國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Master Thesis

基於深度學習方法架構之基因虛擬探針模型 GeneVPNN 於推薦潛在特徵基因組暨預測目標基因群研究

The study of economical genetic testing with GeneVPNN in complex disease recommending latent feature genes and predicting effective target genes

張心和

Hsin-Ho Chang

指導教授：盧子彬 博士

Advisor: Tzu-Pin Lu, Ph.D.

中華民國 112 年 6 月

June 2023

doi:10.6342/NTU202301211

論文口試委員審定書

國立臺灣大學碩士學位論文
口試委員會審定書

論文中文題目:
基於深度學習方法架構之基因虛擬探針模型 GeneVPNN 於推薦潛在特徵基因組暨預測目標基因群研究

論文英文題目:
The study of economical genetic testing with GeneVPNN in complex disease recommending latent feature genes and predicting effective target genes

本論文係 張心和 君（學號 P10849001）在國立臺灣大學流行病學與預防醫學研究所完成之碩士學位論文，於 民國 112 年 6 月 12 日 承下列考試委員審查通過及口試及格，特此證明。

口試委員：

（簽名）

（指導教授）

陳佩君

doi:10.6342/NTU202301211

致謝

本研究涵蓋生物醫學統計及深度學習等領域專業知識，也是本人(張心和)於國立臺灣大學流行病學與預防醫學研究所求學期間之研究成果分享。在進入本所就讀之前，本人的專業背景著重於 ICT 資訊通訊產業，經由本所規劃之專業生物醫學統計與資料科學等課程領域知識及技術實作薰陶，讓本人得以發揮累積多年之資訊專業技能貢獻於人類醫學研究領域。

在此特別感謝我的論文指導教授 盧子彬 博士，以其專業研究經驗分享生醫研究實務，並引領出關鍵生醫研究瓶頸等議題，給於我莫大的研究啟發。研究期間也給予我多次客觀研究內容審查意見，驅使本研究試驗經多元化驗證更趨生醫專業。也特別感謝本論文口試委員 臺大公衛學院 蕭朱杏 副院長 及 國立臺北教育大學數學暨資訊教育學系 陳佩君 老師等給予專業審查意見。與本研究相關學習的課程方面，感謝以下授課老師課堂領域知識的傳授與討論，量性科學課程的 陳秀熙教授、機器學習的生醫應用課程的 張瀚 老師 等學習歷程授課的老師們。另外也特別感謝，引領我進入 AI 領域的國立中央大學資管系研究所的 陳以錚 老師，讓我在深度學習領域有深厚的領域知識。

以上是對於本研究給於我相關領域知識老師的致謝，本研究成果之發表相信並非是研究目標之終點，期以未來生醫研究應用本研究成果，加速複雜疾病生醫藥研發為人類做出貢獻。

張心和 謹致於

國立臺灣大學 流行病學與預防醫學研究所

中 華 民 國　一百一十二 年　六 月

doi:10.6342/NTU202301211

中文摘要

背景：

現今不論是在臨床醫學或是在預防醫學等領域無不追求精準醫療目標，然而精準醫療除需仰賴生醫藥領域之技術精進外，更重要關切的議題為如何在有限研究資源限制下，達到研究目的預期成效堪為重要。然而基因檢測技術及研究工作在精準醫療研究體系中占有舉足輕重地位，基因檢測技術已運行於醫學研究多年至為成熟，且各國學術研究單位也應用基因檢測技術針對指標性疾病、癌症發表過所研究疾病之特徵基因組，隨著醫學基因研究之演進，已有多篇學術文獻證實致病之基因已非單純關鍵少數基因影響所為，故必須將研究觸角延伸至更廣範圍之全基因研究，然而全基因檢測分析所耗之經費及時間，實為醫療研究團隊及病患首要必須面對之議題。

鑒於基因研究日趨重要及解決全基因檢測關切之時間、成本議題，本研究提出 GeneVPNN 基因虛擬探針檢測模型，提供未來生醫藥領域研究致病基因模擬推論參考。

方法：

GeneVPNN (Gene Virtual Probe Neural Network)模型為結合統計及深度學習之複合式科學方法，解決關鍵有效目標基因群推薦檢測議題，GeneVPNN 主要於前階段以變異係數值分析(CV, Coefficient of Variation)挑選出 CV5000 有效目標基因群，進階應用深度學習 Autoencoder (AE)演算架構，挑出更具代表性之 LF-GENESET 潛在特徵基因組，並以 NN 類神經網路推論出完整之 CV5000 基因探針檢測數值全貌。研究試驗採用 GEO 資料平台提供之 GSE102484 資料集，其為台灣和信治癌中心醫院以微陣列晶片檢測亞洲人種乳癌病患之基因探針數值。

doi:10.6342/NTU202301211

結果：

GSE102484 資料集經拆分成訓練及測試兩資料集，GeneVPNN 模型經測試資料集驗證評估後，模型預測之誤差率經統計後，預測 CV5000 基因探針值誤差率小於 30%之數量占比所有預測探針數量可達 96.71 %；預測 CV5000 基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.47 %。

結論：

本研究兩個主要研究產出，一為模型泛用性設計，GeneVPNN 模型所建構之預測流程，在本研究中雖然以女性乳癌術後之追蹤期間的資料集進行研究，但 GeneVPNN 的設計架構是以適用廣泛疾病前提下所設計之模型，可輔助未來醫學研究於各專科複雜疾病之有效目標基因群 CV5000 檢測數值之推論，作為生醫藥領域基因研究前期分析工作之參考數據。另一研究產出為推薦最經濟之基因檢測範圍 LF-GENESET，有助於探索未知病因基因檢測執行規劃，以 GeneVPNN 之 AE 程序所推薦給予最經濟之檢測基因範圍之效益下，醫學研究團隊及病患可省去做全基因檢測之成本，並應用 GeneVPNN 推論研究族群之有效目標基因群 CV5000 進行醫學研究及輔助臨床醫學診斷參考。

關鍵字：基因預測、自編碼器、深度學習、降維、基因組選擇

doi:10.6342/NTU202301211

Abstract

Background：

Nowadays, we are pursuing the goal of precision medicine in the field of clinical medicine and preventive medicine. In addition to relying on the advancement of technology in the field of biomedicine, the more important issue that we must face up to the problem of how to achieve the expected results of research goals under the constraints of limited research resources. However, genetic testing technology plays a pivotal role in the precision medical and has been run in medical research for many years to reach maturity. With the progress of genetic research, much research has confirmed that complex diseases are not simply caused by a few symbol genes. Therefore, it is necessary to extend the research to whole genetic testing. However, the cost and time spent of whole genetic testing are the primary issues that medical research teams and patients must face. In view of that, this study proposes the GeneVPNN model, the virtual probe of genetic testing, to recommend doing real genetic testing range and compute the inference of whole genetic testing.

Method：

GeneVPNN (Gene Virtual Probe Neural Network) is a composite scientific method

doi:10.6342/NTU202301211

combining statistics and deep learning technologies to solve the issue of recommended genetic testing range. At the previous stage, GeneVPNN analyzes the coefficient of variation from RNA microarray raw data to get effective gene group (CV5000); and further, uses Autoencoder (AE) algorithm architecture to select the latent feature genes (LF-GENESET) on the condition of under specified quantity. GeneVPNN also provides prediction function, that uses deep learning neural network to deduce gene probes' value of CV5000 from LF-GENESET. This study uses GEO GSE102484 dataset that is a gene expression array data of Asian breast cancer patients that made by Taiwan Koo Foundation SYS Cancer Center.

Result：

The GSE102484 dataset is split into two parts for training and testing. After the trained GeneVPNN model is validated and evaluated by the test dataset, the results of prediction show that the relative error percentage of less than 30% covers 96.71 % of the test dataset and the relative error percentage of less than 50% covers 99.47 % of the test dataset.

Conclusion：

This study has two main research outputs, one is the versatile design model was made for diseases studied. Although the GeneVPNN model was trained and tested by GSE102484

doi:10.6342/NTU202301211

dataset that is relevant to breast cancer of women, but the framework of GeneVPNN is designed for unspecified diseases analysis. So, it can assist medical research getting a recommended range of genetic testing of unspecified diseases for achieving a more economical genetic testing goal. The other research output is using recommended LF-GENESET to make GeneVPNN deduce the CV5000 gene probe values of unspecified diseases in the early stages of research. Therefore, in addition to saving research's cost and time by GeneVPNN, it can also assist doctors and researchers in diagnosing patients' postoperative conditions.

*

Keywords：Gene prediction, Deep learning, Autoencoder, Dimension reduction, Genomic selection

doi:10.6342/NTU202301211

目錄

doi:10.6342/NTU202301211

圖表資料

圖目錄

doi:10.6342/NTU202301211

doi:10.6342/NTU202301211

表目錄

doi:10.6342/NTU202301211

第一章 導論

1.1 研究背景

現今不論是在臨床醫學或是在預防醫學等領域無不追求精準醫療[1]目標，然而精準醫療除需仰賴生醫藥領域之技術精進外，更重要關切的議題為如何在有限研究資源限制下，達到研究目的預期成效堪為重要。然而基因檢測技術及研究工作在精準醫療研究體系中占有舉足輕重地位，基因檢測技術已運行於醫學研究多年至為成熟，且各國學術研究單位也應用基因檢測技術針對指標性疾病、癌症發表過所研究疾病之特徵基因組[2-5]，隨著醫學基因研究之演進，已有多篇學術文獻證實致病之基因已非單純關鍵少數基因影響所為也包含遺傳基因[6]，尤其在現今複雜疾病研究中，要發掘出根治的病因基因組是件極具挑戰的工作，在過去癌症的基因研究中也發現基因變異除了與腫瘤有因果關係外，也發現與腫瘤無相關的基因突變[7]，如部分致癌基因隱含修復 DNA 功能之缺陷[8]及有些基因也被發現會影響細胞功能，包括轉錄、粘附和侵襲[9]，故需要將研究面向設計更為嚴謹；再者從研究癌症治療方法議題觀之，現今醫學研究單位已發展出多種癌症治療方法，像是抑制癌細胞藥物、核酸藥等，但若要從療效持續性及可根治複雜疾病的觀點出發檢視各癌症治療方法，從過去的研究文獻分析，以採用基因檢測找出基因突變位點，施以體內靶向之核酸基因療法[10]堪為重要，並且考量對所有病患皆有顯著療效，更應從整體免疫宏觀環境的觀點下[11]，研究與癌症相關之免疫組織的基因變化，同樣地，在生殖細胞系相關的 DNA 突變(germline mutation)，也會增加患得癌症的風險[12]，以上論述之觀點皆會關乎致癌基因之研究範圍及複雜度，靶向目標基因如何完整透過研究找出，進而發展出對應之核酸藥進行臨床試驗，是今日先進生醫藥研究單位發展之關鍵技術目標，故必須將研究觸角延伸至更廣範圍之全基因研究，然而全基因檢測分析所耗之經費及時

doi:10.6342/NTU202301211

間，實為醫療研究團隊及病患首要必須面對之議題。

1.2 研究動機與目的

鑒於基因研究日趨重要及解決全基因檢測時間及成本議題，本研究提出以資料科學領域涵蓋之統計方法及深度學習演算法架構之 GeneVPNN 基因虛擬探針檢測模型，其模型演算法組成架構及研究產出資料詳如圖 1 所示，應用 GeneVPNN 具以推薦出執行基因檢測最經濟且關鍵之潛在特徵基因組 LF-GENESET(Latent feature gene set)範圍，作為未來生醫藥領域研究複雜疾病致病基因之參考數據來源。

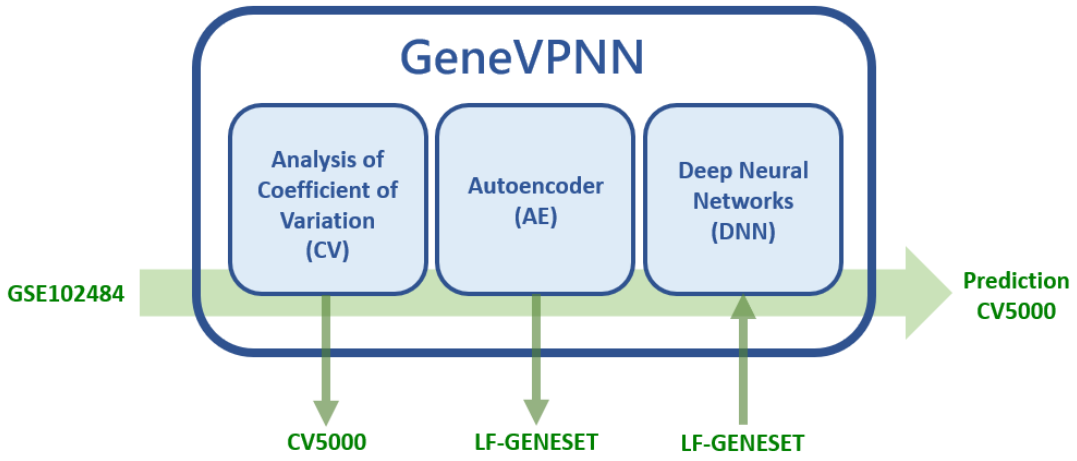

圖 1 GeneVPNN 模型功能圖

在臨床醫學對於疾病之預後推論，醫師往往受限於有限已知的臨床變項而去預後病患可能之病症及其程度，做出後續診療決策，現今在精準醫療[13]的趨勢帶動下，基因檢測分析將有助於輔助醫師對疾病預後做出更可靠之判斷，隨著醫學研究不斷往前邁進，對於病因推論從過往的少數關鍵基因檢測研究，朝向研究更多基因群間交互作用之研究目標發展，在全基因組關聯分析 GWAS (genome-wide association study)[14]中要發掘出所有真正與病症相關聯及交互作用基因群實屬複雜之檢測及分析工作，但首先得要取得完整基因檢測結果數據，而全基因

doi:10.6342/NTU202301211

檢測所耗之資源成本，對病患及醫療研究單位而言是一大限制。

有鑒於此，本研究乃針對完整基因檢測之資源限制議題，提出以科學方法快速推論出研究目標族群之有效目標基因群 CV5000 基因探針數值全貌，以降低檢測成本及時間，為精準醫療做出貢獻。在臨床醫學之癌症治療後長期追蹤病患之術後發展狀態至為重要，藉由 GeneVPNN 推論之 CV5000 可給予醫生在做出預後判斷前提供更多參考資訊。本研究設計以樣本個案癌症術後持續追蹤狀態下之組織切片微陣列基因晶片資料集為主，設計模型產出推薦之潛在特徵基因組 LF-GENESET，以及設計模型具備預測推估研究族群之有效目標基因群 CV5000 基因探針數值全貌功能，並分析模型試驗之預測結果誤差，作為本研究試驗設計之可靠度參考。

doi:10.6342/NTU202301211

第二章 研究材料與方法

2.1 資料來源

本研究試驗所採用之微陣列生物晶片檢測資料集為美國 NCBI 下的 GEO 資料庫平台[15]所發布之 GSE102484 資料集為主，本研究下載的版本為 Jul. 25, 2021 維護更新之版本，該資料集以 GPL570 平台檢測分析，資料涵蓋 683 筆亞洲人種乳癌病患個案，此資料集是臺灣和信醫院收納診斷治療後並進行追蹤之個案所提供之微陣列基因檢測高通量資料[16]，檢測之晶片採用 Affymetrix U133 plus 2.0 arrays，每筆個案皆紀錄 54,627 個微陣列基因檢測探針數值及相關數據，個案資料並包含追蹤標記癌細胞是否遠端移轉(Event_metastasis)作為追蹤結果。

2.2 樣本篩選

本研究在樣本資料篩選及資料前處理階段，採用 R 語言進行 GSE102484 資料集下載，並經程式做資料前處理，包含 quantile normalization 正規化的處理，避免因批次實驗所產生系統性的誤差而影響到模型的預測成效，在此資料集之 683 筆研究個案中，包含有癌細胞轉移/擴散(event_metastasis=1) 101 筆 及 癌細胞未轉移/擴散(event_metastasis=0) 582 筆個案，GSE102484 經資料前處理程序後將資料集格式轉存成通用 CSV 格式檔，提供後續各階段分析方法以 Python 語言建構之各式深度學習模型做研究分析。

