第一章 緒論

1.1 研究背景與動機

1.1.1 三高疾病的公共衛生重要性

三高疾病，即高血壓（Hypertension）、高血糖（Hyperglycemia）與高血脂（Dyslipidemia），是全球主要的慢性疾病，也是心血管疾病、中風、腎臟病等重大疾病的主要風險因子。根據世界心臟聯盟統計，心血管疾病每年造成約 2,000 萬人死亡，占全球死亡人數近三分之一，其中三高是主要的可控風險因子（World Heart Federation, 2023）。

亞洲地區的三高疾病負擔尤為嚴峻。根據 WHO 2023 年全球高血壓報告，西太平洋區域（涵蓋東亞與東南亞）超過四分之一的成年人患有高血壓（WHO, 2023）。在糖尿病方面，中國、日本、印尼等五個亞洲國家的糖尿病患者總數佔全球 48%，且各區域盛行率持續攀升（Ohira & Iso, 2013；JACC: Asia, 2021）。東亞地區的代謝性疾病近數十年急遽增加，其驅動因素包括遺傳易感性、獨特的體脂分布模式，以及都市化與西化飲食的快速轉變（Sun & Zheng, 2025）。值得注意的是，東亞地區的心血管死因以腦中風為主，有別於西亞以缺血性心臟病為主的模式，反映出不同區域三高疾病對心血管系統的影響路徑存在差異（Ohira & Iso, 2013；JACC: Asia, 2021）。

在台灣，三高疾病同樣是重要的公共衛生議題。根據衛生福利部國民健康署 2017-2020 國民營養健康狀況變遷調查，40 歲以上國人高血壓盛行率為 38.3%、高血脂盛行率為 34.1%、高血糖盛行率為 16.4%（國民健康署，2022）。此外，約有 4 至 7 成民眾不知道自己已罹患三高，凸顯早期預測與篩檢的重要性（國民健康署，2022）。

三高疾病往往彼此相關、共同發生，此現象在醫學上被歸納為「代謝症候群（Metabolic Syndrome）」的核心組成。研究顯示，超過 70% 的糖尿病患者同時合併高血壓或高血脂，而糖尿病患者中血脂異常的盛行率更高達 72–85%（Stanciu et al., 2023）。代謝症候群患者的心血管疾病風險為一般人的 2 倍，第二型糖尿病風險則為

5 倍（Alberti et al., 2009）。這種共病現象不僅增加了疾病管理的複雜性，也突顯了同時預測多種疾病風險的重要性。

1.1.2 早期預測與預防的臨床價值

慢性疾病具有幾個重要的流行病學特徵：

1. **早期症狀不明顯**：三高疾病在初期往往沒有明顯症狀，患者經常在例行健康檢查或併發症出現時才發現

2. **進程緩慢但可逆**：從健康到發病可能需要數年，但在這段「前驅期」，透過適當介入仍可逆轉

3. **可預防性高**：研究顯示，生活型態調整（如飲食、運動）可有效降低 50% 以上的發病風險

4. **晚期治療成本高**：一旦發展為嚴重疾病或併發症，治療成本將大幅增加

此外，三高疾病近年呈現明顯的年輕化趨勢。在台灣，30 至 39 歲族群中已有 18.7% 有高血脂、9.7% 有高血壓、2.5% 有高血糖（國民健康署，2022）。為因應此趨勢，國民健康署自 2025 年起將成人預防保健服務年齡從 40 歲下修至 30 歲（國民健康署，2025），反映出早期篩檢與預測的需求已從中高齡族群擴展至青壯年。

基於以上特徵，**早期預測**具有重要的臨床與公共衛生價值：

- 及早識別高風險個體，提供預防性介入
- 實現個人化的健康管理建議
- 降低長期醫療成本
- 提升國民整體健康水平

1.1.3 縱向健檢資料的研究價值

近年來，隨著健康檢查的普及，大量的縱向健檢資料（Longitudinal Health Checkup Data）被累積。這類資料記錄了同一個體在不同時間點的健康狀態，具有獨特的研究價值：

1. **捕捉動態變化**：可追蹤生物標記隨時間的變化趨勢

2. **反映健康軌跡**：呈現從健康到疾病的發展過程

3. **提供預測線索**：健康狀態的變化往往先於疾病確診

然而，傳統的疾病風險評估方法（如 Framingham 風險評分）主要基於**單一時間點**的檢驗數據，未能充分利用縱向資料中蘊含的動態資訊。這是本研究欲填補的重要研究缺口。

1.2 問題陳述

1.2.1 現有預測方法的限制

目前常見的三高疾病風險評估方法存在以下三大限制：

限制一：單時間點評估

傳統方法僅使用當前的檢驗數據進行風險評估，例如以當前空腹血糖值判斷糖尿病風險。這種方式忽略了一個重要事實：**相同的檢驗數值在不同健康軌跡下具有不同的意義**。例如，血糖從 90 上升到 95 mg/dL 的個體，與從 100 下降到 95 mg/dL 的個體，雖然當前血糖值相同，但其未來風險可能截然不同。

限制二：缺乏縱向資訊

大多數現有研究未充分利用歷史健檢資料，也缺少「變化量特徵」（Delta Features）的工程設計。近期研究（如 Yang et al., 2025）已證實，血糖變化量（δ-FPG）在糖尿病預測中具有極高的重要性，但類似的特徵工程方法尚未被系統性地應用於三高疾病的同時預測。

限制三：模型可解釋性不足

許多高準確度的機器學習模型（如深度神經網路）是「黑箱」模型，難以解釋預測背後的原因。這在醫療應用場景中造成兩個問題：

1. 醫療人員難以信任無法解釋的預測結果
2. 無法提供患者具體可行的風險因子改善建議

1.2.2 核心研究問題

基於上述背景，本研究提出以下核心研究問題：

如何利用縱向健檢資料，有效預測個體未來罹患三高疾病的風險，同時兼顧預測準確性與模型可解釋性？

1.2.3 具體研究問題

為回答上述核心問題，本研究設定以下六個具體研究問題：

Q1：變化量特徵的預測價值 在相同的健檢時間點數量下，額外納入變化量特徵（ΔFeatures）是否能顯著提升三高疾病的預測性能？

Q2：模型選擇與比較 在三高疾病預測任務中，哪些機器學習模型表現最佳？傳統統計方法與機器學習方法的性能差異為何？可解釋模型與黑箱模型之間如何權衡？

Q3：多任務學習的效果 同時預測三高疾病（Multi-Task Learning）是否優於分別預測單一疾病（Single-Task Learning）？

Q4：特徵重要性分析 哪些生物標記及其變化量對三高疾病預測最為重要？這些發現如何支持臨床決策？

Q5：時間點選擇策略 使用健檢資料的不同時間區間（如前三次 vs 後三次健檢），對預測性能有何影響？

Q6：健檢次數對預測性能的影響 累積更多次健檢紀錄是否能提升預測準確度？此結果對鼓勵民眾定期健檢及健檢機構的服務規劃有何啟示？

1.3 研究目標

1.3.1 主要目標

本研究的主要目標是：

建立一個基於縱向健檢資料的三高疾病預測系統，此系統能夠：

1. **準確預測**：利用歷史健檢資料預測個體未來罹患三高疾病的風險
2. **利用動態資訊**：透過變化量特徵（ΔFeatures）捕捉健康狀態的動態變化
3. **提供可解釋結果**：識別關鍵風險因子，支持臨床決策

