醫學電資創意整合專題

第一組

主題：使用機器學習模型分析肝癌病患手術後的復發

指導老師：吳沛遠教授

作者：

學號	系級	姓名	工作分配
R12945065	生醫電資一	胡榮澤	資料處理 實作：預測病患術後復發時間 製作簡報：預測病患術後復發時間、結論 口頭報告：預測病患術後復發時間 書面報告：預測病患術後復發時間
R12945079	生醫電資一	蕭智陽	資料過濾與清理 規劃研究方向 協調、聯繫、整合組員間合作 提供臨床建議 製作簡報：背景、目的、方法 口頭報告：回答報告後提問
B10902027	資工三	曹哲維	實作：預測病患是否 2 年內復發(KNN 除外) 書面報告：預測病患是否 2 年內復發
B11902023	資工二	黃梓宏	資料處理 實作：預測病患是否 2 年內復發(KNN 除外) 製作簡報：預測病患是否 2 年內復發 口頭報告：預測病患是否 2 年內復發
B10401034	醫學三	許祐禎	資料前處理 製作簡報：背景、目的、方法 書面報告：統整
B10401022	醫學三	黃胤程	資料前處理 製作簡報：背景、目的、方法 書面報告：統整
B11401016	醫學二	涂凱翔	實作：預測病患是否 2 年內復發(KNN) 製作簡報：背景、目的、方法 口頭報告：研究背景、目的、方法、結論 書面報告：最終統整

中華民國 113 年 6 月

一、研究動機

肝癌是臺灣致命癌症的第二位，在台灣、中國及日本等亞洲國家的發生率比較高，西方國家的發生率相對較低。它的主要成因與 B 型肝炎、C 型肝炎、酒精性肝病、非酒精性脂肪性肝病以及肝硬化有關。肝癌分成多種類型，常見的如下：

1. 由肝細胞轉化形成的肝細胞癌（hepatocellular carcinoma, HCC），為最常見的肝癌種類。
2. 由膽管細胞轉化形成肝內膽管癌。
3. 由其他臟器轉移至肝臟的癌症：如：胃腸道癌症、胰臟、肺臟、乳癌、等。

罹患肝癌的病人的存活時間變異很大，可能小於 3 個月或大於 5 年，而常見的治療方式包含手術或藥物等。但接受手術治療的患者，復發機率仍高達六成。我們這次專題分析對象為最常見的肝細胞癌患者，試圖在病人資料中預測未來癌症復發狀況。

二、研究目的

(一)利用機器學習模型分析現實中肝細胞癌病患的資料。

(二)預測肝細胞癌手術後復發的狀況。

(三)比較不同模型的預測表現。

(四)了解最可能影響復發狀況的參數。

三、文獻探討

根據 2023 年一篇論文[1]，我們整理出肝癌可能影響復發的預後因子，包括：

1. 腫瘤大小
2. 腫瘤數量
3. 組織分化程度
4. 是否侵犯血管
5. 是否有轉移

四、研究方法

(一)病患組成

病患均為曾在台大醫院接受肝細胞癌相關手術，紀錄時間自 2000 年起至 2012 年止，共來自 2314 名病患。

(二)資料參數

我們在原始資料中所提取的數據包括年齡、性別、家族病史、B 型肝炎、C 型肝炎、肝功能(ALBI grade、肝硬化情形)、AFP、腫瘤分期、腫瘤特徵（大小、是否有血管侵犯、病理等級、切除邊緣）等共計 22 個參數。

(三)分析目標

1. 以二元分類(Binary classification)的方式預測個案是否為早期復發，若在 2 年內復發即定義為早期復發，此部分只有追蹤超過 2 年的個案會被納入分析。
2. 預測病患多久會復發，以事件發生所需時間之標準(time-to-event)去觀察。

五、研究過程與結果

(一)資料前處理

在套用相關程式預測前，我們依據組內醫學相關背景的組員經驗，挑選了數個較沒有相關性的 22 個獨立參數進行分析。並去除掉有數據缺失導致演算法無法進行的病患資料。

(二)預測病患術後 2 年內是否會復發

我們總共使用了 6 個不同的模型來進行這一預測，包括：

1. 線性模型：Linear Regression
2. 非線性模型：SVM RBF
3. 基於樹的模型：Decision Tree、Random Forest、AdaBoost
4. 基於鄰近的模型：K Nearest Neighbor

