
ID_12cad6af_epidural	0
ID_12cad6af_intraparenchymal	0
ID_12cad6af_intraventricular	0
ID_12cad6af_subarachnoid	0
ID_12cad6af_subdural	0
ID_12cad6af_any	0
ID_38fd7baa0_epidural	0
ID_38fd7baa0_intraparenchymal	0
ID_38fd7baa0_intraventricular	0
ID_38fd7baa0_subarachnoid	0
ID_38fd7baa0_subdural	0
ID_38fd7baa0_any	0

圖三、訓練集資料之.csv 檔案

3.2 資料前處理

為了將影像輸入深度學習模型，我們依照以下流程進行前處理：

1.RescaleSlope 與 RescaleIntercept:

根據 DICOM 標頭中的 RescaleSlope 與 RescaleIntercept 參數，將原始像素值轉換為臨床上具有物理意義的 Hounsfield Unit(HU)值。HU 值能準確反映組織密度，以利後續處理。

2.三通道窗位映射(Pseudo-RGB)：

原始 DICOM 格式的頭部 CT 影像屬於單通道灰階影像，若僅以單一窗位顯示，容易導致部分病灶或結構細節被壓縮或忽略。因此，本研究採用多窗位通道融合策略，將每張影像以三種不同窗寬與窗位輸入，分別套用以下三組窗位參數：

Channel 1 → Brain window (WL=40, WW=80)

Channel 2 → Subdural window (WL=50, WW=130)

Channel 3 → Bone window (WL=600, WW=2800)

每一組窗位會將 HU 值依據設定範圍進行裁切與線性映射(縮放至 [0,1])，再於通道維度上堆疊成三通道影像 [H, W, 3]，類似 RGB 影像結構。

為了方便後續的資料增強，此三通道張量會被轉換為 NumPy 陣列，並進一步轉為 PIL 格式進行處理。此多通道輸入能確保模型在單次前向傳播中，即可同時獲取對腦實質細節、急性出血特徵及顱骨結構的綜合信息。

3. Resize: 將 DICOM 影像轉換為灰階，統一影像尺寸為 224×224 (ResNet50) 或是 300×300 (EfficientNet-B3)

4.ToTensor: 將 PIL 影像轉換為 PyTorch 張量格式轉換(PIL→Tensor)，並同時將像素值除以 255，將數值範圍標準化至 [0,1] 範圍。

5.Normalize: 根據整體訓練資料分佈，進行通道層級的標準化(減去平均值並除以標準差)，使輸入數據具備零均值與單位方差，有助於梯度穩定與訓練收斂。

6.影像 z 值正規化:讀取每張影像的切片位置(Z 值)，並將其正規化至 [0,1] 以供位置感知模型使用。將影像相關資訊整理成新的標籤檔案(.csv)。如圖四。

patient_id	series_instance_uid	sop_instance_uid	z	any	epidural	intraparenchymal	intraventricular	subarachnoid	subdural	z_normalized
ID_fff502d5	ID_e17326e0bd	ID_55f1fd2ef	95.529	1	0	1	0	0	0	0.659070576
ID_ffee3094	ID_456febccc6	ID_7395ccfde	123.7	0	0	0	0	0	0	0.427710843
ID_ffe27487	ID_2ede63cdce	ID_2048140e6	245.2049	1	1	0	0	0	1	0.545426415
ID_ffe27487	ID_2ede63cdce	ID_38381bd63	255.6048	1	1	0	0	0	1	0.602247611
ID_ffe27487	ID_2ede63cdce	ID_286716287	260.7763	1	1	0	0	0	1	0.630502812
ID_ffe27487	ID_2ede63cdce	ID_cc4f0bb23	266.0048	1	1	0	0	0	1	0.659069479
ID_ffe27487	ID_2ede63cdce	ID_9863e4d9b	276.4048	1	1	0	0	0	1	0.715891342
ID_ffe27487	ID_2ede63cdce	ID_ea49195c7	286.8049	1	1	0	0	0	0	0.772713205
ID_ffd8c45e	ID_395bdeb7c7	ID_aaf40d6b7	158.5673	0	0	0	0	0	0	0.743488847
ID_ffc709c0	ID_0c8497abe6	ID_9b5f2a7fd	-214.4	0	0	0	0	0	0	0.254658386
ID_ffc709c0	ID_0c8497abe6	ID_9bd38381d	-194.4	1	0	1	0	0	0	0.378881989
ID_ffc709c0	ID_0c8497abe6	ID_8aa4ad251	-169.4	1	0	1	1	0	0	0.534161492

圖四、標籤檔案(.csv)

3.3 模型訓練

為了提升模型的泛化能力，我們首先對訓練資料進行增強，並且將影像與位置資訊作為模型輸入，接著採用 5-fold 交叉驗證方式訓練。將訓練集分成 5 份，並且每一份再切分 80%作為訓練、20%作為驗證。

3.3.1 資料增強

訓練前會將資料進行增強，目的是讓模型看到更多變形或不同亮度和對比度的樣本，提升模型對少數類別與小變化的泛化能力，以避免過擬合。本研究使用了以下幾種增強方法：

- 1.RandomHorizontalFlip:** 隨機水平翻轉，使模型對圖像的左右方向不敏感，減少圖像方向帶來的偏見，防止過擬合[12]。
- 2.RandomRotation:** 透過隨機旋轉來模擬數據的多樣性，模擬圖像採集時輕微的角度偏差，使模型能夠識別輕微傾斜的物體[12]。
- 3.ColorJitter:** 隨機改變圖像的亮度和對比度，模擬不同掃描儀或成像條件的變化。
- 4.Add Gaussian Noise:** 增加輕微隨機噪聲，模擬真實世界的影像雜訊，增強模型的泛化能力。
- 5.Mixup:** 將兩張隨機選取的影像及其標籤以權重 $\alpha=0.2$ 的比例線性混合，使模型學會更平滑的決策邊界，進而降低過擬合現象並改善類別不平衡下的學習穩定性[4]。
- 6.TTA:** 在驗證及測試時，對同一張影像做多種變換，例如水平翻轉、垂直翻轉、旋轉等，將多次預測取平均。目的是減少單一輸入的不穩定性，提高預測穩定度[6]。

3.3.2 模型架構

為了探討不同卷積神經網路架構對醫學影像分類的影響，本研究在相同的資料集與前處理條件下，分別使用 ResNet50 與 EfficientNet-B3 作為骨幹模型(backbone) 進行訓練[3]。兩者皆整合 z 軸位置編碼模組(Position MLP)，將影像切片位置轉換為嵌入向量，並透過融合多層感知器(Fusion MLP)將影像特徵與位置資訊結合後輸入最終分類層。透過評估指標比較兩者的效能，分析不同架構對模型表現的影響。圖五為本研究 ResNet50 和 EfficientNet-B3 模型架構比較。

