

(一) 前言

1. 機器學習技術 -- 聯邦式學習

聯邦式學習 (Federated Learning) 是一種機器學習的技術，主要目的是在不交換原始數據的情況下進行模型訓練 [1]。隨著隱私保護觀念的進步，個人資料開始受到嚴格保護，跨機構或跨國取得和整合資料變得極為困難。聯邦式學習恰巧解決了這個問題，使得各機構可以在本地訓練模型，無需將資料集傳送到伺服器，造成資料外洩的風險。透過僅傳輸與分享模型參數的機制，可以在雲端伺服器(Server)建立一個共享的模型，並進行模型權重的更新。因此，聯邦式學習能在不洩漏資料的情況下進行模型訓練 [2]。

然而，跨機構的學習和聯邦式學習也有其缺點。在統整資料時，因為橫跨不同機構，資料特徵常有不同。因此，資料統整成為了聯邦式機器學習的一個非常重要的課題。最常見的方法是將所有的資料結構統一，讓所有的客戶端(Client) 都享有同樣的資料特徵。可是此做法會造成資料移除，或是大量資料填補的問題，可能降低模型的訓練效能和降低模型的準確度 [3]。

2. 二次癌症

二次癌症的定義為曾經罹患癌症的患者再次罹患新的癌症。有別於癌症復發，二次癌症為全新的原發癌，不一定與先前所得到的癌症相同，也並非前次癌症復發造成。

因癌症治療技術的進步，患者存活率逐年上升，造成二次癌症發生比例逐年增加。從本研究的資料集中，目前大約每 20 位肺癌患者，就有 1 位得到二次癌症。某些因素可能增加罹患二次癌症的風險，例如遺傳因素、曾經接受的治療方式（如放射線或化療）、生活方式和環境因素等。有鑑於二次癌症的風險日益增加，透過執行本計畫，建立模型，可以預測該病患得到二次癌症的風險，以及哪些因素可能提升罹患二癌風險，讓醫生能夠及早預防或治療任何新的癌症 [4]。

3. 肺癌

肺癌是一種生長在支氣管或肺泡的惡性腫瘤。根據衛福部資料顯示，在台灣，肺癌曾高居國人癌症死因首位。根據世界衛生組織統計，肺癌同時也是許多歐美先進國家癌症死亡率最高的癌症之一。其風險因素主要包括吸煙、被動吸煙、空氣污染等。因肺癌長年屬於台灣的十大癌症，本研究挑選肺癌患者作為研究對象 [5]。

(二) 研究動機與研究目標：

資料特徵的整合一直是跨機構機器學習(或稱聯邦式機器學習)中一個極具挑戰性的課題，尤其在跨機構的機器學習中更為明顯 [3]。各機構所擁有的獨特特徵使得在資料整合階段需要捨棄，或是大量填補數值。然而，捨棄特徵或過度填補數值可能影響模型的學習效果 [6]。為此，本研究提出了兩種解決聯邦式學習中各機構資料特徵不一致問題的方法，並將此方法與各機構的機器模型和僅利用共同特徵訓練的聯邦式機器學習進行比較。

檢測模型的方法是以肺癌患者罹患二次癌症的風險作為預測依據，評估模型在預測二癌風險的準確度。同時，本研究也會進一步解釋各特徵對模型的重要性，以深入了解模型的訓練預測過程。這項研究的結果有望為處理資料不一致性的方法提供實際解決方案，並為相關健康領域的決策制定提供更可靠的決策支援。

(三) 文獻回顧與探討

1 多模態資料 (Multi-Modalities Data)

在許多應用領域中，數據通常以不同的模態 (Modalities) 呈現，每種模態都攜帶著獨特的資訊。這些模態通常來自不同的感測器、數據來源或測量方式，它們提供了有關同一現象的多視角觀察。在處理這些多模態資料時，如果能夠有效地將各模態訊息融合在一起，就可以從而獲得更為全面、精確的分析結果 [7]。

例如，在醫療領域中，電子健康紀錄 (Electronic Health Record, EHR) 是病人健康管理的重要工具，包含多種不同形式的資料，如病歷、心電圖、醫學影像等資料，而這些資料的記錄方式各不相同 [8]。病歷通常以表格形式存儲，心電圖是時間軸的曲線圖 (Temporal Data) 呈現，而醫學影像則以圖片形式儲存。

2 模型融合 (Model Fusion)

為了充分利用多模態資料 (Multi-Modalities Data)，許多研究透過模型融合 (Model Fusion) 的方法，將來自不同模態的資訊進行整合，從而訓練出更強大且準確的機器學習模型 [9]。透過融合這些不同模態的資料，模型能夠從更廣泛的角度進行學習，並且在許多應用中，比單一模態的模型表現更為優越。常見的模型融合方法可以分成三種主要類型：早期融合 (Early Fusion)、中間融合 (Intermediate Fusion)、晚期融合 (Late Fusion)。

2.1 早期融合 (Early Fusion)：指的是在模型訓練之前，將不同模態的數據進行合併，並將合併後的數據輸入到模型中進行訓練。這種方法通

常將不同模態的數據進行特徵拼接，形成一個統一的特徵向量，然後將其作為模型的輸入。

- 2.2 中間融合 (Intermediate Fusion)：指的是在模型的中間層進行融合。在這種方法中，每個模態首先獨立通過其對應的模型，提取出各自的特徵。接著，這些特徵會在某一中間層進行融合，然後再進行後續的預測或分類 [10]。
- 2.3 晚期融合 (Late Fusion)：指的是在每個模態的預測結果完成後，將不同模態模型的預測結果進行合併，並基於這些結果做出最終決策。這種方法通常涉及將多個模型的輸出進行加權平均、投票等方法進行綜合，獲得最終的預測結果。

3 集成學習 (Ensemble Learning)

集成學習 (Ensemble Learning) 是一種機器學習技術，將多個弱學習器 (weak learner) 結成一個強學習器 (strong learner) [11]。弱學習器通常是由**同一資料集隨機抽樣**訓練出那些表現略優於隨機猜測的模型，但當這些模型進行集成後，能夠產生一個強大的預測模型。集成學習的主要目的是利用多個基礎模型的多樣性來減少偏差，進而改善預測的準確度 [12]。常見的集成學習技術有 Bagging、Boosting、Stacking。

- 3.1 Bagging 是一種集成學習技術，主要通過隨機重複抽樣訓練集來訓練多個基礎模型。每個基礎模型都會訓練在一個稍微不同的數據集上（通過隨機抽樣的方式生成）。最終的預測結果是這些基礎模型預測結果的平均值或投票結果。
- 3.2 Boosting 是一種迭代的集成方法，其中每個新的模型都會根據前一個模型的錯誤進行加權。通過訓練新的模型來集中關注先前模型錯誤分類的樣本。這樣的方式能夠有效減少偏差，並且提高最終預測的準確性。著名的 Boosting 算法包括 AdaBoost、Gradient Boosting 和 XGBoost。
- 3.3 Stacking 是一種將多個基礎模型的預測結果進行融合的集成學習方法。首先，利用同一份資料訓練出多個基礎模型 (base models)（可能是不同算法）分別訓練，然後將這些基礎模型的預測結果作為新的特徵，傳遞給一個次級模型 (second-level model) 進行最終預測。這樣可以進一步改善預測性能，因為次級模型能夠學習到各基礎模型的優勢 [13]。