資料切分出 15%作為 Test data，剩下的部分平分成 5 份進行 5-fold cross validation。並且對於 KNN 以外的模型，都進行 25 次取平均數及標準差。得到的結果如下：

表 1：各種二元分類模型預測正確率

Accuracy	Train	Validation	Test
Linear Regression	0.696±0.009	0.675±0.039	0.667±0.007
SVM RBF	0.742±0.008	0.664±0.034	0.673±0.012
Decision Tree	0.725±0.020	0.634±0.031	0.643±0.017
Random Forest	0.790±0.007	0.666±0.031	0.689±0.013
AdaBoost	0.745±0.010	0.669±0.039	0.631±0.010
KNN	0.699±0.006	0.693±0.009	0.646±0.016

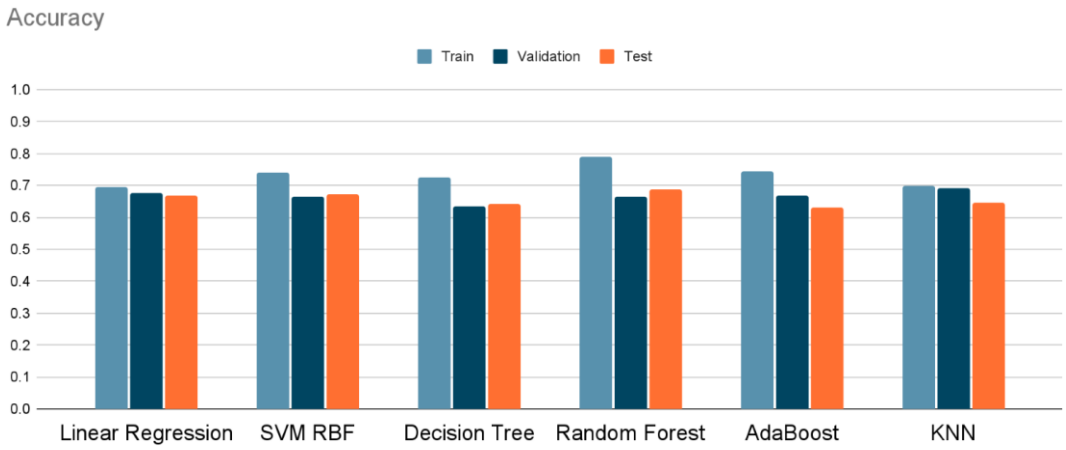

圖 1：各種二元分配模型預測正確率長條圖

我們發現在六種模型當中，表現由好至差為 Random Forest>SVM RBF>Linear Regression>KNN>Decision Tree>Adaboost，但六個模型的表現沒有差很多，對測試集的準確度都在 0.6 至 0.7 之間。其中表現略好的 Random Forest 是一種比較複雜的模型，在處理各種複雜數據集時，通常能夠提供穩健且準確的預測結果，但從我們跑的結果可以看出他存在著 overfitting 的問題（訓練集的準確度明顯較高）。另外，對於 KNN，我們在測試前就猜測它的表現會相對較差，因為它將所有參數視為同等重要，但以醫學的角度來看，病患的各種參數對肝癌復發與否影響的重要性各不相同。

(三)預測病患術後復發時間

我們總共用了 2 個不同的模型進行這一預測，包括：

1. Cox Proportional Hazards Model：使用最廣泛也是最經典的生存回歸模型，假設解釋變數對危險率的影響具有可乘性的並且隨時間恆定。
2. Random survival forest model：是 Random forest 的一種變體，通過創建一組決策樹來處理處理右缺失的數據。

