中華民國第 61 屆中小學科學展覽會

作品說明書

高級中等學校組　動物與醫學科

佳作

052003

肝臟「IN」了沒？以人工智慧評估肝纖維化

學校名稱：臺中市立臺中女子高級中等學校

作者：	指導老師：
高二 陳珂敏 高二 吳鎔均	陳菀菱

關鍵詞：人工智慧、肝硬化、肝纖維化

摘要

嚴重的肝纖維化將導致臺灣十大死因之一的肝硬化，因此肝纖維化的程度評估對肝病的診治及研究極其重要。肝臟病理切片利用特殊染色如馬森三色染色法（Masson’s trichrome stain），可以清楚呈現不同程度而不等量的膠原纖維。臨床上最常使用 Ishak fibrosis score 評估肝臟切片纖維化的程度，可分類為 7 級，從 0 級的正常肝臟、輕微纖維化的 1 級至最嚴重的 6 級(肝硬化)，然而病理醫師常因缺乏第二意見而有診斷不一致的問題。本研究將肝纖維化的檢體照片依照 Ishak Score 分級，訓練 Google 開發的 Teachable Machine (GTM)人工智慧深度學習軟體。再以獨立照片反覆測試人工智慧模型，檢驗病理醫師與模型判讀的差異，加以調整改善。期望本模型可以輔助臨床診斷，在醫療資源相對匱乏的地區，也能簡便地做出即時精確的診斷。

壹、研究動機

根據衛福部統計處的資料顯示，108 年度國人十大死因中，慢性肝病及肝硬化位居第十，每年有 1400 位以上的患者因為慢性肝病或是肝硬化而死亡。肝硬化是末期肝病時，在細胞組織之間呈現嚴重肝纖維化的表現，因此肝纖維化的程度評估對肝病的診治及研究極其重要。但臨床上做肝纖維化程度的評分時，不同病理醫師的判讀結果常存在許多差異。近年來，人工智慧的發展開始介入醫療影像的判讀，我們想針對肝臟組織切片、藉由人工智慧系統輔助分析纖維化的程度，提升肝硬化分級準確度，亦或是提供診斷醫師第二意見。期望能減少病理醫師分析的負擔，更讓偏鄉及偏遠國家能在缺乏醫療資源的困境下，直接使用此影像辨識系統快速評斷。期望不論醫療資源是否充足，都能以高效率的方式提升全民健康福祉。

貳、研究目的

一、利用人工智慧可以準確判斷出肝纖維化與正常肝的檢體照片差異。

二、利用人工智慧可以進一步判斷出肝纖維化的等級。

參、研究設備及器材

一、Google Teachable Machine

二、筆記型電腦

三、病理玻片顯微影像照相系統

肆、研究過程或方法

一、肝纖維化的認識

(一) 肝纖維化

肝纖維化(Fibrosis)起因於肝臟反覆受傷、修復過程中用締結組織取代正常的肝實質組織，導致細胞外基質(ECM)中膠原成分的過度沉積。許多慢性肝臟疾病均可引起肝纖維化，其病因主要分為感染性 (如 B 型、C 型和 D 型病毒性肝炎)、化學代謝缺陷(如慢性酒精性肝病、慢性藥物性肝病、脂肪肝)、先天性代謝缺陷 (如威爾森氏病、鐵質沉積血色病）及自身免疫性肝炎（如自體免疫肝炎、原發性膽汁性肝硬化）等。肝纖維化持續發展的下個階段是肝硬化，會出現其他併發症，如腹水、食道靜脈曲張、肝性腦病變、出血傾向，甚至肝癌等疾病的風險也越高，病人預期存活時間也越短。因此，近年來肝纖維化被認為是肝病治療療效與預後評估的重要決定因子之一。

(二) 肝纖維化的評估

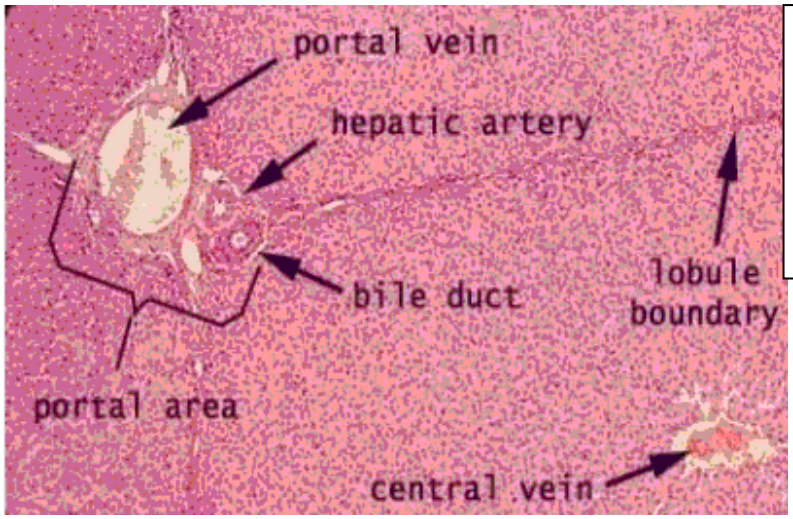
肝纖維化程度和肝臟狀態息息相關，準確診斷纖維化程度有助預後和療程計畫，避免惡化為肝硬化或肝癌。肝纖維化的評估方式有血清學檢測、影像學檢查如腹部超音波，及肝切片判讀等，必定存在誤差，也各有其缺點與限制(表 1)。

本研究是針對肝組織切片，以 Ishak 評分系統判讀肝纖維化程度。肝臟切片會依據肝硬化的嚴重程度，呈現不等程度的纖維化情形。肝纖維化在組織學上會從門脈系統(portal area)開始沉積膠原纖維，再由門脈系統往外延伸纖維化組織，原本平滑的肝臟表面被纖維組織取代，呈現凹凸不平的結節。肝臟切片組織的位置說明如圖一，擷取影像前必須先認識肝組織門脈系統所在位置。組織學上輕微的肝纖維化用常規的蘇木紫伊紅染色會比較不好判讀，借助特殊染色可以染出纖維化區域沉積的膠原蛋白，馬森三色染色法（Masson’ s trichrome stain）是病理醫師常用來判定肝臟纖維化程度的特殊染色，可以把纖維化的區域染成藍色。

國內主要使用的評分系統有 Ishak 分為 F0 ~ F6 共七級，和 METAVIR 分為 F0 ~F4 共五級(另外搭配 Activity Score 共四級)，目前國際間的研究較常使用 Ishak 分級。兩者從檢體有無肝門纖維化(Portal fibrosis)、中隔(septa)、橋狀纖維化(bridging)、肝結節(nodules)等結構進行分級。圖二敘述 Ishak 分級的組織學特徵，並與另一評估肝臟纖維化程度的 METAVIR 作分級比對，圖三是肝臟切片電子影像數位化後，依照 Ishak fibrosis score 1 至 6 分級。

