

融合統計方法與人工智慧之智慧醫療應用

蔡曉婷¹ 吳崇豪^{1,2,3} 盧鴻興^{1,3,4,5,6†}

¹ 高雄醫學大學人工智慧生醫研究院

² 高雄醫學大學基礎科學教育中心

³ 國立陽明交通大學統計學研究所

⁴ 高雄醫學大學人工智慧醫療學科

⁵ 高雄醫學大學附設中和紀念醫院臨床醫學研究部

⁶ 美國康乃爾大學統計與資料科學系

摘 要

數據的價值不只在「分析」而在「轉譯」應用，如何融合統計方法並有效整合進人工智慧系統，以提升診斷準確性與解釋能力是至關重要的課題。團隊整合文獻及臨床運用歸納出透過數據資料轉換臨床實務應用之基礎五大步驟，包含資料前處理（Data Preprocessing）、模型訓練（Model Training）、效能評估（Performance Evaluation）、模型解釋（Model Explainability）、臨床解讀（Clinical Interpretation）。

目的在協助讀者了解從數據轉譯至臨床應用的基礎過程，進一步剖析如何透過統計方法結合傳統統計技術與現代深度學習模型，提高醫療影像分析的準確性和效率，進而改善疾病診斷和治療過程。最後透過團隊過往與臨床醫療團隊實際合作開發之醫療案例作為分析基礎，包含有阿茲海默症進程偵測、心律不整判讀、腦瘤影像識別、ICU EEG 異常偵測、口腔癌細胞化療抗藥性狀態分辨及顫顫關節障礙（TMD）評估等情境做為實例。

每一案例均以真實醫療資料驗證其效能，為未來智慧醫療決策支援系統的發展提供理論與實務參考，讓統計不再只是推論，而是驅動 AI 應用的基石。

關鍵詞：統計、人工智慧、智慧醫療應用。

JEL classification: C45, C52, I10.

[†] 通訊作者：盧鴻興
E-mail: henryhslu@kmu.edu.tw

1. 前言

現今人工智慧 (Artificial Intelligence ; AI) 已成為推動科技創新與產業轉型的關鍵驅動力。AI 通常被定義為一種能夠模擬人類智慧行為的電腦系統，其涵蓋的能力包括學習 (learning)、推理 (reasoning)、規劃 (planning)、感知 (perception)、語言理解 (natural language understanding) 與問題解決 (problem solving) 等 (Russell and Norvig 2021)。近年來，隨著資料收集的門檻降低、演算法設計的進步與運算資源的飛躍發展，AI 應用的深度與廣度更是快速擴展，目前已廣泛導入於醫療 (Huang et al. 2023)、金融 (黃健峯等 2024)、製造 (莊柏逸等 2023) 以及教育 (Wang et al. 2024) 等場域。其中在醫療領域以深度學習 (Deep Learning) 為應用的重點 (Thirunavukarasu et al. 2023)。

「統計」是構成人工智慧演算法的基礎支柱之一，不僅為模型建立與資料處理提供理論支持，也在模型訓練、特徵選取、參數優化與預測推論等重要環節中扮演關鍵角色，例如 Farea Similarity for Feature Selection (FSFS) 是統計理論驅動的特徵選擇方法，可直接影響模型訓練品質、參數穩定性與推論可信度 (Farea et al. 2025)。從早期的回歸分析 (regression analysis)、機率模型 (Probabilistic Graphical Models) 與貝氏網路 (Bayesian networks)，到近年普遍應用的支援向量機 (Support Vector Machine ; SVM)、主成分分析 (Principal Component Analysis ; PCA) 及集成學習 (Ensemble Learning) 等，皆可被視為統計方法的延伸與應用。此外，機器學習本身即涵蓋多種統計方法，使 AI 系統能夠自資料中自我學習與建模，進行推論與決策 (Rajkomar et al. 2019)。同時，統計方法強化了 AI 模型的可解釋性與穩定性，並為其在高風險醫療場域中的臨床應用提供可靠的數據分析架構 (Obermeyer and Emanuel 2016)。然而，相較於國際資料庫的規模優勢，台灣醫療環境仍面臨樣本數有限、資料不平衡與跨院整合困難等挑戰，因此如何在在地情境下開發具臨床適用性的 AI 模型，便成為重要課題。本研究旨在透過五大核心應用步驟與六大案例，與在地的臨床合作發展具體呈現統計思維與人工智慧在智慧醫療中的互補性與整合潛力，並進一步揭示 AI 在臨床應用中的價值與未來發展方向。

2. 統計與人工智慧之融合方法

隨著人工智慧 (Artificial Intelligence ; AI) 模型其高度自動化與複雜化，與背後所倚賴的統計原理與推論機制更顯得核心且不可或缺。而統計與人工智慧該

表 1：統計類別在 AI 中的功能目的、角色及常見應用。

類別	功能與目的	在 AI 中的角色	常見技術與應用舉例
描述性統計 (Descriptive Statistics)	- 資料摘要與概覽 - 偵測資料分佈與偏態	- 資料探索與清理 - 為後續特徵工程與建模做準備	- 均值、中位數、眾數、標準差、分位數 - 分布圖、直方圖、盒鬚圖 - 像素直方圖分析（影像）、詞頻計數（文字）
推論性統計 (Inferential Statistics)	- 從樣本推估母體特性 - 檢定假設與推論置信度	- 少量樣本下之學習與推論 - 提供統計保證（信賴區間、p-value）	- t-test、卡方檢定、ANOVA - 信賴區間、p-value - 機率估計
預測性統計 (Predictive Statistics)	- 建立資料間的因果或關聯模型 - 對未來樣本進行預測	- 與監督學習高度重疊 - 核心為預測與決策支持	- 迴歸分析（線性、邏輯斯） - 決策樹、隨機森林 - 神經網路
多變量統計 (Multivariate Statistics)	- 處理與分析高維資料 - 嘗試揭示變數間關係	- 維度縮減與特徵抽取 - 減少計算成本並幫助可視化	- 主成分分析（PCA） - 多元邏輯斯回歸 - 典型相關分析（CCA） - 因子分析