1.3.2 次要目標

除主要目標外，本研究同時設定以下次要目標：

目標一：縱向特徵工程驗證

- 驗證變化量特徵（ΔFeatures）在三高同時預測場景下的有效性
- 探索不同時間窗口的特徵組合策略

目標二：模型比較研究

- 系統性比較多種機器學習模型，包含傳統統計方法、樹模型、深度學習與符號回歸
- 評估可解釋性與預測性能之間的權衡關係

目標三：多任務學習探索

- 驗證同時預測三高疾病的多任務學習（MTL）架構效果
- 分析三高疾病之間的共享風險因子

目標四：臨床應用指引

- 識別高風險族群的關鍵特徵
- 提供具體可行的個人化健康管理建議

1.4 研究貢獻

本研究預期在學術與應用層面做出以下貢獻：

1.4.1 學術貢獻

貢獻一：縱向變化量特徵的跨疾病驗證

本研究系統性驗證變化量特徵（Δ Features）在三高疾病同時預測場景下的效果。既有研究（如 Kanegae et al. 2020、Yang et al. 2025）已分別在高血壓與糖尿病預測中證實 Δ 特徵的價值，但尚未有研究將此方法同時應用於三高疾病並進行完整的消融實驗。本研究透過系統性的比較，提供 Δ 特徵在不同疾病間適用性的實證依據。

貢獻二：全面的模型比較研究

本研究比較多種類型的機器學習模型，涵蓋：

- 傳統統計方法（Logistic Regression、Naive Bayes、LDA）
- 樹模型（Decision Tree、Random Forest、XGBoost）
- 深度學習（MLP）
- 符號回歸（PySR）

這種跨類型的系統性比較，可為後續研究與臨床應用提供模型選擇的實證依據。

貢獻三：可解釋性與性能的權衡分析

本研究同時關注模型的預測性能與可解釋性，探討兩者之間的權衡關係，為醫療 AI 應用的模型選擇提供指引。

1.4.2 應用貢獻

貢獻一：早期預警系統原型

本研究成果可作為健檢中心部署早期預警系統的基礎，自動標註高風險個體，提供個人化預防建議。

貢獻二：臨床決策支持

透過特徵重要性分析，本研究可識別可干預的風險因子，輔助醫師進行臨床判斷與衛教。

貢獻三：公共衛生效益

長期而言，本研究有助於：

- 降低三高疾病的發生率
- 減少相關醫療支出
- 提升國民整體健康水平

1.5 論文架構

本論文共分為五章，各章內容安排如下：

第一章　緒論

說明研究背景、動機、問題陳述、研究目標與預期貢獻，引導讀者了解本研究的定位與價值。

第二章　文獻探討

回顧三高疾病預測的相關文獻，包括傳統風險評估方法、機器學習應用、縱向資料分析，以及變化量特徵工程的相關研究。透過文獻回顧，識別現有研究缺口，奠定本研究的理論基礎。

第三章　研究方法

詳細說明本研究的方法論，包括資料來源與前處理、特徵工程設計、模型選擇與訓練策略、實驗設計，以及評估指標的選用。

第四章　實驗結果

呈現各項實驗的結果與分析，包括模型比較、消融實驗、特徵重要性分析等，並對結果進行討論與詮釋。

第五章　結論與未來展望

總結本研究的主要發現與貢獻，討論研究限制，並提出未來研究方向的建議。

第二章 文獻探討

本章回顧與本研究相關的文獻，包括三高疾病預測研究、縱向資料分析與變化量特徵工程、機器學習方法，以及相關研究的比較分析。最後定義本研究的問題框架與評估指標。

2.1 三高疾病預測研究

本研究使用的資料集來自 Luo et al. (2024) 之公開資料，三高疾病的確診狀態依據以下標準標記：高血壓定義為 SBP ≥ 140 或 DBP ≥ 90 mmHg，或已確診且正在服用降壓藥物；高血糖定義為 FBG ≥ 7.0 mmol/L 或自我報告糖尿病；高血脂定義為 TC ≥ 6.22 mmol/L。上述閾值與國際通用的診斷標準一致（James et al., 2014；ADA, 2025；NCEP, 2002）。

2.1.1 高血壓預測

Sun et al. (2017) 系統性回顧了 26 篇高血壓預測研究，共涵蓋 48 個預測模型。該回顧指出，常見的風險因子包括 BMI、年齡、血壓水平、吸菸與家族史等，而統計方法以 Logistic Regression（12 篇）、COX Regression（7 篇）和 Weibull Regression（6 篇）為主，顯示傳統統計方法在該領域長期居於主流地位。

近年來，機器學習方法逐漸被引入高血壓預測。Kanegae et al. (2020) 使用日本職場健檢資料（18,258 人）建立高血壓預測模型，採用 XGBoost 和 Ensemble 方法，達到 AUC 0.881。該研究的重要貢獻在於使用縱向變化量特徵（Year(-2) → Year(-1) → Year(0)），證明 Δ 特徵在高血壓預測上的有效性。

Ye et al. (2018) 使用美國 Maine 州的電子健康紀錄（EHR），以 823,627 人的回顧性資料和 680,810 人的前瞻性資料，採用 XGBoost 建立一年期高血壓預測模型，回顧性驗證 AUC 達 0.917，前瞻性驗證 AUC 為 0.870。然而，後續評論指出該研究的前五名重要特徵均為降壓藥物，可能存在資料洩漏問題，提醒研究者在特徵選擇時需審慎避免將結果資訊混入預測因子。

Wang et al. (2024) 使用台灣美兆（MJ）健檢資料進行大規模研究（207,488 人），發現健檢次數越多，預測準確度越高（4 次以上最佳），達到 AUC 0.889。此研究支持多時間點特徵串接的設計理念，與本研究的縱向設計概念一致。

2.1.2 高血糖與糖尿病預測

Liu et al. (2024) 使用台中榮總電子病歷（6,687 人，追蹤 10 年），以 XGBoost 達到 AUC 0.93，關鍵特徵包括 HbA1c、空腹血糖、體重等。

Yang et al. (2025) 同樣使用 MJ 健檢資料（6,247 位 18-35 歲男性），提出雙框架設計同時預測血糖變化量（δ-FPG）與前驅糖尿病風險。研究發現基線空腹血糖（FPGbase）對預測 δ-FPG 的重要性達 100%，遠超第二名體脂肪的 17.64%，顯示縱向血糖變化具有高度可預測性。本研究的 Δ 特徵設計即參考此概念。

2.1.3 高血脂預測

相較於高血壓與糖尿病，高血脂的機器學習預測研究較少。多數研究將高血脂作為心血管疾病的風險因子，而非獨立的預測目標。本研究將高血脂納入三高同時預測的框架中，填補此研究缺口。

2.2 縱向資料分析與變化量特徵

2.2.1 縱向研究設計

縱向研究（Longitudinal Study）追蹤同一群體在不同時間點的變化，相較於橫斷面研究（Cross-sectional Study）具有以下優勢：

1. **捕捉動態變化**：能觀察生理指標隨時間的趨勢
2. **時序因果關係**：可建立預測因子與結果的時間順序
3. **個體內變異**：控制個體間差異，專注於個體內的變化

