封面

**目錄**

參考文獻………………………………………………………………………………………………………… 3

第一篇文章……………………………………………………………………………………………………… 4

第二篇文章……………………………………………………………………………………………………… 9

第三篇文章………………………………………………………………………………………………………. 14

**參考文獻**

1. Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm

by Ningyuan Zhu,Chaoyang Zhu,Liang Zhou,Yayun Zhu and Xiaojuan Zhang

Appl. Sci. 2022, 12(20), 10456; <https://doi.org/10.3390/app122010456>

2. Hyperparameter Optimization of Ensemble Models for Spam Email Detection by Temidayo Oluwatosin Omotehinwa and David Opeoluwa Oyewola Appl. Sci. 2023, 13(3), 1971; <https://doi.org/10.3390/app13031971>

3. Traffic Accident Severity Prediction Based on Random Forest

by Miaomiao Yan and Yindong Shen Sustainability 2022, 14(3), 1729; <https://doi.org/10.3390/su14031729>

**Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm**

**一、研究背景:**

隨著工業控制系統對於網路安全的需求增加，入侵檢測系統（IDS）變得越來越重要。該研究針對電力工業控制系統的入侵檢測，提出了一種改進的隨機森林（Random Forest）模型，並對其超參數進行優化。

**二、相關研究:**

現有的隨機森林（RF）技術在傳統網絡和工業控制系統中的應用，在傳統網絡入侵檢測中，RF方法已較為成熟，具有較高的準確性，但在工業控制系統，尤其是電力工業控制系統中的應用仍然有限。現有的研究多數評估指標不完整，或使用非公開數據集，難以比較不同方法的性能。過去Morris T. 等人針對電力系統入侵檢測使用RF方法二分類的準確率僅為80%，對於電力系統這樣高穩定性要求的系統來說是不夠的。本篇文章研究使用與Morris T.等人研究相同的數據集和RF方法，並且比較兩者的入侵檢測性能

**三、實驗方法：**

該方法是一種基於隨機森林（RF）分類器和超參數優化的入侵檢測方法，具體流程如下：

**1.數據預處理與交叉驗證:** 首先導入電力工業控制系統的原始數據集，進行預處理，並使用5折交叉驗證劃分訓練集與測試集。

**2.超參數配置與RF模型構建**：構建一個超參數配置空間，並基於不同的超參數組合訓練RF分類器，通過測試集驗證其性能。

**3.超參數重要性評估**：使用功能變異分析（Functional ANOVA）來計算每個超參數的重要性，將RF分類器的性能（例如ROC-AUC得分）作為指標，從而為後續的優化提供依據。

**4.改進的網格搜索優化**：根據超參數的重要性，逐一對超參數進行網格搜索優化，優化後的參數值替換基分類器中的對應參數，最終得到最佳的超參數組合。

**5.最佳模型構建與驗證**：基於優化後的超參數，構建最佳的RF分類器，並使用測試集進行驗證，以確保其性能優越。

一張含有 文字, 螢幕擷取畫面, 圖表, 字型 的圖片

自動產生的描述

圖(一):實驗方法流程圖

**四、實驗結果與分析:**

使用功能ANOVA分析超參數對模型方差的貢獻衡量超參數的重要性，如下圖(二)所示。優化順序：max\_depth、min\_samples\_leaf、min\_samples\_split、criterion、n\_estimators、bootstrap 、 max\_features。

**一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述**

圖(二):超參數特徵重要性占比

逐一搜索優化超參數後得到的初步結果如圖(三)所示

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

圖(三):超參數優化結果

調整後的超參數在測試集上準確性roc\_auc得分為0.9935如下圖(四)所示

一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

自動產生的描述

圖(四):隨機森林分類器ROC曲線圖

準確性比較分析:本研究所提出的RF分類模型在準確率、精確度、召回率及F1-score方面均優於SVM、NB、DT。 RF 分類器的精確度分別比其他三個分類器高 38%、29% 和 3%。 RF分類器的召回率比其他三個分類器高21%、36%和4%。其他三類分類器的 F1 分數分別比 RF 分類器低 31%、34% 和 3%，如下圖(五)所示。