由於 GSE102484 資料集紀錄的是 GPL570 平台檢測數據，每個樣本個案基因探針資料維度高達 54,627 維，故有效基因探針資料篩選工作至為關鍵，本試驗於研究方法初期階段設計以可靠之降維方法，篩選出有效目標基因群組成 CV5000 資料集。

doi:10.6342/NTU202301211

2.3 研究方法論述

圖 2 為本研究 GeneVPNN 模型研究試驗流程設計，第 1 階段以 R 語言下載之 GSE102484 微陣列基因檢測資料集，並經第 2 階段資料前處理後轉存成研究目的使用之 CSV 格式資料集，為將預測範圍聚焦至潛在反應病因之有效目標基因群內，於第 3 階段研究方法，試驗針對提取目標族群之研究議題相關有效基因而設計，本階段設計之論述則假定特徵基因群會因不同病患研究之事件發生與否，同一基因會有不同檢測數值變化，故提出結合統計領域可靠之變異係數分析(CV,Coefficient of Variation)方法，挑選出變異係數高之 TOP 5000 基因組成 CV5000 有效目標基因群，作為本研究後續各階段試驗之主體資料集與預測標的基因群。

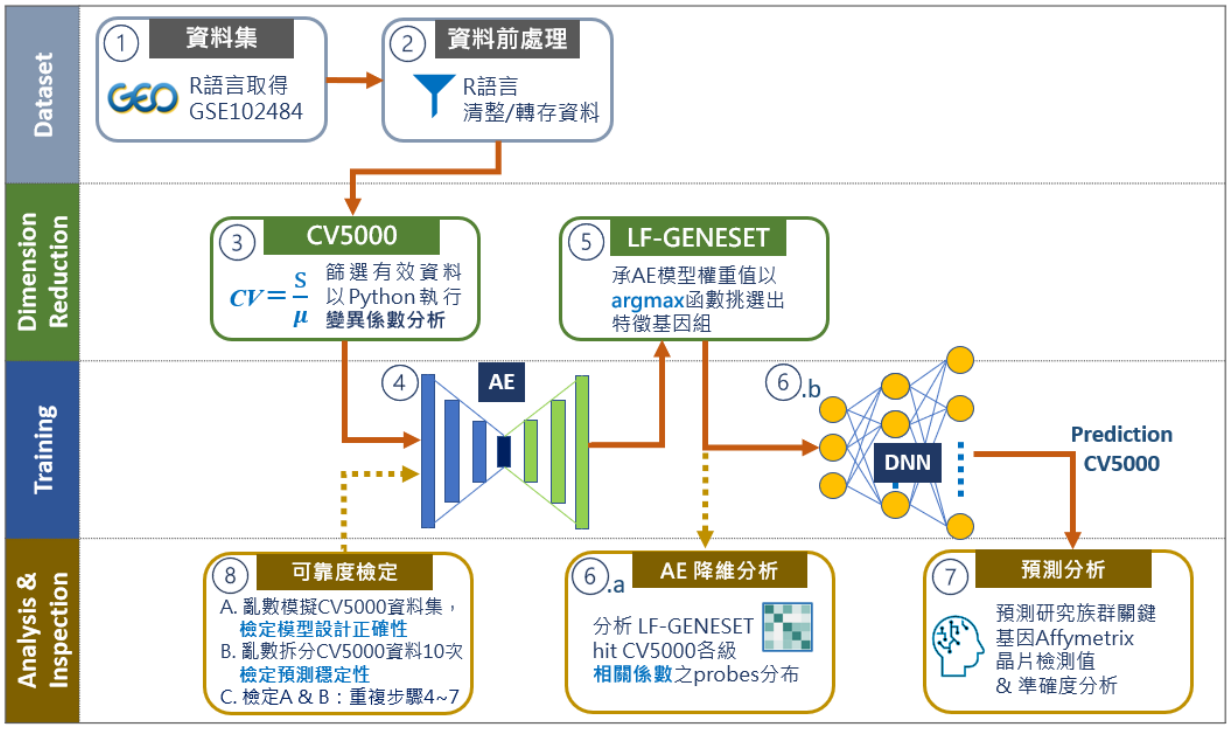

圖 2 GeneVPNN 模型研究試驗流程

圖 2 第 4 階段及第 5 階段試驗則為 AE 模型建構訓練及潛在特徵基因組 LF-GENESET 選取程序，試驗方法則設計以進階非監督式深度學習 Autoencoder 演算

doi:10.6342/NTU202301211

架構推論，以推薦出不超過指定數量內更具代表性之潛在特徵基因組。

圖 2 中第 6.a 階段試驗，為針對 AE 降維取得之 LF-GENESET 檢測其推薦之基因探針分布於 CV5000 各級相關係數探針的分析，以衡量 AE 有效降維能力，第 6.b 階段試驗，為預測有效目標基因群 CV5000 之模型建構及訓練試驗，其試驗方法以 DNN 類神經網路演算法推論，第 7 階段依前階段訓練出之 DNN 模型，以第 5 階段試驗所推薦出之潛在特徵基因組 LF-GENESET 推論出完整之 CV5000 基因探針數值全貌，並分析整體推論之預測誤差，作為本研究方法評鑑 GeneVPNN 模型可靠度之參考。圖 2 中，除第 1 至 2 階段以 R 語言進行資料集下載處理外，其餘試驗流程第 3 至 7 階段皆以 Python 語言自動化連續執行組成 GeneVPNN 基因虛擬探針推論模型，於各試驗階段之目標工作分別推薦產製出 CV5000 與 LF-GENESET 兩關鍵基因組，並於最後階段可以使用未訓練過之資料集達成預測推論出與研究族群相關之有效目標基因群 CV5000 探針數值全貌。第 8 階段主要設計檢定 GeneVPNN 之設計正確性及其預測之穩定性，在模型設計正確性之檢定方法，則以亂數模擬 CV5000 資料集經 GeneVPNN 試驗步驟 4 至 7 程序後，觀測其預測 CV5000 差異，以衡量模型設計正確性，模型設計之預測穩定性檢定方法，則採用機器學習常用之十折交叉驗證方法求證模型預測穩定性。

2.4 變異係數分析方法論述

鑒於 GSE102484 以高通量 GPL570 平台檢測分析，資料多達 54,627 維度，為求研究不受研究族群無效因子之基因干擾及提升機器學習運算效能等目的，試驗設計降維篩選功能以挑選出有效目標基因群 CV5000，試驗在假定癌患在復發與未復發之不同條件下，兩族群代表性基因表量具高變異之論述下，進行變異係數分析(CV,Coefficient of Variation)[17]以篩選出變異係數高之基因探針群，作為研究 GeneVPNN 模型之主體試驗資料集。

doi:10.6342/NTU202301211

公式(1)是 CV 統計分析資料集中各基因檢測探針數值之變異係數，首先統計求出各基因探針之樣本標準差 S 及樣本平均值μ後，以下列 CV 公式求出各基因探針之變異係數。

$$CV = \frac{S}{\mu} \tag{1}$$

公式(1)中的 S 為樣本標準差，μ為樣本平均值。

經前述統計分析出 GSE102484 資料集中各個基因探針之 CV 值後，本研究試驗設計將演算出的各基因探針之 CV 數值進行由大至小排序，再挑選出本研究設計欲篩選出之變異係數高之 TOP 5000 基因探針群，組成 CV5000 資料集。

2.5 Autoencoder 自編碼器試驗方法論述

依據統計推論之精髓--以樣本推論群體之論述，目前人類在研究上已知的基因超過 2 萬個以上[18]，但真正具有明顯基因表現量的僅有少數的 20~30%，故生醫藥研究領域面臨關鍵議題--如何應用現今成熟之高通量微陣列生物晶片所檢測出之基因探針數值，萃取出研究目標族群之代表性潛在特徵基因組資料，實為醫學基因研究必須面臨的關鍵議題。故本試驗在研究問題可應用科學演算法求解的假設前提下，試驗設計以 Hinton 與 Salakhutdinov [19]研究提出的自編碼器 Autoencoder(AE)演算法架構出深度學習模型，以推論出最具代表性之潛在特徵基因組 LF-GENESET，AE 模型試驗過程中經反覆訓練及應用反向傳遞(Backpropagation)推算調整各層神經元之各項權重的梯度參數，以達成縮小預測誤差目的，訓練出基因輸入所對應最佳化之各層神經元權重，並以選取權重變動幅度影響較多的基因探針族群為篩選原則，作為推薦代表群體基因的潛在特徵基因組 LF-GENESET 之推薦標準。此研究設計之立論假說是由各神經元推派代表基因概念為基礎，如同同一選區候選人們(微陣列晶片基因探針群)競選國會代表，

doi:10.6342/NTU202301211

由候選人所屬選區下轄之各開票所(神經元)開出最高票之候選人，作為該投票所之最高票代表，換句話說，也就是依據 AE 演算架構最後經反向傳遞推算調整出各層神經元之各項權重的梯度參數，取 AE 第一隱藏層各神經元權重影響力最多之基因探針，取第一隱藏層各神經元之權重之立論基礎，乃因該層各神經元之權重調整也是經由其後各隱藏層神經元層層反向推算出之故，然而各投票所一定會有選出相同最高票之候選人之情形，在本試驗中也存在相同情況，故試驗設計會將重複之基因探針僅以一個作為代表納入推薦之 LF-GENESET 基因組。

AE 試驗設計

AE 模型中的編碼器(encoder)部分，試驗設計推薦 LF-GENESET 之基因數目以不超過 200、500、1000、2000 四種試驗情境，進而將各情境 AE 模型所推薦出之潛在特徵基因組 LF-GENESET 自動帶入下一預測 CV5000 之 DNN 模型試驗中。有關 AE 模型之基因探針挑選方法部分，試驗設計取第一隱藏層之權重向量矩陣，並以 KERAS API 的 argmax 函數(2)挑選出正負權重影響幅度最大者之特徵基因組。

argmax 函數定義：

$$\underset{x \in X}{\operatorname{argmax}} f(x) := \{x \in X : f(y) \leq f(x) \; for \; all \; y \in X\}. \tag{2}$$

在 argmax 函數定義中，x 為 $f(x)$函數之自變數，X 為可能的 x 的集合，y 為 X 的子集合，argmax 是將給定的函數 $f(x)$將其 $f(x)$對應的自變數之權重係數 x 集合取其最大值。

於 AE 試驗最終階段則應用 argmax 函數定義(2)，將第一隱藏層所有神經元之權重向量矩陣取絕對值後，取出各神經元權重影響最大者之基因探針，作為所推薦之潛在特徵基因組 LF-GENESET。

doi:10.6342/NTU202301211

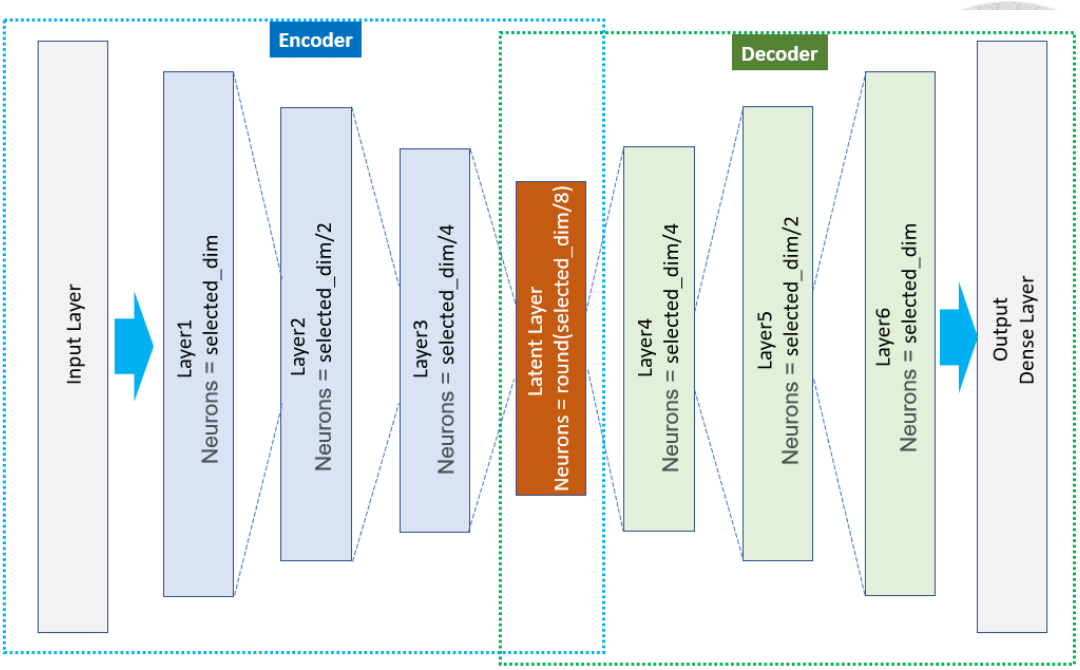

圖 3 GeneVPNN 之 Autoencoder 架構設計

圖 3 為本研究試驗 GeneVPNN 模型中的 Autoencoder 架構設計，主要由 Encoder 及 Decoder 所組成，AE 之階層架構設計會隨 LF-GENESET 所設定欲挑選之潛在特徵基因數，即圖 3 中所設定的變數名稱 selected_dim，自動調整 Encoder 及 Decoder 之 Layer1 至 Layer6 各隱藏層(Hidden layer)的神經元(Neurons)數量，無須手動給予模型太多設定參數，以因應各試驗條件 Input layer 之資料輸入。

AE 試驗藉由 Latent layer 之特徵資訊傳遞給 Decoder 的過程，於訓練階段會不斷優化 Encoder 及 Decoder 各隱藏層之神經元權重值，直至 Decoder 預測的差距收斂至最小，此時 Encoder Layer1 的各層神經元權重值，便是後續挑選 LF-GENESET 的參考數據來源。評量預測差距之損失函數以 MSE 均方誤差公式(6)為主，提供給模型優化器調整各神經元權重之參考依據，AE 整體預測損失 L 定義如下。

AE 整體預測損失 L 定義：

$$L = d\left(X_i, D(E(X_i))\right) \tag{3}$$

doi:10.6342/NTU202301211

其中 X_i 為 encoder 輸入的基因資料矩陣，$d()$為 MSE 均方誤差公式(6)，E()為 encoder 推算 Latent layer 各神經元權重值 Z 矩陣的函數，X_i 的預測值 $\hat{X}_i = D(E(X_i))$，D()為 Decoder 預測$\hat{X}_i$之函數。AE 優化目標就是降低 decoder 預測誤差，表達式如下。

$$\arg\min_{X_i,\hat{X}_i} L(X_i,\hat{X}_i), \quad 其中 \ \hat{X}_i = D(E(X_i)) \tag{4}$$

欲使 Autoencoder 精準預測及快速收斂，程式邏輯架構設計之細節致為關鍵，下圖所展示的是 GeneVPNN 之 Autoencoder 各 Layer 組成。

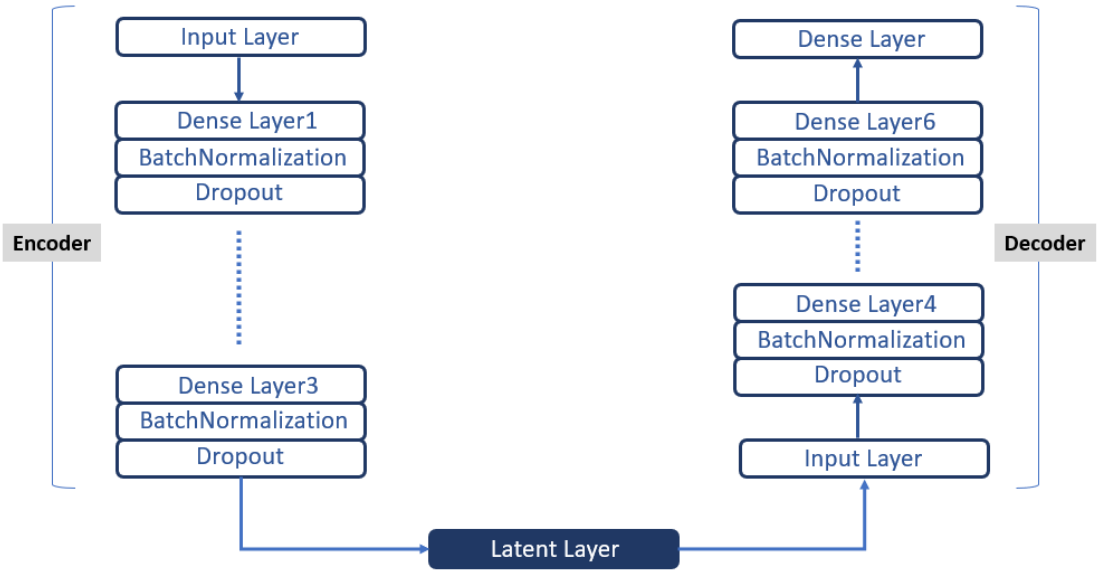

圖 4 GeneVPNN 之 Autoencoder 程式邏輯設計

本研究應用 KERAS API 中的 BatchNormalization layer，其為 Sergey Ioffe 及 Christian Szegedy 於 Google 發表的 Batch Normalization 方法[20]，在各階層輸入批次數據訓練前，先將資料做 scaling 標準化，有效解決訓練期間梯度消失的問題，讓模型訓練達到快速收斂之目的。 並加入了 Dropout layer，防止各層神經元訓練出的權重值過於擬合訓練資料，有效改善預測誤差。

