基於 U-net 框架的急性闌尾炎分類任務

組員姓名	任務分工、合作說明	工作百分比（%）
410315034 何紫妤	nnU-net 模型架構與比較、海報模板、書面報告	25%
412210037 林威岑	2D U-Net 模型架構與比較、海報講解、書面報告	30%
412210010 施柏均	嘗試 3D U-net 模型、海報講解、書面報告	25%
412210043 蔡智丞	書面報告	20%

摘要

本研究針對急性闌尾炎之醫學影像分類任務，比較兩種基於 U-Net 架構的深度學習模型──手動調參的 2D U-Net 與自動化設計的 nnU-Net──在分類與分割效能上的表現差異。資料來源為腹部 CT 影像及其對應之像素級遮罩，任務包括 Scan-level（病人層級）之是否患病分類與 Slice-level（切片層級）之病灶區域分割。

2D U-Net 採用自定義 Dice Loss 層與多階段流程優化，包括訓練集資料的篩選、影像正規化、資料擴增設計以及後處理機制（機率閾值與遮罩面積閾值）提升預測合理性；nnU-Net 則採用 3D full-resolution 設定，搭配自動推導之資料正規化、重取樣、patch 大小與資料擴增策略完成端到端訓練。

在 Scan-level 分類上，兩者皆能準確預測有患病的病患是否患病。其中 2D U-Net 在給定的測試集上達到 F1-score 0.833，2D U-Net 雖然實作上具備較高的彈性，但易受到流程設計細節（如正規化順序、資料切分與後處理）影響，需具備一定的專業性與嚴謹性；相對而言，nnU-Net 在給定的測試集上達到 F1-score 0.666，且需較高運算資源與前期準備，但能提供更穩定與結構性佳的模型效能。

本研究比較顯示，若欲於臨床應用中選擇合適模型，應依據資料量、資源條件與任務需求平衡選擇 U-Net 架構版本。未來亦建議建立一致的實驗標準流程，以避免流程差異影響評估結果，達成公允的跨模型性能比較。

壹、介紹

一、研究動機

急性闌尾炎是常見的急腹症之一，若未能及時診斷與治療，可能導致腹膜炎、敗血症等嚴重併發症，甚至引發休克與死亡。現行臨床診斷主要依賴醫師判讀腹部 CT 或 MRI 影像，然而影像判讀具高度主觀性，加上症狀常不明顯、容易與其他腹部疾病混淆，診斷過程中仍存在誤判或延誤的風險。

因此，本研究希望透過深度學習技術，建構具備自動化辨識能力的模型，以輔助醫師提升急性闌尾炎診斷的效率與準確性，減少人為誤差，並強化臨床診斷的即時性與可靠度。

二、資料集

本研究採用之資料來自 Kaggle 所舉辦的 AOCR 2024 AI Challenge（Classification task for acute appendicitis on contrast-enhanced CT scans）醫學影像競賽，資料集由臺灣新北市亞東紀念醫院提供。完整資料內容如下：

- **訓練集／驗證集資料**：共 1000 筆病患之 CT 掃描圖與對應之標註資料，其中 500 筆為確診急性闌尾炎患者之影像，另 500 筆則為具有急性腹痛症狀但並未診斷為闌尾炎之患者影像。
- **測試集資料**：共 200 筆病患之 CT 掃描圖，病患皆有急性腹痛症狀，但未提供診斷結果，作為最終模型評估使用。
- **三份 CSV 文件**：分別記錄訓練／驗證集的分割資訊、每位病患及其各切片是否具有闌尾炎的真實標註分類，以及提交結果的格式範例檔案。

本資料集之電腦斷層掃描圖（contrast-enhanced CT, CECT）涵蓋過去十年內收集之 18 歲以上成年人的腹部與骨盆腔增強型 CT 軸向切片，每張切片厚度為 5 毫米，大小為 512 × 512，每位病患之切片數量約介於 80 ～ 100 張之間。對應的標註遮罩由專家手動繪製，僅標示異常闌尾發炎區域，不包含周圍組織的發炎反應。上述影像與遮罩皆以 NIfTI 檔案格式（.nii.gz）提供。

考量訓練時間及運算資源限制，本研究從官方提供的訓練資料集中隨機抽取了 300 筆 CT 掃描圖與對應標註資料作為模型訓練集，另外隨機挑選 10 筆資料作為自我驗證測試集。最後，再使用未標註的 20 筆 CT 影像作為測試集進行模型性能的評估。

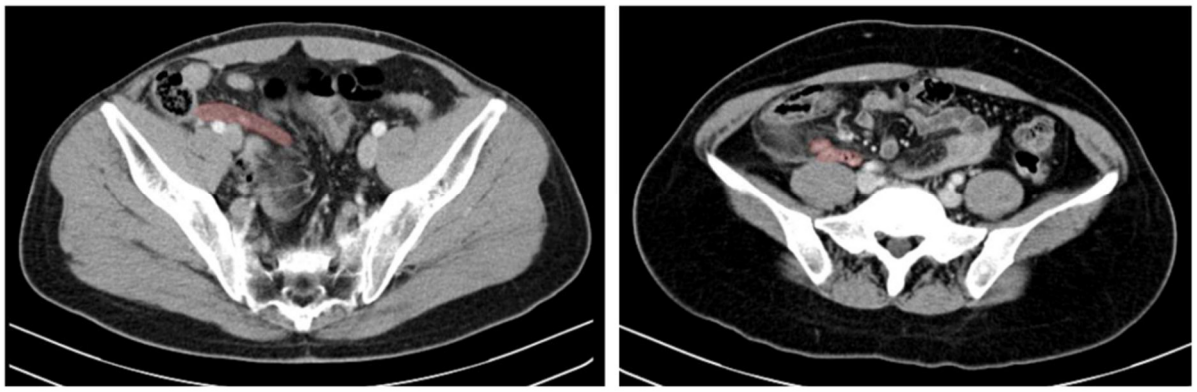

圖 1 為某一病患之 CT 掃描切片與其對應遮罩的合併範例。（圖像來源為 AOCR 2024 AI Challenge 競賽官方所提供之範例）

三、基於 U-Net 架構的網路模型

本研究的模型設計主要基於 U-Net 架構進行語意分割。儘管本研究之最終目的是針對急性闌尾炎進行分類判斷，但我們首先透過 U-Net 模型獲取腹部 CT 影像中闌尾炎區域的分割結果，並以此分割結果作為後續分類任務之依據。