然而，縱向資料也面臨挑戰，包括追蹤期間的樣本流失、時間間隔不一致、以及缺失值處理等問題。

2.2.2 變化量特徵工程

變化量特徵（Delta Features）定義為兩個時間點之間生理指標的差值：

$$\delta_j = x_{j,Y_{-1}} - x_{j,Y_{-2}}$$

其中 x_j 為第 j 個生理指標。Yang et al. (2025) 以 δ-FPG（空腹血糖變化量）作為預測目標，證明縱向變化量具有高度可預測性。Kanegae et al. (2020) 同樣使用 Δ 特徵預測高血壓，證明此方法的跨疾病適用性。

本研究採用八個變化量特徵：ΔSBP、ΔDBP、ΔFBG、ΔTC、ΔCr、ΔUA、ΔeGFR、ΔBMI，分別捕捉血壓、血糖、血脂、腎功能與身體質量指數的動態變化。

2.3 傳統統計方法

根據 Sun et al. (2017) 的系統性回顧，在 26 篇高血壓預測研究所涵蓋的 48 個模型中，Logistic Regression 佔 12 篇（25%）為最大宗，其次為 COX Regression（7 篇）和 Weibull Regression（6 篇），顯示傳統統計方法長期作為疾病風險預測的主流工具。本節介紹本研究所採用的三種傳統統計分類方法。

2.3.1 Logistic Regression

Logistic Regression（邏輯斯迴歸）是疾病預測研究中最常用的基準模型。其模型形式為：

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_nX_n)}}$$

其中 β_i 為迴歸係數，e^{β_i} 可直接解釋為勝算比（Odds Ratio），表示第 i 個特徵每增加一個單位時，疾病風險的倍數變化。此特性使 LR 在臨床應用中具有高度可解釋性，醫療人員可直觀理解各風險因子的貢獻程度。

在 Sun et al. (2017) 回顧的研究中，LR 的 C-statistic（等同 AUC）多落在 0.72–0.85 之間，顯示即使在非線性關係存在的情境下，LR 仍能提供具競爭力的預測效能。然而，LR 假設特徵與對數勝算之間為線性關係，可能無法捕捉複雜的非線性交互作用。

2.3.2 Naive Bayes

Naive Bayes（單純貝氏分類器）基於貝氏定理進行分類：

$$P(Y=k|X)=\frac{P(X|Y=k)\cdot P(Y=k)}{P(X)}$$

其核心假設為各特徵在給定類別下條件獨立，即：

$$P(X|Y=k) = \prod_{j=1}^{n} P\left(X_j|Y=k\right)$$

對於連續型特徵，Gaussian Naive Bayes 假設每個特徵在各類別下服從常態分佈。此方法的優點包括：計算效率極高（時間複雜度為 $O(nd)$，n 為樣本數、d 為特徵數）、無需調參、在小樣本情境下表現穩健。然而，特徵獨立假設在醫療資料中往往不成立，例如收縮壓與舒張壓、血糖與 BMI 之間均存在相關性，此假設的違反可能影響機率估計的校準度，但對分類排序（AUC）的影響通常較小。

Naive Bayes 屬於生成式模型（Generative Model），與 Logistic Regression 的判別式模型（Discriminative Model）形成理論上的互補，兩者的比較有助於理解資料的分佈特性。

2.3.3 Linear Discriminant Analysis

Linear Discriminant Analysis（線性判別分析，LDA）由 Fisher (1936) 提出，是統計學中最經典的分類方法之一。LDA 透過尋找最佳的線性投影方向，最大化類別間變異與類別內變異的比值：

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

其中 S_B 為類別間散布矩陣（Between-class scatter matrix），S_W 為類別內散布矩陣（Within-class scatter matrix），w 為投影方向。LDA 假設各類別的特徵服從多變量常態分佈且共享相同的共變異數矩陣。

相較於 Logistic Regression，LDA 同時考慮特徵的聯合分佈結構，在特徵間存在多重共線性時仍能維持穩定性。此外，LDA 的降維特性（將 d 維特徵投影至最多 $k-1$ 維空間，k 為類別數）使其在高維度資料中具有正則化效果。然而，常態分佈與等共變異數假設在實際資料中可能不完全成立，限制了其對非線性關係的捕捉能力。

2.4 機器學習方法

2.4.1 樹狀模型

Decision Tree（決策樹）透過遞迴分割建立規則，具有高度可解釋性。Random Forest（隨機森林）是決策樹的集成方法，透過 Bagging 降低過擬合風險。XGBoost

採用梯度提升策略，在多項疾病預測競賽中表現優異。Alaa et al. (2019) 使用 AutoPrognosis 自動化機器學習預測心血管疾病，達到 AUC 0.774。Liu et al. (2024) 和 Yang et al. (2025) 皆報告 XGBoost 達到最佳預測效能。Dinh et al. (2019) 使用 NHANES 公開資料集（21,131 筆）以 XGBoost 預測糖尿病，不含實驗室數據時達到 AUC 0.862，並以 Information Gain 進行特徵重要性分析。

2.4.2 支援向量機

Support Vector Machine（SVM）透過尋找最大間隔超平面（Maximum Margin Hyperplane）進行分類，並可使用核函數（Kernel Function）將資料映射至高維空間以處理非線性問題。相較於樹模型依賴特徵的離散分割，SVM 在特徵空間中建立連續的決策邊界，對小樣本與高維資料具有較好的泛化能力。Yang et al. (2025) 在七種機器學習模型的比較中使用 SVM。本研究採用 SVM 作為核方法（Kernel Method）的代表，與線性方法（LR、NB、LDA）、樹模型（DT、RF、XGBoost）及神經網路（MLP）形成四類方法的完整比較架構。

2.4.3 神經網路

多層感知器（MLP）可學習複雜的非線性關係，但面臨可解釋性不足與過擬合風險。Taiwan MTL (2025) 使用 Attention 機制進行多疾病預測，透過注意力分數提供一定程度的可解釋性。

2.4.4 符號回歸

符號回歸（Symbolic Regression）透過遺傳規劃演化出可解釋的數學公式（Cranmer, 2023）。相較於黑盒模型，符號回歸產出的公式可直接理解其醫學意義。然而，符號回歸的搜尋過程具有隨機性，結果穩定性較低。

2.5 類別不平衡處理

三高疾病的發病率通常低於 20%，造成正負類別樣本數量懸殊的類別不平衡（Class Imbalance）問題。在此情境下，模型容易偏向預測多數類（健康），導致少數類（患病）的識別率低落。He & Garcia (2009) 將類別不平衡的處理策略歸納為兩大層面：資料層面與演算法層面。

2.5.1 資料層面方法

資料層面方法透過調整訓練資料的類別分佈來緩解不平衡問題，主要包括過採樣（Over-sampling）與欠採樣（Under-sampling）兩類策略。

過採樣方面，SMOTE（Synthetic Minority Over-sampling Technique）是最具代表性的方法（Chawla et al., 2002）。SMOTE 透過在少數類樣本的特徵空間中進行線性內插，生成合成樣本，避免了簡單複製造成的過擬合問題。其衍生方法包括 Borderline-SMOTE（僅對邊界樣本進行合成）和 ADASYN（根據樣本學習難度自適應生成）。然而，過採樣方法可能引入雜訊樣本，且在高維特徵空間中，合成樣本的品質難以保證。

