科技部補助

大專學生研究計畫研究成果報告

計畫名稱：	結合資料探勘技術及集成架構提昇子宮內膜癌復發預測準確率

執行計畫學生： 吳明蓁

學生計畫編號： MOST 104-2815-C-040-060-E

研 究 期 間 ： 104 年 07 月 01 日至 105 年 02 月 28 日止，計 8 個月

指 導 教 授 ： 張啟昌

處理方式： 本計畫涉及專利或其他智慧財產權,2 年後可公開查詢

執 行 單 位： 中山醫學大學醫學資訊學系

中華民國 105 年 03 月 30 日

(一) 摘要

本研究規劃使用支援向量機 (Support Vector Machine, SVM)、快速學習器 (Extreme learning machine, ELM)、C5.0 決策樹 (Decision Tree)以及隨機森林 (Random Forests , RF)四種資料探勘法，探討子宮內膜癌復發的危險因子並利用集成學習提高整體的準確度。子宮內膜癌在臨床上通常是依據疾病的發展提供適合的進程治療。因此，對於癌症復發徵候的偵測及其後續無症狀復發事件的觀察而言，是與個體的存活率密切相關。過去很多研究缺乏實際觀察個別病患深入特定臨床路徑的移轉、復發和治療的時序關聯樣式，無法提供臨床醫師對可能的病情發展更多資訊可參考。因此，為了提高治癒率與存活率，從實際診療紀錄中找出預測復發因子提供臨床醫師治療的資訊是非常關鍵且重要。

本研究所需病理資料的來源為全民健康保險資料庫之癌症登記資料欄位。初步經由三位資深臨床醫師討論的復發的危險因子有：(1)年齡 (age) (2)組織型態 (histopathology type)(3) 性態碼 (behavior code) (4)手術邊緣 (surgical margin) (5) 病理 T (pathologic T) (6)淋巴轉移 (lymphatic metastasis) (7)腫瘤大小 (tumor grade) (8)病理 N (pathologic N) (9)臨床 M (clinical M) (10) 病理分期 (pathologic stages) (11)臨床分期 (clinical stage) (12) 分化 (differentiation) (13) 臨床 T (clinical T) (14) 臨床 N (clinical N) (15) 病理 M (pathologic M) (16)放射治療手術順序 (sequence of radiotherapy and surgery) (18) 區域淋巴結檢查數目 (regional lymph nodes examined) (17) 區域全身順序 (sequence of locoregional therapy and systemic therapy) (19) 其他放射劑量 (dose to target of other RT) (20) 最高放射劑量臨床標靶體積 (target of CTV_H)。有鑑於過去單一學習方法在分類預測準確性的缺點如：統計問題、計算問題和代表性問題。本研究除了使用 C5.0 決策樹、隨機森林、快速學習器、支援向量機分析外；針對子宮內膜癌復發的特性，本研究加以考量納入集成學習架構，亦即第一階段採用增益比例(Gain Ratio)及資訊量增益比例(Information Gain Ratio)，先行篩選出重要變數後，再進行第二階段支援向量機、快速學習器、C5.0 決策樹、隨機森林分析。結果顯示，各種方法預敏感度與特異度各有不錯的預測結果，進一步整合資料探勘技術與集成學習架構，經由變數篩選後各種方法能夠經由集成學習機制突顯各種方法的分類準確率，可以有效改善單獨資料探勘技術方法的預測結果。

關鍵字：子宮內膜癌復發、分類預測、集成學習

(二)研究動機與研究問題

子宮內膜癌是全球女性第七個常見的癌症，在台灣婦科癌症中發生率占第九位 ，為女性生殖器官第二個常見的腫瘤(衛生福利部，2016)。對於女性來說子宮內膜癌是重要的醫療問題，因子宮內膜癌的發生率正逐年增加，目前粗發生率已超過卵巢癌，晚期的子宮內膜癌復發率很高，一旦復發，能存活的機率很低。大部份的子宮內膜癌早期無明顯的症狀，多數確診個案是因為在做子宮頸檢查或大腸直腸檢查時發現，子宮內膜癌不像子宮頸癌或大腸直腸癌有比較明確且快速的方法可以被篩檢出來，一旦被發現時通常都已經擴散，所以若能早期發現及治療，就能提高治癒率及減少死亡率。治療子宮內膜癌復發是一項臨床挑戰，許多研究試圖找出影響子宮內膜癌復發因素，提高臨床的管理。過去的臨床研究與文獻查證顯示復發因素包括 20 項：(1)年齡 (age) (2)組織型態 (histopathology type)(3) 性態碼 (behavior code) (4)手術邊緣 (surgical margin) (5) 病理 T (pathologic T) (6)淋巴轉移 (lymphatic metastasis) (7)腫瘤大小 (tumor grade) (8)病理 N (pathologic N) (9)臨床 M (clinical M) (10) 病理分期 (pathologic stages) (11)臨床分期 (clinical stage) (12) 分化 (differentiation) (13) 臨床 T (clinical T) (14) 臨床 N (clinical N) (15) 病理 M (pathologic M) (16)放射治療手術順序 (sequence of radiotherapy and surgery) (18) 區域淋巴結檢查數目 (regional lymph nodes examined) (17) 區域全身順序 (sequence of locoregional therapy and systemic therapy) (19) 其他放射劑量 (dose to target of other RT) (20) 最高放射劑量臨床標靶體積 (target of CTV_H))。

隨著資訊技術的發展，資料探勘(Data Mining)技術逐漸成為臨床診療指引及教學研究上最有價值的工具。所謂的資料探勘又稱之為機器學習(Machine Learning)是從儲存於資料庫中的資料表、資料記錄及資料欄位內容裡的大量資料中分析出感興趣而隱藏於資料集內的重要資訊。利用資料探勘

方法的分類技術也已經成為國內外熱門的研究領域，在此種情況下，使用現代的資料探勘方法可找出子宮內膜癌復發重要因子之間的關聯。一般而言，輸出結果只產生一個假說的分類演算法普遍都會遭遇三個問題：統計問題、計算問題和代表性問題。然而，這些問題通常是可以透過集成學習的方法加以解決的。首先對於統計問題部分，當分類演算法的訓練資料數量過於龐大假說時，就會產生所謂的統計問題。因此，若能夠針對所有分類器所進行投票機制，將可有效的降低這種風險。其次是計算問題，常因為在分類演算法不能保證可以從假說找到一個最好可能發生狀態時，就像統計問題一樣，如果可以使用加權方式，是可以有效降低選擇錯誤而陷入本地最佳化的危險。最後關於代表性的問題，當假說不包含任何真實函式f時，代表性問題就會產生。若能提供不同權重的投票方法賦予假說條件，整體分類演算法是可以找到一個非常接近真實函式f的近似值。根據過去的研究報告證實集成學習架構能降低學習演算法的偏差和變異。有鑑於過去單一學習方法在分類預測準確性的缺點：統計問題、計算問題和代表性問題。本研究除了使用支援向量機、快速學習器、C5.0 決策樹、隨機森林分析外；針對子宮內膜癌復發的特性，本研究加以考量納入集成學習架構，亦即第一階段採用增益比例及資訊量增益比例，先行篩選出重要變數後，再進行第二階段支援向量機、快速學習器、C5.0 決策樹、隨機森林分析。預期本計畫將有以下成果：