如何融合呢？從方法論的角度來看，描述性統計（Descriptive Statistics）與推論性統計（Inferential Statistics）、預測性統計（Predictive Statistics）和多變量統計（Multivariate Statistics），此四者分別在 AI 模型的資料前處理、建構、評估與最佳化過程中分別扮演著相當關鍵角色，為了讓讀者能一目了然地掌握統計方法在 AI 流程中的定位，本文彙整成表 1，對照不同統計類別在各建模階段中的功能目的、典型方法與臨床應用範例，作為後續各節案例分析的整體架構。

首先，描述性統計方面主要協助 AI 系統對原始資料進行結構化與摘要化，包括集中趨勢、變異數量化與分布可視化，這對於後續模型訓練中的資料前處理與特徵理解至關重要。舉例而言，在圖像辨識任務中，可藉由像素值分布探索異常點或進行標準化處理；在自然語言處理中，詞頻統計（word frequency count）則可了解語料庫中的語意特徵與上下文模式。而在推論性統計方面則提供了 AI 模型於樣本層級上進行假

設檢定、機率建模與效能評估的理論依據。在小樣本學習 (few-shot learning) 或貝葉斯學習架構下，推論性統計的角色就更加關鍵。而預測性統計 (Predictive Statistics)，已廣泛應用於監督式學習模型，包括線性回歸、決策樹、隨機森林乃至深度神經網路，其核心皆在於建構一個能夠泛化預測未知樣本的統計映射函數。此外，多變量統計方法亦在 AI 實務應用中佔有一席之地。例如主成分分析 (Principal Component Analysis；PCA) 常用於高維資料之維度縮減與特徵壓縮，典型相關分析 (Canonical Correlation Analysis；CCA) 用於探索多資料模態間的潛在關聯性，而多元邏輯斯回歸則應用於分類問題。這些技術不僅提升模型訓練的效率與可解釋性，也有助於改善模型的泛化能力與穩健性。

人工智慧 (Artificial Intelligence；AI) 模型的選擇與設計，通常取決於任務性質 (如分類、回歸、聚類、降維等) 以及數據特性 (例如變數分布、樣本規模、資料型態與資料結構)，而這些關鍵條件往往需透過統計分析來揭示與確認。在模型建構初期，若資料滿足線性關係、變異數齊性與誤差常態性等條件，傳統統計模型如線性回歸與邏輯斯回歸便可成為分類與預測問題的有效基本工具。然而，當資料呈現高度非線性結構或偏態分布，透過 Shapiro-Wilk 等正態性檢定可發現傳統模型假設的違反，此時較適合採用如決策樹 (Decision Tree)、支援向量機 (SVM) 或深度神經網路等非參數模型。在高維度資料分析中，多變量統計方法如主成分分析 (Principal Component Analysis；PCA) 與典型相關分析 (Canonical Correlation Analysis；CCA) 則可作為降維與特徵工程的方法，可有效提升 AI 模型的訓練效率與泛化能力。當面對樣本數遠小於特徵數的「小樣本、高維特徵」問題時，統計方法中的貝葉斯建模策略等手段則能提升模型穩健性，降低過擬合風險。模型評估亦依賴統計指標的支持，包括 AUROC (Area Under the Receiver Operating Characteristic Curve)、F1-score 與 RMSE (Root Mean Square Error) 等，其計算與解釋需結合推論統計與誤差理論，並針對不同任務與資料情境進行適當比較與選擇。其中因預測模型用來提供結果的風險值或分數這在醫療情境是非常重要的評估依據，AUROC 非常適合用於這種模型的效能評估 (模型能夠將正類別的風險值或分數正確排列於負類別之前的能力)。統計方法對於 AI 系統在實務應用中之不確定性控制與可解釋性建構也發揮了重要作用。[Lambert et al. \(2024\)](#) 指出，在醫療影像分析等高風險應用場域中，不確定性量化 (uncertainty quantification；UQ) 技術可降低深度學習模型所產生的「黑箱效應」，進而提升臨床醫師對模型預測結果的信任與接受度。[Matos et al. \(2025\)](#) 則針對臨床 AI 中的公平性評估指標進行 scoping review，發現儘管已有超過 60 種衡量方式，但多數

缺乏臨床驗證，並依賴閾值設計，顯示出統計理論在公平性量化上仍有待整合進臨床應用的標準流程。

綜上所述，人工智慧的發展不僅是演算法與運算資源的技術競賽，更是統計思維、資料素養與不確定性處理能力的深度實踐。統計不僅提供 AI 建模過程中的數理基礎與驗證架構，更在臨床決策輔助、風險預測與公平性演算法設計中扮演關鍵角色，為智慧醫療等領域建立堅實的理論支柱。然而，在實際臨床應用情境中，醫療 AI 仍面臨多重公平性與信賴性挑戰。以機器學習（machine learning；ML）為例，所謂「模型的公平性」是指在基於 ML 模型之自動化決策過程中，盡量避免因演算法偏差而對特定族群造成系統性不利。如果模型在學習與預測時過度偏好被視為敏感的變項（例如性別、種族、性取向或身心障礙）的群體，則其決策結果可能被認為是不公平或不透明，並伴隨資安與隱私風險、外部驗證資料不足等問題。

順應上述挑戰，我國政府於 113 年提出三個以醫療 AI 為核心的專責中心，分別為「負責任 AI 執行中心」、「臨床 AI 取證驗證中心」與「AI 影響性研究中心」。負責任 AI 執行中心主要聚焦於制定資安與隱私保護相關管理辦法，提升 AI 模型之透明度與可解釋性，並建構模型生命週期管理制度；臨床 AI 取證驗證中心則與食品藥物管理署（TFDA）及各層級、跨體系醫療機構合作，建立 AI 軟體之驗證流程與資料蒐集機制；AI 影響性研究中心則與中央健康保險署合作，發展醫療 AI 產品的臨床效益評估與健康經濟學定價模式，作為是否納入健保給付之重要依據。這三個中心共同構成我國在推動醫療 AI 時，兼顧安全、公平與效益評估的制度性基礎，這些議題也凸顯統計方法在偏差偵測、敏感變項處理與不確定性量化上的重要性。

