組別編號Team：#11

資料科學與機器學習成果報告書
Data Science and Machine Learning Report

主題名稱：**基於機器學習與深度學習之台北市PM2.5短期預測模型比較研究**

Topic：The Comparative Study of Short-Term PM2.5 Prediction Models for Taipei City Based on Machine Learning and Deep Learning.

課程主授教師 Supervisors：劉譯閎教授

小組成員**Team Members**：資管三B 高碩辰、資管三B 章祖綸、
資管三B 陳翡甄、資管三B 吳承瑀

日期**Date**：2025.12.31

一、摘要（Abstract）

本研究主要在建立針對台北市細懸浮微粒 PM2.5 的短期預測模型，採用 2018 年至 2024 年之環境監測數據，比較隨機森林（Random Forest）、XGBoost、RNN 與 LSTM 四種演算法之效能。研究針對時間與風向向量化進行的特徵工程。結果顯示，前一小時的 PM2.5 濃度與當下數值呈現**高度相關（r=0.88）**，是預測準確度的關鍵。在模型表現上，LSTM 之解釋變異能力最佳，適合長期趨勢分析；而隨機森林在捕捉突發性高汙染事件上表現優異。本研究亦發現傍晚至夜間時段受大氣擴散條件與交通排放影響，預測誤差顯著增加。

二、問題陳述（Problem Statement）

本研究的核心問題在於提供準確且可驗證的 PM2.5 預測，作為行動與健康管理的依據。此預測具有重要性，因其能提前發出警示，協助民眾減少外出或使用防護措施，降低汙染風險。對於高風險族群，汙染可加劇呼吸與心血管疾病；對學校與運動場館而言，難以及時調整戶外活動；對醫療機構，則可能增加急診負擔並加重慢性病患者的風險。

三、資料來源與預處理（Data Sources and Preprocessing）

本研究之資料集涵蓋 2018 年至 2024 年的小時級觀測數據。主要資料來源包括：

(一)空氣品質資料：取自環境部空氣品質監測網與台北市環境品質資訊網，主要蒐集台北市各行政區之 PM2.5 數值 。

(二)氣象觀測資料：取自 CODIS 氣候觀測資料庫，包含氣溫（AMB_TEMP）、相對溼度（RH）、風速（WIND_SPEED）、風向（WIND_DIREC）、降雨量（RAINFALL）。

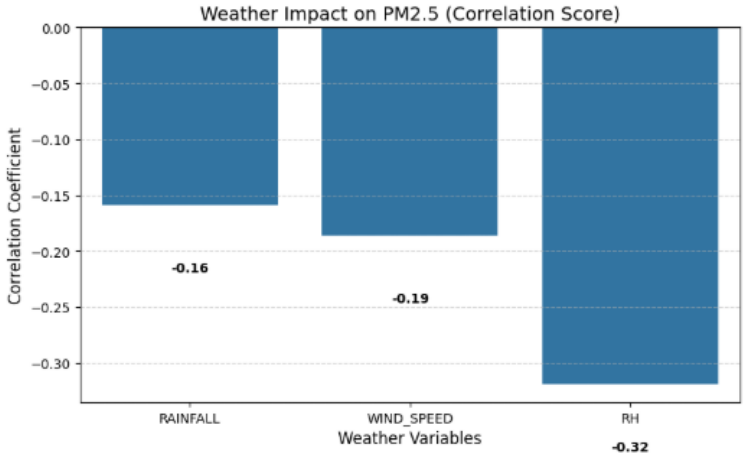
為提升模型訓練效能與收斂速度，本研究執行了以下資料預處理與特徵工程步驟：

- **資料清洗與補值**：首先統一氣象參數格式，剔除異常值（EX:負值）。針對感測器故障造成的數據中斷，採用線性補值法（Linear Interpolation）填補缺失值，確保時序資料的連續性 。
- **特徵標準化**：對所有數值型特徵（氣象變數、PM2.5 過去時間）進行 Z-score 標準化，消除不同物理量綱（EX:溫度與風速）之間的級距差異，避免模型權重受數值大小影響。
- **風向向量化**：考量風向角度（0~360度）在數值上存在不連續性（EX:0度與 360度代表相同物理意義），本研究將風向轉換為正弦（Wind_Sin）與餘弦（Wind_Cos）兩個分量，以解決數值斷層問題 。
- **時空特徵編碼**：
    - 時間特徵：鑑於空氣汙染具有高度持續性，建立了PM25_Lag_1h（前一小時）、PM25_Lag_2h 與 PM25_Lag_24h 等變數，使模型能參考過去狀態預測未來 。

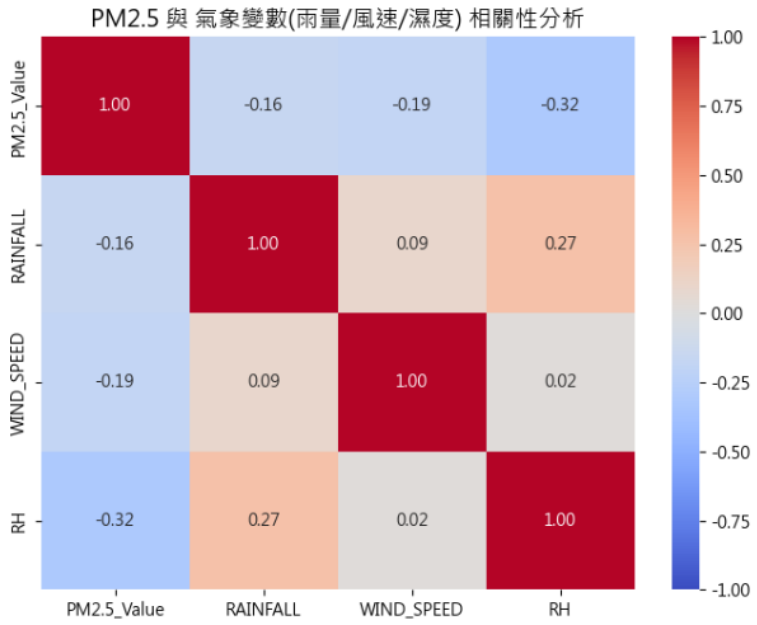
- 測站與日期：針對測站（Station）與星期（Day of Week）進行編碼，以區分不同地理位置的環境特性及平假日的人為活動差異 。

四、特徵相關性與影響因子分析（Feature Correlation and Impact Analysis）

本研究採用皮爾森相關係數進行量化分析。分析重點分為「氣象因子影響」與「歷史滯後效應」兩部分進行探討。首先檢視氣象條件與空氣品質之關聯。如**圖一**與**圖二**所示，主要氣象變數（降雨量、風速、相對濕度）皆與 PM2.5 濃度呈現負相關(Negative Correlation)，顯示這些天氣現象有助於降低懸浮微粒濃度。