欠採樣方面，Random Under-sampling 隨機移除多數類樣本以達到類別平衡，但可能丟失重要資訊。Tomek Links 和 Edited Nearest Neighbours（ENN）等方法則透過移除邊界區域的多數類樣本來清理決策邊界，在保留資訊的同時改善類別分離度。

2.5.2 演算法層面方法

演算法層面方法在不改變資料分佈的前提下，透過修改學習演算法本身來處理不平衡問題。

成本敏感學習（Cost-sensitive Learning） 是最常用的演算法層面策略。其核心思想是對不同類別的誤分類賦予不同的代價（cost），使模型更重視少數類的正確分類。在實務上，scikit-learn 等框架提供 class_weight 參數，設定為 'balanced' 時會自動依據類別頻率的倒數調整權重：

$$w_k = \frac{n}{K \cdot n_k}$$

其中 n 為總樣本數，K 為類別數，n_k 為第 k 類的樣本數。此方法的優點在於不改變訓練資料的原始分佈，避免了合成樣本可能引入的雜訊，且計算成本極低。

決策門檻調整（Threshold Moving） 則在模型訓練後，調整分類的機率門檻（預設 0.5）來平衡 Sensitivity 與 Specificity。此方法不影響模型訓練過程，但需要額外的驗證集來選定最佳門檻。

本研究採用 class_weight='balanced' 作為主要的類別不平衡處理策略，並以 AUC-ROC 作為主要評估指標（因 AUC 不受門檻選擇影響），同時報告 Sensitivity 和 Specificity 以反映臨床應用需求。實驗中亦比較了 SMOTE 與 class_weight 兩種策略的效果差異（詳見第四章）。

2.6 研究缺口與本研究定位

2.6.1 相關研究比較

表 2-1 比較本研究與相關文獻的差異。

研究	預測目標	資料來源	樣本數	最佳模型	AUC	Δ特徵	可解釋性
Ye et al. (2018)	高血壓	Maine EHR(美國)	823,627	XGBoost	0.917	無	特徵重要性
Alaa et al. (2019)	心血管疾病	UK Biobank	423,604	AutoPrognosis	0.774	無	—
Dinh et al. (2019)	糖尿病	NHANES	21,131	XGBoost	0.862	無	Info. Gain
Kanegae et al.(2020)	高血壓	日本職場健檢	18,258	XGBoost	0.881	有	特徵重要性
Hung et al. (2021)	隱匿性高血壓	台灣醫院	1,386	RF	0.851	無	—
Liu et al.	糖尿病	台中榮總EHR	6,687	XGBoost	0.930	無	特徵重要性

研究	預測目標	資料來源	樣本數	最佳模型	AUC	Δ特徵	可解釋性
(2024)							
Wang et al. (2024)	高血壓	台灣美兆	207,488	XGBoost	0.889	無	特徵重要性
Yang et al. (2025)	前驅糖尿病	台灣美兆	6,247	XGBoost	—	有	SHAP
Majcherek et al.(2025)	糖尿病	BRFSS (美國)	253,680	Extra Trees	0.99	無	SHAP
本研究	**三高（同時）**	杭州社區調查	6,056	LR / XGBoost	0.721–0.938	**有**	**SHAP + 符號回歸**

註：Ye、Alaa、Kanegae、Liu、Wang、Yang 及本研究為縱向研究設計；

Dinh、Hung、Majcherek 為橫斷面研究。

2.6.2 研究缺口

綜觀現有文獻，本研究識別以下研究缺口：

1. **多疾病同時預測**：多數研究僅針對單一疾病，缺乏三高疾病的綜合預測框架

2. **變化量特徵的系統性驗證**：雖然 Yang et al. (2025) 和 Kanegae et al. (2020) 證明 Δ 特徵有效，但僅針對單一疾病，缺乏跨疾病的驗證

3. **模型比較的完整性**：現有研究通常只比較少數模型，缺乏傳統統計、樹模型、神經網路、符號回歸的全面比較

4. **可解釋性與效能的平衡**：多數研究偏重預測效能，較少探討臨床可解釋性

2.7 問題定義

本研究使用連續三次健檢紀錄，定義以下時間點：

- Y-2：第一次健檢（最早）
- Y-1：第二次健檢
- Y_0：第三次健檢（預測目標時間點）

2.7.1 特徵定義

給定 d 個基本特徵與 p 個健檢指標，輸入特徵向量定義為：

$$X = [X_{base}, X_{Y-2}, X_{Y-1}, \Delta X] \in \mathbb{R}^{d+3p}$$

其中：

- $X_{base} \in \mathbb{R}^d$：人口學基本資訊
- $X_{Y-2} \in \mathbb{R}^p$：Y-2 時間點的健檢指標
- $X_{Y-1} \in \mathbb{R}^p$：Y-1 時間點的健檢指標
- $\Delta X = X_{Y-1} - X_{Y-2} \in \mathbb{R}^p$：變化量特徵

2.7.2 預測任務

給定輸入特徵 X，學習預測函數 f 使得：

$$\hat{Y} = f(X) = [\hat{y}_{HTN}, \hat{y}_{HG}, \hat{y}_{DL}] \approx Y \in \{0,1\}^3$$

其中：

- $\hat{y}_{HTN}$：高血壓（Hypertension）
- $\hat{y}_{HG}$：高血糖（Hyperglycemia）
- $\hat{y}_{DL}$：高血脂（Dyslipidemia）

第三章 研究方法

3.1 研究架構

本研究旨在建立一個基於縱貫性健康檢查資料的三高（高血壓、高血糖、高血脂）風險預測模型。研究架構如圖 3-1 所示，整體流程分為四個階段：資料前處理、特徵工程、模型建立與評估。

