中華民國第 60 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

團隊合作獎

052506

以人工智慧協助腎臟基因變異之治病預測

學校名稱：國立臺南第一高級中學

作者：	指導老師：
高一　卜德璇 高二　卜德玥	顏永進

關鍵詞：人工智慧、網路爬蟲、生物資訊

摘要

本研究致力於尋找在資料不足的情況下最佳的基因預測模式。

首先，訓練預測腎臟基因(PKD1)變異致病力模型，利用三種訓練方法：「管線」(Pipeline)、SVM 分類器、隨機森林相關分類器訓練資料，並調整相關參數。其中，以隨機森林準確率最高。另外，利用「網路爬蟲」技術整合四個大型網路基因變異註釋工具，製作單一介面方便使用者輸入資料，能夠大幅減少醫生預測病人基因變異致病性的時間，具有極高的實用性。為了嘗試解決基因變異致病力預測資料不足的問題，本研究利用網路工具所預測之致病力分數訓練模型，然基因變異位置複雜，資料量仍有所欠缺。

最後，期望利用本研究「整合網路基因註釋工具」蒐集醫生資料，提高模型資料量同時提高準確率。

壹、 研究動機

以人工智慧訓練基因變異預測疾病模型有數據不足的問題。

未來，「精準醫療」只要檢測一下 DNA，立刻就能對症下藥，但這美好的前景，必須結合眾多分子生物學及人工智慧的研究與合作才能完成。當獲得病人定序後的基因時，如何針對此龐大且複雜的資料進行比對分析，進而應用於臨床或研究上，診斷病人的疾病？大數據整合、機器學習、深度學習及人工智慧等新潮 AI 科技可謂打進「精準醫療」的核心。

對模型進行深度學習需要龐大的資料，如何在數據不足的情況下訓練模型，從變異的基因預測病人的疾病？

因此，本研究致力於尋找在資料不足的情況下最佳的基因預測模式。希望利用「深度學習」技術，藉由網路公開病例資料訓練模型以預測案例的致病程度；為了取得更完整的基因變異預測結果，以「網路爬蟲」整合各網站基因預測技術；並嘗試利用網路工具所預測之致病力分數訓練模型，以期為「精準醫療」領域貢獻心力。

貳、研究目的

一、利用不同模型訓練預測腎臟基因(PKD1)變異致病力模型。

二、整合網路基因變異註釋工具，提升重要且大型的網路工具可用性。

三、嘗試解決訓練模型資料不足的問題，利用網路工具致病力分數訓練模型。

四、由研究結果提出在資料不足情況下最佳的基因預測模式，並期望取得龐大數據量。

參、研究設備及器材

一、軟體環境：Python3.7，ANACONDA3(Python 程式語言的整合開發環境)，Spyder3

二、訓練資料：

(一) 腎臟基因變異資料(網路資料庫)

Autosomal Dominant Polycystic Kidney Disease: Mutation Database

(二) 腎臟基因(PKD1)變異測試資料：由國家衛生研究院黃道揚教授提供

三、硬體規格：

(一) CPU：Intel(R) Core™ i5-8265U CPU @ 1.60GHz 1.80GHz

(二) 記憶體：12.0GB

(三) 作業系統：Windows 10

肆、研究過程與方法

一、基因變異分析流程

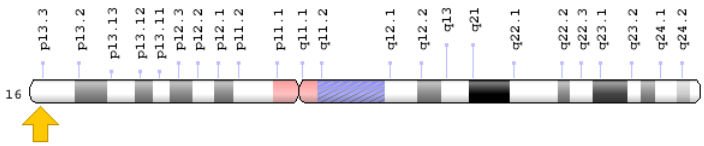
圖一 基因變異分析流

將採檢的病人基因樣本送入實驗室進行萃取，得高純度的 DNA 後再放入定序儀，可得片段式的基因序列。而片段式的基因序列經數位化整理進行「序列組裝」，將每段落的頭尾相接成一個完整的序列，利用電腦軟體幫助有較佳的效率。經過整理的序列需與電腦整合的資料庫內容作比對，執行「變異點偵測」，即可查出和原序列相異之處。再針對「變異點註釋」，探討造成之胺基酸變化、致病程度、變異點的罕見程度等，對於了解此基因變異並幫助醫生診斷至關重要。

本研究針對「變異點註釋」步驟訓練模型、探討特定基因變異特色、並整合網路變異點註釋工具。

二、選擇突變基因：PKD1

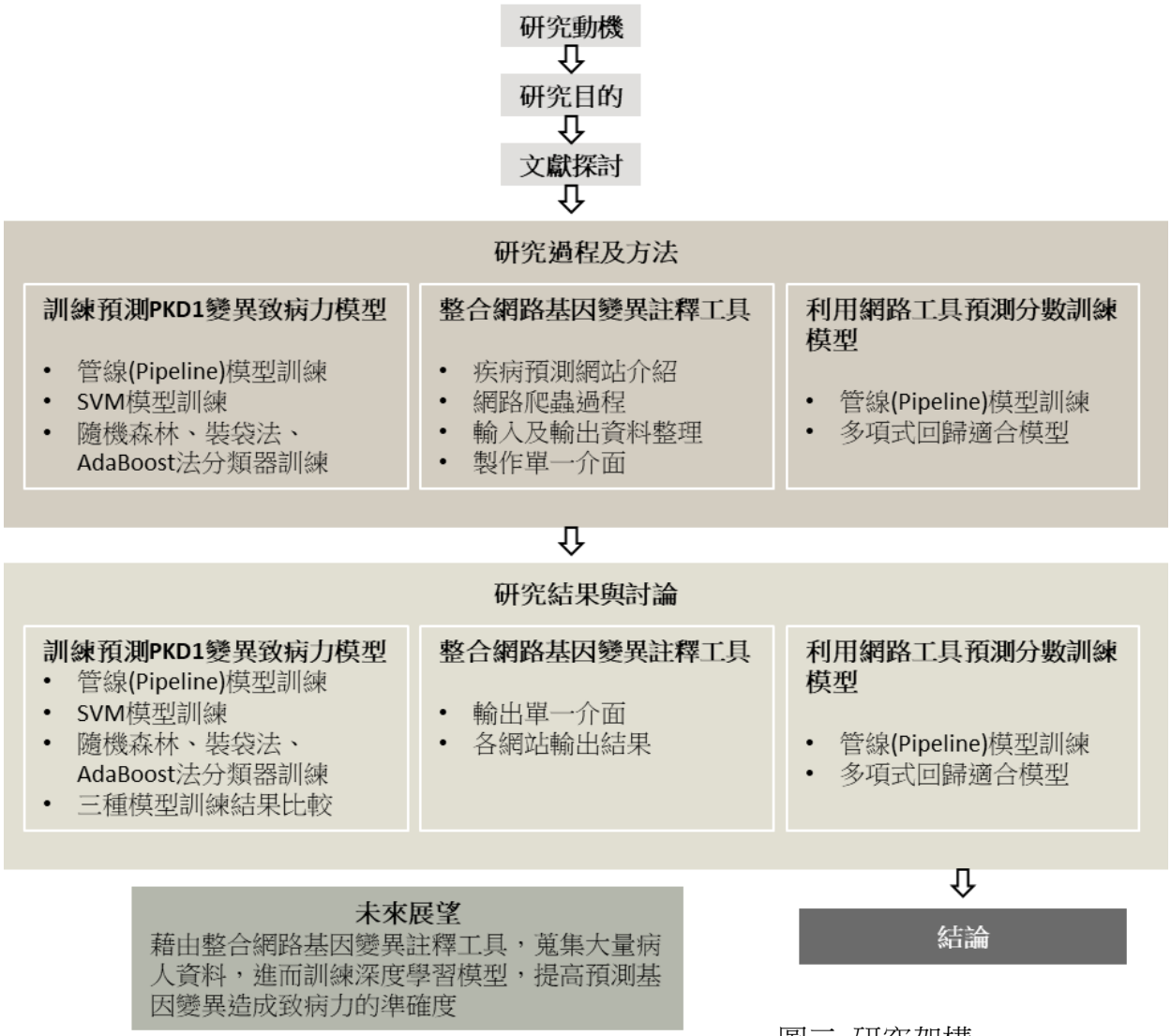
(一) 細胞遺傳學位置(Cytogenetic Location)：16 號染色體短臂(p)位置 13.3。



圖二 PKD1 位置

(二) 基因介紹：PKD1 突變與大多數染色體顯性多囊腎疾病有關，多囊腎是一種遺傳疾病，由於基因的缺陷，使得患者的腎臟會出現大小不等的水泡；多囊腎為體染色體遺傳，故男性以及女性患病的機率是一樣的，且遺傳機率是百分之五十。PKD1 基因編碼決定 PC1 蛋白質。PC1 是一種膜結合蛋白，長度為 4303 個氨基酸，主要表達於初級纖毛，頂膜，粘附連接與橋粒上。

三、研究架構建立



圖三 研究架構

四、訓練預測腎臟基因(PKD1)變異致病力模型

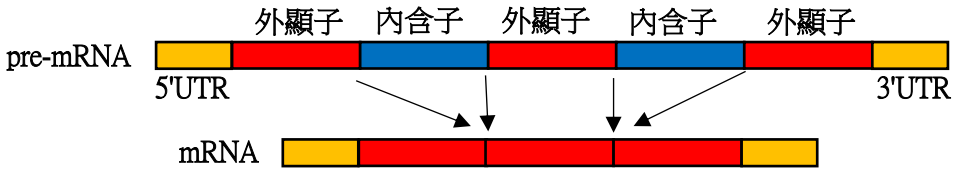
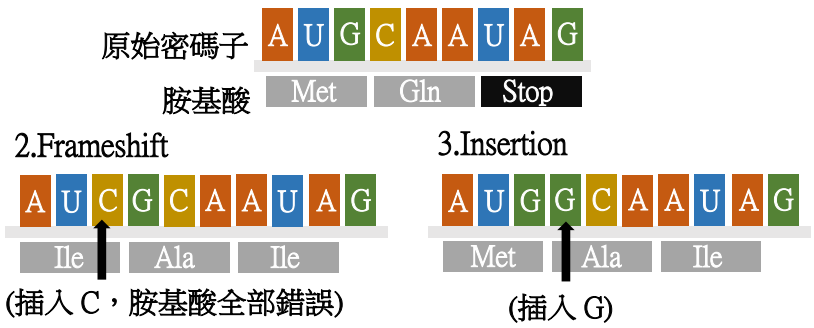
我們自網路公開基因資料庫下載公開的基因變異數據，利用「深度學習」訓練模型，預測腎病基因(PKD1)變異致病力。

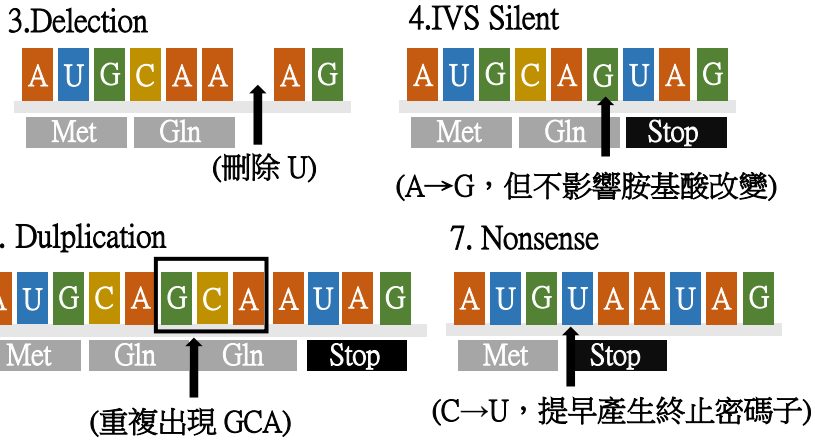
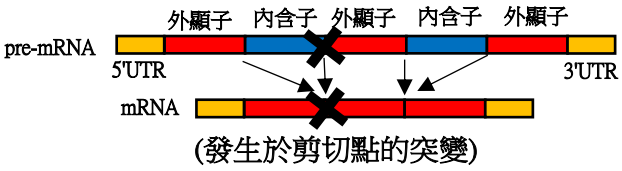
(一) 訓練資料

1. 資料來源：PKD FOUNDATION Autosomal Dominant Polycystic Kidney Disease: Mutation Database，此資料庫收集大量造成多囊腎病（ADPKD）的突變基因 PKD1 及 PKD2，為收集腎病突變基因最多的資料庫。但由於 PKD2 基因資料仍過少(僅 305 筆)，因此本研究只取 PKD1 基因(2332 筆)訓練模型。

2. 資料處理並編碼

表一 資料處理並編碼

<table>
<tr><td rowspan="2">模型說明</td><td>外顯子(Exon)模型</td><td colspan="10">內含子(Intron)模型</td></tr>
<tr><td colspan="6">剪接後仍會被保存下來，並可在轉譯合成過程中被表達成胺基酸。</td><td colspan="5">分開相鄰的外顯子，並阻斷基因線性表現的序列。在 RNA 離開細胞核進行轉譯前被剪除。</td></tr>
<tr><td></td><td colspan="11"> 
</td></tr>
<tr><td rowspan="2">變異位置</td><td>編號</td><td>1</td><td>2</td><td>3</td><td>…</td><td>46</td><td>1</td><td>2</td><td>3</td><td>…</td><td>45</td></tr>
<tr><td>位置</td><td>EX1</td><td>EX2</td><td>EX3</td><td>…</td><td>EX46</td><td>IVS1</td><td>IVS2</td><td>IVS3</td><td>…</td><td>IVS45</td></tr>
<tr><td rowspan="11">變異種類</td><td>編號</td><td colspan="10">變異位置</td></tr>
<tr><td>0</td><td colspan="5">3'UTR</td><td colspan="5">3 端非轉譯區，影響 mRNA 的轉譯，許多其上的突變和特定疾病有關</td></tr>
<tr><td>1</td><td colspan="5">5'UTR</td><td colspan="5">5 端非轉譯區，控制基因表現，其上突變及序列長度已被證實會導致特定疾病</td></tr>
<tr><td>0~1 圖片說明</td><td colspan="10"> 
</td></tr>
<tr><td>2</td><td colspan="3">Frameshift</td><td colspan="7">框移突變，DNA 分子經嵌入或缺失一個或多個（非三之倍數）鹼基，導致多肽鏈中胺基酸轉錄及轉譯全部發生錯誤</td></tr>
<tr><td>3</td><td colspan="3">Insertion or deletion</td><td colspan="7">插入突變(insertion mutation)，增加一個或一段的鹼基/刪除突變(deletion mutation)，DNA 上缺少一個或一段鹼基</td></tr>
<tr><td>4</td><td colspan="3">IVS Silent</td><td colspan="7">同義突變，變異的基因產生和原來相同的胺基酸，因此對生物表型無明顯影響</td></tr>
<tr><td>5</td><td colspan="3">IVS Unknown</td><td colspan="7">未確定其對生物體的健康影響及功能</td></tr>
<tr><td>6</td><td colspan="3">Large Delection or Dulplication</td><td colspan="7">大範圍(>100 個鹼基)的刪除突變或重複基因，特定 DNA 片段發生重複</td></tr>
<tr><td>7</td><td colspan="3">Nonsense</td><td colspan="7">無義突變，發生於點突變，使產生終止密碼子或不完整且通常無功能的蛋白質</td></tr>
<tr><td>2~7 圖片說明</td><td colspan="10"> 
</td></tr>
</table>