表 1 : 不同肝纖維化評估方式的缺點與限制
(表 1 資料來源:自行繪製)

評估方式	缺點與限制
血清學檢測	單一血清學標記檢查 : 敏感性、專一性、關聯性不高 多種項目之血清學標記檢查 : 步驟繁瑣、成本高昂
腹部超音波	腹水存在、肋間空間狹窄、過度肥胖等情況不適用
肝切片	罕見但嚴重併發症(0.5%)、醫師判讀差異、取樣誤差、侵襲性檢查



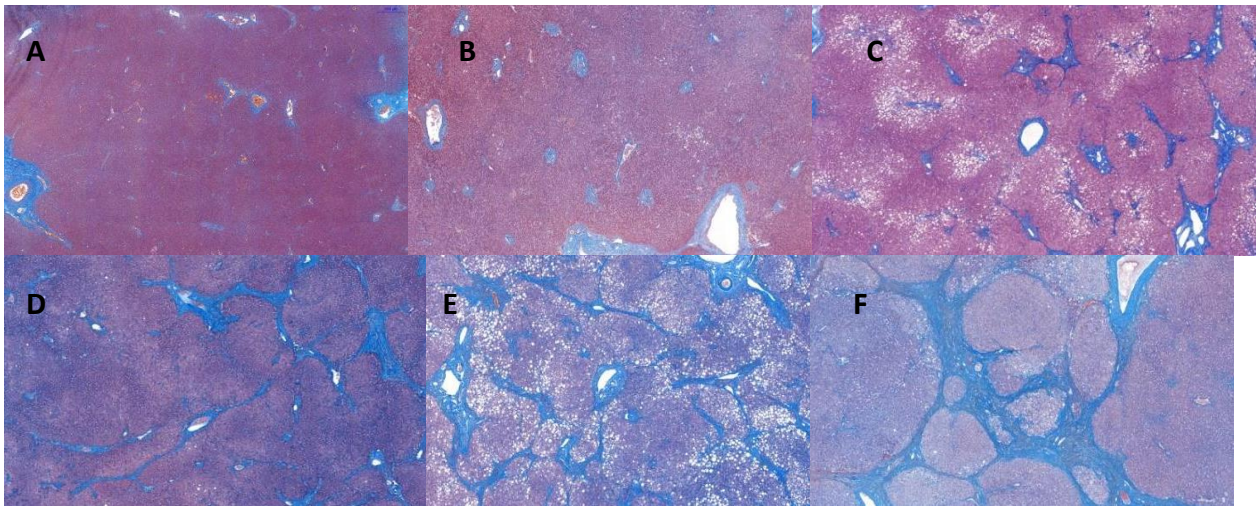
Portal area	門脈系統
Portal vein	肝門靜脈
Hepatic artery	肝動脈
Bile duct	小膽管
Central vein	中央靜脈

圖一 : 肝臟組織切片的部位說明。組織影像擷取必須注意門脈系統，因肝纖維化最先由門脈系統往外擴張。
(圖一來源: http://www.siumed.edu/~dking2/erg/GI152b.htm)

Appearance	Ishak stage: categorical description	Ishak	Metavir
	No fibrosis (normal)	0	F0
	Fibrosis expansion of some portal areas ± short fibrous septa	1	F1
	Fibrosis expansion of portal areas ± short fibrous septa	2	F2
	Fibrosis expansion of most portal areas with occasional portal to portal (P-P) bridging	3	
	Fibrosis expansion of portal areas with marked portal to portal (P-P) bridging as well as portal to central (P-C)	4	F3
	Marked bridging (P-P and / or P-C) with occasional nodules (incomplete cirrhosis)	5	
	Cirrhosis, probable or definite	6	F4

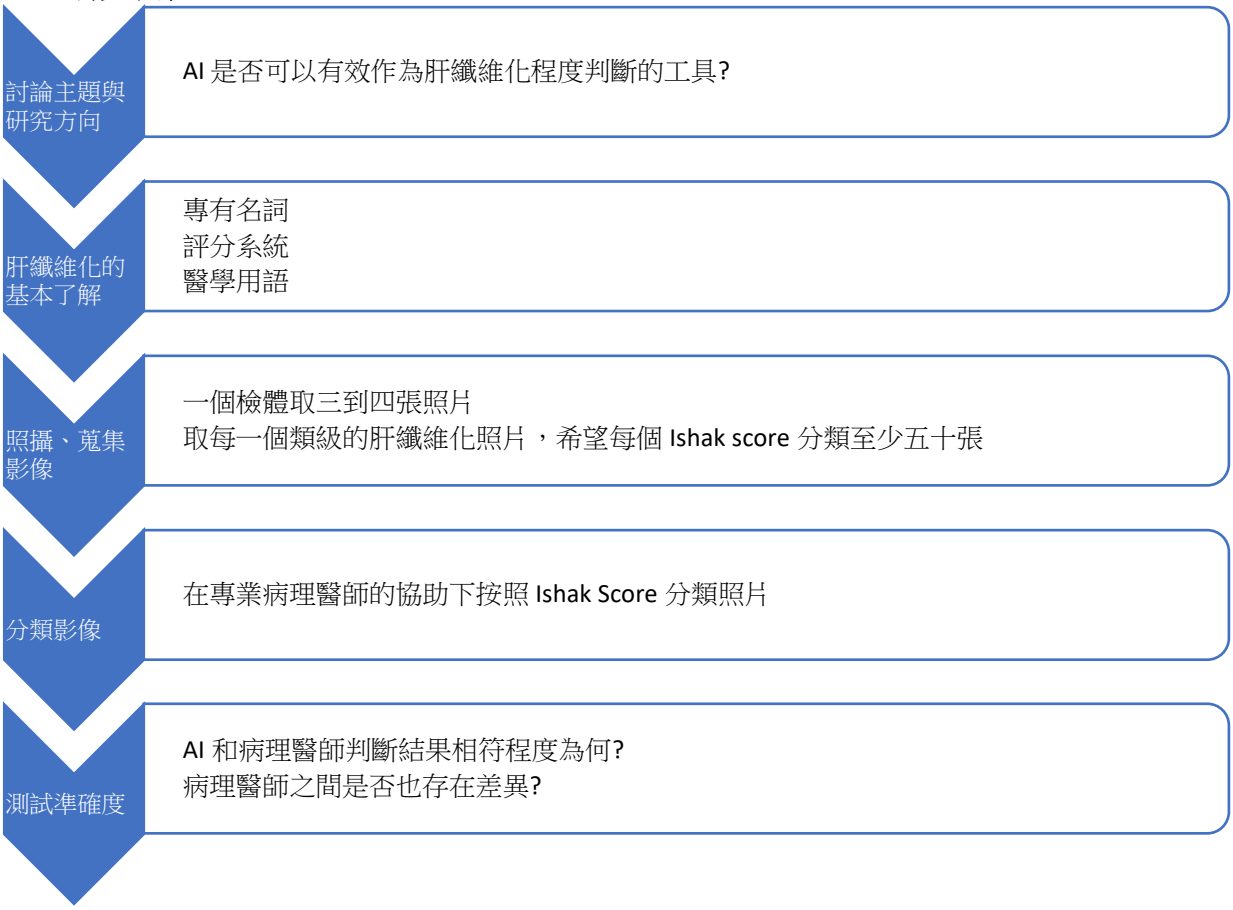
肝臟切片纖維化的組織分級特徵
Ishak 0 正常肝臟
Ishak 1 少部分的肝門系統擴張
Ishak 2 大部分的肝門系統擴張
Ishak 3 大部分的肝門系統擴張合併出現輕微橋狀纖維化
Ishak 4 肝門系統擴張合併明顯的橋狀纖維化
Ishak 5 橋狀纖維化合併結節
Ishak 6 可能或是確定的肝硬化

