## （一）摘要

資料特徵的整合一直是機器學習中一個極具挑戰性的課題，尤其在跨機構學習和聯邦式學習中更為明顯。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。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，並將模型預測效能與以填補數值訓練的模型進行比較。

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，評估模型在評估二癌風險的準確度。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。

## （二）研究動機與研究問題:

1. 研究背景

1.1 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。隨著資料治理與隱私保護觀念的進步，個人資料開始受到嚴格保護，跨機構取得資料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，使得各設備端可以在本地訓練模型，無需將數據傳送到中心化的位置。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。

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［1]。

1.2 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，不同於先前所得到的癌症，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機率，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［2］。

1.3 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［3］。

2. 研究問題與目標

過去使用聯邦式學習預測肺癌患者罹患二次癌症發生機率的研究中，各機構使用同樣的資料特徵完成模型學習［4］。然而，各機構甚至是各國可能會有機構專屬的特徵，其中可能含有對二次癌症預測的有用資訊。當不同機構的資料各有不同特徵時，會導致訓練變得非常麻煩。而一般的解決方法，是直接將不同的特徵刪除，或是將所有機構的特徵取聯集，各機構再針對缺少的特徵進行補值。此方法固然實用，可是會造成資料減少，或是出現大量的填補數值，進而降低模型的效能。不僅如此，各機構中若有類別的特徵，必須先使用 encoding 技術將資料轉換為適合訓練模型的型態，再取不同機構特徵的聯集。如此一來，過多的填補數值很可能會嚴重影響模型的訓練與表現。

**因此，本計畫的目標為提出一個解決各機構擁有不同特徵卻能一起使用聯邦式學習訓練模型的機器學習方法，並利用肺癌患者得到二次癌症風險資料，測試此演算法的效能，最後嘗試解釋特徵的重要性。**

## （三）文獻回顧與探討

1. 各機構特徵不同，如何把獨有的特徵納入模型

在聯邦式學習中，Personalized Layer 是一種特定於各機構的層，它被添加到模型架構中，以允許模型根據各機構不同的特徵進行調整。這種技術常應用於醫療系統，因為地域性的因素，各機構蒐集的資料特徵往往有些差異，Personalized Layer 則恰巧解決了此問題［5］。

儘管 Personalized Layer 在提高模型性能方面具有潛在的優勢，但也存在一些缺點：

- 參數學習困難：由於 Personalized Layer 的參數需要根據各機構的特徵進行調整，因此可能需要更多的數據來訓練這些參數，尤其是在用戶或設備之間的差異較大時。

- 共享學習效率降低：聯邦學習的目的之一是通過合作學習改進整體模型，但 Personalized Layer 可能會減少參與方之間的共享學習效率。

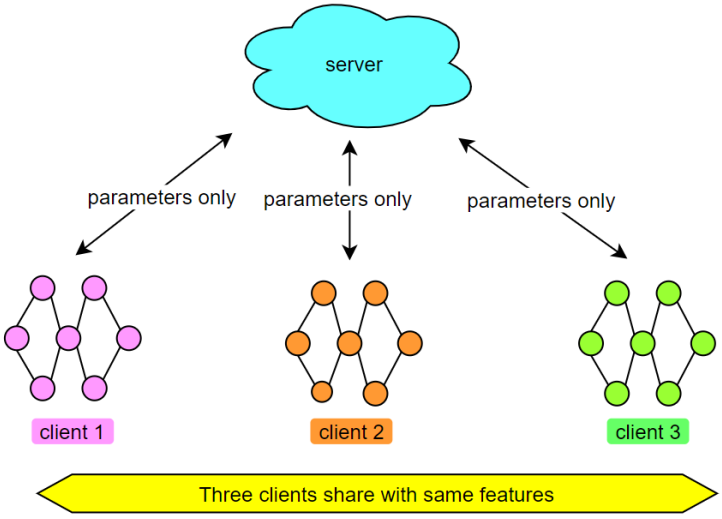
2. 聯邦式學習在醫療中的應用

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。這種方法確保敏感的患者資訊保密，同時允許醫療機構共同提高模型預測準確性［6］。

