科技部補助

大專學生研究計畫研究成果報告

計畫名稱：	預測大腸直腸癌復發的風險因子:原發部位與期別的分層重要性？

執行計畫學生： 吳明蓁

學生計畫編號： MOST 105-2815-C-040-059-E

研 究 期 間 ： 105 年 07 月 01 日至 106 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 106 年 03 月 14 日

(一) 摘要

大腸直腸癌(Colorectal Cancer, CRC)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，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，以提供臨床醫師對可能的病情發展有更多資訊可參考。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。本研究所需的病歷記錄和病理資料的來源為中山醫學大學附設醫院癌症防治中心的癌症登記資料庫。初步經由三位資深臨床醫師討論復發的危險因子有：(1)性別(2)原發位置(3)組織型態(4)性態碼(5)分級(6)區域淋巴腺檢查(7)區域淋巴結侵犯數目(8)外院診斷性及分期性手術處置(9)申報醫院診斷性及分期手術(10)臨床 T (11)臨床 N (12)臨床 M (13)臨床期別組合(14)病理 T (15)病理 N (16)病理 M (17)病理期別(18)原發部位手術邊緣(19)放射治療與手術順序(20)區域治療與全身性治療順序(21)放射劑量(22)外院化學治療(23)申報醫院化學治療(24)生存狀態(25)死亡原因，經過資料清理後共計有效個案 606 筆。首先，藉由使用支援向量機(SVM)、快速學習器(ELM)以及決策樹(C5.0)、隨機森林(RF)、多元適應性雲形回歸(MARS)五種資料探勘法深入分析預測準確率後；我們考量納入變數篩選機制，亦即採用 C5.0 篩選出重要變數後，再進行下一階段支援向量機、快速學習器、C5.0 決策樹、MARS 資料探勘法分析。並以敏感度和特異度繪製出 ROC 曲線之 AUC 曲下面積來驗證方法的鑑別度。研究結果顯示復發危險因子依重要性排序為：原發部位、病理期別、手術邊緣與淋巴結侵犯數，其中原發部位與病理期別是重要的獨立危險因子；在變數篩選前的 10 次執行結果顯示各方法的平均準確率與 AUC 分別為：C5.0(86.825%/0.9405)、MARS(83.701%/0.9165)、RF(77.736%/0.8871)、ELM(86.665/0.9281)、SVM(75.939%/0.8395)。再以原發部位與病理期別進行分層後分析結果，在大腸原發部位：期別<IIb 的準確率以 SVM(91.167%)為最佳；期別≥IIb 的準確率以 ELM(86.119%)最高；在直腸原發部位：期別<IIb 的準確率以 ELM(95.714%)最佳；期別≥IIb 的準確率以 MARS(89.286%)與 ELM(89.286%)最高。本研究結果證實針對大腸直腸癌症病患的復發性預測，可以原發部位與病理期別的分層樣式提供臨床醫師輔助治療。

關鍵字：大腸直腸癌復發、分類預測、集成學習

(二)研究動機與研究問題

大腸直腸癌是西方已開發國家前三大癌症死因(Ong et al., 1997)，在台灣也是國人第一常見的癌症死因(衛生福利部國民健康署, 2016)。治療大腸直腸癌的主要方法是外科手術，然而有超過三分之二原發性疾病的病人接受可行的治癒方式，並將所有的腫瘤切除，仍然高達 50%的病患五年內終將死亡(Walker et al., 2014)，其中多數來自於局部、區域性或遠端的腫瘤復發，但由於腸道裡不同位置的腫瘤細胞可能有不同的復發行為模式，所以很難進行預測分析。大腸直腸癌病患在第一期約有 5-10%；第二期約有 20%以及第三期約有 35%會因產生復發導致死亡(Leung and Liu, 2014)，故早期發現和治療非常重要。面對大腸直腸癌症的復發型態，迄今在文獻中還是無法有一致的結論。因此本研究經過文獻與臨床資深醫師討論後，整理復發因素包括 25 項：(1)性別 (Sex) (2)原發位置 (Primary Site) (3)組織型態 (Cell Type) (4)性態碼 (Tumor Grade) (5)分級 (Tumor Size) (6)區域淋巴腺檢查 (Regional Lymph Nodes Examined) (7)區域淋巴結侵犯數目 (Lymph Node Metastases (LNM)) (8)外院診斷性及分期性手術處置 (Surgical Diagnostic and Staging Procedure at Other Facility) (9)申報醫院診斷性及分期手術 (Surgical Diagnostic and Staging Procedure at This Facility) (10)臨床 T (Clinical T) (11)臨床 N (Clinical N) (12)臨床 M (Clinical M) (13)臨床期別組合 (Clinical Stage Group) (14)病理 T (Pathologic T) (15)病理 N (Pathologic N) (16)病理 M (Pathologic M) (17)病理期別 (Pathologic Stage Group) (18)原發部位手術邊緣 (Surgical Margins of The Primary Site) (19)放射治療與手術順序 (Sequence of RT and Surgery) (20)區域治療與全身性治療順序 (Sequence of Locoregional Therapy and Systemic Therapy) (21)放射劑量 (Target of CTV) (22)外院化學治療 (Chemotherapy at Other Facility) (23)申報醫院化學治療 (Chemotherapy at This Facility) (24)生存狀態 (Vital Status) (25)死亡原因 (Cause of Death)，進一步研究上述潛在危險因子中，分析何者是影響復發的重要變數。

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）就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大腸直腸癌復發重要因子之間的關聯。預期本計畫將有以下成果：