圖二 : Ishak 和 METAVIR 對同一切片的分級比對及圖片說明。
(圖二來源: R A Standish, E Cholongitas, A Dhillon, A K Burroughs,A P Dhillon. AN APPRAISAL OF THE HISTOPATHOLOGICAL ASSESSMENT OF LIVER FIBROSIS. Gut 2006;55:569 – 578. doi: 10.1136/gut.2005.084475)

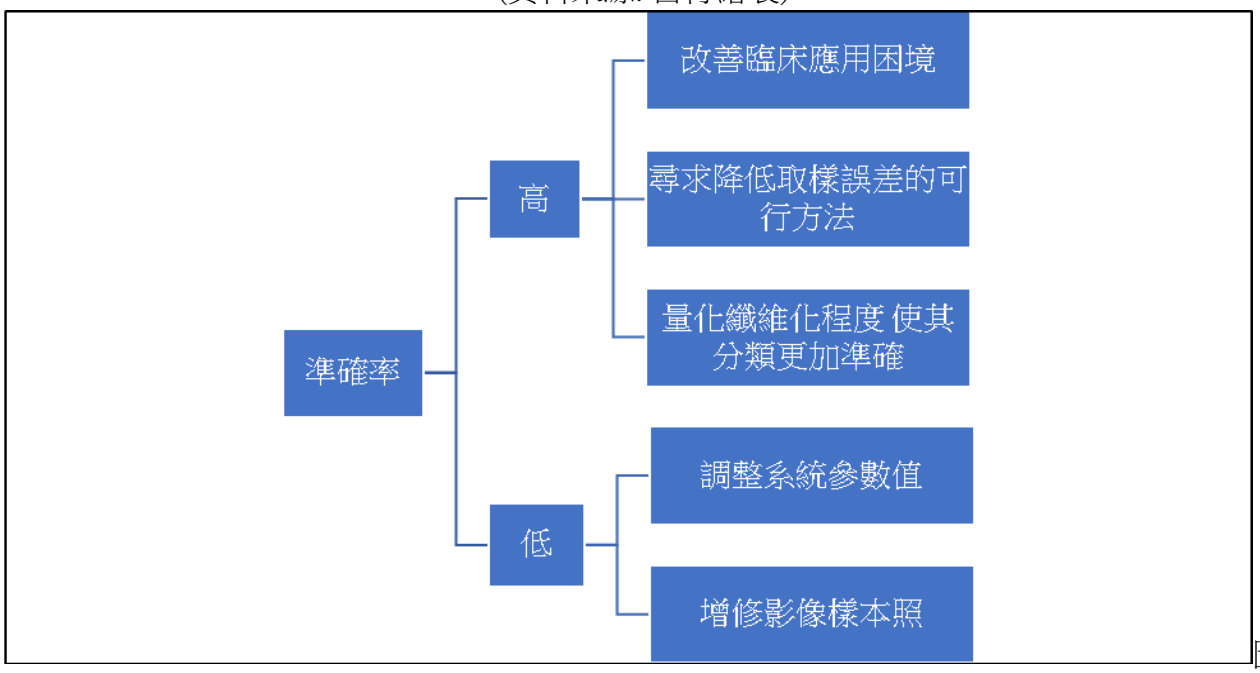


圖三 : A 至 F 分別顯示 Ishak 1 至 6 的馬森三色染色法圖片
(圖三資料來源: 研究者自行拍攝。)

二、研究流程



圖四 : 研究流程圖
(資料來源: 自行繪製)



圖五 : 研究流程圖
(資料來源: 自行繪製)

三、研究方法

(1) 檢體來自臺中榮民總醫院的肝臟病理組織切片，本研究已通過臺中榮民總醫院人體研究倫理審查委員會審查 (編號：CE2045B)。

(2) 先利用病理玻片顯微影像照相系統截取肝組織切片的影像，在擷取完檢體照片後，將照片分成兩部分。

(三) 第一部分是訓練 Google 開發的 Teachable Machine (GTM)人工智慧深度學習軟體的影像，依照病理醫師判讀的分級結果將照片分類到所屬的 Ishak 纖維化類別裡面，並依照分類資料夾上傳至 GTM 進行訓練。訓練用影像包括 72 張 Ishak score 1、112 張 Ishak score 2、117 張 Ishak score 3、52 張 Ishak score 4、41 張 Ishak score 5、及 63 張 Ishak score 6 的肝臟切片影像。

(四) 第二部分是測試用影像，必須與訓練用影像是不同的檔案，測試用影像會請兩位病理醫師各自獨立進行判讀，再針對纖維化程度判別不同的影像做討論以取得共識，同時將測試用影像上傳至 GTM 做判讀。測試用影像則包括 167 張不同等級的肝纖維化切片影像。由於正常肝臟不易取得，訓練用影像及測試用影像皆不包含 Ishak score 0 的正常肝組織切片。

(五) 之後比對 GTM 判讀的結果與病理醫師共識後的分級共識。我們請兩位病理醫師分別對影像進行分級，再請兩位醫師就判別有出入的影像進行討論，獲得共識後再分級。

(六) 第一次的結果，測試用影像病理醫師間的共識度不佳，在兩位病理醫師討論後，決定再將訓練用影像依據共識討論的經驗重新分類，測試用影像則是打亂次序後再分別由兩位病理醫師重新判讀、再共識一次，並重新上傳至新訓練的 GTM 做判讀。

