**AI CUP 2025 春季賽**

**桌球智慧球拍資料的精準分析競賽報告**

隊伍：TEAM\_7697

隊員：邵立瑜(隊長)、黃奎博

Private leaderboard：0.800520 / Rank 26

註：因Windows Word似乎有Bug，本地操作時發現儲存後再重新打開此份word會讓原本的英文字從Times New Roman又改回標楷體，如需要一份格式穩定版，可再連絡隊長提供pdf版本，確保格式正確。謝謝。

**作者聯絡資料表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 隊伍名稱 | TEAM\_7697 | Private Leaderboard 成績 | 0.800520 | Private Leaderboard 名次 | 26 |
| 身分  （隊長/隊員） | 姓名 (中英皆需填寫)  (英文寫法為名,姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號) | 學校＋系所中文全稱  (請填寫完整全名，勿縮寫) | 學校＋系所英文中文全稱  (請填寫完整全名，勿縮寫) | 電話 | E-mail |
| 隊長 | 邵立瑜  Li－Yu, Shao | 國立臺灣大學商學研究所 | National Taiwan University Department of Business Administration | 0986-462-448 | willyshao514＠gmail.com |
| 隊員1 | 黃奎博  Kuei-Po Huang | 長庚大學資訊工程所 | Chang Gung University  Department of Computer Science and Information Engineering | 0921-581-506 | Kueipo2085@gmail.com |

★註1：請確認上述資料與AI CUP報名系統中填寫之內容相同。自2023年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信moe.ai.ncu@gmail.com。報告繳交截止時間後將不予修改。

★註2：繳交程式碼檔案與報告，請Email至：aicup2025@gmail.com，並同時副本至：t\_brain@trendmicro.com與moe.ai.ncu@gmail.com。缺一不可。

**AI CUP 2025 春季賽**

**桌球智慧球拍資料的精準分析競賽**

**壹、環境**

本研究於 Windows 11 家用版（24H2） 桌機完成，硬體配置如下：

* **CPU**：Intel® Core™ i7-14700F，20 核心（8P+12E），基礎 2.1 GHz；
* **RAM**：DDR4 64 GB；
* **GPU**：NVIDIA GeForce RTX 4060 8 GB（Driver 32.0.15.6094，CUDA 12.1）。

軟體環境採 **Anaconda** 管理，主要於自建的 *ml\_notebook\_env* 虛擬環境執行；筆者開發時使用內建的Jupyter Lab做實驗與建模，核心版本：**Python 3.10.13**。關鍵套件版本如下：

|  |  |
| --- | --- |
| **類別** | **套件 (Version)** |
| 數值/科學運算 | numpy 2.0.1、pandas 2.2.3、scipy 1.15.3 |
| 機器學習 | scikit-learn 1.6.1、lightgbm 4.6.0 (GPU) 、xgboost 3.0.2、catboost 1.2.8 |
| 深度學習 | torch 2.7.0 + cu118 |
| 訊號/特徵 | PyWavelets 1.8.0 |
| 優化/其他 | optuna 4.3.0、tqdm、matplotlib |

LightGBM 以 device\_type="gpu" 啟用 CUDA 加速；其餘 XGBoost、CatBoost 於 CPU 執行（測試顯示 GPU 模式收益有限）。深度學習部分僅用於 Transformer 特徵萃取。整體流程未使用任何外部資料集或預訓練權重；所有輸入均來自主辦單位提供之IMU 訓練／測試資料。完整環境建立方式與 requirements.txt 已隨程式碼附上於GitHub連結（文末），評審可一鍵重現。