4 聯邦式學習在醫療中的應用

聯邦式學習在醫療領域中扮演著至關重要的角色，它允許跨越多個醫療機構展開相互訓練資料集，而不必共享敏感的患者資料。這種方法解決了隱私問題，同時輕鬆地增加了訓練資料的數量。通過分散式資料集提高效率，聯邦式

學習在醫療領域比其他領域更合適，能夠在病患隱私和科技進步之間取得平衡。這種方法確保能不侵犯患者的隱私，同時允許醫療機構共同提高模型預測準確性 [14] [15]。

5 醫療領域中常見的資料不平衡

不平衡資料問題常出現在醫療領域，例如本次實驗所研究的二次癌症資料集，罹患機率僅為 3%~7%。文獻提出了一種針對不平衡資料集進行損失分配的方法，其有效性已在多個場景中得到驗證 [16]。本研究採用了該文獻中建議的損失權重分配方式，進一步應用於二次癌症預測問題中，以驗證其在聯邦式學習架構下的適用性。

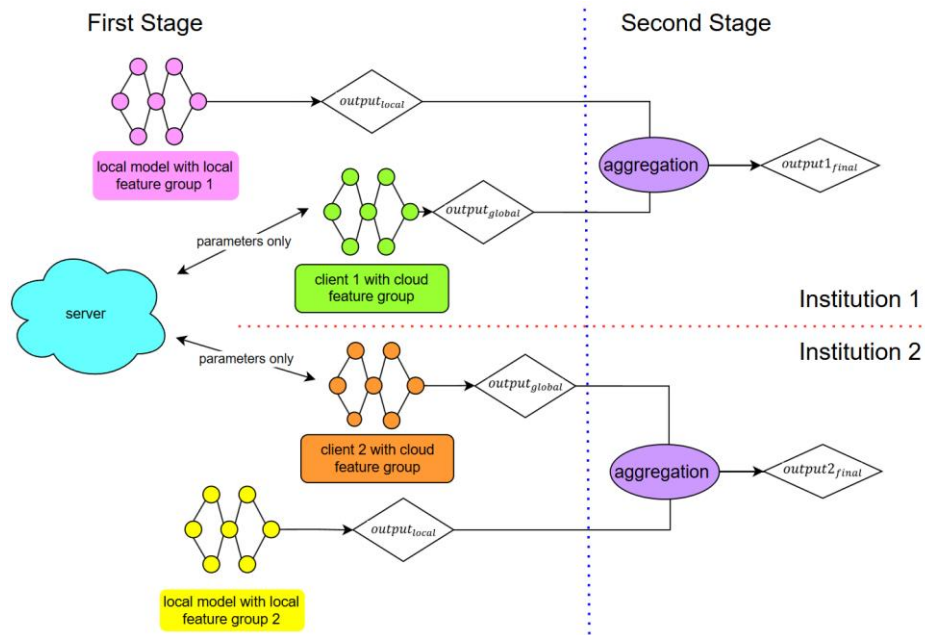
(四) 研究方法及步驟

本研究的研究目標是以肺癌患者罹患二次癌症的風險作為預測依據，評估模型在預測二癌風險的準確度。

模型設計

如圖一所示，本研究提出的演算法是根據特徵相似程度將特徵分成本地組和雲端組。雲端組特徵在客戶端訓練，參數於 Server 端統整，進行聯邦式機器學習(Federated Learning)；而本地組特徵僅在 Client 端進行本地化機器學習(Localized Learning)。透過上述訓練，可分別得到兩個模型的預測結果(Output)。其中，本地組和雲端組的特徵可以重複。最後再利用本研究所提出的兩種聚合(Aggregation)方法將結果合併，得到更精確的答案。

本研究將資料集分成三個不同的子集，訓練集 (Train Set)、驗證集 (Validation Set) 和測試集 (Test Set)。如圖一所示，訓練集(Train Set)的數據用於訓練模型第一階段。訓練過程中，雲端組的特徵數據在各個 Client 進行聯邦學習模型的訓練，並再於伺服器端做全體模型權重的訓練，而本地組的特徵數據則用 Client 端的 Neural Network 神經網路機器模型訓練。訓練完成後，在第二階段使用驗證集(Validation Set)調整本研究提出的聚合方法 (Aggregation Method) 的相關參數。最後，透過測試集(Test Set)評估模型效能。



圖一 完整模型示意圖 $n = 2$ (n 為 client 數目)

實驗流程

1. 資料分組與資料前處理

先整理出各機構中所有的特徵，按下列步驟將特徵放入雲端組和 n 個本地組 (n 為 client 數目)。要注意的是，若特徵同時符合雲端組和本地組的標準，則將該特徵同時放入雲端組和本地組兩個特徵組。在本實驗中，是以不交換任何原始資料為前提設計，僅交換資料特徵與編碼方式。

特徵納入雲端組的條件

將所有特徵依據下列情況處理，若處理過後的資料符合所敘述的標準，則將該筆特徵編入雲端組。

情況 1: 類別特徵

情況 1-1: 在各機構共有的相同特徵且有相同的類別:

以癌症登記檔中的性別(Sex)為例，參考過去研究，依 (1: Male、2: Female)方式編碼。完成編碼後，機構 1 (台灣醫院)的性別資料分佈如表一左，機構 2 (美國資料庫: The Surveillance, Epidemiology, and End Results (SEER) Program)性別的資料分佈如表一右。兩機構均有性別特徵，且在編碼後，類別數目相同。此類特徵編入雲端組。

表一 性別於各機構間的資料分佈範例

性別編碼	機構 1 資料筆數	機構 2 資料筆數
1: Male	5,606	514,904
2: Female	4,939	458,037

情況 1-2: 在各機構的相同特徵卻有不同的類別

以癌症登記檔中的肺癌類別為例，參考過去研究，利用 One hot encoding 完成編碼後，機構 1(台灣醫院 1)的資料分佈如表二左，機構 2(台灣醫院 2)的資料分佈如表二右。兩機構均有肺癌類別特徵，但在編碼後，類別數目不同。此時，由下列公式判定是否編入雲端組。當不等式成立時，將特徵放入雲端組。本公式的意義為檢測種類不同數目的多寡。若各機構僅只缺少或多出些許不同的種類時，我們應該將其放入雲端組，當作共同的特徵。其中， k 為自定義的正整數，本研究以 $k=2$ 作為實驗閾值。

範例中，因 $\frac{6-5}{5} < 2$ ，肺癌類別符合雲端組定義，放入雲端組。

$$\frac{n(\cup_{i=1}^n (Data_{institution\ i}) - \cap_{i=1}^n (Data_{institution\ i}))}{n(\cap_{i=1}^n (Data_{institution\ i}))} < k \quad k \in Z$$

表二 肺癌類別於台灣的資料分佈範例

編碼	機構 1 資料筆數	機構 2 資料筆數
C340	0	68
C341	246	638
C342	34	84
C343	179	344
C348	2	14
C349	8	40