伍、研究結果

一、建立病理醫師間診斷的共識

表 2-1、3-1 是兩次測試病理醫師共識肝纖維化分級診斷和 GTM 判讀的差異，表 2-2、3-2 是兩位病理醫師單獨判讀時的分級診斷差異。第一次測試病理醫師間差異大於或等於 2 級的檢體佔了 21%，依據重新分類訓練的第二次測試，病理醫師間差異大於或等於 2 級的檢體只佔了 7.2%，和第一次相比也沒有差異三級以上的檢體。第一次測試 GTM 判讀結果與病理醫師診斷共識差異大於或等於 2 級的檢體佔 13.8%，第二次測試 GTM 判讀結果與病理醫師診斷共識差異大於或等於 2 級的檢體只佔 9%，而且沒有差異四級的檢體，差異 2 級的檢體也大幅下降。GTM 的表現仍必須依賴良好並且一致性佳的病裡診斷分類，病理醫師之間的差異性較少時，GTM 與共識的差異也跟著變少。

表 2-1 :第一次共識與 GTM 的判斷差異
(表 2-1 資料來源: 研究者整理)

病理醫師共識與 GTM 判讀差異級數	照片數量	比例
0	72	43.1%
1	72	43.1%
2	17	10.2%
3	5	3%
4	1	0.6%

表 2-2 :第一次分類醫師間的診斷差異
(表 2-2 資料來源: 研究者整理)

兩位醫師間的判讀差異級數	照片數量	比例
0	67	40.1%
1	65	39%
2	31	18.6%
3	4	2.3%

表 3-1 :第二次共識與 GTM 的判斷差異
(表 3-1 資料來源: 研究者整理)

病理醫師共識與 GTM 判讀差異級數	照片數量	比例
0	79	47.3%
1	73	43.7%
2	10	6%
3	5	3%

表 3-2:第二次分類醫師間的診斷差異
(表 3-2 資料來源: 研究者整理)

兩位醫師間的 判讀差異級數	照片數量	比例
0	78	46.7%
1	77	46.1%
2	12	7.2%

二、以人工智慧辨別肝纖維化與正常肝的影像

由於正常肝臟很少有切片的檢體，所以我們將 Ishak Score 分數 1(輕微)皆歸類在接近正常肝臟的切片類別裡，Ishak Score 分數 2~6 歸類在非輕微肝纖維化。第二次測試的表現病理醫師間的差異較少，GTM 的分類也比第一次測試好，所以，以第二次測試的結果當此項目的的討論基礎。在 167 張做為測試的照片裡，病理醫師共識分類 Ishak 分數為 1 分的有 11 張、2 分有 38 張、3 分有 39 張、4 分有 28 張、5 分有 19 張、6 分有 32 張。

在 11 張 Ishak 分數為 1 的影像中，有 2 張 GTM 判讀為 1 分，有 6 張判讀成 2 分，有 3 張判讀成 3 分。 在剩餘 156 張非輕微纖維化的影像中，有 7 張影像被 GTM 判讀成 1 分的輕微纖維化，這 7 張影像病理醫師的共識診斷全都是 2 分。纖維化影像誤判的比例是 9.6% (16/167)，正確判斷出有纖維化影像的比例則是 90.4%。在鑑別纖維化與非纖維化影像 GTM 有很高的比例能做出正確診斷 。

三、以人工智慧判斷出肝纖維化的等級

為了讓人工智慧可以判斷肝纖維化的等級，因此將分類的組別進一步分為六組 (Ishak Class 1~6)。類別之間的差異變小，對人工智慧 GTM 來說挑戰更加困難。使用第一次病理醫師們的共識測試時，AI 答對率只有 43.10%，超過一半分到錯誤的類別裡面，其中甚至有誤判超過四級(5 分判成 1 分)的照片 。為解決此問題，兩位病理醫師進行第二次的討論，重新分類較具爭議的訓練用影像。在第二次人工智慧 GTM 模型的表現較佳，除了答對率上升之外，第二次也沒有判斷差異超出四分的照片 。所以，這個問題仍是以第二次的訓練判讀結果來做討論。

表 3-1 顯示總共有 79 張影像 GTM 判讀結果和病理醫師診斷共識一致，佔 47.3%，73 張影像 GTM 判讀結果和病理醫師診斷共識只差一分，佔 43.7%，為輕度誤判。而兩者差異兩分的影像有 10 張，佔 6%，為中度誤判。差異 3 分的影像有 5 張，佔 3%，屬於重度誤判。

表 4 是病理醫師診斷共識與 GTM 判讀結果的比對。從表 4 的 GTM 判讀結果來看，GTM 判讀結果為 1 分、4 分、5 分、和 6 分的影像與病理醫師的共識診斷差異幅度較小，大多只差異 1 分或是 2 分以內。但是 GTM 判讀結果為 2 分和 3 分的影像，與病理醫師共識診斷的差異幅度就非常高。而從病理醫師診斷共識分類來看，共識 5 分的圖片在 GTM 判讀的結果也有比較大的差異。另外，和診斷共識相比，GTM 也呈現較易輕判

的情況。

	GTM 與共識相同
	GTM 較共識嚴重一級
	GTM 較共識嚴重二級
	GTM 較共識輕微一級
	GTM 較共識輕微二級
	GTM 較共識輕微三級

表 4 : 167 張測試用照片病理醫師共識診斷分級與 GTM (AI) 判讀結果比對分析
(表 4 資料來源: 研究者整理)

	共識 1	共識 2	共識 3	共識 4	共識 5	共識 6
GTM 1	2	7	0	0	0	0
GTM 2	6	19	23	2	1	0
GTM 3	3	12	15	16	3	4
GTM 4	0	0	0	10	5	1
GTM 5	0	0	1	0	6	0
GTM 6	0	0	0	0	4	27

表 5 : 病理醫師 A 和診斷共識的差異。
(表 5 資料來源: 研究者整理)

	共識 1	共識 2	共識 3	共識 4	共識 5	共識 6
病理醫師 A 1	11	15	0	0	0	0
病理醫師 A 2	0	23	16	0	0	0
病理醫師 A 3	0	0	23	1	0	0
病理醫師 A 4	0	0	0	26	2	0
病理醫師 A 5	0	0	0	1	17	3
病理醫師 A 6	0	0	0	0	0	29

表 6 : 病理醫師 B 和診斷共識的差異。
(表 6 資料來源: 研究者整理)

	共識 1	共識 2	共識 3	共識 4	共識 5	共識 6
病理醫師 A 1	7	0	0	0	0	0
病理醫師 A 2	4	16	0	0	0	0
病理醫師 A 3	0	22	33	0	0	0
病理醫師 A 4	0	0	6	12	0	0
病理醫師 A 5	0	0	0	16	7	0
病理醫師 A 6	0	0	0	0	12	32

表 5 和表 6 分別代表病理醫師 A 和病理醫師 B 與診斷共識的差異。病理醫師 A 傾向於稍微輕判纖維化，病理醫師 B 傾向於診斷得嚴重些。但是，不管是病理醫師 A 或病理醫師 B 和診斷共識的差距，在任一個纖維化級數都不會大於 1。這代表兩位病理醫