3. 五大核心應用步驟

人工智慧（Artificial Intelligence；AI），特別是以深度學習（Deep Learning）為核心的演算法，已被廣泛應用於心電圖（ECG）判讀、醫學影像分類與預後風險預測等多項臨床情境中，展現其在高維度、非結構化生醫資料處理上的強大能力（Wu et al. 2024）。為使人工智慧系統在醫療領域具備可行性與信賴度，統計的角色不可或缺。本文進一步從五大核心步驟切入，探討統計推論在智慧醫療 AI 系統中的整合方式，包括：資料前處理（Data Preprocessing）、模型訓練（Model Training）、效能評估（Performance Evaluation）、模型解釋（Model Explainability）及臨床解讀（Clinical Interpretation），見圖 1。

第一步驟資料前處理（Data Preprocessing），人工智慧（AI）模型的效能相當依



圖 1：統計方法在智慧醫療 AI 系統中的五大核心步驟。

賴資料前處理策略是否合宜。此過程不僅是提升模型準確率與穩健性（robustness），更關乎模型之泛化能力（generalizability）與可解釋性（interpretability）。面對高維資料，透過適當的維度縮減（dimensionality reduction）與資訊保留可有效抑制雜訊干擾與冗餘特徵，以避免後續模型要面臨維度災難（curse of dimensionality）。在統計基礎理論中變異數分析（Analysis of Variance；ANOVA）、主成分分析（Principal Component Analysis；PCA）與線性判別分析（Linear Discriminant Analysis；LDA）此三項統計之皆已被廣泛應用在醫療資料處理流程（Guyon and Elisseeff 2003）。

於高維非線性醫療資料中導入上述具理論可解釋性的統計轉換工具，有助於提升模型可控性與結構穩定性，亦能有效緩解深度學習演算法常見之過擬合（overfitting）問題，強化模型在未知樣本上之推論可靠性與決策效能。意即，在培訓臨床 AI 專家助手的過程中，須先透過上述統計轉換工具，將臨床資料中具可解釋且經專家審核的重要且正確資訊，整理並轉換為結構化、可重複利用的特徵與規則，作為後續模型學習與知識庫建構的基礎。如此一來，得以在維持模型預測效能的同時，降低純黑箱深度模型在臨床決策上難以解讀與難以驗證之風險。

第二步驟模型訓練（Model Training）：在機器學習流程中，首先須掌握三大模型類型：監督式學習（Supervised Learning）以已標註的資料訓練預測模型；非監督式學習（Unsupervised Learning）用於揭示未標註資料中的內在結構，如聚類與降維；強化學習（Reinforcement Learning）則透過與環境互動和獎勵機制，學習能最大化累積報酬的策略（Jordan and Mitchell 2015）。在明確所面臨的問題該採用什麼類型來處理後，才能進入模型選擇步驟，首先需依據任務類型（例如分類、回歸、聚類）、資料特性（資料量、線性程度）、可解釋性需求與資源限制來篩選候選演算法（如邏輯迴歸、決策樹、隨機森林、SVM、神經網路等）。選定候選模型後，必須使用交叉驗證（如 K-fold Cross Validation）或訓練／驗證／測試集拆分來評估其泛化能力，並利用超參

數調整策略（如網格搜尋、貝葉斯優化）提升表現，再搭配正則化、early stopping、ensemble 等技術控制過擬合，最終選出在精確度、可解釋性與資源效率間取得最佳平衡的模型。模型訓練的目標並非單純追求在訓練資料上的最佳表現，而是為特定臨床情境選出「適配」的臨床 AI 專家助手：該模型需以符合目標族群與照護情境的資料加以建構與驗證，並經過嚴謹的內部與外部驗證程序，方能在真實世界中有效減輕臨床人員的工作負荷。例如，用於兒科重症風險預測的模型，應以兒科病人資料進行訓練與評估；若直接將以成人外科資料訓練之模型套用於兒科族群，即可能產生系統性預測偏差，不僅無法降低臨床負擔，反而增加決策風險。

第三步驟效能評估（Performance Evaluation）：當進行人工智慧或機器學習模型的建構後，效能評估（Performance Evaluation）便成為判斷模型是否具備實用價值的關鍵步驟。根據任務的不同，效能評估的方法與指標也有所差異。以分類任務為例，常見的評估指標包括準確率（Accuracy）、精確率（Precision）、召回率（Recall）與 F1 分數（F1-score），其中 F1 分數特別適用於資料分布不均的情況（Powers 2011；Saito and Rehmsmeier 2015）。此外，AUROC 值可用來衡量模型區分正負類別的能力，而混淆矩陣則可進一步呈現各類預測結果的細節。

對於回歸任務，則常以平均絕對誤差（MAE）、均方誤差（MSE）、均方根誤差（RMSE）以及決定係數（ R^2 ）作為評估指標，用以衡量預測值與實際值之間的差距。而在無監督學習如分群分析中，則常使用 Silhouette Score、Calinski-Harabasz Index 與 Davies-Bouldin Index 等指標來判斷分群效果。實務上，效能評估應搭配交叉驗證（Cross-validation）進行，以提高評估結果的穩定性與代表性。此外，資料切分的比例與方式（如訓練集與測試集的劃分）也會直接影響模型的效能表現。尤其在醫療應用中，更需關注各種評估指標背後所代表的臨床風險與決策後果，例如錯誤分類可能導致病情延誤、錯失治療時機，或造成不必要的檢查與處置。因此，效能評估不僅是演算法層面的技術問題，更關乎模型在真實世界情境中的可靠性與可解釋性。類比於臨床專業人員必須經過國家考試與專科認證，才能確認其具備足夠的專業能力，醫療 AI 模型亦需透過嚴謹的內部驗證、外部驗證以及多中心資料評估，方能被視為具備可被信賴的臨床決策輔助能力。