情況 1-3: 機構獨有特徵

以癌症登記檔案為例，台灣的癌症登記檔案規定記錄患者是否有嚼檳榔習慣，而美國 SEER 資料庫中並無此資料。此類特徵不編入雲端組。

情況 2: 連續特徵

與類別型特徵需要做編碼比對不同，連續型特徵僅需確認特徵內容以及特徵紀錄單位是否相同。若特徵內容相同，則可將該特徵放入雲端組。在雲端組資料部分，需更進一步確認特徵紀錄的單位是否相同。以癌症腫瘤長度大小為例，機構 1 使用釐米紀錄癌症大小，資料分佈區間為 0~9 之間，而機構 2 以

公釐紀錄，資料分佈區間為 10~90 之間，需在訓練前與各機構協調特徵使用單位，以確保模型訓練成效。

特徵納入本地組的條件

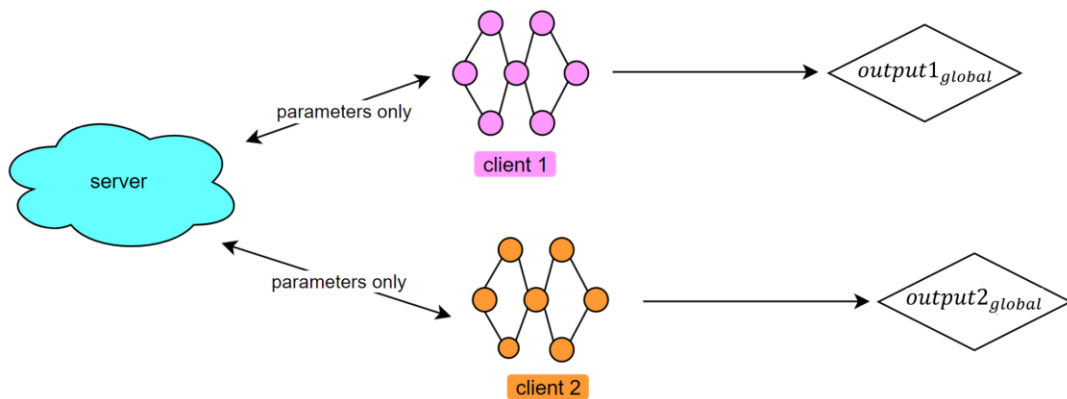
特徵是否納入本地組與實際資料密切相關。在設計本地組特徵時，應以提升各機構本地模型的性能為目標，根據資料特性進行特徵選擇和設計，以確保能達到最佳的訓練效果。以本次研究的二次癌症為例，排除掉在雲端組的特徵，剩下的特徵數量過少導致無法有效訓練模型，因此本次研究將各機構所有的特徵(包括已納入雲端組的特徵)皆編入各機構的本地組特徵集。

2. 設計模型架構

本研究所使用的聯邦式學習架構為神經網路模型。各機構(Client)先在本地端訓練模型，再將模型參數上傳到 Server，Server 將參數加權平均之後，再將平均參數下放給 Clients。如此操作，訓練資料就會留在各機構的裝置中，不用共享資料也可以達到訓練模型的目的。因為二次癌症是不平衡的醫療資料，解決方法將採用文獻之提供的調整損失函數權重方法做處理 [16]。

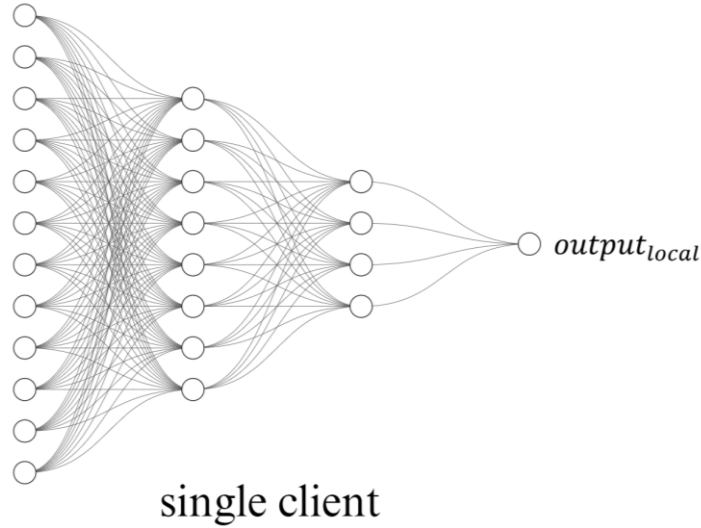
第一階段：使用雲端組資料和本地組資料，各自訓練模型

如圖二，首先利用雲端組特徵作為資料集，並使用上述聯邦式學習架構，與各機構聯合訓練，得到雲端組模型以及相對應的模型預測輸出結果 ($output_{global}$)。



圖二 使用雲端組特徵於聯邦式學習示意圖 $n = 2$ (n 為 client 數目)

如圖三，各機構分別使用本地組特徵作為資料集，以神經網路方法訓練，得到本地組模型以及相對應的輸出結果($output_{local}$)。



圖三 使用本地組特徵於本地端訓練模型架構示意圖

第二階段：合併雲端組模型與本地組模型 (Aggregation Method)

本計畫使用 Seesawing Weights 和 Dual Perception Net 兩種方法合併第一階段的訓練結果。

A. Seesawing Weights Algorithm

使用比例調控機制合併上述雲端組與本地組的模型輸出結果。參考 Adaboost [17] 演算法，結合 Linear Regression 和 Logistic Regression 的想法，提出改良式模型合併演算法 “Seesawing Weights Algorithm”。

將上述的 $output_{global}$ 和 $output_{local}$ 兩種模型預測結果合併，其中 $output_{global}$ 和 $output_{local}$ 分別是兩個 Multi-class 的機率向量。舉例來說，若要預測 3 個類別的發生機率，則 $output_{global}$ 和 $output_{local}$ 可能的形式為 [0.6 0.3 0.1]。由此向量可知，模型預測第一種類別。

首先定義參數 ε_{global} ，用於判斷使用雲端組資料以及聯邦式學習建立的模型是否預測正確。預測正確的結果為 0，預測錯誤的結果為 1。

$$\varepsilon_{global} = \text{ceil}(\max(output_{global}) - (output_{global})_{\text{argmax}(\text{ground truth})})$$

參數 ε_{local} 則用於判斷本地組資料訓練的模型是否預測正確。預測正確的結果為 0，預測錯誤的結果為 1。

$$\varepsilon_{local} = \text{ceil}(\max(output_{local}) - (output_{local})_{\text{argmax}(\text{ground truth})})$$

參數 ε ，用於綜合判斷兩個模型的正確與否。當兩個模型同時預測正確或錯誤時，輸出 0。反之，輸出 1。

$$\varepsilon = \varepsilon_{global} * (1 - \varepsilon_{local}) + \varepsilon_{local} * (1 - \varepsilon_{global})$$