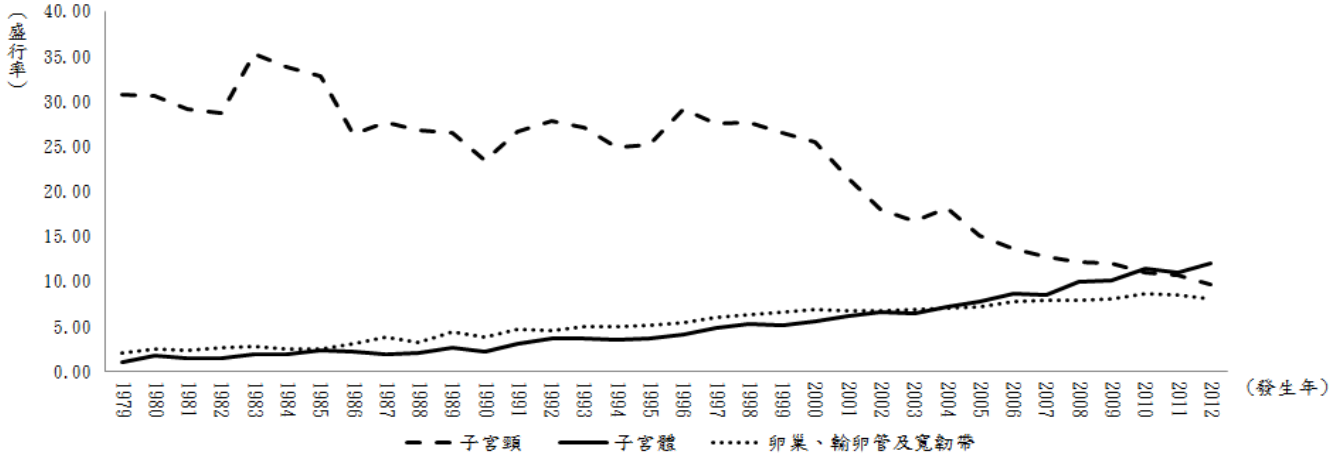
- 透過四種方法分析出來的結果，找出分類準確度較高的預測模型。
- 利用集成學習策略提高整體資料集的分類預測準確度，改善一般學習方法的缺點。
- 準確地預測子宮內膜癌患者的復發因素，提供子宮內膜癌臨床治療更佳的信息。
- 從實際診療紀錄中找出特定癌症的移轉、復發和治療的時序關聯樣式，以讓病患對可能的病情發展有更多資訊可參考，並可提供臨床醫師預測並掌握復發的危險因子。

(三)文獻回顧與探討

本研究文獻探討內容分為三個部分討論:子宮內膜癌盛行率、資料探勘方法與集成學習。

一、子宮內膜癌盛行率

子宮內膜癌在已開發國家是常見的婦科惡性腫瘤之一。國際癌症研究機構(International Agency for Research on Cancer, IARC)指出在 2015 年全球約有 288,387 位女性患有子宮內膜癌和 73,854 位婦女死於這種癌症。 子宮內膜癌在美國和歐洲女性中是第四個常見腫瘤(Ferlaya et al., 2013)；在台灣，子宮內膜癌是女性生殖器官第二個常見的腫瘤，僅次於子宮頸癌(衛生福利部，2015)。根據台灣癌症登記中心資料，在 2010 年子宮內膜癌年發病率為每 10 萬女性有 11.25 件病例，在 1980 年為每 10 萬女性有 1.69 病例(圖一和表一)。據估計 2013 年將增加 2,400 多件病例，發生率有逐年增高的趨勢。



圖一 子宮頸癌、子宮內膜癌、卵巢癌在台灣的發生率，1979-2012(來源:衛生福利部國民健康署，2016)

子宮內膜癌五年的存活率取決於診斷時的癌症期別。一般而言，五年的存活率是 81.5%(SEER，2007-2011 年)，然而 FIGO 第二期到第四期經由多重模式治療後的復發率約 25%(美國癌症協會，2013 年)。對於早期疾病，藉由手術或結合局部治療通常對病情控制是有療效的。一旦初級治療失敗，則次級治療的機會渺茫。

表一 子宮頸癌、子宮內膜癌、卵巢癌在台灣統計趨勢 (1979-2012)

年度	子宮頸					子宮體					卵巢、輸卵管及寬韌帶				
	個案數	平均年齡	年齡中位數	標準化率	癌症百分比(%)	個案數	平均年齡	年齡中位數	標準化率	癌症百分比(%)	個案數	平均年齡	年齡中位數	標準化率	癌症百分比(%)
1979	1,790	50.75	50	30.70	32.20	58	51.24	54	1.00	1.04	129	46.73	49	2.08	2.32
1980	1,827	51.04	51	30.54	28.06	99	53.65	54	1.69	1.52	162	46.62	50	2.50	2.49
1981	1,771	52.37	52	29.09	26.39	89	55.49	55	1.49	1.33	156	47.56	50	2.38	2.32
1982	1,800	52.6	52	28.62	26.98	94	52.51	54	1.47	1.41	175	46.73	49	2.56	2.62
1983	2,279	52.85	53	35.28	28.79	121	53.37	52	1.87	1.53	200	44.17	47	2.78	2.53
1984	2,262	52.51	52	33.74	28.40	124	52.32	53	1.84	1.56	179	48.74	51	2.50	2.25
1985	2,260	53.34	53	32.77	26.79	161	54.02	54	2.34	1.91	191	45.24	47	2.51	2.26
1986	1,892	52.58	53	26.27	23.25	157	53.27	54	2.19	1.93	239	47.7	48	3.09	2.94
1987	2,064	53.27	54	27.66	21.18	143	51.6	53	1.92	1.47	309	45.95	48	3.81	3.17
1988	2,063	53.01	53	26.80	21.52	162	51.05	53	2.06	1.69	265	45.84	48	3.19	2.76
1989	2,119	53.14	54	26.54	18.95	213	53.23	54	2.68	1.90	367	47.29	49	4.33	3.28
1990	1,921	53.82	55	23.33	18.58	181	52.29	52	2.21	1.75	334	46.78	49	3.81	3.23
1991	2,275	53.6	54	26.65	17.55	262	51.44	52	3.02	2.02	415	47.72	49	4.67	3.20
1992	2,480	52.45	52	27.83	17.24	311	52.38	53	3.58	2.16	407	47.94	48	4.45	2.83
1993	2,471	53.59	54	27.05	16.60	325	52.26	53	3.60	2.18	462	48	48	4.91	3.10
1994	2,360	53.56	54	24.93	15.45	327	52.57	53	3.52	2.14	480	48.16	48	4.94	3.14
1995	2,461	53.77	54	25.21	15.30	343	51.9	53	3.58	2.13	514	47.26	47	5.15	3.20
1996	2,942	53.67	54	29.07	15.89	399	52.17	52	4.04	2.16	546	48.04	48	5.34	2.95
1997	2,871	54.26	54	27.52	14.01	495	52.37	52	4.84	2.42	633	49.1	48	6.02	3.09
1998	2,976	54.61	54	27.61	13.38	549	52.5	52	5.21	2.47	677	50.3	49	6.22	3.04
1999	2,965	54.61	53	26.51	12.31	553	52.6	51	5.07	2.30	732	48.82	48	6.52	3.04
2000	2,916	55.22	55	25.42	11.60	627	52.6	52	5.56	2.50	790	49.52	49	6.89	3.14
2001	2,552	55.89	54	21.52	9.96	714	53.79	53	6.17	2.79	789	49.21	49	6.69	3.08
2002	2,193	56.11	54	17.92	8.29	786	52.65	51	6.50	2.97	817	51.26	50	6.75	3.09
2003	2,097	56.03	54	16.66	7.83	815	52.74	52	6.48	3.04	862	50.63	50	6.91	3.22
2004	2,370	56.12	53	18.13	7.87	921	52.73	51	7.11	3.06	889	51.68	51	6.98	2.95
2005	2,025	56.64	54	15.09	6.58	1,031	52.78	52	7.69	3.35	940	50.79	50	7.17	3.05
2006	1,884	56.49	54	13.60	5.81	1,188	53.65	53	8.65	3.67	1,037	51.52	51	7.74	3.20
2007	1,817	56.61	54	12.70	5.33	1,195	52.93	53	8.47	3.51	1,083	51.21	51	7.87	3.18
2008	1,781	56.74	54	12.16	4.99	1,441	53.43	53	9.88	4.04	1,130	51.83	51	7.93	3.17
2009	1,815	57.63	55	12.00	4.63	1,528	53.82	53	10.14	3.90	1,155	51.13	51	8.00	2.95
2010	1,703	57.86	56	10.95	4.19	1,757	54.06	54	11.39	4.32	1,276	51.84	52	8.68	3.14
2011	1,688	57.53	56	10.61	4.09	1,729	54.15	54	10.91	4.18	1,260	51.6	51	8.42	3.05
2012	1,567	57.57	56	9.60	3.63	1,936	54.56	55	11.96	4.49	1,236	52.53	52	8.04	2.87