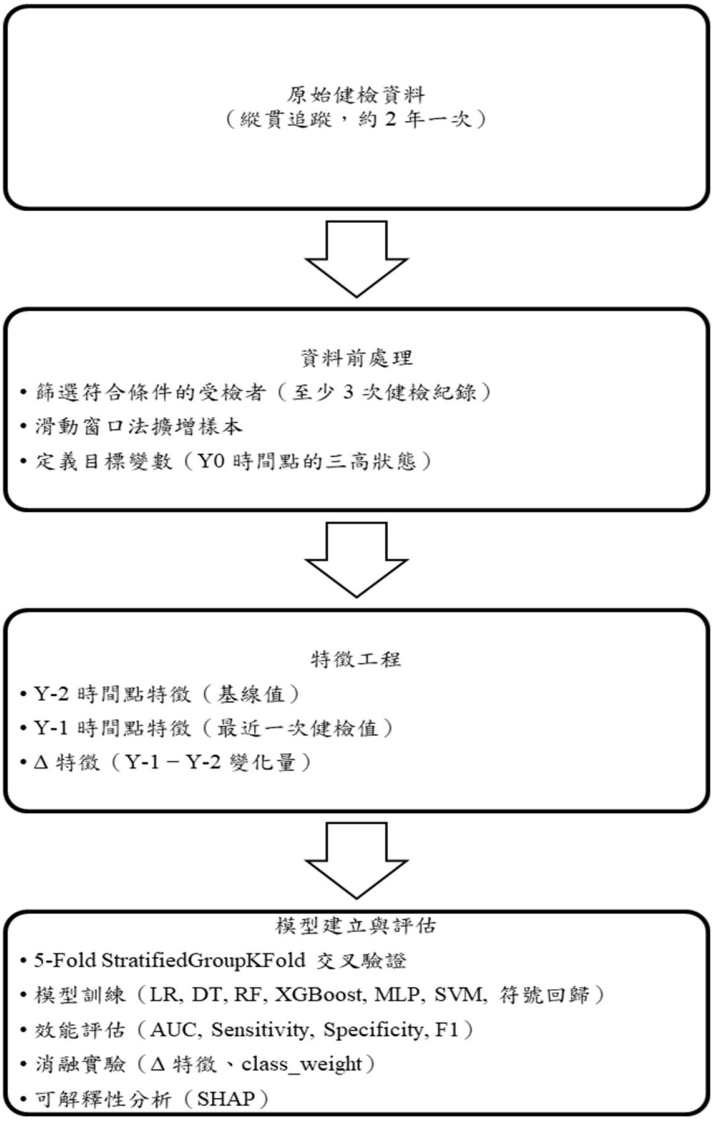

圖 3-1 研究架構圖

3.1.1 研究時間軸設計

本研究採用三個時間點的縱貫設計，如圖 3-2 所示。時間點命名採用相對於預測目標年（Y0）的方式：Y-2 為四年前、Y-1 為兩年前、Y0 為預測目標年。

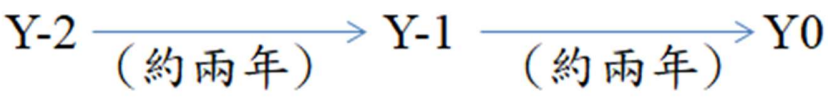

圖 3-2 研究時間軸設計

模型的輸入特徵包含 Y-2 與 Y-1 兩個時間點的健檢指標（SBP、DBP、FBG、TC 等），以及兩時間點之間的變化量（Δ 特徵）。預測目標為 Y0 時間點是否罹患三高（高血壓、高血糖、高血脂）。

選擇 Y0 而非 Y-1 作為預測目標的原因：

1. **避免資料洩漏**：若以 Y-1 為目標，Y-1 的健檢數據與疾病狀態來自同一次檢查，會造成模型「偷看答案」
2. **Δ 特徵可用**：以 Y0 為目標，才能將 Y-1 與 Y-2 的變化量作為有效的預測因子
3. **臨床意義**：提供約 2 年的預警時間窗口，讓醫療人員有足夠時間進行早期介入

3.2 資料來源與處理

3.2.1 資料來源

本研究使用公開於 Dryad 數位資料庫的縱貫性健康檢查資料集（Luo et al., 2024）。

該資料集來自中國浙江省杭州市的社區健康調查，收集期間為 2010 至 2018 年，納入 40 歲以上成人共 6,119 人，多數參與者進行了 3 次以上的健康檢查。

資料特點為僅記錄「第幾次健檢」而無具體日期，追蹤間隔以年齡差推算（例如：55 歲 → 57 歲 = 2 年間隔）。經分析，約 90% 的受檢者維持固定 **2 年間隔**，9.6% 為 1 年間隔（可能為提前回診），平均追蹤間隔為 1.90 年（標準差 0.36 年）。因此，本研究的時間點命名為 Y-2（四年前）、Y-1（兩年前）、Y0（預測目標），反映實際的健檢間隔。由於間隔高度一致，Δ 特徵可直接比較，無需額外的時間校正。

3.2.2 樣本篩選

原始資料集包含 6,119 位參與者共 25,744 筆健檢記錄。由於本研究採用三時間點縱貫設計（Y-2、Y-1、Y0），需要每位參與者至少有 3 次健檢紀錄才能建構完整的特徵集與預測目標。

納入條件： - 至少有 3 次以上的連續健檢紀錄 - 各時間點資料完整，無重大缺失

排除情況： - 共 63 人因僅有 1-2 次健檢紀錄而被排除 - 資料保留率達 98.97%

最終樣本數：6,056 人

篩選後樣本之健檢次數分佈如圖 3-3 所示。約 90% 的樣本健檢次數介於 3 至 5 次之間，其中以 5 次健檢者最多（31.95%），其次為 4 次（29.33%）及 3 次（28.96%）。少數樣本有 6 次以上的健檢紀錄（合計 9.76%）。

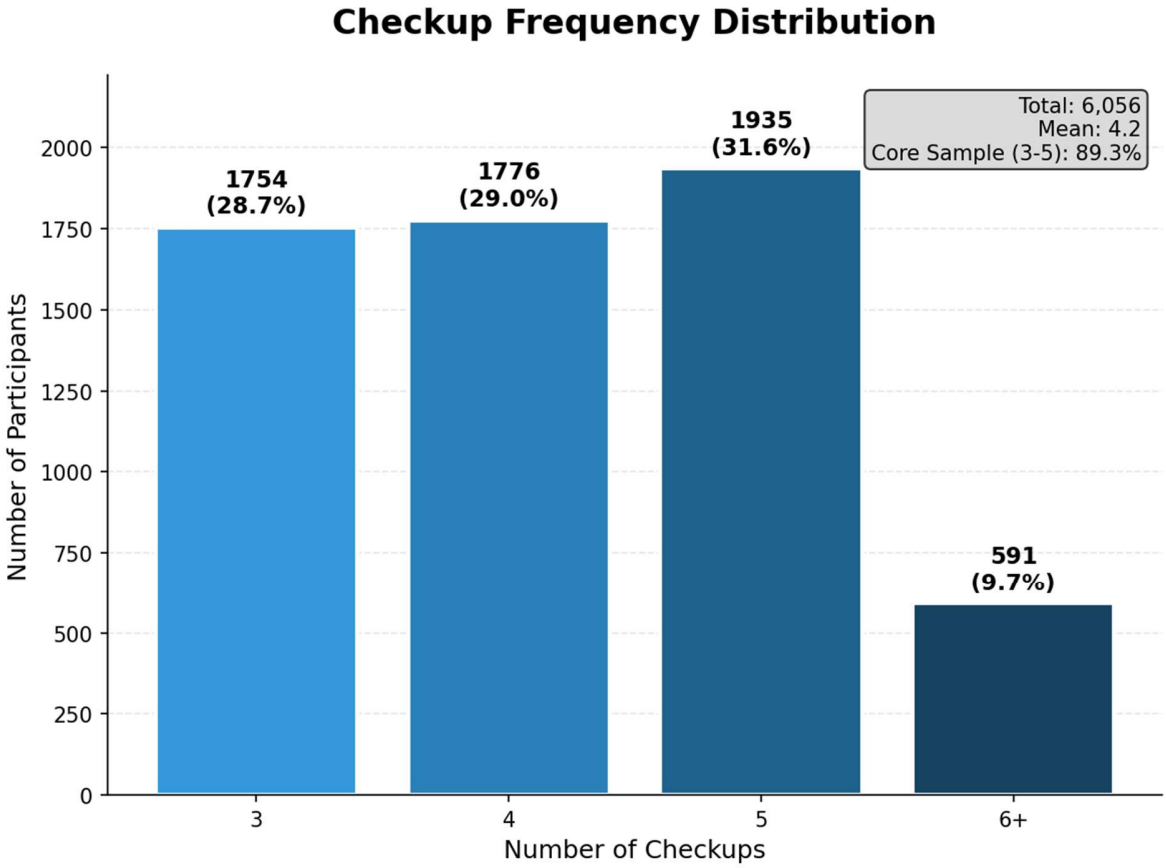

圖 3-3 樣本健檢次數分佈（n = 6,056）

3.2.3 滑動窗口法

為充分利用多次健檢資料，本研究採用滑動窗口（Sliding Window）方法擴增訓練樣本。對於有 N 次健檢紀錄的參與者，可產生 (N-2) 個訓練樣本：

- 3 次健檢 → 1 個樣本：(Y-2, Y-1, Y0)
- 4 次健檢 → 2 個樣本：(Y-3, Y-2, Y-1)、(Y-2, Y-1, Y0)
- 5 次健檢 → 3 個樣本：(Y-4, Y-3, Y-2)、(Y-3, Y-2, Y-1)、(Y-2, Y-1, Y0)