參數 *confidence local* (*cl*)，記錄本地組資料訓練的模型對於實際正確類別的信心水準。因為輸出結果為機率，因此 *cl* 會在 0~1 之間。

$$confidence\ local\ (cl) = (output_{local})_{argmax(ground\ truth)}$$

參數 *confidence global (cg)*，記錄使用雲端組資料以及聯邦式學習建立的模型對於實際正確類別的信心水準。因為輸出結果為機率，因此 *cg* 會在 0~1 之間。

$$confidence\ global\ (cg) = (output_{global})_{argmax(ground\ truth)}$$

使用 Weights Adjustment Function (Δw) 調整模型權重。當其中一個模型預測正確，另一個預測錯誤時，權重會往預測對的模型調整；當兩個模型預測正確或錯誤時，也會因為模型的信心水準有差異，導致權重偏移。不僅如此，偏移的量也會受到兩個模型預測結果而有所改變。當模型同時預測正確或錯誤時，權重改變的量會比較小。相反地，當如兩模型預測結果不相同時，權重的值會比較大。

$$\Delta w = \eta * \left\{ (1 - \varepsilon) e^{\frac{|cl-cg|}{2}} + \varepsilon e^{\frac{|cl+cg|}{2}} \right\} \quad learning\ rate: \eta$$

參數 learning rate 則是採用自定義的 learning rate scheduler 和 convergence number 常數。

$$\eta = \frac{1}{dataset\ size} * e^{\frac{-current\ epoch}{convergence\ number}} \quad convergence\ number \in C$$

將 validation set 輸入至 global 模型和 local 模型，得到兩個輸出的機率向量，再針對以下公式計算得到最後結果。

$$w_{global} * output_{global} + w_{local} * output_{local} = final\ output$$

其中

$$w_{global} + w_{local} = 1$$

初始化權重的方法也會影響到模型收斂的速度，因此，本研究將模型的 Area Under Receiver Operating Characteristic Curve (AUROC) 當作初始化的因素。對於 AUROC 較高的模型，會享有較高的權重。其中，Receiver Operating Characteristic curve 是一種坐標圖式的分析工具，是一種用於在二元分類模型中找尋最佳閾值的方法 [18]。

$$w_{global/local} = \frac{auroc_{global/local}}{auroc_{global} + auroc_{local}}$$

圖四展示了一種更新權重的方法，其概念借鑑了隨機梯度下降 (Stochastic Gradient Descent, SGD)，透過逐筆資料調整權重 (weights)。然而，與傳統神

經網路依賴損失計算（Loss Calculation）及反向傳播（Backpropagation）來更新權重、即時反映模型效能的方式不同，本方法採取了另一種策略。

我們採用了圖四中的方法，當權重聚合（Aggregation）結果出錯時，直接進行更新。這表明，即使不依賴損失函數計算，我們仍能有效判斷模型當前狀態並進行相應調整。

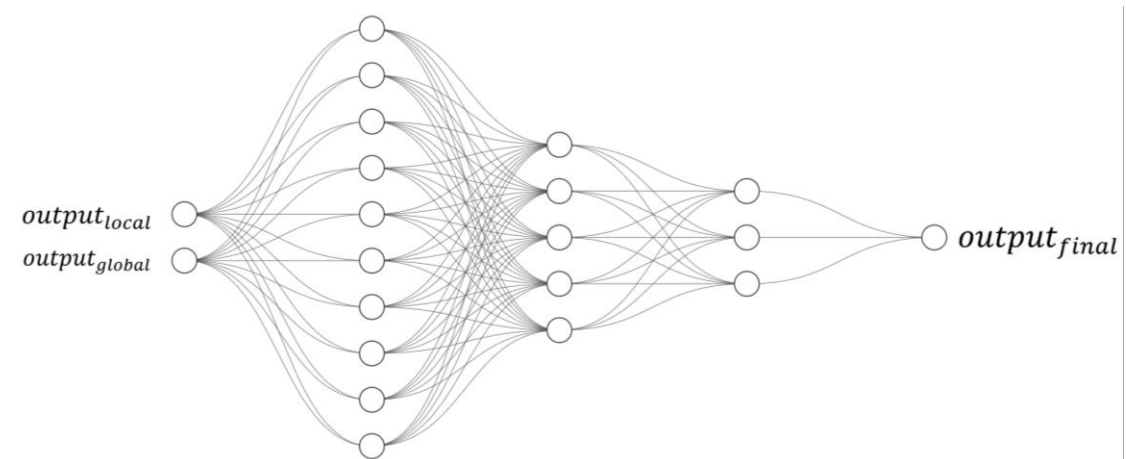
True False 為該模型預測結果正確或錯誤，可以透過 ε_{local} 、 ε_{global} 得知。當 $cg = cl$ 時，表示兩者對於事件皆有相同的信心程度，因此，不調整權重。

		global	
		true	false
local	true	$cl < cg$ $w_{global} = w_{global} + \Delta w$ $w_{local} = w_{local} - \Delta w$	$w_{global} = w_{global} - \Delta w$ $w_{local} = w_{local} + \Delta w$
	false	$cl > cg$ $w_{global} = w_{global} - \Delta w$ $w_{local} = w_{local} + \Delta w$	$cl > cg$ $w_{global} = w_{global} + \Delta w$ $w_{local} = w_{local} - \Delta w$
		$w_{global} = w_{global} + \Delta w$ $w_{local} = w_{local} - \Delta w$	$cl < cg$ $w_{global} = w_{global} - \Delta w$ $w_{local} = w_{local} + \Delta w$

圖四 權重參數調整邏輯

訓練過程使用 Validation Set 做權重的調控。將資料輸入到由本地組和由雲端組訓練的兩個模型後，會得到一連串的預測結果。再利用權重參數調整表和這些結果來反覆調整權重（ $w_{global/local}$ ），進而得到最佳的權重。

B. Dual Perception Net



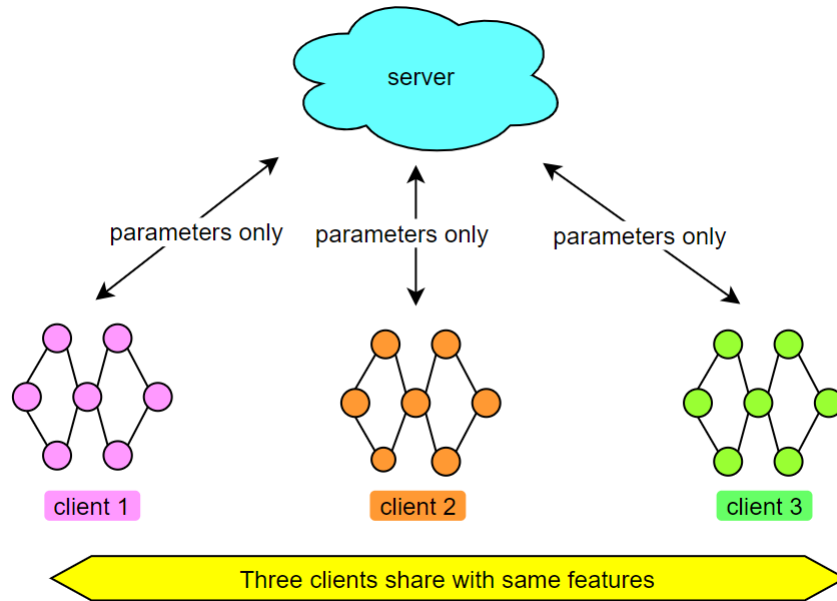
圖五 B. Dual Perception Net 模型示意圖

Dual Perception Net 方法結合了兩個模型的不同視角進行分析，如圖五所示。利用神經網路能模擬多種函數的特性，進一步探索由本地組與雲端組分別訓練的兩個模型之間的內在關聯性。