圖一、氣象變數與 PM2.5 之相關係數長條圖

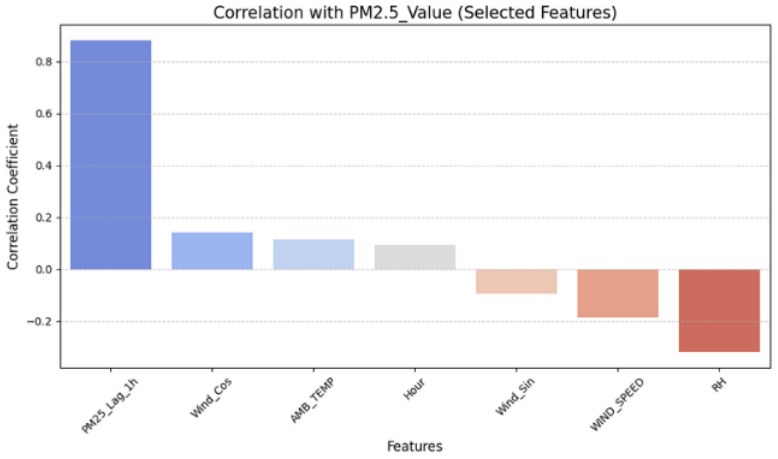


圖二、PM2.5 與氣象變數熱力圖

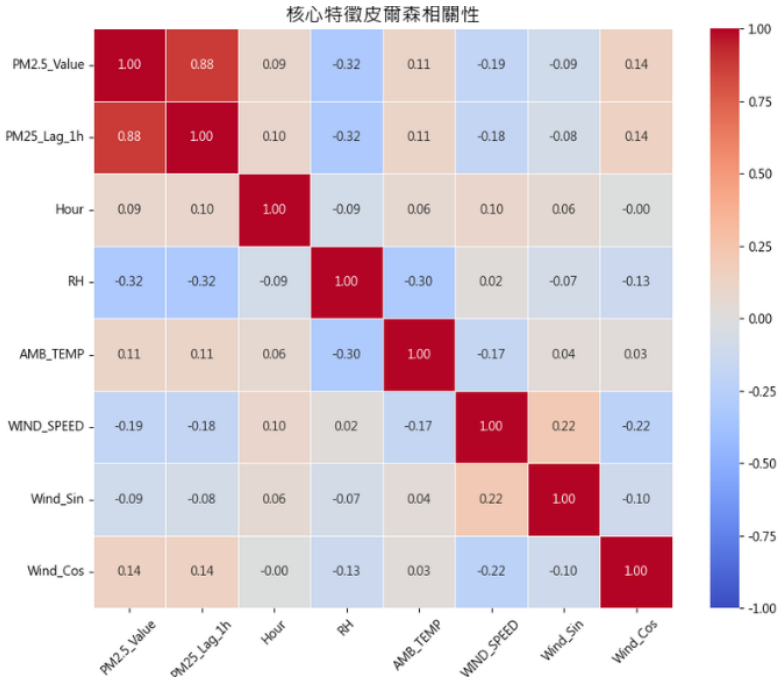
- **相對濕度**：相關係數為 **-0.32**，是所有氣象因子中負相關性最強的變數。因為高濕度環境使懸浮微粒吸濕增重，進而加速沉降作用，有效減少空氣中飄浮的PM2.5。

- **風速**：相關係數為 **-0.19**。較強的風速能增強大氣擴散能力，將聚集的汙染物吹散稀釋；反之，靜風條件下汙染物易累積。
- **降雨量**：相關係數為 **-0.16**。雖然降雨能將汙染物帶離大氣，但在本研究的數據集中，其線性關聯強度略低於濕度與風速。

除氣象因子外，本研究進一步納入時間特徵與時間參數進行綜合分析。**圖三**展示了所有與 PM2.5 的相關係數排序，**圖四**則呈現了核心特徵的相關性矩陣。



圖三、所有變數與 PM2.5 之相關係數排序圖



圖四、核心特徵皮爾森相關性矩陣熱力圖

五、研究步驟與方法（Methods）

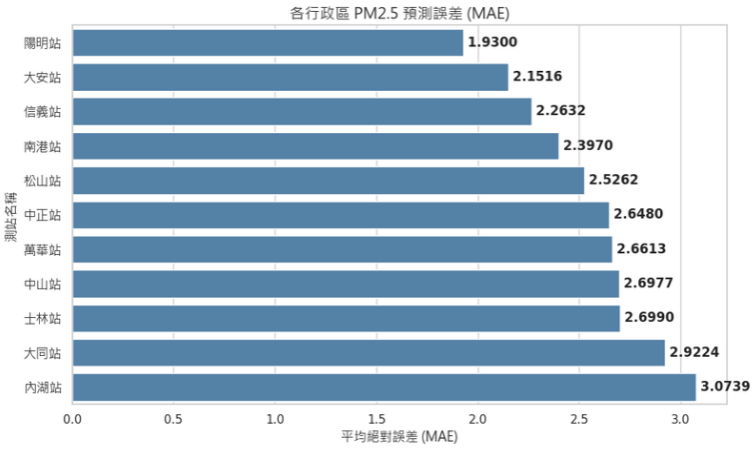
為全面評估模型效能並驗證其在不同情境下的適用性，本研究將實驗驗證過程分為以下三個主要部分進行探討：

1. **模型整體效能評估**：針對 2024 年 12 月 20 日至 29 日區間，分析四種模型（隨機森林、XGBoost、RNN、LSTM）的預測表現。

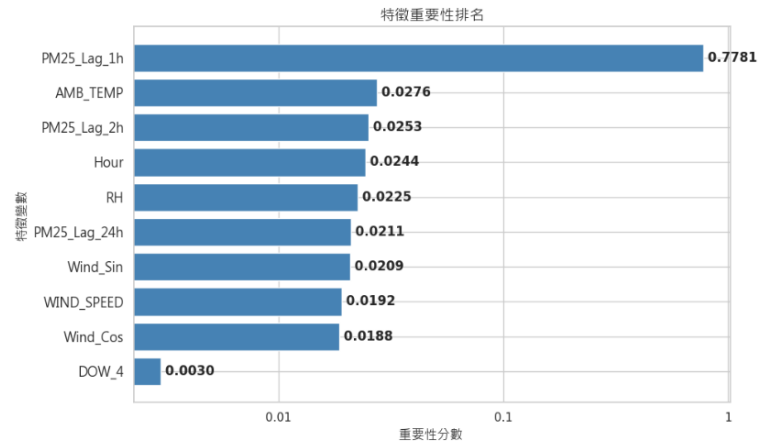
2. **單日個案深入分析**：聚焦於 12 月 31 日（跨年日），檢視各模型在小時制的細部預測能力。
3. **站點別空間分析**：選用表現穩定且準確度最佳的 LSTM 模型，針對不同地理特性的氣象測站（中山、中正、陽明）進行差異化分析。