經滑動窗口處理後，6,056 位參與者共產生 **13,514 筆建模紀錄**。此方法的優點：

1. **充分利用資料**：多次健檢者貢獻更多樣本
2. **捕捉不同階段**：同一人在不同年齡階段的健康變化皆納入分析

需注意的是，由於同一參與者可能產生多筆紀錄，在交叉驗證時必須確保同一人的所有紀錄不會同時出現在訓練集與測試集中（詳見 3.5.1 節）。

3.2.4 變數定義

本資料集包含人口學變數、健檢指標及目標變數三類，各變數說明如表 3-1 所示。

表 3-1 研究變數定義

變數類別	變數名稱	說明	單位/編碼
人口學	Sex	性別	1=男, 2=女
	Age	年齡	歲
健檢指標	BMI	身體質量指數	kg/m²
	SBP	收縮壓	mmHg
	DBP	舒張壓	mmHg
	FBG	空腹血糖	mmol/L
	TC	總膽固醇	mmol/L
	Cr	肌酐	μmol/L
	eGFR	腎絲球過濾率	mL/min/1.73m²
	UA	尿酸	μmol/L

本研究之目標變數為三高疾病狀態（高血壓、高血糖、高血脂），由資料集中的確診欄位直接取得。原始資料中，三項目標變數皆以 1＝ 正常、2＝ 患病 進行編碼，本研究於建模前將其轉換為 0＝ 正常、1＝ 患病 之二元格式。

3.2.5 類別不平衡情況

三高疾病在本資料集中呈現不同程度的類別不平衡。在 6,056 位樣本中，高血壓患者共 1,010 人（16.68%），負正類比例約為 5:1，屬於輕度不平衡；高血糖患者共 335 人（5.53%），負正類比例約為 17:1；高血脂患者共 361 人（5.96%），負正類比例約為 16:1，兩者皆屬於重度不平衡。

此類別不平衡現象反映了真實世界中三高疾病的盛行率特性，但可能導致模型偏向預測多數類（健康），進而降低對少數類（患病）的識別能力。因此，本研究將於模型訓練階段採用 class_weight 方法進行調整，詳見 3.4.6 節。

3.3 特徵工程

3.3.1 特徵集設計

本研究使用的特徵分為四類，共 26 個特徵，如表 3-2 所示。

表 3-2 特徵集設計

特徵類別	包含特徵	特徵數
基本資訊	Sex, Age	2
Y-2 時間點特徵	FBG_Y-2, TC_Y-2, Cr_Y-2, UA_Y-2, eGFR_Y-2, BMI_Y-2, SBP_Y-2, DBP_Y-2	8
Y-1 時間點特徵	FBG_Y-1, TC_Y-1, Cr_Y-1, UA_Y-1, eGFR_Y-1, BMI_Y-1, SBP_Y-1, DBP_Y-1	8
Δ 特徵（Y-1 − Y-2）	ΔFBG, ΔTC, ΔCr, ΔUA, ΔeGFR, ΔBMI, ΔSBP, ΔDBP	8
合計	—	26

3.3.2 Δ 特徵的意義

Δ 特徵代表 Y-1 與 Y-2 之間的變化量：

$$\Delta_i = X_{i,Y-1} - X_{i,Y-2}$$

Δ 特徵的設計理念：

- **捕捉動態趨勢**：某些疾病的發展不僅取決於當前數值，更取決於變化趨勢
- **正值代表上升**：例如 $\Delta FBG > 0$ 表示血糖在兩年間上升
- **負值代表下降**：例如 $\Delta eGFR < 0$ 表示腎功能在兩年間下降

3.4 模型方法

本研究比較多種類型的預測模型，依方法論性質分為以下類別：傳統統計方法、樹狀模型、支援向量機、神經網路，以及符號回歸。各類別模型的詳細說明如下。

3.4.1 傳統統計方法

傳統統計方法具有明確的數學形式與統計假設，可解釋性高，常作為基準模型使用。

Logistic Regression (LR)

Logistic Regression 是一種經典的線性分類模型，適用於二元分類問題。

模型形式：

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\ldots+\beta_nX_n)}}$$

選用原因： - 可解釋性高：係數可直接解讀為風險因子的貢獻 - 計算效率高：適合作為基準模型 - 支援 class_weight：可處理類別不平衡問題

實作參數： - *solver*: 'lbfgs' - *max_iter*: 1000 - *class_weight*: 'balanced'（處理類別不平衡）

Naive Bayes (NB)

Naive Bayes 是一種基於貝氏定理（Bayes' Theorem）的機率分類器，假設各特徵在給定類別下條件獨立。

模型形式：

$$P(Y|X_1,\ldots,X_n)\propto P(Y)\prod_{i=1}^{n}P\left(X_i|Y\right)$$

本研究使用 Gaussian Naive Bayes，假設連續特徵在各類別下服從常態分佈：

$$P(X_i|Y=k)=\frac{1}{\sqrt{2\pi\sigma_{ik}^2}}\exp\left(-\frac{(X_i-\mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

選用原因： - 計算效率極高：僅需估計各特徵的均值與變異數 - 理論基礎明確：基於機率推論框架 - 適合作為基準：在特徵獨立性假設合理時表現穩健

實作參數：

- 使用 scikit-learn 的 *GaussianNB*
- *priors*: None（依訓練資料自動估計類別先驗機率）

注意：Naive Bayes 的條件獨立假設在實務中通常不完全成立（如 SBP 與 DBP 高度相關），但即便假設違反，其分類表現仍可作為有意義的參考基準。

Linear Discriminant Analysis (LDA)

線性判別分析（Linear Discriminant Analysis, LDA）透過最大化類別間變異與類別內變異的比值，尋找最佳線性投影方向進行分類。

演算法原理：

尋找投影方向 w，使得 Fisher 準則最大化：

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

其中 S_B 為類別間散佈矩陣，S_W 為類別內散佈矩陣。

選用原因： - 兼具降維與分類功能：可同時降低特徵維度 - 考慮類別分佈結構：利用共變異數矩陣進行判別 - 計算效率高：無需迭代優化

實作參數：

- 使用 scikit-learn 的 *LinearDiscriminantAnalysis*
- *solver*: 'svd'（奇異值分解，適合特徵數多於樣本數的情況）
- *priors*: None（依訓練資料自動估計）