實驗中採用手動調整參數的 **2D U-Net** 與自動化調整參數的 **nnU-Net** 兩種模型，比較兩種模型對於急性闌尾炎辨識準確度的影響，並探討其各自的適用情境與限制。

貳、文獻

一、選擇 U-Net 的動機

由於 U-Net 由 Encoder（編碼器）和 Decoder（解碼器）所組成，而這兩個結構能進行精確定位和分割，並且過往研究指出使用 U-Net、DenseNet、Res-U-Net 去偵測闌尾炎，可從 DSC 中看出 U-Net 相比其他兩個是最佳的[1]，應用領域也與本篇研究相符，因此選擇 U-Net 去做闌尾炎的分類。

二、基於 U-Net 框架的模型介紹

接下來我們會對這兩個模型進行簡單的介紹與說明。

首先 U-Net 是專為醫學影像分割設計的一個神經網路模型，其模型架構因長得像 U 字形而得名，主要透過 Encoder 將資料的特徵進行萃取，接著透過 bridge 連接 Encoder 與 Decoder，再透過 Decoder 重建回原本圖片的大小時，同時經由中間的跳接層保留了前面 Encoder 同一層級的定位資訊與細節。 U-Net 基於這樣的設計使得它在小型的資料集中能有效地進行像素級的精準分割，所以常用於做腫瘤、器官等的醫學影像分割任務。在這邊我們使用論文最初版的

2D U-Net 作為我們參考的模型架構。

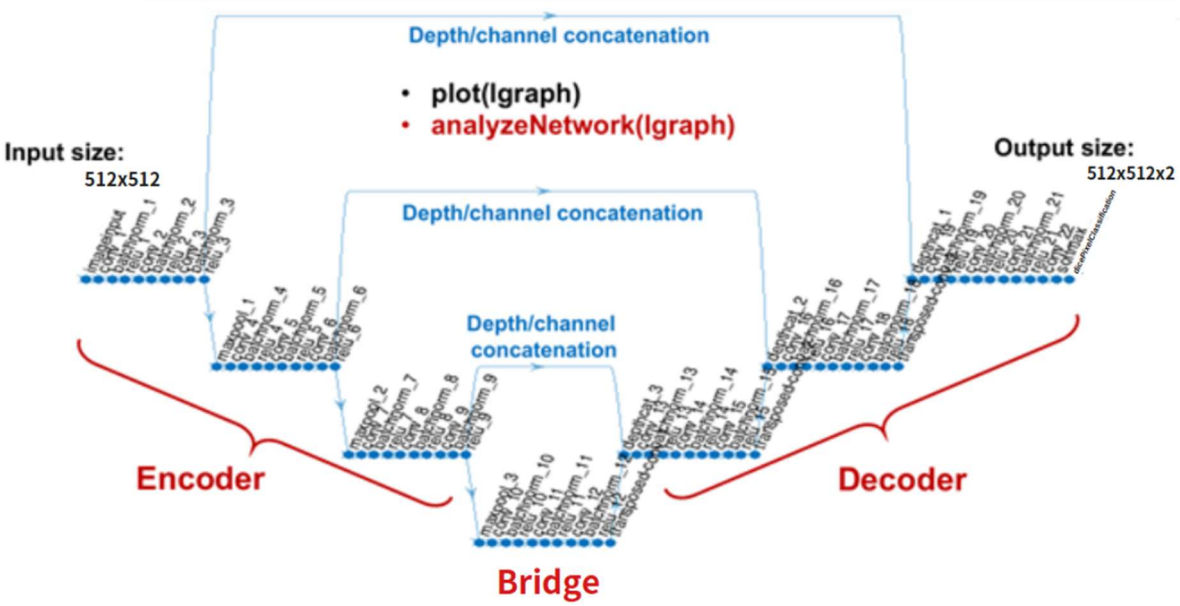

圖 2 為基於陽交大盧老師的上課影片內所撰寫的 U-Net 架構[5]去建構的 2D U-Net 模型

而 nnU-Net 也就是 no-new-Net，意思就是它不是新的神經網路架構，其提出「無需手動調參」的 U-Net 自動調整流程，顯示 U-Net 架構具高泛化能力，只要建構好穩定的資料結構使其自我訓練模型，就可穩定應用各類任務。在這邊我們選擇使用 3D U-Net 作為我們的模型。

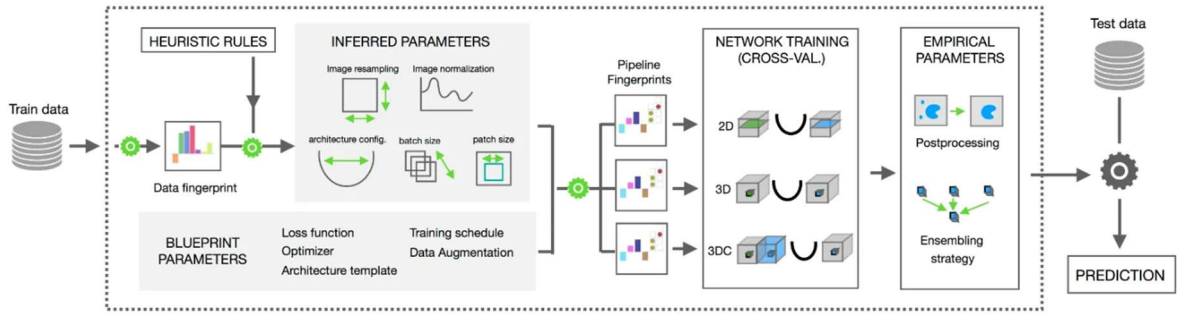

圖 3 為 nnU-Net 模型的基本架構流程[7]