本階段以判定係數（R^2 Score）與平均絕對誤差（MAE）作為指標，比較四種模型在連續 10 天內的預測能力。

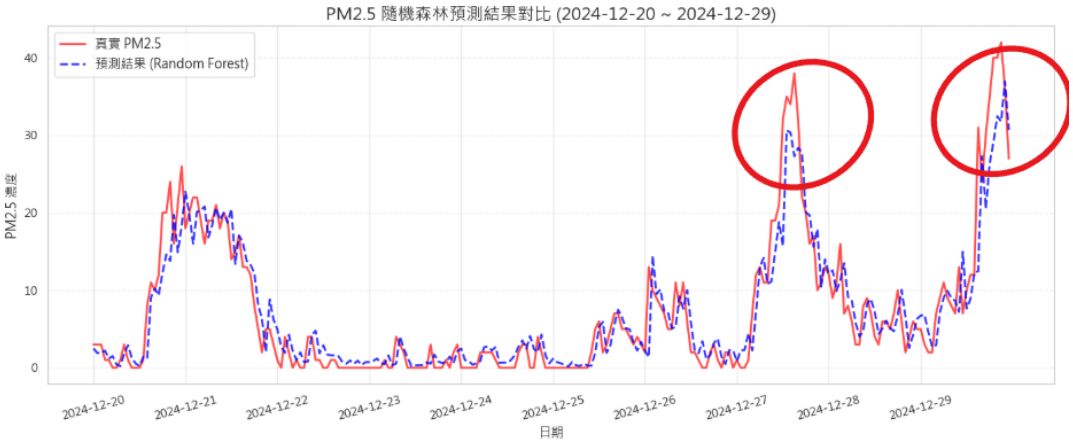
- **隨機森林:** 隨機森林模型在本階段展現了(R^2 Score) 85.48% 與（MAE）2.2986 的效能。由於使用多棵決策樹並行投票的特性，該模型在捕捉「突發性高汙染數值」上表現優異。如預測趨勢圖所示，當真實值出現劇烈波峰時，隨機森林能有效反應，這使其非常適合應用於汙染預警系統。



圖五、各行政區PM2.5預測誤差(隨機森林)

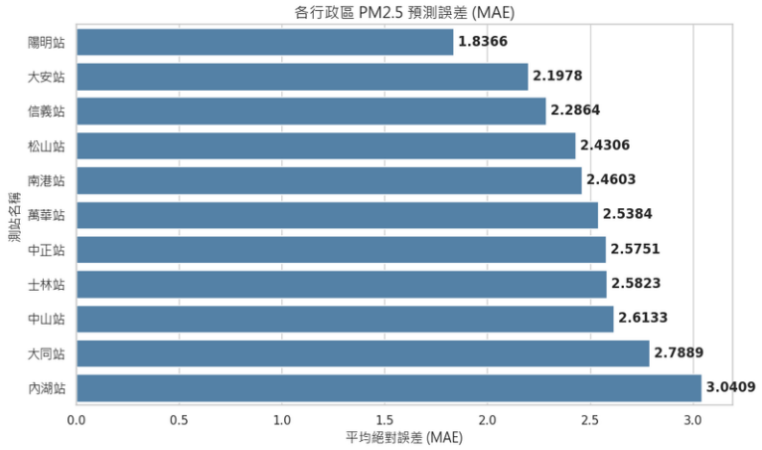


圖六、特徵重要性排名(隨機森林)

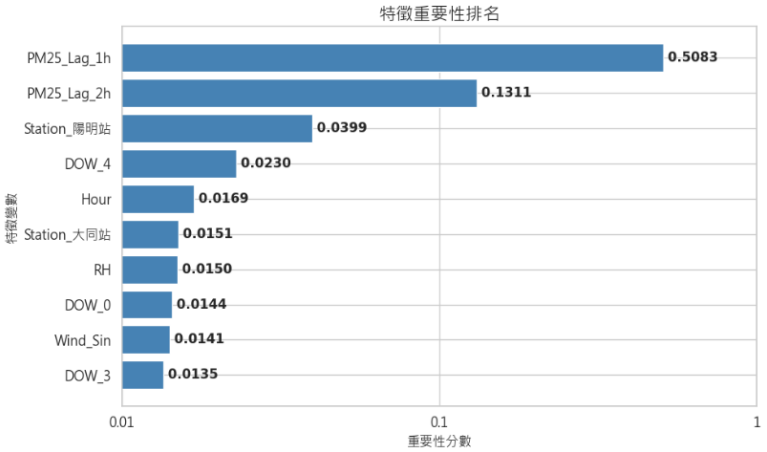


圖七、模型預測 & 評估(隨機森林)

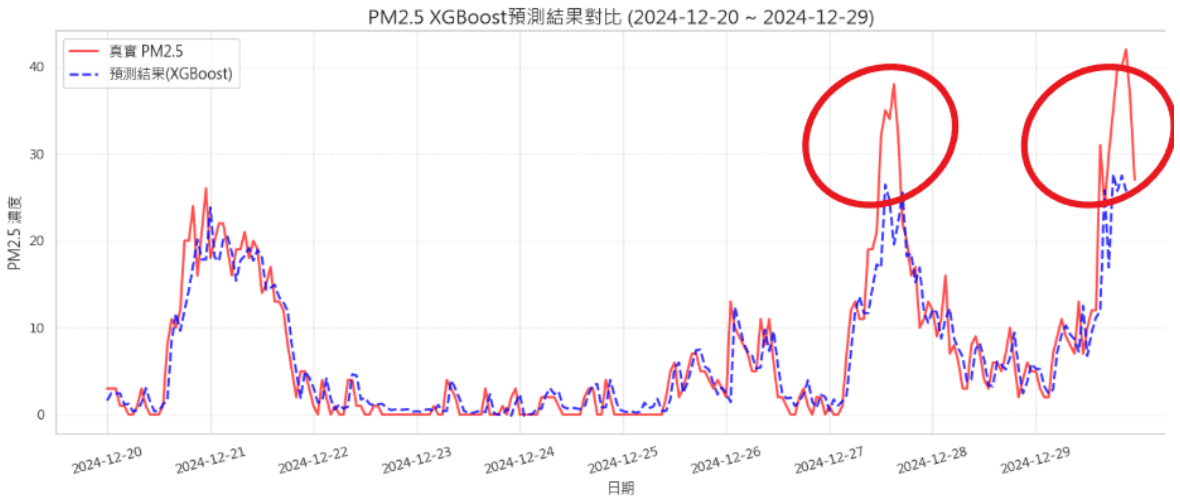
- **XGBoost:** XGBoost模型的表現相對較差，**(R^2 Score) 為 82.43%**，**(MAE) 為 2.3596**。分析其時間序列圖可發現，當 PM2.5 濃度平穩時預測準確，但在數值突然飆升時，模型因過度依賴前一小時特徵（Lag_1h），傾向於預測接近上一刻的數值（如前一小時 20，預測 22），導致預測曲線出現明顯的「Lag」現象，難以即時追上真實濃度的變化。



圖八、各行政區PM2.5預測誤差(XGBoost)

