

謝辭

脫離學業十年有餘，回想當初考上心目中的臺北科技大學時，抱著由你玩四年的心態，並沒有將重心放在課業上。然而，大學時期工程數學課堂上，楊士萱教授曾說過一句話：「不要對自己太好」，這句話一直銘記在心，卻是出了社會、經歷了職場的磨練後，才真正深刻體悟其中的意義。十多年後，帶著截然不同的心態重返學術殿堂，這一次，我格外珍惜每一堂課、每一次學習的機會。

首先，誠摯感謝指導教授許揚老師的悉心指導。當初選擇軟體工程作為研究方向，正是因為這個領域與我十多年的軟體開發職涯高度契合，這是一個不會後悔的選擇。教授深知在職學生的處境，鼓勵我從工作實務中尋找研究題目，讓學術研究與職場經驗得以相互印證，這樣的指導方式讓我受益良多。

感謝公司大安聯合醫事檢驗所提供的在職進修福利，這份制度成為我重返學術領域的重要推力。公司全額補助學費，大幅降低了經濟上的顧慮，使我能夠專注於學業。同時，順利取得碩士學位後的額外加薪制度，更體現了公司對員工自我提升的重視。我始終相信，所學到的知識是別人帶不走的，而公司投資員工成長、員工回饋所學於工作，正是一個雙贏的局面。

最後，我要特別感謝我的妻子。她是一位獨立且有能力的人，在我每天早出晚歸的求學期間，默默承擔了許多家庭的責任與付出。更令人欣慰的是，在我就讀碩士的第一年，她也報名了國外的碩士進修課程，大兒子進入國小一年級，小兒子進入幼稚園——一家四口同時都是學生，整個家庭充滿了學習的氛圍。這份共同成長的經歷，是這段求學旅程中最珍貴的收穫。

謹以此論文，獻給所有支持我的人。

中文摘要

論文題目：縱向健檢資料與變化量特徵之三高疾病風險預測：多模型比較研究

三高疾病（高血壓、高血糖、高血脂）是全球主要的慢性疾病，也是心血管疾病的關鍵可控風險因子。然而，現有風險評估方法多仰賴單一時間點的檢驗數據，未能充分利用縱向健檢資料中蘊含的動態資訊。

本研究使用公開於 Dryad 資料庫的縱向健檢資料集（Luo et al., 2024），涵蓋 6,056 位 40 歲以上社區成人，追蹤期間為 2010 至 2018 年。研究採用三時間點縱貫設計（Y-2、Y-1、Y0），以健檢指標及其變化量（ Δ 特徵）預測三高疾病狀態，透過滑動窗口法產生 13,514 筆建模紀錄。本研究系統性比較八種模型（傳統統計、樹模型、SVM 及神經網路），並以符號回歸探討可解釋性，實驗採用 StratifiedGroupKFold 五折交叉驗證確保無資料洩漏。

主要研究發現：(1) Logistic Regression 表現穩定優異，高血糖預測 AUC 達 0.938；(2) Δ 特徵可帶來 1.5% – 2.3% 的 AUC 提升，且在 Top 10 重要特徵中佔比達 30 – 50%；(3) SHAP 分析揭示疾病特異性預測因子；(4) 符號回歸發現極簡公式即可達到 AUC 0.943；(5) 累積更多健檢紀錄有助於提升預測性能；(6) 僅用 Top 5 特徵，AUC 降幅小於 0.5%。

本研究證實：以縱向健檢資料透過簡單特徵工程與線性模型，即可達到臨床可用的預測性能，適合在基層醫療單位實施早期預警系統。

關鍵詞：三高疾病、高血壓、高血糖、高血脂、機器學習、縱向資料、變化量特徵、SHAP 可解釋性、符號回歸

Abstract

Title: Predicting Hypertension, Hyperglycemia, and Dyslipidemia Using Longitudinal Health Checkup Data with Delta Features: A Multi-Model Comparative Study

Hypertension, hyperglycemia, and dyslipidemia—collectively known as the "three highs"—are major chronic diseases and key modifiable risk factors for cardiovascular disease. Conventional risk assessment methods rely on single-timepoint data, failing to capture the dynamic health trajectories embedded in longitudinal records.

We utilized a publicly available longitudinal dataset (Luo et al., 2024) comprising 6,056 adults aged 40+ followed from 2010 to 2018. A three-timepoint design (Y-2, Y-1, Y0) was adopted, using health indicators and delta features (Δ) to predict disease status. A sliding window approach generated 13,514 records. Eight models (traditional statistics, tree-based, SVM, neural network) were compared, with symbolic regression exploring interpretability. StratifiedGroupKFold 5-fold cross-validation ensured no data leakage.

Key findings: (1) Logistic Regression achieved AUC 0.938 for hyperglycemia; (2) Δ features improved AUC by 1.5%–2.3% and comprised 30–50% of top 10 features; (3) SHAP revealed disease-specific predictors; (4) symbolic regression achieved AUC 0.943 with a minimal formula; (5) more checkup records improved performance; (6) top 5 features showed <0.5% AUC decrease.

This study demonstrates that longitudinal health checkup data with simple feature engineering can achieve clinically useful predictions, suitable for early warning systems in primary healthcare.

Keywords: Hypertension, Hyperglycemia, Dyslipidemia, Machine Learning, Longitudinal Data, Delta Features, SHAP Interpretability, Symbolic Regression