圖 5、圖 6 為使用 Python 建構 AE 模型之摘要輸出，以推薦 LF-GENESET 小於 2000

doi:10.6342/NTU202301211

個基因為例，其為訓練 AE 預測 CV5000 試驗情境所產出之執行摘要。

```
Model: "encoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 encoder_input (InputLayer)  [(None, 5000)]            0

 selected_Layer (Dense)      (None, 2000)              10002000

 batch_normalization (BatchN  (None, 2000)             8000
 ormalization)

 dropout (Dropout)           (None, 2000)              0

 L2 (Dense)                  (None, 1000)              2001000

 batch_normalization_1 (Batc  (None, 1000)             4000
 hNormalization)

 dropout_1 (Dropout)         (None, 1000)              0

 L3 (Dense)                  (None, 500)               500500

 batch_normalization_2 (Batc  (None, 500)              2000
 hNormalization)

 dropout_2 (Dropout)         (None, 500)               0

 latent_out (Dense)          (None, 250)               125250

=================================================================
Total params: 12,642,750
Trainable params: 12,635,750
Non-trainable params: 7,000
```

圖 5 AE-Encode 摘要(LF-GENESET_u2000)

```
Model: "decoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 decoder_input (InputLayer)  [(None, 250)]             0

 dense (Dense)               (None, 500)               125500

 batch_normalization_3 (Batc  (None, 500)              2000
 hNormalization)

 dropout_3 (Dropout)         (None, 500)               0

 dense_1 (Dense)             (None, 1000)              501000

 batch_normalization_4 (Batc  (None, 1000)             4000
 hNormalization)

 dropout_4 (Dropout)         (None, 1000)              0

 dense_2 (Dense)             (None, 2000)              2002000

 batch_normalization_5 (Batc  (None, 2000)             8000
 hNormalization)

 dropout_5 (Dropout)         (None, 2000)              0

 dense_3 (Dense)             (None, 5000)              10005000

=================================================================
Total params: 12,647,500
Trainable params: 12,640,500
Non-trainable params: 7,000
```

圖 6 AE-Decoder 摘要(LF-GENESET_u2000)

doi:10.6342/NTU202301211

2.6 DNN 線性回歸試驗方法論述

GeneVPNN 的 DNN(Deep Neural Network)試驗主要以章節 2.5 Autoencoder 自編碼器試驗方法所挑選出指定數量下之 LF-GENESET 基因組，應用 DNN 神經網路演算架構將 CV5000 基因群的探針檢測值推論出。因所推論之目標值屬於連續型資料型態，故設計以 DNN 線性回歸神經網路作為本階段試驗模型。

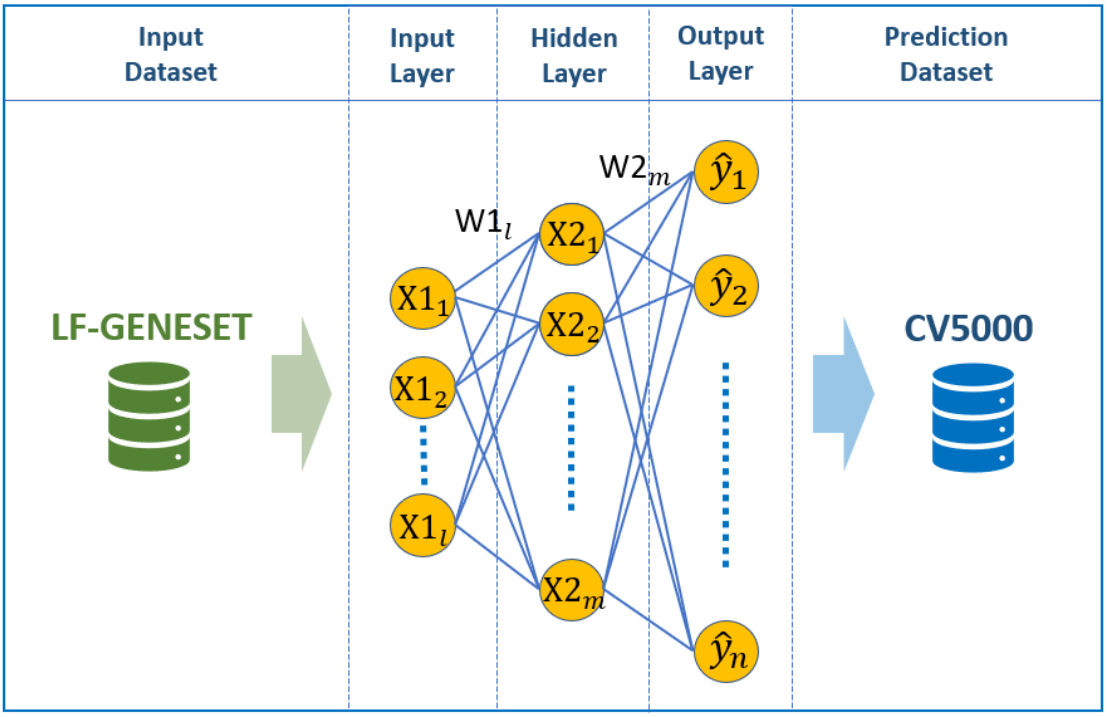

圖 7 DNN 預測 CV5000 架構

圖 7 中，Input layer 的維度(同 LF-GENESET 基因探針數)即是隱藏層 Hidden Layer 中各神經元之線性回歸自變數的個數，Output layer 的神經元數則設計以目標推論之 CV5000 基因數為主，由於 DNN 神經網路之演算架構後級階層之各神經元會全連結前級階層所有神經元或 Input layer 中各輸入變數，因而組成圖 8 線性回歸式之推論架構，其為本試驗 GeneVPNN 中 DNN 各層單一神經元間演算原理，並由本研究設定之 Adam[21]優化器(optimizer)依所設定之損失函數(Loss function) MSE (6)於各訓練階段運算推論結果差距後，依反向傳播梯度下降法調整各項輸入權重值(weight)，並給予線性回歸式合理之偏誤值(bias)，依此原

doi:10.6342/NTU202301211

理重複運算調整權重值及偏誤值，直至評估 DNN 整體推論誤差最小化，至此各神經元權重將收斂至穩定值而組成本階段試驗訓練後之 DNN 模型。

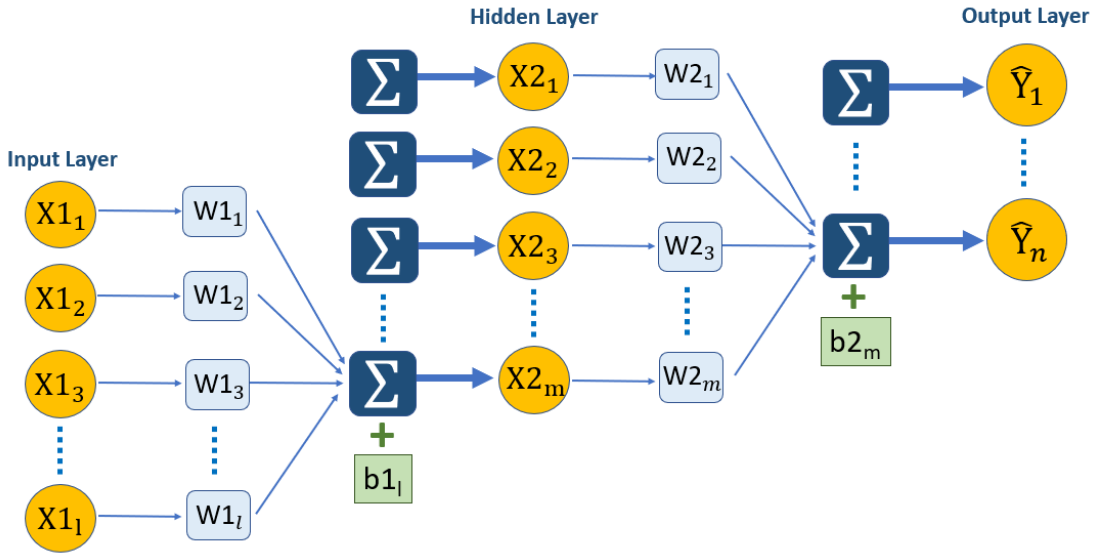

圖 8 GeneVPNN 中 DNN 各層單一神經元間演算原理

CV5000 各基因檢測值$\hat{y}_n$的推論公式：

$$\hat{y}_n = (\sum_{1}^{m}((\sum_{1}^{l}(X1_l * W1_l)) + b1_l) * W2_m) + b2_m \quad (5)$$

公式(5)中，n 為 CV5000 之基因微陣列探針數量 5,000；l 為 LF-GENESET 所設定欲挑選之潛在特徵基因數；m 為 hidden layer 神經元數。

$W1_l$ 及 $W2_m$ 權重值，經由 Adam 優化器於每批次訓練結束後，自動更新各$\hat{y}_n$所對應之各階層各神經元之權重值($W1_l$、$W2_m$)與偏誤值($b1_l$、$b2_m$)，經由所設定之模型訓練次數(epoch)不斷更新權重，依所設定之 MSE 損失函數(6)檢視各次訓練後之預測 $\hat{y}_n$值與 ground truth 實際值之差異程度，優化器會再度更新各權重值，使預測之$\hat{y}_n$逼近 ground truth 實際值達到訓練結果收斂目的。

圖 9 為 Python 建構 DNN 模型之輸出摘要，以試驗 LF-GENESET 不超過 2000 個基因為例，訓練 DNN 預測 CV5000 試驗情境所產出之執行摘要。

doi:10.6342/NTU202301211

```
Model: "DNNmodel"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 PredictRealGeneDNN_input (I  [(None, 862)]            0
 nputLayer)

 dense_4 (Dense)             (None, 4820)              4159660

 batch_normalization_6 (Batc  (None, 4820)             19280
 hNormalization)

 dropout_6 (Dropout)         (None, 4820)              0

 dense_5 (Dense)             (None, 5000)              24105000

=================================================================
Total params: 28,283,940
Trainable params: 28,274,300
Non-trainable params: 9,640
```

圖 9 DNN 摘要(LF-GENESET_u2000)

2.7 評量預測準確度方法論述

GeneVPNN 研究方法所推論出之擬真微陣列生物晶片基因檢測數值之準確度評估，採用深度學習模型的預測值與實際值之間常見的誤差評量方法，設定以均方誤差 MSE(Mean-Square Error)(6)作為 GeneVPNN 之損失函數(loss function)，以評量模型各項權重值及偏誤值等參數是否調適至最佳化。但對於非屬機器學習或資料科學領域的生醫藥領域專業人士而言，並非能直觀地剖析預測之準確程度，故本研究直接將預測結果資料以程式處理成生醫藥領域常見之預測誤差級別分布直條圖(圖 15、圖 17、圖 19、圖 21)及預測誤差級別之累積預測涵蓋率曲線(圖 16、圖 18、圖 20、圖 22)，讓分析工作更直觀。主要將預測之相對誤差百分比(7)以每 10%作為統計解析誤差之級距，並分別整體統計相對誤差百分比範圍小於 30%及 50%內之資料占比(%)，作為評估研究方法推論準確度之參考。

均方誤差公式：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (6)$$

doi:10.6342/NTU202301211

公式(7)中，Y_i為訓練資料集 614 筆個案之 CV5000 之真實探針數值，$\hat{Y}_i$則為 GeneVPNN 最後階段預測出的微陣列探針數值，n 為實際參與估算的樣本數。

相對誤差百分比公式：

$$Relative\ Error(\%) = \left| \frac{(Y_i - \hat{Y}_i)}{Y_i} \right| \quad (7)$$

公式(7)中，Y_i為測試資料集 69 筆個案之 CV5000 真實基因探針數值，$\hat{Y}_i$則為 GeneVPNN 最後階段預測出對應的 CV5000 微陣列基因探針數值。

公式(7)中，Y_i的真實基因探針數值，試驗分析顧及網路資料庫平台部分的資料集會因實務上微陣列探針量測該基因位點之訊號雜訊比 SNR < 1 時，而將該位點探針量測值設為 0，致使公式(7)分母為零無法給出正確之相對誤差百分比，本研究對於此情境下之相對誤差百分比設計，則將該個案之探針位點 0 值，以測試資料集中所有個案對應之相同探針位點值經排序後取其第二小值且為非 0 之值替代。

doi:10.6342/NTU202301211

第三章 結果

3.1 資料展示

資料集下載及資料前處理工作，主要以 R 語言至 GEO 資料庫平台下載 GSE102484 資料集，並經資料前處理清整等程序後轉存成 CSV 通用資料檔案格式後，轉由後續深度學習模型之 Python 程式處理。GSE102484 資料集摘要，詳如下表所示：

表 1 GSE102484 樣本數摘要

資料集組成	樣本數 N
乳癌_遠端移轉 (event_metastasis=1)	101
乳癌_未遠端移轉 (event_metastasis=0)	582
Total：	683

下圖為原始資料集之資料分布比例情形。

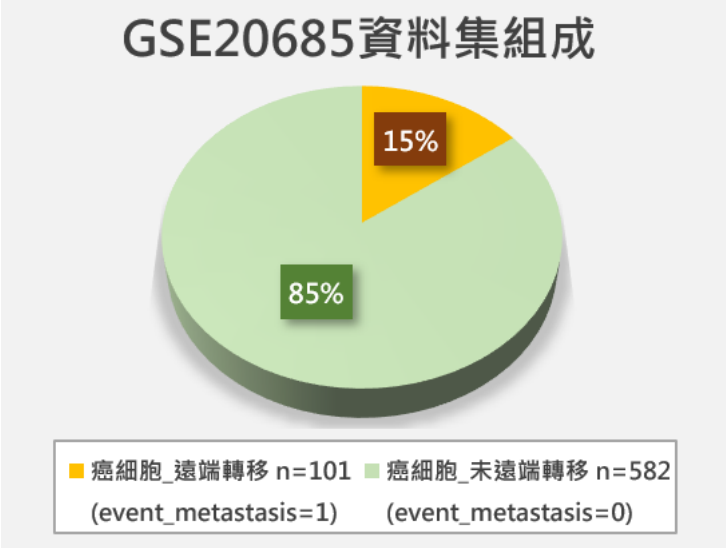

圖 10 GSE102484 資料集組成

本研究方法採用之 Autoencoder 方法屬非監督式機器學習，故製作 AE 模型之資

doi:10.6342/NTU202301211

料集時，將不另行處理標註乳癌遠端移轉與否之標籤於資料集中。但因 Affymetrix 微陣列晶片檢測每個個案探針數值之維度高達 54,627 維，為求降低試驗之模型對非相關基因之干擾，故將 GSE102484 資料集先經統計方法中之變異係數分析法，將原始資料集降維篩選出前 5000 個變異係數 CV 值較高之探針作為 CV5000 資料集，提供給 GeneVPNN 模型作為 AE 訓練之資料集，並作為 GeneVPNN 模型 DNN 預測有效目標基因群的 ground truth 比對來源，如下圖步驟 2 之資料處理程序。

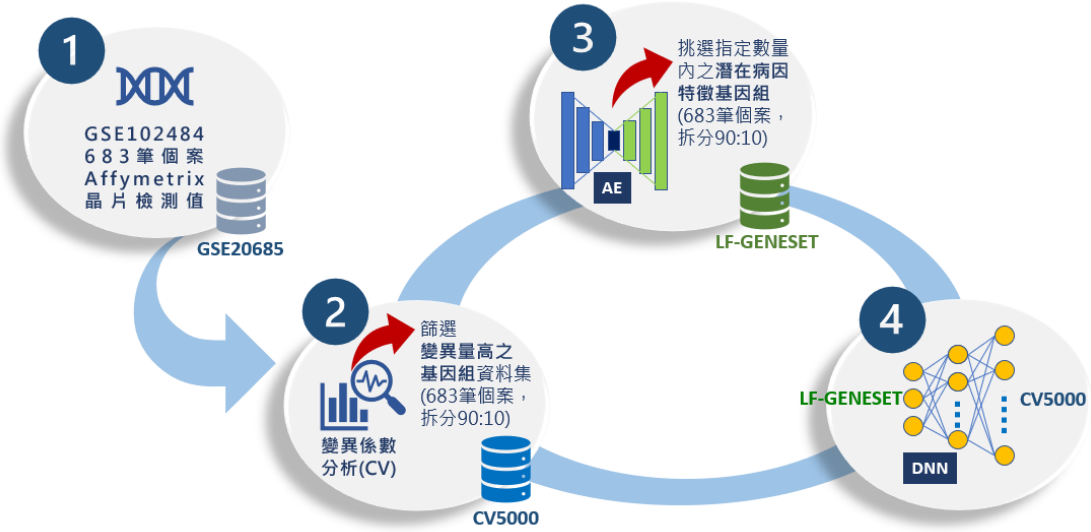

圖 11 GeneVPNN 各階段資料集處理及產出流程

各階段試驗使用之資料集於進入各深度學習模型訓練及預測前，會經由 Python 程式做資料正規化處理，相對地，並於模型預測結果的輸出階段，會將正規化資料再度轉換回原始資料數值以利研究判讀分析。