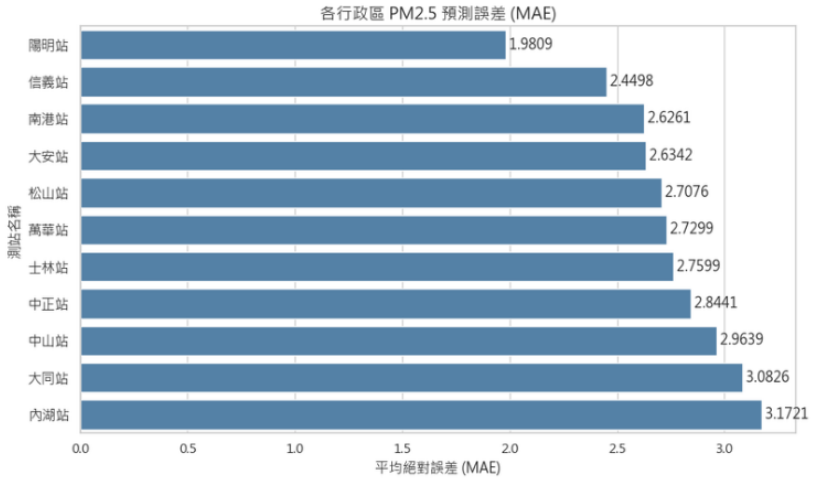


圖九、特徵重要性排名(XGBoost)

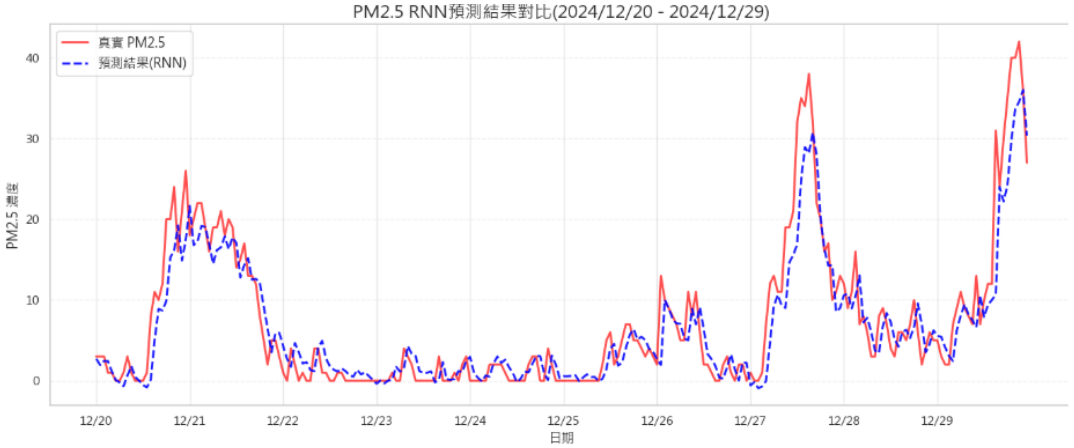


圖十、模型預測 & 評估(XGBoost)

- **RNN:** RNN模型因為自身的記憶機制，能有效處理時間序列資料，展現了 **(R^2 Score) 86.08%** 與 **(MAE) 2.2400** 的優異成績。其預測曲線與真實值的高度貼合，證明了遞歸架構在捕捉空氣品質連續變化上的優勢。

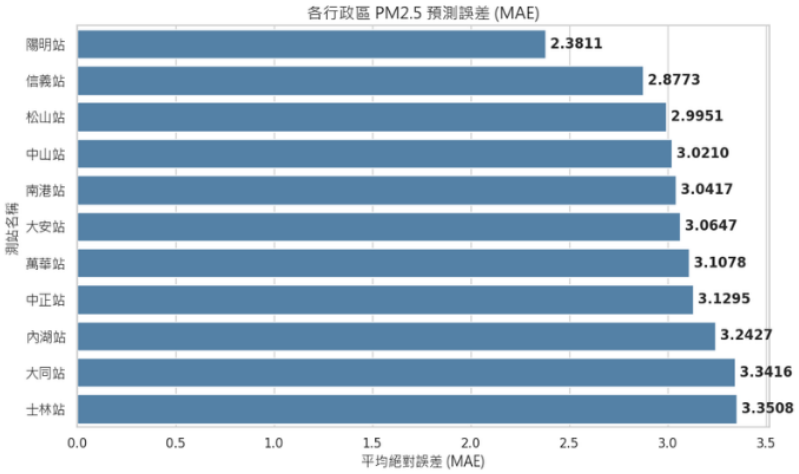


圖十一、各行政區PM2.5預測誤差(RNN)

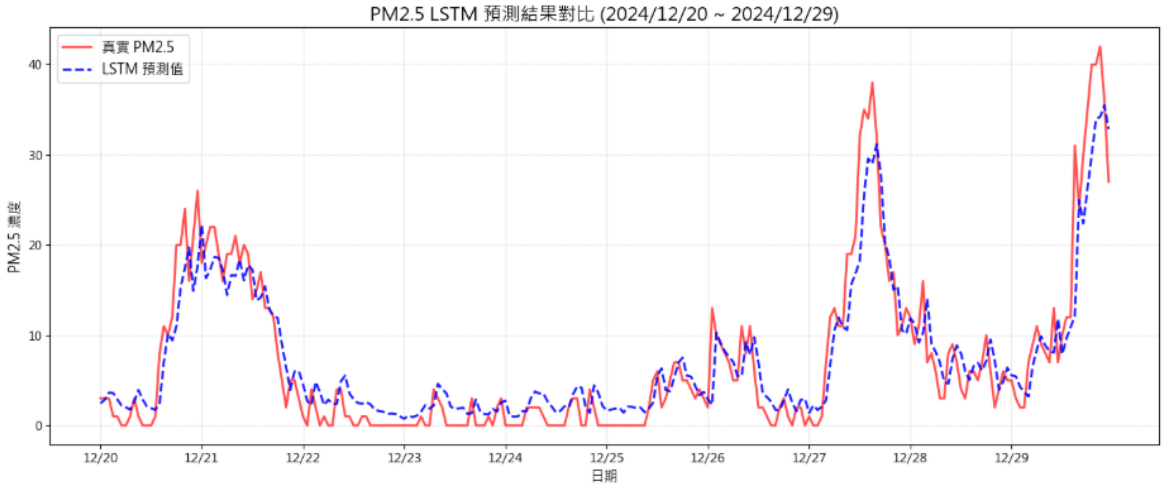


圖十二、模型預測 & 評估(RNN)

- **LSTM:** LSTM模型在本階段的解釋變異能力最強，(R^2 Score) **高達 87.48%** (MAE 2.2609)。相較於其他模型，LSTM 的預測軌跡最為平滑且穩定，顯示其長短期記憶單元能有效過濾雜訊並掌握長期趨勢，適合用於趨勢監控。