			
	8	Splice	剪切位點突變，發生於 RNA 剪接處，包括插入、刪除或序列的改變
	8 圖片說明	 (發生於剪切點的突變)	
	9	Substitution	DNA 中某鹼基被另一鹼基所取代，可導致突變
	10	Synonymous	DNA 序列的改變不會影響蛋白質序列
變異型態	編號	變異型態	
	0	Somatic 體細胞突變：可能造成癌症，不遺傳給後代	
	1	Germline 生殖細胞系突變：遺傳給後代	
致病程度(輸出結果)	編號	致病力	
	A	Definitely Pathogenic 絕對致病	
	B	Highly Likely Pathogenic 高度致病	
	C	Indeterminate 未確定	
	D	Likely Hypomorphic 可能為亞效型(使基因的表現或是基因產物的活性減弱，但不會消失)	
	E	Likely Neutral 可能不影響	
	F	Likely Pathogenic 可能致病	

3. 輸入資料格式

表二 輸入資料格式

類別	ID	變異位置	變異種類	變異型態	致病程度(輸出結果)
舉例	EX1-EX39del	1	6	1	A

(二) 訓練預測致病程度模型

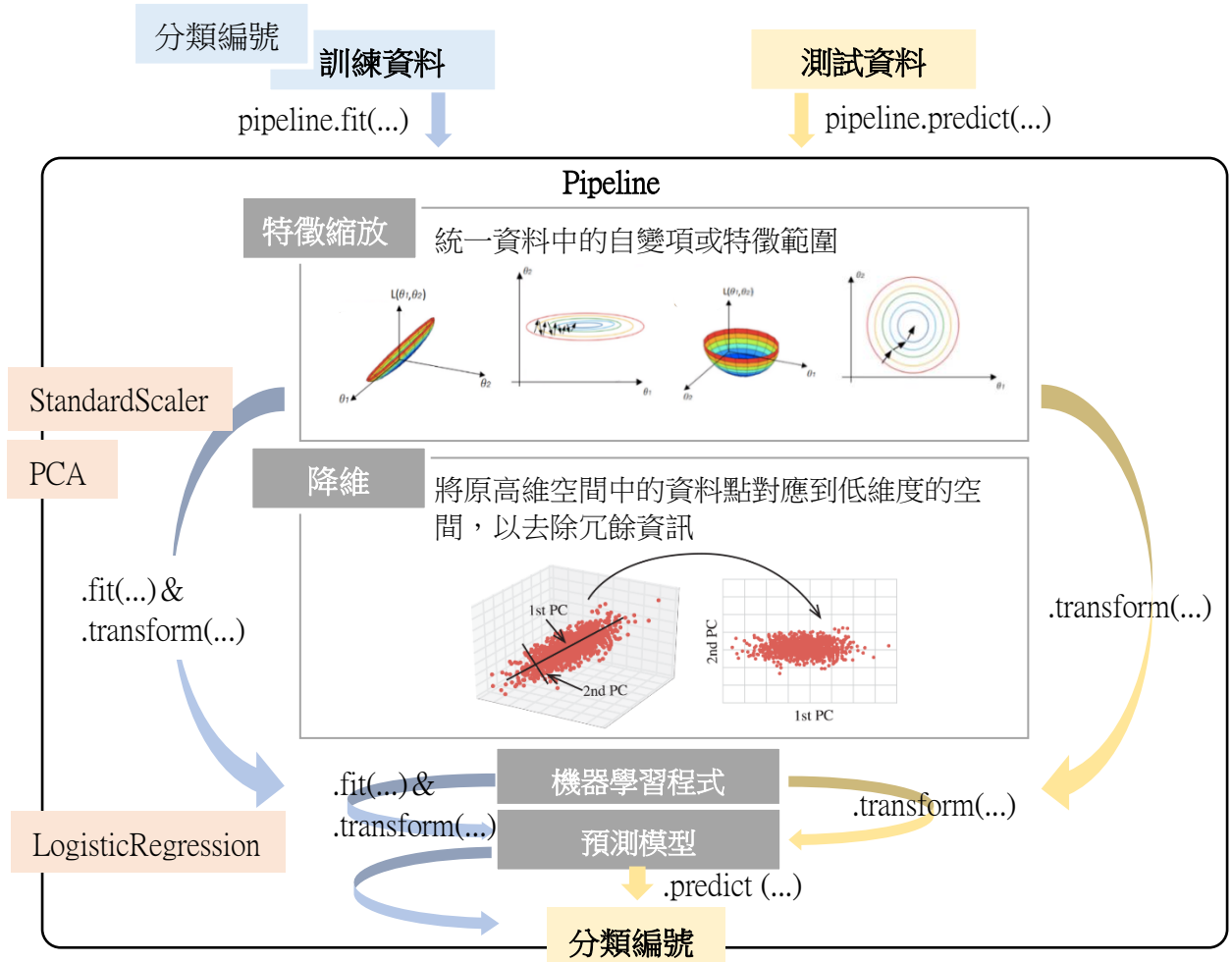
預測致病力模型程式碼
https://github.com/shaniabu/Science-fair-code

1. 利用「管線」(Pipeline)製作訓練模型

(1) 載入腎臟基因數據集

將三個特徵：變異位置、變異種類、變異型態(1-3)指派給 x 陣列；輸出結果：6 種致病程度指派給 y 陣列，並編碼成 0-5。再將全數據拆成「訓練數據集」(80%)及「測試數據」(20%)。

(2) 利用「管線」(Pipeline)將 StandardScaler、PCA、LogisticRegression 等物件串聯起來



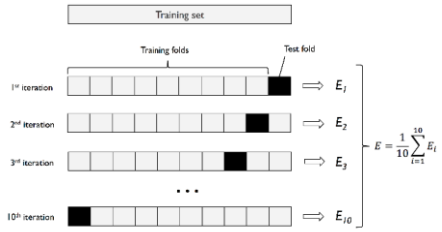
圖四 管線(Pipeline)訓練資料步驟

StandardScaler 和 PCA 對「訓練數據集」套用 fit 與 transform 函數，StandardScaler 計算用於「特徵縮放」的標準差與平均值，擬合並轉換數據；再藉由 PCA(主成分分析)壓縮到二維子空間。經由 StandardScaler 和 PCA 的轉換之後，最後 LogisticRegression「估計器」會「擬合」出訓練模型。

類似上述的 fit 方法訓練模型，Pipeline 同樣可以使用 predict 方法。將數據提供給 Pipeline 的 predict 函數，藉由 transform 函數通過上述中間步驟後，同樣進入最後一步的「估計器」，並回傳對變換後數據的預測。

(3) 利用 k 折交叉驗證法(k-fold cross validation)評估模型效能

我們使用「分層 k 折交叉驗證法」(stratified k-fold cross validation)，每折數據中類別大小比率和原數據是相同的，並用 k=10、15、20 來訓練數據集。

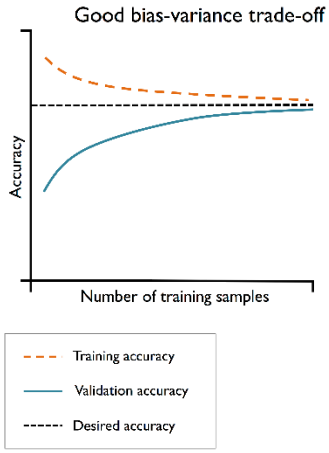
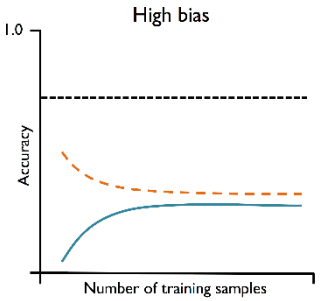
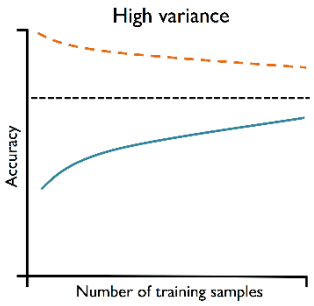


圖五 k 折交叉驗證法

(4) 運用「學習曲線」診斷高偏誤或高方差

透過製作不同訓練集數目下的學習曲線：訓練數據集及驗證數據集正確率的變化，可以檢驗該模型是否具有「高偏誤」或「高方差」的現象。

表三 診斷高偏誤或高方差方法

效能佳模型	高偏誤模型	高方差模型
	「訓練正確率」與「驗證正確率」都很低，表示它「低度適合」(underfitting)	「訓練正確率」和「驗證正確率」相差太多，表示它「過度擬合」(overfitting)
		

(5) 運用「驗證曲線」討論低度適合或過度擬合

透過製作不同 LogisticRegression 參數 C 下「訓練正確率」及「驗證正確率」變化，可以檢驗該模型是否具有過度擬合或低度適合的現象。並找出 C=0.001、0.01、0.1、1.0、10.0、100.0 時，何者擁有最佳的模型。

參數 C 是正則化係數 λ 的倒數，控制正則化程度的「超參數」，正則化是用來防止過度擬和的過程。因此，C 越小，表示有越大的正則化。

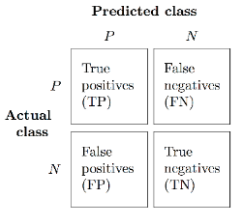
$$J(\theta)_{L2} = C \times J(\theta) + \sqrt{\sum_{j=1}^{n}(\theta_j)^2 (j \geq 1)}$$

(參數θ(w)向量中的每個參數的平方和的開方值)

(6) 讀取混淆矩陣(confusion matrics)

混淆矩陣包含「分類器」對「測試數據集」所犯下的不同類型的各種錯誤，讀取混淆矩陣可以幫助釐清模型需改善的部分。實際類別及預測類別為 6 種致病程度，編號為 0 至 6。

真陽(True Positive, TP)：預測為 Positive 且預測準確 True
真陰(True Negative, TN)：預測為 Negative 且預測準確 True
偽陽(Flase Positive, FP)：預測為 Positive 但預測錯 False
偽陰(Flase Negative, FN)：預測為 Negative 但預測錯 False

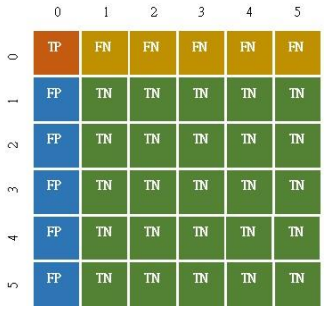


圖六 混淆矩陣

接著計算混淆矩陣各項的 F1 值，它是衡量二分類模型的一種指標。先算出各項準確率(Precision, PRE)及召回率(Recall, REC)，其調和平均數即 F1 值。再計算整體的「宏觀平均」(macro average)和「微觀平均」(micro average)的 F1 值。

$$\boldsymbol{Precision} = \frac{TP}{TP + FP}, \boldsymbol{Recall} = \frac{TP}{TP + FN}$$

$$\mathbf{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$



圖七 (編號 0)混淆矩陣

(7) 用外顯子和內含子資料由上述步驟各自訓練，得兩個預測基因致病力模型。

2. 利用 SVM(支援向量機)訓練分類器

(1) 資料處理

將基因資料中致病程度的編號 A 及類別 B 合併，將編號 C 和編號 F 合併，以及編號 E，共分為三個類別。並將 30%歸類為測試數據集，70%為訓練數據集。

(2) 訓練支援向量機(SVM)分類器

藉由調整參數 c 的值，尋找支援向量機的最大化邊界，以獲取最佳化模型。接著，藉由 sklearn.svm 套件的 SVC 類別，參數 kernel='linear'，訓練 SVM 分類器，並繪製決策區域。

(3) 訓練核支援向量機(kernel SVM)分類器

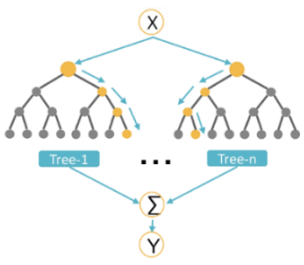
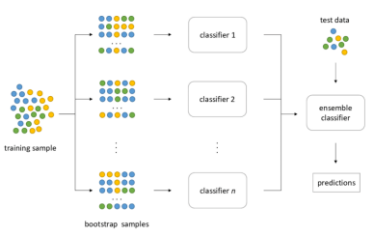
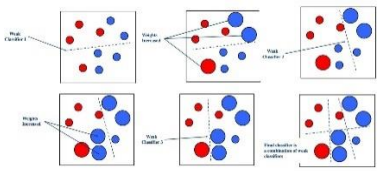
核心(kernel)為成對相似樣本的的「相關度函數」，利用「核技巧」(kernel trick)在高為空間中找到分離超平面，同樣套用 sklearn.svm 套件的 SVC 類別，參數 kernel='rbf'， 訓練核支援向量機來繪製「非線性決策邊界」。