Autoencoder 模型訓練/測試資料集

AE 模型試驗流程之資料集拆分，應用 Python sklearn 套件之 train_test_split 函數，將已經過變異係數分析法篩選出 CV5000 資料集 shuffle 隨機打亂處理後，以 90:10 比例拆分成訓練資料集與測試資料集。

doi:10.6342/NTU202301211

DNN 模型訓練/測試資料集

DNN 模型試驗流程之資料集組成，主要是依前項 AE 模型訓練後收斂之模型權重值，從變異係數高之 TOP5000 基因組資料集中，再進階挑選指定數量內之潛在特徵基因組 LF-GENESET，如圖 11 步驟 3 所示，並將 683 筆個案樣本同樣經 shuffle 隨機打亂處理後，以 90:10 比例拆分成 DNN 模型之訓練與測試資料集。

GeneVPNN 不論在 AE 或 DNN 模型訓練階段，皆會設定 KERAS 之 validation ratio=0.1，也就是將前階段 sklearn 套件 train_test_split 函數 90:10 比例拆分出的 90% 訓練資料集再另外將其保留 10%作為每一訓練回合(epoch)之 validation 資料集，故前階段 90:10 比例拆分的 10%測試資料集並未參與模型訓練，而是獨立用於模型訓練完畢後公正地測試預測誤差。

GeneVPNN 各試驗階段-輸出/入資料集摘要結果，詳如表 2 所示。

表 2 GeneVPNN 各試驗階段-輸出/入資料集摘要

		Input		***Output***
		Train (validation ratio)	Test	
CV	Cases	683		683
	Features	54,627		5,000
AE	Cases	614 (0.1)	69	614 / 69
	Features	5,000	5,000	Specified
DNN	Cases	614 (0.1)	69	614 / 69
	Features	AE specified probes		5,000

* Specified ：under 200, 500, 1000, 2000 probes

doi:10.6342/NTU202301211

3.2 CV 值試驗方法結果

GSE102484 資料集經 Python 以 CV 變異係數分析後，挑選出 CV 值排序較高基因探針，如圖 12 之 CV 變異係數分析試驗結果摘要所示，圖中的 Indices 為依變異係數 CV 值排序過(由大至小)之 GSE102484 資料集的基因探針索引編號，而 Values 為實際對應之 CV 值以 List 資料型態呈現 CV 值試驗結果。

```
CV-TOP (5000,) Indices: [ 6622 39598 50896 ... 29409 29333 46137]
CV-TOP (5000,) Values: [138.83610472 123.69987699 117.56902152 ...  34.86656728  34.86621044
  34.86152559]
```

圖 12 CV 變異係數分析試驗結果摘要

表 3 為依據 GSE102484 變異係數 CV 值所排序(由大到小)出對應之基因探針名稱簡表。以 CV5000 之 TOP1 至 10 及 TOP4990 至 5000 為例，展示篩選出之 TOP5000 結果摘要。

表 3 GSE102484_CV 值試驗方法挑選 CV5000 基因探針名稱簡表

TOP_1-10		TOP_4991-5000	
TOP_1	243146_at	TOP_4991	226622_at
TOP_2	243337_at	TOP_4992	234314_at
TOP_3	236538_at	TOP_4993	205821_at
TOP_4	243800_at	TOP_4994	226846_at
TOP_5	243520_x_at	TOP_4995	1556284_at
TOP_6	232377_at	TOP_4996	236222_at
TOP_7	243015_at	TOP_4997	235746_s_at
TOP_8	219612_s_at	TOP_4998	1560788_at
TOP_9	243734_x_at	TOP_4999	229391_s_at
TOP_10	214612_x_at	TOP_5000	235278_at

doi:10.6342/NTU202301211

3.3 AE 試驗方法結果

將 CV 變異係數分析後挑選出的 CV5000 資料集，經 AE 試驗程序訓練 400 回合(epoches)後，整體模型預測損失 loss 趨於收斂穩定，並以驗證資料集(Validation dataset)驗證 AE 模型可靠度是否趨近於訓練資料集之預測誤差 loss，下圖為分別展示 AE 試驗階段於選取 LF-GENESET 不超過 200、500、1000、2000 條件下(分別對應下圖 A、B、C、D 子圖)之模型訓練階段與驗證階段之 loss 曲線圖。

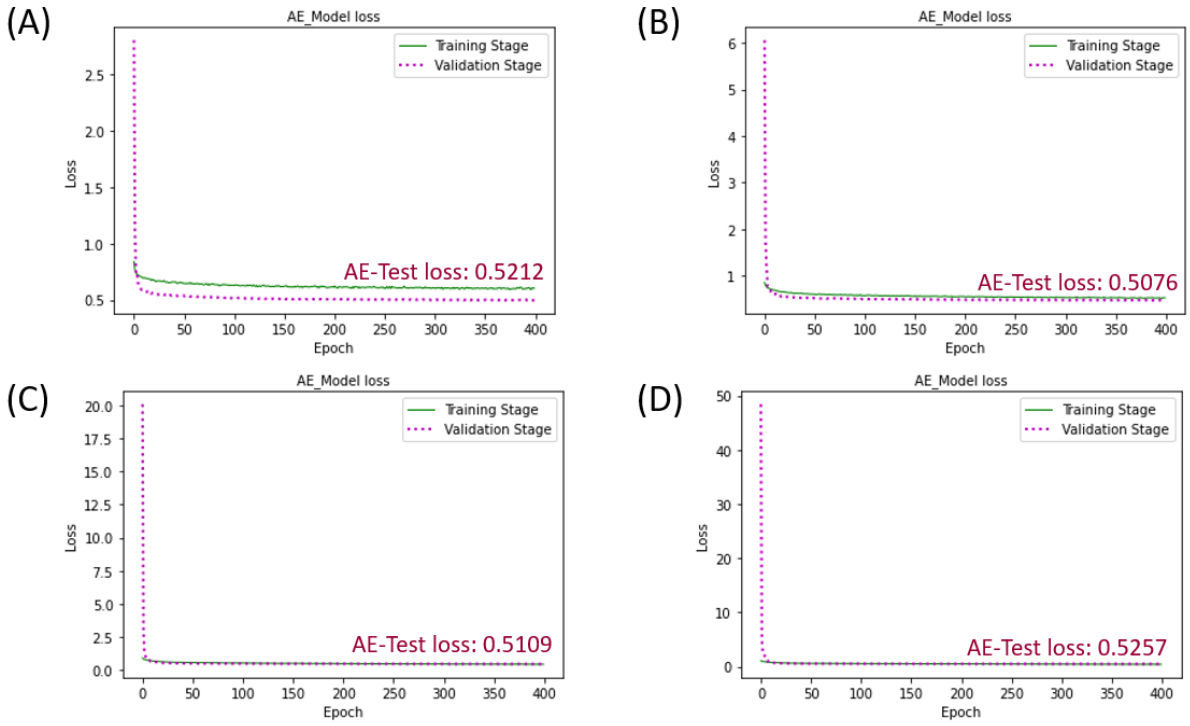

圖 13 GSE102484_AE 試驗階段-Loss 曲線圖

由上圖各測試條件(A)、(B)、(C)、(D)之 Loss 曲線圖觀察得知，在 AE 階段若 LF-GENESET 設定之挑選數量愈多，則相對地 Latent layer 神經元數量也相對增加，故會傳遞給 AE 試驗階段 decoder 更多特徵資訊，因而使 decoder 輸出預測能力更準確。

doi:10.6342/NTU202301211

表 4 所顯示的數據是 AE 模型於不同試驗條件下，所挑選出的基因探針結果。

表 4 GSE102484_AE 試驗-挑選 LF-GENESET 潛在特徵基因組結果

AE 試驗條件	試驗結果	
設定 LF-GENESET 上限	LF-GENESET 組數	Index & Probe name list
Under 200	156	2 236538_at 7 219612_s_at 75 214087_s_at 90 243643_x_at 131 206204_at ... ... 4671 244565_at 4757 1558888_x_at 4825 240188_at 4854 230601_s_at 4876 242901_at
Under 500	341	35 243753_at 90 243643_x_at 120 1553622_a_at 135 243254_at 189 243175_at ... ... 4924 1565602_at 4928 209904_at 4930 242931_at 4981 236219_at 4986 236414_at
Under 1000	560	26 230117_at 29 216238_s_at 39 236308_at 83 219643_at 90 243643_x_at ... ... 4949 1562516_at 4951 223484_at 4958 225660_at 4966 232573_at 4990 226622_at

doi:10.6342/NTU202301211

AE 試驗條件	試驗結果	
設定 LF-GENESET 上限	LF-GENESET 組數	Index & Probe name list
Under 2000	851	9 214612_x_at 17 238047_at 26 230117_at 66 243549_at 73 1562821_a_at ... ... 4966 232573_at 4974 241826_x_at 4986 236414_at 4990 226622_at 4993 226846_at

上表所呈現是 AE 試驗方法結果摘要，表中的 LF-GENESET 基因組數及探針名稱為各試驗條件下之結果，AE 模型依試驗條件設計之上限值(under 200, 500, 1000, 2000 probes)，推薦出不同 LF-GENESET 潛在特徵基因組，其資料集產製乃源自各條件下訓練 AE 模型使其神經元權重收斂後，將 AE 前半部 Encoder 神經元權重值經排序(由大至小)並統合各神經元最高權重後而推薦出。潛在特徵基因組 LF-GENESET 推薦之基因探針數量，主要依試驗條件設計之挑選上限值(under 200, 500, 1000, 2000 probes)而異外，也會因訓練資料集之資料組成差異，即模型訓練後當下權重會有些微變異，而推薦之數量也會有些微差異。

doi:10.6342/NTU202301211

3.4 DNN 試驗方法準確度分析結果

將 AE 試驗階段所挑選出的 LF-GENESET 資料集，經 DNN 模型試驗程序訓練 400 回合(epoches)後，使模型整體預測 CV5000 真實探針值之損失 loss 趨於收斂穩定，在模型訓練各回合中的驗證階段，同樣以 LF-GENSET 驗證資料集(Validation dataset)驗證 DNN 模型可靠度是否趨近於訓練資料集之預測誤差 loss，下圖為分別展示 DNN 試驗階段於選取 LF-GENESET 不超過 200、500、1000、2000 條件下(分別對應下圖 A、B、C、D 子圖)之模型訓練階段與驗證階段之 loss 曲線圖。

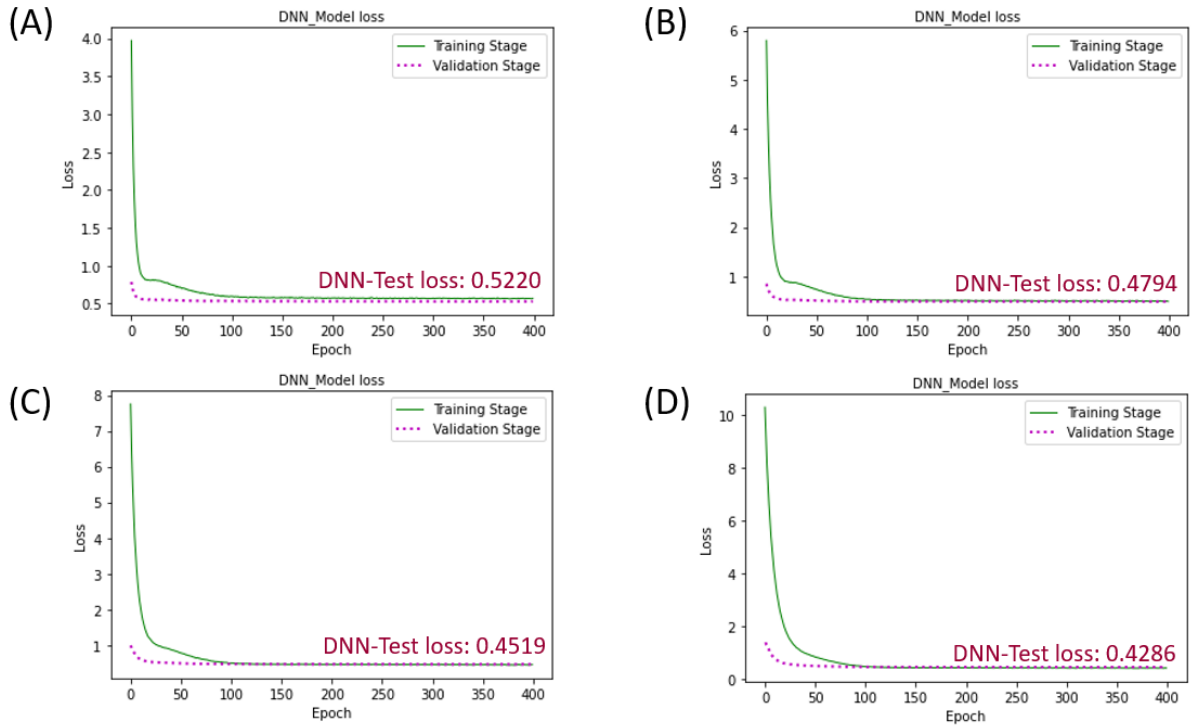

圖 14 GSE102484_DNN 試驗階段-Loss 曲線圖

上圖(A)、(B)、(C)、(D)各試驗條件之 DNN 經由演算最佳化後 Validation 之 Loss 皆已調整趨近 Training 階段之預測誤差 loss，最後載入已訓練完成之模型並將 Test 資料集傳送入模型檢測預測誤差之損失值，皆與訓練階段 loss 相近。

GeneVPNN 模型總體最後階段之 DNN 預測 CV5000 微陣列探針數值結果之評估，使

doi:10.6342/NTU202301211

用未經參與模型訓練的測試資料集(Test dataset)，共計有 69 筆個案之 CV5000 資料作為預測結果分析對照之 Groun truth 來源，即評估預測探針數值共計有 345,000 筆探針數值參與評比；以下各試驗結果，分別為 AE 模型於設定推薦之 LF-GENESET 不超過 200(u200)、不超過 500(u500)、不超過 1000(u1000)及不超過 2000(u2000)個基因探針等試驗條件下，繼表 4 推薦之 LF-GENESET 所訓練之 DNN 模型後，其產製的預測試驗結果圖，分別以各誤差級別之分布直條圖與累積預測涵蓋率曲線分析呈現。

