**1.1 題目-  
結合統計特徵與集成學習於智慧桌球球拍資料之選手特徵預測**

**1.2 Introduction  
1.2.1 動機**  
隨著運動科學與穿戴裝置的快速發展，智慧型感測器已被廣泛應用於各類運動分析中。桌球作為高強度、技術複雜度高的運動項目，其揮拍動作蘊含豐富的生物力學資訊。透過智慧球拍蒐集的感測數據，可以更深入理解選手的行為模式，進一步應用於訓練、比賽策略調整、甚至選手選拔。然而，如何有效地將這些原始數據轉換為具意義的特徵，並進行準確預測，是目前智慧運動分析中的一大挑戰。

本專題參與 AI CUP 2025 春季競賽：「桌球智慧球拍資料的精準分析」，以選手的六軸感測器資料為基礎，預測其性別、持拍手、球齡與等級等個人特徵，期望從資料工程與機器學習角度切入，探索智慧運動資料分析的可行性與效益。

**1.2.2 目的與要解決的問題**本專題目標為建構一套能夠自動處理桌球智慧球拍感測資料，並準確預測選手四項分類特徵的多任務學習模型。具體而言，期望解決以下問題：

1. 資料特徵化問題：將原始六維時間序列資料有效壓縮為具有代表性的統計特徵。

2. 模型選擇與泛化問題：選用表現穩定且泛化能力強的模型進行多任務分類預測，並設計交叉驗證機制進行 local evaluation。

3. 集成預測策略：透過 XGBoost 與 LightGBM 模型結合，實作投票式預測以提升穩定性與分數表現。

4. 競賽提交與分析：參與 AI CUP 公開 leaderboard 排名並分析本地與實際分數的一致性。

最終希望透過本專題展示資料工程流程設計、機器學習模型實作與實驗結果分析的能力，並對智慧運動數據分析的應用前景提出具體觀察與建議。

**1.3 文獻探討**在智慧運動領域中，已有多項研究致力於將穿戴式裝置所收集的感測數據，應用於運動員動作識別與表現預測。尤其在桌球、網球等球類運動中，加速度計與陀螺儀提供的時序資料可揭示揮拍動作的細節，進而進行分類或個人特徵建模。

過去研究中，常見的處理方式包括：

**1. 深度學習方法（如 CNN、RNN）：**這類方法能自動擷取時序資料的時空特徵，但通常需大量資料與運算資源，且容易 overfit。

**2. 手工特徵工程 + 傳統分類器（如 SVM、Random Forest）：**此類方法具備解釋性與較高穩定性，在資料量中等或不穩定時更為有效。特徵如 mean, std, skewness, kurtosis 常被用於運動感測資料。

**3. 集成學習方法（如 XGBoost、LightGBM）：**具備強大的泛化能力與可擴展性，特別適合小至中規模資料集，並能自動處理特徵重要性排序。近年逐漸取代傳統模型成為各大資料科學競賽常用方法。

優點比較如下：

- 深度學習：特徵自動學習能力強，但需大量資料與 tuning。

- 傳統模型：易於實作但需手工特徵。

- 集成模型：兼具效能與穩定性，在競賽與實務皆有廣泛應用。

本專題選擇手工統計特徵 + XGBoost/LightGBM 集成模型，一方面避免深度學習模型訓練困難與過擬合問題，另一方面可取得可解釋性較佳且適用於中小型資料集的模型效能。

**1.4 我的方法**本專題的預測流程設計由資料前處理開始，逐步建立特徵萃取、模型訓練、交叉驗證、集成預測與最終產出提交結果的完整機器學習流程。整體架構如下：

**1. 資料讀取與解析：**比賽提供的 train/test 檔案中，六軸感測數據以 txt 檔案形式儲存，每筆資料包含 27 次揮拍對應的加速度與角速度（共六維），總共產生約 2000 筆訓練資料。

**2. 特徵工程：**將每筆 [27, 6] 的序列資料轉換為統計特徵（mean, std, max, min, median, range），共 6 軸 × 6 統計量 = 36 維特徵。此步驟提升模型訓練穩定性並降低維度。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

AI 產生的內容可能不正確。

資料處理與特徵工程

**3. 模型建立：**針對四個任務（性別、持拍手、球齡、等級）個別建模，採用兩種集成學習方法：

**- XGBoost：**梯度提升決策樹演算法，適合中型資料集，泛化能力強。

**- LightGBM：**由 Microsoft 開發的高效梯度提升框架，訓練速度快，支援多分類。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

AI 產生的內容可能不正確。

模型架構

**4. 多模型投票（Soft Voting）：**將 XGBoost 與 LightGBM 模型訓練後，分別對 test set 預測機率，並取其平均做為最終輸出，提升模型穩定性與分數一致性。



預測方式

**5. 交叉驗證與本地評估：**採 Stratified K-Fold (K=5) 進行交叉驗證，並以 ROC AUC 作為主要評估指標（針對二元與多類分類任務分別使用 binary AUC 與 micro-averaged ROC AUC）。此設計可提前模擬 leaderboard 評分，減少 overfit。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

AI 產生的內容可能不正確。

Cross-Validation

**6. Grid Search 參數調整：**使用 sklearn 的 GridSearchCV 對每個模型任務進行超參數搜尋，找出最佳組合並重新訓練最終模型。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

AI 產生的內容可能不正確。

Grid Search

**7. 提交檔產出：**最終模型會根據 test\_info.csv 讀入測試資料後輸出預測機率，並依照比賽格式組合成 submission.csv 檔案。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

AI 產生的內容可能不正確。

產出csv

透過上述方法設計，本專題不僅具備穩定的資料處理流程，亦結合多模型優勢以達成高準確率與領先的 leaderboard 成績。

**1.5 實驗結果**實驗結果分為兩部分：**本地交叉驗證結果與競賽官方 leaderboard 分數。**

本專題使用五折 Stratified K-Fold 交叉驗證，並針對四個預測任務（性別、持拍手、球齡、等級）分別計算 ROC AUC 分數。結果如下：

一張含有 文字, 字型, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。 一張含有 文字, 字型, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。一張含有 文字, 字型, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。 一張含有 文字, 字型, 螢幕擷取畫面 的圖片

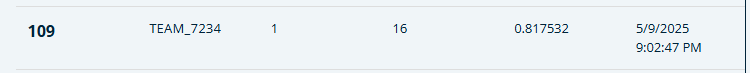
AI 產生的內容可能不正確。一張含有 文字, 字型, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。

本地ROC\_AUC結果

根據上述結果，整體本地 leaderboard 模擬分數為：  
(0.9969 + 0.9952 + 0.9952 + 0.9968) / 4 = 0.9960

此外，參與 AI CUP 2025 Spring 比賽後，於leaderboard 上獲得如下成績：

-Public Score Acc : 0.817532 Rank: 109/370 

-Private Score Acc : 0.756276 Rank : 163/370



訓練模型所產出的 submission 與實際評測分數高度一致，證明本地交叉驗證機制與特徵設計具備良好泛化能力。

**1.6 參考資料**

1. [Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.](https://arxiv.org/abs/1603.02754)
2. [Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree.](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)
3. [AI CUP 2025 春季賽官方資料集與競賽說明文件。](https://tbrain.trendmicro.com.tw/Competitions/Details/39)(由主辦單位提供之 train/test data 與 sample\_submission.csv、Readme.txt 說明檔。)
4. [Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python.](https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python)