(資料來源: 衛生福利部國民健康署，2016)

二、資料探勘方法

在醫療領域中，資料探勘的應用可以被用來預測疾病模式，並可預測不同群體之間的重要因子。本研究將應用以下四種不同的資料探勘方法來預測子宮內膜癌復發的重要因子：

(1) Support Vector Machine (SVM):支援向量機是由 Vladimir Vapnik 從 1995 年開始發展的一種分類方法，被視為最具成效的監督式學習方法之一，現在成為資料探勘熱門工具之一(Shutao et al., 2003)。SVM 的特性是將輸入空間(Input Space)先使用非線性的對應 Mapping 轉換到高維度的特

徵空間(Feature Space)再做分類。其中 SVM 所使用的 Mapping 在選擇所對應的核心函數上有很大的彈性，且需為非線性的函數，之後將 Mapping 到高維度的特徵空間中的資料建構線性分類式子，選擇能使分類錯誤降到最小的權重，得到最大化邊界超平面(Maximal Margin Hyperplane)以完成分類(Mao et al., 2005)。SVM 相關研究如 David(2004) 研究一個支援向量機對醫學實際資料做分類，利用量測位於螢光上交雜(Fluorescence In-Situ Hybridization, FISH) 影像細胞發生的訊號，去診斷發生的併發症狀，研究中突出測試圖樣距離的門檻值，從 SVM 分割超平面去拒絕錯誤分類的圖樣，因此可減少誤差的發生，研究結果與其他先進方法比對，指出基於 SVM 發展診斷系統的優勢潛力(David and Lerner, 2004)。

(2) C5.0：C5.0 演算法又稱為規則推理模型(rule-based reasoning model)，是 C4.5 演算法的修訂版，屬於監督式學習的一種，適用在處理大資料集，採用 Boosting 方式提高模型準確率，又稱為 Boosting Trees，在軟體上的計算速度比較快，佔用的記憶體資源較少，主要能解析連續型變數與類別型變數，結果可產生決策樹(decision tree)或規則集(rule sets) 。張惟智(2009)使用 C5.0 決策樹及類神經網路找出腹主動脈瘤手術併發症三大類併發症的分類規則及重要因子，再利用貝氏網路，找出重要因子間的因果關係並計算出其聯合條件機率。

(3) Extreme learning machine (ELM):快速學習器(ELM)是一種新型態的類神經網路架構，「快速學習的理論與應用」一文於 2006 年由 Huang 等人共同發表於 "*Neurocomputing*"上。有別於其他類神經網路，快速學習器採用截然不同的演算規則，屬於單一隱藏層的前饋式類神經網路模式(Single hidden Layer Feed-forword neural Network, SLFN)，其輸入層到隱藏層間的權重稱之為輸入權重，是隨機產生的，而隱藏層到輸出層之間的權重則稱為輸出權重，是由 MP 轉置矩陣(Moore-Penrose inverse)分析後得到，ELM 的學習速度相較於傳統的陡坡降法(gradient-based)明顯快速許多，許多文獻皆已證實此一特點(歐宗殷，2010)。Vani 等人(2010) 使用 ELM 方法應用於乳房 X 光檢查異常分類，結果表明 ELM 在分類乳房 X 光檢查異常的效能優於其他演算法。

(4) Random Forests (RF):隨機森林是 Breiman(2001)提出的一個新式決策樹演算法。是一整合多決策樹進行分類預測與重要變數(variable importance)，(Breiman, 2001 ; Liaw and Wiener, 2002; Svetnik et al., 2003)。採用分類迴歸樹(Classification and Regression Trees, CART)作為元分類器，將變數隨機投入，以 Gini 方式進行子節點分裂，Bagging 方式得出整合分類結果。隨機森林不同於傳統決策樹是，傳統決策樹，僅以單一決策樹為單位作出決策，隨機森林則以多個決策樹整合得出分類結果。對於分類與規則上相較於舊有的 CART、CHAID 與 C5.0…等擁有精確的分類預測能力(許智宇，2010)。李放歌(2011)等人認為隨機森林方法已經被用來研究乳腺癌和哮喘，研究顯示交互作用對疾病發生有影響。

三、集成學習

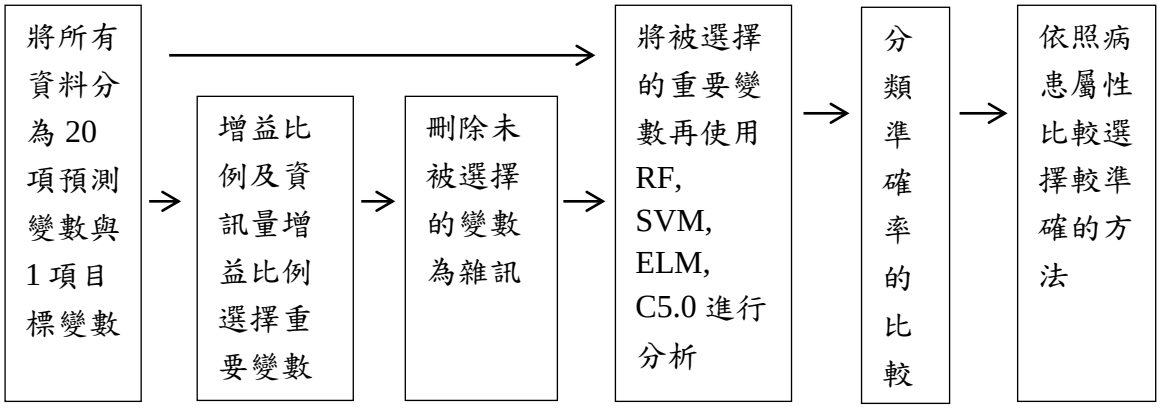
集成學習(Ensemble Learning)是透過多個分類工具加以整合成為一個新的綜合分類器，它的優點是能提供給預測模型不錯的泛化能力，進而成為一個強學習器。整體學習演算法的運作是透過多次執行基礎學習演算法，並且針對每次產生的假說進行投票，最後整合投票的結果構成一致同意的假說。一般而言，設計整體學習演算法的技巧有兩種主要的方法。第一種方法是「使用獨立的模式去建造每一個假說」，每一單獨的假說對於新資料點的預測，具有某一個合理低的出錯率，但是假說和假說彼此之間，在大多數預測裡常常是不一致的。如果能夠統合單獨假說的預測，並建立一個具有整體性時，會比起任何一個單獨或個別的分類器更具有高準確度的預測；第二種設計整體學習的方法是「採用連接模式來建造假說」。此一連接模式是把權重高的票投給和實際資料誤差小的假說，然後把權重低的票投給和實際資料誤差大的假說，藉由不同權重的投票方式結合所有的假說，並產生一個比任何單獨假說都逼近實際資料的整體假說。國際機器學習界權威 Dietterich 指出當前常見的三種集成學習策略，分別是 Bagging，boosting 和 stacking。Bagging 針對相同的演算法，去訓練出多個分類器，使用非加權的方法進行投票，即採用多數決的方法作為最後集成模型的決策。而 Boosting 利用類似 bagging 的作法，皆選用相同的演算法去訓練出多個分類器，兩者差別在於 Boosting 是採用各分類器的預測結果作加權投票準確率也較 bagging 高。Stacking 和前兩種策略最主要的不同在於可以使用不同的演算法去得到多元的分類器，在決策結果上，則可使用加權或不加權投票的處理方式(洪智力與陳勁宏，2007)。從另一個角度來看，整體學習也可以是一種附加模型(additive model)。所謂的附加模型通常是指一個新增的資料點，最後所指定的類別標籤，是由部分