師之間的診斷和共識的一致性仍是比 GTM 還要來得好，完全沒有嚴重誤判的情形。

四、判讀結果的交叉分析

我們針對病理醫師共識的標準答案(Consensus)與人工智慧(AI)比較的分析模型，做判讀結果的交差分析，得到表 7 的結果。資料分析定義如下:

(TP) True Positive：模型準確預測情況為真
(TN) True Negative：模型準確預測情況為假
(FP) Flase Positive：模型錯誤預測情況為真
(FN) Flase Negative：模型錯誤預測情況為假
Precision=TP/(TP+FP)
Recall=TP/(TP+FN)
f1-score=2*Precision *Recall /(Recall +Precision)

表 7 : Precision、Recall 和 f1-score 在 Consensus and AI 分析模型的結果
(表 7 資料來源: 研究者整理)

	Precision	Recall	f1-score
class 1	0.22	0.18	0.2
class 2	0.37	0.5	0.43
class 3	0.28	0.38	0.33
class 4	0.62	0.36	0.45
class 5	0.86	0.32	0.46
class 6	0.87	0.84	0.86

透過表 7 的三項指標能鑑別我們設計的模型在診斷各等級的好壞。Precision（精確率）代表模型預測結果屬於某等級時，實際的「精準度」是多少，Recall（召回率）代表實際屬於某等級的情況下，模型能「預測回多少」該等級，f1-score 則是精確率和召回率的調和均值。從表 7 可得知，我們的模型在較高等級 (較嚴重纖維化) 的各項指標值均較高，即診斷成效較佳。而臨床上極忌諱重度誤判為輕微纖維化，也就更在乎較高等級的預測效果如何，我們的模型能符合這點要求。

陸、討論

一、 導致人工智慧誤判之因素

(一) 病理醫師之間的判斷差異

由於 Ishak Score 並沒有確切的數值作為評斷標準，而只是依照肝纖維化嚴重程度進行分級，不同病理醫師在肝纖維化程度的模糊地帶(Ishak Score 2~3、3~4)很容易會有不一樣的答案。一開始病理醫師間的判斷歧異較大，因此訓練出來的 GTM 模組對於診斷的分類表現也較不準確。人工智慧若要在臨床上被應用，嚴重的誤判便要儘量減少，以免低估了疾病的嚴重程度，無法為病人及早診斷及治療。然而，這也必須仰賴病理醫師間能有高度判讀共識。

(二) 訓練模組的參數

在 GTM 的模型訓練(Model Training)裡面，有兩個會影響模型表現的參數: 批大小(Batch Size) 與訓練回合數(Epochs)。

1. 批大小 (Batch Size)

批大小的數字便是樣本總數除以總共有幾批 (Eg. 100 張樣本照/5 批=20 批大小)，適當的批大小有助於模型的學習效率。

2. 訓練回合數 (Epochs)

訓練回合數則是全部批數總共被人工智慧學習幾次 (Eg. 5 批樣本照全部被 AI 學習過一次時便是 1 個 Epochs)。訓練回合次數越大，所花的時間就越多。一般來說，較高的訓練回合次數有助於 AI 在判斷上的表現，因為這代表資料及被 AI 學習的次數越高。但是過高的回合次數則會使模型產生過度擬合(Overfitting)的情況。過適的定義便是 AI 在學習的資料集中判斷表現良好，但是在實際測試時準確度便下降，其中原因便是因為 AI 過度學習資料集，使 AI 將我們不想列為參數的特徵一併學了起來。反之，過低的訓練回合次數則會造成低度擬合(Underfitting)的情況，使模型無法有效擷取資料特徵。

二、改善人工智慧模型表現的方法

(一) 縮減病理醫師之判斷差異

當兩位醫師初次進行影像評分時，許多答案都有落差。由此可知，不同的專業醫師在相同的檢體前仍有不一樣的評分標準。在共識不高的情況下，GTM 的表現便會下降，因為模糊地帶的照片分級容易出錯，導致誤判率上升。第一次訓練 GTM 的資料集只有被一位病理醫師評分過，這亦可能造成 GTM 有盲點。而在兩位病理醫師重新將有爭議的樣本照進行分類才進行模型訓練，表現的結果較為理想。病理醫師之間共識訓練越多次也有助於訓練模組的一致性。

(二) 簡化照片樣本的分類

因為 Ishak 1-3 的特徵區分極細微，交叉分析的結果也顯示為此模型將來最需要加強改善的地方。之後將把 GTM 判別目標減少，如目前的六組改為較簡單的兩組，分成 Ishak 0-3 級和 4-6 級(advanced fibrosis)，並把樣本分成訓練 AI 組、驗證組，這樣就能評估 AI 訓練的成果。

柒、結論

我們的研究顯示，在適當的訓練之下，人工智慧模型的確可以輔助病理醫師，為肝臟病理組織切片的肝纖維化程度做初步的診斷分級。雖然本人工智慧模型判斷的結果，仍存在著約 3%的顯著誤差率，但是和病理醫師診斷的一致性(相同或是輕微誤差)，可以達到 91%。再加上我們以病理醫師的診斷為標準建立人工智慧模型，相較以往的人工智慧模型，更具有臨床診斷的參考價值。若能以本人工智慧模型為基礎，持續改善各纖維化等級的正確率，將來可能成為病理組織影像診斷的有用工具。

捌、參考資料及其他

衛生福利部統計處 108 年十大死因統計 (民 109 年 9 月 7 日)。檢自 https://dep.mohw.gov.tw/DOS/np-1776-113.html (民 109 年 9 月 19 日)

病理學基礎認識。檢自 https://allentube211.files.wordpress.com/2017/02/ch01.pdf (民 109 年 9 月 29 日)

組織染色法(IHC)方法大全 (民 108 年 9 月 19 日)。檢自 https://www.toolsbiotech.com/news/detail/id/245.html (民 109 年 10 月 18 日)

Google Teachable Machine 訓練模組參數 (民 109 年 4 月 17 日)。檢自 https://www.rs-online.com/designspark/google-teachable-machine-raspberry-pi-4-cn (民 109 年 10 月 25 日)

YC Note : 資料科學技術，如何辨別機器學習模型的好壞？秒懂 Confusion Matrix (2017)。檢自 https://www.ycc.idv.tw/confusion-matrix.html (民 109 年 11 月 11 日)

林其斌 (譯) (2003)。彩色圖文對照系列 : 胃腸與肝臟疾病 (原作者: Richard J. Aspinall, Simon Taylor-Robinson)。台北市：藝軒圖書。(原著出版年：1826)

R A Standish, E Cholongitas, A Dhillon, A K Burroughs,A P Dhillon. (2006). An appraisal of the histopathological assessment of liver fibrosis. *Gut, 55(4),* 569-78.