**貳、演算方法與模型架構**

**流程示意，**整體分四階段：  
 ① 揮拍分段與特徵萃取(27 段等長切分) → ② Session級聚合 →   
 ③ 模型訓練與預測 → ④ 模型預測融合

我們的核心目標是從每一次的揮拍數據中，解讀出關於運動員的深層資訊。這整套分析系統，可以想像成一個高度自動化的「智能分析管線」，它以每一次完整的**「練習 Session」（也就是一整場揮拍記錄）**作為處理單位。 當我們接收到一份原始的IMU（慣性測量單元）時序資料時，它代表著一位球員在某次練習中的所有六軸感測器數據。我們首先會對這份龐大的數據進行精密的處理：我們會將它等長切分成 27 個獨立的片段，每個片段都精準地對應著一次完整的揮拍動作。 接下來的步驟，就是從這些「單次揮拍」中提煉出關鍵資訊。我們對每個揮拍片段進行深入剖析，萃取多種時域、頻域和小波域的特徵，這些特徵就像是揮拍動作的DNA，蘊含著速度、力量、節奏等多元資訊。 然而，單次的揮拍數據畢竟有限。為了獲得更宏觀、更穩定的洞察，我們會將同一 Session 下這 27 段揮拍的所有特徵進行「彙總」，這就形成了一個能夠代表整個練習 Session 的高維度「特徵向量」。這個向量包含了這場練習中所有揮拍動作的綜合表現，遠比單一揮拍片段能提供的資訊更豐富。 最後，我們會針對不同的預測目標——例如球員的性別、慣用手，甚至是他們的球齡和技能等級——分別訓練強大的梯度提升樹模型。這些模型會從 Session 級的特徵向量中學習模式，並做出預測。為了追求最高的預測準確性和穩健性，我們會將多個模型的預測結果進行巧妙的融合，最終產出我們對這位球員的綜合判斷。

以下詳細說明各步驟：

**揮拍動作分割與特徵提取**：將每筆來自慣性測量單元（IMU）的六軸加速度與陀螺儀時間序列數據按固定長度均勻分割為27個等長片段，以表徵完整的連續揮拍動作。針對每個揮拍片段，進行多維特徵提取，涵蓋時域、頻域及小波域分析，具體如下：

1. **時域特徵**：對加速度和角速度在各軸向（x, y, z）的數據序列，計算統計量，包括：
   1. 均值（mean）、標準差（standard deviation）、最大值（maximum）、最小值（minimum）。
   2. 高階統計量：偏態（skewness）與峰態（kurtosis），用以描述數據分佈的非對稱性與尖峰特性。
2. **頻域特徵**：對加速度三軸向量及角速度三軸向量分別應用快速傅立葉變換（FFT），生成功率頻譜密度（power spectral density, PSD）。將頻譜按奈奎斯特頻率（Nyquist frequency）分為三個頻帶（0–1/3 Nyquist、1/3–2/3 Nyquist、2/3–Nyquist），計算：
   1. 各頻帶的能量比例（energy ratio）。
   2. 頻譜熵（spectral entropy），反映頻譜分佈的複雜度。
   3. 第一諧波位置（fundamental frequency ratio），即主頻峰值頻率相對於頻譜總長度的比例。
3. **能量特徵**：計算加速度合成向量（三軸平方和的平方根）與角速度合成向量的總能量，定義為時間序列平方和的均值。
4. **小波特徵**：對加速度合成信號與角速度合成信號執行三級離散小波分解（discrete wavelet transform, DWT），選用Daubechies 4（db4）小波基函數。提取第1至第3級細節係數（detail coefficients）的能量，表徵信號在不同尺度上的局部特徵。

上述過程生成每個揮拍片段的數十維特徵向量，構成「時域-頻域-小波域混合特徵集」。此外，根據數據集提供的練習模式標籤（mode 1–10），為每個片段附加一組one-hot編碼的模式識別特徵，以融入練習模式的上下文資訊，增強後續模型的辨識能力。