或所有的附屬模型(component model)經由賦予不等的權重後，再加總所得到的結果。Freund 和 Schapire(1996；1997)提出的 Adaboost 演算法，可以說是建造附加模型極有效的方法。透過學習演算法，極盡可能地將分類錯誤減少到最小的方式去產生一個假說，每次增加一個假說到整體學習之中，分類錯誤就相對的降低。在多數的研究實驗中(Freund and Schapire, 1996；Bauer and Kohavi, 1999；Dietterich, 2000) 都說明了 Adaboost 確實可以提供大部分數據資料最好的表現結果。若針對包含較多貼錯標籤(mislabeled)的訓練資料來說，Adaboost 把非常高的權重放在雜訊的資料點上，然後生成一個非常差的整體分類器。目前確實有許多的研究工作，著重在如何延伸 Adaboost 的功能，使之能夠在處理較高雜訊的訓練資料(莊永裕，2006)。因應子宮內膜癌復發的臨床反應，本研究將採取第二種設計整體學習的方法：採用連接模式來建造假說，迫使學習演算法產生多樣化特性的目的是在每次呼叫學習演算法時，都採用一個具有不同輸入特徵的子集合。利用隨機森林整合多決策樹去選取輸入特徵的復發因素，最後形成群體特徵的重要變數後，再進行病患特性更深入的臨床解釋。

(四)研究方法與步驟

研究架構:

本研究所需的病歷記錄和病理資料的來源為全民健康保險資料庫之癌症登記資料欄位。為了比較重要變數篩選的差異，研究設計架構如圖二所示：在圖二中，本研究依據文獻查證與臨床醫師討論後決定 20 項預測變數：(1)年齡 (age) (2)組織型態 (histopathology type)(3) 性態碼 (behavior code) (4)手術邊緣 (surgical margin) (5) 病理 T (pathologic T) (6)淋巴轉移 (lymphatic metastasis) (7)腫瘤大小 (tumor grade) (8)病理 N (pathologic N) (9)臨床 M (clinical M) (10) 病理分期 (pathologic stages) (11)臨床分期 (clinical stage) (12) 分化 (differentiation) (13) 臨床 T (clinical T) (14) 臨床 N (clinical N) (15) 病理 M (pathologic M) (16)放射治療手術順序 (sequence of radiotherapy and surgery) (18) 區域淋巴結檢查數目 (regional lymph nodes examined) (17) 區域全身順序 (sequence of locoregional therapy and systemic therapy) (19) 其他放射劑量 (dose to target of other RT) (20) 最高放射劑量臨床標靶體積 (target of CTV_H))進行復發的預測。在圖二上方研究流程中未經變數篩選直接以 SVM、C5.0 決策樹、ELM、RF 方法進行預測；在圖二下方研究流程中則是藉由增益比例(Gain Ratio)及資訊量增益比例(Information Gain Ratio)篩選變數後，再以 SVM、C5.0 決策樹、ELM、RF 方法進行預測。在進一步比較兩個流程所分析分類準確率；最後針對所分析的變數結果，依照病患屬性完成臨床後續預測子宮內膜癌復發重要因子的建議。



圖二 研究流程圖

研究方法：

在醫學衛生領域中，資料探勘應用已大幅度地被用來直接取得預測不同群體之間患者的相關資訊。然而，探勘方法分類技術尚未被利用於分析子宮內膜癌復發。因此，本研究試圖利用五種資料探勘方法由子宮內膜癌的資料庫中進行分類並進一步分析集成學習架構的優勢。

一、支援向量機 (Support Vector Machine, SVM)

支援向量機廣泛被使用來處理統計分類及回歸分析，適合應用於解決具有較小範圍、非線性及高維度等特性的問題。從有限的訓練樣本中學習得到決策規則，對獨立的測試集合仍能夠得到較小的預測誤差。支援向量機將資料映射至高維空間當中，希望從映射過後的結果找出一個可將資料分

隔成兩組不同集合的超平面(hyperplane)。透過此超平面分類方法對資料進行分類，區分出互不重疊的分類集合。支援向量機從二維空間中找出一條分隔線區分兩種類型資料，且此分隔線與兩集合之距離越大越好，藉由此分隔線對資料進行分類。以分隔線將資料分隔成兩組不互相重疊之集合，並可找出集合中最鄰近分隔線且各自平行於分隔線的兩條平行線。SVM 算法如下：假設 $\{(x_i, y_i)\}_{I=1}^{N}, x_i \in R^d, y_i \in \{-1,1\}$資料集合為可輸入向量之訓練組，N 為樣本數量，而 d 為每一觀測值之維度。y_i是已知的目標。此算法為了求超平面(hyperplane) $w \cdot x_i + b = 0$其中 w 為超平面向量，b為偏移量，區分兩超平面的最大寬度為$2 / \|w\|^2$，所有在範圍內的點皆稱為支援向量(Vapnik, 2000)。

$$\text{Min}\Phi(x) = \frac{1}{2}\|w\|^2 \tag{1}$$

$$\text{S.t.}\, y_i(w^T x_i + b) \geq 1, i = 1,2,\dots,N$$

由於(1)式較難解，需透過拉格朗乘數法(Lagrange method)將理想化問題轉換成對偶問題。拉格朗乘數法的數值為非負實係數，(1)式被轉換為以下形式：

$$\text{Max}\Phi(w, b, \xi, \alpha, \beta) = \sum_{i=1}^{N} \alpha_I - \frac{1}{2}\sum_{i=1, j=1}^{N} \alpha_i \alpha_j y_i y_j {x_i}^T x_j \tag{2}$$

$$\text{S.t.} \sum_{j=1}^{N} \alpha_j y_j = 0, 0 \leq \alpha_i \leq C, i = 1,2,\dots,N$$

在(2)式中C為懲罰因子並決定懲罰的權重，被視為可調整參數，用於控制最大極限與分類誤差之間的交換。一般情況下，在所有可應用的數據無法找到線性分離的超平面，最佳的解決方法為將原始非線性數據轉換為更高線性分離的維度。最常見的核心函數為線性、多項式、半徑式函數(RBF)。雖然核心函數具多種選擇且可被利用的，但 RBF 仍較被廣泛使用。其定義為：$K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2), \gamma \geq 0$, (Vapnik, 2000)其$\gamma$表示 RBF 寬度。本研究將使用二元分類 SVM 方法(Hsu and Lin, 2003)。