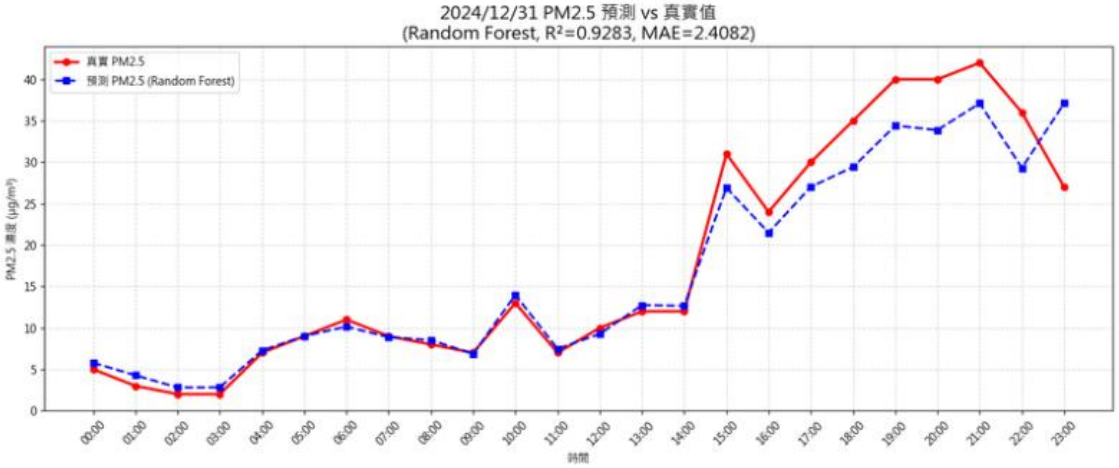
圖十三、各行政區PM2.5預測誤差(LSTM)



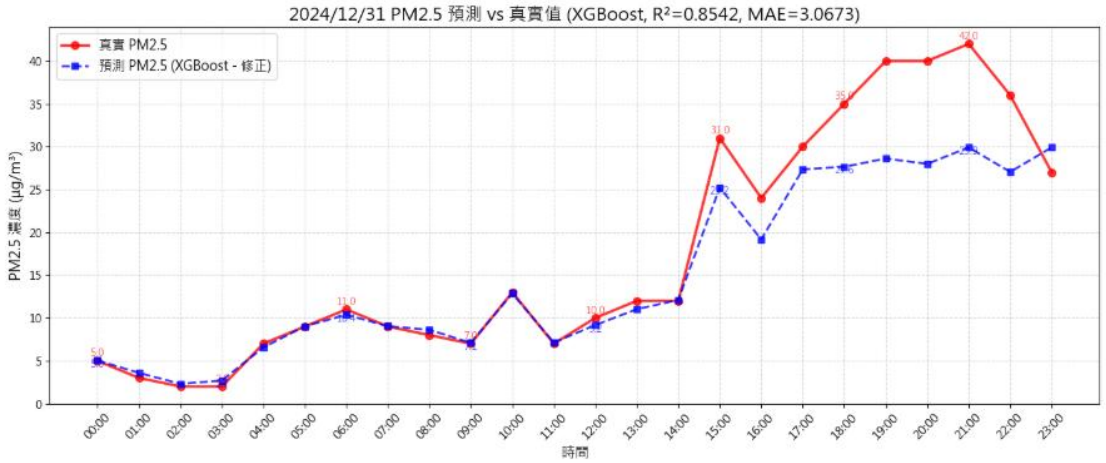
圖十四、模型預測 & 評估(LSTM)

為驗證模型在短時間內的精細度，本節聚焦分析 12 月 31 日當天的每小時預測結果。

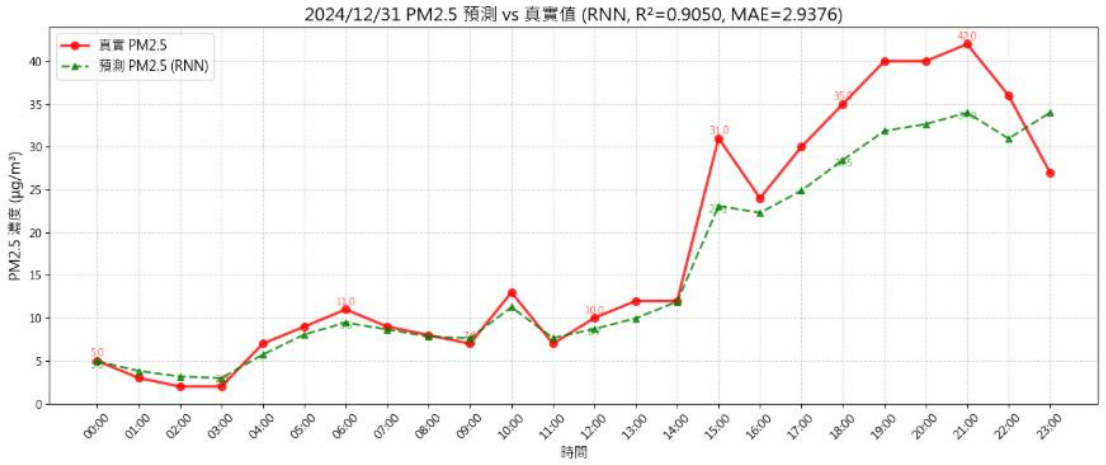
- **模型表現比較：** 在單日預測中，**隨機森林**以 (R^2 Score) 92.83%的極高準確度奪冠，再次證實其對非線性突波的捕捉能力。RNN (R^2 Score) 90.5%與 LSTM (R^2 Score) 89.26% 維持穩定水準，而XGBoost (R^2 Score) 85.42% 則修正能力較弱。