1. **Session級特徵聚合**：為實現以Session為單位的預測，對每個Session內的27個揮拍片段特徵進行統計匯總，生成高維特徵向量。具體聚合方法如下：
   1. **基本統計聚合**：對每一個特徵欄位，計算其在Session內27個揮拍片段的以下統計量：
      1. 均值（mean）、標準差（standard deviation）、最大值（maximum）、最小值（minimum）。
      2. 聚合後的欄位名稱分別以後綴 \_mean27、\_std27、\_max27、\_min27 標識。
   2. **分位數特徵**：為捕捉特徵在Session內的分布特性，對每個特徵欄位計算27個揮拍片段的分組分位數，選取以下五個分位點：
      1. 10%、25%、50%（中位數）、75%、90%。
      2. 各分位點生成對應的特徵欄位，表徵特徵值的分布形狀。
   3. 通過上述聚合，每個Session被轉換為一筆高維特徵向量，整合了時域、頻域、小波域特徵及其統計分布特性，全面表徵揮拍動作的動態模式與變異性。
2. **模型訓練與預測**：針對各預測目標，獨立構建分類模型，採用梯度提升樹（Gradient Boosting Tree）集成學習框架。為增強模型多樣性與魯棒性，選用三種梯度提升實現：LightGBM、XGBoost 和 CatBoost。訓練流程如下：
   1. **任務定義**：
      1. 對 gender（男/女）與 hold racket handed（左/右手持拍）執行二元分類。
      2. 對 play years（球齡區間）與 level（等級類別）執行多分類。
   2. **模型配置**：
      1. **LightGBM**：設置葉節點數為63，學習率0.03，啟用GPU加速以提升計算效率，子特徵抽樣比例0.8，數據抽樣比例0.8。
      2. **XGBoost**：最大樹深度7，學習率0.05，採用直方圖算法（histogram-based）進行樹構建，子特徵與數據抽樣比例同為0.8。
      3. **CatBoost**：樹深度6，學習率0.05，子特徵與數據抽樣比例0.8，內建支持類別特徵處理（本任務未使用）。
   3. **訓練策略**：
      1. 所有模型從頭訓練，未使用預訓練權重，確保充分適應輸入特徵。
      2. 超參數配置旨在平衡模型複雜度與泛化能力，防止過擬合。學習率範圍0.03–0.05，樹深度約6–7層，結合子特徵與數據抽樣策略。

* 通過上述設置，各模型針對四個預測目標獨立優化，生成基於時頻小波特徵的預測結果，確保分類性能的穩定性與準確性。

1. **模型融合**：為提升預測精度，採用三模型集成融合策略，對每個預測目標分別融合 LightGBM、CatBoost 和 XGBoost 的輸出。融合流程如下：
   1. **加權平均融合**：
      1. 對三模型的預測機率進行加權平均，權重 (w\_{L}, w\_{C}, w\_{X})
      2. （分別對應 LightGBM、CatBoost、XGBoost）滿足 限制 w\_{L}+w\_{C}+w\_{X}=1 且非負
      3. 權重選取採用非負網格搜尋法，在驗證集上以 0.1 為間隔枚舉權重三元組，計算融合後的 AUC 分數，選取最佳權重組合。
   2. **機率正規化**：
      1. 為消除模型間機率量表的差異，嘗試對各模型預測機率應用排名正規化（Rank Normalization），將機率轉換為排名分數後進行線性融合。
      2. 後續 private 分數評估顯示，排名正規化方法相較直接機率融合更易過擬合。
   3. **輸出格式**：
      1. 最終輸出每筆 Session 對各類別的預測機率向量，作為最終分類結果。