第四步驟模型解釋（Model Explainability）：在高階醫療人工智慧模型的應用中，「模型解釋性」（model explainability）不僅是實現結果透明（transparency）與可信度（trustworthiness）的關鍵，更是臨床應用能否獲得醫師信賴的核心要素。對統計而言，模型可解釋性亦代表一種可驗證性與科學性，並非僅止於視覺化呈現，更

需建立於嚴謹的機率邏輯與關聯性推論之上。傳統統計模型如邏輯斯迴歸 (Logistic Regression) 與 Cox 比例風險模型 (Cox Proportional Hazards Model) 在醫學風險預測上具備高度可解釋性。前者提供各個自變項與二元結果間的對數勝算比 (log-odds ratio)，便於臨床人員理解風險因子之方向與強度；後者則透過風險比 (hazard ratio) 評估時間相關事件 (如疾病復發、生存時間) 之影響，並廣泛應用於預後模型建構。此外，貝式推論 (Bayesian inference) 架構亦日益受到重視，因其可融合先驗知識與資料證據進行參數估計，提供具不確定性量化的風險預測結果，特別適合資料稀疏或不平衡的臨床場景。然而，深度學習與集成模型 (如隨機森林、XGBoost) 之複雜結構，欠缺直接的係數解釋能力，故需輔以現代解釋技術來提升透明度。其中，SHapley Additive exPlanations (SHAP) 為基於合作博弈論的特徵歸因方法，能針對單一預測案例計算各變數對預測結果的邊際貢獻度，其理論基礎與 Shapley value 相連結，具公平性、公理化與一致性特質，適合應用於高風險臨床決策系統。Local Interpretable Model-Agnostic Explanations (LIME) 則採用局部線性逼近策略，針對目標預測點附近進行擾動取樣，構建可解釋的線性模型以近似原模型在該區域的預測邏輯，適用於不同模型架構之比較與可視化分析。結合具統計推論架構之傳統模型與後設可解釋方法，不僅可提升 AI 模型在臨床應用場景中的解釋性與可信度，更有助於支持以風險因子為導向之個人化醫療策略制定，促進智慧醫療系統之倫理性與法規遵從。換言之，在臨床情境中培育與應用「臨床 AI 專家助手」時，除了關注其預測結果本身的準確度之外，更必須藉由上述可解釋技術檢視其判斷所依據的特徵與推論邏輯是否合理、一致且足以被臨床醫師信賴。

第五步驟臨床解讀 (Clinical Interpretation)：是指將人工智慧模型或統計分析結果轉化為具備臨床意義的資訊，進而支援醫療決策與病人照護的過程。臨床解讀的過程通常包括：分析預測結果的臨床意涵、將模型輸出與現行臨床指引進行比對、考量病人個別差異與共病條件，以及評估模型不確定性所帶來的潛在風險。再者，臨床解讀強調「人機協作」的原則，AI 模型應作為臨床判斷的輔助工具，而非取代專業經驗。雖然在模型訓練過程中，臨床應用常面臨「準確度」與「可解釋性」取捨，然而如何權衡兩者的重要性，筆者認為只要在臨床可接受的準確度之下，來提供合理解釋性，便能發展出最貼近臨床需求的模型。臨床解讀步驟的關鍵，在於釐清這類「臨床 AI 專家助手」應如何被嵌入實際照護流程——由誰在何時使用、以何種方式與臨床判讀結果相互驗證——並持續監測其對工作效率、照護品質與醫療成本的實際影響，方能確保 AI 的導入真正為臨床帶來可被信賴且具成本效益的價值。

綜合以上，統計不僅為 AI 系統提供建模與評估工具，更擔任 AI 模型可信性與倫理應用的「橋梁角色」。

表 2：五大步驟 vs. 六大案例應用。

案例	資料前處理 (Data Pre-processing)	模型訓練 (Model Training)	效能評估 (Performance Evaluation)	模型解釋 (Model Explainability)	臨床解讀 (Clinical Interpretation)
心房顫動 AF 預測	PVCT 影像切片標準化、臨床變數清理與 I-Score 特徵選擇	深度學習模型 (PVCT 影像) + 機器學習模型 (臨床變數) + 集成學習	AUROC、敏感度、特異度	I-Score 選出具鑑別力變數，分析術前預測因子	根據模型風險分層，提供術前個別化治療建議
腦腫瘤 fDDFT 無監督分割	Fermi normalization、變異數正規化處理影像特徵	fDDFT 演算法 + 幾何深度學習 (GDL 架構) + 圖結構度量 (GNN metrics)	Dice score、運算時間 $O(N \log N)$ vs. $O(N^3)$	顯著性特徵子圖 + 幾何穩定性解釋分割結果	用於輔助放射科醫師快速區分腫瘤與正常組織
阿茲海默症 (AD) 病程偵測	多模態資料 (MRI、臨床變項) 前處理	分類／預測模型 (多分類架構) + 集成學習	Accuracy、敏感度	Grad-CAM 解釋萎縮腦區 (如海馬迴) 對預測影響	用於早期 AD 病程分層，輔助臨床追蹤決策
ICU EEG 異常偵測	EEG 信號濾波、標準化 + 時窗化特徵切片	GRU 模型 + class weights 處理不平衡	Accuracy、敏感度、特異性、Balanced Accuracy	以醫師標註對齊 GRU 模型輸出，回溯異常 EEG 時段，提升預測結果的臨床可解釋性	ICU 神經監控決策支援，降低臨床判讀負擔
口腔癌細胞抗藥性分類	細胞影像增強、上下取樣平衡各類別比例	EfficientNet-B3 CNN 模型 + 集成學習	Accuracy、RMSE、交叉驗證	Grad-CAM 顯示 CNN 決策關注於細胞輪廓，間接解釋形態變化與抗藥性關聯	預測抗藥細胞族群比例，輔助臨床化療規劃
顳顎關節障礙 (TMD) 分類	YOLO 自動截取影像 ROI、HSV 空間轉換、LDA 臨床變數壓縮	CNN 多類別分類架構 (6-class)	Accuracy	以 Grad-CAM 標示 CNN 關注的關鍵關節結構，解釋模型在多階診斷中的決策依據。	提高 TMD 多階診斷效率，減少人工辨識誤差