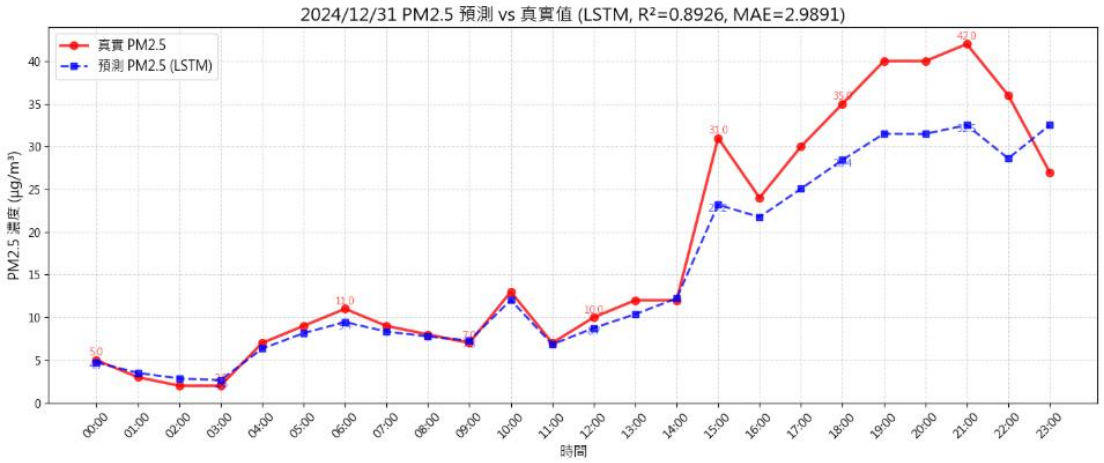
圖十五、12/31模型預測 & 評估（隨機森林）



圖十六、12/31模型預測 & 評估（XGBoost）



圖十七、12/31模型預測 & 評估（RNN）

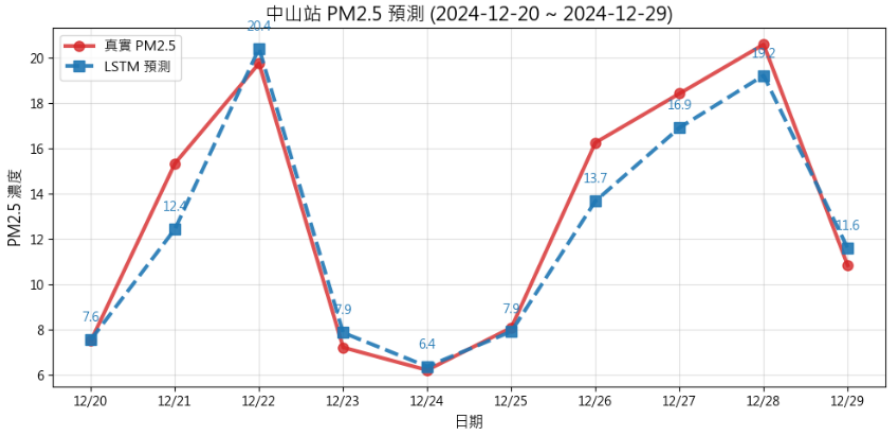


圖十八、12/31模型預測 & 評估（LSTM）

- **傍晚誤差分析：** 所有模型在 **16:00 至 23:00** 時段的預測誤差均顯著增加。本研究推測，此時段正值日落後大氣邊界層高度下降，擴散條件轉差，疊加下班尖峰時段的交通排放累積，形成了極難預測的非線性汙染暴增現象。

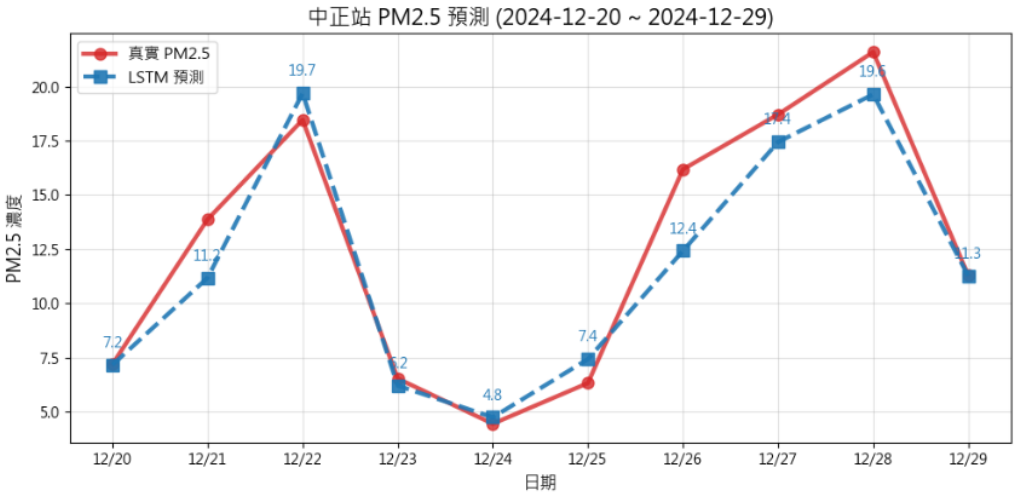
鑑於 LSTM 模型在整體評估中的高準確度與穩定性，本節選用 LSTM 針對台北市三個不同特性的測站進行深入分析，以探討地理空間對預測的影響。

- **中山站**：位於市中心交通繁忙區，PM2.5 濃度波動較大。LSTM 模型在此站展現了良好的追蹤能力，能準確預測出早晚的濃度變化峰值。



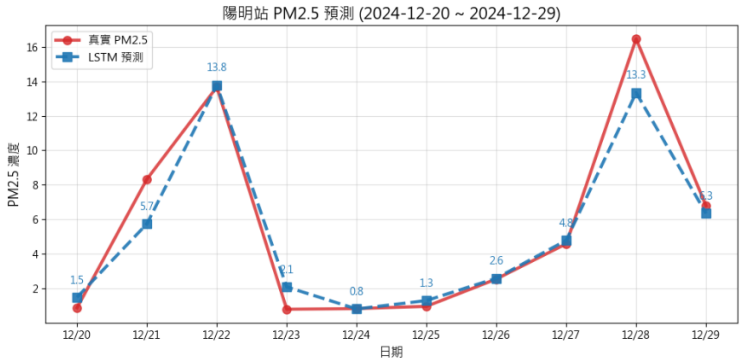
圖十九、PM2.5站點預測（中山站）

- **中正站**：同樣位於核心都會區，模型預測結果與真實值高度吻合，顯示 LSTM 對於都會型態的汙染特徵具有良好的泛化能力。



圖二十、PM2.5站點預測（中正站）

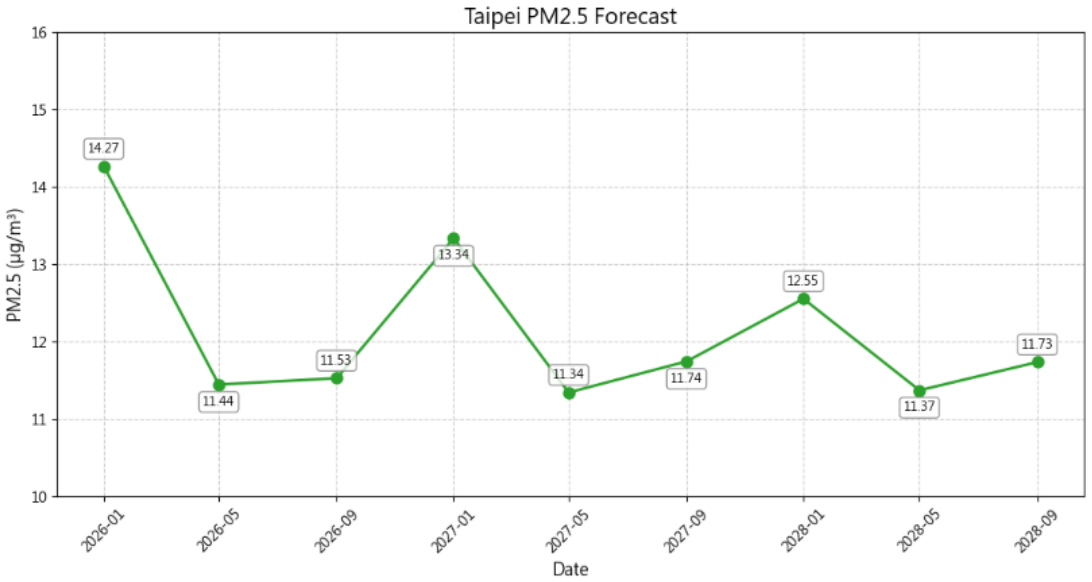
- **陽明站**：位於山區，環境較為單純且汙染濃度普遍較低。LSTM 模型在此站的預測曲線極為平滑，準確捕捉了低濃度背景下的微幅波動。