- 透過五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。
- 準確地預測大腸直腸癌患者的復發因素，提供大腸直腸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。

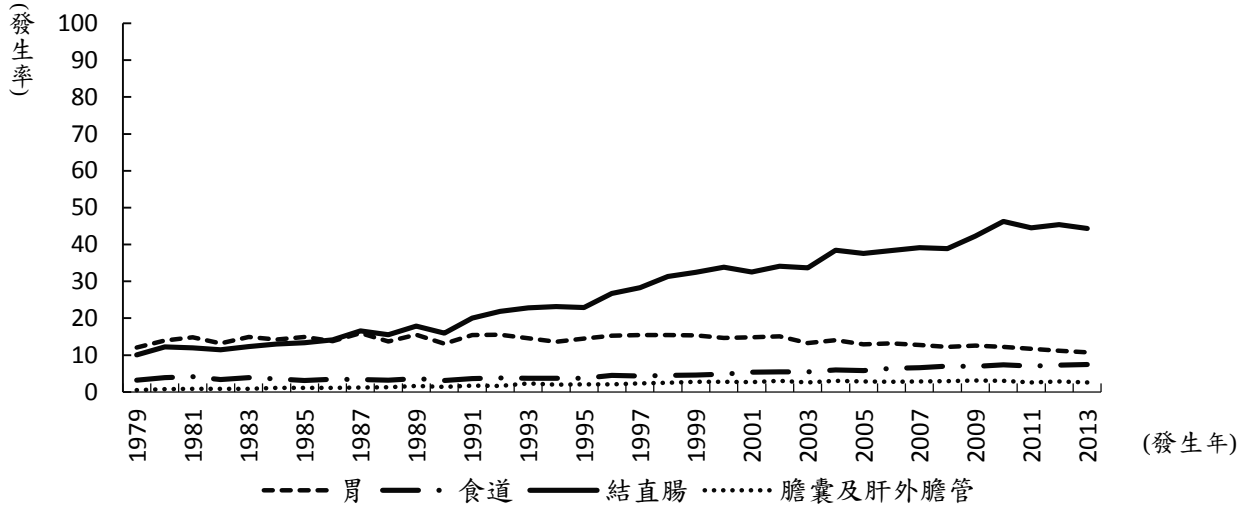
(三)文獻回顧與探討

本研究文獻探討內容分為兩個部分討論: 大腸直腸癌盛行率、資料探勘方法。

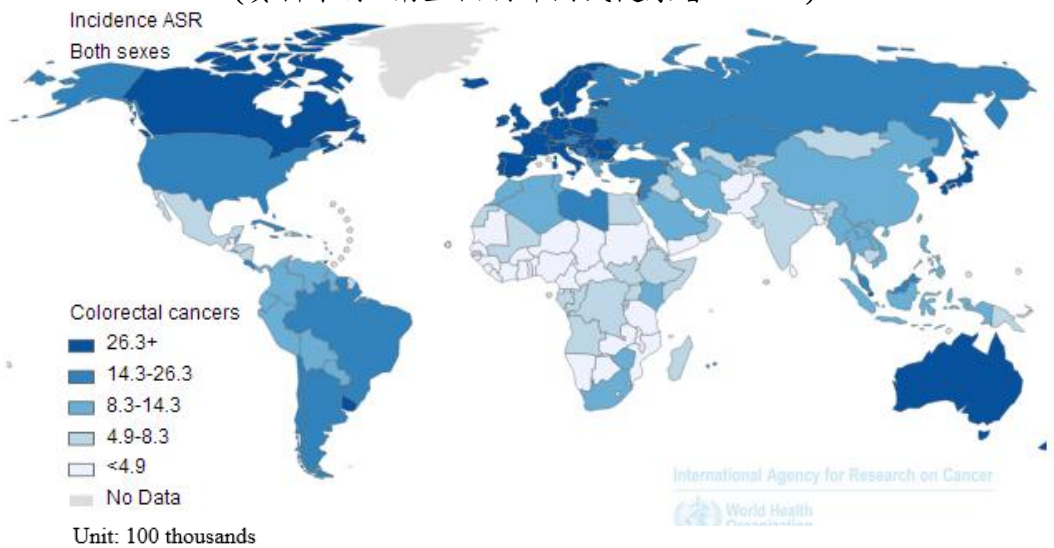
一、大腸直腸癌盛行率

根據 2012 年 International Agency for Research on Cancer (IARC)的資料統計顯示，大腸直腸癌是全球男性第三常見的癌症(746,000 個案例，佔罹患癌症人口的 10%)和全球女性第二常見的癌症(614,000 個案例，佔罹患癌症人口的 9.2%)，將近 55%的案例發生在開發國家。

在台灣，根據衛生福利部國民健康署統計，2013 年台灣大腸直腸癌發生率約為每 10 萬人口 15,130 例，平均年齡約在 66.02 歲，而死亡率約為每 10 萬人口 5,603 例，平均年齡約在 70.95 歲，在十大癌症發生率中占第一位，且發生率逐年增加。



圖一、胃癌、食道癌、結直腸癌及膽囊及肝外膽管癌在台灣的發生率(1979-2013)
(資料來源: 衛生福利部國民健康署，2016)



圖二、大腸直腸癌-全球不分性別發生區域
(資料來源：International Agency for Research on Cancer (IARC), 2015)