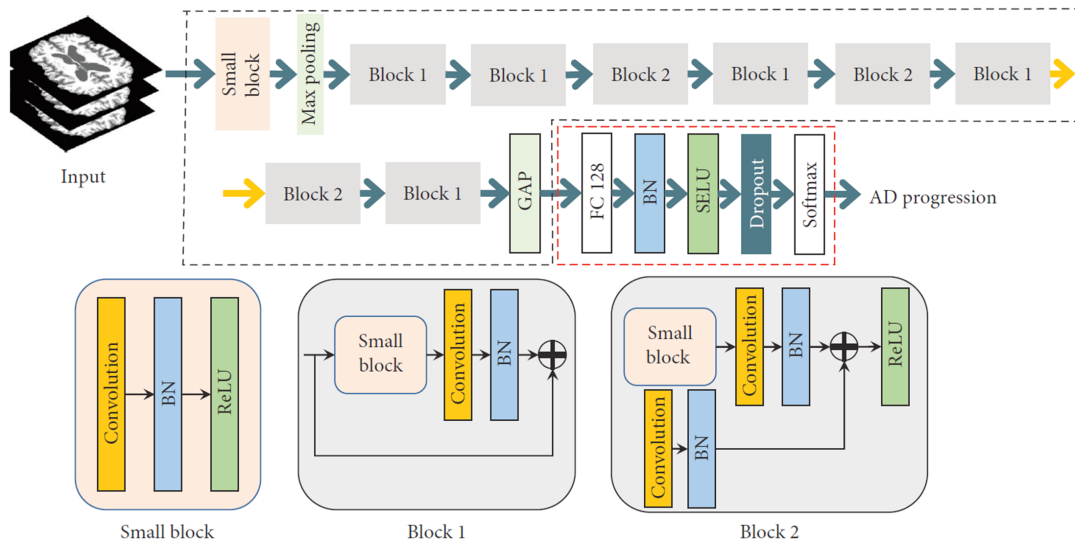


圖 2：模型訓練過程（參考來源：Wu et al. 2025）。

4. 六大案例應用

在醫療領域中，AI 系統能自大量結構化與非結構化資料中學習臨床模式，以下介紹盧鴻興教授團隊過往與臨床醫療團隊實際合作的六項任務，透過高階統計方法和先進的人工智慧，提高醫療影像分析的準確性和效率，進而改善疾病診斷和治療過程之應用及挑戰，並運用五大核心應用步驟及結合六項任務，具體展現從統計到 AI 的重要連結，詳見表 2。

第一項任務為阿茲海默症進程偵測的研究，為從傳統統計模型轉向以影像資料為基礎的機器學習方法。研究在模型訓練中所提出的模型架構由兩個主要部分組成：第一部分為 ResNet3D-18 模型的特徵萃取模組（以黑色虛線框標示），並採用預訓練權重；第二部分則為針對阿茲海默症（AD）疾病進展偵測任務重新設計的分類模組（以紅色虛線框標示）。最終輸出層採用 softmax 函數，其輸出節點的數量等於需分類的目標類別數（見圖 2），在所有模型訓練設定中，研究採用的損失函數 focal loss 可視為對交叉熵損失（cross-entropy loss）的一種改良形式。以降低那些容易辨識之樣本在訓練中的影響程度。換言之，focal loss 會較關注那些較難區分的訓練樣本。其數學式為

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t),$$

其中 p_t 為正確類別的預測機率， α_t 為自定義樣本權重。

研究利用 ADNI 資料庫進行簡單多類別分類，模型在準確率與靈敏度上表現優

異。為提升模型可解釋性，作者透過 Grad-CAM 視覺化，指出部分 MRI 區域（如海馬迴、杏仁核）之萎縮特徵為辨識疾病進展的重要指標。此研究展現出統計與 AI 方法融合於高維影像與臨床變項整合分析的潛力，並突顯模型可解釋性在臨床應用與決策輔助中所扮演的關鍵角色（Wu et al. 2025）。過程中本研究共蒐集 4658 份 MRI 影像資料，並以分層抽樣（stratified split）方式將資料分為 64% 訓練集、16% 驗證集，以及 20% 測試集。每一筆 MRI 影像皆透過 FSL 進行前處理，以分離出灰質（GM）、白質（WM）與腦脊髓液（CSF），用以呈現腦部結構特徵。所有模型皆以 PyTorch 框架實作。為提升訓練資料的多樣性與增強模型泛化能力，研究採用了影像增強技巧。於 ResNet3D-18 模型的特徵萃取部分，使用由 Kinetics 資料集訓練所得的預訓練權重，此預訓練權重已內建於 PyTorch，可直接啟用。研究在所有訓練設定中皆採用參數 $\gamma = 2$ 的 focal loss 作為損失函數，並依據各類別的資料量動態調整 class weights，以處理類別不平衡問題。學習率與批次大小分別設定為 0.00005 與 16。優化器部分採用 Adam，使模型參數能以自適性學習率更新。模型在訓練集上進行 300 個 epoch 的訓練，並以驗證集準確率的變化趨勢監測模型表現，分別挑選每個模型於訓練過程中表現最佳的參數組（及其對應的 epoch）。其相關測試評估指標見表 3。