將 Validation Set 輸入到本地組和雲端組兩個模型中，得到兩模型的輸出結果後，將預測結果串接(Concatenate) 成新的資料集。例如 $output_{global} = [0.8 \ 0.1 \ 0.1]$ 、 $output_{local} = [0.6 \ 0.3 \ 0.1]$ \rightarrow $New \ input = [0.8 \ 0.1 \ 0.1 \ 0.6 \ 0.3 \ 0.1]$ 。訓練中，資料不平衡問題也將採用文獻之提供的調整損失函數權重方法做處理 [16]。

過去研究使用的方法 – 標準的聯邦式學習模型

在過去的研究中，當處理各機構中不同類別特徵時，先利用 One hot encoding 將類別資料轉變成數值。遇到連續性的特徵時，將資料的單位統一。接下來，取得所有機構的特徵聯集，將各機構中缺少的特徵數補零或是以其他方式填補。當所有 client 均享有同樣特徵後，使用聯邦式學習訓練模型，並預測結果 (圖六)。本研究會將各機構獨特的特徵捨棄以統一各機構的特徵數，實作聯邦式學習，並與上述提出的方法做比較。



圖六 聯邦式模型示意圖 $n = 3$ (n 為 client 數目)

3. 特徵重要性

本研究採用 Shapley Additive Explanations (SHAP) 作為量測特徵重要性的方法 [19]。SHAP 借鑒了博弈論中的 Shapley 值，通過計算各特徵對模型輸出的貢獻，量化每個特徵在整體決策過程中的影響。採用此方法的原因是因為在計算 SHAP 值時，考慮了所有可能的特徵組合，確保結果對所有特徵具有公平性。

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

ϕ_i : 表示特徵 i 的 SHAP 值，即該特徵對模型輸出的邊際貢獻

S 是特徵的子集， N 是所有特徵的集合

$f(S)$ 是僅使用特徵子集 S 訓練的模型預測值

本研究實驗資料集 -- 肺癌患者罹患二癌資料

利用台灣衛生福利部的癌症登記資料庫和美國的癌症資料庫-- Surveillance, Epidemiology, and End Results Program (SEER) 作為此研究的資料集。台灣的資料包含五家醫院的癌症登記檔。SEER 資料庫則採用 Incidence - SEER Research Data, 18 Registries, Nov 2020 Sub。台灣 5 家醫院共有 10545 筆肺癌患者資料，SEER 則有 972941 筆肺癌患者資料。台灣的資料統計是從 2011~2020；美國的 SEER 資料則是從 2000~2018，且包含不同種族的資料。經上述實驗流程的資料特徵分組方法得出雲端組（表三）與本地組（表四、表

五) 特徵，並用表格列出該組別所有的特徵。表四、表五中，本地組特徵中以灰底表示的特徵和雲端組特徵重複。

表三 雲端組特徵列表

Encode Name	Description
Laterality	Laterality
Age	Age at Diagnosis
Gender	Sex
SepNodule	separate tumor nodules in the ipsilateral lung
PleuInva	visceral pleural invasion/elastic layer
Tumorsz	tumor size
LYMND	positive regional lymph nodes
AJCC	combine stage
Radiation	radiology treatment
Chemotherapy	systemic therapy
Surgery	Surgery (treatment)
Target (Y1)	Having second primary cancer or not (Target)

表四 美國本地組特徵列表

Encode Name	Description
Laterality	Laterality
Age	Age at Diagnosis
Gender	Sex
SepNodule	separate tumor nodules in the ipsilateral lung
PleuInva	visceral pleural invasion/elastic layer
Tumorsz	tumor size
LYMND	positive regional lymph nodes
AJCC	combine stage
Radiation	radiology treatment
Chemotherapy	systemic therapy
Surgery	Surgery (treatment)
Income	Money earned by the patient
Area	Residence (Rural-Urban Continuum Code)
Race	Races (White, Black, Other)
Target (Y1)	Having second primary cancer or not (Target)

表五 台灣本地組特徵列表

Encode Name	Description
Laterality	Laterality
Age	Age at Diagnosis
Gender	Sex
SepNodule	separate tumor nodules in the ipsilateral lung
PleuInva	visceral pleural invasion/elastic layer
Tumorsz	tumor size
LYMND	positive regional lymph nodes
AJCC	combine stage
Radiation	radiology treatment
Chemotherapy	systemic therapy
Surgery	Surgery (treatment)
PleuEffu	malignant pleural effusion (SSF4)
EGFR	EGFR (Epidermal growth factor receptor) gene mutation (SSF6)
ALK	ALK gene mutation (SSF7)
MAGN	surgical margins of the primary site
DIFF	pathologic differentiation
BMI_label	BMI
CIG	Smoking
BN	Betel nut chewing
ALC	Alcohol consumption
Target (Y1)	Having second primary cancer or not (Target)

(五) 實驗結果

模型驗證

本研究實驗不同的資料集切割比例，將資料集依照 6:3:1 和 7:2:1 分為訓練集 (Train Set)、驗證集 (Validation Set) 和測試集 (Test Set)，進行共計二組實驗。模型預測的評估指標為 Area Under Receiver Operating Characteristic Curve (AUROC，表六、表八) [18] 和 Area Under Precision Recall Curve (AUPRC，表七、表九)，平均使用 30 個隨機種子 (Random Seed) 切分資料與訓練模型所計算出的 AUROC 與 AUPRC 做為最終結果，在兩個地區（台灣和美國）的資料集上進行對比。

表六 二次癌症分析結果(AUROC)--資料切割比例 6:3:1

Model	Taiwan	USA
Federated Learning	0.6959 (0.0176)	0.7159 (0.0216)
Localized Learning	0.7135 (0.0153)	0.7365 (0.0055)
Seesawing Weights	0.7279 (0.0362)	0.7287 (0.0136)
Dual Perception Net	0.7280 (0.0379)	0.7349 (0.0093)

表七 二次癌症分析結果(AUPRC)--資料切割比例 6:3:1

Model	Taiwan	USA
Federated Learning	0.0912 (0.0090)	0.1282 (0.0127)
Localized Learning	0.1027 (0.0102)	0.1410 (0.0085)
Seesawing Weights	0.1083 (0.0229)	0.1371 (0.0092)
Dual Perception Net	0.1144 (0.0300)	0.1441 (0.0310)

表八 二次癌症分析結果(AUROC)--資料切割比例 7:2:1

Model	Taiwan	USA
Federated Learning	0.7014 (0.0276)	0.7209 (0.0282)
Localized Learning	0.7130 (0.0189)	0.7387 (0.0061)
Seesawing Weights	0.7156 (0.0451)	0.7354 (0.0095)
Dual Perception Net	0.7064 (0.0473)	0.7382 (0.0095)

表九 二次癌症分析結果(AUPRC)--資料切割比例 7:2:1

Model	Taiwan	USA
Federated Learning	0.0937 (0.0115)	0.1281 (0.0147)
Localized Learning	0.1023 (0.0090)	0.1405 (0.0044)
Seesawing Weights	0.1069 (0.0203)	0.1376 (0.0082)
Dual Perception Net	0.1019 (0.0217)	0.1512 (0.0331)