本研究同時嘗試使用 3D U-Net，是基於三維資料本身具有體積資訊。與 2D 模型相比，3D 模型可以處理橫切面、冠狀面與矢狀面資料，能更完整捕捉

病灶空間特徵，所以 3D U-Net 能提升對闌尾區域的辨識準確性與穩定性，特別適合本研究面對的多切面腹部影像。

參、模型設計

一、2D U-Net

（一）研究流程

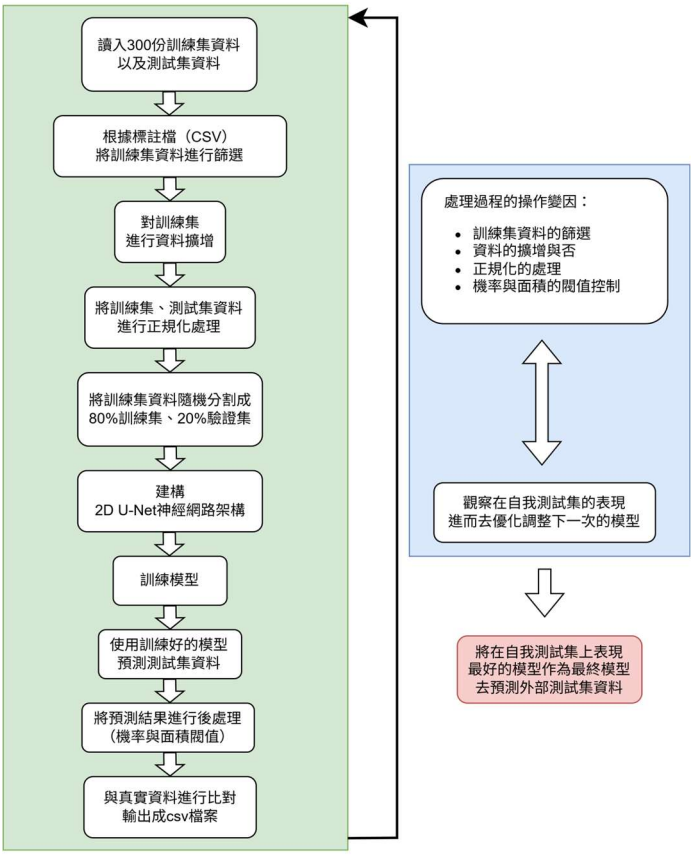

圖 4 為進行手動調參 2D U-Net 的研究流程圖。

（二）架構環境

此研究流程全程於 MATLAB 平台上進行實作，過程中使用了多項特殊函數與模組，部分來自 MATLAB 提供的工具箱，部分則為自定義實作，亦有取自第三方資源（如 MATLAB File Exchange）的輔助工具。以下列出實作所需的環境與依賴項：

需特別安裝的工具箱：

- **Image Processing Toolbox**：此為影像處理相關操作的核心工具箱，幾乎所有影像讀寫與轉換處理皆依賴於此，例如：imwrite, mat2gray, imwarp, logical, niftiread, niftiinfo 等函數。
- **Computer Vision Toolbox**：主要用於標註與儲存影像資料，如使用 pixelLabelDatastore 處理 segmentation 任務的遮罩資料。
- **Deep Learning Toolbox**：提供建構與訓練 U-Net 模型所需的層級（如 convolution2dLayer, batchNormalizationLayer, trainNetwork 等）及訓練選項。
- **Parallel Computing Toolbox**（非必要）：若使用 GPU 執行訓練（'ExecutionEnvironment','gpu'），則需依賴此 toolbox 配合支援 CUDA 的 NVIDIA GPU，以提升訓練效率。

自定義函數：

- **DicePixelClassificationLayer.m**：為我們自行實作的損失函數，用於模型輸出層，計算方式參考 Dice Loss，能有效處理醫學影像中常見的不均衡 segmentation 問題。
- **compute_f1.m**：根據輸入的 TP、FP、FN 值計算對應的 F1-score，應用於 slice-level 與 scan-level 預測評估。
- **natsort.m** 與 **natsortfiles.m**：由於 MATLAB 預設的字串排序非自然排序（例如 file_1, file_10, file_2），因此使用 MATLAB File Exchange 提供的自然排序工具，使檔名可依一般自然排序。

（三）實驗設計

在本研究的 2D U-Net 模型訓練中，我們透過反覆手動調整多項參數（包括影像正規化策略、資料擴增方法、前後處理機制，以及訓練資料的篩選方式），並搭配自建的測試集進行迭代驗證與優化，最終選定表現最穩定之模型進行最終預測。詳細的版本測試紀錄可參見附錄之實驗日誌（圖 6），其中以螢光標記的結果為加入後處理策略後的模型表現。

本研究所使用之 2D U-Net 模型以論文原始架構為基礎[2]，並也參考了陽交大盧家鋒老師在 MATLAB 進階程式語言的上課影片內所撰寫的 U-Net 架構[5]，建構出模型深度為四層結構（三層 Encoder 與三層 Decoder，並以中間一層 Bridge 相連接），特徵通道數從輸入的 2 開始，逐層擴增至 16、32、64，於 Bridge 處達到 128，並於解碼階段對稱還原至最終 2 通道之 segmentation 輸出。此外，針對醫學影像中常見的不平衡類別問題，我們模型之最後輸出層採用自定義的 DicePixelClassificationLayer 作為損失函數。

模型的輸入影像尺寸固定為 [512, 512]，訓練過程使用 Adam optimizer，初始學習率設為 0.0001，訓練共進行 30 個 epoch，每次 mini-batch 的大小為 32，整個過程皆於 GPU 環境下執行。

以下分別說明各項參數手動調整之策略與實施細節：

1. 訓練資料的篩選策略

   研究初期，我們僅使用含有闌尾炎遮罩的 CT 切片資料作為訓練資料。然而經過數次訓練後發現，雖然此方式能獲得較高的真陽性（TP）率，但同時產生了大量的偽陽性（FP）結果。因此為有效降低 FP，我們逐次增加健康的 CT 切片資料，並透過實驗結果證明，此方法能夠有效提升模型整體表現，降低誤判率。

2. 資料擴增設計

   由於訓練資料中具有闌尾炎遮罩的 CT 切片數量（共 1314 張）與健康切片數量嚴重不均衡，為強化模型對闌尾炎病灶區域的辨識能力，我們針對含有闌尾炎遮罩的資料進行資料擴增，操作包括：

   - **隨機旋轉**：-10°至 +10°
   - **隨機伸縮**：0.8 倍至 1.2 倍

   考量闌尾炎病灶在 CT 影像中的結構特性，本研究未採用隨機垂直或水平翻轉之擴增方式，以避免破壞影像的真實性與臨床上的一致性。

3. 前處理策略：影像強度正規化

原始影像資料為 12-bit 灰階 CT 影像，為確保模型穩定且有效訓練，我們在影像輸入模型前，採用正規化方式將影像強度縮放至 0–1 區間內。我們比較了以下兩種正規化方式：

- **mat2gray**：根據單張影像的極值，線性拉伸影像強度至 [0, 1] 範圍內。此方法雖能凸顯單一影像的對比細節，但整體影像資料間的亮度與對比不一致。