鑒於簡單多類別分類的表現不甚滿意（僅 0.924 的準確度），該研究進一步提出整合式學習架構（Ensemble Learning）以提升預測表現，該研究整合簡單多類別分類、階層式分類（Hierarchical Classification）、一對一分類（One-Versus-One Classification）及一對多分類（One-Versus-All Classification）的結果共同進行預測，

表 3：簡單多類別分類的測試評估指標。

疾病進程	準確率 (Accuracy)	敏感度 (Sensitivity)	特異度 (Specificity)	F1 分數 (F1-score)	ROC 曲線下面積 (AUROC)
CN	0.924	0.932	0.964	0.94	0.991
EMCI		0.942	0.938	0.889	0.984
LMCI		0.889	0.991	0.914	0.98
AD		0.906	0.999	0.948	0.997
平均值		0.917	0.973	0.923	0.988

註 1：(1) CN：Cognitive Normal，認知正常；(2) EMCI：Early Mild Cognitive Impairment，早期輕度認知障礙；(3) LMCI：Late Mild Cognitive Impairment，晚期輕度認知障礙；(4) AD：Alzheimer’s disease，阿茲海默症。

註 2：平均值為四個疾病進程類別的算術平均值（參考來源：Wu et al. 2025）。

以辨識阿茲海默症不同病程階段（正常、早期認知障礙、晚期認知障礙與 AD），最終實現 0.95 的測試準確率，媲美神經內科專科醫師的診斷水準。

第二項任務在腦腫瘤的識別上，研究團隊提出一種融合密度泛函理論（Density Functional Theory；DFT）與幾何深度學習（Geometric Deep Learning；GDL）的新穎影像分割框架，稱為快速資料密度泛函轉換（fast Data Density Functional Transform；fDDFT），以實現對三維腦瘤影像的無監督分割。此方法以醫學影像中像素強度矩陣為基礎，構造出能量泛函景觀（如動能泛函 KEDF 與位能泛函 PEDF），進一步透過圖神經網路架構將局部結構映射至拓撲子圖，藉此辨識腫瘤與正常組織間的結構異質性。fDDFT 的關鍵統計概念包括：以變異數正規化與費米函數（Fermi normalization）定義特徵子圖的邊界與相似性收斂（similarity convergence），再以梯度上升演算法實現幾何穩定性（geometric stability）校正偏差。此方法不僅將原先 $O(N^3)$ 的卷積運算時間降至 $O(N \log - N)$ ，更在無監督學習情境下於 BraTS 2020 腦瘤資料集達到中位 Dice score > 0.75 的表現，展示統計建模與深度網路在高維醫學影像領域中的高度融合潛力（Huang *et al.* 2023）。

第三項任務在醫療預測應用中，我們團隊利用人工智慧（AI）技術，結合消融術前肺靜脈電腦斷層（PVCT）影像與臨床資料，預測心房顫動（AF）於導管消融術後的復發風險。共納入 638 位對藥物無效的陣發性心房顫動患者，皆接受導管消融（catheter ablation）治療。模型訓練所使用的影像來自消融前之 PVCT 影像，共計 126,288 張左心房（LA）切片圖像。臨床資料共蒐集 29 項變數，涵蓋基線特徵、病史、實驗室檢查結果、經胸心臟超音波參數及三維重建的左心房體積。研究中使用統計方法 I-Score 進行變數選擇以訓練模型。針對一年內 AF 復發的預測，分別建立了基於 PVCT 的深度學習模型與基於臨床變數的機器學習模型，並進一步將兩者進行集成學習（Ensemble Learning）以提升預測效能。在測試資料集中，PVCT 單獨模型的 AUROC 為 0.63。I-Score 選出的臨床變數組合可達 AUROC 0.72，顯著優於使用全部變數或以非參數統計選出的特徵（AUROC 為 0.66 至 0.69）。最終結合 PVCT 影像與臨床變數的集成模型預測效能最高，AUROC 達 0.76，敏感度為 86.7%。在導管消融術前，運用 AI 技術分析 PVCT 影像並結合 I-Score 所選變數，可有效預測陣發性心房顫動患者術後的復發風險。I-Score 可自多項預測變數中篩選出具鑑別力的特徵，有助於術前風險分層與個別化治療規劃（Liu *et al.* 2024）。

第四項任務在重症加護病房（ICU）中，腦波圖（EEG）是評估神經異常的重要指標，然而其高度複雜且時變的特性使得傳統人工判讀既耗時又負擔沉重。為提升神經

照護效率，團隊開發一套能考慮長短期記憶資訊的深度學習模型，以進行自動異常偵測。該系統針對即時偵測任務與高頻、資料不平衡等特性，最終選擇 GRU 為主體模型，搭配資料標準化、多通道輸入與前處理策略。為處理不平衡資料，團隊進行 class weight 調整，並透過 early stopping 與正則化避免過擬合。模型展現優異的泛化能力與穩定性，能有效辨識癲癇樣放電與背景活動低落等異常狀況，在精確度、靈敏度與 F1-score 皆優於傳統機器學習方法 (Wu et al. 2024)。此研究顯示，深度模型的選擇與優化應基於任務需求、資料結構與部署條件整體考量，具體展現統計與 AI 模型整合於臨床應用的價值與潛力。