3.4.2 樹狀模型

樹狀模型透過遞迴分割特徵空間進行預測，從單一決策樹到集成方法，兼具可解釋性與預測能力。

Decision Tree (DT)

決策樹是一種基於規則的分類模型，透過遞迴地將資料依特徵值分割成子集，最終形成樹狀結構。

演算法原理：

1. 選擇最佳分割特徵（依 Gini 或 Entropy 指標）
2. 依該特徵的閾值將資料分成兩個子集
3. 遞迴執行直到滿足停止條件（如深度限制或樣本數不足）

選用原因：

- 高度可解釋：分類規則可直接呈現為 if-then 規則
- 計算效率高：訓練與預測速度快
- 支援 class_weight：可處理類別不平衡問題

實作參數：

- *criterion*: 'gini'
- *max_depth*: None（完全生長）
- *class_weight*: 'balanced'

注意：單一決策樹容易過擬合，預測效能通常低於集成方法，但因其高可解釋性，仍納入比較。

Random Forest (RF)

Random Forest 是一種基於 Bagging 的集成學習方法，透過多棵決策樹的投票產生預測結果。

演算法原理：

1. 從原始資料中有放回地抽樣（Bootstrap）產生多個子資料集
2. 在每個子資料集上訓練一棵決策樹，且每次分裂時只考慮部分特徵
3. 最終預測為所有樹的多數決（分類）或平均（回歸）

選用原因：

- 抗過擬合：Bagging 降低變異數
- 穩定性高：對異常值和雜訊較不敏感
- 支援 class_weight：可處理類別不平衡問題

實作參數：

- *n_estimators: 100*
- *max_depth: None（完全生長）*
- *class_weight: 'balanced'*

XGBoost

XGBoost (eXtreme Gradient Boosting) 是一種基於梯度提升的集成學習方法。

演算法原理： 透過逐步加入決策樹，每棵新樹專注於修正前面樹的預測誤差：

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

選用原因：

- 預測效能強：在許多醫學預測任務中表現優異
- 可處理非線性關係：能捕捉特徵間的複雜交互作用
- 支援特徵重要性評估

實作參數：

- *n_estimators: 100*
- *max_depth: 6*
- *learning_rate: 0.1*
- *scale_pos_weight: 自動計算（處理類別不平衡）*

3.4.3 支援向量機

支援向量機（Support Vector Machine, SVM）基於統計學習理論，透過尋找最大間隔超平面進行分類，並可使用核函數處理非線性問題。

演算法原理： 尋找一個超平面 $w^T x + b = 0$，使得兩類樣本之間的間隔最大化。對於非線性問題，使用核函數將資料映射到高維空間。

選用原因：

- 理論基礎扎實：基於統計學習理論
- 適合中小型資料集：在樣本數有限時表現良好
- 支援 class_weight：可處理類別不平衡問題

實作參數：

- *kernel*: 'rbf'（徑向基函數）
- *C*: 1.0
- *gamma*: 'scale'
- *class_weight*: 'balanced'

3.4.4 神經網路

Multi-Layer Perceptron (MLP)

多層感知器（MLP）是一種前饋神經網路，透過多層神經元的非線性轉換學習複雜的特徵表示。本研究使用 scikit-learn 的 MLPClassifier 實作。

網路架構：

- 輸入層：26 個特徵
- 隱藏層：2 層，每層 64 個神經元
- 輸出層：二元分類

選用原因：

- 非線性建模：可學習複雜的特徵交互
- 彈性高：可調整網路深度與寬度
- 實作簡便：scikit-learn 提供統一的 API 介面

實作參數：

- *hidden_layer_sizes*: (64, 64)
- *activation*: 'relu'
- *solver*: 'adam'
- *max_iter*: 500

注意：MLPClassifier 不直接支援 class_weight 參數，本研究透過手動調整樣本權重處理類別不平衡問題。

3.4.5 符號回歸

符號回歸（Symbolic Regression）透過遺傳規劃（Genetic Programming, GP）演化出可解釋的數學公式。

演算法原理：

1. 初始化：隨機生成一群數學公式（個體）
2. 評估：計算每個公式的預測誤差（適應度）
3. 選擇：保留表現較好的公式
4. 演化：透過交叉、突變產生新公式
5. 重複直到收斂

選用原因：

- 完全透明：產出的公式可直接理解
- 領域知識驗證：可檢驗公式是否符合醫學邏輯
- 輕量部署：簡單公式不需複雜運算資源

使用套件：本研究初期使用 Python 原生的 gplearn 套件進行符號回歸實驗，但因其不支援 class_weight 且搜尋效率有限，後改用基於 Julia 的 PySR 套件（Cranmer, 2023）。PySR 支援 sample_weight 且搜尋效能更佳，為本研究最終採用之符號回歸工具。

3.4.6 類別不平衡處理

由於三高疾病的患病率較低（高血壓 16.68%、高血糖 5.53%、高血脂 5.96%），本研究採用 class_weight='balanced' 作為主要的類別不平衡處理策略（原理與公式詳見第二章 2.5.2 節）。各模型的具體設定如下：

- **LR、DT、RF、SVM**：設定 *class_weight='balanced'*，由 scikit-learn 自動依類別頻率倒數計算權重
- **XGBoost**：設定 *scale_pos_weight* 為負正類比例，效果等同 balanced
- **MLP**：透過手動調整 *sample_weight* 實現加權訓練
- **NB、LDA**：不支援 class_weight，以原始資料分佈訓練
- **PySR**：透過 *sample_weight* 參數加權

3.5 模型評估

3.5.1 交叉驗證策略

本研究採用 scikit-learn 提供的 **StratifiedGroupKFold** 進行 5-Fold 交叉驗證。此方法結合了分層抽樣（Stratified）與群組控制（Group）兩項特性：

1. **分層抽樣**：確保每個 fold 中各類別（患病/健康）的比例與整體資料集一致
2. **群組控制**：確保同一參與者的所有紀錄（由滑動窗口產生）不會同時出現在訓練集與測試集中

此設計的重要性在於：由於滑動窗口法使同一參與者可能貢獻多筆紀錄，若不進行群組控制，模型可能在訓練時學習到某位參與者的特徵模式，而在測試時遇到同一人的其他紀錄，造成評估結果過度樂觀（資料洩漏）。

交叉驗證資料規模

項目	數量
參與者數	6,056 人
建模紀錄數（滑動窗口後）	13,514 筆
Fold 數	5
每 Fold 約測試紀錄數	~2,703 筆

3.5.2 評估指標

AUC-ROC (Area Under the ROC Curve)