根據實驗結果，和 Baseline Model 的 Federated Learning 和 Localized Learning 模型相比，Dual Perception Net 神經網路模型在實驗中均達到良好的 AUROC 表現，特別是在台灣資料集上的預測效能提升顯著。此外，Seesawing Weights Algorithm 緊隨其後，展現了一定的穩定性。

時間成本比較

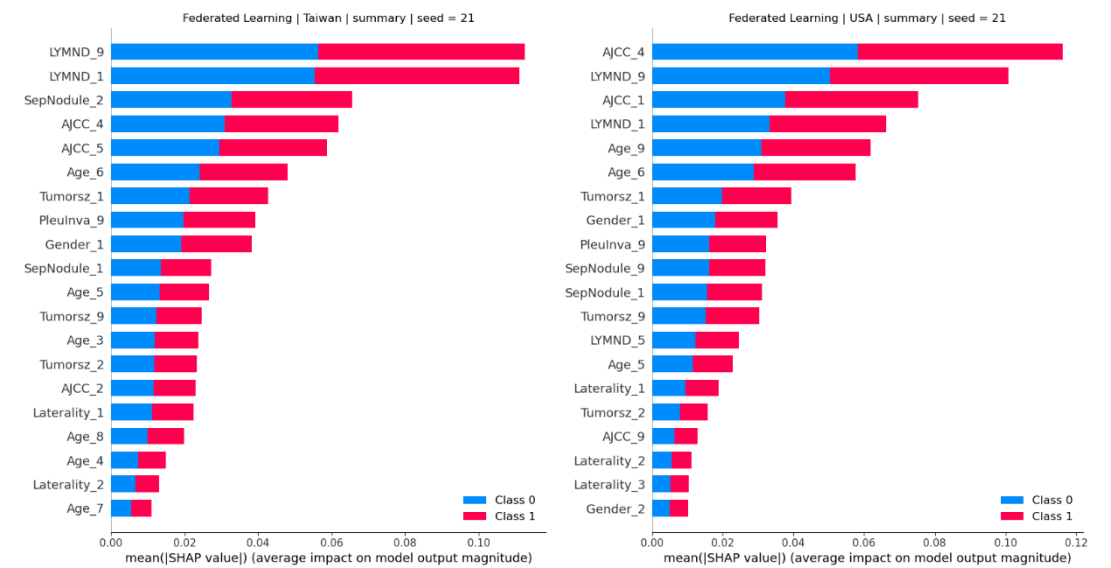
各模型的訓練時間亦為重要考量。本研究計算了 Dual Perception Net 神經網路模型與 Seesawing Weights Algorithm 調整權重模型演算法的訓練時間比值，再平均各機構所計算出的比值，其結果如下(表十)：

	Taiwan	USA
$\text{Ratio}(\frac{\text{TIME}_{\text{DPN}}}{\text{TIME}_{\text{SSW}}})$	5.5751	5.8520

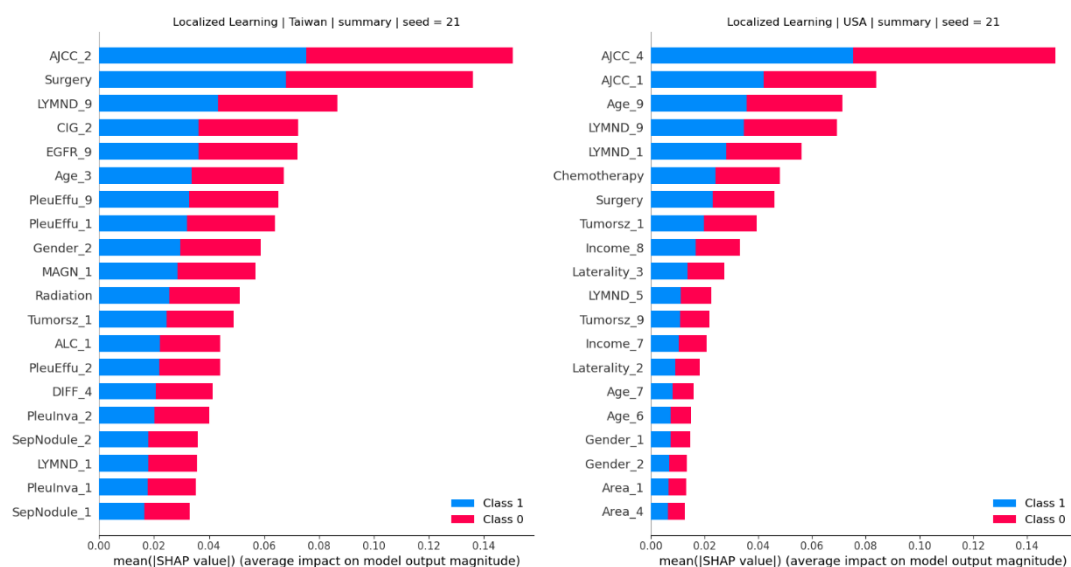
結果顯示，Dual Perception Net 神經網路模型雖然在效能上稍有優勢，但其訓練時間成本相對於 Seesawing Weights Algorithm 高出許多。

特徵重要性解釋

為進一步探討模型差異，本研究利用 SHAP 值分析模型的特徵重要性。根據實驗結果，當隨機種子（Random Seed）設定為 21 時，雙方模型表現最佳，因此以該條件下的 SHAP 值作為分析基礎。此分析有助於了解不同模型對特徵的依賴程度及其解釋能力，並進一步探討不同模型間的決策差異。