表 1 台灣不分性別 每 10 萬人口標準化發生率 (2000 年世界標準人口)，1979-2013 年

年度	胃					食道					結直腸癌					膽囊及肝外膽管				
	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%	個案數	平均年齡	年齡中位數	標準化率	癌症百分比%
1979	1,471	57.31	59	12.09	11.33	377	59.23	59	3.16	2.9	1,255	54.11	57	10.09	9.67	64	56.95	57.5	0.51	0.49
1980	1,734	58.02	60	13.96	11.15	478	59.6	60	3.87	3.07	1,552	55.14	57	12.20	9.98	94	56.32	56.5	0.75	0.6
1981	1,859	59.04	60	14.84	11.63	503	61.41	61	4.11	3.15	1,550	55.64	57	11.94	9.7	104	60	61.5	0.86	0.65
1982	1,706	59.79	61	13.17	11.15	434	61.94	62	3.38	2.84	1,528	56.41	58	11.48	9.99	107	59.88	63	0.84	0.7
1983	1,985	59.54	61	14.92	11.15	502	62.42	62	3.88	2.82	1,652	57.59	59	12.31	9.28	115	57.6	58	0.87	0.65
1984	1,940	60.61	62	14.18	10.66	480	63.23	63	3.55	2.64	1,819	57.92	60	12.97	9.99	148	61.15	62	1.13	0.81
1985	2,095	61.14	62	14.92	10.8	451	62.52	63	3.14	2.32	1,929	58.57	60	13.35	9.94	155	59.61	61	1.07	0.8
1986	2,026	60.96	63	13.76	10.37	502	62.33	63	3.43	2.57	2,099	58.94	60	14.12	10.74	165	62.55	63	1.14	0.84
1987	2,458	61.09	63	15.96	10.5	516	63.25	64	3.41	2.2	2,578	59.34	61	16.58	11.01	193	60.22	61	1.23	0.82
1988	2,156	61.18	63	13.69	9.5	502	62.03	63	3.21	2.21	2,461	59.69	62	15.51	10.85	204	60.76	63	1.27	0.9
1989	2,495	62.05	65	15.54	9.25	580	63.8	64	3.64	2.15	2,935	59.91	63	17.88	10.88	261	62.23	63	1.61	0.97
1990	2,194	62.36	65	13.09	8.96	522	62.12	63	3.12	2.13	2,687	60.64	63	16.00	10.97	219	62.23	64	1.33	0.89
1991	2,654	62.95	65	15.42	8.79	622	62.72	64	3.62	2.06	3,473	61.05	63	20.06	11.5	297	64.31	66	1.73	0.98
1992	2,785	62.65	65	15.52	8.39	681	63.35	65	3.84	2.05	3,923	61.53	63	21.85	11.81	294	63.87	66	1.67	0.89
1993	2,680	63.65	66	14.56	7.77	673	63.15	64	3.69	1.95	4,220	61.74	64	22.84	12.23	422	65.48	67	2.34	1.22
1994	2,581	63.73	66	13.59	7.23	701	63.9	65	3.74	1.96	4,411	62.5	65	23.17	12.36	372	64.27	66	1.99	1.04
1995	2,849	64.41	67	14.52	7.59	723	62.98	64	3.75	1.93	4,483	62.56	65	22.87	11.95	397	66.05	68	2.06	1.06
1996	3,077	64.78	67	15.26	7.14	881	63.56	64	4.49	2.04	5,346	63.4	66	26.69	12.4	415	65.77	67	2.09	0.96
1997	3,194	65.13	68	15.41	6.77	895	62.87	64	4.35	1.9	5,845	63.66	66	28.26	12.39	478	66.06	67.5	2.34	1.01
1998	3,291	65.55	68	15.43	6.34	962	62.63	64	4.52	1.85	6,679	63.74	66	31.34	12.86	533	65.79	67	2.53	1.03
1999	3,386	65.74	69	15.33	6.01	1,009	61.69	63	4.58	1.79	7,124	64.16	66	32.42	12.64	594	67.68	69	2.73	1.05
2000	3,351	66.18	69	14.64	5.69	1,082	61.63	62	4.82	1.84	7,668	64.53	67	33.88	13.02	618	68.25	70	2.76	1.05
2001	3,502	66.54	70	14.79	5.82	1,257	61.41	62	5.38	2.09	7,640	64.88	67	32.56	12.69	630	67.38	69	2.69	1.05
2002	3,694	66.76	70	15.12	5.85	1,310	60.83	61	5.47	2.08	8,251	65.12	68	34.07	13.07	720	67.06	68	3.00	1.14
2003	3,360	66.91	70	13.30	5.3	1,356	60.88	61	5.42	2.14	8,391	65.19	68	33.68	13.24	650	67.99	70	2.61	1.03
2004	3,686	66.61	70	14.08	5.19	1,537	60.14	59	6.00	2.16	9,873	65.18	67	38.45	13.9	775	67.46	70	3.00	1.09
2005	3,506	67.43	70	12.93	4.87	1,530	59.42	58	5.77	2.13	9,938	65.56	67	37.56	13.82	748	68.33	69	2.83	1.04
2006	3,683	67.75	70	13.16	4.85	1,766	59.34	57	6.44	2.33	10,524	65.23	67	38.37	13.86	767	68.43	70	2.75	1.01
2007	3,702	67.53	70	12.78	4.64	1,863	59.15	57	6.59	2.34	11,085	65.77	67	39.13	13.9	835	69.52	72	2.88	1.05
2008	3,657	67.75	70	12.21	4.44	2,035	58.98	57	6.98	2.47	11,397	65.98	68	38.90	13.84	872	69.59	71	2.94	1.06
2009	3,890	68.03	70	12.56	4.34	2,076	58.89	57	6.94	2.32	12,769	66.05	67	42.28	14.24	959	69.07	71	3.14	1.07
2010	3,922	68.21	70	12.25	4.22	2,285	58.38	56	7.39	2.46	14,350	65.59	66	46.30	15.43	948	69.48	71	2.98	1.02
2011	3,869	68.1	70	11.73	4.1	2,221	58.7	57	6.99	2.35	14,331	65.76	66	44.53	15.18	856	69.57	71	2.59	0.91
2012	3,824	68.4	70	11.20	3.91	2,382	58.88	57	7.30	2.44	15,072	65.74	66	45.41	15.43	946	68.99	70	2.81	0.97
2013	3,768	68.58	70	10.74	3.8	2,496	58.74	57	7.46	2.52	15,140	66.02	66	44.32	15.27	923	69.89	71	2.63	0.93