二、C5.0 決策樹

C5.0 分類器是將一龐大數據分類與分析出隱藏資料的方法，亦可用決策樹呈現出有用的資料(Larose, 2005)。此算法採用決策樹，由循環式劃分與採用選擇的方式在訓練組主要部分中取得方法。C5.0 由 C4.5 改善了一些問題。如：變得更快、記憶效率更高、透過更小的決策樹區分較相似的結果、準確度更高、權重不同與分類錯誤的型態、降低干擾(Larose, 2005)。C4.5 中 Quinlan(1993) 利用具信息熵概念的 ID3 演算法(Iterative Dichotomiser 3)由一組已分類的訓練組建立決策樹，訓練組資料擴大由每個樣本所屬類別包括屬性向量，每個資料屬性可以用來做決策。C4.5 在決策樹的每個節點上使用資訊獲取量(Information Gain)來選擇測試屬性，選擇最高資訊獲取量的屬性作為節點的測試屬性。該屬性使得對產生之劃分中的樣本分類所需的資訊量最小，能反應劃分的最小隨機性與不純性(impurity) (Han and Micheline, 2001)。以計算A的屬性為例，計算資訊獲取率 $GainRatio(A)$，S表一資料樣本集，p_i為屬於B_i的任意樣本概率。假設有 n 個不同類B_i的值,其中(i = 1, ... ,n)，假設S_i為類別 B 的樣本數，$Info(S)$表示在現有樣本內的信息熵，計算過程如下：

$$Info(S) = \sum_{i=1}^{n} p_i \log(p_i) \tag{3}$$

假設 A 屬性有 n 個不同值$\{A_1, A_2, \dots, A_n\}$，使用 A 將 S 劃分為 n 個子集合$\{S_1, S_2, \dots, S_n\}$，S_j為A_j在 A 子集合中的樣本數，S_{ij}為S_j子集合中B_i類別的樣本數，$Info(S, A)$為要計算的信息熵。計算過程如下：

$$Info(S, A) = \sum_{j=1}^{n} \frac{S_{1j} + S_{2j} + \cdots + S_{nj}}{S} Info(A) \tag{4}$$

以分割的信息$SplitInfo(A)$是S裡每個屬性A的熵值，用來消除有大量屬性值誤差。計算過程如下：

$$SplitInfo(A) = -\sum_{i=1}^{n} \frac{|S_j|}{|S|} log\left(\frac{|S_j|}{|S|}\right) \tag{5}$$

$$Gain(A) = Info(S) - Info(S,A) \tag{6}$$

$$GainRatio(A) = Gain(A)/SplitInfo(A) \tag{7}$$

三、Extreme learning machine (ELM)

快速學習器(ELM)是由 Huang 於 2004 年提出的單隱藏前饋式類神經網路(SLFNs)演算法(Huang et al., 2006)，可隨機輸入權重與分析輸出權重。本節將介紹單一隱藏層網路的矩陣數學描述，並說明快速學習器演算法。給定 N 個任意的輸入輸出樣本(x_i, t_i)，i =1,...,N，其中：$x_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in R^n$以及$t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in R^m$，標準的單一隱藏層網路$\tilde{N}$個隱藏節點以及激活函數(Activation function) $g(x)$可以近似 N 個樣本達到平均零誤差。數學模型為以下式子：

$$H\beta = T \text{，} \tag{8}$$

其中

$$H(w_1, \dots, w_{\tilde{N}}, b_1 \dots, b_{\tilde{N}}, x_1 \dots, x_N) = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}};$$

$$\beta_{\tilde{N} \times m} = \left(\beta_1^T, \dots, \beta_{\tilde{N}}^T\right)^t; T_{N \times m} = \left(T_1^T, \dots, T_N^T\right)^t$$

其中$w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$, i =1,2,...,$\tilde{N}$，為權重向量連接第 i 連接第 i 個隱藏節點和輸入節點 $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$為權重向量連接第 i 個隱藏節點和輸出節點，b_i為第 i 個隱藏節點的開端，$w_i \cdot x_j$ 表示w_i和x_j的內積。H 被稱作網路隱藏層輸出矩陣(Hidden layer output matrix of neural network)；H 的 i 行是 i 個隱藏節點的輸出向量跟輸入樣本$x_1, x_2, \dots, x_N$之間的關係，而 H 的 j 列是隱藏層輸出向量跟輸入樣本x_j之間的關係。因此，測定輸出權重(連結隱藏層到輸出層)與找到最小平方解法得到線性系統一樣簡易。透過最低標準 LS 解法得到線性系統需利用以下式子：

$$\hat{\beta} = H^{\Psi}T \tag{9}$$

H^{Ψ}是根據 Rao(1971)和 Serre(2002)的 Moore-Penrose 廣義逆矩陣H，而具有最低的標準的 LS 解法是獨一無二的。快速學習器算法步驟如下：

給一訓練樣本集合$\aleph = \{(\mathrm{x_i}, \mathrm{t_i}) | \mathrm{x_i} \in \mathrm{R^n}, \mathrm{t_i} \in \mathrm{R^m}, \mathrm{i} = 1, \dots, N\}$、激活函數 g(x)，以及隱藏節點數$\tilde{N}$。

步驟 1. 隨機給一輸入權重w_i以及閥值b_i，, $\mathrm{i} = 1, \dots, \tilde{N}$

步驟 2. 計算隱藏層輸出矩陣H

步驟 3. 計算輸出權重$\hat{\beta}$。$\hat{\beta} = H^{\Psi}T$其中T=$[\mathrm{t_{1,\dots,}t_n}]^{\mathrm{T}}$。

五、 Random Forests (RF)

隨機森林演算法是將多數類樣本劃分為數個獨立的子集合；再將每一個獨立子集合進行交叉組合以構成不同的訓練樣本集，並針對不同的訓練樣本集利用決策樹分類器加以學習；最後根據平均加權法產成隨機森林，進而獲得決策規則(吳華芹，2013)。計算方法為給定 K 個分類器以及隨機向量 x、y，定義邊際函數如下：(張華偉等人，2006)

$$-max_{j \neq y} \mathrm{a}v_k I(h_k(\mathrm{mg(x,y)} = \mathrm{a}v_k I(h_k(x) = y)x) = j) \tag{15}$$

其中，I()是可能性函數，邊際函數顯示向量 X 所得到正確分類 y 的平均得票數超過其它任何類平均得票數的程度。由此可知邊際越大分類的可信度就越高。分類器誤差定義：

$$PE^* = P_{x,y}(\mathrm{mg(x,y)} < 0)$$

將上面的結論推廣到隨機森林函數： $h_k(X) = h(X, \theta_k)$

邊際函數如下：

$$\mathrm{mr(x,y)} - P_\theta(h(x,\theta)=y) - max_{j\neq Y}P_\theta(h(x,\theta)=j) \quad (16)$$

隨著樹的數目增加，PE^*就會趨向於

$$\mathrm{P}_{x,y}(P_\theta(h(x,\theta)=y) - max_{j\neq Y}P_\theta(h(x,\theta)=j) < 0) \quad (17)$$

而分類器{h(X,θ)}的強度可以表示為

$$\mathrm{s} = \mathrm{E}_{X,Y}mr(x,y) \quad (18)$$

假設 $s \geq 0$，根據契比雪夫不等式，(16) (17) 兩式可以得到：

$$PE^* \leq (\mathrm{var(mr)}) \,/\, \mathrm{s}^2 \quad (19)$$

根據 Breiman(2001)可推導出

$$\begin{aligned}\mathrm{var(mr)} &= \bar{\rho}\big(E_\theta sd(\theta)\big)^2 \\ &\leq \bar{\rho}E_\theta var(\theta) \\ &\geq 1 - s^2\end{aligned} \quad (20)$$