Daisuke Komura, Shumpei Ishikawa. (2018). Machine Learning Methods for Histopathological Image *Analysis. Computational and Structural Biotechnology Journal, 16,* 34 – 42.

Markos G. Tsipouras, Nikolaos Giannakeas, Alexandros T. Tzallas , Zoe E. Tsianou, Pinelopi Manousou , Andrew Halld , Ioannis Tsoulos , Epameinondas Tsianos. (2017). A methodology for automated CPA extraction using liver biopsy image analysis and machine learning techniques. *Computer Methods and Programs in Biomedicine ,140,* 61 – 68.

【評語】052003

研究目的 一、利用人工智慧可以準確判斷出肝纖維化與正常肝的檢體照片差異。 二、利用人工智慧可以進一步判斷出肝纖維化的等級。實驗設計及執行嚴謹，雖原始構想顯然來自原實驗室或老師提供之想法，但自己對其知識之掌握良好，在 AI 執行上有創意，有初步結果顯示可行性。

1. 此作品的研究目標明確但不夠深入，目前所得的結果尚屬初步階段，對相關研究領域的貢獻度不高。
2. 結合 AI 及醫學影像來預測或判斷疾病的進程，已是目前疾病診斷的趨勢之一，但多採用非侵入性的醫學影像檢查。由目前整體內容評估，新穎性有限。應有文獻回顧的段落，並說明此作品探討的方向與過去研究有哪些不同。
3. 此研究所使用的實驗策略及材料方法不是很明確，例如不知道此研究使用了多少病例，只看到不同等級肝纖維化的切片影像張數(其等於病例數嗎?)。也不知道在這些影像上有多少特徵點被鑑定出來，又哪些特徵點具有辨識能力，所建立的模型包含幾項特徵點。此外，資料分析似乎未使用統計檢定，以確定該分析是否具顯著性差異。除了 GTM 之外，建議也可採用其他的機器學習法並互相比較之。
4. 簡報資料編排不是很恰當，呈現的字數太多，有些圖太小、內容不清晰。
5. 在建立病理醫師間診斷的共識階段，宜增加病理醫師的人數，一同討論形成共識之後， 讓機器學習比較能做出準確的判斷。

作品簡報

肝臟「IN」了沒？以人工智慧評估肝纖維化

科　別:動物與醫學學科

組　別:高中職組

摘要

- 臺灣十大死因之一的肝硬化
- 檢體照片依照Ishak Score分級
- 訓練Google開發的Teachable Machine (GTM)人工智慧深度學習軟體
- 獨立照片反覆測試人工智慧模型

研究目的

- 肝纖維化有等級之分
- 現今臨床的檢體切片為病理醫師以肉眼診斷
- 不同病理醫師判讀存有差異
- 藉由人工智慧的協助減少判讀誤差，並減輕醫師負擔

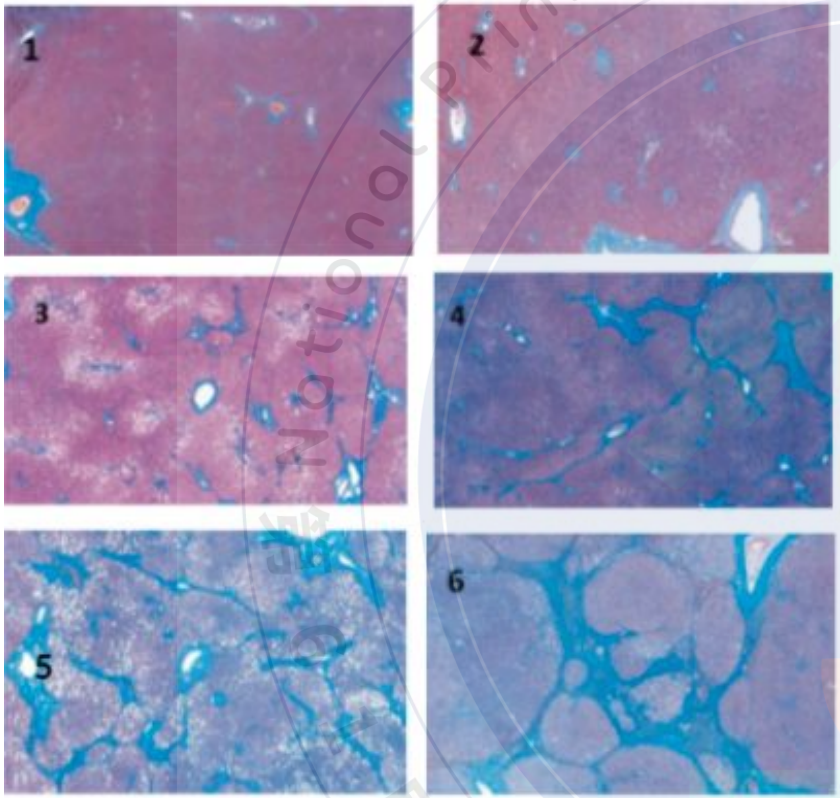
研究方法與過程

- 討論主題及研究方向
- 肝纖維化的基本認識
  - 病因與評估
  - 專有名詞
  - 評分系統
- 切片實作
  - 組織切片學習
- 照攝與蒐集影像
  - 每個Ishak Score分類至少50張
- 在專業病理醫師的指導下學習如何分類檢體照片
- 測試AI準確度
  - AI與病理醫師的相符程度?
  - 病理醫師間是否也存在著差異?

研究方法與過程

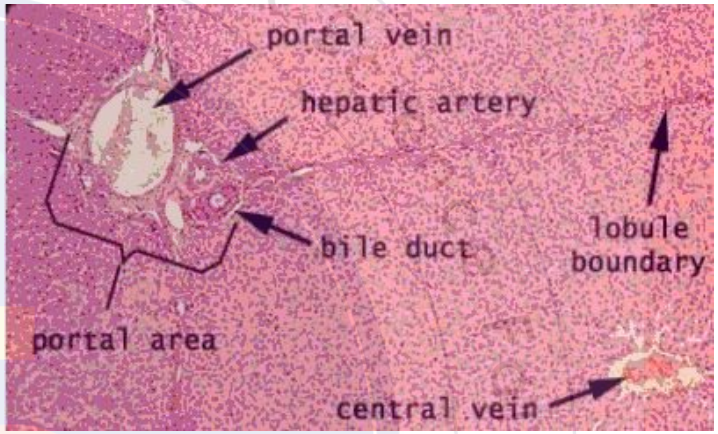
- 先利用病理玻片顯微影像照相系統截取肝組織切片的影像。
- 第一部分是訓練Google開發的Teachable Machine (GTM)人工智慧深度學習軟體的影像，依照病理醫師判讀的分級結果將照片分類到所屬的Ishak纖維化類別裡面，並依照分類資料夾上傳至GTM進行訓練。
- 第二部分是測試用影像，必須與訓練用影像是不同的檔案
  - 會請兩位病理醫師各自獨立進行判讀，再取得共識
  - 將測試用167張不同等級的肝纖維化切片影像上傳至GTM做判讀。
- 比對GTM判讀的結果與病理醫師共識後的分級共識。請兩位病理醫師分別對影像進行分級，再請兩位醫師就判別有出入的影像進行討論，獲得共識後再分級。
- 第一次的結果，測試用影像病理醫師間的共識度不佳，在兩位病理醫師討論後，決定再將訓練用影像依據共識討論的經驗重新分類，測試用影像則是打亂次序後再分別由兩位病理醫師重新判讀、再共識一次，並重新上傳至新訓練的GTM做判讀。