## （四）研究方法及步驟

**過去研究使用的方法 -- 利用 NN 神經網路的聯邦式學習模型**

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，取所有機構的特徵聯集，將各機構中缺少的特徵數補零。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（圖一）。



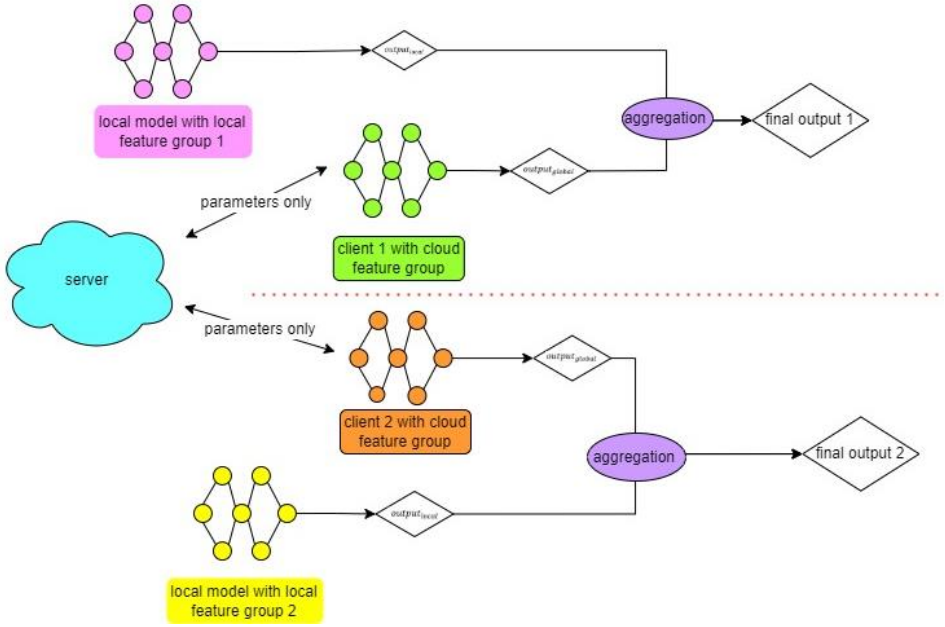

圖一 聯邦式模型示意圖

**本研究實驗模型設計**

本研究提出的演算法是根據特徵相似成度將特徵分成本地組和雲端組。雲端組特徵在 client 端訓練，參數於 server 端統整，而本地組特徵僅在 client 端訓練，得到兩個模型的預判結果（output）。再利用本研究所提出的兩種 aggregation 方法將結果合併，得到更精準的答案。

如圖二所示，訓練過程是使用各 client 雲端組特徵的資料，以聯邦式學習訓練模型 (global model)，local model 則是依照各機構獨有特徵在 client 端訓

練。訓練結束之後，利用 validation set 調整本研究所提出 aggregation method 中使用的參數。最後，利用 test set 檢測模型的效能。




圖二 完整模型示意圖 n = 2（n 為 client 數目）

**資料前處理:**

**資料前處理步驟 1: 處理類別特徵**

先整理出各機構中所有的類別特徵，依照特徵的相似程度，按下列步驟分成雲端組和 n 個本地組(n 為 client 數目)。此實驗中，是以不交換任何原始資料為前提實測，僅交換資料特徵與編碼方式。

**情況 1: 在其他機構中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 SEER 資料庫)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。此類特徵定義為雲端組。

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** | 5606 | 514904 |
| 2: **Female** | 4939 | 458037 |

**情況 2: 在其他機構中的相同特徵卻有不同的種類**

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，由下列公式判定加入雲端組或是本地組。當不等式成立時，將特徵放入雲端組，反之，放入本地組。本公式的意義為檢測種類不同數目的多寡。雖然特徵相同，可是種類數目卻很大不一致時，應將資料放入本地組，當作各機構獨有的特徵。相反的，若各機構僅只缺少或多出些許不同的種類時，我們應該將其放進雲端組，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，我們可以清楚看到 $\frac{6-5}{5} < 2$，因此，將肺癌類別放入雲端組。