3.4.1 DNN 試驗 1：AE 推薦之 LF-GENESET-under 200

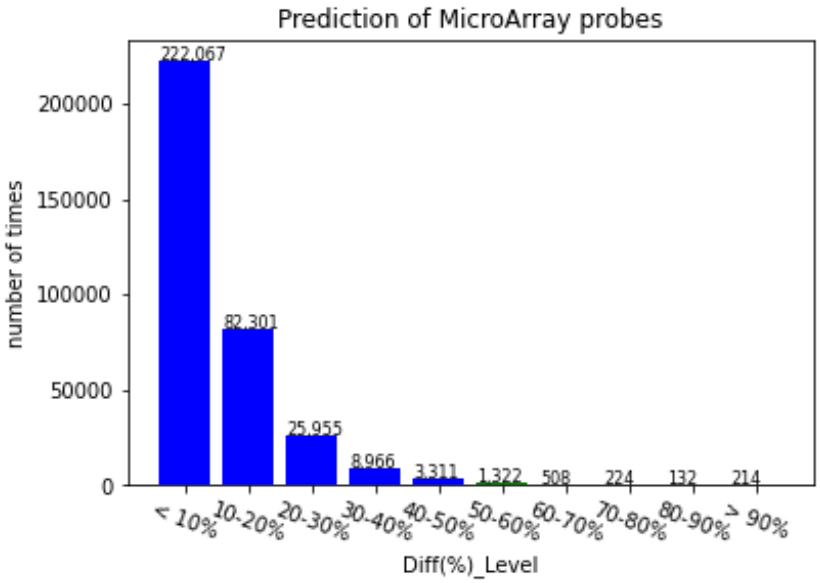

圖 15 DNN LF-GENESET(u200)預測試驗-各誤差級別分布直條圖

doi:10.6342/NTU202301211

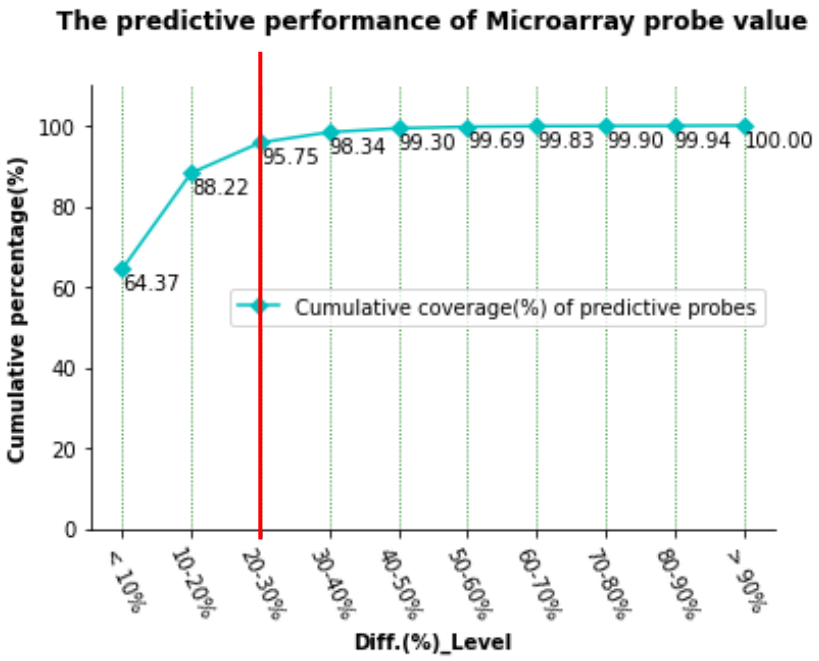

圖 16 DNN LF-GENESET(u200)預測試驗-各誤差級別之累積預測涵蓋率曲線

依上圖 DNN 試驗 1 之預測誤差率統計後，預測基因探針值誤差率小於 30%之數量占比所有預測數量可達 95.75 %；預測基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.30 %。

3.4.2 DNN 試驗 2：AE 推薦之 LF-GENESET-under 500

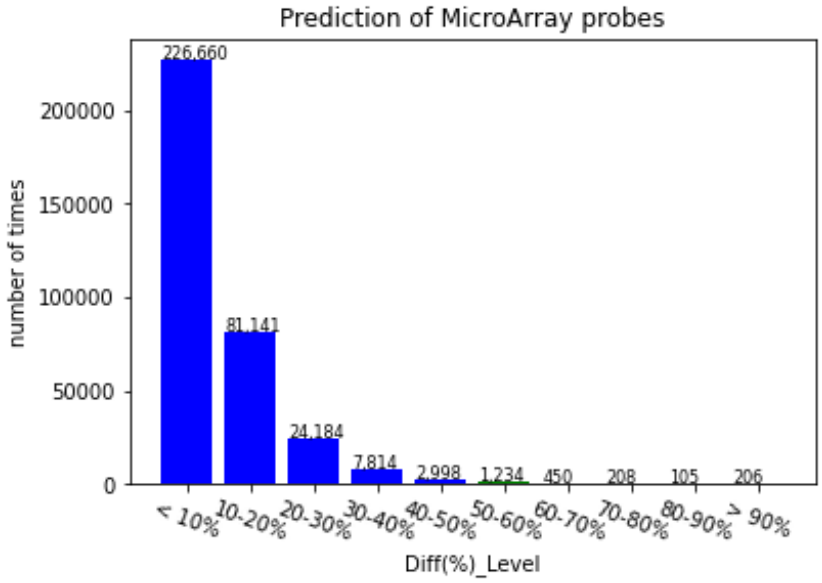

圖 17 DNN LF-GENESET(u500)預測試驗-各誤差級別分布直條圖

doi:10.6342/NTU202301211

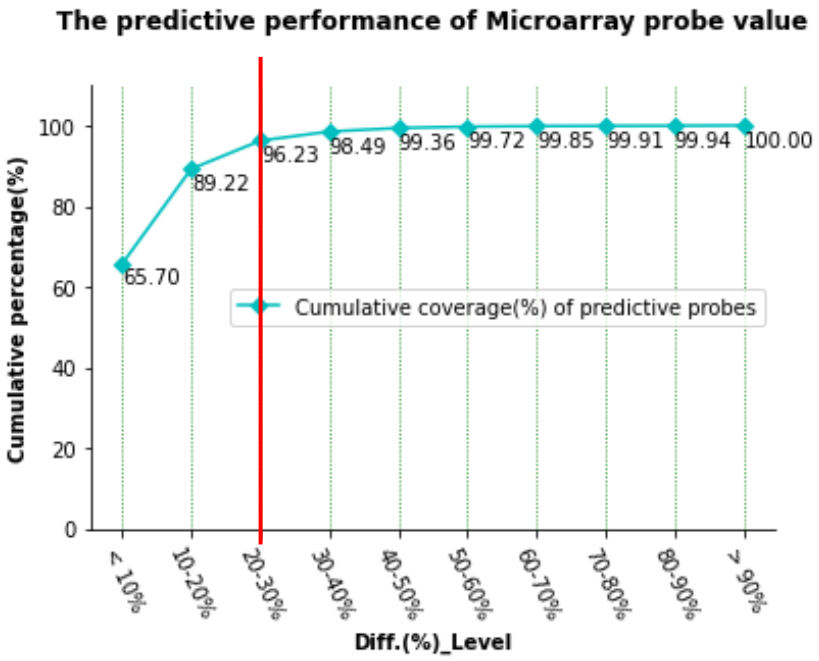

圖 18 DNN LF-GENESET(u500)預測試驗-各誤差級別之累積預測涵蓋率曲線

依上圖 DNN 試驗 2 之預測誤差率統計後，預測基因探針值誤差率小於 30%之數量占比所有預測數量可達 96.23 %；預測基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.36 %。

3.4.3 DNN 試驗 3：AE 推薦之 LF-GENESET-under 1000

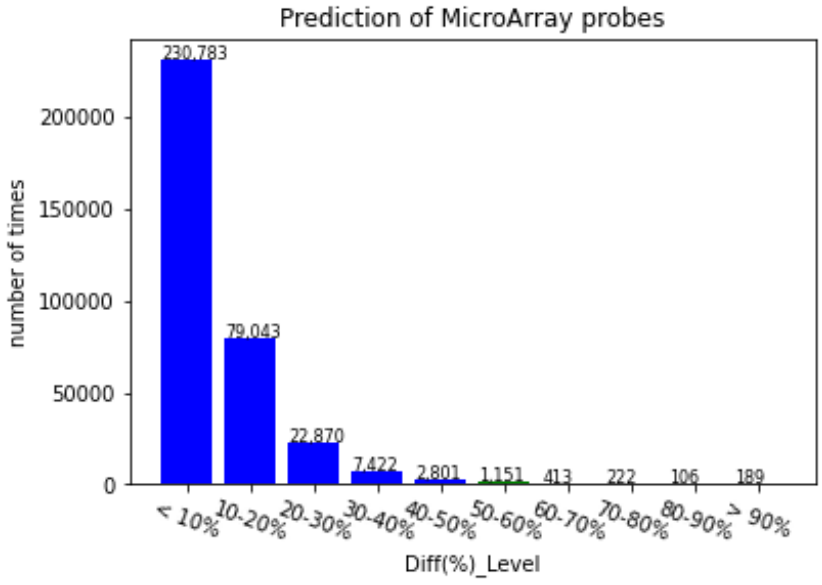

圖 19 DNN LF-GENESET(u1000)預測試驗-各誤差級別分布直條圖

doi:10.6342/NTU202301211

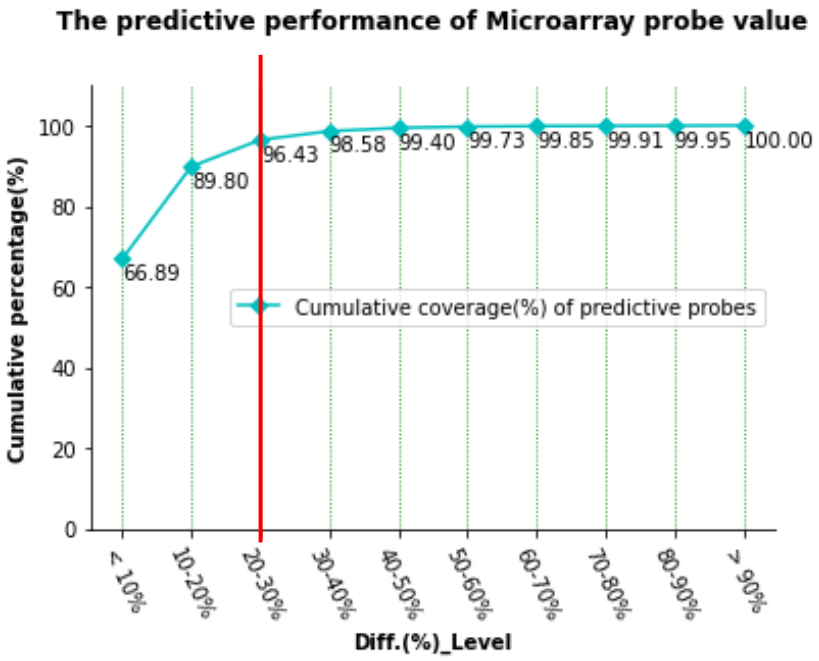

圖 20 DNN LF-GENESET(u1000)預測試驗-各誤差級別之累積預測涵蓋率曲線

依上圖 DNN 試驗 3 之預測誤差率統計後，預測基因探針值誤差率小於 30%之數量占比所有預測數量可達 96.43 %；預測基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.40 %。

3.4.4 DNN 試驗 4：AE 推薦之 LF-GENESET-under 2000

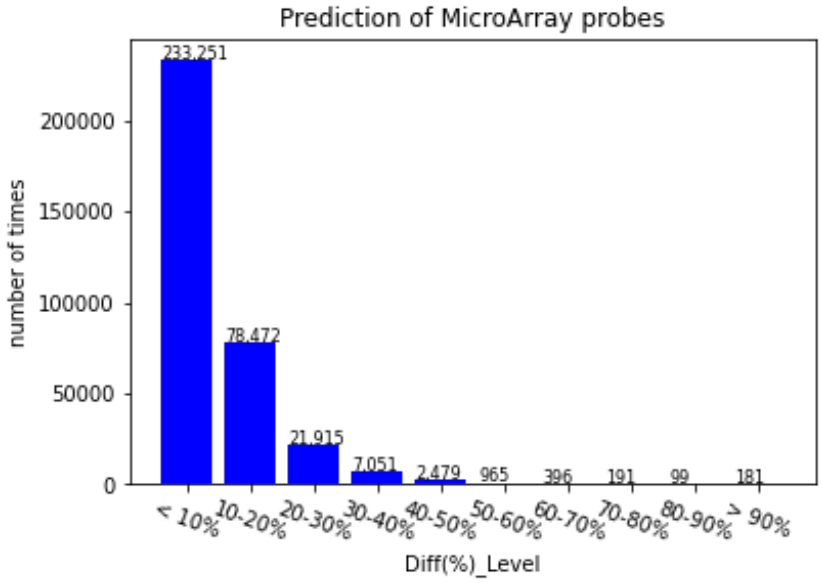

圖 21 DNN LF-GENESET(u2000)預測試驗-各誤差級別分布直條圖

doi:10.6342/NTU202301211

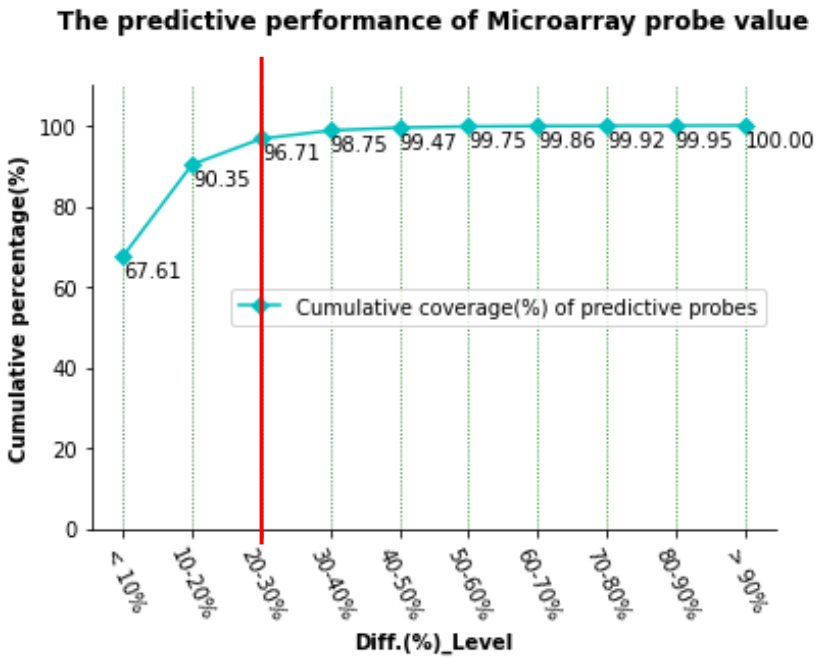

圖 22 DNN LF-GENESET(u2000)預測試驗-各誤差級別之累積預測涵蓋率曲線

依上圖 DNN 試驗 4 之預測誤差率統計後，預測基因探針值誤差率小於 30%之數量占比所有預測數量可達 96.71 %；預測基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.47 %。

3.4.5 DNN 試驗結果分析

統合以上 DNN 各條件試驗，歸納出以下相對誤差率占比分析表，由分析結果可觀察出不論是誤差率<30% 或 <50%之預測涵蓋占比，皆隨試驗條件 AE 推薦之 LF-GENESET 組數逐步增加而遞增。

表 5 DNN 試驗結果摘要

	LF-GENESET 組數	*誤差率< 30%占比*	*誤差率< 50%占比*
DNN 試驗 1	under 200	95.75%	99.30%
DNN 試驗 2	under 500	96.23%	99.36%
DNN 試驗 3	under 1000	96.43%	99.40%
DNN 試驗 4	under 2000	96.71%	99.47%

doi:10.6342/NTU202301211

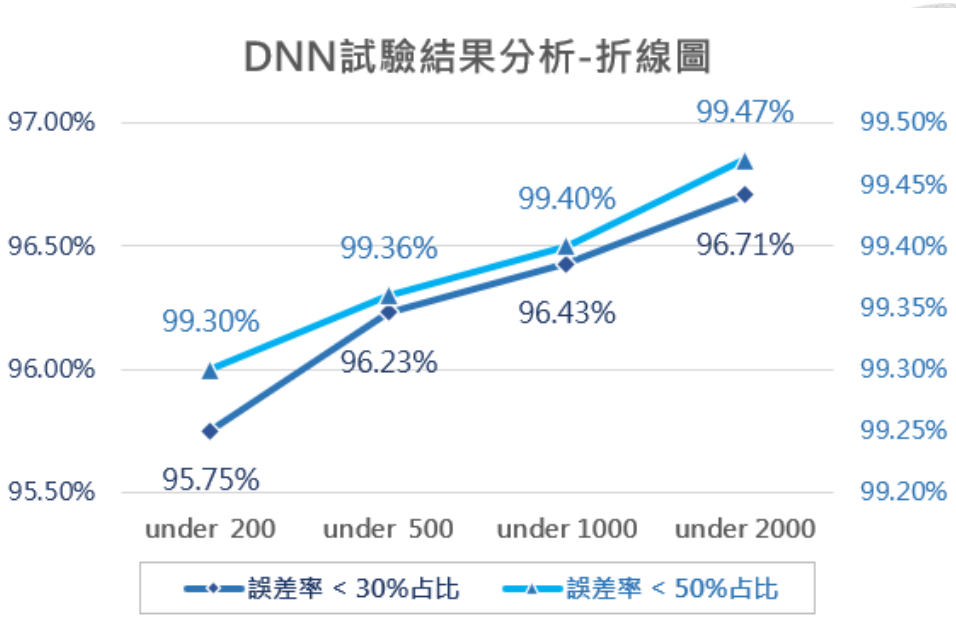

圖 23 DNN 試驗結果分析-折線圖

由上圖 DNN 試驗結果分析-折線圖可觀察出，不論是誤差率 < 30%占比 或是 於誤差率 < 50%占比，皆呈現同步正向遞增趨勢，故在基因檢測研究預算允許下，本試驗建議以不高於檢測基因組數 2000 前提下，進行本研究設計架構之模型流程，做最經濟性及前瞻性基因預測。