- **除以 4095**：直接將所有像素值除以固定值 4095，保持各影像間灰階強度比例的一致性，有效穩定整體 CT 圖像的對比與細節。

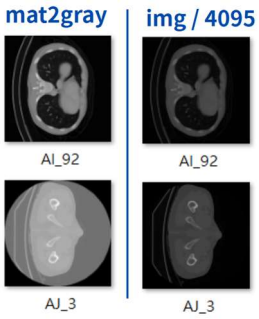

圖 5 為不同正規化處理之影像視覺效果比較。

4. 後處理設計：機率與面積閾值控制

在模型預測的後處理階段，除了基本的機率閾值－label = 1（闌尾炎）機率需大於 label = 0（正常）外，本研究額外加入以下兩項策略，以提升實用性與準確性：

- **機率控制**：
  模型預測區域要被認定為闌尾炎遮罩，其對應 label = 1（闌尾炎）之預測機率需大於 0.77 才視為有效區域。

- **面積控制**：
  臨床上闌尾炎區域通常具一定大小，因此我們透過統計訓練集中真實病灶的遮罩面積，原先以最小值作為過濾門檻，但發現極端值影響效果較大。為此，我們改用第 9 百分位數作為遮罩面積閾值，能更穩定有效排除面積過小之偽陽性預測。

透過觀察模型在自我測試集上的結果（請參閱圖 6 的第 2 次訓練），我們發現主要控制的偽陽性（FP）數量並非模型機率閾值的設定，而是面積閾值的控制。

最終，為選出最能兼顧影像分割準確度與臨床應用可行性的最佳模型，我們綜合考量兩種不同層級的表現指標：slice-level 與 scan-level 的 F1-score。前者反映模型在切片層級的正確性，後者則對應病患層級的實際診斷準確性。我們設計最終評分函數如下（同 kaggle 競賽的評分標準）：

Final score = 0.5 × (scan-level F1-score) + 0.5 × avg(slice-level F1-score)

此方式可平衡模型在不同應用情境下的重要性，並輔助我們從多個版本中選出最具穩定性與實用價值的模型。

而從自我測試集訓練成果的表格（圖 6）上來看，經過後處理設計的第 13 次訓練擁有最高的 final score 0.7468，所以我們選擇使用這個模型作為我們最終丟入給定測試集評估的模型。

	餵入的資料		前處理		網路架構設計 & 訓練超參數		後處理		Self test						
訓練次數	training_data	資料數量	資料擴增	正規化	MaxEpoches	Learning rate	機率閾值	面積閾值	TP	FP	FN	Scan - TP	Scan - FP	Scan - FN	Final score
1	只有闌尾炎遮罩的資料	1314	0	mat2gray	30	1.00E-04	0	0	38	319	22	6	4	0	0.522285553
							0.77	9%	20	128	40	6	4	0	0.508769913
2	只有闌尾炎遮罩的資料	1314	0	/4095	30	1.00E-04	0	0	47	296	13	6	4	0	0.556817169
							0.7	0	45	275	15	6	4	0	0.558283275
							0.7	5%	28	105	32	6	4	0	0.585032389
							0.77	9%	25	85	35	6	4	0	0.582352433
3	只有闌尾炎遮罩的資料	1314	1	mat2gray	30	1.00E-04	0	0	59	333	1	6	4	0	0.579981623
							0.77	9%	45	112	15	6	4	0	0.657281643
4	只有闌尾炎遮罩的資料	1314	1	/4095	30	1.00E-04	0	0	53	338	7	6	4	0	0.563321167
							0.77	9%	44	147	16	6	4	0	0.625582211
5	闌尾炎資料後30張	4645	0	mat2gray	30	1.00E-04	0	0	47	113	13	6	4	0	0.655695899
							0.77	9%	33	54	27	6	4	0	0.634047619
6	闌尾炎資料後30張	4645	0	/4095	30	1.00E-04	0	0	53	191	7	6	4	0	0.624882423
							0.77	9%	46	80	14	6	4	0	0.693470041
7	闌尾炎資料後30張	9290	1	mat2gray	30	1.00E-04	0	0	47	98	13	6	4	0	0.673192675
							0.77	9%	39	46	21	6	4	0	0.669750967
8	闌尾炎資料後30張	9290	1	/4095	30	1.00E-04	0	0	51	112	9	6	4	0	0.7037017
							0.77	9%	37	52	23	6	4	0	0.687388769
9	闌尾炎資料前後20張	6141	0	mat2gray	30	1.00E-04	0	0	45	51	15	6	4	0	0.703715957
							0.77	9%	39	19	21	6	3	0	0.74480254
10	闌尾炎資料前後20張	6141	0	/4095	30	1.00E-04	0	0	46	44	14	6	4	0	0.728238916
							0.77	9%	40	18	20	6	3	0	0.740648148
11	闌尾炎資料前後20張	7485	1	mat2gray	30	1.00E-04	0	0	39	29	21	6	4	0	0.706395688
							0.77	9%	26	9	34	5	2	1	0.691198018
12	闌尾炎資料前後20張	7485	1	/4095	30	1.00E-04	0	0	44	54	16	6	4	0	0.698687683
							0.77	9%	34	21	26	6	2	0	0.738209707
13	闌尾炎資料前後20張 + 5個健康的人部分遮罩(3n)	6289	0	mat2gray	30	1.00E-04	0	0	46	44	14	6	4	0	0.725694445
							0.77	9%	39	19	21	6	3	0	0.746847694
14	闌尾炎資料前後20張 + 5個健康的人部分遮罩(3n)	6289	0	/4095	30	1.00E-04	0	0	48	58	12	6	4	0	0.729150955
							0.77	9%	42	30	18	6	4	0	0.738575036
15	闌尾炎資料前後20張 + 5個健康的人部分遮罩(3n)	7633	1	mat2gray	30	1.00E-04	0	0	43	98	17	6	4	0	0.663024691
							0.77	9%	35	21	25	6	3	0	0.723412698
16	闌尾炎資料前後20張 + 5個健康的人部分遮罩(3n)	7633	1	/4095	30	1.00E-04	0	0	39	74	21	6	4	0	0.655034086
							0.77	9%	31	31	29	6	3	0	0.670515411