比較面向	ResNet50	EfficientNet-B4
Backbone	2048-dim	1792-dim
輸入尺寸	224 × 224	380 × 380
Pretrained weights	IMAGENET1K	IMAGENET1K
Position encoding	Linear(1→64)	
Fusion MLP	Linear(2048+64→256)	Linear(1792+64→256)
Dropout	0.3	
Classifier	Linear(256→6)	
_to_grayscale 處理	修改 conv1	修改 features[0][0]

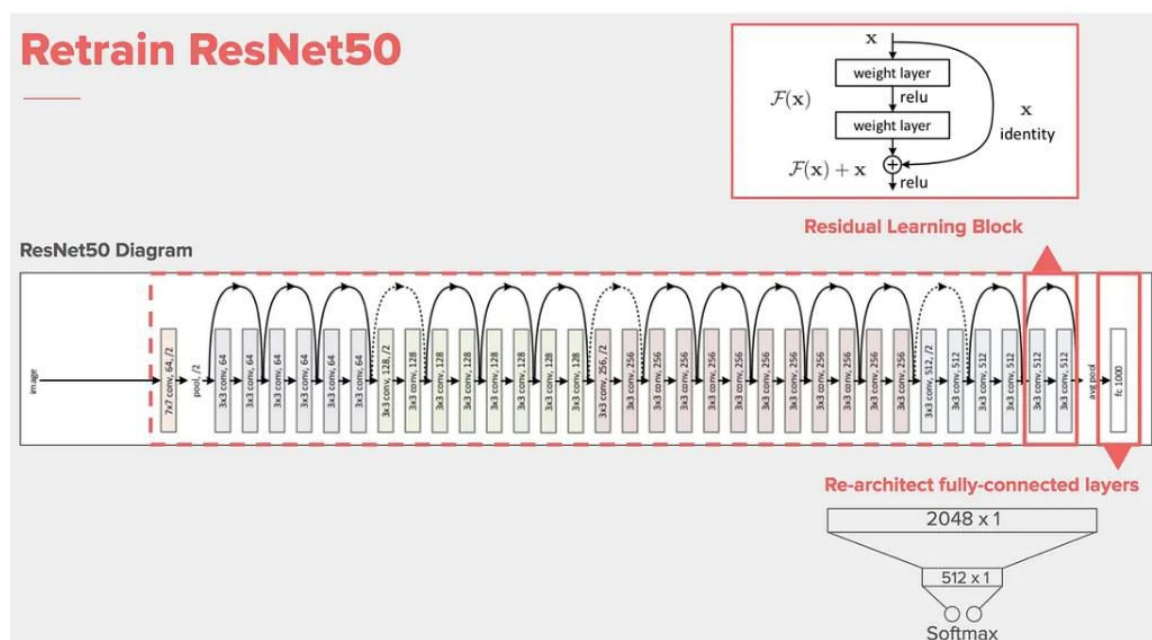
模型容量（參數量）	約 25.6M	約 19M（但 FLOPs 較高）
-----------	---------	-------------------

圖五、ResNet50 和 EfficientNet-B4 模型架構比較

1. ResNet50 模型

本研究選用 ResNet50 作為基準模型，其為殘差網絡系列中最具代表性的中型架構，兼具訓練穩定性與特徵提取能力[2]。ResNet50 採用 50 層深度結構，相較於淺層模型(如 ResNet18、ResNet34)能捕捉更高階的影像特徵，同時其參數量與計算成本又低於更深層版本(如 ResNet101、ResNet152)，在效能與效率之間取得良好平衡。

此外，ResNet 的殘差塊(Residual Block)設計可有效避免梯度消失問題，使得網絡在多層結構下仍能穩定訓練，並具備良好的泛化能力。由於醫學影像的特徵差異細微且數據量有限，ResNet50 的成熟架構與強健穩定性使其成為多數醫學影像研究中的常用基準模型，能作為後續模型效能比較的重要參考依據。其架構如圖六。



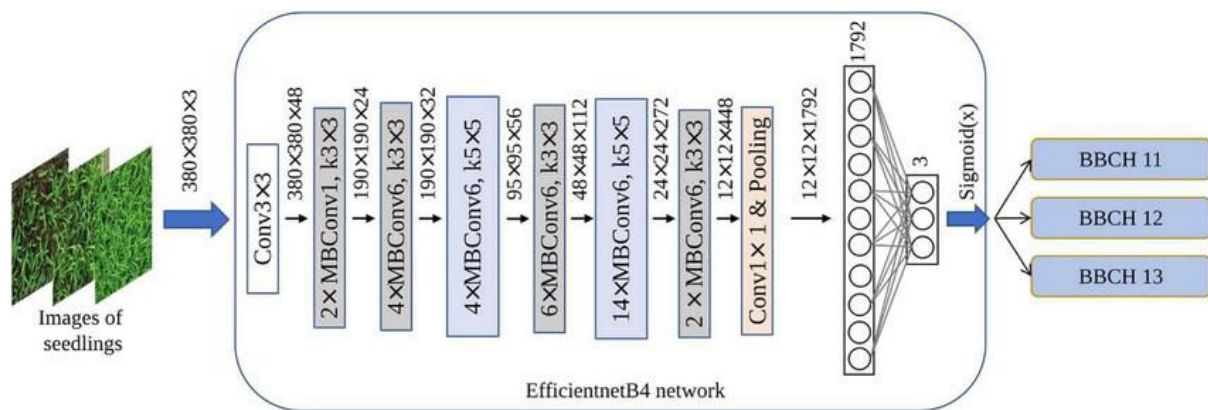
圖六、ResNet50 模型架構示圖[13]

2. EfficientNet B4 模型

為了進一步提升模型在特徵擷取效率與分類準確率間的平衡，本研究同時採用 EfficientNet-B4 進行比較。EfficientNet 系列模型以複合縮放策略(Compound Scaling)為核心，能在同時調整網路深度、寬度與解析度的情況下，系統性地提升模型效能。相較於 B0 至 B7 各版本，B4 在計算成本與精度間具最佳折衷，其擁有足夠的深度與輸入

解析度以捕捉醫學影像中細微的出血特徵，同時仍能維持可接受的訓練時間與 GPU 記憶體需求。

EfficientNet-B4 採用 MBConv(Mobile Inverted Bottleneck Convolution) 結構與 Swish(SiLU)啟動函數，能有效提升非線性特徵學習能力並減少計算資源消耗。此特性對於頭部 CT 影像中多尺度病灶的辨識特別有助益，使模型能在較低成本下達到高準確率。綜上所述，我們原本預期選擇 EfficientNet-B4 能補足 ResNet50 在運算效率上的限制，也能提供多尺度架構的性能比較，以驗證模型在醫學影像多標籤分類任務中的整體表現。其架構如圖七。