3.4.6 GeneVPNN 對不同族群預測試驗結果比較分析

本研究試驗資料 GSE102484 資料集由乳癌遠端移轉 metastasis (Y_event=1)及乳癌遠端未移轉(Y_event=0)兩族群提供之組織切片基因表現量資料所組成，為了進一步分析了解 GeneVPNN 針對不同族群的預測能力，本項試驗設計以挑選 LF-GENESET 不超過 2000 探針數的條件下，統計分析兩族群之 GeneVPNN 預測能力。

總體預測結果分析

圖 24、圖 25 所示，其為以測試資料集包含兩族群之 CV5000 總體預測結果，試驗執行方法以 Random seed=123 將 683 筆樣本數之 CV5000，以 90:10 比例依樣本個案拆分出訓練(614 筆)及測試(69 筆)資料集，並經 AE 階段推薦出 LF-

doi:10.6342/NTU202301211

GENESET 共計有 879 組潛在特徵基因組，依據此 LF-GENESET，經 DNN 階段試驗預測出 CV5000，並分析 GeneVPNN 對各族群預測能力，即在測試資料 69 筆樣本數下，預測總探針數共計達 345,000 probes，並記錄下各級預測誤差之數量及各誤差級別之累積預測涵蓋率曲線，如後所示。

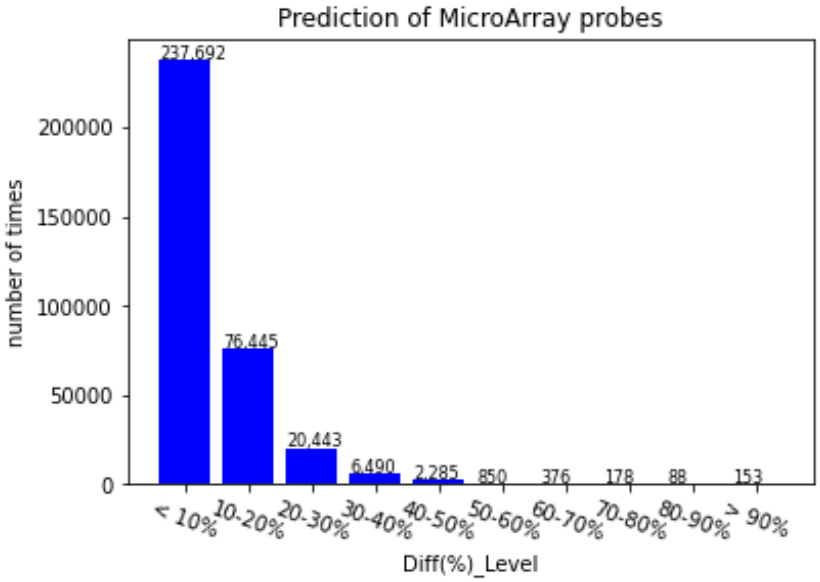

圖 24 GeneVPNN 不同族群總體預測結果- 各誤差級別分布直條圖

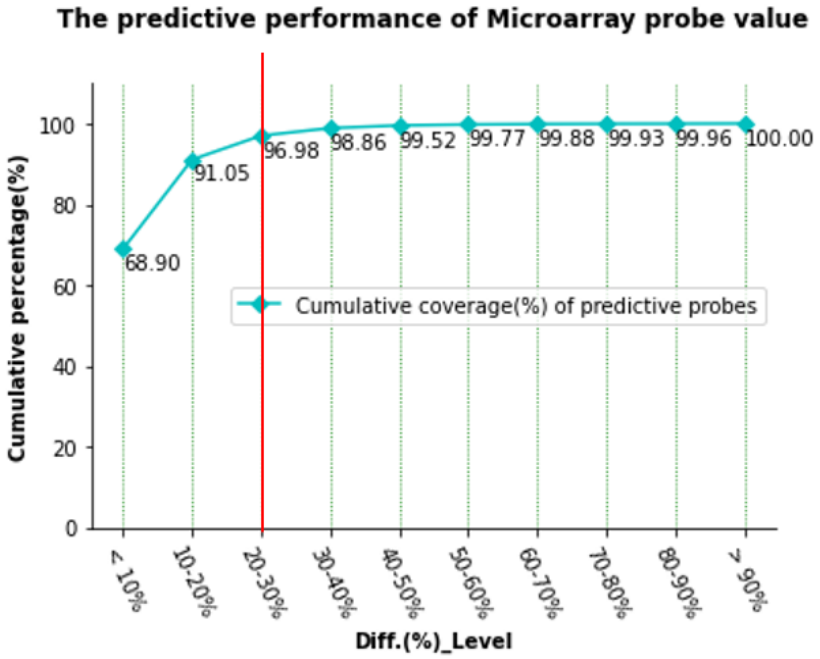

圖 25 GeneVPNN 不同族群總體預測結果- 各誤差級別之累積預測涵蓋率曲線

doi:10.6342/NTU202301211

依各族群資料預測結果分析

承上兩族群 CV5000 之總體預測結果，以下將進一步分析乳癌遠端未移轉(Y_event=0)與遠端移轉(Y_event=1)各自之預測能力。在測試 69 筆 CV5000 的測試資料集中，遠端未移轉(Y_event=0)的族群有 61 筆，即共計有 305,000 probes，遠端移轉(Y_event=1)的族群有 8 筆，即共計有 40,000 probes 被預測分析，各族群預測結果如下所示。

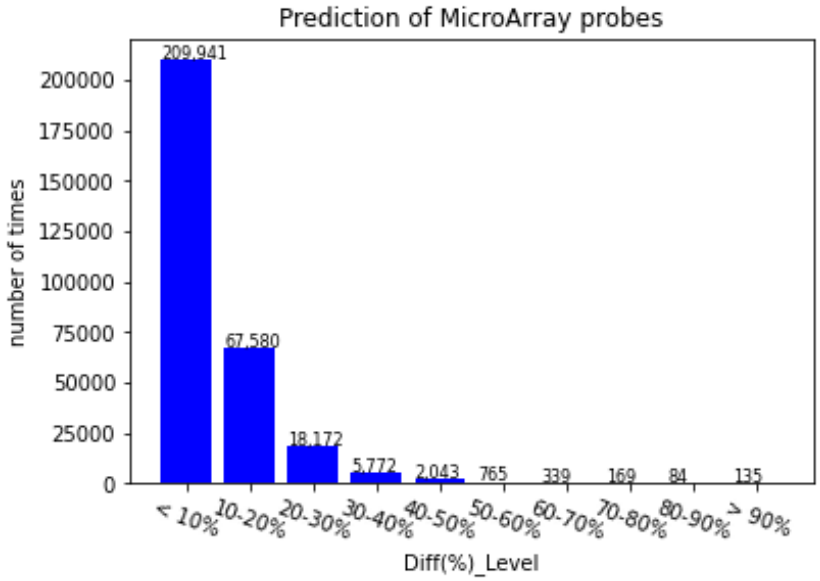

圖 26 GeneVPNN 對乳癌遠端未移轉族群預測- 各誤差級別分布直條圖

doi:10.6342/NTU202301211

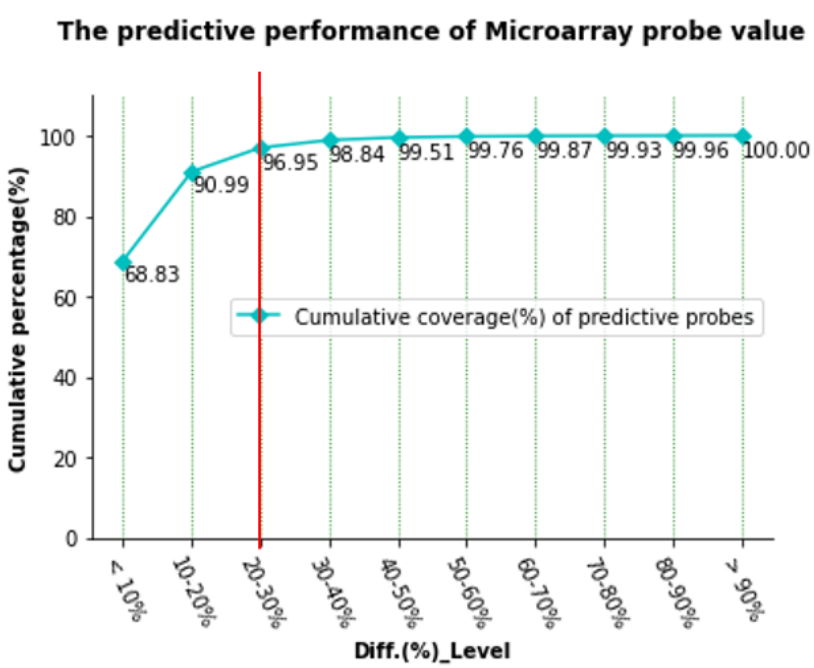

圖 27 GeneVPNN 對乳癌遠端未移轉族群預測- 各誤差級別累積預測涵蓋率曲線

針對乳癌遠端移轉(Y_event=1)族群的預測結果，如後所示。

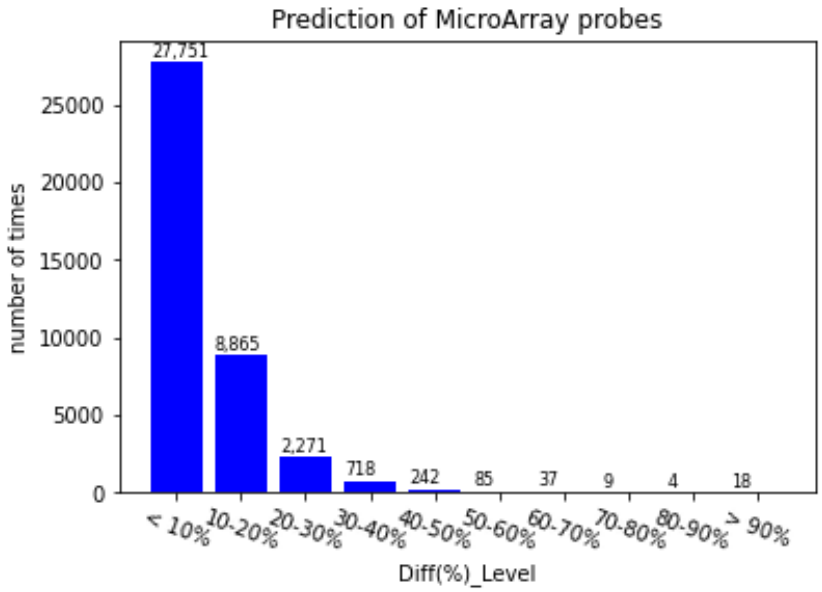

圖 28 GeneVPNN 對乳癌遠端移轉族群預測- 各誤差級別分布直條圖

doi:10.6342/NTU202301211

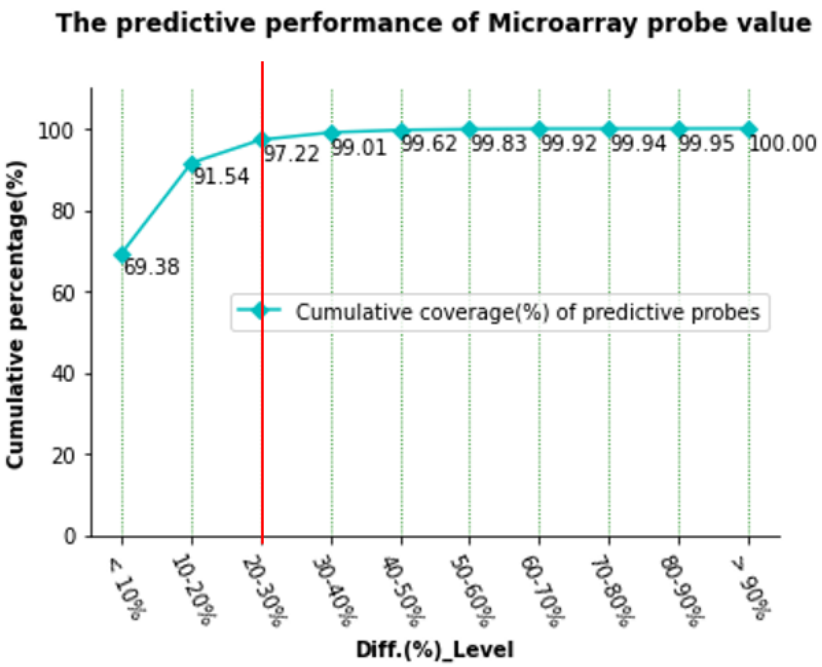

圖 29 GeneVPNN 對乳癌遠端移轉族群預測- 各誤差級別累積預測涵蓋率曲線

試驗分析小結

統整 GeneVPNN 對不同族群預測試驗結果比較，將各族群試驗結果彙整如下表。

表 6 GeneVPNN 對不同族群預測試驗結果比較

	總體預測	乳癌遠端未移轉 (Y_event=0)	乳癌遠端移轉 (Y_event=1)
預測總探針數 (A)	345,000	305,000	40,000
預測誤差< 30% 之累計探針數 (B)	334,580	295,693	38,887
預測誤差< 30% 之累計涵蓋率 (B/A)	**96.98 %**	**96.95 %**	**97.22 %**

由表 6 之預測結果分析可知 GeneVPNN 對乳癌遠端未移轉族群之 CV5000 預測誤差小於 30%之累計涵蓋率為 96.95%，而對乳癌遠端移轉族群之 CV5000 預測誤差小於 30%之累計涵蓋率為 97.22%，此 GeneVPNN 對兩族群之預測力皆與總體預測 96.98%差異不大，故 GeneVPNN 對兩族群預測表現是一致的。

doi:10.6342/NTU202301211

3.5 GeneVPNN 總體可靠度檢定

3,5.1 模型設計正確性檢定

試驗假設

針對模型設計正確性檢定，此試驗假設 GeneVPNN 具有挑選關鍵之潛在特徵基因 LF-GENESET 能力，並依此具代表性之基因樣本可推論出較佳之 CV5000 預測效果，相反地，若給予模型的資料集若不具病因相關之數據，就算模型設計正確理應受到無相關之資料影響其預測力應不如前者佳，在此試驗假設前提下進行試驗及檢定觀測結果。

試驗設計

將 GSE102484 的 CV5000 有效目標基因群資料集改採亂數模擬 CV5000 資料集，並對模型預測 CV5000 誤差分布進行分析。為求模擬合理之基因表現量，試驗針對 GSE102484 資料集進行總體表現量質分析，其探針數值介於 2 至 16 間，故以程式取亂數小數值介於此數值區間，以組成(683, 5000)維度之亂數 CV5000 資料集。