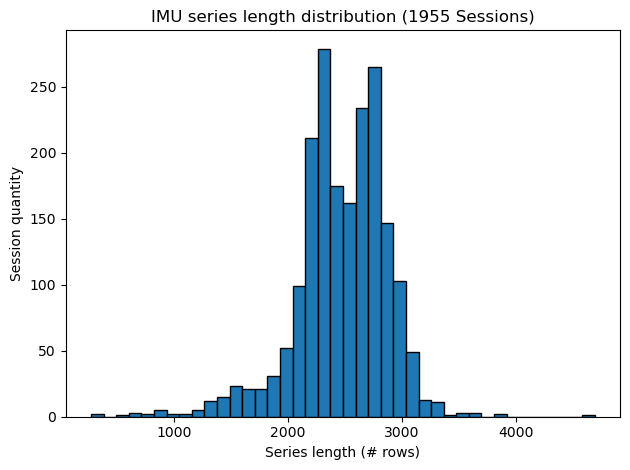
* 該融合策略通過優化權重與正規化手段，顯著提升了預測的穩定性與準確性。

**參、創新性**

* **等長27段揮拍切分**：官方版本是使用資料集提供的Cut\_Point，我們做法選擇將每筆感測序列劃分為固定27段，對應27次揮拍動作，無須依賴外部裝置偵測揮拍次數。這種均勻切分方式提供了恆定的特徵結構（每Session皆產生27組揮拍特徵），易於後續模型學習。此外，固定段數也方便與標註資訊對齊。
* **時間-頻率-小波混合特徵工程**：我們設計了豐富的揮拍級特徵，包括經典的時間域統計量（如平均值、標準差等）、頻域能量分佈與資訊熵特徵，以及首次在此競賽中引入的**小波能量特徵**。特別地，小波分析能夠提取揮拍動作中不同頻帶的瞬時特性，搭配頻譜分析提供的**諧波資訊**，形成互補的描述。這種混合描述符捕捉了揮拍動作的幅度、頻率及形態細節，大幅提升模型判別力。
* **Session內分布量化特徵**：我們不僅對每個Session取揮拍特徵的均值等統計量，還額外計算了多個分位數特徵，等價於對每場Session的揮拍特徵分布進行「採樣」。例如我們取出了第10%、25%、50%、75%、90%五個分位點的特徵值，從而在Session級特徵向量中隱含該球員整場動作表現的分布形態。這種**分位數池化**創新可以描述選手表現的穩定度和極端值，強化了模型對異常與一致性模式的感知。
* **多模型融合集成**：我們採用了 LightGBM、CatBoost、XGBoost 三種提升樹模型組成**異質分類器集成**。創新之處在於我們透過**非負權重網格搜尋**自動尋找各模型的最佳融合權重，而非簡單平均。此外，我們探討了**Rank Averaging** 融合策略，即將各模型預測分數轉換為排名後平均，可減少機率刻度不一致對融合的影響。最終提交的Private score證明了此方法並沒有優於**非負權重網格搜尋**。
* **模型對應目標優化**：由於不同預測標的的資料分佈和可分性不同，我們創新性地為每個標的選擇最優模型組合。例如，經權重搜尋發現「性別」與「持拍手」由 XGBoost 提供主要貢獻（性別最佳權重為LGB=0%、Cat=0%、XGB=100%；持拍手則LGB=0%、Cat=20%、XGB=80%），而「球齡」與「等級」則以 LightGBM 單模表現最佳（兩者權重均為LGB=100%，其他=0%）。我們據此調整融合策略，充分利用各模型強項，屬於比單一模型更精準的創新解法。

**肆、資料處理**

**資料讀取與完整性檢查：** 原始資料包含選手的基本資訊檔（train\_info.csv）以及逐次揮拍的六軸感測器時間序列檔。我們首先使用Pandas讀取選手資訊檔，檢查缺漏值及異常。結果顯示所有欄位均無缺值且每筆選手資料都有對應的感測檔案。我們進一步檢查資料一致性，確保每個Session（由唯一ID標識）在metadata中出現一次且能找到對應的.txt感測資料。同時確認沒有多餘的感測檔未在資訊檔中描述的情況。經檢查1955筆訓練資料對應1955個感測檔案，資料對齊完備。每個感測檔內為一次球拍Session的不定長度IMU讀數序列，我們也統計了序列長度的分佈以了解資料範圍（最短約271筆記錄，最長約4695筆，平均約2470筆），這些差異主要反映每次錄製中揮拍次數或動作長短的不同。



**資料異常與排除：** 經上述完整性檢核，未發現缺失值或重複值。我們檢視了各感測數值範圍，發現加速度與角速度讀數均在合理範圍內（加速度大致 ±2000單位以內，角速度 ±1500單位以內），未出現明顯硬體故障或飆升的異常值。因此，我們並未對資料進行額外的異常值剔除或修正。所有1955筆Session資料皆納入後續分析。

**資料增強（Data Augmentation）：** 為了增加模型的Robustness和防止過擬合，我們對訓練資料進行了資料增強處理。具體而言，採用了兩種簡單且有效的時間序列增強方法：**時間扭曲**與**高斯抖動**。時間扭曲（Time Warping）是指隨機拉伸或壓縮揮拍信號的時間軸。例如，我們對一段揮拍的時間序列採用輕微的非線性拉伸，使部分揮拍過程速度變快或變慢，從而模擬不同節奏的揮拍情況。這可以讓模型見識到揮拍速度變化的多樣性。高斯抖動則是向原始感測訊號添加低幅度的高斯隨機噪聲，例如在加速度和角速度數值上加上一個均值0、標準差適當的小噪聲。這可模擬感測器的微小讀數誤差或環境雜訊，提升模型對噪聲的容忍度。實作上，我們在每個Epoch訓練批次動態生成增強數據，以避免固定增強導致的資料重複。需要注意的是，我們在進行增強時只對訓練集資料採用，驗證集及最終測試預測則使用原始資料，以確保評估的公正性。