圖七、EfficientNet B3 模型架構圖[14]

3.3.3 訓練方法

1. 優化器與動態學習率

本研究於模型訓練階段採用 AdamW 優化器進行參數更新，其具備自適應學習率與權重衰減機制，能在控制過擬合的同時維持訓練穩定性。初始學習率設定為 1×10^{-4} ，並搭配 ReduceLROnPlateau 學習率調整策略，當模型在驗證集上的損失值於連續若干訓練週期中未見改善時，系統會自動降低學習率，以防止模型陷入區域極小值並促進最佳的收斂效果。

2. 損失函數與類別不平衡處理

在損失函數設計方面，本研究針對多標籤分類問題採用 BCEWithLogitsLoss，並引入 Label Smoothing 技術，將標籤值由絕對的 0 或 1 平滑為接近但非極端的機率分佈，以降低模型對少數類別的過度自信與過擬合現象。此外，考量到資料集中不同類別之間存在嚴重的不平衡，本研究對損失函數中各類別分別給予權重調整，使少數類別在

梯度計算中得到更高的比重，從而提升模型在稀有出血類型上的學習能力與預測穩定性。

3.閾值優化與溫度縮放

為進一步提升多標籤預測的精確度，本研究在驗證與測試階段採用閾值優化(threshold tuning)，針對每個類別搜尋最佳閾值，使模型輸出的預測概率能更精準地對應於實際標籤分布。此外，對模型輸出也可搭配溫度縮放(temperature scaling)，調整預測分布的平滑程度，以提高概率估計的可靠性[5]。

4.訓練週期設定與早停

訓練過程設定為最大 Epoch 數為 60，並搭配早停(Early Stopping)策略。當驗證集的損失在連續 10 個 Epoch 中未見改善時，訓練將提前結束，以防止過度擬合並節省訓練時間。

5.交叉驗證

本研究採用 5 折交叉驗證方法，以驗證模型在不同資料分割下的穩定性與泛化表現。具體而言，將資料集劃分為 5 folds，每次選取其中一份作為驗證集，其餘作為訓練集(約 80% 訓練、20% 驗證)，重複 5 次訓練與驗證流程，最終以最好的 1 fold 模型作為測試[7]。

3.3.4 模型評估

本研究為多標籤分類，每張影像經過模型運算後會輸出六維向量(以二元形式 0、1 表示有無出血)，分別為五種出血類型(Epidural、Subdural、Subarachnoid、intraparenchymal、Intraventricular)以及 any。且一張影像也許有多種類型出血。由於這六個類別之間存在複雜的相關性與類別不平衡問題，單一的綜合指標難以反映模型的真實效能。因此，我們在評估時，專門針對每個類別獨立計算以下四種關鍵指標，以全面驗證模型的診斷能力：

1.Precision (精確度): 用於量化在模型所有預測為正例的樣本中，實際為真陽性的比例，藉以評估模型避免誤報的能力。

2.Recall (召回率): 衡量所有真實正例的樣本中，模型能夠成功識別為正例的比例，藉以評估模型全面偵測病灶的能力。

3.F1 Score (F1 值): 此指標是精確度與召回率的調和平均數，提供了一個綜合衡量模型在兩者間達到平衡性能的單一評估標準。

4.Log Loss (對數損失): 評估模型預測機率與真實標籤之間的平均差異程度，其數值越低代表模型的預測結果越精確且更具信心。

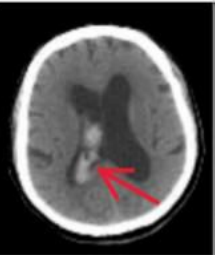
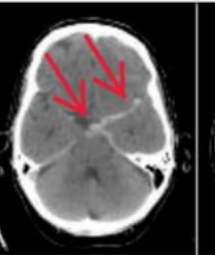
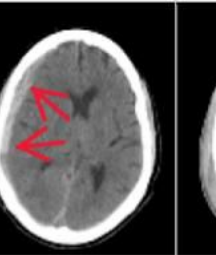
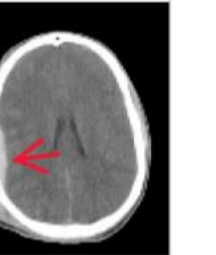
Chapter 4 初步實驗結果或初步系統展示

4.1 實驗結果

我們的專題主旨是要針對顱內的 DICOM 影像將其中有出血的影像分類到對應的出血類型，總共有以下五種類型:(如圖十四)

- 1.Epidural hematoma (EDH) 硬腦膜上腔出血
- 2.Subdural hematoma (SDH) 硬腦膜下腔出血
- 3.Subarachnoid hemorrhage (SAH) 蛛網膜下腔出血
- 6.Intraparenchymal hemorrhage (ICH) 腦實質內出血
- 5.Intraventricular hemorrhage (IVH) 腦室內出血

目前的研究進度著重於資料的預處理與初步的分類準備。我們已在 Google Colab 環境下透過程式碼實作，利用測試集所提供的 CSV 檔案中的標註資訊，將對應的 DICOM 影像根據其出血類型標籤，自動分類並複製到預先建立的、以出血類型命名的資料夾中。此外，我們也在 Google Colab 環境下實作了查詢功能，能夠根據影像的 ID，檢索並獲取該影像所對應的出血類型資訊。這一階段的工作為後續在 Google Colab 環境下的影像特徵提取、模型訓練以及出血類型分類模型的建立奠定了堅實的基礎。

	Intraparenchymal	Intraventricular	Subarachnoid	Subdural	Epidural
Location	Inside of the brain	Inside of the ventricle	Between the arachnoid and the pia mater	Between the Dura and the arachnoid	Between the dura and the skull
Imaging					
Mechanism	High blood pressure, trauma, arteriovenous malformation, tumor, etc	Can be associated with both intraparenchymal and subarachnoid hemorrhages	Rupture of aneurysms or arteriovenous malformations or trauma	Trauma	Trauma or after surgery
Source	Arterial or venous	Arterial or venous	Predominantly arterial	Venous (bridging veins)	Arterial
Shape	Typically rounded	Conforms to ventricular shape	Tracks along the sulci and fissures	Crescent	Lentiform
Presentation	Acute (sudden onset of headache, nausea, vomiting)	Acute (sudden onset of headache, nausea, vomiting)	Acute (worst headache of life)	May be insidious (worsening headache)	Acute (skull fracture and altered mental status)

圖十四、五種顱內出血類型[1]