(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。本研究將應用以下五種不同的資料探勘方法來預測大癌直腸癌復發的重要因子：

(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) Multivariate Adaptive Regression Splines (MARS)：多元適應性雲形回歸(MARS)是由 Friedman 等人提出來處理多元複雜資料問題的新方法。MARS 目前較被廣泛運用的領域，大部份是資料探勘的分類問題，以及預測。Falk 等人(2006)比較了 CART、MARS 和 GMR 方法在預測經歷癌症切除患者乳腺癌復發時間。結果顯示，GMR 算法證明相較於 CART 和 MARS 有較好的效率。

(5) 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

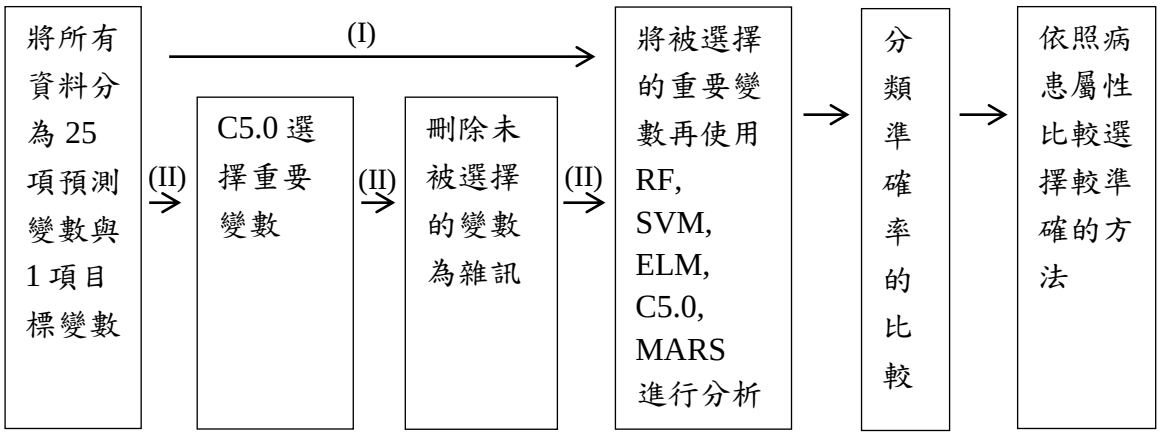
(additive model)。所謂的附加模型通常是指一個新增的資料點，最後所指定的類別標籤，是由部分或所有的附屬模型(component model)經由賦予不等的權重後，再加總所得到的結果。Freund 和 Schapire(1996；1997)提出的 Adaboost 演算法，可以說是建造附加模型極有效的方法。透過學習演算法，極盡可能地將分類錯誤減少到最小的方式去產生一個假說，每次增加一個假說到整體學習之中，分類錯誤就相對的降低。在多數的研究實驗中(Freund and Schapire, 1996；Bauer and Kohavi, 1999；Dietterich, 2000) 都說明了 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)。因應大腸直腸癌復發的臨床反應，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)研究流程與研究方法

本研究以 SVM、C5.0 決策樹、ELM、MARS、RF 五種模型等相關研究基礎，建立大腸直腸癌復發的重要因子，並探討五種資料探勘方法預測之準確度。

研究流程:

為了比較重要變數篩選的差異，研究設計架構如圖三所示：在圖三中，首先依據文獻查證與臨床醫師討論後決定 25 項預測變數((1)性別(2)原發位置(3)組織型態(4)性態碼(5)分級(6)區域淋巴腺檢查(7)區域淋巴結侵犯數目(8)外院診斷性及分期性手術處置(9)申報醫院診斷性及分期手術(10)臨床 T (11)臨床 N (12)臨床 M (13)臨床期別組合(14)病理 T (15)病理 N (16)病理 M (17)病理期別(18)原發部位手術邊緣(19)放射治療與手術順序(20)區域治療與全身性治療順序(21)放射劑量(22)外院化學治療(23)申報醫院化學治療(24)生存狀態(25)死亡原因)進行復發的預測。在圖三(I)為第一階段研究流程:未經變數篩選直接以 SVM、C5.0 決策樹、ELM、MARS、RF 方法進行預測；在圖三(II)為第二階段研究流程:藉由變數重要性篩選，經過變數篩選後進行分層再以五種資料探勘方法進行預測。最後，進一步比較兩個流程所分析分類準確率，針對所分析的變數結果，依照病患屬性完成臨床後續預測大腸直腸癌復發重要因子的建議。



圖三、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大腸直腸癌復發。因此，本研究試圖利用五種資料探勘方法由大腸直腸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γ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$\widetilde{N}$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

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

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

其中$w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$, i =1,2,...,$\widetilde{N}$，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$\widetilde{N}$。