接著，調整參數 γ 值，藉由增加 γ 值能夠產生較「顛簸」的邊界，本研究輸出 γ =0.2 及 100 的決策邊界。

3. 利用隨機森林、裝袋法、AdaBoost 法訓練模型

(1) 三種分類器介紹

表四 三種分類器介紹

隨機森林(Random Forest)	裝袋法(Bagging)	AdaBoost 法
決策樹分類器是藉由特徵值將數據分割到不同群組後，分別計算其資訊增益(Imformation gain, IG)，再藉由設定目標函數(objective function)，尋找出準確率最高的分割方式。「隨機森林」則是藉由組合多高變異的深度決策樹，將他們的結果平均。	從初始訓練集中抽取「自助式樣本」(「放回式」的隨機樣本)來做適合。接著，每個「自助樣本」被用來適合分類器，當各分類器完成預測，再進行投票，以「多數決」來做預測。	有別於「裝袋法」，此種「強化法」以「不放回式隨機抽樣」來做預測，並且讓學習器從「誤判訓練樣本」之中去學習，強化整體效能。一開始訓練一個單層決策樹，每個訓練樣被賦予相同權重；下一輪中，給定錯誤分類樣本最大的權重值，同時降低錯誤樣本的權重；最後再以多數決做預測。
		

(1) 資料處理

將基因資料中致病程度的編號 A 及類別 B 合併，將編號 C 和編號 F 合併，以及編號 E，共分為三個類別。

(2) 利用「隨機森林」(Random Forest)、裝袋法(Bagging)、AdaBoost 法訓練模型，80%訓練數據集，20%測試數據集。

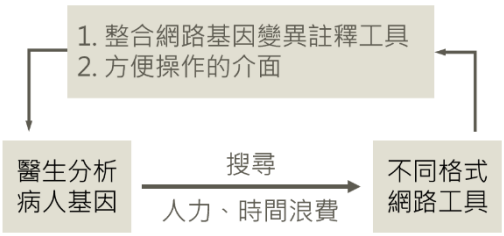
(3) 改變隨機森林的決策樹分類器數量，以分類器數量為 1 至 300 進行測試。

(4) 改變不純度計算公式，並利用「Gini 不純度(Gini Impurity)」及「熵(entropy)」訓練模型，尋找能最佳化預測結果的模式。

五、整合網路基因變異註釋工具

整合基因變異註釋工具程式碼
https://github.com/shaniabu/combine-gene-annotation-tools

網路上有許多針對基因變異的註釋工具，包括 The Mutation Significance Cutoff (MSC) Server、PROVEAN HUMAN GENOME VARIANTS、Fathmm (Functional Analysis through Hidden Markov Models (v2.3))和 Mutation Assessor。然而，工具眾多使醫生難以快速使用，因此，我們利用「網路爬蟲」技術使基因資料統一填入單一介面，即可輸出所有網站資料，製作具實用性且方便快速的基因註釋工具。

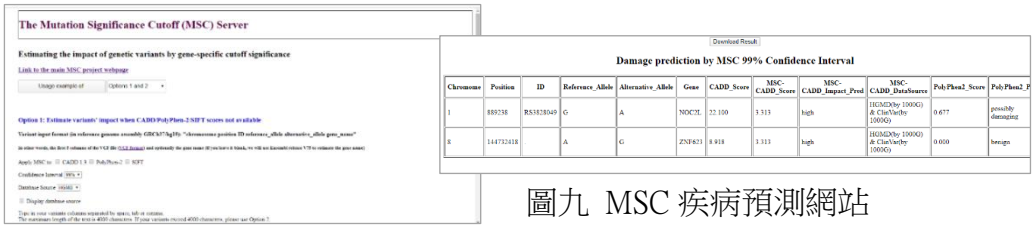


圖八 網路基因註釋工具使用目的

(一) 疾病預測網站

1. The Mutation Significance Cutoff (MSC) Server

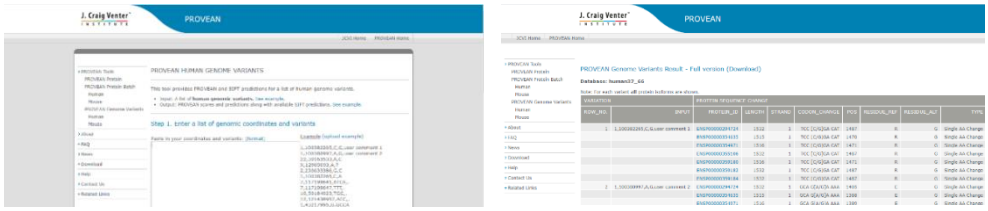
該網站能透過突變顯著性臨界值（MSC）、信心區間基因庫的揀選，多方面比對基因，並取得當特定基因高表現量時，或是基因發生特定突變，容易引起的疾病。



圖九 MSC 疾病預測網站

2. PROVEAN HUMAN GENOME VARIANTS

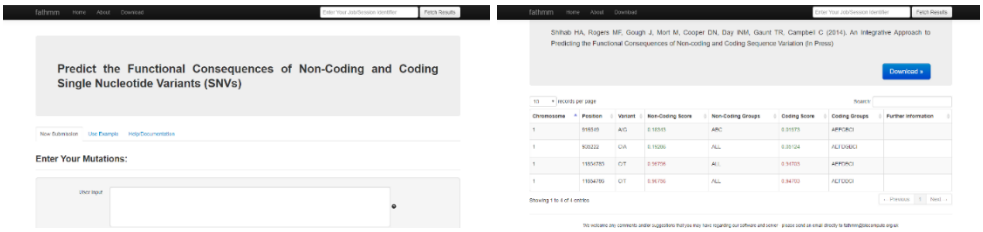
此網站利用數個網站及分析方式，通過多方的比較及權重分配，揀選出當特定基因變異時，最有可能導致的疾病。



圖十 PROVEAN 疾病預測網站

3. Fathmm (Functional Analysis through Hidden Markov Models (v2.3))

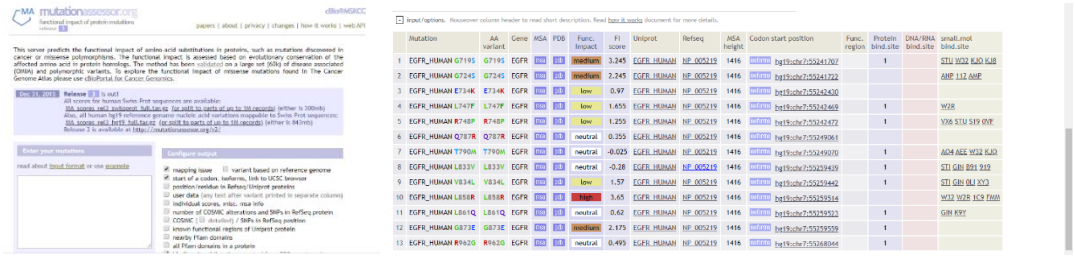
此網站利用數個網站及分析方式，通過多方的比較及權重分配，揀選出當特定基因變異時，最有可能導致的疾病。



圖十一 Fathmm 疾病預測網站

4. Mutation Assessor

此網站利用數個網站及分析方式，通過多方的比較及權重分配，揀選出當特定基因變異時，最有可能導致的疾病。透過輸入變異基因所在的染色體編號、基因的位置、及變異前後的鹼基，預測此基因變異後可能致病的機率。



圖十二 Mutation Assessor 疾病預測網站

(二) 網路爬蟲過程

1. The Mutation Significance Cutoff (MSC) Server

(1) 資料整理：由於每個網站要求的輸入格式各不相同，因此必須先將資料整理為如下形式：(變異基因所在染色體　變異基因編號 備註 變異基因原始鹼基 變異後的鹼基)(範例： 1　916549　A　G)

(2) 爬蟲過程：本研究將經過自動化整理好的資料輸入網站後，在揀選需要的設定後，就能送出後並經網站分析後獲得資料結果。不過有限於此網站之設定，需先在爬蟲前新增一 csv 檔以紀錄資料。

```
url = 'http://pec630.rockefeller.edu:8080/MSC/'

# pretend as a real browser
headers = {'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.70 S

# request the website for cookie
req = requests.get(url, headers=headers)

cookie = req.headers['Set-Cookie']
print(cookie)
cookie = cookie[:cookie.index(';')]
headers['Cookie'] = cookie

# put the query info such as "Confidence Interval", "Apply MSC to" and  "variants" into the url, and needs to be encoded for url, see
variantTextArea = urllib.parse.quote(variantTextArea)
url = 'http://pec630.rockefeller.edu:8080/MSC/VariantTextServlet?measure0=CADD&measure0=PolyPhen/2&measure0=SIFT&confidenceInterval0=c

# this action is like pressing the "Submit" button
req = requests.get(url, headers=headers)

# getting the "Download Result", this needs the cookie from above
url = 'http://pec630.rockefeller.edu:8080/MSC/DownloadServlet'
req = requests.post(url, headers=headers)
```

圖十三 MSC 網站爬蟲程式碼

(3) 結果整理：由於並不是所有爬蟲下來的資料都是我們需要的資料，故在爬蟲後仍需進行資料的篩選。本研究在此網站所擷取的表格資料中指選「CADD_Score」、「MSA-CADD_Score」、「PolyPhen2_Score」、「hvar_prediction」、「SIFT_Score」及「SIFT_Pred」等欄位

2. PROVEAN HUMAN GENOME VARIANTS

(1) 資料整理由於每個網站要求的輸入格式各不相同，因此必須先將資料整理為如下形式：變異基因(所在染色體, 變異基因編號, 變異基因原始鹼基, 變異後的鹼基)。(範例 1,916549,A,G)

(2) 爬蟲過程：本研究將經過自動化整理好的資料輸入網站後，在揀選需要的設定後，就能送出後並經網站分析後獲得資料結果。不過有限於此網站之設定，本研究只能等結果出來後再手動點選「Download」就能獲取需要的資料了。

```

url = 'http://provean.jcvi.org/genome_submit_2.php?species=human&fbclid=IwAR2kfnydOI6wsWab0CAfNw0zIeaRVeTJsCrSmdWvM1CwmBNoUo8IAHdQ_JQ'

driver = webdriver.Chrome(executable_path='C:\Program Files (x86)\Google\chromedriver.exe')
driver.get(url)
driver.maximize_window()

driver.find_element_by_id('CHR').send_keys(variantTextArea)

driver.find_element_by_xpath("//input[@value='gene_id']").click()
driver.find_element_by_xpath("//input[@value='gene_name']").click()
driver.find_element_by_xpath("//input[@value='transcript_id']").click()
driver.find_element_by_xpath("//input[@value='transcript_status']").click()
driver.find_element_by_xpath("//input[@value='description']").click()
driver.find_element_by_xpath("//input[@value='gc_content']").click()
driver.find_element_by_xpath("//input[@value='chr_band']").click()
driver.find_element_by_xpath("//input[@value='family_id']").click()
driver.find_element_by_xpath("//input[@value='family_desc']").click()
driver.find_element_by_xpath("//input[@value='uniprot_id']").click()
driver.find_element_by_xpath("//input[@value='refseq_protein_id']").click()
driver.find_element_by_xpath("//input[@value='mim_accession']").click()
driver.find_element_by_xpath("//input[@value='pfam_id']").click()
driver.find_element_by_xpath("//input[@value='tigrfam_id']").click()
driver.find_element_by_xpath("//input[@value='interpro_id']").click()
driver.find_element_by_xpath("//input[@value='go_term_accession']").click()
driver.find_element_by_xpath("//input[@value='go_slim_goa_accession']").click()

```

圖十四 PROVEAN 網站爬蟲程式碼

(3) 結果整理：由於並不是所有爬蟲下來的資料都是我們需要的資料，故在爬蟲後仍需進行資料的篩選。本研究在此網站所擷取的表格資料中指揀選「SCORE」及「PREDICTION(cutoff = -2.5)」二欄位。

3. Fathmm (Functional Analysis through Hidden Markov Models (v2.3))

(1) 資料整理：由於每個網站要求的輸入格式各不相同，因此必須先將資料整理為如下形式：變異基因(所在染色體, 變異基因編號, 變異基因原始鹼基, 變異後的鹼基)。(範例：1,916549,A,G)。

(2) 爬蟲過程：本研究將經過自動化整理好的資料輸入網站後，在揀選需要的設定後，就能送出後並經網站分析後獲得資料結果。不過有限於此網站之設定，本研究只能手動點選「Sumbit」，等結果出來後再手動點選「Download」就能獲取需要的資料了。

```
38 url = 'http://fathmm.biocompute.org.uk/fathmmMKL.htm'
39
40
41 #f = open('FATHMN_data.txt', "r")
42
43 driver = webdriver.Chrome(executable_path='C:\Program Files (x86)\Google\chromedriver.exe')
44 driver.get(url)
45 driver.maximize_window()
46
47 driver.find_element_by_id('batch').send_keys(variantTextArea)
48 # 手動點擊submit 以及download
49 # 複製網頁內容
```

圖十五 Fathmm 網站爬蟲程式碼

(3) 結果整理：由於並不是所有爬蟲下來的資料都是我們需要的資料，故在爬蟲後仍需進行資料的篩選。本研究在此網站所擷取的表格資料中指揀選「Coding Score」及「Coding Groups」二欄位。

4. Mutation Assessor

(1) 資料整理：由於每個網站要求的輸入格式各不相同，因此必須先將資料整理為如下形式：(Hg19, 變異基因所在染色體, 變異基因編號, 變異基因原始鹼基, 變異後的鹼基)。(範例：hg19,1,916549,A,G)

(2) 爬蟲過程：本研究將經過自動化整理好的資料輸入網站後，在揀選需要的設定後，就能送出後並經網站分析後獲得資料結果。不過有限於此網站之設定，本研究只能等結果出來後再手動點選「submit」就能獲取需要的資料了。

```
75 url = 'http://mutationassessor.org/r3/'
76
77 driver = webdriver.Chrome(executable_path='C:\Program Files (x86)\Google\chromedriver.exe')
78 driver.get('http://mutationassessor.org/r3/')
79 driver.maximize_window()
80
81 driver.find_element_by_id('vars').send_keys(variantTextArea)
82 driver.find_element_by_name('tableQ').click()
83 """ 自己按送出 """
84 #driver.find_element_by_xpath("//input[@value=' submit ']").click()
85
```

圖十六 Mutation Assessor 網站爬蟲程式碼

(3) 結果整理：由於並不是所有爬蟲下來的資料都是我們需要的資料，故在爬蟲後仍需進行資料的篩選。本研究在此網站所擷取的表格資料中指揀選「Func. Inpact」、「Msa Height」及「FI Score」三欄位。