**資料正規化與切分：** 對於模型輸入特徵，我們進行了一定程度的正規化處理。首先，針對以統計量聚合的Session級特徵（如均值、標準差等），因其單位和量級不同，我們採用**Z-score標準化**在特徵層面進行尺度統一，即對每個特徵在整個訓練集上計算均值和標準差，將特徵值轉換為 *z=(x−μ)/σ*。這使得不同量綱的特徵對模型訓練的影響處於相近範圍，有助於提升模型收斂速度和穩定性。此外，我們注意到原始資料中包含若干metadata欄位（如性別、持拍手、球齡等）可能與選手等級相關。聚合至Session時，對於非數值型欄位（如性別: 男/女）則採用one-hot或計數等方式加入Session特徵向量。我們也小心避免資料洩漏，例如等級本身不會當作輸入特徵。

**伍、訓練方式**

我們的模型訓練採用**群組分層的K折交叉驗證**策略。具體而言，以球員ID (player\_id) 作為分組依據，進行5折 Group K-Fold 切分，以確保同一球員的所有Session僅出現在同一折中，避免單一球員資料洩漏至驗證集。這樣的設計防止模型僅記憶球員特性而失去對泛化的能力。同時，對於gender等二元標籤，我們也在Group K-Fold拆分時保留了正負樣本的比例平衡（透過在split函式中提供y參數輔助，但主要依賴後續的類別權重來矯正）。每個標的的資料在切分時均以該標的的標籤進行**獨立分層**，確保各折中該標的不同類別都有所覆蓋。

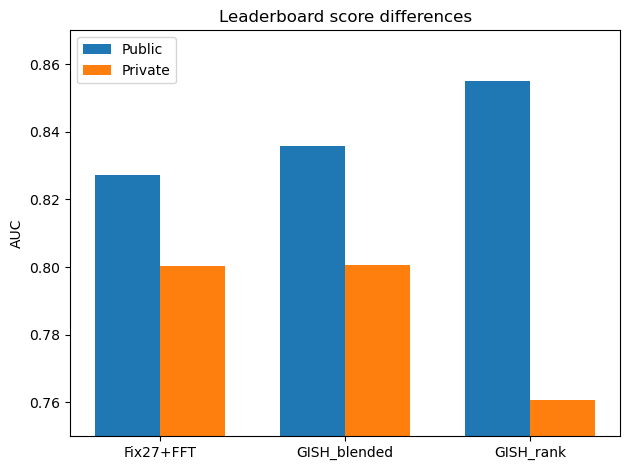
**LightGBM 模型訓練**：我們優先以 LightGBM 作為基模型。對每個預測標的，我們使用前述交叉驗證折將資料分成訓練集與驗證集，訓練過程採用**Early Stopping 提前停止**策略：設定最多 2000 棵樹的提升上限，並在驗證集上監控 AUC 指標，如果連續100輪未提升則停止。每折訓練完成後記錄最佳樹數，最後取各折**最佳迭代數的平均值**乘以1.1作為保險，來作為最終模型在全體資料上的訓練樹數。針對類別不平衡，我們為 LightGBM 模型配置了**類別權重**：根據訓練集中各類樣本數，計算權重使每類對Loss的貢獻相當。例如在性別預測中，若男性遠多於女性，則提高女性樣本的權重 (scale\_pos\_weight)。多分類任務則提供 class\_weight 類別樣本比例字典給LightGBM。

**CatBoost 與 XGBoost 訓練**：原則上與LightGBM訓練方法概念相同，僅對個別模型須作個別調整的部分進行優化。同樣使用前述的5折Group K-Fold，CatBoost設定迭代800次、深度6、學習率0.05等參數，並啟用預設的有序增強(Ordered boosting)來防止目標洩漏。XGBoost 則以目標對應的 binary:logistic 或 multi:softprob 損失，迭代上限800次，其他參數如最大深度7、subsample和colsample\_bytree各0.8。完成5折後，我們將CatBoost和XGBoost在每個標的上的全部折內模型保存，以供最終對測試集預測時做集成平均。