圖 6 為每一次在自我測試集上訓練優化的成果。

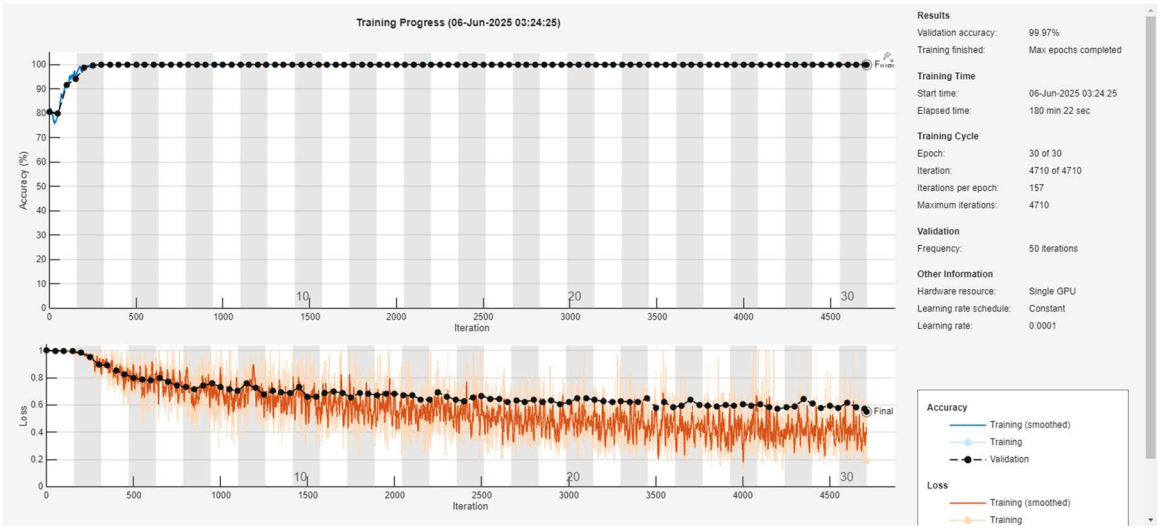

圖 7 為第 13 次模型訓練過程的正確率以及損失曲線。

二、nnU-Net

（一）研究流程

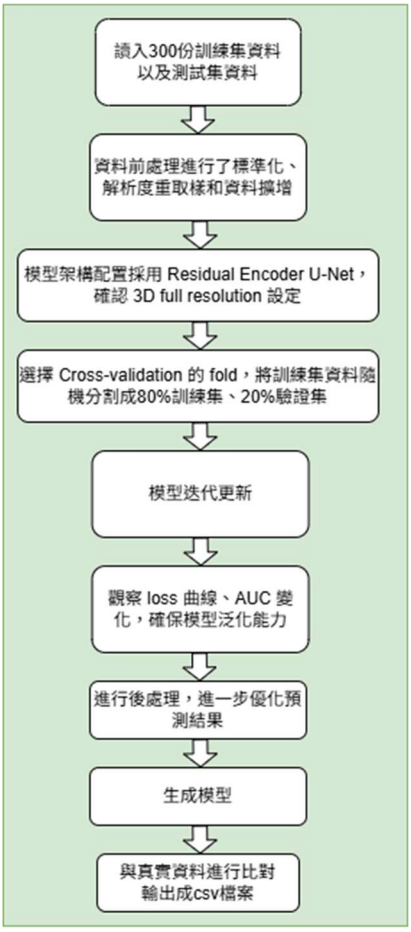

圖 8 為進行 nnU-Net 模型的研究流程圖。

（二）架構環境

採用 PyTorch 為主要深度學習框架，並基於 nnU-Net 進行模型訓練與推論。系統執行於 Windows 作業系統，搭配 Python 3.8 以上版本與相關函式庫（如 torch, monai, SimpleITK），以便處理醫學影像資料，並配合 nnUNet_preprocessed 目錄結構進行標準化管理，以確保數據一致性。

（三）實驗設計

選用 nnU-Net 偵測闌尾炎區域。nnU-Net 為一種自動化深度學習影像分割框架，具備根據資料特性自動設定前處理流程、網路架構、訓練策略與後處理規則的能力。該方法在無需人工干預的情況下，即可建立高效能的模型結構，並已被證實在多數醫學影像資料集上皆具優異表現。

模型架構使用來自 nnU-Net 的 3D 高解析配置（3d_fullres），網路基礎為 Residual Encoder U-Net，具有七層深度的編碼與解碼結構，從 32 通道逐層擴增至 320，並應用了三維卷積操作與 LeakyReLU 非線性函數。為進行影像標準化，採用 CTNormalization 方法處理灰階值，而空間解析度則統一重取樣為 [5.0, 0.6836, 0.6836] 毫米，輸入 patch 尺寸設定為 [56, 320, 256]，以確保資料在不同患者間保持一致的維度特性。

資料分割使用預設之 cross-validation 分割設定，選擇 Fold 1 進行訓練，其中包含 240 筆訓練樣本與 60 筆驗證樣本。為增強模型泛化能力，開啟虛擬 2D 資料擴增機制以提高數據多樣性。此外，訓練過程的主要超參數亦經細緻調整以避免過擬合現象。模型訓練採批次大小為 1，初始學習率設定為 0.0002，並引入步進式學習率調整策略，每 20 epoch 降低學習率，其衰減係數 gamma 經由觀察驗證集性能曲線細緻調控。模型效能的穩定性與泛化能力亦透過觀察訓練與驗證 loss 曲線及分類表現（如 AUC）加以驗證，進一步確立最適參數組合。

整體訓練過程與資料處理流程皆依據 nnU-Net 自動生成之規劃檔（plans.json）與預設前處理器（DefaultPreprocessor）進行，確保操作一致性與實驗重現性。

此外，儘管 nnU-Net 框架預設訓練上限為 1000 個 epoch，惟本研究受限於時間與運算資源，最終僅執行前 33 次訓練週期。雖已挑選驗證集表現最佳之模型版本進行後續比較，但整體模型尚未收斂，其效能表現仍呈現高度波動。如圖 9 所示，雖訓練與驗證損失（loss）有逐步下降趨勢，惟 Dice 分數（pseudo dice）波動劇烈，且移動平均線（mov. avg.）未能穩定上升，顯示模型尚未有效學習到具泛化能力的特徵表示。此外，學習率呈線性衰減，推論其尚未進入最佳學習階段，而 epoch 執行時間維持穩定，顯示硬體效能足以支撐完整訓練流程。未來若延長訓練週期，預期模型表現仍具進一步優化空間。