步驟 1. 隨機給一輸入權重w_i以及閾值b_i，, $\mathrm{i} = 1, \dots, \widetilde{N}$

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=$[\mathrm{t}_{1,\dots,}\mathrm{t_n}]^{\mathrm{T}}$。

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：

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

隨著樹的數目增加，PE^*就會趨向於

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

而分類器{h(X,θ)}的強度可以表示為

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

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

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

根據 Breiman(2001)可推導出

$$\begin{aligned}\mathrm{var}(\mathrm{mr}) &= \bar{\rho}\left(E_\theta sd(\theta)\right)^2 \\ &\leq \bar{\rho}E_\theta var(\theta)\end{aligned}$$

$$\geq 1 - s^2 \tag{20}$$

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

評估方法:

對癌症登記資料庫各欄位研究的評估方法採用的 ROC(Receiver Operating Characteristic)之 AUC(Area Under Curve)作為評估結果。ROC 曲線和 AUC 常被用來評價一個二值分類器的優劣。對於分類器或者分類演算法，評估指標主要有精確度(Precision)、召回(Recall)、F 值。ROC 曲線的橫坐標是 FPR(False Positive Rate)，縱坐標是 TPR(True Positive Rate)。FPR 與 TPR 的具體定義如下：

	P	N
Y	True Positives (TP)	False Positives (FP)
N	False Negatives (FN)	True Negatives (TN)

$$FPR = \frac{FP}{FP+TN} \tag{21}$$

$$TPR = \frac{TP}{TP+FN} \tag{22}$$

$$Precision = \frac{TP}{TP+FP} \tag{23}$$

$$Recall = \frac{TP}{TP+FN} \tag{24}$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{25}$$

$$F_1 = \frac{2}{1/Precision + 1/Recall} \tag{26}$$

(五) 實證研究

在研究中，我們由中山醫學大學附設醫院癌症防治中心癌症登記資料庫提供的大腸直腸癌數據集，使用 C5.0、MARS、RF、SVM、ELM 驗證其敏感度與特異度，並預測大腸直腸癌復發之重要因子。數據集中共包含 25 個預測變數，分別為(1)性別(2)原發位置(3)組織型態(4)性態碼(5)分級(6)區域淋巴腺檢查(7)區域淋巴結侵犯數目(8)外院診斷性及分期性手術處置(9)申報醫院診斷性及分期手術(10)臨床 T (11)臨床 N (12)臨床 M (13)臨床期別組合(14)病理 T (15)病理 N (16)病理 M (17)病理期別(18)原發部位手術邊緣(19)放射治療與手術順序(20)區域治療與全身性治療順序(21)放射劑量(22)外院化學治療(23)申報醫院化學治療(24)生存狀態(25)死亡原因以及 1 個目標變數為復發型態（Type of Recurrence），共 606 筆資料，隨機選取 182 筆資料為測試樣本，其餘 424 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：沒有復發。因此{1-1}代表：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：原始判定為沒有復發，經由模式判定亦為沒有復發。由表 2 知 C5.0 的整體正確判別率為 86.825%，而個別的判別正確率以{2-2}的比率最高，為 93.62%：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 92.31%。其中有 1 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 3 個原本群體為第 2 類的樣本，被錯分為第 1 類的群體中。

表2 使用C5.0分類結果

	分類	
類別	1 (復發)	2 (無復發)
1 (復發)	12(92.31%)	1(7.69%)
2 (無復發)	3(6.38%)	44(93.62%)

平均分類準確率：86.825%

由表3可知MARS的整體正確判別率為83.701%，而個別的判別正確率以{1-1}的比率最高，為92.31%：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92.31%；而{2-2}的判別正確率較差，為91.49%。

表3 使用MARS分類結果

	分類	
類別	1 (復發)	2 (無復發)
1 (復發)	12(92.31%)	1(7.69%)
2 (無復發)	4(8.33%)	43(91.49%)

平均分類準確率：83.701%

由表4可知RF的整體正確判別率為77.736%，而個別的判別正確率以{1-1}的比率最高，為100%：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100%；而{2-2}的判別正確率為70.21%。

表4 使用RF分類結果

	分類	
類別	1 (復發)	2 (無復發)
1 (復發)	13(100%)	0(0%)
2 (無復發)	14(%23.33)	33(70.21%)

平均分類準確率：77.736%

由表5可知SVM的整體正確判別率為75.939%，而個別的判別正確率以{2-2}的比率最高，為82.98%：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92.31%；而{2-2}的判別為82.980%。

表5 使用SVM分類結果

	分類	
類別	1 (復發)	2 (無復發)
1 (復發)	12(92.31%)	1(7.69%)
2 (無復發)	8(17.02%)	39(82.98%)

平均分類準確率：75.939%

由表6可知ELM的整體正確判別率為86.665%，而個別的判別正確率以{1-1}的比率最高，為92.31%：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92.31%；而{2-2}的判別正確率較差，為95.74%。

表6 使用ELM分類結果

	分類	
類別	1 (復發)	2 (無復發)
1 (復發)	12(92.31%)	1(7.69%)
2 (無復發)	2(4.26%)	45(95.74%)