我們以兩種方式衡量模型預測的表現，包括：

1. C-index：衡量模型預測的風險排序與實際生存時間排序的一致性。如果模型預測的高風險個體確實具有較短的生存時間，那麼模型的一致性就高，C-Index 的值也就越高。

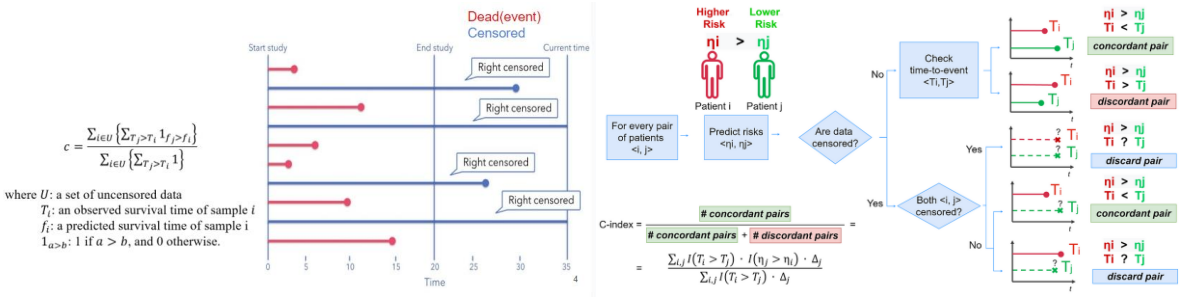

圖 2：C-Index 計算方式示意圖

若在一個群體中，隨機取出兩個人，若觀測值 j 的實際生存時間大於 i 的生存時間，並且模型預測 j 的生存時間也大於 i 的生存時間，則說明該組預測是正確，C-index 可以理解為在一個觀測群體中，所有組別中正確組別所占的比例。C-Index 為 0.5 表示模型的預測效果與隨機猜測相當，而接近 1 表示模型的預測排序與實際排序高度一致，說明模型性能優越。若在臨床上記錄病人數據時，並沒有紀錄到全程，而是中途因故無法繼續蒐集該病患資料，這種情況被稱為數據的「右缺失」，若因而導致無法確定生存時間的大小關係時，該組預測就捨去不計。

2. 生存曲線

並且我們繪製生存曲線(Kaplan-Meier curve)作為衡量模型優劣的另一個指標。我們將病患資料依據模型預測結果，以中位數拆分成高風險群和低風險群，理想情況下，高風險群的生存曲線應該要位在低風險群的下面，若兩條曲線差距越顯著(p-value 越小)，就表示模型越好。

我們將資料切分出 15%作為 Test data，剩下的部分平分成 5 份進行 5-fold cross validation。然後運行這個模型 25 次。記錄下平均值和標準差。

1. 得到 C-Index 的結果如下：

表 2：Cox Proportional Hazards 及 Random Survival Forest 的 C-Index

	Train	Validation	Test
Cox Proportional Hazards	0.670±0.004	0.657±0.005	0.649±0.020
Random Survival Forest	0.828±0.002	0.655±0.005	0.657±0.017

我們發現 Cox Proportional Hazards model 和 Random Survival Forest model 在 testing data 的 perfomance 都大概在 0.65 左右。並且 Random Survival Forest model 的 C-index 略高於 Cox model。

2. AUC 對時間的關係

因為復發時間是具有時間依賴性的數據，在生存分析中，我們通常不僅關注事件是否發生，還關注何時發生。因此我們畫出 AUC 對時間的關係圖，以得知如果以該時間作為分界，預測該時間時病患是否會復發的二元分類問題的預測情況。