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

表二 診斷年齡於機構間的資料分佈範例

| 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 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。此類機構獨有特徵對該機構的預測可能有幫助，因此放進各機構的本地組訓練集。

**資料前處理步驟 2: 處理連續性的特徵**

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等。若特徵內容相同，則可將該特徵對應的資料放入雲端組，若特徵不相同，則將該特徵的資料放入本地組。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。

**聯邦式學習使用架構**

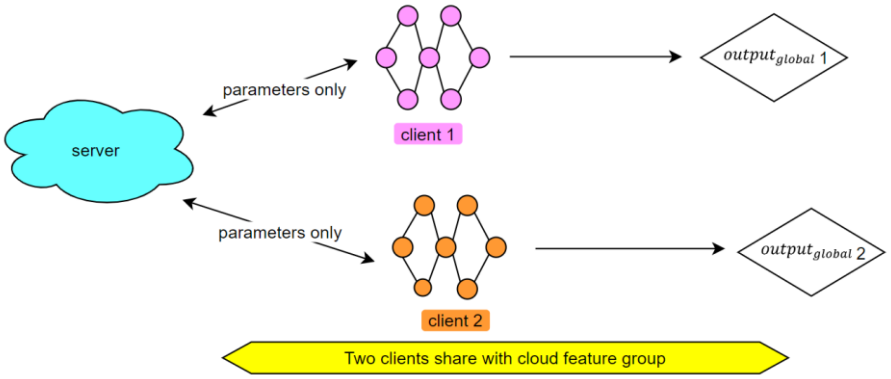
本研究所使用的聯邦式學習架構為 NN 的神經網路模型。各機構(client)先在本地端訓練模型，再將模型參數上傳到 server，server 將參數加權平均之後，再將平均參數下放給 clients。如此操作，訓練資料就會留在各機構的裝置中，不用共享資料也可以達到訓練模型的目的。

**模型訓練**

將資料依照 7:2:1 分別拆成 training set、validation set 和 test set。

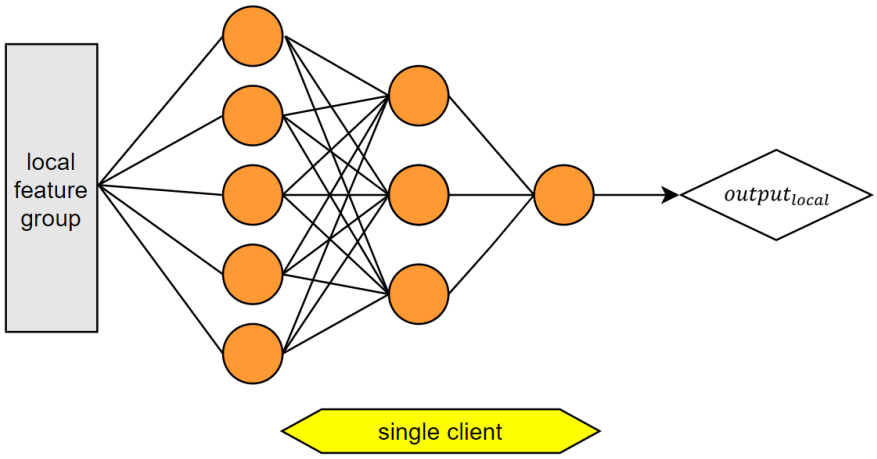
**模型訓練步驟 1：使用雲端組資料和本地組資料，各自訓練模型**

如圖三，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（$output_{global}$）。




圖三 使用雲端組特徵於聯邦式學習示意圖

如圖四，各機構分別使用本地組獨有的特徵作為資料集，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($output_{local}$)。