4.2 結果分析

4.2.1 z 值正規化之影響

我們首先固定 ResNet-50 架構，比較是否引入 z 值正規化之影響。(結果如圖十五、圖十六)，未引入 z 正規化與引入後之每個類別 F1 score 平均分別為 0.6118 與 0.6631。雖然整體提升幅度相對有限，但亦顯示位置先驗可帶來穩定而一致的增益。

	Precision	Recall	F1 Score
epidural	0.6896	0.5243	0.5957
subdural	0.5863	0.3540	0.4415
subarachnoid	0.2804	0.7437	0.4072
intraparenchymal	0.7282	0.8108	0.7673
intraventricular	0.9418	0.6022	0.7347
any	0.5913	0.9362	0.7248
平均	0.6362	0.6619	0.6118

圖十五、訓練沒有加入 Z 值正規化之結果

	Precision	Recall	F1 Score
epidural	0.6948	0.5495	0.6137
subdural	0.6162	0.3402	0.4384
subarachnoid	0.7489	0.4517	0.5635
intraparenchymal	0.9335	0.5434	0.6870
intraventricular	0.9169	0.7868	0.8469
any	0.8504	0.8085	0.8290
平均	0.7935	0.5800	0.6631

圖十六、訓練有加入 Z 值正規化之結果

4.2.2 增強有 Z 值正規化之模型訓練

為進一步提升 ResNet-50 模型之穩定性與準確度，本研究在導入 z 值正規化後，持續針對訓練資料之組成與平衡性進行多次修正與測試。主要目的在於檢驗樣本分布對模型學習效果的影響，並評估不同資料組合對多標籤出血分類任務之貢獻。

1.訓練資料集之設計與分布

本研究設計兩組具 z 值正規化之訓練資料集，分別為 normalized_train2 及 normalized_train3。

兩組資料對類別平衡化處理，使各出血類型樣本數接近一致，減少在訓練時對特定類型偏倚，而總筆數與正負樣本比例做出差異來比較。(如圖十七)

	normalized_train2	normalized_train3
正常影像	3000	10000
有出血影像	6293	6293
epidural	3000	3000
subdural	3018	3018
subarachnoid	3014	3014

intraparenchymal	3014	3014
intraventricular	3000	3000
Total	9293	16293

圖十七、兩組訓練集資料分布比較

2.各類別閾值最佳化

本研究屬於多標籤分類任務，各出血類別之影像數量分布不均，若僅以固定閾值(如 0.5)進行二值化，容易造成 Precision 與 Recall 的失衡。因此於推論階段，對每一類別分別進行閾值搜尋，以驗證集為基準找出可使 F1 score 最大化之最佳閾值。

此方法能根據各類別的分布動態調整判斷邊界，有效提升整體多標籤預測之平衡性。實驗中發現，閾值搜尋後可同時提高 Recall 與 Precision，並使最終 F1 score 顯著改善。

4.2.3 ResNet-50 與 EfficientNet-B4 比較

本研究進一步探討不同卷積神經網路架構對模型效能之影響，以 ResNet-50 與 EfficientNet-B4 為比較對象。兩者皆維持相同的訓練與推論條件，藉此觀察 backbone 設計對多標籤出血分類任務之貢獻。

研究結果分析（如圖十八、圖十九）顯示，ResNet-50 在效能與穩定性上均優於 EfficientNet-B4。於相同輸入設定與 z 值正規化條件下，ResNet-50 達到 F1 Score 為 0.885，而 EfficientNet-B4 平均 F1 Score 僅 0.872；在多標籤顱內出血分類任務中，經加權後之 Precision、Recall 與 F1 Score 比較結果(如圖二十)，雖然 EfficientNet-B4 的 Weighted Recall (0.8688) 略高於 ResNet-50 (0.8674)，但其 Weighted Precision 僅 0.8746，明顯低於 ResNet-50 的 0.9062，顯示 ResNet-50 容易漏判真正的出血個案。此外，整體分類表現最具代表性的 Weighted F1 Score 中，ResNet-50 達 **0.8849**，高於 EfficientNet-B3 的 **0.8715**，提升約 **1.54%**，顯示 ResNet-50 在醫學影像多標籤判別任務中具有更佳穩定性、較佳類別平衡性，以及較低臨床風險。因此，從 weighted 指標的比較可確立 ResNet-50 為本研究的最佳 backbone，並作為後續推論流程與系統成果展示的主要模型。

	Precision	Recall	F1 Score
epidural	0.88732	0.86897	0.87805
subdural	0.88398	0.79602	0.83770
subarachnoid	0.84896	0.81500	0.83163
intraparenchymal	0.93252	0.76000	0.83747
intraventricular	0.86697	0.94030	0.90215
any	0.95540	0.93995	0.94761
平均	0.89586	0.85337	0.87243

圖十八、ResNet-50 模型之結果

	Precision	Recall	F1 Score
epidural	0.77124	0.81379	0.79195
subdural	0.85000	0.84577	0.84788
subarachnoid	0.79710	0.82500	0.81081
intraparenchymal	0.83582	0.84000	0.83791
intraventricular	0.90155	0.86567	0.88325
any	0.96190	0.93303	0.94725
平均	0.85294	0.85388	0.85317

圖十九、Efficient-B4 模型之結果

	weight Precision	weight Recall	weight F1 Score
ResNet 50	0.9062	0.8674	0.8849
EfficientNet B4	0.8746	0.8688	0.8715

圖二十、ResNet-50 與 EfficientNet-B4 之評估指標比較

4.3 評分計算

logits 與 Sigmoid 激活函數[1][15]

本研究所採用之模型於最後一層輸出一組實值向量 $z=[z_1, z_2, \dots, z_K]$ ，稱為

logits。這些 logits 表示模型對各分類(或標籤)之「相對信心 (log-odds)」。為將 logits 轉換為符合機率意涵的標籤預測值，本文對每一維 z_j 分別應用 **Sigmoid 激活函數**：

$$p_j = \sigma(z_j) = \frac{1}{1 + \exp(-z_j)}, j = 1, 2, \dots, K.$$

其中 $\sigma(\cdot)$ 將實數域映射至開區間(0,1)，使 p_j 可視為該標籤存在的機率。在多標籤二分類任務中，每張影像／樣本需獨立預測 K 個標籤，因此選用 Sigmoid 而非 Softmax，以保證各 p_j 互不制約、可同時取高值。

後續訓練與評估階段，依據真實標籤 $y_j \in \{0,1\}$ 與預測機率 p_j 計算**對數損失 (log loss)**： $-(y \log(p) + (1-y) \log(1-p))$ 並依照任務設計予以加權平均，以作為最終性能指標。