隨機森林的目標誤差上界是$PE^* \leq \bar{\rho}(1-S^2)/S^2$

研究步驟:

1. 取得子宮內膜癌資料庫數據為研究對象。為確保資料的完整性、一致性，將進行資料編碼，不同的數值型態與臨床醫師討論進行轉換，並做分類。
2.將資料分為 20 項預測變數與 1 項目標變數。
3.利用增益比例及資訊量增益比例方法選擇重要變數。
4.利用 RF,SVM,ELM,C5.0 再次分析，其他未被選擇的重要變數則被視為影響分析的雜訊而從資料中刪除。
5.比較各種方法的預測準確率，以及比較集成學習方法與一般機器學習。
6.最後，與臨床醫師討論並證實重要變數的可信度，並且可以依照病患的屬性決定只用何種方法預測最為準確。

(五) 實證研究

在研究中，我們由全民健康保險資料庫之癌症登記資料欄位，使用C5.、RF、SVM、ELM驗證其敏感度與特異度，並預測子宮內膜癌復發之重要因子。數據集中共包含20個預測變數，分別為(1)年齡 (age) (2) 組織型態 (histopathology type)(3) 性態碼 (behavior code) (4)手術邊緣 (surgical margin) (5) 病理T (pathologic T) (6)淋巴轉移 (lymphatic metastasis) (7)腫瘤大小 (tumor grade) (8)病理N (pathologic N) (9) 臨床M (clinical M) (10) 病理分期 (pathologic stages) (11)臨床分期 (clinical stage) (12) 分化 (differentiation) (13) 臨床T (clinical T) (14) 臨床N (clinical N) (15) 病理M (pathologic M) (16)放射治療手術順序 (sequence of radiotherapy and surgery) (18) 區域淋巴結檢查數目 (regional lymph nodes examined) (17) 區域全身順序 (sequence of locoregional therapy and systemic therapy) (19) 其他放射劑量 (dose to target of other RT) (20) 最高放射劑量臨床標靶體積 (target of CTV_H)，以及1個目標變數為復發型態(type of recurrence)，共268筆資料，隨機選取80筆資料為測試樣本，其餘188筆資料為訓練樣本，進行重複取樣十次。

以{1}代表復發；{2}則代表未復發。因此{1-1}代表敏感度：原始的判定為復發，而經由模式判定後亦為復發；而{2-2}則表示特異度：原始判定為沒有復發，經由模式判定亦為沒有復發。由表二可知變數未經篩選前C5.0的整體正確判別率為91.2%，而個別的判別正確率{1-1}的比率為82.4%：即原始群體為第1類的樣

本正確的被判別到第1類的比率為82.4%。其中有3個原本群體為第1類的樣本，被錯分為第2類的群體中；而有4個原本群體為第2類的樣本，被錯分為第1類的群體中。

表二 C5.0之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	14 (82.4%)	3 (17.6%)
2 (未復發)	4 (6.3%)	59 (93.7%)
整體平均預測準確率	91.2 %	

由表三可知變數未經篩選前RF的整體正確判別率為92.5%，而個別的判別正確率{1-1}的比率為81.8%：即原始群體為第1類的樣本正確的被判別到第1類的比率為81.8%；而{2-2}的判別正確率為96.6%。其中有4個原本群體為第1類的樣本，被錯分為第2類的群體中；而有2個原本群體為第2類的樣本，被錯分為第1類的群體中。

表三 RF之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	18 (81.8%)	9 (18.2%)
2 (未復發)	2 (3.4%)	61 (96.6%)
整體平均預測準確率	92.5 %	

由表四可知變數未經篩選前SVM的整體正確判別率為77.5 %，而個別的判別正確率{1-1}的比率為0%：即原始群體為第1類的樣本正確的被判別到第1類的比率為0%；而{2-2}的判別正確率為100%。其中有原本群體為第1類的樣本，全部被錯分為第2類的群體中。

表四 SVM之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	0 (0%)	18 (100%)
2 (未復發)	0 (0%)	62 (100%)
整體平均預測準確率	77.5 %	

由表五可知變數未經篩選前ELM的整體正確判別率為88.8%，而個別的判別正確率以{1-1}的比率最高，為88.8%：即原始群體為第1類的樣本正確的被判別到第1類的比率為88.2%；而{2-2}的判別正確率為88.9%。其中有2個原本群體為第1類的樣本，被錯分為第2類的群體中；而有7個原本群體為第2類的樣本，被錯分為第1類的群體中。

表五 ELM之預測結果

實際類別	預測類別	
	1 (有復發)	2 (未復發)
1 (有復發)	15 (88.2%)	2 (11.8%)
2 (未復發)	7 (11.1%)	56 (88.9%)
整體平均預測準確率	88.8%	

為了評估各方法的穩定度，我們使用 10 組獨立資料集進行 C5.0、RF、SVM、ELM 四種模式的測試。根據表六的結果，我們可以觀察到 C5.0 模式在{1-1}產生最高平均分類準確率，為 77.65%；而在{2-2}最高平均分類準確率為 RF，為 94.05%。在整體情況下，我們可以看到 C5.0 模式優於 RF、ELM 和 SVM 模式，這表明 C5.0 模式針對資料集整體結果確實比其他四種方法提供更好的分類準確度。

表六 C5.0、RF、SVM、ELM模式預測評估

	預測有復發，實際有復發（敏感度%）{1-1}				預測無復發，實際無復發（特異度%）{2-2}				整體平均預測準確率%			
模組	C5.0	RF	ELM	SVM	C5.0	RF	ELM	SVM	C5.0	RF	ELM	SVM
1	66.67	66.67	68.42	00.00	100.0	100.0	91.80	100.0	92.50	92.50	86.25	77.50
2	82.35	66.67	65.00	100.0	93.65	96.23	93.33	00.00	91.25	86.25	86.25	21.25
3	69.23	46.15	44.44	00.00	95.52	95.52	98.11	100.0	91.25	87.50	80.00	83.75
4	81.48	51.85	79.17	00.00	86.79	90.57	85.71	100.0	85.00	77.50	83.75	66.25
5	72.73	68.18	71.43	00.00	96.55	94.83	96.15	100.0	90.00	87.50	87.50	72.50
6	80.95	57.14	80.00	100.0	94.92	94.92	86.15	00.00	91.25	85.00	85.00	26.25
7	79.17	37.50	78.95	100.0	75.00	85.71	85.25	00.00	76.25	71.25	83.75	30.00
8	77.27	72.73	88.24	100.0	86.21	86.21	88.89	00.00	83.75	82.50	88.75	27.50
9	100.0	81.82	71.43	00.00	74.14	96.55	88.14	100.0	81.25	92.50	83.75	72.50
10	66.67	61.11	81.25	100.0	100.0	100.0	92.19	00.00	92.50	91.25	90.00	22.50
平均	77.65	60.98	72.83	50.00	90.28	94.05	90.57	50.00	87.50	85.38	85.50	50.00

本研究第二階段使用集成學習投票策略，相關貢獻率之排序如表七所示，在增益比例(Gain Ratio)，依被選取次數將重要變數排名依序為手術邊緣、病理分期、pT、淋巴轉移、放射手術順序、cM、pN、腫瘤大小、pM、區域全身順序、cN、臨床分期、淋巴檢查、cT、組織型態、分化、其他放射劑量、最高劑量、性態碼；