認識肝纖維化及評估方式



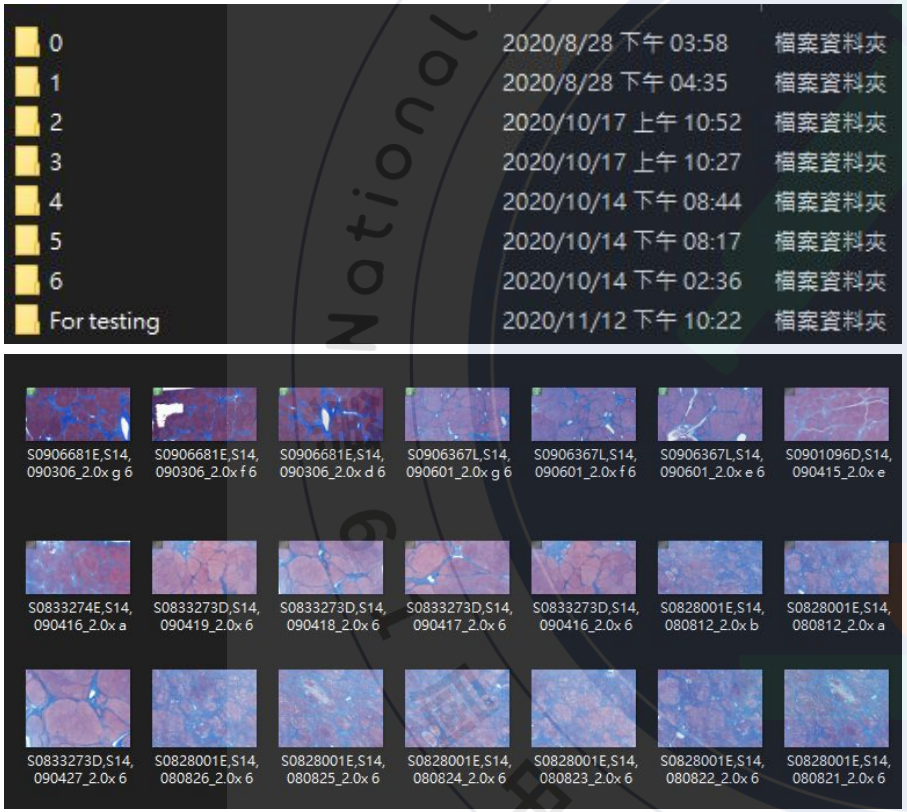
Ishak 0 正常肝臟
Ishak 1 少部分的肝門系統擴張
Ishak 2 大部分的肝門系統擴張
Ishak 3 大部分的肝門系統擴張合併出現輕微橋狀纖維化
Ishak 4 肝門系統擴張合併明顯的橋狀纖維化
Ishak 5 橋狀纖維化合併結節
Ishak 6 可能或是確定的肝硬化

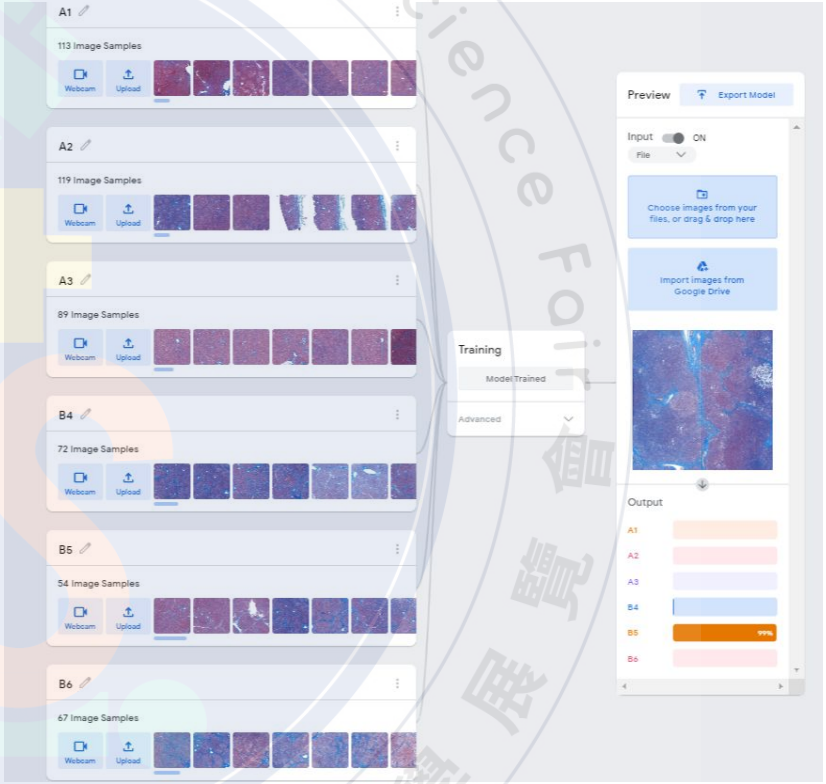
Portal area 門脈系統
Portal vein 肝門靜脈
Hepatic artery 肝動脈
Bile duct 小膽管
Central vein 中央靜脈



- 肝臟反覆受傷、修復，用締結組織取代正常的肝實質組織，導致細胞外基質(ECM)中膠原成分的過度沉積
- 將導致肝硬化及其它併發症，因此肝纖維化的程度評估對肝病的診治及研究極其重要。
- 評分系統有Ishak分七級，和METAVIR分五級。

研究過程





- 建立肝纖維化分級照片的資料夾
- 依序將訓練用照片放入各自的分類裡

- 使用訓練用照片訓練模組
- 以測試用照片測試模型

研究結果

病理醫師共識與GTM判讀差異級數	照片數量	比例
0	72	43.1%
1	72	43.1%
2	17	10.2%
3	5	3%
4	1	0.6%

病理醫師共識與GTM判讀差異級數	照片數量	比例
0	79	47.3%
1	73	43.7%
2	10	6%
3	5	3%

- 以人工智慧辨別肝纖維化與正常肝的影像
  - 正常肝臟很少有切片的檢體，所以我們將Ishak Score分數1(輕微)皆歸類在接近正常肝臟的切片類別裡
  - Ishak Score分數2~6歸類在非輕微肝纖維化
  - 纖維化影像誤判的比例是9.6% (16/167)，正確判斷的比例則是90.4%