圖四 使用本地組特徵於本地端訓練模型示意圖

**模型訓練步驟 2：合併雲端組模型與本地組模型**

本計畫預計使用兩種方式合併步驟 1 訓練所得雲端組和本地組模型。

A. **合併方法一：**

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 [7] 演算法，結合 Linear Regression，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**。

Ex: 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$\varepsilon_{global}$，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。

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

參數 $\varepsilon_{local}$ 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。

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

參數 $\varepsilon$，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 ($\Delta w$) 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。

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

將 validation set 放入 global 模型和 local 模型，得到兩個結果，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 F-SCORE 當作初始化的因素。對於 F-SCORE 較高的模型，會享有較高的權重。

$$F - SCORE = \frac{(1+\beta^2)precision * recall}{\beta^2 precision + recall} \qquad \beta \in R^+$$

$\beta$ 的選擇也會是本研究探討的問題之一。選擇 $\beta$ 較大的值，$F - SCORE$ 會有趨近於 recall 的趨勢。在此研究中的(五)預期結果中也會探討為什麼模型的 recall 值越高越合適。

$$w_{global/local} = \frac{f - score_{global/local}}{f - score_{global} + f - score_{local}}$$

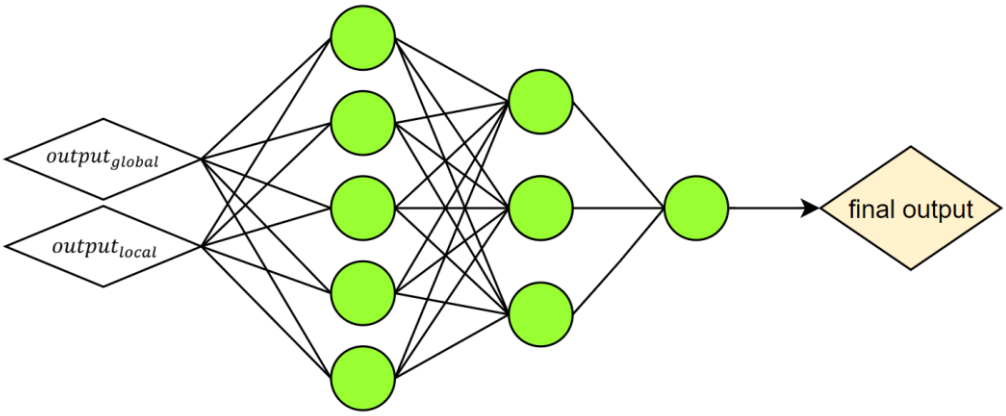
圖五顯示更新權重的方法。True False 為該模型預測結果正確或錯誤，可以透過 $\varepsilon_{local}$、$\varepsilon_{global}$ 得知。當 $cg = cl$ 時，表示兩者對於事件皆有相同的信心程度，因此，不調整權重。

<table>
<tr><td colspan="2"></td><td colspan="2">global</td></tr>
<tr><td colspan="2"></td><td>true</td><td>false</td></tr>
<tr><td rowspan="3">local</td><td rowspan="2">true</td><td>$cl < cg$<br>$w_{global} = w_{global} + \Delta w$<br>$w_{local} = w_{local} - \Delta w$</td><td rowspan="2">$w_{global} = w_{global} - \Delta w$<br>$w_{local} = w_{local} + \Delta w$</td></tr>
<tr><td>$cl > cg$<br>$w_{global} = w_{global} - \Delta w$<br>$w_{local} = w_{local} + \Delta w$</td></tr>
<tr><td>false</td><td>$w_{global} = w_{global} + \Delta w$<br>$w_{local} = w_{local} - \Delta w$</td><td>$cl > cg$<br>$w_{global} = w_{global} + \Delta w$<br>$w_{local} = w_{local} - \Delta w$<br><br>$cl < cg$<br>$w_{global} = w_{global} - \Delta w$<br>$w_{local} = w_{local} + \Delta w$</td></tr>
</table>

圖五 權重參數調整表

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。再利用這些結果來反覆操作步驟二，進而得到最佳的權重。

B. **合併方法二:**




圖六 模型示意圖

將 Validation set 輸入到兩個模型中，得到兩模型的結果後，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] → $New\ input$ = [0.8 0.1 0.1 0.6 0.3 0.1]。