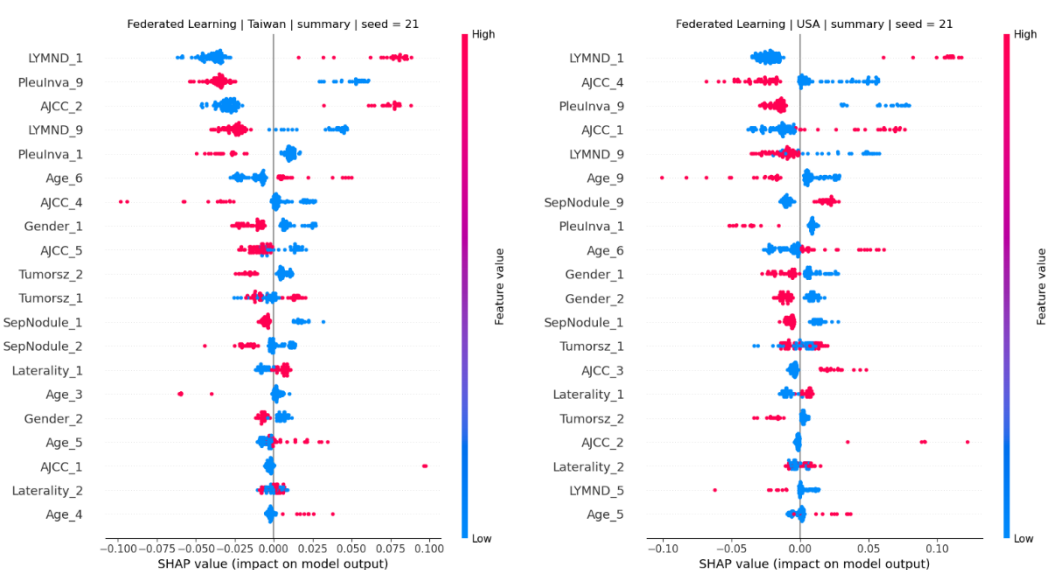


圖七：聯邦式機器學習預測二次癌症罹患與否的特徵重要性分布圖

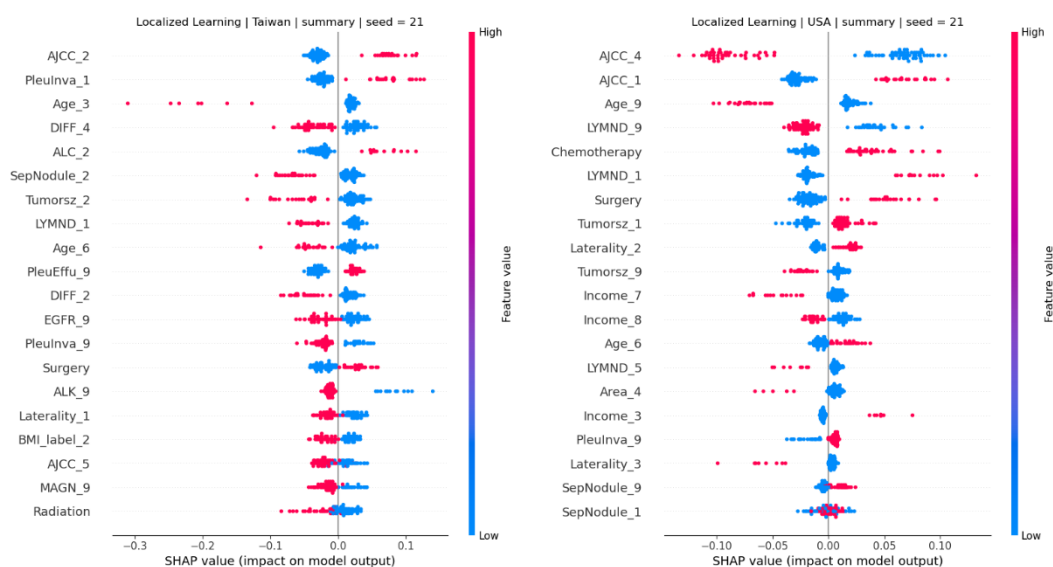


圖八：本地化機器學習預測二次癌症罹患與否的特徵重要性分布圖

從聯邦式學習(Federated Learning，圖七)與本地化學習(Localized Learning，圖八)的結果可以看出，AJCC 癌症分期是一個關鍵特徵，且兩種模型都成功捕捉到了其重要性。然而，本地數據中獨特的特徵也同樣不可忽視。如前文所假設的，若直接忽略這些特有特徵，可能會導致資訊的流失。以台灣地區的數據為例，肋膜積水(PleuEffu)、表皮生長因子受體基因突變(EGFR)、身體質量指數(BMI) 等特徵在模型中被視為重要特徵。若僅依賴聯邦學習，可能無法充分利用這些關鍵資訊。

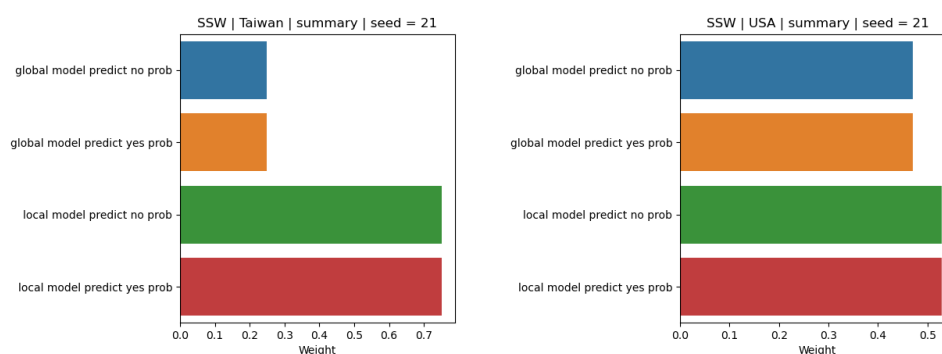


圖九：聯邦式機器學習對於罹患二次癌症特徵重要性分布圖

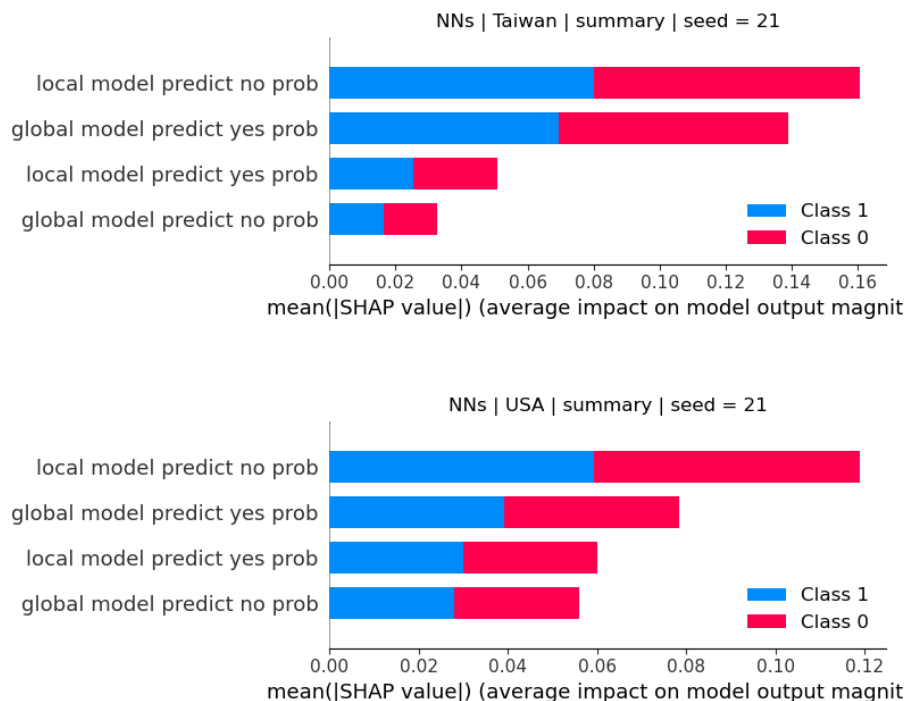


圖十：本地化機器學習對於罹患二次癌症特徵重要性分布圖

圖九與圖十描述特徵對於罹患二次癌症的正負相關性。以 AJCC 癌症第四分期為例(AJCC_4) 為例，和罹患二次癌症呈現負相關。也就是說，若病人處在這一階段，則有較大的可能不會罹患二次癌症。從圖十可以看出，在一般生活的可控因子當中，台灣人若有飲酒習慣罹患二次癌症的機率較高。



圖十一：Seesawing Weights Algorithm 預測二次癌症罹患與否的特徵重要性分布圖



圖十二: Dual Perception Net 預測二次癌症罹患與否的特徵重要性分布圖

本研究提出的兩種聚合方法中，Seesawing Weights Algorithm 能清楚解釋本地化模型（Localized Learning）的重要性，且能利用加權平均來計算出本地組特徵的影響力（圖十一）。而 Dual Perception Net 則因為其背後運用了神經網路（Neural Networks）來構建模型，使得無法直接量化或解釋單一模型的重要性，同時也難以精確定位原始特徵的影響力（圖十二）。