對於 Cox Proportional Hazards，結果如下：

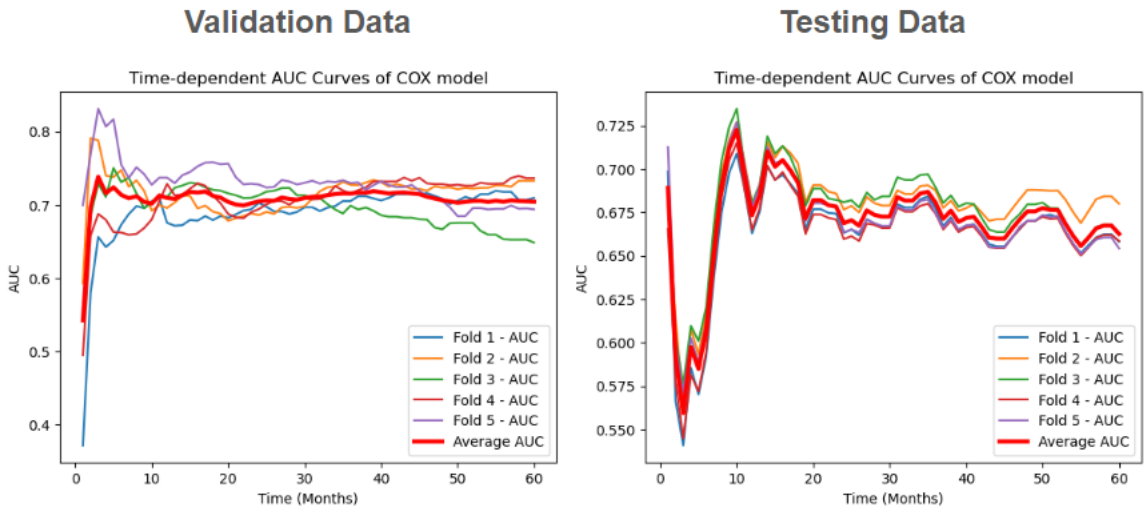

圖 3：Cox Proportional Hazards 的 AUC 對時間折線圖(左：驗證資料、右：測試資料)

在 Cox Proportional Hazards 的模型中，我們發現驗證資料的二元分類預測狀況在時間很短的時候很低，但很快就增加穩定到 0.7 左右；測試資料則可以看到預測狀況在時間較短時有一個突然的減低，但之後還是增加我們發現驗證資料的二元分類預測狀況在時間很短的時候很低，但很快就增加穩定到 0.7 左右；測試資料則可以看到預測狀況在時間較短時有一個突然的減低，但之後還是增加並穩定在 0.65 左右。

對於 Random Survival Forest，結果如下：

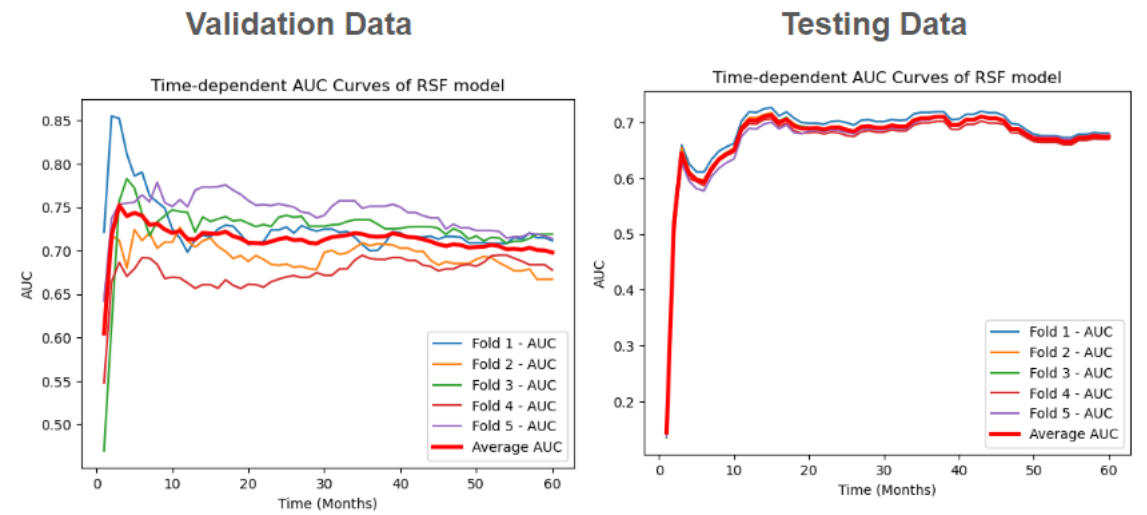

圖 4：Random Survival Forest 的 AUC 對時間折線圖(左：驗證資料、右：測試資料)

在 Random Survival Forest 的模型中，我們發現驗證資料的二元分類預測狀況在時間很短的時候很低，但很快就增加，並也是穩定到略高於 0.7；測試資料則可以看到預測狀況在時間較短時也有一個突然的減低，但沒有 Cox Proportional Hazards 明顯，之後還是增加並穩定在接近 0.7。

比較兩種模型以後，我們發現 Random sruvival forest 的結果比 Cox Proportional Hazards 略好一些。兩者在時間足夠長的情況下，AUC 均達到 0.7 左右。

3. 生存曲線分析

在繪製生存曲線時，首先也是將資料分割成出測試資料，並將剩餘資料進行 5-fold cross validation 分成訓練資料與驗證資料。資料分別通過 Cox Proportional Hazards model 和 Random Survival Forest model 回歸分析得到 risk score，以 risk score 中位數將資料拆分成高風險組和低風險組。然後再將這兩個組別各自進行生存曲線的繪製。結果如下：