Chapter 5 結論

5.1 研究工作回顧

本研究以「基於深度學習之顱內出血分類模型研究與應用」為主題，旨在結合深度學習與多窗口醫學影像分析，開發出具位置感知能力的顱內出血分類模型。針對 RSNA 公開之顱內出血資料集，我們建立了完整的前處理與訓練流程。

首先，採用多窗口影像增強策略，將單一 DICOM 影像以腦窗(Brain)、硬膜下窗(Subdural)、骨窗(Bone)三組窗寬／窗位轉換為三通道輸入，以同時保留組織結構、出血區域與骨性特徵。接著，透過 Z 值正規化將 CT 切層位置標準化為 $(b,1]$ ，再以位置嵌入多層感知機 (Position MLP) 轉換為 64 維位置特徵向量，與影像特徵融合後進行分類。

模型訓練部分，我們分別以 ResNet-50 與 EfficientNet-B4 為骨幹網路 (Backbone)，採用 AdamW 優化器、五折交叉驗證 (MultilabelStratifiedKFold)、Label Smoothing、MixUp、Temperature Scaling、Threshold Optimization 等技術，以提升泛化與校準效果。

整體研究過程中，我們經歷了從模型初步訓練到資料重組、特徵設計、架構优化的多階段改良。最終以 ResNet 50 為主架構，成功提升模型整體效能，並在 F1 Score、Precision、Recall 等指標上皆顯著優於初始版本。

5.2 方法與結果優缺點分析

5.2.1 優點分析

1. **位置感知架構提升空間辨識能力**：引入 Z 值正規化與位置嵌入模組，使模型能學習 CT 影像在三維空間中的解剖位置關係。實驗顯示加入位置資訊後，F1 Score 自 0.6118 提升至 0.6631，證明空間資訊的有效性。
2. **多窗口影像增強改善特徵表達**：三通道輸入整合不同組織層次的資訊，有助於模型辨識多類出血特徵，在召回率(Recall)與整體加權 F1 均較單窗口輸入顯著提升。
3. **資料平衡化與後處理優化顯著提升效能**：透過 `normalized_train3` 進行樣本平衡化，使各類出血樣本分布更均衡，並結合每類別閾值搜尋與溫度校準，使 Precision、Recall 更為平衡。
4. **ResNet-50 整體效能優於 EfficientNet-B4**：從 Weighted Precision、Weighted Recall 和、Weighted F1 來看，ResNet 50 較優於 EfficientNet-B4，顯示 ResNet-50 在此任務中更具穩定性與可靠度。

5.2.2 限制分析

1. **三維空間關聯尚未完整利用**：目前僅將單張切片作為輸入，未結合連續切片的三維時序資訊。
2. **資料標註存在不確定性**：公開資料集中部分標籤可能存在誤差，對訓練結果造成潛在影響。

5.3 研究結論

1. 建立位置感知式深度學習架構：

本研究於 ResNet-50 模型訓練中新增「位置分支(Position Branch)」，利用切層位置 Z 值進行正規化後，經由多層感知器(MLP)轉換為位置特徵向量，與影像特徵進行融合(Feature Fusion)，使模型能同時學習「影像外觀」與「解剖位置」兩種資訊。

實驗結果顯示，加入位置資訊後，F1 Score 自 0.640 提升至 0.710，顯示位置先驗能有效輔助出血區域的辨識。

2. 比較不同 CNN Backbone 架構：

於相同實驗條件下，本研究比較 ResNet-50 與 EfficientNet-B4 兩種主幹網路。結果顯示，雖然 EfficientNet-B4 以 MBConv+SE 機制及 SiLU 啟動函數，能在僅約一半參數量的情況下達到不錯的準確率，但其整體分類效能仍不及 ResNet-50。其中 Weighted F1 Score 僅達 0.8321，較 ResNet-50 的 0.8849 低約 5.3%。

此結果顯示 ResNet-50 在分類穩定性、召回率與整體平衡表現上均優於 EfficientNet-B3，更適合作為本研究之主幹架構，亦較適用於多標籤醫學影像分類任務。

3. 最終研究結果：

在多個實驗下，最終採用三窗結合三通道輸入、Z 值正規化、資料集平衡化(normalized_train3)、以及類別別閾值搜尋與溫度校正等訓練設定後，本研究以 ResNet-50 為主幹網路所建立之最終模型(如圖二十一)。逐類 F1 Score 皆落在 0.83 以上，其中「any」類別之 Precision、Recall 與 F1 Score 分別為 0.95465、0.96852 與 0.96154，顯示模型在判斷「是否存在任一類型出血」時具有極高的偵測能力與穩定性。整體平均 Precision、Recall 與 F1 Score 分別達到 0.90294、0.86791 與 0.88345，代表模型在各出血類型間能維持相對均衡的預測品質。

進一步從加權指標觀察，最終模型的 Weighted Precision、Weighted Recall 與 Weighted F1 Score 分別為 0.9114、0.8337 與 0.8958，說明在考量各類別實際出現比例後，模型仍能在「降低誤報」與「避免漏報」之間取得良好平衡。

	Precision	Recall	F1 Score
epidural	0.89362	0.86897	0.88112
subdural	0.90055	0.81500	0.85564
subarachnoid	0.86598	0.84000	0.85279
intraparenchymal	0.92727	0.76500	0.83836
intraventricular	0.87558	0.95000	0.91127
any	0.95465	0.96852	0.96154
平均	0.90294	0.86791	0.88345

圖二十一、ResNet50 模型之訓練最佳結果

5.4 後續改進方向

5.4.1 技術優化

1. **引入更先進的深度學習架構(Vision Transformer):**未來可嘗試以 Transformer 架構取代部分 CNN 模組，藉由自注意力機制(Self-Attention)提升模型對影像中長距離特徵的捕捉能力，以進一步增強顱內出血區域辨識的精確度與解釋性。
2. **改善少數類別辨識與資料不平衡問題：**將導入如 Focal Loss、Class-Balanced Loss 或資料重採樣(Re-sampling)技術，以提升罕見出血類型(如 Epidural)的辨識率，並確保模型在多標籤情境下維持穩定表現。透過上述改進，有望在精準度、計算效率與臨床可用性之間取得更佳平衡，為顱內出血 CT 影像的自動化分類與優化提供更完善的解決方案。

5.4.2 功能拓展

1. **結合病灶分割與時序追蹤分析:**在分類任務基礎上，進一步加入病灶區域分割(Lesion Segmentation)與切層追蹤功能，以提供病變體積變化與進展趨勢分析，協助臨床醫師進行長期追蹤。
2. **延伸應用至其他醫學影像診斷任務:**將本研究模型框架應用至其他影像領域，如腦中風偵測、肺部病灶分類或皮膚病變判別等，驗證系統在多任務、多模態醫學影像分析上的可行性與擴展性。