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

圖(五):與先前研究的準確性對比圖

超參數搜尋方法效率分析:比較了網格搜尋、遺傳演算法搜尋和改進網格搜尋法(IGSA)的計算時間和roc\_auc分數。 IGSA的roc\_auc得分接近全域最優且高於遺傳演算法搜尋。更重要的是，超參數優化速度比網格搜尋提高了165倍，比遺傳演算法搜尋快了5倍，證明了此改進方法的有效性，如下圖(六)所示

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

圖(六):改進網格搜尋方法與其他方法對比圖

**五、結論:**

本研究針對電力工業控制系統入侵偵測能力不足的問題，提出了一種基於隨機森林 (RF) 模型的入侵偵測方法，針對RF模型的七個超參數，依照其重要性順序逐一調整，實現分類器的最佳效能，並且通過改進網格搜尋法進行超參數最佳化，提升了RF分類器的性能與效率，比傳統網格搜尋速度快165倍，同時能穩定準確率，且測試結果相較先前研究使用的傳統機器學習演算法，準確性、精確度、召回率、F1-score，以及ROC-AUC指標有明顯提升，且略優於其他數模型演算法。

**Hyperparameter Optimization of Ensemble Models for Spam Email Detection**

**一、研究背景:**

垃圾郵件的發送目前仍為全球網路安全最大威脅之一， 2020年10月至2021年9月期間，全球共發送了3,364.1億封電子郵件，其中約84%（超過一半）是垃圾郵件，電子郵件詐欺勒索造成相當大的經濟損失，因此有許多研究開始使用機器學習的方式來偵測並且阻擋垃圾郵件發送，然而目前技術多依賴於單一模型分類，模型容易出現判斷錯誤或是過度擬合訓練資料，而集成式學習的方式整合了多個模型的預測，可以提高模型的偵測準確性合泛化能力，但尚未被廣泛應用在偵測垃圾郵件應用上，本研究採用基於隨機森林和極限梯度增強（XGBoost）的集成模型，並使用網格搜尋交叉驗證技術進行超參數優化。研究結果顯示，調整後的模型在準確度和靈敏度上均有顯著提升，其中XGBoost模型表現更佳，能有效提升偵測垃圾郵件的準確程度。

**二、相關研究:**

有研究使用基於決策樹C4.5演算法的技術，專注於電子郵件標頭而非內容，提升對垃圾郵件結構變化的適應能力，準確率達96%​。另一項研究使用貝葉斯過濾與眾包機制區分了完全垃圾郵件與半垃圾郵件，實現95.1%的準確率​。也有研究用語義方法與特徵選擇WordNet將電子郵件分類，並透過PCA和CFS技術降低特徵維度，準確率超過90%​。也有研究使用元啟發式演算法概念將人工蜂群演算法與邏輯回歸結合，在多個資料集上測試，取得最高分類精度98.91%。也有研究使用深度學習針對影像垃圾郵件檢測，開發CNN模型，F1分數達88%。先前研究展示了在標題、語意、影像方面有不同的研究偵測方向。

**三、實驗方法：**

**1.資料預處理:**

使用Enron1資料集32,860封電子郵件其中49%為正常郵件，51%為垃圾郵件，並進行雜訊處理去除非ASCII字符、HTML標籤、空格、URL、標點符號等，並刪除停用詞，以提高模型性能，資料分割訓練集佔70%，測試集佔30%

**2.選用的分類器:**

隨機森林（Random Forest）：用bagging方法，通過建立多棵決策樹來進行預測，並透過voting方式能有效減少過度擬合問題。極限梯度提升（XGBoost）：用boosting方法，後續模型的目標是優化先前模型的損失函數或殘差，以提高可預測性。

**3.模型參數調整:**

通過網格搜尋方法(Grid Search)這種技術來遍歷尋找最佳的超參數設定，以提高模型性能，並且透過10折交叉驗證方法多次訓練驗證降低模型過擬合及偏差，以提高模型的泛化能力。

一張含有 文字, 圖表, 方案, 工程製圖 的圖片

自動產生的描述

圖(七):實驗方法流程圖

**四、實驗結果與分析:**

透過網格搜尋調整隨機森林基本模型參數ntree是決策樹數量，mtry 是決策樹節點處分裂的特徵數量參數，經過兩輪優化在ntree = 1500和mtry = 7時，獲得了最高的訓練精度0.9750725。如下圖(八)所示。