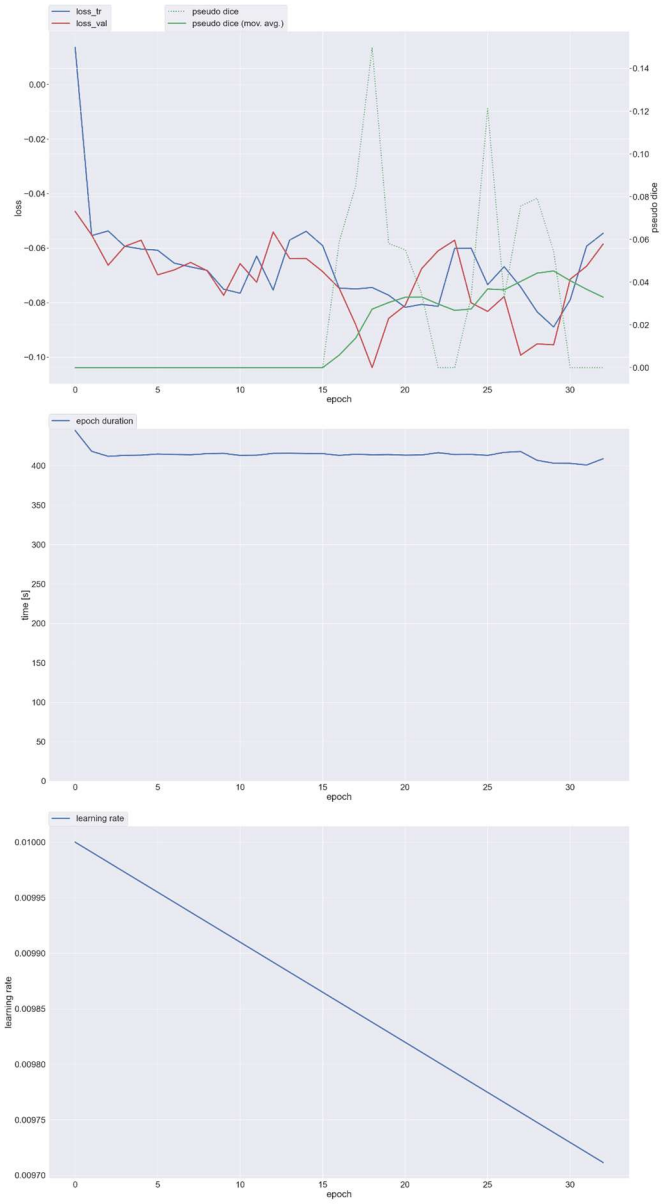

圖 9 為 nn U-Net 的模型訓練趨勢圖。

三、嘗試使用 3D U-Net

當我們嘗試訓練 3D U-Net 模型時，發現處理資料的遮罩（mask）是一個相當複雜的步驟。原本我們使用 MATLAB 的 categorical 類別來進行標註分類，但在將處理後的資料合併並輸入模型進行訓練時，卻發現模型的預測結果與我們預期的輸出不一致。因此，我們改採用自定義的 one-hot 編碼方式來表示分類遮罩和自定義的 Dice loss 才成功解決了這個問題。然而，模型在訓練時仍遇到 GPU 記憶體不足的問題。為了解決這個問題，我們將原始影像尺寸從 [512 512 96] 降低至 [256 256 48]還有減少編碼器和解碼器的層數從 4 層到 3 層，以減少運算與記憶體負擔。調整後，模型才能順利完成訓練流程。然而基於時間上的限制，導致測試集的資料尚未處理完，因此無法將模型應用到測試集上進行比較與探討。

肆、成果

一、模型評估的指標

在本研究中，我們採用 F1-score 作為主要的模型效能評估指標，以衡量模型在急性闌尾炎分類任務中對病灶區域的識別能力與整體分類準確性。F1-score 為精確率（Precision）與召回率（Recall）之調和平均，特別適用於處理正負類不平衡的醫學影像資料情境。

在模型預測結果中，**真陽性**（True Positive, TP）表示實際為陽性（有闌尾炎）且被模型正確判斷為陽性的樣本；**偽陽性**（False Positive, FP）則為實際為陰性（無闌尾炎）但被模型誤判為陽性的樣本；**偽陰性**（False Negative, FN）為實際為陽性但模型未能偵測出的樣本。

根據上述定義，精確率與召回率的計算公式如下：

$$\text{Precision} = TP / (TP + FP)$$
$$\text{Recall} = TP / (TP + FN)$$

而 F1-score 的計算公式為：

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

為了更全面地評估模型效能，我們分別針對**掃描層級**（scan-level）與**切片層級**（slice-level）進行分析。掃描層級評估模型能否準確辨識每位病人是否

患有闌尾炎；切片層級則進一步衡量模型對於病灶區域在單張切片中的定位準確性。此雙層次指標設計有助於兼顧整體分類與區域分割任務的評估需求。

二、比較各模型的 F1 Score

下方試算表（圖 10, 11）為模型在 20 位病人測試集上的推論結果分析，模型任務是判斷病患是否有闌尾發炎，根據掃描級和切片級的 CT 影像判別結果，利用 TP、FP、FN 計算精確率及召回率進而算出 F1 score。

欄位解讀：

id：病人代碼，每一列代表一位病人；最後一列 all 是整體（scan level）評估。

TP（True Positive）：模型正確偵測到有發炎的切片數。

FP（False Positive）：模型誤判為發炎的切片數。

FN（False Negative）：模型未偵測到發炎的切片數。

F1 score：根據每位病人所有切片計算的 F1 分數，衡量模型在該病人切片中的準確程度，也就是精確率與召回率的調和平均[6]。

（一）2D U-Net

	A	B	C	D	E
1	id	TP	FP	FN	f1score
2	AA	2	2	8	0.285714
3	AB	0	1	4	0
4	AC	0	0	0	0
5	AD	0	0	0	0
6	AE	4	0	4	0.666667
7	AF	8	1	4	0.761905
8	AG	0	0	0	0
9	AH	6	0	10	0.545455
10	AI	2	2	4	0.4
11	AJ	0	0	0	0
12	AK	5	1	3	0.714286
13	AL	0	0	0	0
14	AM	0	2	4	0
15	AN	10	2	4	0.769231
16	AO	7	4	3	0.666667
17	AP	0	0	0	0
18	AQ	0	5	0	0
19	AR	0	6	0	0
20	AS	0	1	0	0
21	AT	0	3	0	0
22	all	10	4	0	0.833333