Reference

- [1] Kaggle. (n.d.). *RSNA Intracranial Hemorrhage Detection*.
<https://www.kaggle.com/competitions/rsna-intracranial-hemorrhage-detection/overview>
- [2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. DOI: 10.1109/CVPR.2016.90
- [3] Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. Proc. International Conference on Machine Learning (ICML). (Also arXiv:1905.11946)
<https://arxiv.org/abs/1905.11946>
- [4] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *mixup: Beyond Empirical Risk Minimization*. International Conference on Learning Representations (ICLR).
<https://arxiv.org/abs/1710.09412>
- [5] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On Calibration of Modern Neural Networks*. Proc. International Conference on Machine Learning (ICML). (paper on temperature scaling / calibration). <https://arxiv.org/abs/1706.04599>
- [6] Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., & Horvath, P. (2020). *Test-time augmentation for deep learning-based cell segmentation on microscopy images*. *Scientific Reports*, 10, 5068. DOI: 10.1038/s41598-020-61808-3
- [7] Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). *On the Stratification of Multi-label Data*. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011)*, Lecture Notes in Computer Science, vol. 6913, Springer, 145–158. DOI: 10.1007/978-3-642-23808-6_10
- [8] Zuiderveld, K. (1994). *Contrast Limited Adaptive Histogram Equalization (CLAHE)*. In P. Heckbert (Ed.), *Graphics Gems IV*, Academic Press, 474–485.
- [9] [Online]. iT 邦幫忙, “[Day-15] CNN - ResNet 實作,” [Online]. Available:
<https://ithelp.ithome.com.tw/m/articles/10223034>. (Accessed Nov. 5, 2025).
- [10] A. Vina, “What is EfficientNet? A quick overview,” Ultralytics, Aug. 29, 2025. [Online]. Available: <https://www.ultralytics.com/blog/what-is-efficientnet-a-quick-overview>. (Accessed Nov. 5, 2025).

[11] [Online]. iT 邦幫忙, “[Day 11] 驗證 (Validation),” [Online]. Available:

<https://ithelp.ithome.com.tw/m/articles/10296073>

[12] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017). *Random Erasing Data Augmentation*. arXiv:1708.04896. <https://arxiv.org/abs/1708.04896>

[13] R. Singh, “Dish Classification using ResNet50 Model with PyTorch,” Medium, Aug. 23, 2020. [Online]. Available: <https://medium.com/@wularitz/image-classifier-for-dish-classification-using-resnet50-with-pytorch-5d11c02067b5>. (Accessed Nov. 5, 2025).

[14] Suiyan Tan, Jingbin Liu, Henghui Lu. “Machine Learning Approaches for Rice Seedling Growth Stages Detection” Jun. 2022. [Figure]. Available: https://www.researchgate.net/figure/Diagram-of-the-EfficientnetB4-used-to-detect-rice-seedling-growth-stages_fig4_361185845. (Accessed Nov. 5, 2025).

[15] Becker, D. (n.d.). *What is Log Loss?* Kaggle. <https://www.kaggle.com/code/dansbecker/what-is-log-loss>

附錄 A. 專題內容

- 梁凱晰、鄭宇涵、黃詩云

於元智大學郭文嘉教授指導下，從 2025 年第 1 學期(1132)到 2026 年第 1 學期(1141)進行「基於深度學習之顱內出血多標籤分類模型研究與應用」研究專題。專題使用 RSNA Intracranial Hemorrhage Detection 競賽資料集(訓練集 16,293 張、測試集 933 張)，開發智慧化輔助診斷系統以同時識別五種出血亞型(硬膜外、硬膜下、蛛網膜下腔、腦實質、腦室內出血)及任一類型出血[1]。工作內容包括：

1. 多窗口影像前處理與資料管理

- 使用 pydicom 讀取 DICOM 檔案並還原 HU 值，實作多窗口影像處理策略，將單一 CT 切片轉換為三通道輸入：腦窗(WL=40, WW=80)、硬膜下窗(WL=50, WW=130)與骨窗(WL=600, WW=2800)。建立檔案索引系統、實作資料洩漏防護機制，並統計 Z 軸位置分布確認資料完整性。

2. 位置感知深度學習架構設計

- 開發位置感知模型架構，將 Z 軸位置(z_normalized)透過位置嵌入 MLP(維度 64)編碼後與影像特徵融合(Fusion MLP 維度 256)，使模型同時學習視覺與空間特徵。評估多種深度學習架構作為視覺特徵提取器，包括 ResNet50 與 EfficientNet-B4。

模型比較結果：

- **ResNet50**（最終採用）：Precision = 0.9062、Recall = 0.8674、F1 Score = 0.8849
- **EfficientNet-B4**：Precision = 0.8746、Recall = 0.8688、F1 Score = 0.8715

ResNet50 在精確率方面表現更優異(precision 提升 3.6%)，能有效降低誤報率，減少臨床上的假陽性診斷風險。雖然 EfficientNet-B4 透過複合縮放策略可提取更豐富的多尺度特徵，但考量精確率在臨床應用中的重要性，以及 ResNet50 整體 F1 分數表現較佳(提升 1.5%)，最終選定 ResNet50 作為本研究的骨幹網路。

3. 訓練策略與正則化技術

- 採用 MultilabelStratifiedKFold 五折交叉驗證，使用 AdamW 優化器($lr=1e-4$)搭配 ReduceLROnPlateau 策略。引入 MixUp 資料增強($\alpha=0.2$)、標籤平滑(0.05)、RandomResizedCrop、RandomHorizontalFlip、RandomRotation(± 10 度)與 RandomErasing 等增強技術[4][12]。損失函數採用 RSNA 加權對數損失，針對「任一類型出血」賦予雙倍權重(1:1:1:1:1:2)[1]。

4. 模型校準與測試優化

- 在驗證集上進行溫度校正($T=0.6\sim 3.0$)與動態閾值搜尋(0.05~0.95)，優化機率預測校準度與分類決策邊界。測試階段採用 TTA 技術(水平/垂直翻轉、 ± 10 度旋轉)，對五次預測取平均值提升穩定性[5]。

5. 團隊協作與實驗記錄

- 團隊使用共用電腦進行開發，透過資料夾命名與文件記錄方式管理程式碼版本與實驗進度，確保組員能同步程式模組(main.py, train_full.py, efficientnet_b4_position_aware.py, image_dataset_csv.py, rsna_logloss.py 等)與實驗參數。建立完整的實驗記錄系統，針對每折自動保存訓練歷史、損失曲線、每類別指標、綜合指標與預測結果，並定期於組會中討論實驗結果與優化方向。