表七 各機器學習法之貢獻率比較

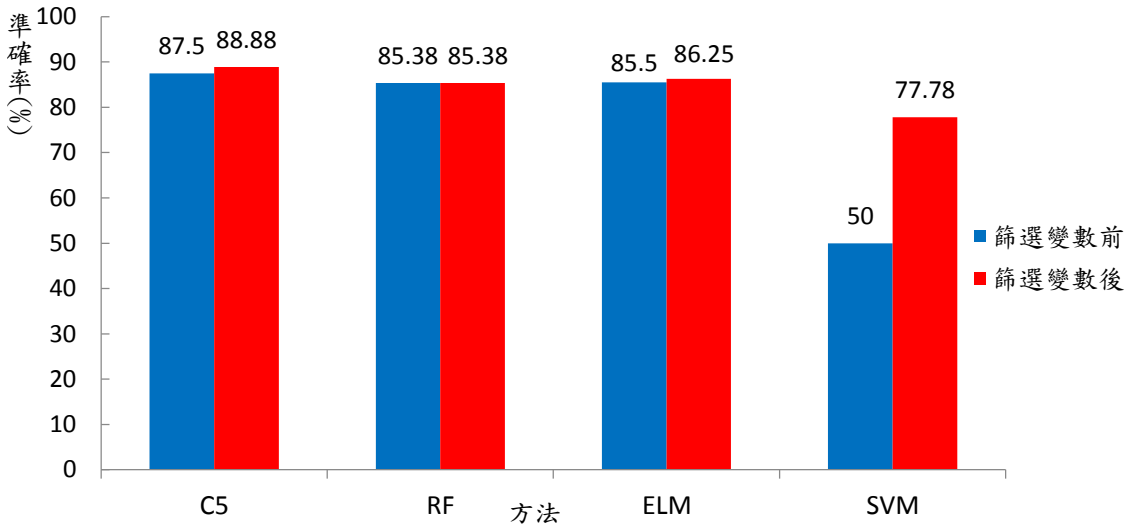
增益比例 (Gain Ratio)		資訊量增益比例 (Information Gain Ratio)	
變數	貢獻率	變數	貢獻率
手術邊緣	0.1757	病理分期	0.3722
病理分期	0.1314	pT	0.3490
pT	0.1282	手術邊緣	0.2660
淋巴轉移	0.1076	淋巴轉移	0.1881
放射手術順序	0.0984	臨床分期	0.1784
cM	0.0943	cT	0.1605
pN	0.0903	pN	0.1539
腫瘤大小	0.0902	區域全身順序	0.1291
pM	0.0877	cM	0.1158
區域全身順序	0.0847	放射手術順序	0.1152
cN	0.0711	cN	0.1113
臨床分期	0.0709	組織型態	0.1077
淋巴檢查	0.0699	淋巴檢查	0.0955
cT	0.0678	pM	0.0900
組織型態	0.0495	腫瘤大小	0.0840
分化	0.0313	分化	0.0674
其他放射劑量	0.0133	其他放射劑量	0.0100
最高劑量	0.0109	最高劑量	0.0080
性態碼	0.0000	性態碼	0.0000

而在資訊量增益比例(Information Gain Ratio)依貢獻率排名為病理分期、pT、手術邊緣、淋巴轉移、臨床分期、cT、pN、區域全身順序、cM、放射手術順序、cN、組織型態、淋巴檢查、pM、腫瘤大小、分化、其他放射劑量、最高劑量、性態碼。表七為經過投票機制選擇重要變數，根據結果依名次排序為病理分期、手術邊緣、pT、淋巴轉移、pN、cM、放射手術順序、臨床分期、區域全身順序、cT、cN、pM、淋巴檢查、腫瘤大小、組織型態、分化、其他放射劑量、最高劑量、性態碼。所有機器學習法經集成學習投票後病理分期為貢獻率最高之變數，因此可以得知病理分期是最重要的子宮內膜癌復發因子。另外，組織型態、分化、其他放射劑量、最高劑量、性態碼為貢獻率最低的五個變數，因此我們將這五項變數從資料中刪除，且再次以四種機器學習法進行分析。

表八 重要變數之篩選結果

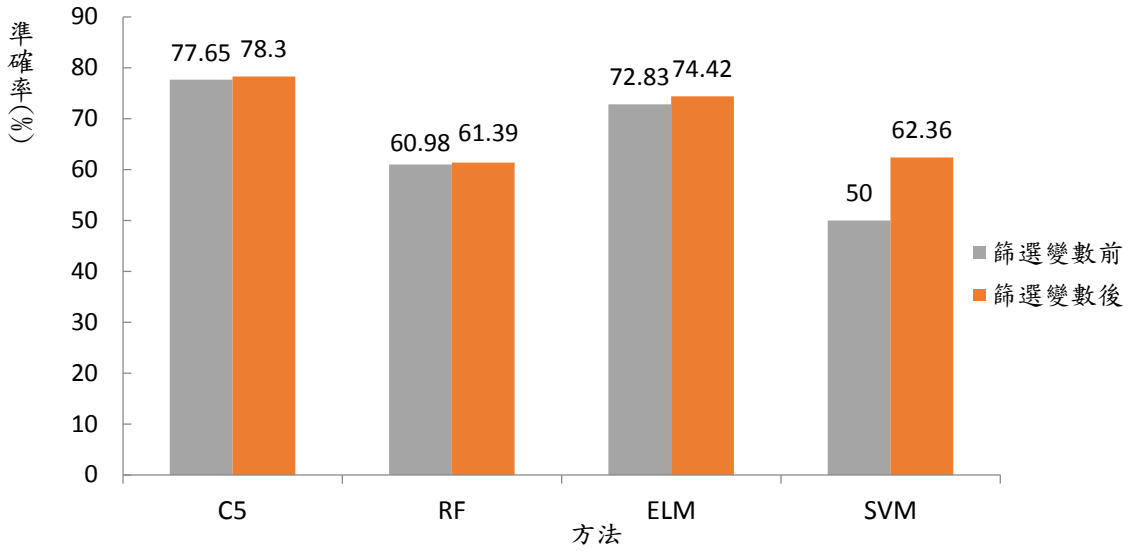
變數	貢獻率排名	變數	貢獻率排名
病理分期	1	cN	11
手術邊緣	2	pM	12
pT	3	淋巴檢查	13
淋巴轉移	4	腫瘤大小	14
pN	5	組織型態	15
cM	6	分化	16
放射手術順序	7	其他放射劑量	17
臨床分期	7	最高劑量	18
區域全身順序	9	性態碼	19
cT	10		

經過變數篩選後分類準確率結果如圖三所示，C5.0、RF、ELM與SVM分類準確率分別為88.88%、85.38%、86.25%與77.78%，經過變數篩選後。可以發現各方法經過變數篩選後，其分類準確率明顯提高，C5.0之分類準確率由87.5%提高至88.88%，ELM分類準確率由85.5%提高至86.25%，SVM之分類準確率由50%提高至77.78%，唯RF經過集成學習投票機制篩選變數後分類準確率不變，皆為85.38%。



圖三 各機器學習篩選變數前後之比較

敏感度分析結果見圖四，經過集成學習篩選變數後，C5.0、RF、ELM與SVM敏感度依序為78.3%、61.39%、74.42%、62.36%。四種方法經過集成學習變數篩選後敏感度皆有提高。