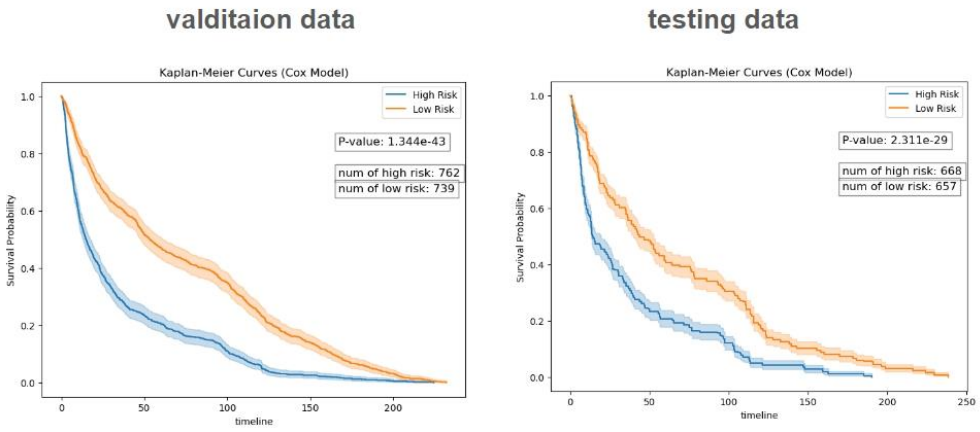

圖 5：Cox Proportional Hazards 的生存曲線(左：驗證資料、右：測試資料)

可以看出在 Cox Proportional Hazards model 中，validation data 的表現顯著高於 testing data 的結果。

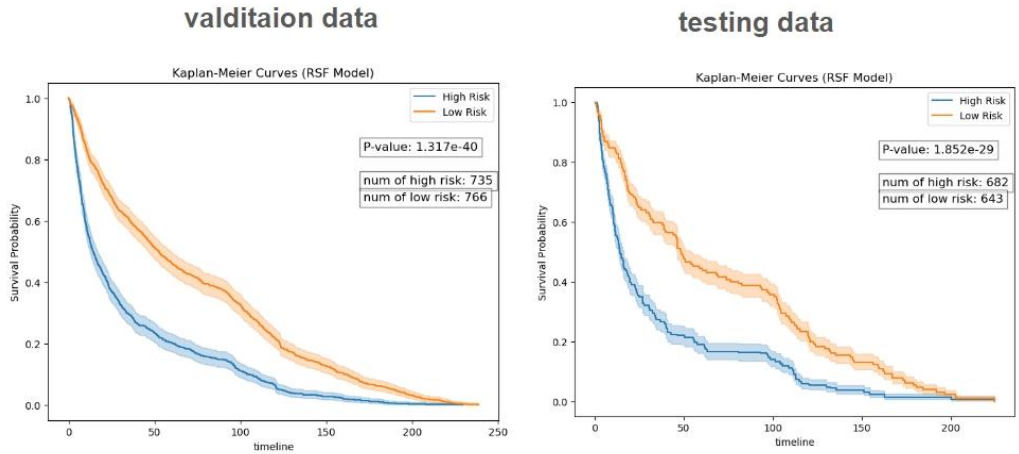

圖 6：Random Survival Forest 的生存曲線(左：驗證資料、右：測試資料)

可以看出在 Cox Proportional Hazards model 中，validation data 的表現顯著高於 testing data 的結果。

4. 各參數的 Hazard ratio 比較

最後我們想單獨去分析這些參數對生存回歸分析影響的權重大小，因此我們將這些參數單獨以 Cox Proportional Hazards model 進行生存回歸分析，並且計算其 Hazard ratio 及其 95%信賴區間。得到的結果如下：