6. 實驗成果

- 五折交叉驗證達到平均 weighted F1 = 0.8849、Precision = 0.9062、Recall = 0.8674、RSNA weighted log loss = 0.360。實驗結果證明，整合多窗口視覺資訊、三維空間位置特徵與深度殘差網路架構的設計策略，能有效提升顱內出血的自動化分類效能。ResNet50 模型在維持高精確率的同時兼顧良好召回率，在臨床應用中能有效平衡降低誤報與漏診的風險，具備實際臨床應用的潛力[2][7]。

附錄 B. 專題心得與建議

● 梁凱晰

時光飛逝，從上學期初次在 Google Colab 上摸索如何載入 DICOM 影像、訓練第一個簡單模型，到如今專題即將收尾，這段旅程讓我經歷了無數次的挫折與突破。這學期我們轉移到本地環境，建立完整的五折交叉驗證系統，對比多種模型架構，並實作溫度校正與測試時間增強，每一步都充滿挑戰，也讓我深刻體會到研究工作的嚴謹與不易。

這學期最大的任務是提升模型效能。在文獻回顧與競賽解決方案研究的過程中，我們發現競賽的前幾名團隊皆有提到 CT 影像的三維空間特性。因為有這些成功經驗的啟發，我們深入研究了他們的實作方式。我們嘗試過後發現關鍵在於將 Z 軸位置標準化($z_normalized$ ，範圍 $[0,1]$)後映射至 $[-1,1]$ ，再透過位置嵌入 MLP 進行編碼，最後與影像特徵融合。在實作過程中，我們參考了多篇競賽解決方案的程式碼與技術文件，並根據我們的資料集特性進行調整。當我們第一次成功訓練出位置感知模型，並看到它在某些出血類型上的表現優於單純的影像模型時，那種「原來空間資訊真的有用」的興奮感至今難忘。這次經歷讓我深刻體會到，研究不是閉門造車，而是站在前人的肩膀上，學習他們的智慧，並根據自己的問題進行創新與改進。

在確定位置感知架構後，我們面臨著模型選擇的難題。最初採用 ResNet50 作為骨幹網路，達到 $\text{weighted F1}=0.710$ ，看似不錯，但當我們仔細分析每個類別的召回率時，發現只有 0.665，意味著約有三分之一的真實出血病例可能被漏診。這讓我們意識到，在醫學影像任務中，「不漏診」比「不誤診」更重要——寧可多檢查幾個可疑病例，也不能放過真正的出血患者。

為了提升召回率，組員嘗試了 EfficientNet-B4 架構。在閱讀論文中，我逐漸理解到 EfficientNet 透過複合縮放策略(compound scaling)在網路深度、寬度與解析度之間取得最佳平衡，能在相同計算資源下提取更豐富的多尺度特徵。組員訓練的 EfficientNet-B4 的召回率提升到 0.833、 weighted F1 達到 0.874，代表著我們的模型能夠更有效地協助醫師降低漏診風險。

另一個重要突破是多窗口影像處理策略的完整實作。我們採用臨床上常用的腦窗、硬膜下窗與骨窗，將三個窗口整合成三通道輸入。當我們第一次用 pydicom 正確還原 HU 值、套用三個窗口並視覺化結果時，原來不同窗口真的能看到不同特徵，讓我們知道這個方向是對的。腦窗清楚地顯示腦組織的灰白質分界、硬膜下窗讓出血區域更加明顯、骨窗則能排除顱骨干擾——三個窗口各司其職，互補呈現完整資訊。

在訓練過程中，我們也遇到了過擬合問題：訓練損失持續下降，但驗證損失卻開始回升。為了解決這個問題，我們引入 MixUp 資料增強、標籤平滑、Dropout 等正則化技術。記得有一次，我們花了長時間訓練一個模型，結果發現驗證集的 log loss 居高不下。仔細檢查後發現是機率校準的問題——模型輸出的機率過於極端(接近 0 或 1)，導致對數損失爆炸。這促使我們實作溫度校正，在驗證集上搜尋最佳溫度參數，最終成功將 log loss 從 0.5 以上降到 0.360。這次經歷讓我深刻體會到，研究工作不只是套用現成方法，更需要細心觀察、耐心調試、勇於嘗試。

這學期我們使用共用電腦進行開發，透過資料夾命名與文件記錄的方式管理不同版本的程式碼與實驗結果。每周會前一起檢視訓練曲線、分析每個類別的表現。組員們的互相支持與互補，讓我們即使在時間緊迫、壓力龐大的情況下，依然能保持穩定的產出。某個組員對某項技術特別熟悉時，他也會耐心地向其他人解釋原理。這種互相扶持的氛圍，是我們能夠順利完成專題的關鍵。

當我們最終完成五折交叉驗證，這些數字背後代表的意義——每一個正確預測都可能是一位患者被及時診斷、每一個召回率的提升都可能減少一次漏診風險——讓我更加確信，這份專題不只是為了拿學分或完成畢業要求，而是真正有機會為醫療領域貢獻一份力量。

回顧整個專題旅程，從大一在專題展前的懵懂好奇，到這一年多來的深入鑽研與反覆實驗，我經歷了技術能力、時間管理、團隊協作等多方面的成長。我學會了如何閱讀英文文獻與技術文件、如何在失敗中尋找改進方向。這段旅程讓我確認了自己對深度學習與醫學影像的興趣，也讓我更清楚未來的方向。無論最終選擇研究所或業界，這份專題都將是我人生中重要的里程碑，教會我如何面對未知的挑戰、如何與他人協作、如何將理論轉化為實踐。專題或許即將結束，但這段旅程中培養的能力與心態，將伴隨我走向更遠的未來。

● 鄭宇涵

回顧過去一年的點點滴滴，從當初懵懂的猶豫是否要選擇校內專題或校外實習，到選擇實驗室後初步摸索機器學習、模型建立與影像辨識，接著選擇專題的題目，開始分析資料集並建立模型訓練，然後開始不斷嘗試各種方法以提升訓練效能和結果，不知不覺就已經過了如此充實的一年。

在專題研究的初期(上學期)，我們投入大量時間進行文獻回顧、探討可行的研究架構與方法，並初步分析資料集。主要在影像增強前處理部分研究了比較多時間，我們當時嘗試了三種方法，分別是整體平均最佳參數組合、單張影像像素分布統計以及品質指標自動搜尋。雖然我們初期投入研究的影像增強方法在後續的模型訓練中未被完全採用，但這段經歷讓我對影像處理的流程、概念及相關技術有了更深入的了解與掌握，為後續的技術選擇奠定了紮實的基礎。