試驗結果

試驗結果之數據以圖 30 及圖 31 展示亂數資料集檢定 GeneVPNN 預測 CV5000 誤差之涵蓋占比分析，並與真實 CV5000 資料集之試驗結果進行對照比較。

doi:10.6342/NTU202301211

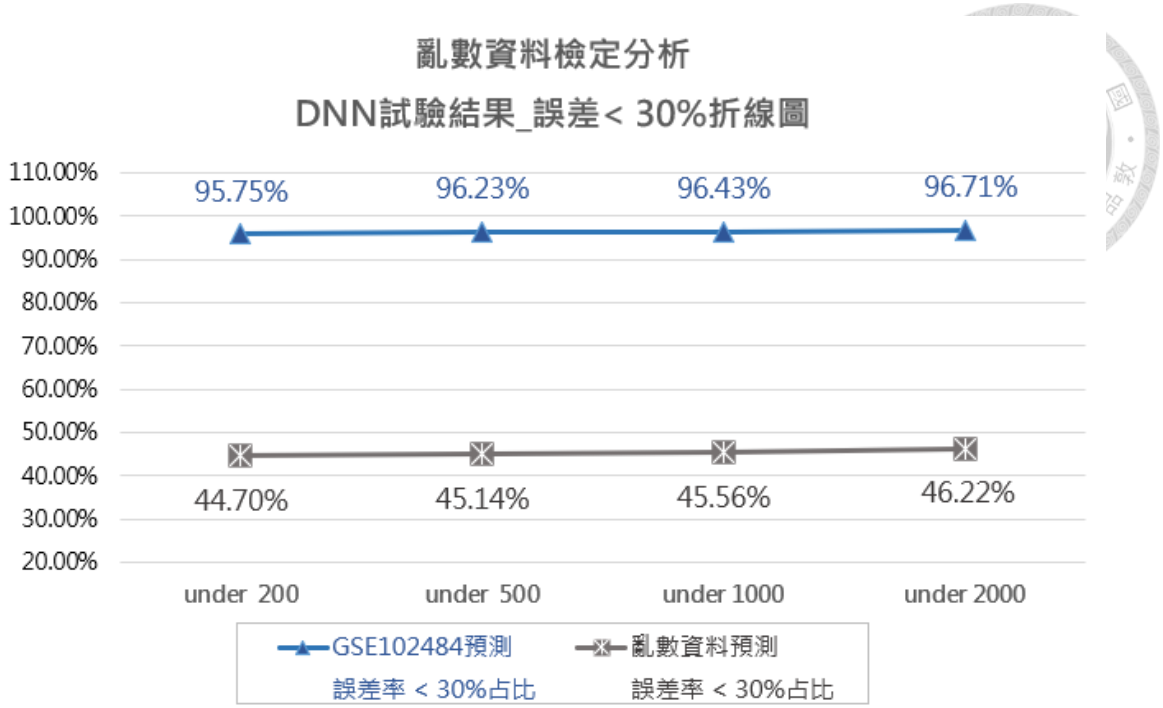

圖 30 亂數資料集檢定 GeneVPNN 預測 CV5000 誤差< 30%占比分析

圖 30 是以本研究相同的 GeneVPNN 設計架構，以亂數 CV5000 資料集分別檢定在 LF-GENESET 限定推薦不超過 200、500、1000、2000 四種不同試驗條件下，統計預測亂數 CV5000 誤差< 30%占比 CV5000 的涵蓋率(圖中的灰色線段)，並與 GSE102484 真實 CV5000 資料集所訓練之模型的預測誤差占比(圖中的藍色線段)進行比較。圖 31 另外也針對兩者誤差< 50%占比 CV5000 涵蓋率進行比較。

doi:10.6342/NTU202301211

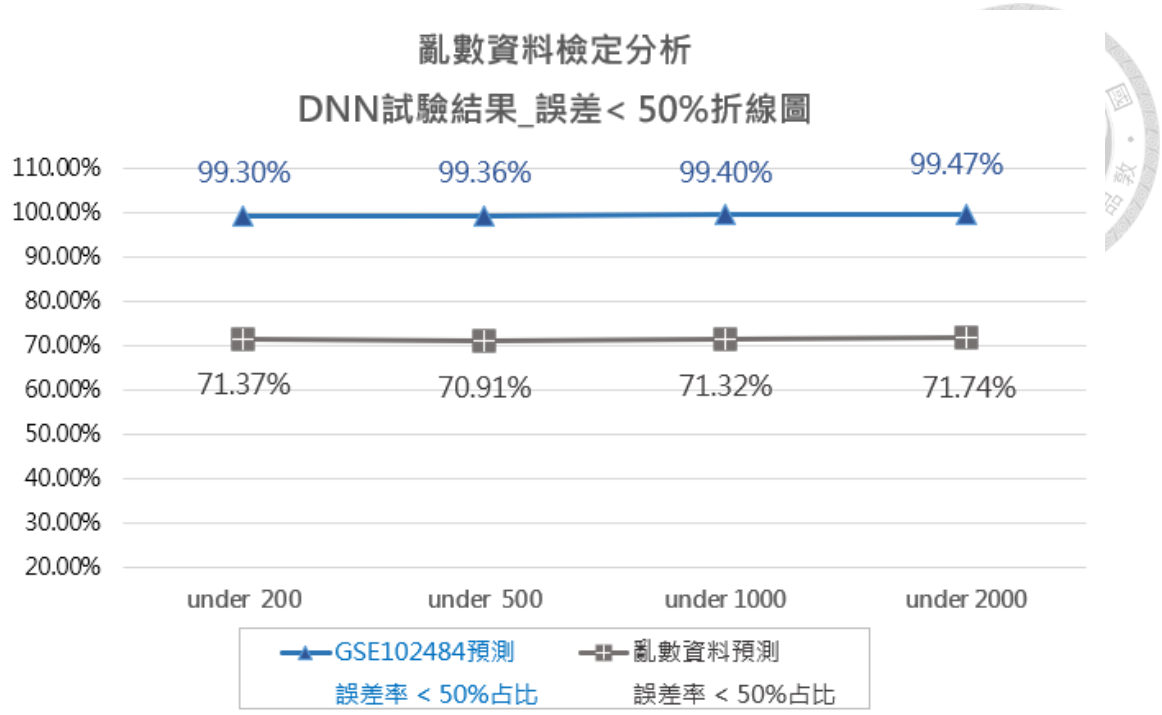

圖 31 亂數資料集檢定 GeneVPNN 預測 CV5000 誤差< 50%占比分析

試驗結論

從本項模型設計正確性檢定之結果剖析，因研究之 GSE102484 資料集具有真正代表病因之基因表現量存在，故 GeneVPNN 採用真實的 CV5000 所推薦出的 LF-GENESET 基因組具有篩選一定程度代表性潛在特徵基因能力，作為 GeneVPNN 預測 CV5000 推估運算之樣本來源，試驗結果顯示出其預測誤差< 30%皆達 96%以上涵蓋占比。以病因無相關之亂數 CV5000 檢定模型，其預測誤差< 30% 僅可達 46%左右涵蓋占比，其能力不如真實數據來得佳。故符合試驗立論，即模型依研究目標設計正確。

doi:10.6342/NTU202301211

3.5.2 預測穩定性檢定

試驗假設

本項試驗設計最主要檢定模型之預測穩定性，不會因資料取樣差異而使預測能力偏差過大，也就是期望透過試驗來檢定 GeneVPNN 在 GSE102484 不同取樣拆分的訓練資料及測試資料集前提下，經過本研究的模型設計程序所產出的預測表現，都應該呈現穩定之預測效果。

試驗設計

同樣應用圖 2 中的程序 3 所產製的 CV5000 有效目標基因群資料集，以 Python sklearn 套件設定 10 次不同的亂數因子，將 683 筆個案之 CV5000，依亂數因子挑選出個案之 CV5000，並從 683 筆個案以 90:10 比例拆分訓練(614 筆)及測試(69 筆)資料集，依此試驗規則，由 10 次不同的亂數因子組成 10 份不同內容之訓練及測試資料集，提供給 GeneVPNN 做預測穩定性檢定，且設計各批次試驗應用圖 2 中的程序 4、5 推薦之 LF-GENESET 在設定以推薦不超過 2000 個潛在特徵基因之條件下執行 10 次資料交叉檢定模型，並推論之預測 CV5000 試驗結果紀錄分析。

試驗結果

表 7 為 10-Fold cross validation 試驗 10 次之試驗結果紀錄表，並將各次預測誤差占比 CV5000 基因探針涵蓋率以折線圖方式呈現於圖 32。

表 7 GeneVPNN 10-fold cross-validation result

Random seed of split data	Number of LF-GENESET	Under 30% diff. cover rate (%)	Under 50% diff. cover rate (%)
111	856	96.74%	99.48%
222	830	96.71%	99.46%
333	874	96.83%	99.50%
444	846	96.76%	99.46%

doi:10.6342/NTU202301211

Random seed of split data	Number of LF-GENESET	Under 30% diff. cover rate (%)	Under 50% diff. cover rate (%)
555	837	96.78%	99.45%
666	863	96.67%	99.44%
777	911	96.81%	99.49%
888	865	96.78%	99.48%
999	877	96.87%	99.50%
678	859	96.73%	99.48%

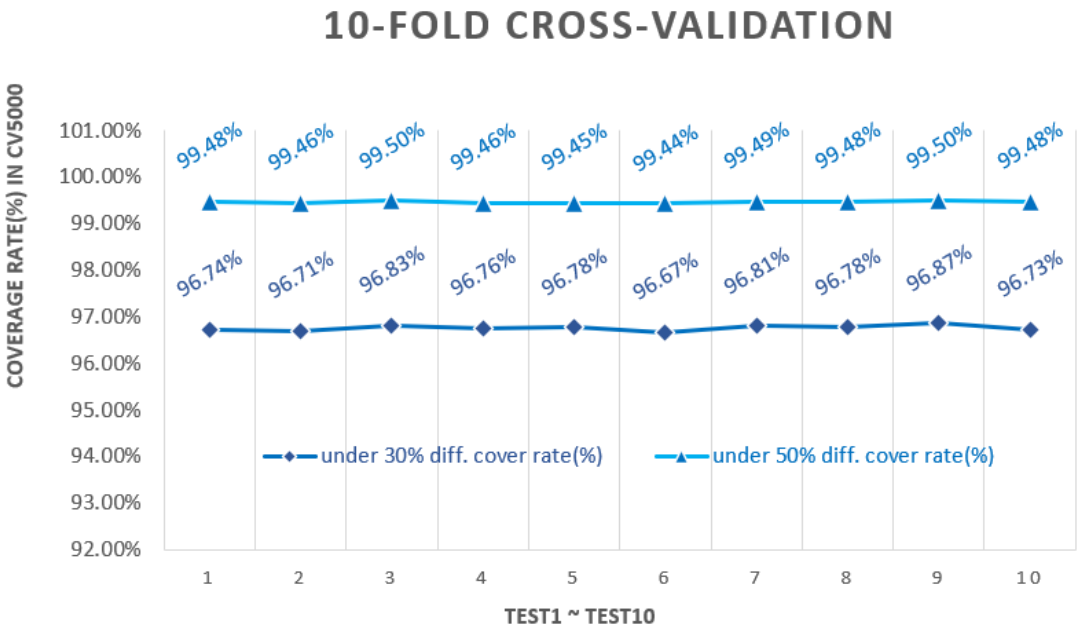

圖 32 模型穩定性檢定分析(10-fold cross-validation)

試驗結論

試驗結果依據試驗假設目標檢定分析，在 10 次的交叉驗證模型預測表現如同本研究章節 3.4.4 GeneVPNN 之 DNN 試驗 4 在 AE 推薦 LF-GENESET under 2000 的試驗結果，在預測基因探針值誤差率小於 30%之數量占比所有預測數量可達 96.6 %左右；預測基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.45 %，故本研究之 GeneVPNN 模型具有預測穩定之表現。

doi:10.6342/NTU202301211

3.6 GeneVPNN 試驗方法適用非乳癌疾病泛用性檢定

試驗假設

鑑於 GeneVPNN 研究之基礎原理是經由 CV 演算法先降維篩選出研究族群基因較具表現量之探針資料 CV5000，並經進階 AE 降維演算法推薦出更具代表 CV5000 的潛在特徵基因組 LF-GENESET，生醫研究單位即可應用 LF-GENESET 執行較為經濟快速之基因檢測，達成預測研究族群的有效目標基因群 CV5000 之階段研究任務。然而本研究模型設計是在未針對任何複雜專科疾病為專屬應用的假設前提下所規劃出，期以泛用於任何複雜疾病及人種。故本項泛用性檢定，期望藉由 GEO 資料平台提供之非乳癌疾病及非亞洲人種 GSE 研究資料集進行泛用性驗證。

試驗設計

本項檢定採用 GSE37745 資料集，其研究對象為北歐瑞典人種的非小細胞肺癌（Non-small cell lung cancer, NSCLC）復發與否之研究資料，其包含 196 個於 1995 年至 2005 年癌症治療術後之追蹤研究樣本[22]。其樣本數據是經由 Affymetrix microarrays HG-U133-Plus2 基因微陣列晶片於 GPL570 平台分析後紀錄組成資料集。試驗則將資料集依 90:10 比例拆分成訓練資料集及測試資料集，其中訓練資料集於模型各批次訓練期間再依 90:10 比例保留 10%作為訓練期間優化器參考誤差損失 loss 之 validation 資料集來源，而模型泛用性檢定程序仍依照 圖 2 GeneVPNN 模型研究試驗流程的第 1 階段至第 7 階段進行驗證程序，AE 推薦之 LF-GENESET 設定挑選不超過 2000 個基因探針為試驗條件。

試驗結果

表 8 為依據 GSE37745 變異係數 CV 值所排序(由大到小)出對應之基因探針名稱簡表。以 CV5000 之 TOP1 至 10 及 TOP4990 至 5000 為例，展示篩選出之 TOP5000

doi:10.6342/NTU202301211

結果摘要。

表 8 GSE37745_CV 值試驗方法挑選 CV5000 基因探針名稱簡表

TOP_1-10		TOP_4991-5000	
TOP_1	209988_s_at	TOP_4991	212806_at
TOP_2	235075_at	TOP_4992	228455_at
TOP_3	205649_s_at	TOP_4993	209228_x_at
TOP_4	217561_at	TOP_4994	1556429_a_at
TOP_5	211361_s_at	TOP_4995	211648_at
TOP_6	217528_at	TOP_4996	211564_s_at
TOP_7	1567912_s_at	TOP_4997	244015_at
TOP_8	216238_s_at	TOP_4998	223785_at
TOP_9	205625_s_at	TOP_4999	215307_at
TOP_10	206165_s_at	TOP_5000	202756_s_at

承前一階段執行變異係數 CV 值分析產製出有效目標基因群 CV5000，圖 36(A)為應用 CV5000 訓練 GeneVPNN 模型之 AE 最佳化後的 loss 曲線圖與圖 36(B)訓練 GeneVPNN 模型之 DNN 最佳化後的 loss 曲線圖，不論 AE 或 DNN 皆經由演算最佳化後訓練期 Validation 階段之 loss 皆已調整趨近 Training 階段之 Loss，最後載入已訓練完成之模型並將 Test 資料集傳送入模型檢測預測誤差之損失值，AE 測試階段 loss 為 0.6155，DNN 測試階段 loss 為 0.5203，皆與訓練階段 loss 相近。

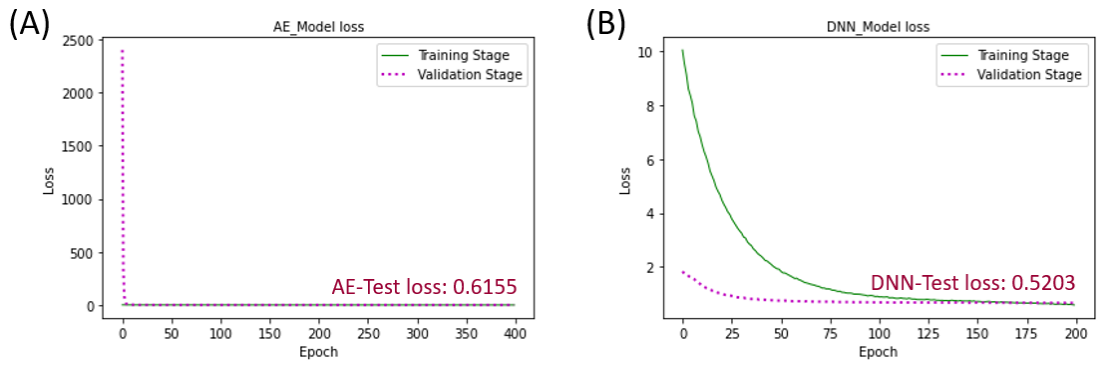

圖 33 GSE37745_AE & DNN 試驗階段-Loss 曲線圖

doi:10.6342/NTU202301211

下表所顯示的數據是 AE 模型於 LF-GENESET 設定挑選不超過 2000 個基因探針的試驗條件下，所挑選出的基因探針結果。

表 9 GSE37745_AE 試驗-挑選 LF-GENESET 潛在特徵基因組結果

AE 試驗條件	試驗結果	
設定 LF-GENESET 上限	LF-GENESET 組數	Index & Probe name list
Under 2000	1151	6 1567912_s_at 13 224590_at 14 209987_s_at 15 214254_at 19 214218_s_at ... ... 4983 205522_at 4988 1556021_at 4993 1556429_a_at 4994 211648_at 4995 211564_s_at

經本試驗測試資料集 20 筆樣本，每筆樣本有 5000 個基因探針數值(CV5000)作為 ground truth 檢定後，GeneVPNN 預測 CV5000 之誤差分布占比直條圖與累積分布占比曲線圖如下圖所示。