- 以人工智慧判斷出肝纖維化的等級
  - 將分類進一步分為六組(Ishak Class 1~6) 類別差異變小，對GTM更加困難
  - 共識後GTM與醫師判讀差異≤一級的達91%，≥二級的檢體13.8%降至9%，而且沒有差異四級的檢體，差異2級的檢體也大幅下降

研究結果

	共識1	共識2	共識3	共識4	共識5	共識6
病理醫師A 1	11	15	0	0	0	0
病理醫師A 2	0	23	16	0	0	0
病理醫師A 3	0	0	23	1	0	0
病理醫師A 4	0	0	0	26	2	0
病理醫師A 5	0	0	0	1	17	3
病理醫師A 6	0	0	0	0	0	29

	共識1	共識2	共識3	共識4	共識5	共識6
病理醫B 1	7	0	0	0	0	0
病理醫師B 2	4	16	0	0	0	0
病理醫師B 3	0	22	33	0	0	0
病理醫師B 4	0	0	6	12	0	0
病理醫師B 5	0	0	0	16	7	0
病理醫師B 6	0	0	0	0	12	32

- 病理醫師間的共識有助GTM模型優化
  - Ishak Score沒有確切的數值作為評斷標準，不同病理醫師在肝纖維化程度的模糊地帶(Ishak Score 2~3、3~4)易有不同答案
  - 起初病理醫師間的判斷歧異較大，訓練出的GTM模組對於診斷的分類表現也較不準確

研究結果

	Precision	Recall	f1-score
class 1	0.22	0.18	0.2
class 2	0.37	0.5	0.43
class 3	0.28	0.38	0.33
class 4	0.62	0.36	0.45
class 5	0.86	0.32	0.46
class 6	0.87	0.84	0.86

- 針對病理醫師共識的標準答案(Consensus)與人工智慧(AI)比較的分析模型，做判讀結果的交叉分析
  - Precision（精確率）代表模型預測結果屬於某等級時，實際的「精準度」
  - Recall（召回率）代表實際屬於某等級的情況下，模型能「預測回多少」
  - f1-score則是精確率和召回率的調和均值
  - 我們的模型在較高等級 (較嚴重纖維化) 的各項指標值均較高，即診斷成效較佳，符合臨床上極忌諱重度誤判為輕微纖維化的要求

研究結果

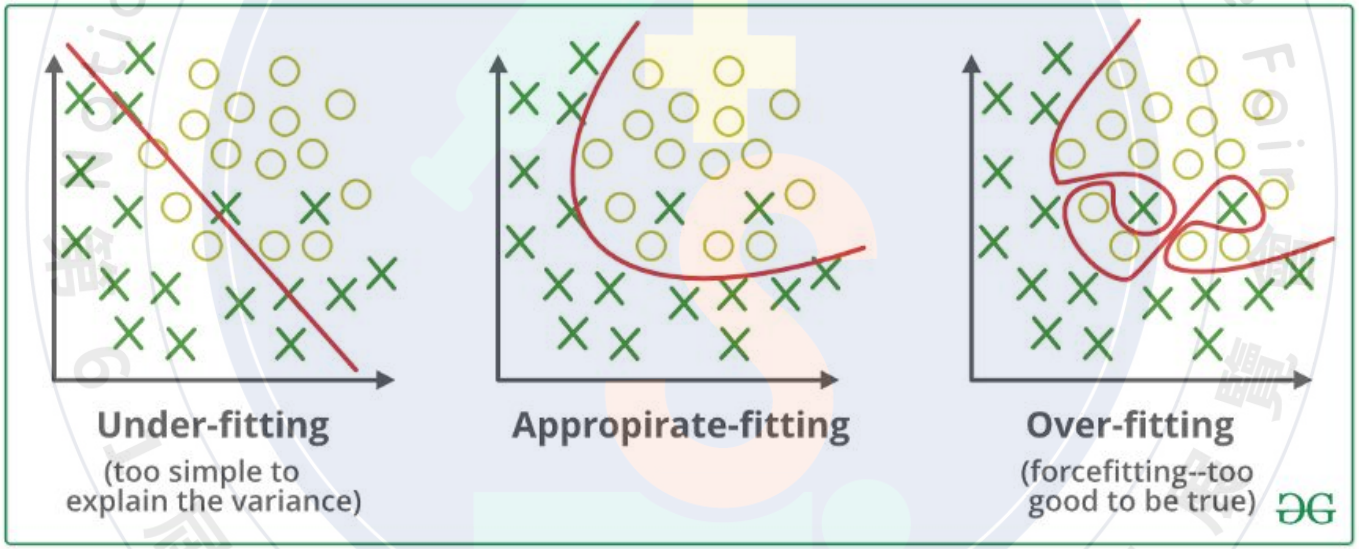
	GTM與共識相同
	GTM較共識嚴重一級
	GTM較共識嚴重二級
	GTM較共識輕微一級
	GTM較共識輕微二級
	GTM較共識輕微三級

	共識1	共識2	共識3	共識4	共識5	共識6
GTM 1	2	7	0	0	0	0
GTM 2	6	19	23	2	1	0
GTM 3	3	12	15	16	3	4
GTM 4	0	0	0	10	5	1
GTM 5	0	0	1	0	6	0
GTM 6	0	0	0	0	4	27

- 167張測試用照片 病理醫師共識診斷分級與GTM (AI) 判讀結果比對分析
  - 我們的模型在2、3等級的精確率較低
  - 我們的模型在5等級的召回率較低

改善的方法

- 調整訓練模組的參數
  - 批大小 (Batch Size)：樣本總數除以總共有幾批 ，適當的批大小有助於模型的學習效率。
  - 訓練回合數 (Epochs)：全部批數總共被人工智慧學習幾次 。



- 縮減病理醫師之判斷差異
- 簡化模型複雜度：照片樣本的分類

參考資料

病理學基礎認識。檢自 https://allentube211.files.wordpress.com/2017/02/ch01.pdf (民109年9月29日)

組織染色法(IHC)方法大全 (民108年9月19日)。檢自

https://www.toolsbiotech.com/news/detail/id/245.html (民109年10月18日)

R A Standish, E Cholongitas, A Dhillon, A K Burroughs,A P Dhillon. (2006). An appraisal of the histopathological assessment of liver fibrosis. Gut, 55(4), 569-78.

Daisuke Komura, Shumpei Ishikawa. (2018). Machine Learning Methods for Histopathological Image Analysis. Computational and Structural Biotechnology Journal, 16, 34–42.