一張含有 文字, 圖表, 繪圖, 行 的圖片

自動產生的描述

圖(八):隨機森林參數優化圖

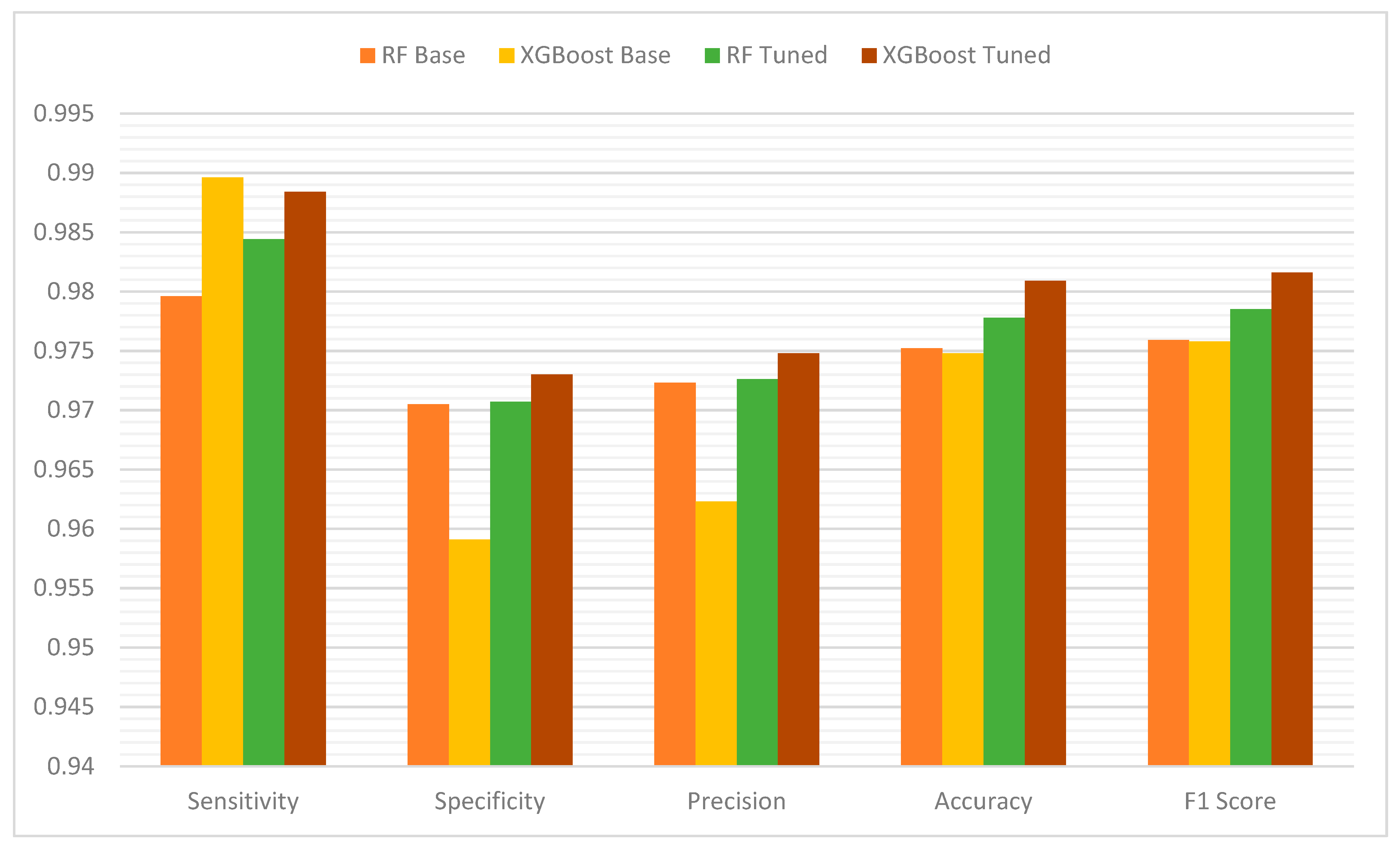
XGBoost模型透過網格搜尋法Boosting 迭代次數為 560 次後，第三面網格的準確度達到峰值 0.9781017，其中最大深度為 7，eta 為 0.25。如下圖(九)所示。