平均分類準確率：86.665%

本研究使用集成學習投票策略，相關貢獻率之排序如表7所示，依被選取次數將重要變數排名前五名依序為手術邊緣、pM、pN、淋巴結侵犯數、cT。

表7 變數重要性排名

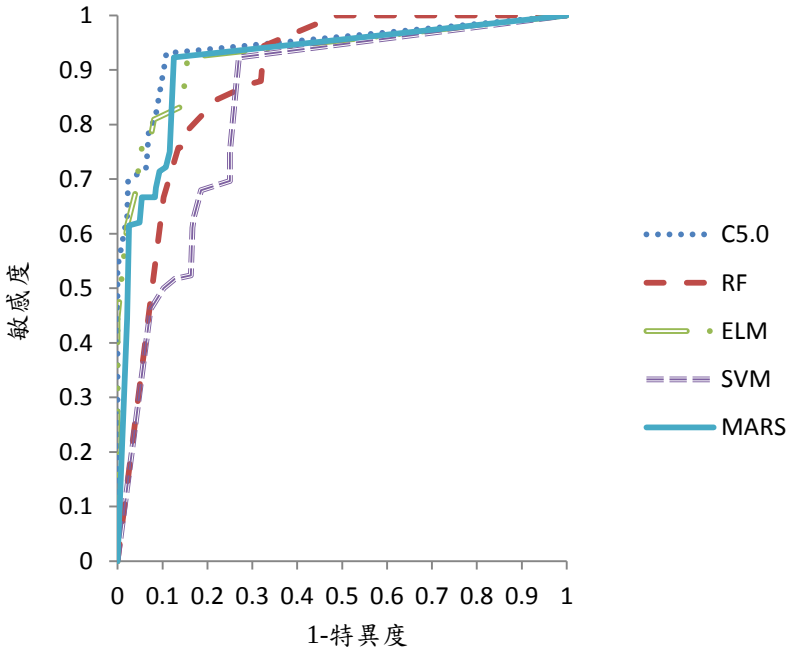
變數	貢獻率排名	變數	貢獻率排名
手術邊緣	1	cN	14
pM	2	有放射	14
pN	3	分級與分化	16
淋巴結侵犯數	4	pT	17
cT	5	原發部位	18
pStage	6	有手術	18
本院處置方式	7	外院處置方式	20
Cm	7	性態碼	21
全身治療	7	性別	22
本院化學治療	10	外院化學治療	23
放射+手術	11	組織型態	24
淋巴結檢查數	12	放射治療劑量	25
cStage	12		

在整體預測結果中，如表 8 所示，MARS 模式在{1-1}產生最高平均分類準確率，為 83.27%；而 C5.0 在{2-2}也產生最高平均分類準確率，為 94.60%；在整體情況下，我們可以看到 C5.0 模式優於 MARS、RF、SVM 和 ELM 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8 C5.0、MARS、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	MARS	RF	SVM	ELM	C5.0	MARS	RF	SVM	ELM	C5.0	MARS	RF	SVM	ELM
1	92.31	92.31	100	92.31	92.31	93.62	91.49	70.21	82.98	95.74	93.33	91.67	76.67	85.00	95.00
2	80.95	71.43	80.95	52.38	71.43	100	97.44	89.74	89.74	100	93.33	88.33	86.67	76.67	93.33
3	53.85	61.54	84.62	46.15	61.54	97.87	97.87	68.09	87.23	97.87	88.33	90	71.67	78.33	90.00
4	63.16	68.42	78.95	63.16	68.42	100	95.12	85.37	92.68	95.12	88.33	86.67	83.33	83.33	86.67
5	79.17	75.00	87.5	75.00	75.00	89.19	94.59	86.49	72.97	91.89	85.25	86.89	86.89	73.77	85.25
6	72.22	72.22	94.44	61.11	72.22	93.02	90.7	51.16	83.72	95.35	86.89	85.25	63.93	77.05	91.80
7	93.10	62.07	75.86	51.72	62.07	93.75	87.5	84.38	75.00	84.38	93.44	75.41	80.33	63.93	78.69
8	27.78	44.44	66.67	50.00	44.44	97.67	88.37	67.44	81.40	93.02	77.05	75.41	67.21	72.13	78.69
9	72.00	66.67	88.00	68.00	66.67	91.67	91.67	77.78	83.33	94.44	83.61	80.33	81.97	77.05	86.89
10	69.70	66.67	75.76	69.70	66.67	89.29	89.29	82.14	75.00	85.71	78.69	77.05	78.69	72.13	80.33
平均	70.42	83.27	73.01	62.95	68.07	94.60	92.40	76.28	82.40	93.35	86.82	83.70	77.73	75.93	86.66

接著，本研究使用ROC曲線下方的面積大小來評估C5.0、MARS、RF、SVM、ELM模式五種機器學習方法的效能和鑑別度(如圖四)，C5.0有最大的曲線下的面積AUC為0.9405。