(六) 結果討論:

表十一: Seesawing Weights 和 Dual Perception Net 的優劣比較

	優點	缺點
Seesawing Weights	<ol style="list-style-type: none"> 快速求得特徵對於模型輸出的貢獻度 降低時間訓練成本 	<ol style="list-style-type: none"> 只能找出簡單的線性關係，無法處理複雜的關係
Dual Perception Net	<ol style="list-style-type: none"> 可以模擬複雜的函數，提升模型的精準度 	<ol style="list-style-type: none"> 容易發生過擬和(Overfitting) 時間成本大幅提升 求得特徵對於模型輸出的貢獻度難度提升

模型預測結果

在基準模型（Baseline Models）中，聯邦式機器學習（Federated Learning）透過整合多個資料集，成功訓練出高效能的模型，而本地化機器學習（Localized Learning）則能精準捕捉區域性的重要特徵。這兩者在最終的聚合方法中皆扮演了關鍵角色。

在模型聚合方法中，Seesawing Weights Algorithm 與 Dual Perception Net 神經網路模型皆展現出穩定且卓越的表現。特別是在台灣地區數據的應用上，兩種模型相較於基準模型皆有顯著提升，其中 Dual Perception Net 的表現略勝一籌。值得注意的是，在資料集切割策略上，7:2:1 切割法可能導致驗證集（Validation set）過小，使得模型在第二階段訓練時容易出現 underfitting。為改善此問題，未來研究可嘗試不同的切割策略，例如讓驗證集與訓練集部分重疊，以期獲得更佳的模型表現。

總結而言，這兩種方法有效解決了跨機構領域中非共同特徵的挑戰，為未來跨機構機器學習的應用提供了有力的參考依據。表十一 進一步統整並比較了兩種模型的優劣表現。

特徵分析

從聯邦式學習（Federated Learning）與本地化學習（Localized Learning）的結果可以觀察到，兩種模型皆能成功捕捉某些關鍵特徵，例如：AJCC 癌症分期。本地化學習（Localized Learning）則捕捉了更多區域性的特徵重要性。Seesawing Weights Algorithm 的結果清楚地反映出本地化學習在權重分配中的重要性，而 Dual Perception Net 神經網路模型則通過多層次的非線性映射展現了特徵與基準模型(Baseline Models)貢獻的深層關係。

值得注意的是，在美國的本地化學習特徵分析中(圖八)，高收入（Income_7、Income_8）被模型識別為重要特徵之一。圖十，則進一步顯示出高收入的人罹患二次癌症的機率會比一般收入的人低。這可能反映高收入人群能獲得更完善的醫療資源，且從第一次癌症康復後，有更多的金錢來保養與照顧身體。這也進一步突顯生物醫療領域加速發展的必要性，如果人工智慧能在診療過程中輔助醫生，抑或是在生活中給予民眾提高生活品質的建議或是重要的保健知識，則有望彌平財富帶來的不公。讓醫療資源的分配更加均衡，從而實現每個人都能有健康的生活。

(七) 結論

肺癌作為台灣乃至全球的重大健康威脅，高發病率與死亡率進一步突顯出其被研究診斷的重要性。本研究以肺癌患者罹患二次癌症的風險為例，在保護病患資料隱私的前提下，提出的模型解決各機構資料特徵不一致問題的方法，減少資源的浪費，並展現其在臨床應用中的潛力。隨著技術的進步，未來智慧

醫療將能更精確地協助醫生進行診斷與預測，最終實現提升醫療資源公平性與普及性的目標。

(八) 參考文獻

- [1] McMahan et al. “Communication-Efficient Learning of Deep Networks from Decentralized Data” [Online] Available: <https://arxiv.org/pdf/1602.05629>
- [2] Adrian Nilsson et al. “A Performance Evaluation of Federated Learning Algorithms” [Online] Available: <https://dl.acm.org/doi/pdf/10.1145/3286490.3286559>
- [3] Priyanka Mary Mammen, “Federated Learning: Opportunities and Challenges” [Online] Available: <https://arxiv.org/pdf/2101.05428>
- [4] Lois B. Travis, “The-Epidemiology-of-Second-Primary-Cancer” [Online] Available: <https://aacrjournals.org/cebp/article/15/11/2020/275025/The-Epidemiology-of-Second-Primary-Cancers>
- [5] 衛生福利部國民健康署 [Online] Available: <https://www.hpa.gov.tw/Pages/List.aspx?nodeid=4050>
- [6] University of Colorado Boulder “Data Analytics Best Practices: Best Machine Learning Model for Sparse Data” [Online] Available: <https://www.colorado.edu/business/ms-programs/ms-blog/2023/01/24/data-analytics-best-practices-machine-learning-sparse-data>
- [7] Luis R. Soenksen et al. “Integrated multimodal artificial intelligence framework for healthcare applications” [Online] Available: <https://www.nature.com/articles/s41746-022-00689-4>
- [8] Zhen Xu et al. “MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records” [Online] Available: <https://arxiv.org/pdf/2102.02340>
- [9] Adrienne Kline1 et al. “Multimodal machine learning in precision health: A scoping review” [Online] Available: <https://www.nature.com/articles/s41746-022-00712-8>
- [10] Valerio Guarrasi et al. “A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications” [Online] Available: <https://arxiv.org/pdf/2408.02686>
- [11] IBM. “What is ensemble learning?” [Online] Available: <https://www.ibm.com/think/topics/ensemble-learning>
- [12] Palak Mahajan et al. “Ensemble Learning for Disease Prediction: A Review” [Online] Available: <https://www.mdpi.com/2227-9032/11/12/1808>

- [13] Subasish Mohapatra et al. “Heart Diseases Prediction based on Stacking Classifiers Model” [Online]. Available:
<https://www.sciencedirect.com/science/article/pii/S1877050923001400>
- [14] J.-F. Hong and Y.-J. Tseng, “Performance vs. Privacy: Evaluating the Performance of Predicting Second Primary Cancer in Lung Cancer Survivors with Privacy-preserving Approaches,” in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Sep. 2022, pp. 01–04.
- [15] “Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions” [Online]. Available:
<https://arxiv.org/pdf/2208.03392.pdf>
- [16] Yin Cui et al. “Class-Balanced Loss Based on Effective Number of Samples” [Online] Available: <https://arxiv.org/pdf/1901.05555>
- [17] Yoav Freund and Robert E. Schapire,” A Short Introduction to Boosting” [Online]. Available:
<https://cseweb.ucsd.edu/~yfreund/papers/IntroToBoosting.pdf>
- [18] Francis Sahngun Nahm,” Receiver operating characteristic curve: overview and practical use for clinicians” [Online] Available:
<https://synapse.koreamed.org/articles/1159825>
- [19] Scott M. Lundberg and Su-In Lee,” A Unified Approach to Interpreting Model Predictions” [Online] Available: <https://arxiv.org/pdf/1705.07874.pdf>

補充說明:

本計畫的演算法是以集成學習為發想基礎，解決各機構獨有特徵問題，並以肺癌患者罹患二次癌症作為檢驗模型效能的資料集。與陽明交通大學資訊工程學系張育安同學分析二次癌症台灣與美國各機構特徵分布並利用模型融合解決各機構獨有特徵的解法不同。