(三) 輸入及輸出資料統整

1. 輸入資料

表五 輸入資料格式

第一欄	第二欄	第三欄	第四欄	第五欄	第六欄	第七欄
ID	Chormosome	Region	Reference	Allele	Type	Homo_sapiens_refseq

2. 最終輸出資料

表六 最終輸出資料

網站	索取資料		
MSC	CADD_Score	MSC-CADD_Score	PolyPhen2_Score
	提供的值為「排名」，30 分以上表示「可能有害」，30 分以下表示「可能有益」，分數越高表示越可能有害。	針對特定基因，預期最低臨床 CADD 臨界值	>0.908 較可能有害 0.446~0.908 可能有害 <0.446 良性 未知
	hvar_prediction	SIFT_Score	SIFT_Pred
	根據 PolyPhen2_Score 的各類有效範圍輸出 probably damaging、possibly damaging、或 benign	<0.05 較無害的 >0.05 有害的	根據 SIFT_Score 的各類有效範圍輸出 「D」Deleterious(有害) 「T」Tolerated(無害)
FATHMM	Fathmn_Coding Score	Fathmn_Coding Groups	
	>0.5 deleterious(有害的) <0.5 Neutral or benign(不顯著的或良性)	A~I 表示不同預測方法，同一個基因可以使用多種方法 A 46-way conservation 8 continuous B Histone ChIP-Seq 190 continuous C TFBS PeakSeq 443 continuous D Open chromatin DNase-Seq 122 discrete E 100-way conservation 8 continuous F GC content 1 discrete G Open chromatin FAIRE 19 continuous H TFBS SPP 443 continuous I Genome segmentation 6 categorical J Footprints 41 continuous	
PROVEAN	SCORE(cutoff = -2.5)	PREDICTION (cutoff = -2.5)	
	≦-2.5	Deleterious(有害的)	
	>-2.5	Neutral(不顯著)	

	SCORE(cutoff = 0.05)		PREDICTION (cutoff = 0.05)
	≦0.05		Damaging(有害的)
	>0.05		Tolerated(無害的)
MUTATION ASSESSOR	FI Score	Func.Impact	mutation_MSA height
	F1≦0.8	neutral	可用基因庫數據資料數
	0.8<FI≦1.9	low impact	
	1.9<FI≦3.5	medium impact	
	FI>3.5	high impact	

(四) 製作操作介面

製作上傳檔案方框，使操作者輸入檔案路徑以讀取 excel 檔，並製作 MSC、FATHMN、PROVEAN、MUTATION ASSESSOR 及 EXPORT 按鈕，讓使用者可以選取各網站或全部資料。

並且註明各網站填入注意事項：

表七 各網站填入注意事項

網站名	注意事項
MSC	需自行新增 MSC_result 的 excel 檔
FATHMM	需要手動按下送出以及複製結果網頁成文字檔，檔名為 Fathmn_result
PROVEAN	需要手動點選所需結果檔，並將之存成 Provean_result 的 csv 檔
Mutation Assessor	需要手動按下送出

```python
from PyQt5 import QtCore, QtGui, QtWidgets

class Ui_MainWindow(object):
    def setupUi(self, MainWindow):
        MainWindow.setObjectName("MainWindow")
        MainWindow.resize(1024, 636)
        self.centralwidget = QtWidgets.QWidget(MainWindow)
        self.centralwidget.setObjectName("centralwidget")
        self.label = QtWidgets.QLabel(self.centralwidget)
        self.label.setGeometry(QtCore.QRect(30, 50, 111, 31))
        font = QtGui.QFont()
        font.setPointSize(14)
        self.label.setFont(font)
        self.label.setObjectName("label")
```

```python
    def retranslateUi(self, MainWindow):
        _translate = QtCore.QCoreApplication.translate
        MainWindow.setWindowTitle(_translate("MainWindow", "MainWindow"))
        self.label.setText(_translate("MainWindow", "Upload file : "))
        self.label_2.setText(_translate("MainWindow", "資料格式為:  第一欄  第二欄  第三欄  第四欄  第五欄  第六欄  第七欄"))
        self.label_3.setText(_translate("MainWindow", "ID"))
        self.label_4.setText(_translate("MainWindow", "Chromosome"))
        self.label_5.setText(_translate("MainWindow", "Region"))
        self.label_6.setText(_translate("MainWindow", "Reference"))
        self.label_7.setText(_translate("MainWindow", "Allele"))
        self.label_8.setText(_translate("MainWindow", "Type"))
        self.label_9.setText(_translate("MainWindow", "Homo_sapiens_refseq"))
        self.label_10.setText(_translate("MainWindow", "輸入資料路徑，檔案需為excel"))
```

圖十七 操作介面設置程式碼

```python
    def download_file(self):
        ans = []
        F = open('./web_result/Fathmn_result.txt', 'r')
        while True:
            line = F.readline()
            if not line:
                break
            fathmn = list(line.split("\t"))
            count = len(ans)
            temp = 0
            while count>0:
                if fathmn[1] == ans[count-1][1]:
                    temp=1
                    break
                else:
                    count = count -1
                    temp=0
            if temp==0:
                f = [fathmn[0:4] + fathmn[6:8]]
                ans.extend(f)
        F.close()
```

```python
    def fathmn(self):
        path = self.file_path.text()
        path = path.replace('\\','/')
        path = path + '.xlsx'
        FATHMM_web.catchData(path)

    def msc(self):
        path = self.file_path.text()
        path = path.replace('\\','/')
        path = path + '.xlsx'
        MSC_web.catchData(path)

    def provean(self):
        path = self.file_path.text()
        path = path.replace('\\','/')
        path = path + '.xlsx'
        PROVEAN_web.catchData(path)

    def mutation(self):
        path = self.file_path.text()
        path = path.replace('\\','/')
        path = path + '.xlsx'
        Mutation_Assessor_web.catchData(path)
```

圖十八 製作各網站及操作介面程式碼

(五) 以病人腎臟基因變異案例作測試

將有 86 筆基因變異資料 excel 檔在 Upload File 處填入位址作為測試，且基因變異必須符合格式。

六、利用網路基因變異致病力預測結果訓練模型

為了嘗試解決基因預測資料不足的問題，我們將基因變異資料輸入多個網站取得致病力預測分數，為連續型資料，再以「管線」(Pipeline)及「多項式回歸」訓練模型。

(一) 資料來源

1. 腎臟基因(PKD1)變異資料(國家衛生研究院黃道揚教授提供)共 808 筆
2. 將 808 筆資料放入網路基因變異預測工具：CADD_phred、CONDEL、GERP_RS_rankscore、HVAR_rankscore、LRT_converted_rankscore、Polyphen2_HDIV_rankscore、REVEL、SIFT_converted_rankscore、VEST3_rankscore，每一筆基因變異資料分別得出 9 筆預測分數

(二) 資料處理

1. 每一筆預測分數調整至 1-100 的值，並以 100 為致病力最高，再將 9 個分數平均，為模型輸出值
2. 以基因變異位置、基因變異種類為模型輸入值

(三) 利用「管線」(Pipeline)訓練預測模型

1. 將輸入資料的 80%作為訓練資料，20%作為測試資料
2. 利用「管線」(Pipeline)將 StandardScaler、PCA、LogisticRegression 等物件串聯起來
3. 輸出模型準確率

(四) 利用「多項式回歸」訓練預測模型

1. 將基因變異位置和致病力分數塑模，使用線性、二階和三階多項式做比較
2. 分別輸出線性、二階、三階 R^2 值
3. 輸出線性、二階、三階曲線和訓練樣本關係圖

伍、研究結果

一、預測腎臟基因(PKD1)變異致病力模型訓練結果

(一) 管線(Pipeline)模型訓練結果

1. 外顯子(exon)模型

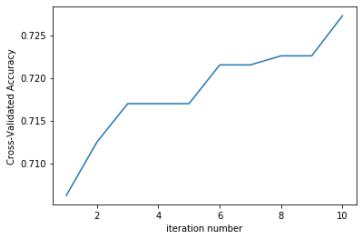
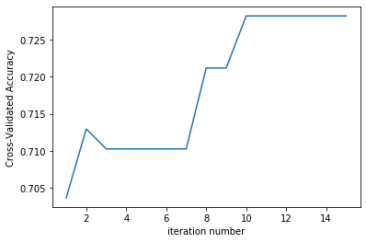
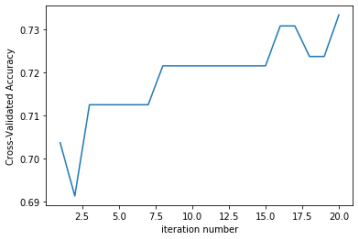
(1) 利用「管線」(Pipeline)訓練模型，80%訓練數據集，20%測試數據集

準確率：0.719

(2) 利用 k 折交叉驗證法(k-fold cross validation)評估模型效能

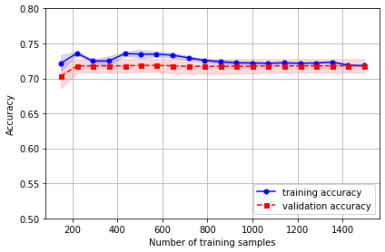
由下表可知，不同 k 值準確率差異甚小，但 k=10 時標準差最小。

表八 不同 k 值下之準確率及標準差

k 值	10	15	20
各 k 值驗證準確率			
每折平均準確率	0.719	0.719	0.719
標準差	0.006	0.009	0.009

(3) 運用「學習曲線」診斷高偏誤或高方差

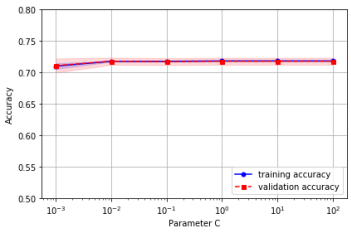
由訓練及驗證數據集準確率和樣本數關係曲線可知，在不到 200 個訓練樣本時，準確度較低，可能有「低度適合」的傾向；400-600 個樣本時，則是訓練數據集和驗證數據集差異較大，可能有「過度擬合」的傾向；600 個樣本以上則接近較佳的模型



圖十九 訓練及驗證數據集準確率和樣本數關係曲線

(4) 運用「驗證曲線」討論低度適合或過度擬合

由訓練及驗證數據集準確率和 LogisticRegression 參數 C 關係曲線可知，訓練數據集和驗證數據集在任何情況下皆無準確率偏低或差異過大的現象，因此應無「過度擬合」或「低度適合」。且參數 C 在 0.001 至 100 沒有太大差

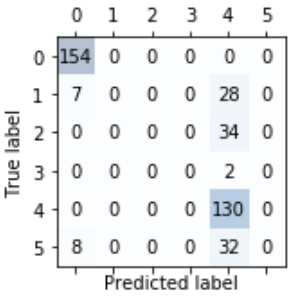


圖二十 訓練及驗證數據集準確率和 LogisticRegression 參數 C 關係曲線

異，但 0.01、0.1、1、10、100 仍有較佳的準確率，因此此五種參數 C 應能產生較佳的學習模型。

(5) 讀取混淆矩陣(confusion matrics)及宏觀微觀 F1 值

由混淆矩陣及不同類別下 F1 值及宏觀微觀 F1 值可知，類別 0(絕對致病)及類別 4(可能不影響)預測正確較其他類別高出許多，且類別 1(高度致病)、2(未確定)、5(可能致病)容易被預測為類別 4(可能不影響)。而宏觀 F1 值平均各項 F1 值較低，微觀 F1 值則較高。



圖二十一 混淆矩陣

表九 不同類別下 F1 值及宏觀微觀 F1 值

類別	0	1	2	3	4	5
致病程度	絕對致病	高度致病	未確定	可能亞效型	可能不影響	可能致病
F1 值	0.9535604	0.	0.	0.	0.73033708	0.
宏觀 F1 值	0.2806495750281189					
微觀 F1 值	0.7189873417721518					

2. 內含子(intron)模型

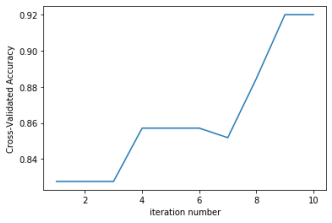
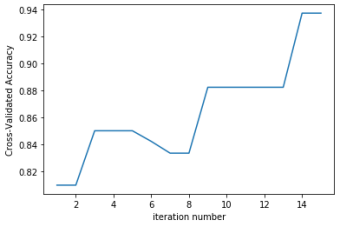
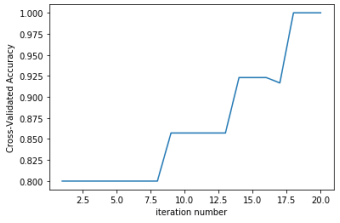
(1) 利用「管線」(Pipeline)訓練模型，80%訓練數據集，20%測試數據集

準確率：0.870

(2) 利用 k 折交叉驗證法(k-fold cross validation)評估模型效能

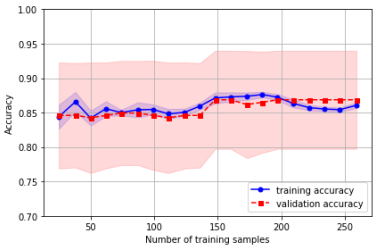
由下表可知，各 k 值準確率差異不大，然 20 折時準確率較高，但標準差最大。

表十 不同 k 值下之準確率及標準差

k 值	10	15	20
各 k 值驗證準確率			
每折平均準確率	0.863	0.864	0.869
標準差	0.033	0.038	0.071

(3) 運用「學習曲線」診斷高偏誤或高方差

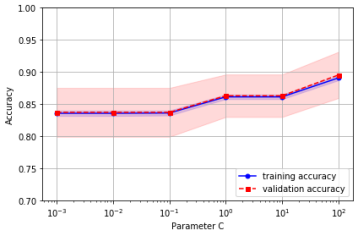
由訓練及驗證數據集準確率和樣本數關係曲線可知，在不到 150 個訓練樣本時，準確度較低，可能有「低度適合」的傾向；但訓練數據集和驗證數據集在各訓練樣本數下皆很接近，應較無「高度擬合」現象。值得一提的是，內含子模型驗證數據集樣本標準差(紅色透明區塊)明顯較外顯子模型大許多。