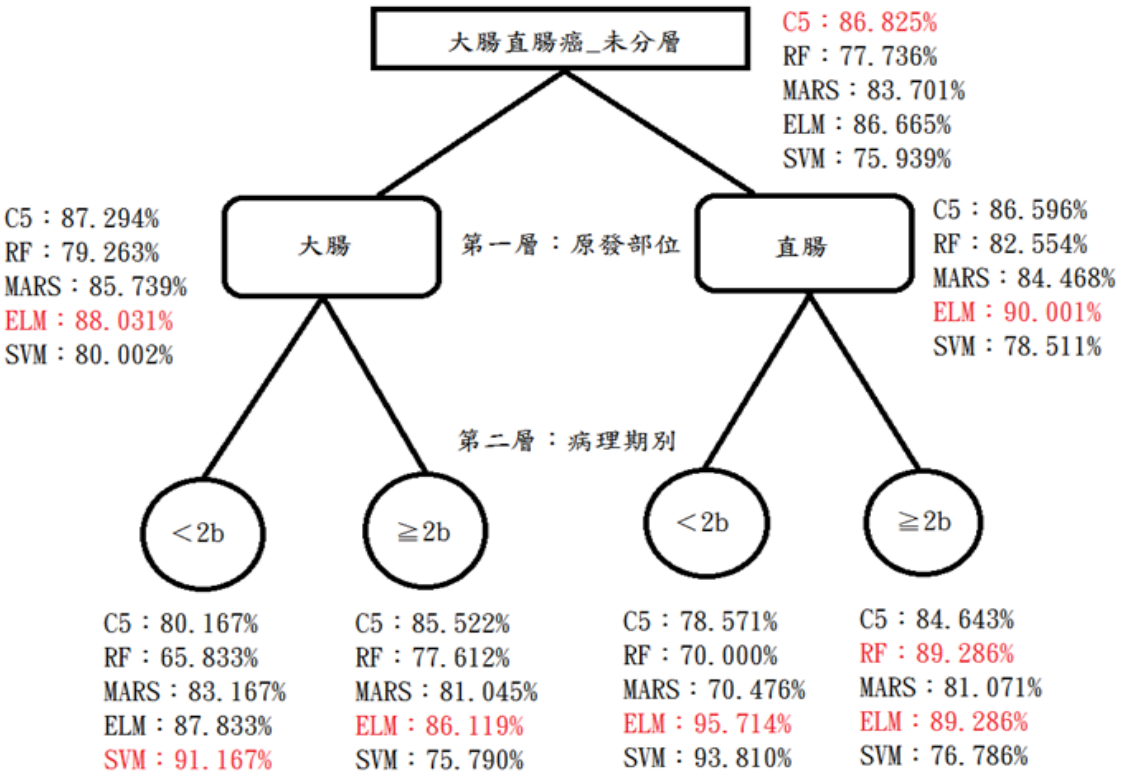


圖四、 ROC曲線

	AUC
C5.0	0.9405
MARS	0.9165
RF	0.8871
SVM	0.8395
ELM	0.9281

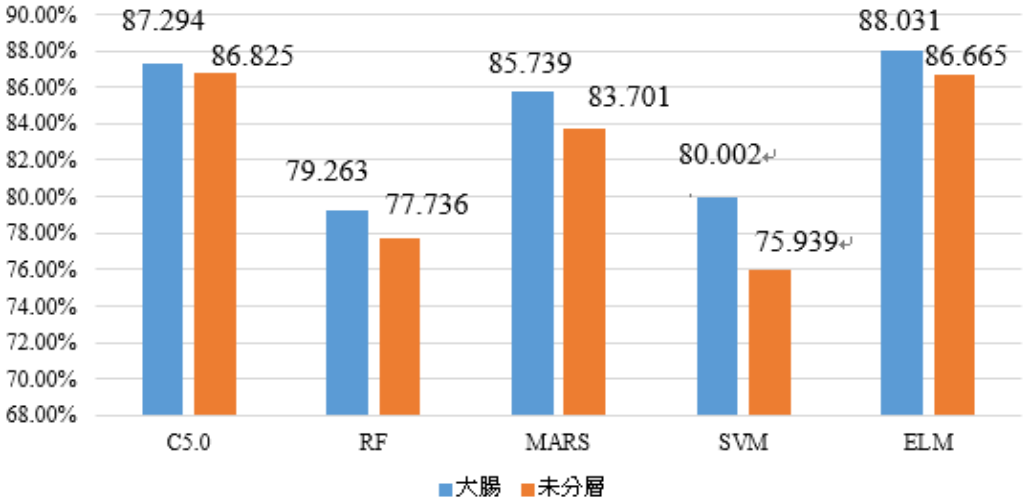
表9 AUC曲下面積

第二階段，首先經由C5.0分析後發現病理期別的貢獻率最高，因此本研究以病理期別與原發部位進行分層後分析結果，在大腸原發部位：期別<IIb的準確率以SVM(91.167%)為最佳；期別≥IIb的準確率以ELM(86.119%)最高；在直腸原發部位：期別<IIb的準確率以ELM(95.714%)最佳；期別≥IIb的準確率以MARS(89.286%)與ELM(89.286%)最高(如圖五所示)。本研究結果可以針對大腸直腸癌症病患原發部位與病理期別的分層提供臨床醫師輔助預測。整體資料分析，針對(1-1)敏感度與(2-2)特異度進一步做分析 在位分層前之敏感度以RF為最高準確率83.275%，經過分層分別在大腸以及直腸也是以RF預測的敏感鑑別度最高，分別是83.314%以及90.470%，進一部針對病理期別在<IIb時，RF敏感度的鑑別率可以達到100%。另外針對特異度而言，在未分層前以C5.0平均的預測準確率最高，經過分層後 ELM在大腸與直腸的預測準確率為最高，分別是95.050%以及93.584%，進一步再針對病理期別<IIb時，ELM在特異度中亦是最高，分別是大腸原發部位98.18%以及直腸原發部位98.45%。