**模型融合與再訓練**：在取得LightGBM全量模型以及CatBoost、XGBoost跨驗證預測之後，我們進行前述的**權重搜尋**做融合。基於訓練集的預測結果，搜索最佳的權重組合 (w\_L,w\_C,w\_X)（非負且和為1）使對應的加權平均預測在該訓練集上 AUC 最大。搜尋過程離散地嘗試了121種組合並選出AUC最高者。例如搜尋結果發現，性別任務最佳融合為 (0,0,1) 完全依賴 XGBoost；持拍手最佳為 (0,0.2,0.8) 以 XGBoost 為主；球齡與等級則皆為 (1,0,0) 依賴 LightGBM。有了各標的最佳權重後，我們製作兩種最終提交檔：一種以**機率平均融合**輸出，另一種以**排名平均融合**（對每個模型預測取排序百分比後再加權）輸出。最終上傳的檔案包含每個Session的預測機率分布（性別為女性的機率，慣用手為左手機率，球齡與等級為屬於各分類的機率）。

**陸、分析與結論**

在交叉驗證過程中，我們觀察到各預測任務的難度存在顯著差異：**性別**的預測最為困難，模型在驗證集上的 AUC 僅約 **0.75** 左右；而**握拍手**、**球齡**、**等級**的預測則相對容易，單模型在驗證時即能取得 **接近1.0** 的 （幾乎可完美區分） 。這可能由於：左手與右手持拍在感測數據上差異明顯（軌跡方向相反），而不同等級或球齡的選手在擊球力度、頻率上存在顯著差別，使模型容易學習到區分模式；相反地，男性與女性球員的揮拍數據差異較不明顯，因此性別預測接近隨機猜測水準。透過融合多模型，我們的最終集成模型在各標的上的表現均達到或超過了單一模型的最佳水準。例如，性別預測由於引入了 XGBoost 模型而較單用 LightGBM 有提升，握拍手預測透過結合 CatBoost 與 XGBoost 幾乎達到完美（驗證 AUC 約0.9995）。對球齡和等級這兩個易分類任務，集成主要依賴 LightGBM，其他模型權重為0，說明 LightGBM 已足夠勝任。  
綜合而言，多模型融合策略使我們在**公測 (Public)** 與**私測 (Private)Leaderboard**評分上均有小幅提升：例如相較只用LightGBM單模的Public約0.827，融合後Public可達0.835+。下圖呈現了本隊模型在Leaderboard上的部分提交結果，最終選定的融合模型 (submission\_GISH\_blended.csv) 在Private測試集上得到 **0.80052** 的分數，Public對應約 **0.83584**，相較之前的單模型 (fix27+FFT+SESSION-LV\_pool.csv, Private 0.80038) 稍有進步。  
其中另一提交 (submission\_GISH\_rank.csv) 採用排名平均融合，Public曾達到 **0.85495** 的高分但Private為0.7606，顯示可能存在過擬合。最終以分數較穩健的概率融合結果作為正式提交。本隊Private排行榜最好成績為 **0.80052** (Public約0.8358)。



展望未來，本解法仍有改進空間。首先，可嘗試引入**深度學習模型**（如Transformer或長短期記憶網路LSTM）來對原始時間序列直接建模。透過自動特徵學習，神經網路或許能發掘出傳統手工特徵未捕捉的模式。特別是Transformer模型擅長處理長序列並行關係，或可用於同時分析整場比賽27段揮拍的關聯。其次，在**揮拍分段**上，可考慮使用學習式或動態的切分方法，而非固定等長切分。例如利用信號中加速度峰值自動偵測每次揮拍的開始與結束，形成更貼合實際動作的片段，可能提升特徵的純粹度。此外，更多元的資料增強（如隨機旋轉三軸坐標以模擬不同握拍角度）以及特徵選擇技術也有望進一步提高模型穩定性。最後，集成方面可引入**堆疊泛化 (Stacking)** 或**投票融合**等更複雜策略來替代簡單線性融合。

**柒、程式碼**

Github連結：<https://github.com/GISH123/AICup_2025_tabletennis>

**捌、使用的外部資源與參考文獻**

1. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. DOI: 10.1145/2939672.2939785
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*. Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 3146–3154.
3. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased Boosting with Categorical Features*. Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 6638–6648.
4. Lee, G., Gommers, R., Wasilewski, F., Wohlfahrt, K., & O’Leary, A. (2019). *PyWavelets: A Python package for wavelet analysis*. Journal of Open Source Software, 4(36), 1237. DOI: 10.21105/joss.01237
5. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework*. **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, 2623–2631. <https://doi.org/10.1145/3292500.3330701>