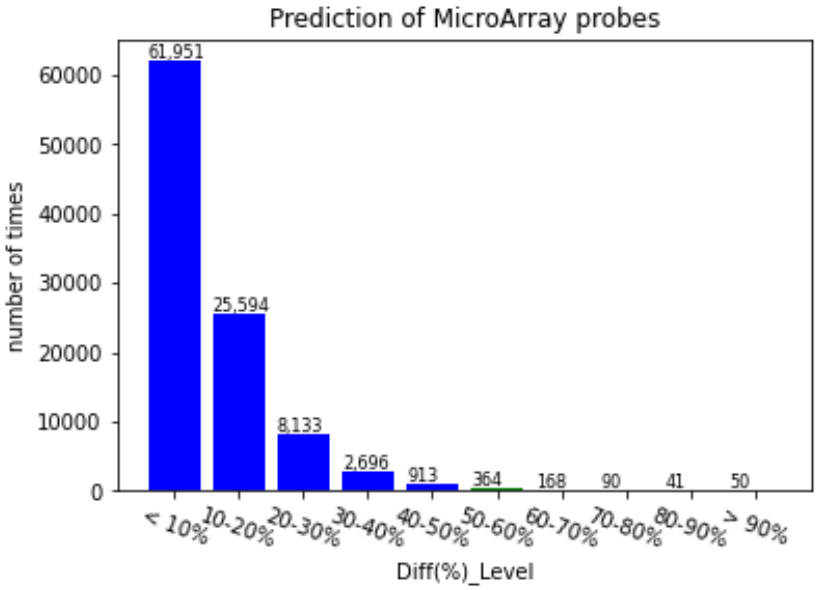

圖 34 GSE37745_DNN LF-GENESET(u2000)預測試驗-各誤差級別分布直條圖

doi:10.6342/NTU202301211

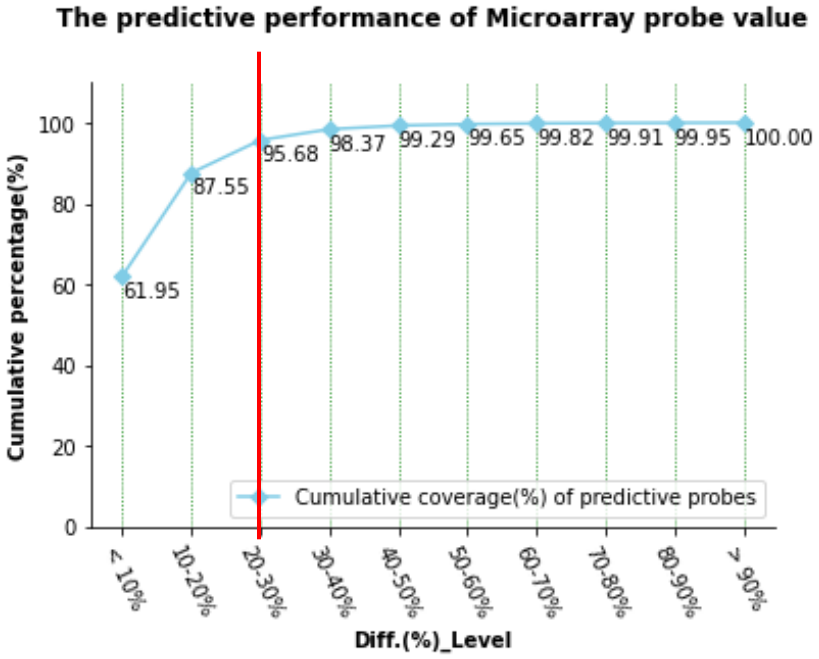

圖 35 GSE37745_DNN LF-GENESET(u2000)預測試驗-各誤差級別之累積預測涵蓋率曲線

依上圖 DNN 試驗之預測誤差率統計後，預測基因探針值誤差率小於 30%之數量占比所有預測數量可達 95.68 %；預測基因探針值誤差率小於 50%之數量占比所有預測數量可達 99.29 %。

試驗結論

由表 10 試驗結果分析得知使用 GSE37745 資料集訓練之 GeneVPNN 模型預測力，不論其預測基因探針值誤差率小於 30%之數量占比，還是預測基因探針值誤差率小於 50%之數量占比，皆與 GSE102484 資料集訓練之 GeneVPNN 模型預測力差距約 1%，故也為 GeneVPNN 驗證了可適用於非乳癌及非亞洲人種之研究泛用性。

doi:10.6342/NTU202301211

表 10 GSE37745 & GSE102484 GeneVPNN 預測誤差累計占比分析表

誤差級別	GeneVPNN of GSE102484 (A)	GeneVPNN of GSE37745 (B)	預測力差距 (A)-(B)
預測基因探針值誤差率小於 30%之 CV5000 占比	96.71 %	95.68 %	1.03 %
預測基因探針值誤差率小於 50%之 CV5000 占比	99.47 %	99.29 %	0.18 %

doi:10.6342/NTU202301211

第四章 結論與討論

4.1 主要發現

4.1.1 GeneVPNN 模型可實現基因預測工作

依據本研究設立之假說，有效目標基因群 CV5000 之微陣列基因檢測值可經由統計科學方法及深度學習演算篩選出指定數量之 LF-GENESET 潛在特徵基因組做 CV5000 推估，經試驗設計架構出的 GeneVPNN 基因虛擬探針檢測模型推論證實，被推估的有效目標基因群 CV5000 探針檢測數值，在推薦 LF-GENESET 以不超過 2000 基因探針數量條件下，其相對誤差小於 50%的基因探針數占總推估總數約 99 %，更進階統計於縮小相對誤差範圍至小於 30%之基因探針數則占總推估總數約 96 %，對於目前複雜之基因預測研究工作提供具有簡易、經濟、快速等特性的科學推估方法。

4.1.2 GeneVPNN 之 AE 特徵基因挑選方法具有效降維能力

人體基因研究領域已知的基因數量超過 2 萬，然則要在數以萬計的基因中剔除與研究族群非相關之基因或重複類似相同功能基因，實為模型訓練發展前期關鍵重要工作，在 GeneVPNN 降維最終階段是由 AE 產製潛在特徵基因組 LF-GENESET，為瞭解其降維關連與 CV5000 基因間相關性之分布情形，故本研究針對 LF-GENESET 與 CV5000 各個皮爾森相關係數[23]級別探針做交集試驗進行分布分析。表 11 依相關係數級別分別統計 LF-GENESE 交集 CV5000 各相關係數級別基因探針占比 LF-GENESET 分布情形，表中的 LF-GENESET 是經由亂數因子 777 拆分之資料集所訓練的 GeneVPNN 推薦之探針組，以挑選不超過 2000 個探針為試驗條件，其結果由 911 個潛在特徵基因所組成，而相關係數分析不重複探針數統計是由 CV5000 先進行各探針對(pair)皮爾森相關係數分析，CV5000 共組合出 2,5000,000 個分析結果，再將相關係數分析表左下直角三角形區重複分析探針

doi:10.6342/NTU202301211

對係數及左上至右下對角線上探針本身相關係數為 1 的探針對係數剔除，保留右上角直角區內之探針對係數，再進行相關係數分級並將各級所屬探針對取不重複探針名統計其數量，即為表中之相關係數分析不重複探針數欄位所對應之數值。

表 11 GeneVPNN LF-GENESET 交集 CV5000 各相關係數級別探針分布摘要

相關係數	*相關係數分析不重複探針數(A)*	*LF-GENESET 探針數(B)*	*(B)交集(A) Hit Probes (C)*	*Hit Rate (%) (C) / (B)*
0.9	1895	911	251	27.55%
0.8	1930	911	255	27.99%
0.7	2042	911	247	27.11%
0.6	2497	911	298	32.71%
0.5	3359	911	436	47.86%
0.4	4183	911	634	69.59%
0.3	4809	911	854	93.74%
0.2	4980	911	908	99.67%
L0.2	5000	911	911	100.00%

依表 11 之數據將 LF-GENESET 與 CV5000 各相關係數級別探針交集分布情形，以圖 36 將其分布統計數據視覺化呈現。

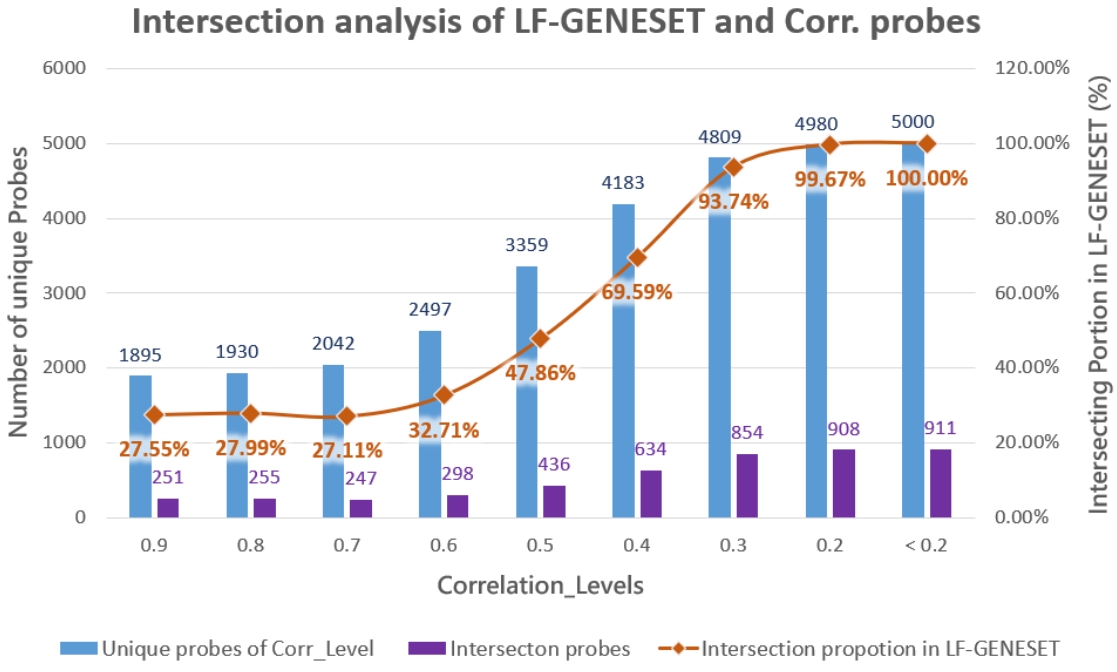

圖 36 GeneVPNN LF-GENESET 交集 CV5000 各相關係數級別探針分布圖

doi:10.6342/NTU202301211

由圖 36 之統計圖得知， LF-GENESET 與高度相關係數級別(0.9~0.7)不重複探針交集數占其比重約 27%，也就是說，高度相關係數的基因對(gene pair)不會在 LF-GENESET 中占比太高，乃因 GeneVPNN 會剔除掉相關性高的同類基因僅挑選其一代表，並且 LF-GENESET 交集 CV5000 各相關係數級別探針分布百分比，會隨相關係數下降逐漸上升，符合降維對應之相關係數分布常態表現，證實 GeneVPNN 之 AE 特徵基因挑選方法具有效降維能力。

4.2 研究討論

在本試驗結果中，仍有一定比例之基因群探針值無法被 GeneVPNN 模型有效的預測推估出，從過往許多文獻研究中也證實同種人種中，人體中有部分基因序列除致病後造成基因序列變異外，也會因個體差異而呈現單核苷酸多態性(SNP, Single Nucleotide Polymorphism)[18]，故也或許在 GeneVPNN 模型運算前期挑選推估有效目標基因 CV5000 時，挑出因個體差異之 SNP 型態基因群所致，但試驗基於複雜疾病之致病基因來自多元基因組之間相互關聯而致病觀點下，仍將 CV5000 所有基因群檢測值納入進行試驗，作為 GeneVPNN 推估之目標基因來源。

4.3 結論

GeneVPNN 兩個主要研究產出結論，一為模型泛用性設計，可輔助各類複雜疾病之醫學研究，另一試驗產出為輔助推薦出最經濟之基因檢測範圍，達成精準醫療階段性研究任務。

模型泛用性設計

GeneVPNN 基因虛擬探針檢測模型所建構之預測流程，在本研究中雖然以女性乳癌術後化療後之追蹤期間的資料集進行研究，但 GeneVPNN 的設計架構是以適用廣泛疾病前提下所設計之模型，可輔助未來醫學研究於各專科複雜疾病之有效目標基因群 CV5000 推論，解決全基因檢測耗費時間及資源議題，可作為生醫藥領域基因研究前期分析工作之參考數據來源。

doi:10.6342/NTU202301211

推薦最經濟之基因檢測範圍

再者本研究試驗結果有助於探索未知病因基因檢測執行範圍規劃，在 GeneVPNN 基因虛擬探針檢測模型之 AE 程序推薦 LF-GENESET 給予最經濟之基因檢測範圍之效益下，除了醫學研究團隊及病患可省去做全基因檢測之成本及時間外，並可應用 GeneVPNN 推論研究族群之有效目標基因群 CV5000，協助醫學研究分析工作及輔助臨床醫學診斷參考。

4.4 研究限制

資料限制

於模型訓練階段，除應盡可能採用收案樣本量充足之資料集外，並應在同樣收案研究條件下，試驗組與對照組兩類別資料量理想比例應盡可能趨近平衡為宜。本研究試驗以亞洲女性乳癌術後癌症復發追蹤之資料集為主，然而此類亞洲女性之乳癌復發追蹤研究之資料集甚少，再則欲找到相同癌症且經相同療程術後追蹤多年研究之資料集進行合併更甚為困難，在 GSE102484 資料集所收案之樣本 683 筆中，癌症復發資料比例占整體資料集比率約 15%，故欲進一步優化 GeneVPNN 模型以改善整體預測誤差率，則建議應盡量排除此資料限制產生之模型訓練瓶頸。

未來延續研究

本研究目的以設計驗證 GeneVPNN 模型推薦 LF-GENESET 為始，並發展出 GeneVPNN 模型預測 CV5000 目標基因群為最終研究目的。未來延續研究除在盡可能在排除本研究限制條件下發展外，可將 GeneVPNN 所推薦之 LF-GENEESET 潛在特徵基因組結合醫學主流基因篩選之統計方法，發展出進階致病基因推薦系統輔助生醫藥研究之決策執行。或發展如基因富集分析 GSEA (Gene Set Enrichment Analysis) 平台[24]，將 LF-GENESET 潛在特徵基因組納為 GSEA C4 等級(Computational Gene Set) pathway data base 之一，使 Leading edge 範圍內所篩選出的基因組研究面向更多元。

doi:10.6342/NTU202301211

參考文獻

1. Aronson, S.J. and H.L. Rehm, *Building the foundation for genomics in precision medicine.* Nature, 2015. **526**(7573): p. 336-42.
2. Sporikova, Z., et al., *Genetic Markers in Triple-Negative Breast Cancer.* Clin Breast Cancer, 2018. **18**(5): p. e841-e850.
3. Castilla, L.H., et al., *Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer.* Nat Genet, 1994. **8**(4): p. 387-91.
4. Wooster, R., et al., *Identification of the breast cancer susceptibility gene BRCA2.* Nature, 1995. **378**(6559): p. 789-92.
5. *PALB2 is a novel breast cancer susceptibility gene.* Nature Clinical Practice Oncology, 2007. **4**(5): p. 271-271.
6. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
7. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.
8. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**(7605): p. 47-54.
9. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.
10. Kulkarni, J.A., et al., *The current landscape of nucleic acid therapeutics.* Nat Nanotechnol, 2021. **16**(6): p. 630-643.
11. Hiam-Galvez, K.J., B.M. Allen, and M.H. Spitzer, *Systemic immunity in cancer.* Nat Rev Cancer, 2021. **21**(6): p. 345-359.
12. Jackson, S.P. and J. Bartek, *The DNA-damage response in human biology and disease.* Nature, 2009. **461**(7267): p. 1071-8.
13. Hodson, R., *Precision medicine.* Nature, 2016. **537**(7619): p. S49.
14. Uffelmann, E., et al., *Genome-wide association studies.* Nature Reviews Methods Primers, 2021. **1**(1): p. 59.
15. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.
16. Cheng, S.H.-C., et al., *Validation of the 18-gene classifier as a prognostic biomarker of distant metastasis in breast cancer.* PloS one, 2017. **12**(9): p. e0184372.
17. Kesteven, G.L., *The coefficient of variation.* Nature, 1946. **158**: p. 520.
18. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.*

doi:10.6342/NTU202301211

Nature, 2001. **409**(6822): p. 860-921.
19. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks.* Science, 2006. **313**(5786): p. 504-7.
20. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015. pmlr.
21. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
22. Botling, J., et al., *Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation.* Clin Cancer Res, 2013. **19**(1): p. 194-204.
23. Pearson, K., *On an Extension of the Method of Correlation by Grades or Ranks.* Biometrika, 1914. **10**(2/3): p. 416-418.
24. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

doi:10.6342/NTU202301211