如圖六，利用 NN 神經網路模型能模擬多種函數的特性，進一步找出由本地組和由雲端組訓練的兩個模型間的關係。

表三 方法一和方法二的優劣比較

| | 優點 | 缺點 |
|---|---|---|
| **方法一** | 1. 快速求得特徵對於模型輸出的貢獻度<br>2. 降低時間訓練成本 | 1. 只能找出簡單的線性關係，無法處理複雜的關係 |
| **方法二** | 1. 可以模擬複雜的函數，提升模型的精準度 | 1. 容易發生過擬和(overfitting)<br>2. 時間成本大幅提升<br>3. 求得特徵對於模型輸出的貢獻度難度提升 |

**模型效能驗證方法**

**1. 肺癌患者罹患二癌資料 -- 資料來源**

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，分別為長庚、中山、中慈、亞東、仁愛。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 筆資料，SEER 則有 972941 筆資料。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。此次研究，是以肺癌患者罹患二次癌症的風險為分析與預測目標。

**2. 特徵重要性解釋**

Shapley Additive explanations (Shap) 為一種量測特徵重要性的方法。透過下列公式計算參數的 shap 值，得到該參數對於整體模型輸出的貢獻程度［8］。

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\cdots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{x_j\}) - val(S))$$

## （五）預期結果:

**檢測模型與演算法效能**

將本計畫提出的兩種模型架構以及權重調整方式和單純使用聯邦式學習的模型比較，檢測預測二次癌症發生的效能是否提升。效能的檢測將利用 accuracy、precision、recall、f-score、mcc、AUC 等多種方式檢測。本次資料集的類別高度不平衡，因此綜合考量 f-score、mcc、AUC 等效能評估方法，綜合討論模型優劣。

對於醫療的模型來說，高度的資料不平均是常有的事情，因此，模型不能意味的輸出不可能發生。本研究也會考慮 recall 值高，但 accuracy 較低的模型，針對他做討論。

最終，本計畫也會計算各種訓練模式的時間成本，作為模型效能好壞的評判標準之一。此外，本計畫也預計分析比較各模型間的特徵重要性，並討論其差異。

**特徵重要性解釋與討論:**

- 僅考慮共同特徵，利用 NN 神經網路的聯邦式學習模型: 透過 shap 公式計算出
- 考慮各機構獨有的特徵，使用合併方法 1 “Seesawing weights algorithm”: 過 shap 公式與權重計算出
- 考慮各機構獨有的特徵，使用合併方法 2 “Using NN network to merge the result”: 試圖利用 shap 與多層推導的方式得到特徵重要性

## （六） 需要指導教授指導內容

- 研究執行進度追蹤控管
- 文獻探討方向
- 醫療知識援助
- 研究報告撰寫指導
- 演算法優化建議

## （七） 參考文獻

[1] Adrian Nilsson et al. “A Performance Evaluation of Federated Learning

Algorithms” [Online] Available: https://dl.acm.org/doi/pdf/10.1145/3286490.3286559

[2] 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

[3] 衛生福利部國民健康署 [Online] Avaliable: https://www.hpa.gov.tw/Pages/List.aspx?nodeid=4050

[4] 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.

[5] Manoj Ghuhan Arivazhagan and Vinay Aggarwal,” Federated Learning with Personalization Layers” [Online] Available: https://arxiv.org/pdf/1912.00818.pdf

[6] “Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions” [Online]. Available: https://arxiv.org/pdf/2208.03392.pdf

[7] Yoav Freund and Robert E. Schapire,” A Short Introduction to Boosting” [Online]. Available: https://cseweb.ucsd.edu/~yfreund/papers/IntroToBoosting.pdf

[8] 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:

本計畫的核心為實作新提出之模型演算法以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。與陽明交通大資訊工程學系張育安同學的個人化層處理非共用特徵及分布不同特徵的計畫有所不同。