第五項任務口腔癌細胞化療抗藥性狀態分辨任務細胞形狀反映細胞與其環境訊號間達到平衡後的空間構型，是評估細胞功能與生物特性的重要指標之一。對癌細胞而言，包括化療在內的多種生理與環境挑戰，會引發細胞狀態的轉變，並伴隨連續性的形態改變，而這些改變往往即使在顯微鏡下亦難以辨識。為探討深度學習影像分析是否能夠辨識代表關鍵細胞狀態轉變的形態變化，本研究以一株對化療具抗藥性的口腔癌細胞株作為模型，其形態與未具抗藥性的母株細胞極為相似。研究採用基於 EfficientNet-B3 的卷積神經網路 (CNN) 架構，並結合上取樣與下取樣技術，建立自動化影像分類模型，以區分非癌細胞、無抗藥性癌細胞與具抗藥性癌細胞三類。結果顯示，透過過採樣與欠採樣強化的 CNN 模型，可達到至少 98.33% 的分類準確率。此外，針對異質性細胞族群，最佳模型可準確估算具與不具抗藥性的癌細胞比例，其預測誤差經集成學習 (Ensemble Learning) 處理後，RMSE 降至 0.16。本研究證實，CNN 架構可辨識即使在視覺上難以察覺的細胞形態變化，支持深度學習作為未來細胞狀態分析與分類的可行自動化影像分析工具 (Hsieh et al. 2024)。

第六項任務為顫顎關節障礙 (TMD) 評估，TMD 是常見的口腔疾病，會影響咀嚼功能與生活品質。過去藉由核磁共振影像 (MRI) 進行診斷雖具精確性，但判讀需依賴經驗豐富的放射科醫師，且人工辨識具有高度主觀性。本研究旨在開發一個基於卷積神經網路 (CNN) 的自動化模型，用以分類不同階段的 TMD 疾病，提升診斷效率與客觀性。研究團隊建構一個六分類的 CNN 影像分類架構，以 MRI 影像進行訓練與測試。模型前處理部分包括：利用 YOLO 模型自動截取顫顎關節相關區域；影像轉換至 HSV 色彩空間後擷取明度 (V) 資訊；使用 LDA (線性判別分析) 壓縮與年齡、性別等臨床變項結合處理，最後使用 CNN 進行分類。CNN 模型成功將 MRI 影像分類為六個 TMD 階段，其分類準確率達 84.73%。實驗亦顯示，模型能有效處理資料異質性與影像雜訊問題，並具穩定的泛化能力。本研究證實 CNN 模型結合影像預

處理與統計方法（如 LDA）可有效支援顫顫關節障礙的多階分類任務，具有臨床應用潛力，可用於輔助醫師進行快速與客觀的 TMD 診斷（Su et al. 2025）。

在技術細節方面，團隊提到了幾個關鍵挑戰，包括如何處理醫學資料中的高維數據、如何挑選模型以適應不同類型的醫學資料，以及如何進行轉移學習，使模型能夠將在一種資料上學到的知識應用到另一種完全不同的資料上。這要求研究人員不僅要有深厚的人工智慧知識，還需要對醫學資料的特性有深入理解。為了克服這些挑戰，團隊採用諸如卷積神經網絡（CNN）和遞歸神經網絡（RNN）等先進的深度學習架構，並開發了專門針對醫學資料特性的訓練策略。這些技術使得 AI 系統能夠從大量的醫學影像數據中學習，並於不同的醫學診斷任務中達到人類專家級別的表現。此外，研究團隊在實作過程中亦面臨跨院與跨國驗證以及臨床落地的實質挑戰。一方面，模型必須在其他醫院甚至不同國家的資料上進行外部驗證，以檢驗其在異質族群與多中心資料下的泛化能力；另一方面，如何將 AI 系統真正嵌入既有臨床工作流程，包含臨床人員教育訓練、與醫院資訊系統整合以及相關流程的自動化處理，亦需要跨團隊長期協調。其突破之處，在於透過多層面的跨國與跨院合作與連結，逐步建構多中心驗證機制與臨床導入方案，使系統不僅在統計指標上達到專家級表現，更能在真實醫療現場中穩定運作並支援臨床決策。

5. 結論

經由智慧醫療應用案例，可了解統計方法在資料前處理、特徵選擇、模型建構與效能評估等環節中皆扮演關鍵角色，不僅有效提升 AI 模型的預測準確性與泛化能力，亦強化模型之透明度與臨床解釋性。智慧醫療發展的核心將朝向跨領域協作與資料倫理導向的可解釋 AI 演進，臨床人員須逐步建立統計素養與人工智慧素養（AI literacy），理解統計推論於機器學習建模中的角色與限制，進而與工程、統計及資訊專業者協同合作，共同建構具臨床可行性與實證基礎的智慧決策支援系統。特別是在複雜且模糊的臨床推論情境中，結合規則導向（rule-based）邏輯推理與資料驅動（data-driven）模式辨識的雙軌式推論架構，將有助於打造具人機協作特質的診斷模型。

在實際落地使用經驗中，團隊與臺中榮總共同開發之 Healthcare GPT，結合大型語言模型於臨床場域，支援四大應用領域：病歷摘要（Medical Record Agent）、AI 問答諮詢、Email 草稿生成，以及會議紀錄與技術文件整合等知識管理。該系統可於 15 秒內完成病歷摘要生成，顯著提升醫師與護理師的紀錄效率與照護品質。另一方面，

與高雄醫學大學醫療體系自主研發生成式 AI 平台 KMU Genie，結合五大形象代言人展現應用落地潛力，包括：臨床熊哥（智慧毒物診斷與解毒劑指引）、護理邦尼（AI 護理紀錄自動生成與健康問題篩選）、醫事小狐（AI 藥物諮詢與智慧營養衛教）、行政小松（行政流程優化）、以及導師歐文（教學與知識管理輔助）。這些應用不僅大幅減輕醫護人員負擔，亦有效提升診斷決策效率與臨床精準度，實現數位轉型的智慧醫療藍圖。

未來應進一步探索統計方法於 AI 模型可解釋性（explainability）、偏誤偵測（bias detection）與臨床信任建立（trust calibration）等面向之應用潛力，並透過可重複驗證的實證設計與跨機構資料整合，提升智慧醫療模型的實務轉譯能力。在 AI 時代，統計不僅是一項工具，更是連結數據科學與臨床判斷的橋梁，值得於智慧醫療發展藍圖中持續深化其核心價值。