圖 10 為手動調參 2D U-Net 的模型測試成果

在 Scan-level 的預測結果中，模型針對 20 位病患成功辨識出所有實際患有闌尾炎的案例（FN＝0），僅將 4 位無病者（AQ、AR、AS、AT）錯誤標記為陽性，造成偽陽性（FP）數量為 4。整體 F1-score 達 0.833，顯示模型在病患層級的分類能力具備良好準確性與實用性。

然而，進一步觀察 Slice-level 的分割結果，則可見模型在病灶區域定位方面仍有不足。例如 AB 與 AM 兩位病患雖整體被正確判定為患病，但其所有預測切片均未與真實病灶位置重疊（TP＝0），導致其 slice-level F1-score 為 0。這顯示模型雖具備辨識患病個案的能力，但對病灶實際區域的掌握仍顯不足。另一方面，模型在部分個案中表現良好，例如 AE、AF、AK、AN、AO 五位病患，其 slice-level F1-score 皆高於 0.66，其中 AN 更達 0.769，代表模型在這些病患中能夠正確預測大部分病灶切片，具備穩定的區域辨識能力。亦有表現中等偏低的案例，例如 AA、AH、AI 三位病患，F1-score 落於 0.28 至 0.54 間，主要是因

為 FN 值偏高，顯示模型無法有效捕捉大多數病灶區塊，仍有偵測遺漏的問題。

綜合而言，模型在 Scan-level 的分類效能已具實用價值，但在 Slice-level 的病灶定位準確性與穩定性上仍顯不足。F1-score 的個體差異亦顯示模型泛化能力尚未穩定，可能與訓練資料數量有限、個體影像特性差異過大，或模型於開發過程中過度針對自我測試集進行優化有關，進而導致在外部資料上的泛化能力下降。未來可透過資料擴增、多樣性樣本引入與後處理機制等方式進行改進，以進一步提升模型於病灶區域辨識與臨床應用上的實用性與可靠性。

（二）nnU-Net

	A	B	C	D	E
1	id	TP	FP	FN	f1score
2	AA	4	33	6	0.170213
3	AB	4	29	0	0.216216
4	AC	0	61	0	0
5	AD	0	76	0	0
6	AE	3	44	5	0.109091
7	AF	12	40	0	0.375
8	AG	0	73	0	0
9	AH	9	53	7	0.230769
10	AI	5	60	1	0.140845
11	AJ	0	65	0	0
12	AK	3	47	5	0.103448
13	AL	0	50	0	0
14	AM	4	32	0	0.2
15	AN	10	39	4	0.31746
16	AO	10	48	0	0.294118
17	AP	0	58	0	0
18	AQ	0	53	0	0
19	AR	0	59	0	0
20	AS	0	36	0	0
21	AT	0	45	0	0
22	all	10	10	0	0.666667

圖 11 為 nnU-Net 的模型測試成果

nnU-Net 模型在 scan level 上表現尚可（F1-score 為 0.666），代表對於整體是否有闌尾炎的判斷具有一定準確性。然而在 slice level 的細節表現上則顯得不穩定，不同個體的 TP、FP、FN 落差明顯，像是 AT、AF、AN 等病人達到相對較高的 F1-score，顯示模型能準確偵測出其發炎區域；但也有許多樣本（如 AC、AD、AG 等）完全無法偵測出發炎

（TP = 0），導致 F1-score 為 0。這些現象可能與模型尚未完全收斂、訓練資料數量不足，或個體之間的生理解剖差異有關。因此，儘管整體分類結果具一定參考價值，模型仍需進一步優化，以提升在細部影像層級的偵測穩定性與泛化能力。

伍、討論

一、「2D U-Net」中研究流程的潛在設計問題

在本研究的 2D U-Net 模型訓練過程中，透過多次訓練與結果觀察，我們意識到流程中存在一些潛在的設計問題，可能影響模型的穩定性與泛化能力。以下針對五個主要面向進行討論與反思：

（一）正規化與資料擴增的先後順序

我們觀察到，在有進行資料擴增（如仿射變換）時，若採用 /4095 作為正規化方式，其模型表現反而不如使用 mat2gray。分析後推測，若影像尚未經正規化就直接進行幾何變形，會使原始灰階強度在插值與裁切過程中失真，導致擴增後影像亮度異常或灰階比例錯亂，進而影響模型學習的穩定性。

（二）資料切分方式對評估穩定性的影響

本研究訓練與驗證資料劃分採固定比例（8：2）隨機切分，雖操作簡便，但在資料量有限的情況下，驗證集樣本的選擇可能對模型效能評估造成顯著影響。特別是當驗證集樣本分布與整體資料略有偏差時，可能導致評估結果缺乏代表性，進而影響模型表現的判斷準確性。

（三）資料擴增與資料切分的順序錯誤

在原始流程中我們先進行資料擴增再劃分訓練與驗證集。這樣的順序可能導致同一張影像的變形版本同時出現在訓練與驗證集中，形成資訊洩漏，進而導致驗證表現過於樂觀，無法真實反映模型在未見資料上的泛化能力。

（四）模型後處理的改進方向：引入條件隨機場、位置約束過濾機制

本研究所採用的 2D U-Net 架構會對每張切片進行獨立的影像分割，並未考慮影像在整體 3D 體積中的相對位置或相鄰切片之間的結構連續性。此限制導致部分預測結果出現位置不連貫、分散，甚至落於非臨床合理區域（如腹部上方）等現象。

為了提升空間連貫性與生理解剖邏輯的合理性，未來可導入**條件隨機場（Conditional Random Field, CRF）**作為後處理策略。CRF 能根據影像中相鄰像素／切片間的相似度與距離關係，優化初步分割結果，使其在語意與形態上更為一致，並減少偽陽性遮罩的產生。

此外，亦可引入**位置約束過濾機制**，根據闌尾炎通常出現在腹部下段的特性，僅保留切片序列後半部的預測結果，進一步排除空間上不合常理的預測錯誤。

（五）模型對自我測試集的過度優化風險

本研究雖採用自建測試集作為最終評估依據，且該資料未參與模型訓練，具一定的獨立性，但實驗過程中曾多次根據測試集結果進行後處理策略與模型參數的調整。此種反覆參照測試集進行優化的方式，可能導致模型學習偏向該特定資料分布，進而出現過度擬合現象，降低其在其他外部資料上的泛化能力。這也可能是造成 F1-score 個體落差顯著的原因之一。