ROC 曲線以 False Positive Rate 為 X 軸，True Positive Rate 為 Y 軸，AUC 為曲線下面積。

- AUC = 0.5：隨機猜測
- AUC = 0.7-0.8：可接受
- AUC = 0.8-0.9：良好
- AUC > 0.9：優秀

特點：與分類閾值無關，反映模型的整體排序能力。

Sensitivity（敏感度/召回率）

$$Sensitivity = \frac{TP}{TP + FN}$$

代表模型正確識別患病者的能力。在疾病篩檢中，高 Sensitivity 意味較少漏診。

Specificity（特異度）

$$Specificity = \frac{TN}{TN + FP}$$

代表模型正確排除健康者的能力。高 Specificity 意味著較少誤診。

F1-Score

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

精確率與召回率的調和平均，適用於類別不平衡的情況。

混淆矩陣

混淆矩陣（Confusion Matrix）為評估分類模型效能的基礎工具，用於呈現模型預測結果與實際類別之間的對應關係。在二元分類問題中，混淆矩陣包含四個元素：

- **True Positive（TP，真陽性）**：實際為患病且模型正確預測為患病的樣本數
- **True Negative（TN，真陰性）**：實際為健康且模型正確預測為健康的樣本數
- **False Positive（FP，偽陽性）**：實際為健康但模型錯誤預測為患病的樣本數
- **False Negative（FN，偽陰性）**：實際為患病但模型錯誤預測為健康的樣本數

上述 Sensitivity、Specificity 及 F1-Score 皆由混淆矩陣計算而得。在健康篩檢應用中，漏診比誤診後果更嚴重，故本研究特別重視 Sensitivity。

未採用 PR-AUC 之說明

在極度類別不平衡（正樣本比例 ＜5%）的情境下，ROC-AUC 可能因大量真陰性而產生過度樂觀的評估，此時 Precision-Recall AUC（PR-AUC）被認為是更適合的指標（Saito & Rehmsmeier, 2015）。然而，本資料集三項疾病的盛行率分別為高血壓 16.68%、高血糖 5.53%、高血脂 5.96%，皆高於 5% 的極度不平衡門檻，且本研究已搭配 class_weight 調整與 Sensitivity/Specificity 報告，足以反映模型在少數類上的辨識能力，故本研究以 AUC-ROC 作為主要評估指標。

3.6 實驗設計

本研究設計一系列消融實驗（Ablation Study）與比較實驗，以驗證研究問題。實驗設計總覽如表 3-3 所示。

表 3-3 實驗設計總覽

實驗	目的
Δ 特徵消融	驗證變化量特徵的貢獻
模型比較	比較簡單與複雜模型效能
特徵選擇消融	驗證精簡特徵集的可行性
類別不平衡處理比較	比較不同不平衡處理方法
符號回歸實驗	探索可解釋數學公式

3.6.1 消融實驗

Δ 特徵消融實驗

為驗證 Δ 特徵對預測效能的貢獻，本研究設計五組特徵組合進行消融實驗。

表 3-4 Δ 特徵消融實驗設計

實驗組	特徵組合	特徵數	說明
Full	Y-2 + Y-1 + Δ	26	完整特徵集
No-Δ	Y-2 + Y-1	18	移除 Δ 特徵
Y-2-Only	Y-2	10	僅使用基線值
Y-1-Only	Y-1	10	僅使用最近值
Δ-Only	Δ	10	僅使用變化量

特徵選擇消融實驗

為驗證精簡特徵集的可行性，基於 SHAP 重要性排序設計消融實驗。

表 3-5 特徵選擇消融實驗設計

實驗組	特徵數	說明
Top 3	3	僅使用最重要的 3 個特徵
Top 5	5	僅使用最重要的 5 個特徵
Top 10	10	使用前 10 個重要特徵
All	26	使用全部特徵（基準線）

各疾病的 Top 5 特徵由 SHAP 分析結果決定，詳見第四章。

3.6.2 類別不平衡處理比較

由於三高疾病的患病率較低（5-17%），模型容易偏向預測多數類（健康），導致 Sensitivity 偏低。本研究比較五種類別不平衡處理方法，如表 3-6 所示。

表 3-6 類別不平衡處理方法比較

方法	類型	原理	特點
Baseline	無處理	使用原始資料分佈	作為對照基準
class_weight	權重調整	調高少數類損失函數權重	不改變資料分佈
SMOTE	過採樣	特徵空間中合成少數類樣	增加訓練樣本數
ADASYN	自適應過採樣	針對邊界少數類合成新樣	關注邊界樣本
RandomUnderSampler	欠採樣	隨機移除多數類樣本	可能損失資訊

class_weight 權重計算：

$$w_i = \frac{n_{samples}}{n_{classes} \times n_{sampl\ \ i}}$$

其中 w_i 為第 i 類的權重，$n_{samples}$ 為總樣本數，$n_{classes}$ 為類別數，$n_{samples_i}$ 為第 i 類的樣本數。

SMOTE 演算法原理：

1. 對每個少數類樣本，找出其 k 個最近鄰（預設 k=5）
2. 隨機選擇一個最近鄰
3. 在原樣本與選定鄰居之間的連線上隨機生成新樣本
4. 重複直到少數類與多數類樣本數平衡

3.6.3 符號回歸實驗

符號回歸旨在從資料中自動發現可解釋的數學公式。本研究使用 PySR 套件進行符號回歸實驗，實驗設計如表 3-7 所示。

表 3-7 符號回歸實驗設計

參數	設定	說明
套件	PySR	基於 Julia 的符號回歸套件
二元運算子	+, -, *, /	基本四則運算
一元運算子	exp, log, abs, square	數學轉換函數
最大複雜度 (maxsize)	35	控制公式長度上限
迭代次數 (niterations)	200	遺傳演算法迭代次數
複雜度懲罰 (parsimony)	0.0001	避免過於簡單的常數解
族群數 (populations)	20	平行搜索的族群數量
族群大小 (population_size)	100	每個族群的個體數

實驗流程：

1. 使用 5-Fold StratifiedGroupKFold 交叉驗證（按 patient_id 分組）
2. 對訓練集進行標準化（StandardScaler）
3. 在每個 fold 的訓練集上執行 PySR
4. 從 Pareto 前沿選擇最佳公式（平衡複雜度與準確度）
5. 將預測值限制在 [0, 1] 區間作為機率估計
6. 使用訓練集正樣本比例作為分類閾值
7. 在測試集上評估公式的 AUC

公式評估標準：

- **預測效能**：AUC 與 Logistic Regression 相近（差距 < 5%）
- **穩定性**：多個 fold 產出類似的公式結構
- **可解釋性**：公式符合臨床直覺（如：SBP ↑ → 高血壓風險 ↑）

3.6.4 可解釋性分析

本研究使用 SHAP (SHapley Additive exPlanations) 進行模型可解釋性分析：

- **SHAP 值**：量化每個特徵對預測結果的貢獻
- **特徵重要性排序**：識別最具影響力的風險因子
- **交互效應**：分析特徵間的協同或拮抗作用

3.7 實驗環境

本研究之實驗環境與使用套件如表 3-8 所示。

表 3-8 實驗環境與工具

類別	項目	規格/版本
硬體環境	處理器	Intel Core i7-11700 @ 2.50GHz (8 核心)
	記憶體	32 GB DDR4 3200 MHz
	顯示卡	NVIDIA GeForce RTX 3050 (6 GB VRAM)
	儲存裝置	SSD (ADATA SX8200PNP + WDC WDS200T2B0A)
軟體環境	作業系統	Windows 10 專業版
	程式語言	Python 3.10
	開發環境	Jupyter Notebook, VS Code
主要套件	機器學習	scikit-learn, XGBoost
	神經網路	MLPClassifier (scikit-learn)
	符號回歸	PySR（初期曾使用 gplearn）
	可解釋性	SHAP
	資料處理	pandas, numpy
	視覺化	matplotlib, seaborn

-