參考文獻

- [1] Farea, A. H., Askerzade, I., Alhazmi, O. H., and Takan, S. (2025). FSFS: A novel statistical approach for fair and trustworthy impactful feature selection in artificial intelligence models. *Computers, Materials and Continua*, 84(1), pages 1457-1484.
- [2] Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, pages 1157-1182.
- [3] Hsieh, H. C., Chen, C. Y., Chou, C. H., Peng, B. Y., Sun, Y. C., Lin, T. W., Chien, Y., Chiou, S. H., Hung, K. F., and Lu, H. H. S. (2024). Deep learning-based automatic image classification of oral cancer cells acquiring chemoresistance in vitro. *PLoS ONE*, 19(11), e0310304.
- [4] Huang, S. J., Chen, C. C., Kao, Y., and Lu, H. H. S. (2023). Feature-aware unsupervised lesion segmentation for brain tumor images using fast data density functional transform. *Scientific Reports*, 13, 13582.
- [5] Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pages 255-260.
- [6] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., and Dojat, M. (2024). Trustwor-

- thy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150, 102830.
- [7] Liu, C. M., Chen, W. S., Chang, S. L., Hsieh, Y. C., Hsu, Y. H., Chang, H. X., Lin, Y. J., Lo, L. W., Hu, Y. F., Chung, F. P., Chao, T. F., Tuan, T. C., Liao, J. N., Lin, C. Y., Chang, T. Y., Kuo, L., Wu, C. I., Wu, M. H., Chen, C. K., ... , Chen, S. A. (2024). Use of artificial intelligence and I-Score for prediction of recurrence before catheter ablation of atrial fibrillation. *International Journal of Cardiology*, 402, 131851.
- [8] Matos, J., Van Calster, B., Celi, L. A., Dhiman, P., Gichoya, J. W., Riley, R. D., Russell, C., Khalid, S., and Collins, G. S. (2025). Critical appraisal of fairness metrics in clinical predictive AI. URL: <https://arxiv.org/abs/2506.17035>
- [9] Obermeyer, Z., and Emanuel, E. J. (2016). Predicting the future —Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), pages 1216-1219.
- [10] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pages 37-63.
- [11] Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), pages 1347-1358.
- [12] Russell, S. J., and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- [13] Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432.
- [14] Su, T. Y., Wu, J. C. H., Chiu, W. C., Chen, T. J., Lo, W. L., and Lu, H. H. S. (2025). Automatic classification of temporomandibular joint disorders by magnetic resonance imaging and convolutional neural networks. *Journal of Dental Sciences*, 20(1), pages 393-401.

- [15] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), pages 1930-1940.
- [16] Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., and Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252(4), 124167.
- [17] Wu, J. C. H., Chien, T. C., Chang, C. C., Chang, H. I., Tsai, H. J., Lan, M. Y., Wu, N. C., and Lu, H. H. S. (2025). Learning Based Progression Detection of Alzheimer's Disease Using 3D MRI Images. *International Journal of Intelligent Systems*, 2025(1), 3981977.
- [18] Wu, J. C. H., Liao, N. C., Yang, T. H., Hsieh, C. C., Huang, J. A., Pai, Y. W., Huang, Y. J., Wu, C. L., and Lu, H. H. S. (2024). Deep-Learning-Based Automated Anomaly Detection of EEGs in Intensive Care Units. *Bioengineering (Basel, Switzerland)*, 11(5), 421.
- [19] 莊柏逸、鄭憲君、沈意傑、林順傑 (2023)。超效率佈署 AI-BOT 流程自動化機器人技術與製造應用實例。《機械工業雜誌》，479 卷，頁 70-79。
- [20] 黃健峯、林彥頌、洪宗貝 (2024)。利用生成式 AI 的金融預測：Bing Chat 在股票市場的評估與應用。《前瞻科技與管理》，12 卷 2 期，頁 12-28。

[Received August 2025; accepted December 2025.]

Integrating statistical methods and artificial intelligence for smart healthcare applications

Shiau-Ting Tsai¹, Jacky Chung-Hao Wu^{1,2,3} and Henry Horng-Shing

Lu^{1,3,4,5,6†}

¹ Biomedical Artificial Intelligence Academy, Kaohsiung Medical University,
Kaohsiung, Taiwan

² Center for Fundamental Science, Kaohsiung Medical University, Kaohsiung, Taiwan

³ Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

⁴ Department of Artificial Intelligence in Medicine, Kaohsiung Medical University,
Kaohsiung, Taiwan

⁵ Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung,
Taiwan

⁶ Department of Statistics and Data Science, Cornell University, Ithaca, New York

ABSTRACT

In rapidly evolving smart healthcare, the value of data lies not only in analysis but also in its translation into clinical applications. A key challenge is how to integrate robust statistical methodologies into artificial intelligence (AI) systems to improve diagnostic accuracy and interpretability. This paper summarizes five fundamental steps for translating data into clinical applications: Data Preprocessing, Model Training, Performance Evaluation, Model Explainability, and Clinical Interpretation. These steps provide a systematic framework for developing translational AI solutions.

We further demonstrate how traditional statistical techniques can be effectively combined with modern deep learning models. This hybrid approach enhances both the accuracy and efficiency of medical image analysis, thus supporting improved diagnosis and treatment planning.

Finally, real-world case studies are presented, developed in collaboration with medical teams. These cases cover diverse clinical applications, including Alzheimer's disease progression monitoring, arrhythmia detection, brain tumor imaging, abnormal EEG Detection in the ICU, chemoresistance identification in oral cancer cells, and tem-

[†]Corresponding to: Henry Horng-Shing Lu
E-mail: henryhslu@kmu.edu.tw

poromandibular joint disorder (TMD) assessment. Each case study was validated using real-world medical data, providing both theoretical and practical insights for the development of next-generation intelligent clinical decision support systems.

By integrating statistical methods with advanced AI techniques, this work demonstrates how statistical thinking can empower AI-enabled healthcare innovation.

Key words and phrases: statistical, artificial intelligence, smart healthcare applications.

JEL classification: C45, C52, I10.