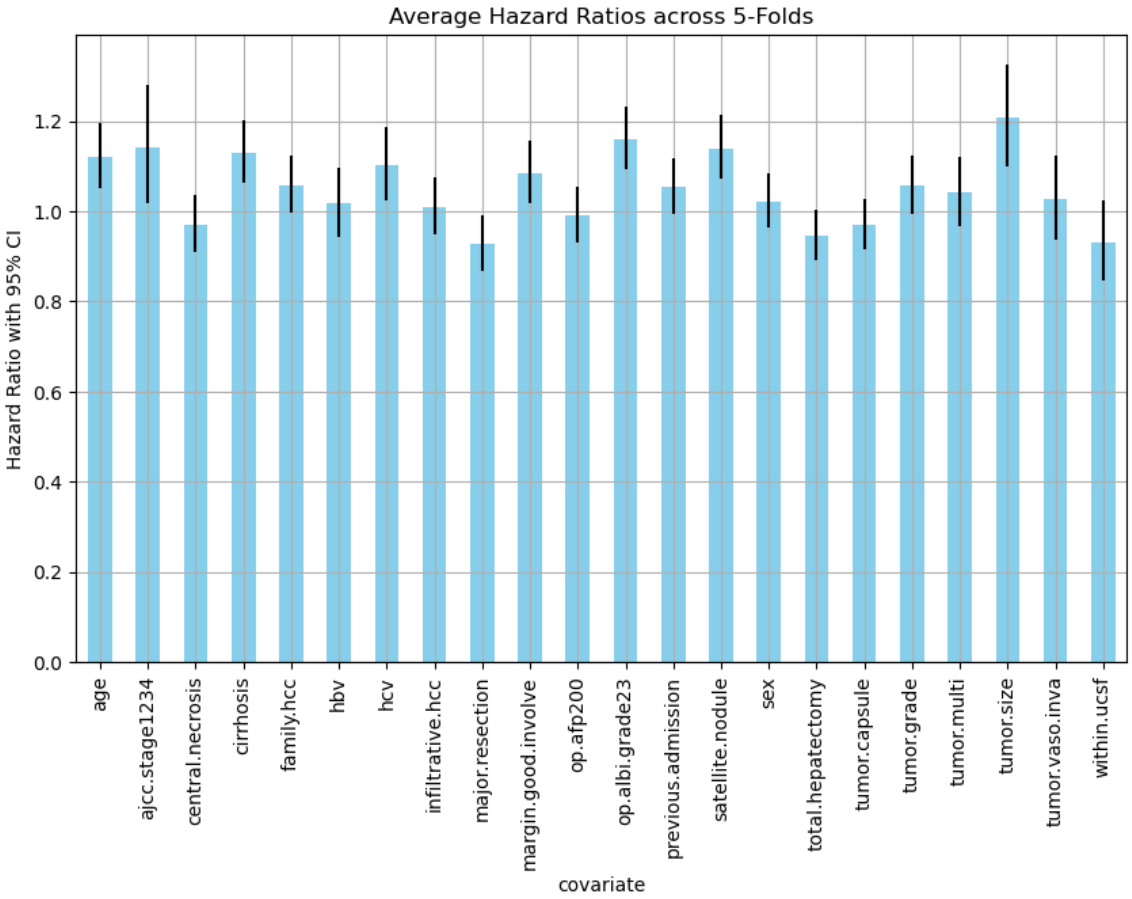

圖 7：各參數之 Hazard ratio 的 95%信賴區間

其中年齡、腫瘤分期、肝硬化、C 型肝炎、切除邊緣、肝指數中的 ALBI 指數(與病患抽血中白蛋白與膽紅素的濃度有關的指數)、是否有衛星轉移、腫瘤大小的 Hazard ratio 顯著大於 1。可以看出在這些參數對模型結果的影響較大，也是最有可能影響肝細胞癌病患術後復發的參數。其中又以腫瘤大小最為重要。

六、討論

1. 為什麼 AUC 在時間很短的時候會比較低？

我們認為部分患者的肝細胞癌復發的原因是肇因於操作手術的醫師的人為疏失，導致病患的腫瘤並沒有被完整清除，這些病患或許復發率本來其實比較低，模型也從該病患的參數預測出復發率低的結果，只是因為醫師的疏失導致病患快速復發。這些資料可能影響模型的早期預測。

2. 為什麼預測正確率大約只有 0.7？

我們認為是因為癌症的復發本來就容易受到很多狀況的影響而難以預測。若將本次研究中分析的病患資料以 t-SNE 降至 2 維後視覺化於座標平面上，其中越接近紅色的點表示越晚復發的病患，越接近藍色的點表示越早復發的病患，共分成 10 個等級，而左圖額外加入了灰色方形則代表研究中沒有復發紀錄的病患。可以稍微粗略得看到較晚復發的患者多偏向圖的左側，而較早復發的患者則多偏向圖的右側，但彼此有很嚴重的重疊，導致進行二元分類或預測時並沒有辦法很好的將復發早或晚的患者分開。