一張含有 文字, 螢幕擷取畫面, 繪圖, 圖表 的圖片

自動產生的描述

圖(九) XGBoost模型參數優化圖

透過最終測試集計算分類模型評判指標分別對調整超參數的模型，以及預設值給出如下圖(十)模型對比，展示出調參後XGBoost模型準確和穩定性。



圖(十):隨機森林、XGBoost 預設調整參數模型的效能評估指標比較圖

**五、結論:**

XGBoost調整參數後的模型敏感度為最高，在實驗結果中展現出最高垃圾郵件偵測率，在特異性、精密度、準確性、F1 score效能評判指標優於其他模型，且使用10折驗證方法進行檢驗，展示出調整參數後的XGBoost模型較隨機森林模型在垃圾郵件偵測分類的有效性以及高泛化能力，能夠有效改善收到垃圾郵件造成的網路安全隱患。

**Traffic Accident Severity Prediction Based on Random Forest**

**一、研究背景:**

由於交通事故的發生受到多種因素的影響，導致數據具有高維度和非線性特徵，開發準確的預測模型具有挑戰性。當前模型主要分為統計模型和AI模型，統計模型對於解釋變數有嚴格假設限制，而AI模型靈活性高且準確度高，如神經網路多層感知機可解釋高度非線性資料，但其結果往往缺乏可解釋性。而隨機森林演算法（RF），能有效處理非線性和高維度變量，並提供變量的重要性分析。本文採用混合模型BO-RF，結合RF與貝葉斯優化（BO），提升城市道路交通事故的預測準確性，並識別影響事故嚴重性的關鍵因素，能有效評估改善預防措施。

**二、實驗方法:**

**1.資料預處理:**

研究使用的數據集為US-Accidents，預處理移除資料缺失過多變數、缺失值填補和編碼，最終採用30426條事故記錄，每條記錄包含15個特徵變數和1個目標變數。

**2.分割資料集:**

將預處理好的資料分別拆為80%的訓練集用於模型訓練，20%測試集用於驗證。

**3.初步RF模型預測:**

使用基本的RF模型進行交通事故嚴重程度預測。

**4.用BO調整RF超參數:**

透過交叉驗證F1結果值，對n\_estimators、max\_depth 、max\_feature超參數進行優化

**5.輸入最佳超參數及模型解釋:**

算出調整最佳參數後的feature importance找出影響大因子提供預防措施參考。

一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

自動產生的描述

圖(十二)實驗方法流程圖

**三、實驗結果:**

整體架構使用Python完成，模型訓練採用Scikit-learn 函式庫和 Hyperopt 套件進行超參數優化，並且使用10折交叉驗證法F1 score作為BO評估的目標函數。下圖(十三)呈現了RF超參數的搜尋空間範圍取值，以及最佳化參數。

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

圖(十三)超參數搜尋空間及最佳化值圖

和傳統的機器學習模型對比，BO-RF模型在Precision、Recall、F1 Score、AUC等四種評估指標皆有所提升，但精確度略低於RF，如下圖(十四)所示。

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

圖(十四)模型評估指標對比圖

調完參數的模型中能看到各項特徵的重要性程度，交通、時間、天氣和POI特徵對事故嚴重程度預測的貢獻率分別為68.89%、13.78%、13.69%和3.64%，交通事故的特徵是影響事故嚴重程度最重要的因素，而POI是最不重要的影響因素，如下圖(十五)所示。

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

圖(十五)特徵重要性程度百分比圖

**五、結論:**

這篇研究提出BO-RF模型用於預測交通事故的嚴重程度，並與ANN、KNN、SVM和RF進行比較。結果顯示，RF在準確率、召回率、F1 分數和 AUC 上均優於其他傳統機器學習模型，且 BO-RF在召回率、F1 分數和 AUC 方面進一步超越基本的RF模型，顯示BO優化超參數能顯著提升RF的性能，透過不同資料集劃分測試，以及10折交叉驗證法BO-RF在所有評分標準上仍保持優勢，顯示其泛化能力高，透過特徵重要性列出如 Start\_lat 和 Start\_lng對交通事故嚴重程度有顯著影響，展現出了模型對特徵的可解釋性，也對於未來預防交通事故的原因提供改進的參考方向。