圖四 集成學習策略前後之敏感度結果比較

(六) 結論

子宮內膜癌在婦科癌症中致死率相當高，晚期的子宮內膜癌復發率很高，一旦復發，能存活的機率很低，過去很多研究將變因的觀察以全民健保資料庫抽樣檔的門診處方及治療明細檔作為資料分析，至今仍缺乏以資料探勘方法分析之相關研究。因此本研究使用了資料探勘分析，且更進一步加入投票機制刪除變數來改善一般機器學習法的缺點。經由結果比較後可以驗證變數經過篩選後用於子宮內膜癌復發預測之成效。本研究結果顯示：1.使用C5.0、RF、ELM和SVM四種方法預測子宮內膜癌復發的準確度，我們發現C5.0為準確度最高之機器學習方法; 2.經過機器學習預測子宮內膜癌復發之風險因子，並針對風險因子，篩選變數來改善一般機器學習之缺點。此方法是有效的，大多數之機器學習法再經過篩選變數改善，其分類準確率皆提高。本研究臨床實務建議：針對病患個案需鑑別遭遇復發之預測可以使用C5.0方法進行敏感度分析；相對地，對於針對病患個案需識別無復發之預測可以使用RF方法進行特異度分析。另外對於病理分期、手術邊緣、pT、淋巴轉移、pN、cM、放射手術順序、臨床分期、區域全身順序、cT、cN、pM、淋巴檢查、腫瘤大小是否扮演復發重要的預後因子，建議未來可以深入分析。最後，重要的考量是個案資料不完整可能造成的數據缺失問題，但是若能提高樣本數，相信也能具體反應子宮內膜癌復發之重要變數。

(七)文獻參考

Bauer E. and Kohavi R. (1999) An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learning*, Vol. 36, pp. 105-139.

Breiman L. (2001) Random forests, *Machine Learning*, Vol. 45, no. 1, pp. 5-32.

Breiman L (1996) Bagging predictors, *Machine Learning*, Vol. 24, no. 2, pp. 123-140.

Craven P. and Wahba G. (1979) Smoothing Noisy Data with Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation, *Numberische Mathematik*,, Vol. 31, pp. 317-403.

David A. and Lerner L. (2004) *Pattern classification using a support vector machine for genetic disease diagnosis, Electrical and Electronics Engineers in Israel*, 23rd IEEE Convention of Prceedings, pp. 289-292.

Dietterich T. G. (2000) *Ensemble methods in machine learning*, Proceedings of the First International Workshop on Multiple Classifier Systems (MCS00), pp. 1-15.

Ferlaya J., Steliarova-Foucher E., Lortet-Tieulent J. (2013) Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012, *Eur J Cancer*, Vol. 49, pp.1374-1403.

Freund Y. and Schapire R. E. (1997) A decision-theoretic generaliation of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, Vol. 55, no. 1, pp. 119-139.

Han J. and Micheline K. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann, New York.

Ho S. H., Jee S. H. and Lee J. E., Park J.S. (2004) Analysis on risk factors for cervical cancer using induction technique, *Expert Systems with Applications*, Vol. 27, no. 1, pp. 97-105.

Hsu C. W., Chang C. C. and Lin C. J. (2003) *A practical guide to support vector classification*. Taipei, Taiwan: Department of Computer Science and Information Engineering, National Taiwan University.

Huang G. B., Zhu Q. Y. and Siew C. K. (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. School of Electrical and Electronic Engineering, Nanyang Technological University, *Nanyang Avenue*, Vol. 2, pp. 985-990.

Huang G. R., Zhu Q. Y., Siew C. X. (2006) Extreme learning machine: theory and applications, *Neurocomputing*, Vol. 70, pp. 489-501.

Kim H. S., Park N. H. and Kang S. B. (2008) Rare Metastases of Recurrent Cervical Cancer to the Pericardium and Abdominal Muscle, *Archives of Gynecology and Obstetrics*, Vol. 278, pp. 479-482.

Larose D. T. (2005) *Discovering Knowledge in Data: An Introduction to Data Mining*, New Jersey: John Wiley & Sons, Inc.

Liaw A. and Wiener M. (2002) Classification and Regression by RandomForest, *R news*, Vol. 2, no. 3, pp. 18-22.

Li S. T., James T. K., Zhu H. and Wang Y. (2003) Texture classification using the support vector machines, *Pattern Recognition*, Vol. 36, pp. 2883-2893.

Li F. G., Wang Z. P., Hu G. and Li H. (2011) Current status of SNPs interaction in genome-wide association study, *Hereditas (Beijing)*, Vol. 33, no. 9, pp. 905.

Mao Y., Zhou X., Pi D., Sun Y. and Wong T. C. (2005) Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection, *Journal of Biomedicine and Biotechnology*, Vol. 2005, no. 2, pp. 160-171.

Quinlan J. R. (1993) *C4.5: programs for machine learning*, San Mateo, CA: Morgan Kaufinann.

Rao C. R. and Mitra S. K. (1971) *Generalized inverse of matrices and its applications*, New York: Wiley.

Schapire R. E. (1990). The Strength of weak learnability, *Machine Learning*, Vol. 5, no. 2, pp. 197-227.

See5: An Informal Tutorial (2007) Retrieved from http://www.rulequest.com/see5-win.html.

Serre D. (2002). *Matrices: Theory and applications*, New York: Springer.

Steinberg D., Bernard B., Phillip C. and Kerry M. (1999) *MARS User Guide*, San Diego, CA: Salford Systems.

Sun Z. L. and Choi T. M. (2008) Sales forecasting using extreme learning machine with applications in fashion retailing, *Decision Support Systems*, Vol. 46, pp. 411-419.

Svetnik V., Liaw A., Tong C., Culberson J. C., Sheridan R. P. and Feuston B. P. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling, *Journal of chemical information and computer sciences*, Vol. 43, no. 6, pp. 1947-1958.

Thangavel K., Jaganathan P. P. and Easmi P. O. (2006) Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique, *Asian Journal of Information Technology*, Vol. 5, no. 4, pp. 413-417.

Tiago H. F., Hagit S. and Chan W. P. (2006) *Breast Cancer Prognosis via Gauss-ian Mixture Regression*. Queen's University Kingston, ON, Canada.

Vani G., Savitha R. and Sundararajan N. (2010) *Classification of Abnormalities in Digitized Mammograms using Extreme Learning Machine*, Automation, Robotics and Vision Singapore.

Vapnik V. N. (2000) *The Nature of Statistical Learning Theory*, Springer, Berlin.

衛生署福利部國民健康署(2015)，1979-2011 台灣子宮頸癌、子宮內膜癌、卵巢癌發生率，取自 http://www.hpa.gov.tw/。

美國癌症協會(2013)，取自 http://www.cancer.org/cancer/endometrialcancer/overviewguide/endometrial--uterine--cancer-overview-survival-rates。

莊永裕(2006)，整體學習(Ensemble Learning)入門，取自 http://www.csie.ntu.edu.tw/~cyy/。

張惟智(2009)，運用資料探勘分類模型對腹主動脈瘤術後併發症之探討與研究，國立台北護理學院資管系研究所碩士論文。

歐宗殷(2010)，資料探勘為基礎之零售業銷售預測模式以連鎖超商鮮食商品為例，國立清華大學工業工程與工程管理研究所博士論文。

許智宇(2010)，整合KMV模型、約略集合及隨機森林應用於企業信用評等之研究，國立台北科技大學商業自動化與管理研究所碩士論文。

洪智力和陳勁宏(2007)，破產預測選擇性集成模型比較，中原大學資訊管理研究所。
吳華芹(2013)，基於訓練集劃分的隨機森林算法。科技通報，2013年，第10期，第124-126頁。
張華偉，王明文和甘麗新(2006)，基於隨機森林的文本分類模型研究，山東大學學報(理學版)， 第 41 卷，第 3 期，第 5-9 頁。