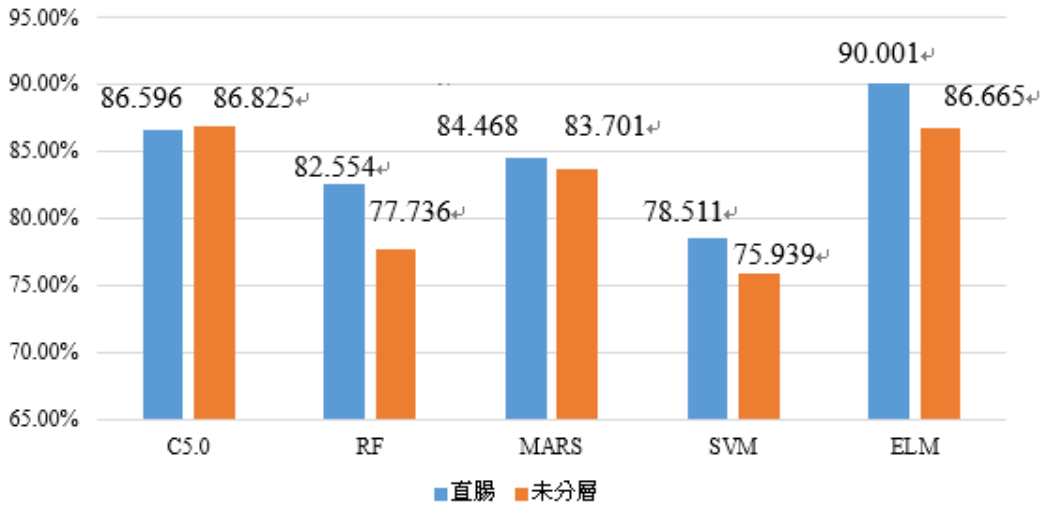
圖五、分層結果圖

在未分層之原發部位在大腸，我們可以發現ELM的準確值最高，且原發部位在大腸的準確值亦是五種方法準確值最高，如圖六所示。



圖六、原發部位為大腸之分層前後方法準確率比較

在未分層之原發部位在直腸，我們可以發現ELM的準確值最高，且可以觀察出五個方法中只有C5.0在未分層的準確率比病理期別在直腸來的高，如圖七所示。



圖七、原發部位為直腸之分層前後方法準確率比較

(六) 結論

為了獲得更佳的大腸直腸癌復發之重要因子，本專題研究使用多種資料探勘方法找出復發的危險因素。研究結果支持原發部位和病理期別是重要的預測復發影響因子。進一步比較C5.0、RF、MARS、ELM和SVM在大腸直腸癌的預測準確度。在未分層C5.0(86.825%)表現優於其他方法；在分層後依照原發部位與病理期別進行分析，在大腸原發處：期別<2b，以SVM預測復發準確率最高，其次為ELM、MARS、C5.0、RF，期別≥2b，以ELM預測復發準確率最高，其次為C5.0、MARS、RF、SVM，在直腸原發處：期別<2b狀況，以ELM預測復發準確率最高，其次為SVM、C5.0、MARS、RF，期別≥2b狀況，以ELM與RF最高，其次為C5.0、MARS、SVM。本研究結果證實針對大腸直腸癌症病患的復發性預測，可以原發部位與病理期別的分層樣式提供臨床醫師輔助治療。另外對於手術邊緣、pM、pN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，相信也能具體反應大腸直腸癌復發之重要變數。

(七)文獻參考

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

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, *Numerische 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 Proceedings, pp. 289-292.

Friedman J. H. (1990) *Multivariate Adaptive Regression Splines*, Department of Statistics, Stanford University, Technical Report 102 Rev.

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

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. and Siew C. X. (2006) Extreme learning machine: theory and applications, *Neurocomputing*, Vol. 70, pp. 489-501.

Hveem T. S., Merok M. A., Pretorius M. E., Novelli M., Bæ vre M. S., Sjo O. H. and Danielsen H. E. (2014) Prognostic impact of genomic instability in colorectal cancer, *British Journal of Cancer*, Vol. 110, pp.

2159-2164.
Larose D. T. (2005) *Discovering Knowledge in Data: An Introduction to Data Mining*, New Jersey, John Wiley & Sons, Inc.
Leung W. H. and Liu C. K. (2014) Chemotherapy and Targeted Therapy in Colorectal Cancer: The Current Status, *Journal of Cancer Research and Practice*, Vol. 30, no. 1, pp. 11-20.
Li F. G., Wang Z. P., Hu G. and Li H. (2011) Current status of SNPs interaction in genome-wide association study, *Yi Chuan*, Vol. 33, no. 9, pp. 901-10.
Li S., James T. K., Zhu H., and Wang Y. (2003) Texture classification using the support vector machines, *Pattern Recognition*, Vol. 36, pp. 2883-2893.
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. 2, pp. 160-171.
Ong L. S., Shepherd B., Tong L. C., Seow-Choen F., Ho Y. H., Tang C. L. and Tan K. (1997) The colorectal cancer recurrence support (CARES) system, *Artificial Intelligence in Medicine*, Vol. 11, no. 3, pp. 175-188.
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.
See5: (Accessed May 10, 2007) An Informal Tutorial <http://www.rulequest.com/see5-win.html>.
Steinberg D., Bernard B., Phillip C. and Kerry M. (1999) MARS User Guide. San Diego, CA, *Salford Systems*.
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.
Walker A. S., Johnson E. K., Maykel J. A., Stojadinovic A., Nissan A., Brucher B. and Steele S. R. (2014) Future Directions for the Early Detection of Colorectal Cancer Recurrence, *Journal of Cancer*, Vol. 5, no. 4, pp. 272-280.
International Agency for Research on Cancer (IARC)(2016)，取自 http://globocan.iarc.fr/Pages/Map.aspx.
中央健保局(2006)，全民健保預防保健服務，取自 http://www.nhi.gov.tw/.
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。
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。
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。
衛生署福利部國民健康署(2016)，取自 http://www.hpa.gov.tw/。