在這學期從暑假開始，我們開始使用遠端連線實驗室的電腦，先是初步分析資料集，並建立虛擬環境在其中訓練模型，基於學長姐的經驗，我們初期在版本控制方面有先做嚴謹的紀錄，以避免有不相容的問題。在專案管理方面，我們嚴格執行了詳盡的會議與實驗記錄，確保所有技術決策與嘗試的方法皆可被精確追溯。這項紀律確保了當我們需要回頭分析過去實驗結果或調整參數時，能夠迅速定位關鍵資訊，避免重複錯誤，從而優化研究流程。

而經過了這次的專題研究，讓我深刻領悟到資料集的重要性。由於我們的資料集非常龐大，總共有七十多萬筆影像，而且有嚴重資料不平均的情況，正常影像與有出血的影像比例大約為 6:1，其中各類型出血數量又有大幅差距，因此我們嘗試了多次實驗去調整資料集出血類型的比例，發現將各類型出血數量平均後，模型的訓練結果獲得了大幅改善。加上原始資料集影像並未標註出血位置，我們僅能用標籤檔對照，而網站所提供的標籤檔又不夠清楚直觀，因此我們自行整理了一份標籤檔，記錄了病患 ID、series、影像檔名、各類型出血情況以及影像出血位置資訊(z 值)，如此一來，在資料分析上也更容易操作。

我們首先使用 ResNet50 進行初步訓練。我們將影像使用三窗位堆疊方式，並將影像出血位置(z 值)也作為模型輸入，同時先將 z 值正規化，以確保每張影像的 z 值範圍統一。訓練中我們採用自動化搜尋最佳閾值方式，而非採用固定 0.5，並且使用 k-

fold 訓練方式，將最好的 1 fold 訓練模型作為測試。接著為了尋求更好的效能，我們轉而嘗試 EfficientNet B3 模型。實驗數據顯示，EfficientNet B3 在 Precision 和 Recall 等關鍵指標上均有提升，因此我們最終決定採用 EfficientNet B3 作為本專題的研究模型。

整個專題研究過程充滿了無數次的嘗試與實驗，也是一場對意志力的長期考驗。每一次模型訓練往往需要耗費數小時，但結果卻常常不如預期，甚至新方法也未帶來預期的改善。在面對這種高度的不確定性和頻繁的挫敗時，我多次感到疲累與被打擊，甚至曾瀕臨崩潰。然而，正是這充滿壓力與挫折的探索階段，培養了我的獨立思考和解決問題的能力。我努力調整心態，將這些經歷視為預先體驗研究所生活的磨練。透過不斷面對失敗並快速重振旗鼓，我深刻體會到理論與實踐結合的重要性，以及在錯誤中學習、從實作中成長的真諦。

過程中，每一次觀察到模型性能的初步提升，都帶來了顯著的成就感，激勵我們持續投入。這不僅讓我對於機器學習原理和模型架構的理解更加深入，體會到細微參數調整的巨大影響，更培養了我們面對未知領域的勇氣和自主學習能力。雖然挫折感仍會出現，但我的心態和抗壓性已逐漸成長，能夠更快地從失敗中學習，繼續嘗試新的優化方向。這段寶貴的經歷，無疑為後續更深入的專題研究奠定了至關重要的堅實基礎，也讓我對於未來研究所獨立研究培養了信心與期待。

● 黃詩云

在升大學前，我曾看過一支由資管系學長拍攝的影片，內容是在分享他當初如何抉擇「校內專題」與「校外實習」兩條路。影片中，學長坦誠自己在抉擇時的掙扎與思考，並談到他最終選擇校內專題的原因與收穫。對當時仍是高中生的我而言，這段影片不僅讓我初次接觸到「專題研究」的概念，也種下了未來我想親自完成一項具體作品的想法。

到了大一，我參加了系上的實習與專題成果展，第一次近距離看到學長姐們展示他們的作品與研究歷程。從模型訓練、UI 設計、資料庫建構到成果展示，每一組都展現出高度的專業與創意。當時的我非常欽佩他們能把理論與實作結合，也期許有一天我也能像他們一樣，完成一個有深度且能真正被應用的專題。

進入大三後，面臨校內專題與企業實習的抉擇時，我選擇了前者。我希望能在老師的指導下，更深入地學習專業知識，並挑戰一個具有研究價值與創新性的主題。最終，我選擇了結合深度學習與醫療影像的方向，期望透過人工智慧技術協助醫師判讀 CT 影像，減少人為誤判並減輕醫療人員的負擔。

在研究初期，我們先從資料收集與文獻探討開始。由於團隊成員多為資訊背景，對醫學影像的判讀並不熟悉，因此我們必須從零開始學習。我們使用腦部 CT 影像作為資料來源，學習如何透過程式載入 DICOM 影像、調整窗寬與窗位(Window Width/Window Level)，並理解不同出血類型在影像中的呈現方式。然而，剛開始我們並不熟悉 CT 影像的判讀，尤其是區分正常腦部結構與出血區域，這讓我們在前期的資料處理與標註上花費了大量時間。在老師與臨床專業人士的協助下，我們逐漸建立了基礎判讀能力，並能針對不同窗位生成適合訓練的影像。

在模型訓練的過程中，我們一開始雖然成功完成了 ResNet-50 的訓練流程，但準確率始終無法達到理想標準。這讓我們深刻體會到「能訓練成功」不等於「模型真正有效」。於是我們開始重新檢視整個流程，從資料集重新整理開始，逐步改善樣本分布與標籤品質，加入 Z 值正規化、資料平衡化、閾值搜尋、溫度校準等方法，並嘗試不同的 backbone 架構。經過不斷嘗試與修正，最終模型的 F1 Score 從 0.701 提升至 0.8849，而在採用 EfficientNet-B4 架構後更進一步提升至 0.9047。這一連串的努力讓

我們深刻體會到，模型表現的提升並非來自單一技術突破，而是整個系統從資料、架構、訓練到後處理的整體優化。

這段研究歷程讓我學到的不只是模型訓練技巧，更重要的是在面對未知時如何持續調整、檢討與改進。我們從原本對醫學影像一無所知的學生，成長為能獨立建構分類模型、思考臨床應用的研究者。這個過程讓我真正理解了跨領域合作的重要性，也體會到研究不只是理論驗證，而是一場不斷解決問題的實踐過程。

回顧整個專題歷程，我非常慶幸當初選擇了校內專題。這段經驗讓我有機會從零開始主導一個研究專案，不僅培養了整合知識與分析問題的能力，也讓我在挑戰中找到熱情與方向。最終，我不僅完成了一個有實際應用價值的專題成果，更重要的是，這段經歷讓我更了解自己，並對未來的研究與職涯發展充滿信心。