圖二十二 訓練及驗證數據集準確率和樣本數關係曲線

(4) 運用「驗證曲線」討論低度適合或過度擬合

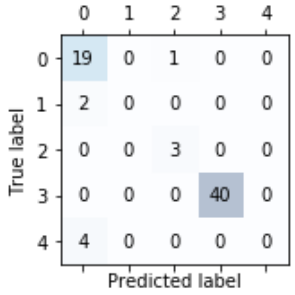
由訓練及驗證數據集準確率和 LogisticRegression 參數 C 關係曲線可知，參數 C 為 0.001、0.01、0.1 時，準確率較低，可能有「低度適合」傾向。且參數 C 為 100 時模型有最高的準確率。值得一提的是，內含子模型驗證數據集樣本標準差(紅色透明區塊)同樣明顯較外顯子模型大許多。



圖二十三 訓練及驗證數據集準確率和 LogisticRegression 參數 C 關係曲線

(5) 讀取混淆矩陣(confusion matrics)及宏觀微觀 F1 值

由混淆矩陣及不同類別下 F1 值及宏觀微觀 F1 值可知，類別 0(絕對致病)及類別 2(未確定)及類別 3(可能亞效型)預測準確率較高。而宏觀 F1 值直接將各類 F1 值平均較低；微觀 F1 值則較高。



圖二十四 混淆矩陣

表十一 不同類別下 F1 值及宏觀微觀 F1 值

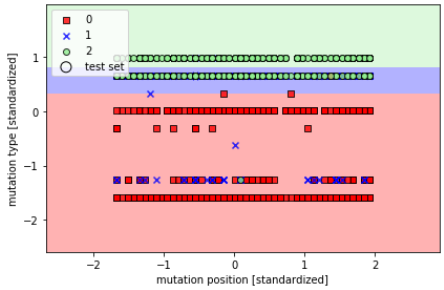
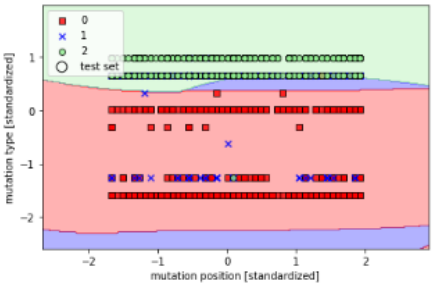
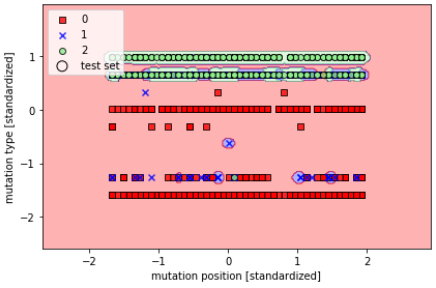
類別	0	1	2	3	4
致病程度	絕對致病	高度致病	未確定	可能不影響	可能致病
F1 值	0.8444444	0.	0.8571	1.	0.
宏觀 F1 值	0.5403174603174603				
微觀 F1 值	0.8985507246376812				

(二) 利用 SVM(核支援向量機)訓練分類器結果

1. 外顯子模型

三種分類輸出「決策區域」結果：由表可知，類別 0(紅色)及類別 2(綠色)有較佳的決策結果，比較不同模型則可知，「核支援向量機」中使用參數 γ =100 有較「顛簸」的邊界，能夠較準確的分類不同樣本。

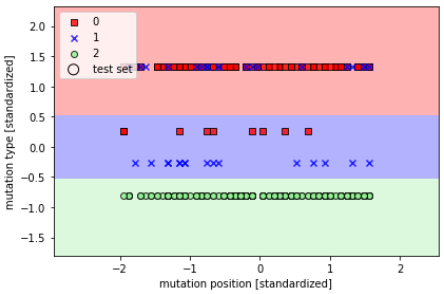
表十二 三種模型輸出結果

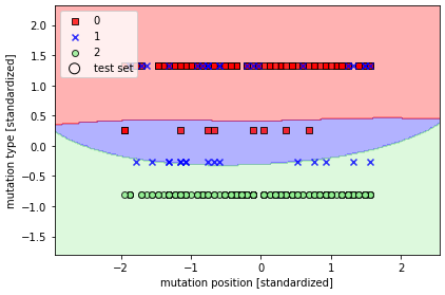
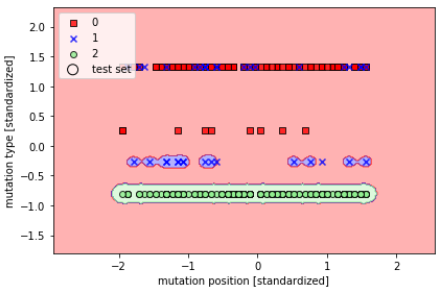
支援向量機(SVM)	
	
核支援向量機(kernel SVM)	
γ =0.2	γ =100
	

2. 內含子模型

三種分類輸出「決策區域」結果：由表可知，類別 0(紅色)及類別 2(綠色)仍有較佳的決策結果，然而，類別 1(藍色)相較外顯子模型有較佳的決策能力。比較不同模型則可知，「核支援向量機」中使用參數 γ =100 有較「顛簸」的邊界，能夠較準確的分類不同樣本。

表十三 三種模型輸出結果

支援向量機(SVM)


核支援向量機(kernel SVM)	
γ =0.2	γ =100
	

(三) 利用隨機森林(Random Forest)訓練模型結果

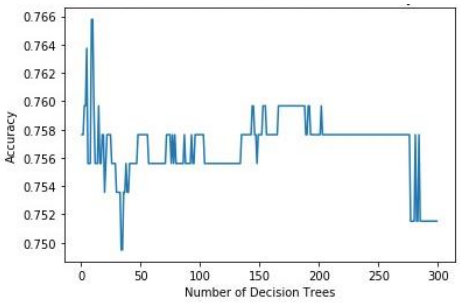
1. 外顯子模型

(1) 利用「隨機森林」(Random Forest)訓練模型

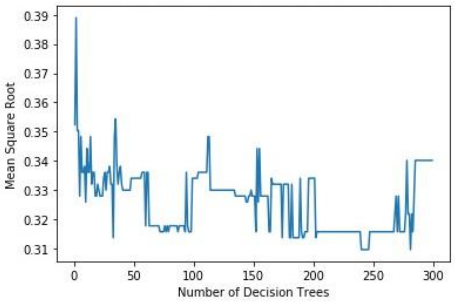
準確率：0.765

(2) 改變「隨機森林」中的「決策樹分類器」數量

由下圖二十六結果可知，在分類器數量介於 0 至 20 之間，較容易有較高的準確率，其中又以分類器數量為 8 或 9 時準確率最高，而分類器超過 50 個後，分類器數量對準確率不造成太顯著的改變。由下圖二十七可知，其均方誤差於分類器數量為 60 至 100 時和 200 至 270 時最低。



圖二十五 準確度和決策樹分類器數量關係



圖二十六 均方誤差和決策樹分類器數量關係

(3) 改變不純度計算公式

本研究以「Gini 不純度(Gini Impurity)」及「熵(entropy)」訓練模型，但對準確度並未造成顯著影響。

(4) 利用裝袋法(Bagging)訓練模型

準確率：0.756

(5) 利用 AdaBoost 法訓練模型

準確率：0.756

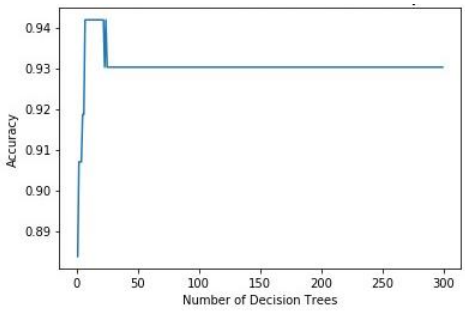
2. 內含子模型

(1) 利用「隨機森林」(Random Forest)訓練模型

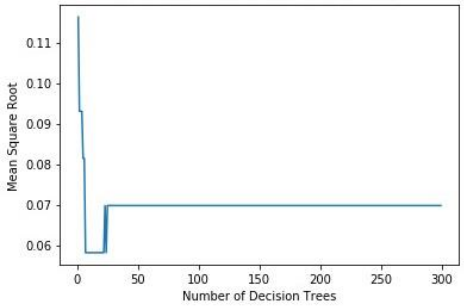
準確率：0.942

(2) 改變「隨機森林」中的「決策樹分類器」數量

本研究以「決策樹分類器」數量為 1 至 300 訓練模型，由下圖二十八結果可知，在分類器數量介於 6 至 20 之間，具有較高的準確率，而分類器超過 50 個後，分類器數量對準確率不造成太顯著的改變，可能是資料數量過少所致。由下圖二十九可知，其均方誤差同樣於分類器數量為 6 至 20 時最低。



圖二十七 準確度和決策樹分類器數量關係



圖二十八 均方誤差和決策樹分類器數量關係

(3) 改變不純度計算公式

本研究以「Gini 不純度(Gini Impurity)」及「熵(entropy)」訓練模型，但對準確度並未造成顯著影響。

(4) 利用裝袋法(Bagging)訓練模型

準確率：0.930

(5) 利用 AdaBoost 法訓練模型

準確率：0.930

二、整合網路基因變異註釋工具結果

(一) 資料輸入介面

在 Upload File 處輸入符合格式的資料(需為 excel 檔)路徑，點選 MSC、FATHMM、POVEAN、Mutation Assessor 可輸出對應網站資料，點選 Export 則輸出所有網站整合資料。

圖二十九 資料輸入介面

(二) 點選 Export(所有網站結合)結果

開啟 result.csv 檔案，即可得所有網站較重要資料整合

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Fathmn_Coding Score	Fathmn_Coding Groups	CADD_Score	MSC-CADD_Impact	Pred_PolyPhen2_Score	hvar_prediction	SIFT_Score	SIFT_Pred	mutation_Func	mutation_FI sc	mutation_MSA	Provean_SCC	Provean_PREI	Provean_SCC	Provean_PREDIC
2	Coding Score	Coding Groups													
3	0.66809	AEGCI	13	high	0.048	benign	0.77	T	low	1.435	23	-1.31	Neutral	0.409	Tolerated
4	0.05707	AEFBI	4.737	high	0.001	benign	0.62	T	low	1.1	19	-0.4	Neutral	0.527	Tolerated
5	0.59807	AEFI	23	high	0.026	benign	0.11	T	low	0.895	14	-0.08	Neutral	0.045	Damaging
6	0.92455	AEFBI	24.6	high	0.956	probably damaging	0	D	medium	2.015	14	-5.94	Deleterious	0.003	Damaging
7	0.99386	AEFGBI	15.84	high	0.032	benign	0.15	T	low	1.445	34	-0.88	Neutral	0.169	Tolerated
8	0.94032	AEFGBIJ	34	high	0.94	probably damaging	0	D	medium	2.88	173	-5.42	Deleterious	0.001	Damaging
9	0.98823	AEFDGBI	15.94	high	0.269	benign	0.17	T	low	1.285	287	-1.59	Neutral	0.163	Tolerated

圖三十 所有網站輸出資料

三、利用網路基因變異致病力預測結果訓練模型結果

(一) 利用「管線」(Pipeline)訓練預測模型結果

準確率：0.753

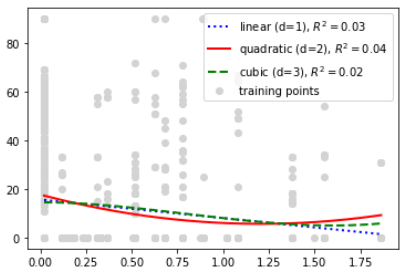
(二) 利用「多項式回歸」訓練預測模型結果

1. 線性、二階、三階回歸 R^2 值和回歸線與訓練樣本關係圖

線性 R^2：0.028834517079321498

二階 R^2：0.03872674208855176

三階 R^2：0.019149538396007193



圖三十一 回歸線與訓練樣本關係圖

2. 由以上結果可知基因變異位置和致病力預測分數無法以線性、二階、或三階表示

陸、討論

一、預測腎臟基因(PKD1)變異致病力模型訓練結果討論

(一) 管線(Pipeline)模型訓練結果討論

1. 準確度大約八成，各致病程度樣本數不一，導致各類準確度差異大

對於外顯子模型，因為「絕對致病」和「可能不影響」兩類別樣本數較其他類別樣本數多出許多，因此混淆矩陣中此二類別的「真陽」較高，而其他類別的預測準確度則明顯較低。對於內含子模型，雖然整體樣本數較少，但是各類樣本數較平均，因此整體預測準確度較高，但是樣本少的「高度致病」、「未確定」、「可能不影響」三個類別的預測準確度仍偏低。

除此之外，此資料庫複雜度高，外顯子模型的「變異位置」特徵即有 46 種(EX1~EX46)，內含子模型的「變異位置」特徵也有 45 種(IVS1~IVS45)，此原因也可能導致準確度偏低。

我們雖然使用最大的 PKD1 基因收集資料庫，但是準確度仍因為資料不夠且複雜度過高而偏低。因此，若要增加預測準確度，應以擴大基因資料庫為先。

2. 內含子模型驗證數據集準確率標準差較大

由內含子模型的學習曲線即驗證曲線可知，其標準差明顯較外顯子模型大許多，可推知是因為內含子基因變異數據較少，若再將其分為 10、15、20 等分以利用 k 折交叉驗證法將數據，進行訓練、驗證，每一等份樣本數將會更少，使標準差極大。因此，若要使標準差變小，增加樣本數應為首要方法。

3. LogisticRegression 最佳參數

對於外顯子模型，由驗證曲線可知，參數 C=0.01、0.1、1、10、100 時無「低度適合」或「過度擬合」，此應為其最佳參數。對於內含子模型，由驗證曲線可知，參數 C=100 時無「低度適合」或「過度擬合」且準確度最高，此應為其最佳參數。

(二) 利用 SVM(支援向量機)訓練分類器結果討論

1. 針對類別 0、2 有較佳的決策分類能力及預測準確度

由外顯子與內含子模型所輸出的 LogisticRegression、支援向量機(SVM)、核支援向量機(kernel SVM)決策區域圖可知，類別 0 與 2 的決策區域較能有效區分不同樣本，而類別 1 決策區域則較容易與另二類別重疊，使用線性與非