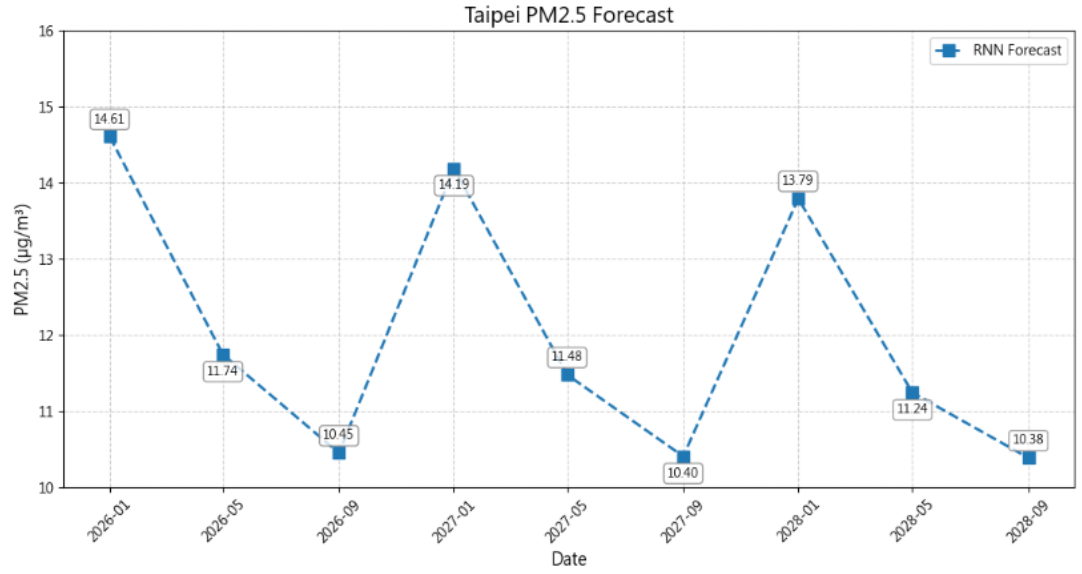
圖二十一、PM2.5站點預測（陽明站）

六、2026-2028 年長期趨勢預測（Long-term Trend Forecast（2026-2028））

鑑於 RNN 與 LSTM 模型在驗證階段展現較高的準確度與穩定性，本研究選用此兩模型對台北市 2026 年至 2028 年的 PM2.5 濃度進行長期模擬預測 。預測結果如 **圖二十二**（LSTM 模型）與 **圖表二十三**（RNN 模型）所示。



圖二十二、2026-2028 台北市 PM2.5 使用 LSTM 預測



圖二十三、2026-2028 台北市 PM2.5 使用 RNN 預測

綜合兩張趨勢圖的分析結果如下：

- **季節性波動**：

  未來三年的 PM2.5 濃度將持續呈現顯著的季節性循環。觀察 圖表 與 圖表 的波峰，每年 1 月至 3 月（如橫軸所示 2026-01、2027-01）均為汙染高峰期，預測數值約落在 **13-14** 區間（例如 RNN 預測 2026 年 1 月達 14.61）；而夏季受對流旺盛影響，擴散條件較佳，數值將回落至 **10-11** 左右（例如 LSTM 預測 2026 年 5 月降至 **11.44**）。

- **長期趨勢**：

  整體而言，從兩張折線圖的走勢可見，模型預測台北市未來三年的 PM2.5 平均濃度將維持在 **11-14** 的區間內震盪。曲線並未呈現持續上升或大幅下降的單調趨勢，顯示在現有環境條件假設下，空氣品質將維持穩定的週期性變化。

- **模型比較**：

對比 圖 1（綠線）與 圖 2（藍虛線）可發現，RNN 與 LSTM 的長期預測趨勢高度一致。兩者皆成功捕捉了冬季汙染升高與夏季下降的週期性特徵，且峰值與谷值的發生時間點幾乎完全吻合，進一步驗證了本研究預測結果的可靠性。

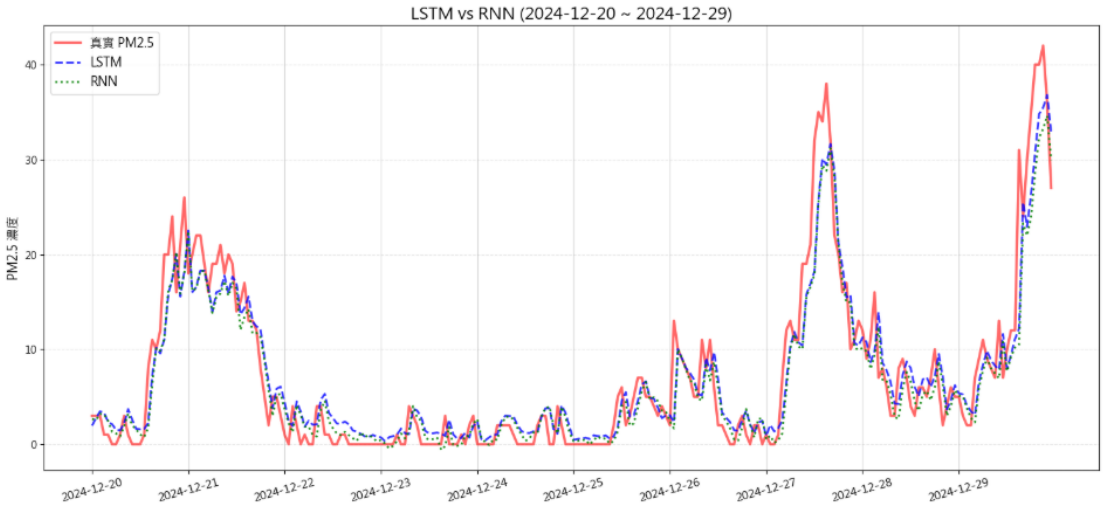
七、結果與分析（Results and Analysis）

本研究首先透過皮爾森相關係數矩陣檢視各特徵變數與目標變數（PM2.5 濃度）之關聯性。分析結果顯示，時間特徵呈現極強的正相關性，其中「前一小時 PM2.5 濃度（PM25_Lag_1h）」的相關係數高達 **0.88**，顯示空氣汙染具有高度的時間延續性與相關性。

在氣象因子方面，相對濕度（Relative Humidity, RH）、風速（Wind Speed）與降雨量（Rainfall）皆呈現負相關。其中，相對濕度的相關係數為 **-0.32**，為氣象變數中影響最大者，這可能歸因於高濕度環境下水氣易吸附懸浮微粒並產生濕沉降作用，進而降低空氣中 PM2.5 濃度。風速係數為 **-0.19**，顯示強風有助於大氣擴散，降低汙染物堆積。