（六）後續修正與未來建議

若時間與資源允許重新設計實驗流程，建議可優先採取以下改進方向：

1. 正規化與資料擴增順序修正：影像讀取階段即進行正規化處理，再進行資料擴增的操作，以確保灰階資訊保持一致性並提高模型效能；

2. 資料劃分：導入交叉驗證，例如使用 K-fold Cross-validation，使所有資料皆能輪流作為驗證樣本，提升模型效能評估的可信度；

3. 優化擴增流程：劃分完訓練與驗證集之後，再針對訓練集中有標註病灶的樣本進行擴增，並避免對驗證集做任何資料增強，以維持模型評估的公平性。也可以搭配統計量分析（如平均值、標準差、熵值、直

方圖）進一步驗證擴增後資料與原始資料的特性一致性，避免模型偏離真實分布；

4. 後處理優化：加入 CRF 與位置約束機制補足 2D 架構的空間感知限制，強化分割結果的連續性與臨床合理性。

透過上述流程優化與技術補強，未來可望建構出更具穩定性、泛化能力與臨床實用價值的醫學影像分割模型。

二、「nnU-Net」中研究流程的潛在設計問題

在本研究的 nnU-Net 訓練過程中，希望藉由其自動化配置與高效能的特性提升醫學影像分割表現，然而在實作過程中仍遭遇多項限制，影響了模型效能與研究結果的穩定性。以下整理兩項關鍵問題與對應的未來修正方向：

（一）電腦效能不足導致訓練參數調整

由於本研究執行環境的 GPU 記憶體有限，無法支援 nnU-Net v2 原建議的 batch size 設定。因此在訓練階段，將 batch size 手動調整為 1，以避免記憶體溢位（out of memory）錯誤。雖然這使訓練得以進行，但也可能導致梯度更新不穩定、模型收斂速度變慢，進而影響最終模型的準確性與泛化能力。

（二）小資料集限制了標準訓練流程的完整性

nnU-Net 預設支援以 k-fold 的方式切分資料集、訓練多個模型後自動選出最佳版本。然而由於本研究的資料集樣本數較小，若再依照標準流程分割將導致單一訓練集不足，模型更容易過擬合。因此本研究僅以單一分割方式訓練模型，未能善用 nnU-Net 的完整自動化流程。這使得模型缺乏對不同資料切分條件的穩定測試，也難以保證當前結果為最優版本。

（三）後續修正與未來建議

為提升後續研究的可靠性與模型效能，建議可從以下幾方面進行修正：

1. 硬體方面，可嘗試使用雲端平台（如 Google Colab Pro、AWS EC2 GPU）或校內具備高效能 GPU 的伺服器，以支援更合適的 batch size 與完整資料切分流程；

2. 資料方面，應考慮擴增資料集數量，或進行資料增強（data augmentation）以模擬更多變異情況，使模型更具泛化能力；

3. 訓練策略方面，未來可重新啟用 nnU-Net 的 cross-validation 功能，自動選出最佳模型版本，並透過學習率調整、early stopping 等技巧穩定模型訓練流程。

三、「3D U-Net」中研究流程的潛在設計問題

（一）在處理資料時資料型態的問題

在處理 mask 的 one-hot 編碼時若沒有確保每一張 mask 都轉換為 single，後續在轉換為 4D 或 5D array 時會發生「Brace indexing is not supported」的錯誤。若處理後仍然是 cell array 就需要透過 cat(4, …) 或 cat(5, …) 轉換為 array，然後再傳入到 arrayDatastore 中。

（二）未來建議

若時間允許，後續還需要處理測試集的資料讀取與 F1 score 的評估，但是 3D 模型有一個重大優勢，就是能夠保留每張影像在 Z 軸上的空間連貫性。我們希望透過此處理方式，更完整地捕捉病灶的空間資訊，以獲得更準確且穩定的分割結果。

陸、結論

本研究顯示，在小型醫療影像資料集上，透過手動調參與流程優化所建立的 2D U-Net 模型，能以較少資源取得穩定成效，展現高度的彈性與實作效率。然而，其效能高度依賴於設計過程中每一環節（如資料擴增、正規化策略、後處理機制等）的合理性與嚴謹性，需投入大量人工驗證與經驗判斷，對使用者專業度與流程掌控力具有一定要求。

相較之下，nnU-Net 具備高度自動化的建模能力與強大架構設計，能在中大型資料集中快速收斂，具備良好的泛化潛力，但其前期準備較為繁複，且對計算資源要求較高，短期內在資料規模有限或資源受限情境下較難發揮完整效益。

未來若欲公允比較 U-Net 與 nnU-Net 於醫療影像分割任務之表現，宜建立統一的實驗設計與控制條件。目前雖兩者使用相同資料集進行訓練，但在前處理、後處理、資料增強策略等方面並無一致規範，實際上評估結果更反映整體實作流程的優化程度，而非模型架構本身的效能差異。為確保比較的科學性與代表性，後續研究應排除非架構因素之干擾，建立可重複、可驗證的標準化流程，以呈現兩者在相同條件下之真實性能表現。

柒、參考資料

[1] Baştuğ, B. T., Güneri, G., Yıldırım, M. S., Çorbacı, K., & Dandıl, E. (2024). Fully Automated Detection of the Appendix Using U-Net Deep Learning Architecture in CT Scans. *Journal of Clinical Medicine*, *13*(19), 5893. https://doi.org/10.3390/jcm13195893

[2] Ronneberger, Olaf, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation." Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, 2015, pp. 234–41. Crossref, https://doi.org/10.1007/978-3-319-24574-4_28.

[3] Isensee, Fabian, et al. "Abstract: NnU-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation." Bildverarbeitung Für Die Medizin 2019, Springer Fachmedien Wiesbaden, 2019, pp. 22–22. Crossref, https://doi.org/10.1007/978-3-658-25326-4_7.

以下為網路資料：

[4] Kaggle 的 AOCR 2024 AI Challenge 競賽
https://www.kaggle.com/competitions/aocr2024/overview

[5] Chia-Feng Lu. (2022, May 19). *[2021.05.19 Lesson13-Session3] Image Segmentation Using U-Net.* https://youtu.be/drF4L0jXFhY?si=IQS1DbD32HZwbfnE

[6]瞭解醫學影像分割中的評估指標
https://medium.com/@nghihuynh_37300/understanding-evaluation-metrics-in-medical-image-segmentation-d289a373a3f

[7]nnU-net 網路簡介資料
https://medium.com/miccai-educational-initiative/nnu-net-the-no-new-unet-for-automatic-segmentation-8d655f3f6d2a