線性訓練方式皆有相同問題，此應與類別 1 樣本較不突出因此容易混淆有關，其中，內含子模型表現較佳。預測準確度也同樣針對類別 0、2 較佳。

2. 核支援向量機(kernel SVM)γ越大，能將不同數據分類的越好

由外顯子及內含子模型輸出的核支援向量機(kernel SVM)可知，γ=100 時能判斷出極佳的決策分類區域，然而，當分類器在處理未見過的數據時，「一般化誤差」也越高，有可能造成錯誤判斷。

(三) 利用隨機森林(Random Forest)、裝袋法、AdaBoost 方法訓練模型結果討論

1. 模型訓練準確度及決策樹分類器的最佳數量

外顯子模型訓練準確度為隨機森林訓練準確度為 0.765；裝袋法為 0.756；AdaBoost 方法為 0.756，隨機森林分類器數量介於 0 至 20 之間，較容易有較高的準確率，其中又以分類器數量為 8 或 9 時準確率最高。而內含子模型為隨機森林訓練準確度為 0.942；裝袋法為 0.930；AdaBoost 方法為 0.930，隨機森林在分類器數量介於 6 至 20 之間，具有較高的準確率。決策樹分類器越多應有較高的準確率，然而，本研究使用樣本維度較低，增加決策樹分類器時，可能造成過度擬合，當各決策樹判斷結果錯誤較多時，再輸出所有決策樹平均值時即導致較低的準確度。

2. 三種分類器訓練結果比較

由輸出準確度可知，隨機森林(Random Forest)在外顯子及內含子訓練皆有較高準確率，可以推測此訓練資料組成較為單純，錯誤分類資料較少，使用裝袋法及 AdaBoost 方法並未提高準確度。

(四) 三種不同模型訓練方式比較討論

表十四 三種不同模型訓練方式比較討論

模型種類 / 內容	Pipeline	SVM 支援向量機	隨機森林、裝袋法、AdaBoost 法
準確率	(外)0.719、(內)0.870 整體最低	(外)0.756、(內)0.922	(外)0.765、(內)0.942 整體最高
分析	若單筆資料中內容較多，擬合使用管線(Pipeline)模型預測基因變異疾病，因為其包含降維、特徵縮放、再擬合出模型，若資料充足，能較高效產出結果，並藉由調整參數 C 能夠更有效避免過度擬合的問題。	若資料量較少，應使用 scikit-learn 中的 SVM 或隨機森林等訓練模式，因為此二方法為類似「分類器」的訓練方式，適合本研究所使用的訓練資料。而前者能夠針對資料維度輸出線性或非線性的決策邊界，後者則能有效的處理缺失值，即使數據少量也能有高精確度，然而，若資料量過大，可能有過度擬合以及運作時間過長的問題。	

二、整合網路基因變異註釋工具

(一) 能夠大幅減少醫生預測病人基因變異致病性的時間，具有極高的實用性

醫生使用網路工具進行疾病預測時，需要利用大量的時間一筆一筆剪貼病人資料至網路基因註釋工具進行預測，需要 4 個小時以上才能夠完成繁雜的預測步驟，因此，整合網路基因變異註釋工具、製作單一介面，僅需不到 10 分鐘即能輸出結果，能有效解決醫生耗費過多時間預測基因變異，大幅提升註釋工具的實用性。

(二) 符合第一線醫界人員或研究人員需求，幫助增加網路基因變異註釋工具的使用者

基因變異預測疾病工具是第一線醫界人員以及研究人員進行臨床上的參考或研究相關領域的重要助手，然而，目前醫院中較無此類整合網路基因預測工具的開發人員及工具，可能使醫界人員因為推廣不足而未使用到這些新興領域的實用工具，若整合這些工具，或許能夠增加此類工具的使用者。

三、利用網路基因變異致病力預測結果訓練模型結果討論

(一) 使用連續型數據作為預測，然基因變異資料複雜，資料量仍然不足

本研究嘗試提高模型訓練的準確率，因此蒐集網路工具所輸出的致病力預測結果，然而因為基因位置繁多，基因變異資料複雜，導致資料量仍然不足。由「多項式回歸」也可知，基因變異位置和致病力分數並非線性、二階、或三階關係，因其基因變異位置複雜，因此應為更高階關係。

四、未來展望

(一) 藉由整合網路基因變異註釋工具，蒐集大量病人資料，進而訓練深度學習模型，提高預測基因變異造成致病力的準確度

「深度學習」需要極為龐大的資料量才能訓練高準確度的模型。然而，生物醫學領域資料複雜性、多樣性極高，難以在各個類別基因變異型態皆有大量的數據進行模型訓練，也造成本研究預測腎臟基因模型準確度問題。雖然大型國外網站(Clinvar、HGMD 等)有由不同醫生上傳的各式基因突變資料，但若只看單一基因變異種類(由基因種類、變異位置、採樣人種等分類方法)的資料量，仍不足以進行機器學習。因此，若能藉由整合網路基因變異註釋工具，蒐集大量病人資料，以得更多更深入更豐富的資料，或許能夠提高預測基因變異致病力模型的準確度。

(二) 在蒐集醫生上傳資料的同時，為臺灣建立基因資料庫

每一筆病人的資料都十分珍貴，都可作為未來研究的樣本，然而，目前臺灣並沒有一個較大的資料庫提供醫生蒐集病歷資料，因此，我們期望可以藉由「整合網路基因註釋工具」，在醫生上傳基因變異資料的同時，蒐集並公開這些資料，為臺灣建立較完整的基因變異資料庫。

柒、結論

一、模型訓練、網路工具整合及簡易基因註釋器結果

訓練腎臟基因(PKD1)變異致病力模型方面，利用管線(Pipeline)模型訓練得準確度為八成左右，因為公開基因變異資料量不足且不平均，導致部分準確度仍提升空間，而外顯子模型之最佳參數 C=0.01、0.1、1、10、100 時最佳，內含子模型則是 C=100。利用 scikit-learn 訓練三種不同決策分類器，得核支援向量機(kernel SVM) γ 越大，擁有最佳的決策邊界。利用隨機森林(Random Forest)訓練模型，得外顯子模型決策樹數量在 8~9 個有最高準確度 0.765，內含子模型在決策樹數量 6~20 個有最高準確度 0.942。而資料較多擬合使用管線(Pipeline)訓練模型，資料較少則擬合使用 scikit-learn 中的訓練方法及隨機森林(Random Forest)。

接著，為了提高網路基因變異註釋工具的可用性，製作單一介面整合各網站，能夠大幅減少醫生預測病人基因變異致病性的時間，具有極高的實用性。

利用網路基因註釋工具所輸出致病力分數訓練模型，期望解決資料不足的問題，然而基因資料太過複雜，仍需更大量資料以達高準確率模型。

二、資料不足的情況下最佳的基因預測模式

為了解決公開基因變異資料量的不足，我們整合網路基因註釋工具，增加網路工具的實用性，再利用網路公開基因變異資料訓練模型，並且期望藉由此網路整合工具蒐集大量病人資料，以提升在資料不足的情況下最佳的基因預測準確度。

捌、參考資料

1. Sebastian Raschka、Vahid Mirjalili(著)，劉立民、吳建華(譯)・2018・Python 機器學習－使用 Python 的 scikit-learn 和 Tensorflow 進行機器學習和深度學習(二版)・(第六章)學習模型評估和超參數調校的最佳實作・新北市：博碩文化

2. 網路爬蟲整合網路基因預測工具

(1) The Mutation Significance Cutoff (MSC) Server・The Rockefeller University・St. Giles Laboratory of Human Genetics of Infectious Disease・取自：http://pec630.rockefeller.edu:8080/MSC/

(2) Predict the Functional Consequences of Non-Coding and Coding・Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C (2014)・取自：http://fathmm.biocompute.org.uk/fathmmMKL.htm

(3) PROVEAN HUMAN GENOME VARIANTS・National Institutes of Health・取自：http://provean.jcvi.org/genome_submit_2.php?species=human

(4) mutationassessor.org-functional impact of protein mutations・The Cancer Genome Atlas・取自：http://mutationassessor.org/r3/

3. Autosomal Dominant Polycystic Kidney Disease: Mutation Database・PKD FOUNDATION・取自：https://pkdb.mayo.edu/index.html

4. Evaluating the quality of a germline short variant callset・GATK・取自：https://gatk.broadinstitute.org/hc/en-us

5. 圖片來源

(1) 圖二：PKD1 gene・Chromosomal Location・取自：https://ghr.nlm.nih.gov/gene

(2) 圖四：Data Preprocessing Steps for Deep Learning in Python・Feature Scaling・取自：https://www.deeplearning-academy.com/p、機器學習-特徵工程-降維・取自：http://www.taroballz.com/2018/07/06/ML_DecreaseFeature/

(3) 圖五、表三、圖六：Python 機器學習－使用 Python 的 scikit-learn 和 Tensorflow 進行機器學習和深度學習(二版)・(第六章)學習模型評估和超參數調校的最佳實作

(4) 表四：隨機森林（Random Forest）・取自：https://medium.com/marketingdatascience/ 、裝袋法(Bagging)・取自：https://www.researchgate.net/figure/ 、AdaBoost 法・取自：https://zhuanlan.zhihu.com/p/39972832

【評語】052506

本作品有兩大部分，第一部分探討不同機器學習訓練模型對於腎臟基因變異致病之預測，是不錯的前期機器學習應用研究。第二部分則製作單一介面方便醫生輸入資料用以預測基因變異致病性的時間，這部分屬於工具開發，雖具有實用性但欠缺科學實驗精神，資訊科學探究較為不足。

作品海報

研究動機

以人工智慧訓練基因變異預測疾病模型有數據不足的問題。

未來，「精準醫療」只要檢測一下DNA，立刻就能對症下藥，但這美好的前景，必須結合眾多分子生物學及人工智慧的研究與合作才能完成。當獲得病人定序後的基因時，如何針對此龐大且複雜的資料進行比對分析，進而應用於臨床或研究上，診斷病人的疾病？大數據整合、機器學習、深度學習及人工智慧等新潮AI科技可謂打進「精準醫療」的核心。

對模型進行深度學習需要龐大的資料，如何在數據不足的情況下訓練模型，從變異的基因預測病人的疾病？

因此，本研究致力於尋找在資料不足的情況下最佳的基因預測模式。希望利用「深度學習」技術，藉由網路公開病例資料訓練模型以預測案例的致病程度；為了取得更完整的基因變異預測結果，以「網路爬蟲」整合各網站基因預測技術；並結合蛋白質變化及真實案例之致病分數，以期為「精準醫療」領域貢獻心力。

研究目的

一、利用不同模型訓練預測腎臟基因(PKD1)變異致病力模型。

二、整合網路基因變異註釋工具，提升重要且大型的網路工具可用性。

三、嘗試解決訓練模型資料不足的問題，利用網路工具致病力分數訓練模型。

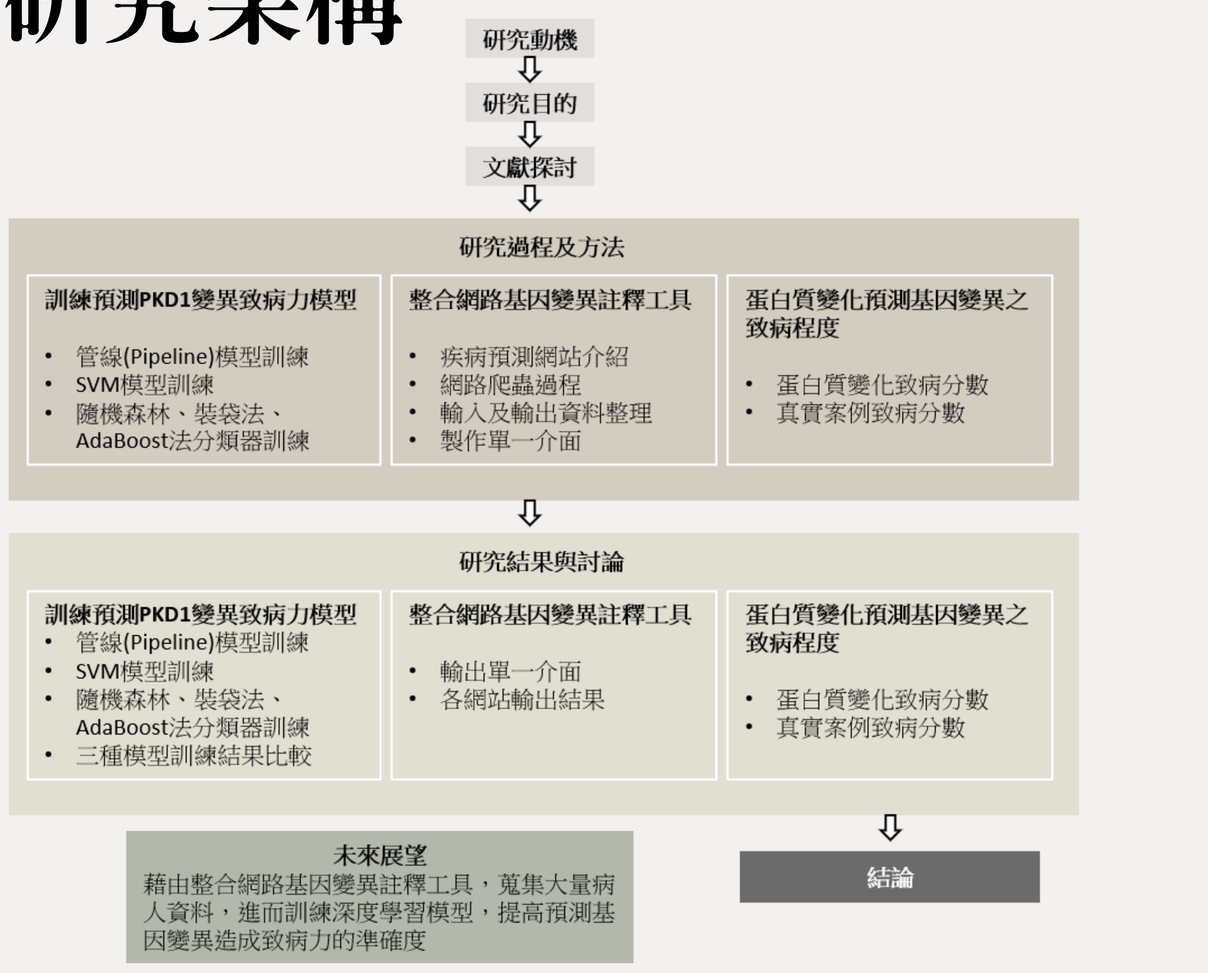
四、由研究結果提出在資料不足情況下最佳的基因預測模式，並期望取得龐大數據量。

研究過程與結果

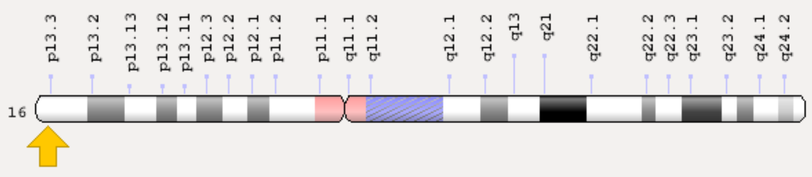
基因變異分析流程

濕實驗室(Wet Lab)　　乾實驗室(Dry Lab)