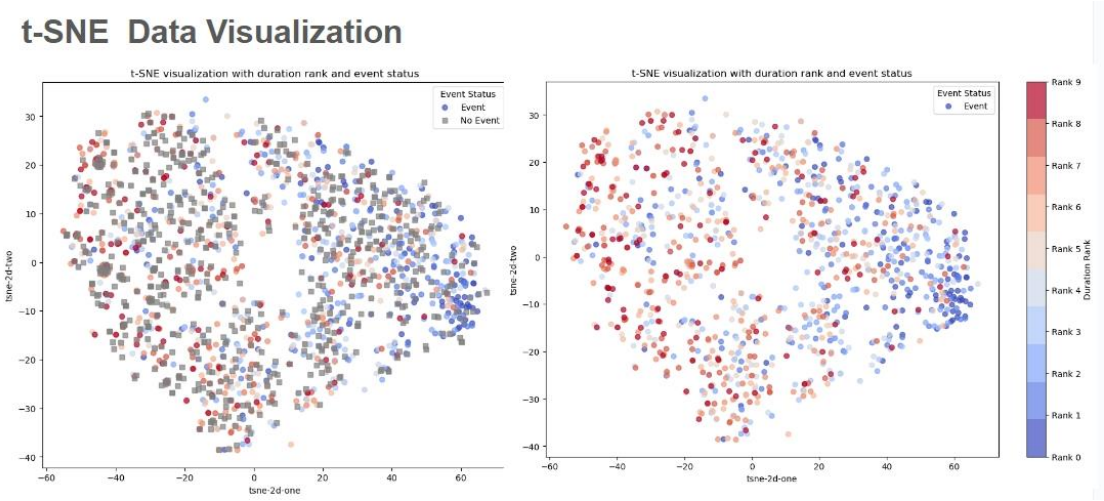

圖 8：病患各參數壓縮成 2 個維度以後的視覺化分布

在臨床上即便是醫師，也難以從病患的檢查數據或其他參數，就精準預測該病患是否會在 2 年內復發，更不用說精準預測病患復發時間，最多也只能依據參數大概猜一下病患復發風險高或低，因此模型預測正確率 0.7，或許已經達到等同於臨床醫師的標準，因此並不能因為正確率僅 0.7，就否定了這份研究的模型。在臨床上，這個研究也能起到輔助醫師預測該病患的復發風險的作用，雖然不是完全準確，但醫師可以大概知道哪些是高復發風險的病患，可能可以安排更頻繁的定期追蹤，或建議患者在家時自行關心自己的身體狀況，提醒病患若有異狀即須回診。當然被預測為低復發風險的病患，因為模型有其不準確性，還是必須要求病患依據正常時程回診追蹤。

七、結論

1. 對於預測肝細胞癌患者術後 2 年內的早期復發狀況，以 Random Forest 的表現最佳，但其他模型的表現也沒有明顯的差異。
2. 對於預測肝細胞癌患者術後的復發時間，發現 Random Survival Forest 模型的整體表現優於傳統 Cox Proportional Hazards 模型。考慮到 Cox Proportional Hazard 是目前臨床是最常被使用的模型，我們認為我們發現了一個可能比臨床上的主要模型更好的一種模型。並且當篩選出越多特徵，模型的準確度及表現也隨著篩選出的特徵數量增加而增加。
3. 最後我們用 Cox Proportional Hazards model 篩選出與肝細胞癌患者術後復發最相關的因子，包含年齡、腫瘤分期、肝硬化、C 型肝炎、切除邊緣、ALBI 指數、是否有衛星轉移、腫瘤大小。

八、參考資料

[1]Liu R, Wu S, Yu HY, Zeng K, Liang Z, Li S, Hu Y, Yang Y, Ye L. Prediction model for hepatocellular carcinoma recurrence after hepatectomy: Machine learning-based development and interpretation study. Heliyon. 2023 Nov 19;9(11):e22458. doi: 10.1016/j.heliyon.2023.e22458. PMID: 38034691; PMCID: PMC10687050.