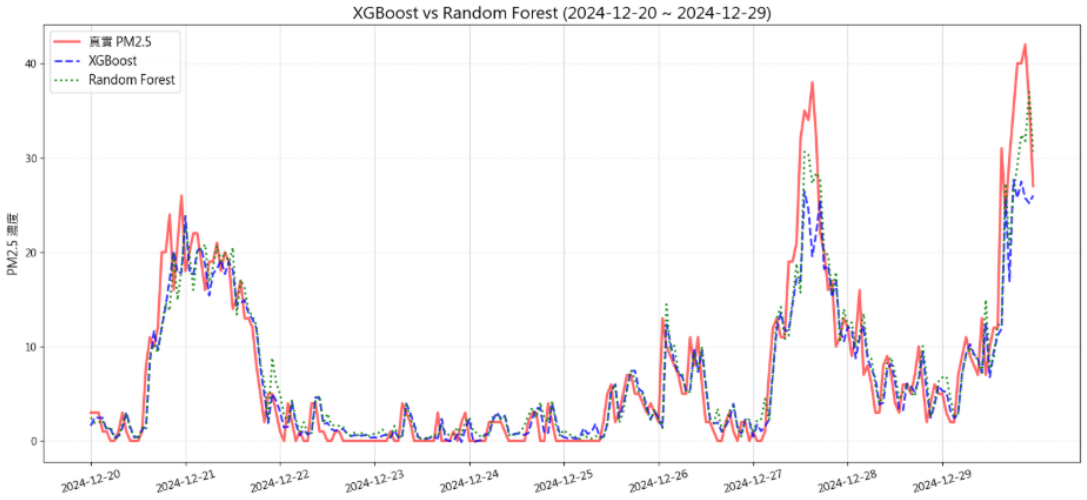
本研究運用隨機森林（Random Forest）、XGBoost、RNN 與 LSTM 四種模型進行預測，並採用判定係數（R^2 Score）與平均絕對誤差（MAE）作為評估指標。實驗數據顯示：

- **深度學習模型表現穩健**：LSTM 模型在（R^2 Score）上表現最佳，達到 87.48%，顯示其解釋變異的能力最強；RNN 模型則在平均絕對誤差（MAE）上表現最優，數值為 2.24。這證實了循環神經網路架構在處理長序列時間資料上的優勢。



圖二十四、RNN 與 LSTM 之比較圖

- **集成學習模型之特性**：隨機森林模型的（R^2 Score）為 85.48% ，略低於深度學習模型，但優於 XGBoost（82.43%）。XGBoost 的 MAE 較高（2.3596），原因在於其過度依賴前一小時特徵，導致在汙染濃度劇烈變化時出現預測滯後現象，無法即時捕捉數值的快速爬升。



圖二十五、XGBoost 與 Random Forest 之比較圖

各模型效能評估比較表

模型（Model）	（R^2 Score）	MAE（平均絕對誤差）
LSTM	87.48%	2.2609
RNN	86.08%	2.2400
Random Forest	85.48%	2.2986
XGBoost	82.43%	2.3596

當我們進一步分析 2024 年 12 月 31 日之單日預測結果，發現隨機森林模型在捕捉「突發性高汙染峰值」的能力上優於其他模型，能夠有效反應極端數值的變化。相對地，各模型在傍晚至夜間時段（16:00-23:00）的預測誤差普遍增加。此現象推測與日落後大氣邊界層高度下降導致擴散條件轉差，加上下班尖峰時段交通排放源增加，形成非線性的汙染累積有關。

在空間維度上，不同測站的預測誤差存在顯著差異。陽明站（山區）因環境單純，MAE 最低（1.93）；而內湖站、大同站等位於交通繁忙或盆地地形之測站，MAE 則高達（3.0）以上。這顯示地理特徵與局部排放源的複雜度是影響模型準確度的關鍵外部變數。

八、結論(Conclusions)

本研究旨在建立適用於台北市之 PM2.5 短期濃度預測模型，整合了環境部與氣象局之歷史監測數據，並針對資料特性進行了滯後特徵工程與風向向量化處理。研究證實，透過機器學習與深度學習技術，能夠有效掌握空氣品質的變化趨勢。

綜合實驗結果，我們提出以下核心發現：

1. **歷史數據的重要性**：前一小時的 PM2.5 濃度是短期預測中最關鍵的解釋變數，其影響力遠高於單一氣象因子。
2. **模型適用性差異**：LSTM 與 RNN 等深度學習模型在整體趨勢預測上展現了高度的

穩定性與準確度，適合用於長期的空氣品質監測與趨勢分析。隨機森林模型則在應對數值劇烈波動的突發事件上具有優勢，適合應用於即時的汙染警示系統。

3. **環境因素的限制**：傍晚交通尖峰與大氣擴散條件不佳之時段，仍是目前模型預測的挑戰所在，顯示單純依賴歷史數值與基礎氣象資料在特定情境下仍有不足。

基於模型對 2026 年至 2028 年的預測趨勢，台北市 PM2.5 濃度預計將維持季節性波動模式，冬季（1-3月）仍為汙染高峰期，平均數值約落在 13-14 之間。建議未來研究可朝向混合模型發展，例如結合 LSTM 的長期記憶能力與隨機森林的峰值捕捉能力，並納入更細緻的交通流量數據或大氣邊界層高度資料，以進一步提升尖峰時段與高汙染事件的預測精準度。

九、分工表(Division of labor)

12156206 章祖綸	Word製作、製作、評估模型預測PM2.5、特篩篩選
12156217 陳翡甄	收集資料、資料預處理、處理程式碼
12156223 吳承瑀	ppt製作、處理程式碼
12156229 高碩辰	Word製作、製作、評估模型預測PM2.5、特篩篩選

十、文獻(References)

1. **環境保護署 (EPA)，台灣**

- 透過政府公開平台獲得的空氣品質資料，包含PM2.5濃度。

監測站歷史資料查詢 - 臺北市環境品質資訊網

2. **Codis氣候觀測資料查詢**

- 用於參考與提升預測準確度的氣候相關資料。

全國細懸浮微粒手動監測資料 | 政府資料開放平臺

3. **A Machine Learning-Based Ensemble Framework for Forecasting PM2.5 Concentrations in Puli, Taiwan**

- 該研究提供了相關資料和方法，供模型比較與驗證使用。

A Machine Learning-Based Ensemble Framework for Forecasting PM2.5 Concentrations in Puli, Taiwan

4. **基於深度學習技術應用於空氣品質PM2.5預測**

- 採用深度學習技術進行空氣品質預測的資料與方法。

基於深度學習技術應用於空氣品質PM2.5預測 | Airiti Library 華藝線上圖書館

5. 台北市部分行政區PM2.5數值

環境部空氣品質監測網