研究架構



選擇突變基因：PKD1



PKD1突變與大多數染色體顯性多囊腎疾病有關，多囊腎是一種遺傳疾病，由於基因的缺陷，使得患者的腎臟會出現大小不等的水泡；多囊腎為體染色體遺傳，故男性以及女性患病的機率是一樣的，且遺傳機率是百分之五十。

訓練預測腎臟基因變異致病力模型

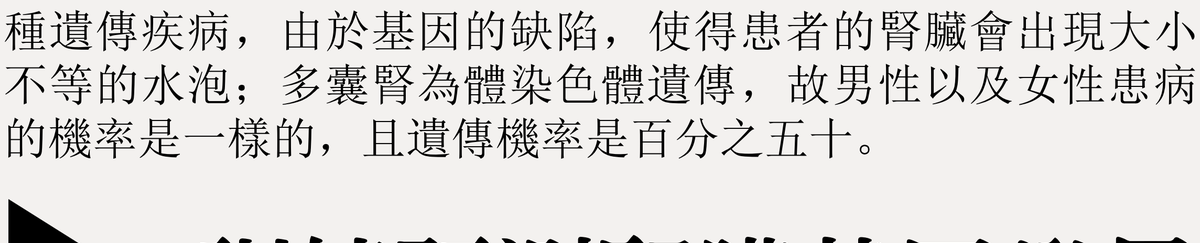
(一)研究過程

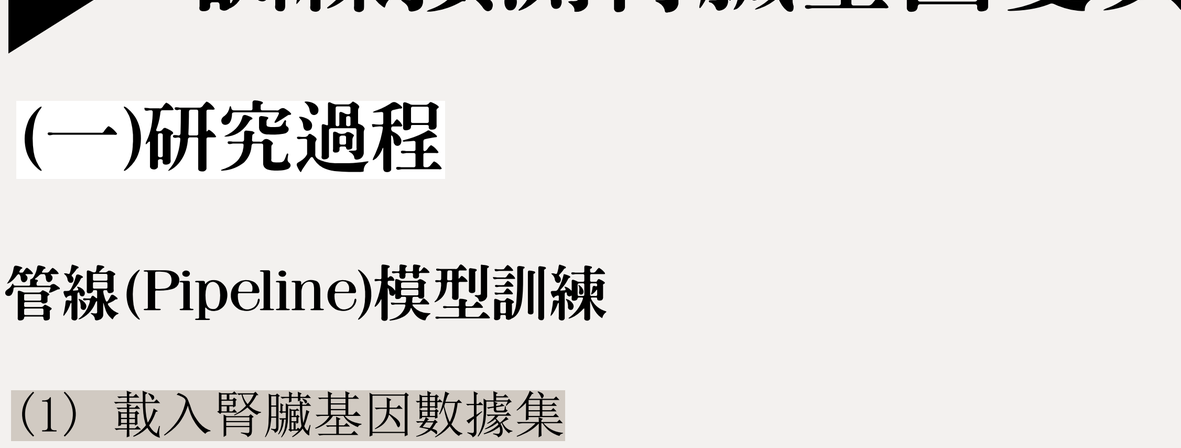
管線(Pipeline)模型訓練

(1) 載入腎臟基因數據集

輸出結果：6種致病程度指派給y陣列，並編碼成0-5。再將全數據拆成「訓練數據集」(80%)及「測試數據」(20%)。

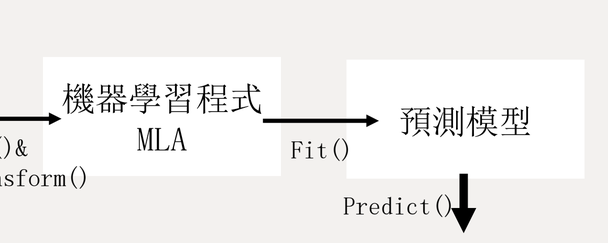
(2) 利用「管線」(Pipeline)將StandardScaler、PCA、LogisticRegression等物件串聯起來



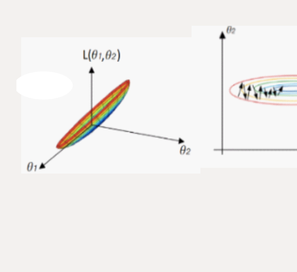
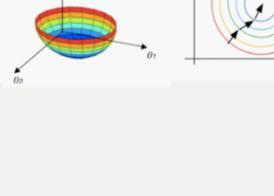
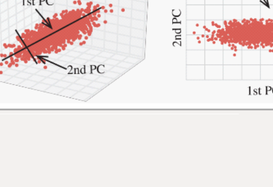


(3) 利用k折交叉驗證法(k-fold cross validation)評估模型效能

每折數據中類別大小比率和原數據是相同的，並用k=10、15、20來訓練數據集。



(4) 運用「學習曲線」診斷高偏誤或高方差

效能佳模型	高偏誤模型	高方差模型
	「訓練正確率」與「驗證正確率」都很低，表示它「低度適合」(underfitting)	「訓練正確率」和「驗證正確率」相差太多，表示它「過度擬合」(overfitting)
		

1. 資料來源：PKD FOUNDATION Autosomal Dominant Polycystic Kidney Disease: Mutation Database，取PKD1基因(2332筆)訓練模型。
2. 輸入資料格式：

類別	ID	變異位置	變異種類	變異型態	致病程度(輸出結果)
舉例	EX1-EX39del	1	6	1	A

(5) 運用「驗證曲線」討論低度適合或過度擬合

找出C=0.001、0.01、0.1、1.0、10.0、100.0時，何者擁有最佳的模型，C越小，有越大的正則化。

$$J(\theta)_{L2} = C \times J(\theta) + \sqrt{\sum_{j=1}^{n}(\theta_j)^2 (j \geq 1)}$$

(參數 θ (w)向量中的每個參數的平方和的開方值)

(6) 讀取混淆矩陣(confusion matrics)

$$\textbf{Precision} = \frac{TP}{TP+FP}, \textbf{Recall} = \frac{TP}{TP+FN}, \textbf{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

(7) 用外顯子和內含子資料由上述步驟各自訓練，得兩個預測基因致病力模型

SVM(核支援向量機)訓練分類器

(1) 資料處理

將基因資料中致病程度的編號A及類別B合併，將編號C和編號F合併，以及編號E，共分為三個類別。將30%歸類為測試數據集，70%為訓練數據集。

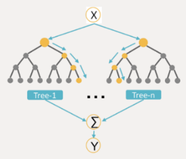
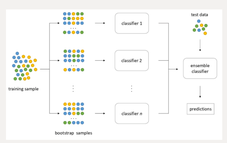
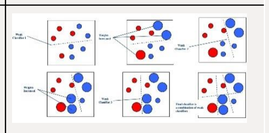
(2) 支援向量機(SVM)分類器

調整參數c的值，尋找最大化邊界，以獲取最佳化模型。參數kernel='linear'。

(3) 核支援向量機(kernel SVM)分類器

利用「核技巧」(kernel trick)在高為空間中找到分離超平面，參數kernel='rbf'。調整參數γ值，藉由增加γ值能夠產生較「顛簸」的邊界，本研究輸出γ=0.2及100的決策邊界。

隨機森林、裝袋法、AdaBoost法

隨機森林 (Random Forest)	裝袋法 (Bagging)	AdaBoost法
決策樹分類器是藉由特徵值將數據分割到不同群組後，分別計算其資訊增益 (Information gain, IG)，尋找出準確率最高的分割方式。	從初始訓練集中抽取「自助式樣本」(「放回式」的隨機樣本)，再進行投票，以「多數決來做預測。	以「不放回式隨機抽樣」來做預測，並且讓學習器從「誤判訓練樣本」學習。逐次降低錯誤樣本的權重；最後再以多數決預測。
		

(1) 資料處理

將基因資料中致病程度的編號A及類別B合併，將編號C和編號F合併，以及編號E，共分為三個類別。80%訓練數據集，20%測試數據集。

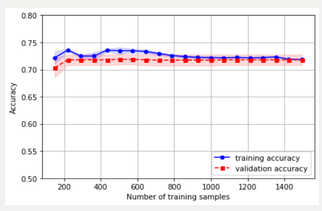
(2) 隨機森林

分類器數量為1至300進行測試。並改變不純度計算公式，利用「Gini不純度(Gini Impurity)」及「熵(entropy)」訓練模型。

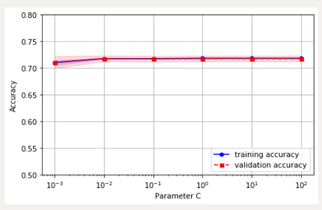
(二)研究結果

管線(Pipeline)模型訓練

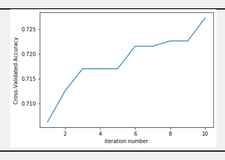
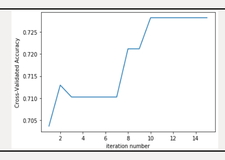
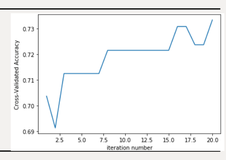
外顯子(exon)模型　準確率：0.719

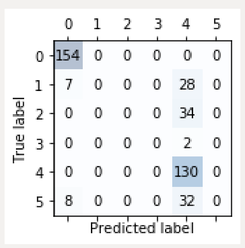


「學習曲線」
不到200個訓練樣本有低度適合的傾向；400-600個樣本時，有過度適合的傾向；600個樣本以上則接近較佳的模型

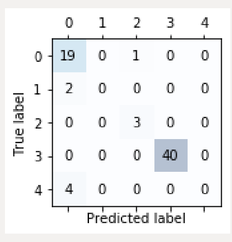


「驗證曲線」
0.01、0.1、1、10、100有較佳的準確率，因此此五種參數C應能產生較佳的學習模型

k值	10	15	20
各k值驗證準確率			
準確率	0.719	0.719	0.719
標準差	0.006	0.009	0.009

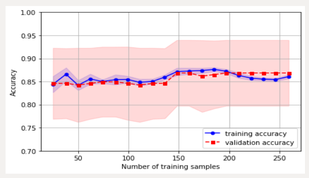


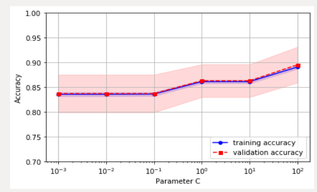
外顯子
混淆矩陣



內含子
混淆矩陣

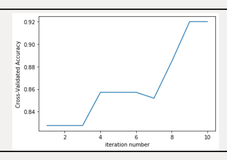
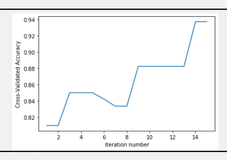
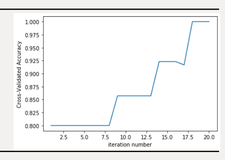
內含子(intron)模型　準確率：0.870





「學習曲線」
不到150個訓練樣本有低度適合傾向；但應較無「高度適合」現象。標準差(紅色透明區塊)明顯較外顯子模型大許多

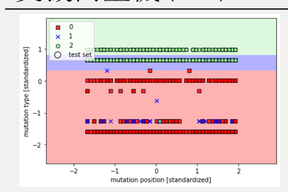
「驗證曲線」
參數C為0.001、0.01、0.1時有低度適合傾向。且參數C為100有最高的準確率，標準差同樣較大

k值	10	15	20
各k值驗證準確率			
準確率	0.863	0.864	0.869
標準差	0.033	0.038	0.071

SVM(核支援向量機)訓練分類器

外顯子(exon)模型

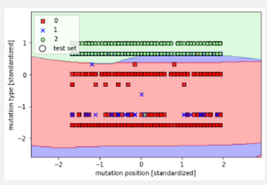
支援向量機(SVM)



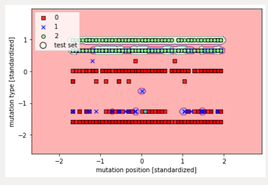
類別0(紅色)及類別2(綠色)有較佳的決策結果。「核支援向量機」中使用參數γ=100有較「顛簸」的邊界，能夠較準確的分類不同樣本。

核支援向量機(kernel SVM)

γ=0.2

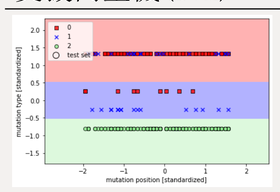


γ=100



內含子(intron)模型

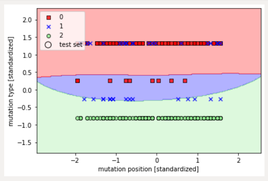
支援向量機(SVM)



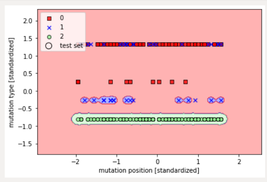
類別0(紅色)及類別2(綠色)仍有較佳的決策結果，但類別1(藍色)相較外顯子模型有較佳的決策能力。

核支援向量機(kernel SVM)

γ=0.2



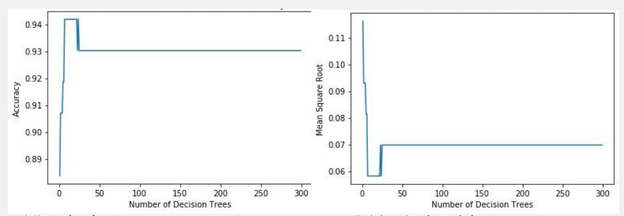
γ=100



「核支援向量機」中使用參數γ=100有較「顛簸」的邊界，能夠較準確的分類不同樣本

隨機森林、裝袋法、AdaBoost法

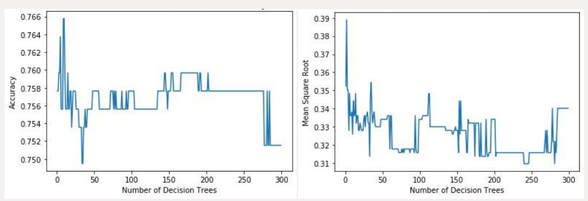
隨機森林外顯子(exon)模型



「準確度」　「均方誤差」

在分類器數量介於0至20之間，較容易有較高的準確率，又為8或9時準確率最高，超過50個後，分類器數量對準確率不造成太大的改變。

隨機森林內含子(intron)模型



「準確度」　「均方誤差」

分類器數量介於6至20之間，具有較高的準確率，超過50個後，分類器數量對準確率不造成太顯著的改變，可能是資料數量過少所致。

	隨機森林	裝袋法	AdaBoost法
外顯子	0.765	0.756	0.756
內含子	0.942	0.930	0.930

蛋白質變化預測基因變異之致病程度

資料來源：
1. NC_000016.10 2135898-2088708共47191個鹼基：NCBI Homo sapiens chromosome 16, GRCh38.p13 Primary Assembly
2. 4319個胺基酸：NCBI Homo sapiens chromosome 16, GRCh38.p13 Primary Assembly CDS
3. 449項真實變異案例：ClinVar PKD1[gene]

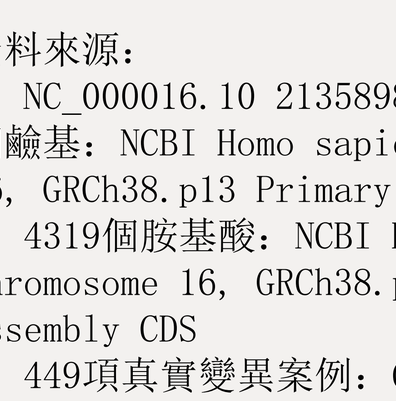
(1) 研究過程

1. 輸入47191項所對應鹼基，並擷取資料中確定能轉變為mRNA之範圍

0	1	2	3	4	5	6	7	……	47188	47189	47190
G	C	A	C	T	G	C	A		C	A	A

2. 使用者input變異位置(2088708-2135898)及變異後鹼基(ATCG)

3. 由蛋白質變化判斷



依照鹼基→胺基酸變異規則，判斷變異後生成之胺基酸

無胺基酸變化：score 1=0
形成終止密碼子：score 1=1
產生胺基酸變化(非終止)：score 1=0.9

4. 真實案例判斷

由ClinVar網站共449筆真實案例資料產生。

Conflicting interpretations of pathogenicity			0.5
Benign	0	Benign/Likely benign	0.01
Pathogenic	1	Pathogenic/Likely pathogenic	0.9
Likely benign	0.05	Likely pathogenic	0.8

整合網路基因變異註釋工具

(1) 輸入資料

第一欄	第二欄	第三欄	第四欄	第五欄	第六欄	第七欄
ID	Chormosome	Region	Reference	Allele	Type	Homo_sapiens_refseq

(2) 輸出資料

網站	索取資料		
MSC	CADD_Score	MSC-CADD_Score	PolyPhen2_Score
	30分以上表示「可能有害」，30分以下表示「可能有益」，分數越高表示越可能有害。	針對特定基因，預期最低臨床CADD臨界值	>0.908 較可能有害 0.446~0.908 可能有害 <0.446 良性
	hvar_prediction	SIFT_Score	SIFT_Pred
	根據PolyPhen2_Score的各類有效範圍輸出probably damaging、 possibly damaging、或benign	<0.05 較無害的 >0.05 有害的	根據SIFT_Score的各類有效範圍輸出「D」Deleterious(有害)、「T」Tolerated(無害)
FATHMM	Fathmn_Coding Score	Fathmn_Coding Groups	
	>0.5deleterious(有害的) <0.5 Neutral or benign(不顯著/良性)	A~I表示不同預測方法，同一個基因可以使用多種方法	
PROVEAN	SCORE / PREDICTION(cutoff = -2.5) ≦-2.5 / Deleterious(有害的) >-2.5 / Neutral(不顯著)	SCORE / PREDICTION(cutoff = 0.05) ≦0.05 / Deleterious(有害的) >0.05 / Neutral(不顯著)	
MUTATION ASSESSOR	FI Score(Func.Impact) F1≦ 0.8(neutral；0.8<FI≦1.9(low impact)；1.9<FI≦3.5(medium impact)；FI>3.5(high impact)		

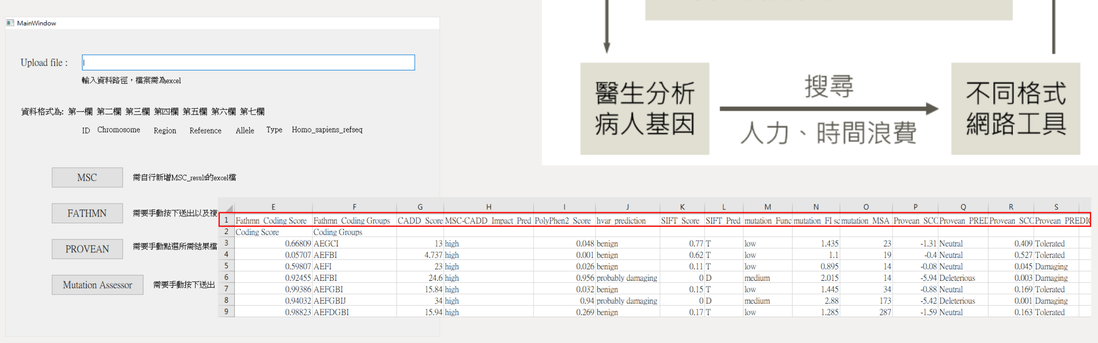
(3) 製作操作介面

製作上傳檔案方框，使操作者輸入檔案路徑以讀取excel檔，並製作四個網站及EXPORT按鈕，讓使用者可以選取各網站或全部資料。

(4) 以病人腎臟基因變異案例作測試

將有86筆基因變異資料excel檔在Upload File處填入位址作為測試，且基因變異必須符合格式。

(5) 研究結果



討論

預測腎臟基因(PKD1)變異致病力模型訓練討論

(一) 管線(Pipeline)模型訓練結果討論

1. 準確度大約八成，各致病程度樣本數不一，準確度差異大

外顯子模型「絕對致病」和「可能不影響」樣本數較其他類別樣本數多出許多，此二類別的「真陽」較高，而其他類別的預測準確度則明顯較低。對於內含子模型，整體預測準確度較高，但是樣本少的「高度致病」、「未確定」、「可能不影響」三個類別的預測準確度仍偏低。除此之外，此資料庫複雜度高也可能導致準確度偏低。因此，若要增加預測準確度，應以擴大基因資料庫為先。

2. 內含子模型驗證數據集準確率標準差較大

因為內含子基因變異數據較少，若再將其分為10、15、20等分以利用k折交叉驗證法將數據，進行訓練、驗證，每一等份樣本數將會更少，使標準差極大。因此，若要使標準差變小，增加樣本數應為首要方法。

3. LogisticRegression最佳參數

外顯子模型，參數C=0.01、0.1、1、10、100時無「低度適合」或「過度適合」，此應為其最佳參數。內含子模型，參數C=100時無「低度適合」或「過度適合」且準確度最高，此應為其最佳參數。

(二) 利用SVM(支援向量機)訓練分類器結果討論

1. 針對類別0、2有較佳的決策分類能力及預測準確度

由外顯子與內含子模型所輸出的LogisticRegression、支援向量機(SVM)、核支援向量機(kernel SVM)決策區域圖可知，類別0與2的決策區域較能有效區分不同樣本，使用線性與非線性訓練方式皆有相同問題，此應與類別1樣本較不突出因此容易混淆有關，其中，內含子模型表現較佳。

2. 核支援向量機(kernel SVM) γ 越大，能將數據分類的越好

由外顯子及內含子模型輸出的核支援向量機(kernel SVM)可知，γ=100時能判斷出極佳的決策分類區域，然而，當分類器在處理未見過的數據時，「一般化誤差」也越高，有可能造成錯誤判斷。

(三) 利用隨機森林(Random Forest)、裝袋法、AdaBoost方法訓練模型結果討論

1. 模型訓練準確度及決策樹分類器的最佳數量

外顯子模型訓練準確度為隨機森林訓練準確度為0.765；裝袋法為0.756；AdaBoost方法為0.756，隨機森林分類器數量介於0至20之間有較高的準確率，又以8或9時準確率最高。內含子模型為隨機森林訓練準確度為0.942；裝袋法為0.930；AdaBoost方法為0.930，隨機森林在分類器數量介於6至20之間，有較高的準確率。因樣本維度較低，增加決策樹分類器時，可能造成過度擬合，導致較低的準確度。

2. 三種分類器訓練結果比較

由輸出準確度可知，隨機森林(Random Forest)在外顯子及內含子訓練皆有較高準確率，可以推測此訓練資料組成較為單純，錯誤分類資料較少，使用裝袋法及AdaBoost方法並未提高準確度。

(四) 三種不同模型訓練方式比較討論

模型種類 / 內容	Pipeline	SVM 支援向量機	隨機森林、裝袋法、AdaBoost法
準確率	(外)0.719、 (內)0.870 整體最低	(外)0.756、 (內)0.922	(外)0.765、 (內)0.942 整體最高
分析	若單筆資料中內容較多時使用，因為其包含降維、特徵縮放、再擬合出模型，若資料充足，能較高效產出結果。	若資料量較少時使用，因為此二方法為類似「分類器」的訓練方式，適合本研究所的訓練資料。然而，若資料量過大，可能有過度擬合以及運作時間過長的問題。	

整合網路基因變異註釋工具

(一) 能夠大幅減少醫生預測病人基因變異致病性的時間，具有極高的實用性

醫生使用網路工具進行疾病預測時，需要利用大量的時間一筆一筆剪貼病人資料至網路基因註釋工具進行預測，需要4個小時以上才能夠完成繁雜的預測步驟，因此，若不到10分鐘即能輸出結果，能有效解決醫生耗費過多時間預測基因變異，大幅提升註釋工具的實用性。

(二) 符合第一線醫界人員或研究人員需求，幫助增加網路基因變異註釋工具的使用者

基因變異預測疾病工具是第一線醫界人員以及研究人員進行臨床上的參考或研究相關領域的重要助手，然而，醫界人員因為推廣不足而未使用到這些新興領域的實用工具，若整合這些工具，或許能夠增加此類工具的使用者。

蛋白質變化預測基因變異之致病程度

(一) 利用程式快速判斷蛋白質變化且直接影響致病程度，藉此提高預測準確度

由於蛋白質變化直接影響人體機能，進而影響基因變異所造成之致病程度。利用程式能夠快速計算不同鹼基變化所產生的不同蛋白質，再由不同種類的蛋白質變化方式判斷可能造成的致病程度，並且結合真實數據之致病力分數。

(二) 藉由資料的擴充能夠增加真實資料的數據量，並期望未來能有非單點單變異之致病預測模式

由於目前本研究程式中有一項致病分數來自真實資料，因此若增加現有的數據量，能使預測更具準確度。另外，由於目前此程式用於判斷的基因變異種類屬於單點單變異，但基因突變方法尚有插入突變、刪除突變、單點多變異等種類，因此期望未來能夠增加此類變異方式之致病預測。

未來展望

(一) 藉由整合網路基因變異註釋工具，蒐集大量病人資料，進而訓練深度學習模型，提高預測基因變異造成致病力的準確度

「深度學習」需要極為龐大的資料量才能訓練高準確度的模型。然而，生物醫學領域資料複雜性、多樣性極高，難以在各個類別基因變異型態皆有大量的數據進行模型訓練，也造成本研究預測腎臟基因模型準確度問題。雖然大型國外網站(Clinvar、HGMD等)有由不同醫生上傳的各式基因突變資料，但若只看單一基因變異種類(由基因種類、變異位置、採樣人種等分類方法)的資料量，仍不足以進行機器學習。因此，若能藉由整合網路基因變異註釋工具，蒐集大量病人資料，以得更多更深入更豐富的資料，或許能夠提高預測基因變異致病力模型的準確度。

(二) 在蒐集醫生上傳資料的同時，為臺灣建立基因資料庫

每一筆病人的資料都十分珍貴，都可作為未來研究的樣本，然而，目前臺灣並沒有一個較大的資料庫提供醫生蒐集病歷資料，因此，我們期望可以藉由「整合網路基因註釋工具」，在醫生上傳基因變異資料的同時，蒐集並公開這些資料，為臺灣建立較完整的基因變異資料庫。

結論

一、模型訓練、網路工具整合及簡易基因註釋器結果

訓練預測腎臟基因變異致病力模型	管線(Pipeline)	準確度為八成左右，因為公開基因變異資料量不足且不平均，導致部分準確度仍提升空間。外顯子模型之最佳參數C=0.01、0.1、1、10、100時最佳，內含子模型則是C=100。
	scikit-learn訓練	核支援向量機(kernel SVM) γ 越大，擁有最佳的決策邊界。
	隨機森林 (Random Forest)	外顯子模型決策樹數量在8~9個有最高準確度0.765 內含子模型在決策樹數量6~20個有最高準確度0.942。
	資料較多擬合使用管線(Pipeline)訓練模型，資料較少則擬合使用scikit-learn中的訓練方法及隨機森林(Random Forest)。	
網路基因變異註釋工具	製作單一介面整合各網站，大幅減少醫生預測病人基因變異致病性的時間，具有極高的實用性。	
蛋白質變化預測基因變異之致病程度	快速判斷基因變異造成的蛋白質變化，進而直接影響致病分數。	

二、資料不足的情況下最佳的基因預測模式

為了解決公開基因變異資料量的不足，我們整合網路基因註釋工具，增加網路工具的實用性，再利用網路公開基因變異資料訓練模型，並且期望藉由此網路整合工具蒐集大量病人資料，以提升在資料不足的情況下最佳的基因預測準確度。