機器學習建構 PM2.5 濃度模型之研究-以台中市為例

陳大田[1] 陳美心[2] 周天穎[2] 陳耀鐘[2*] 甯方璽[3]

[1]逢甲大學建設學院建設規劃與工程博士學位學程博士生
[2]逢甲大學地理資訊系統研究中心
[3]國立政治大學地政學系

摘 要

本研究探討 6 種機器學習(Machine Learning)演算法建構台中地區 PM2.5 濃度建模效能，並分析引入時空因子對模型精度的效益。成果顯示，同時加入時空因子或僅加入時間因子皆能提升建模精度，而後者對建模精度提升更為顯著，反映出，時間因子在 PM2.5 濃度建模時扮演相當重要角色。隨機森林(Random Forest, RF)在不同的自變數模型均具有最佳表現，相較於其他方法，M16、M14 及 M12 三種自變數模型在 RMSE 精度提升率分別介於 1.90%~22.41%、0.24%~22.50%與 1.39%~21.15%，而在 MAE 部分，精度提升率分別介於 4.66%~22.47%、1.29%~22.28%與 1.85%~22.06%。顯示採用隨機森林(RF)並額外引入時間因子，為本研究中最適 PM2.5 濃度建模方式。

關鍵詞：PM2.5濃度，機器學習，隨機森林

Machine Learning-Based Modeling of PM2.5 Concentration: A Case Study in Taichung City

Tai-Tien Chen[1], Mei-Hsin Chen[2], Tien-Yin Chou[2], Yao-Chung Chen[2*], and Fang-Shii Ning[3]

[1]Ph.D. Program for Infrastructure Planning and Engineering, Feng Chia University
[2]GIS Research Center, Feng Chia University
[3]Department of Land Economics, National Chengchi University

ABSTRACT

This study evaluated six machine learning algorithms for modeling PM2.5 concentration in the Taichung region and analyzed the impact of introducing spatiotemporal factors. Incorporating both spatiotemporal and temporal factors improved accuracy, with temporal factors exhibiting significant enhancement. Random Forest (RF) consistently outperformed other methods, with RMSE improvements of 1.90%-22.41%, 0.24%-22.50%, and 1.39%-21.15% for M16, M14, and M12 models, respectively. MAE improvements ranged from 4.66%-22.47%, 1.29%-22.28%, and 1.85%-22.06%. These results highlight RF with the additional inclusion of temporal factors as the optimal approach for PM2.5 concentration modeling in this study.

Keywords: pm2.5 concentration, machine learning, Random Forest

文稿收件日期 112.03.09; 文稿修正後接受日期 112.○○.○○;*通訊作者
Manuscript received Mar 09, 2023; revised…;* Corresponding author

一、前　言

全球暖化與氣候變遷為目前聯合國主要重視的議題之一，因為這將影響人類未來的生存與發展，空氣汙染被視為造成地球環境損害的重要因素，此外，對於人類的健康與否有決對的相關性。環境顆粒物(Particulate Matter, PM)為空氣汙染的主要指標之一，泛指懸浮在空氣中的小固體或液體顆粒，故又稱大氣懸浮微粒(atmospheric particulate matter)，其顆粒微小至難以用肉眼辨識，惟仍有尺度的差異，一般以懸浮微粒空氣動力學直徑小於或等於 2.5 微米(μm)的懸浮微粒稱為細懸浮微粒(PM2.5)，而小於或等於 10 微米(μm)的懸浮微粒稱為懸浮微粒(PM10)[1]。

研究證實空氣污染會導致人類健康、生活環境和社會經濟的惡化，已被列為全球性的危害。依據世界衛生組織報告，“環境(室外)和家庭空氣污染的綜合影響每年導致大約 700 萬人過早死亡”。引起的原因有中風[2]、心臟病[3]、慢性阻塞性肺病[4]、肺癌[5]和急性呼吸道感染[6]，世界衛生組織的統計數據亦顯示，99%的人呼吸著高度污染的空氣，超過了低收入和中等收入國家的空氣品量限值[7]。基此，如何有效地建構 PM2.5 濃度模型，進行預警與管控，為現今各國政府所努力的目標。

傳統對於 PM2.5 濃度的建模方法可分為物理方法和統計方法，物理方法利用物理學、化學、氣象學和生物學等原理建構數學模型，基於物理方法的 PM2.5 濃度代表性模型有：區域多尺度空氣品質模型(Community Multiscale Air Quality Mode, CMAQ)[8]、嵌套空氣品質預測模型系統(Nested Air Quality Prediction Modeling System, NAQPMS)[9]、化學天氣研究與預報模型(Weather Research and Forecasting Model with Chemistry, WRF-Chem)[10]和衛星數據反演(Satellite Data Inversion)[11, 12]。然而，前述模型須使用已知的理論模型來描述因素之間的關係進而估計 PM2.5 濃度。不幸地，複雜的理論模型需要大量的觀測數據和長時間的試驗週期，須耗費較高的資源、人力與時間成本。統計方法與前述物理方法相比，統計方法不需要預先確定理論模型，統計方法中如克里金(kriging)、地理統計(Geostatistics)及反距離加權（inverse distance weighting，IDW）等[13-17]，藉由對歷史污染物數據的調查和分析來建構模型。隨著電腦計算能力與硬體的進步，機器學習(Machine Learning)方法在建模與預測領域的應用更為引人注意，有別於傳統統計方法，機器學習更強調對數據的預測能力及對複雜模式的學習，因此已被應用於空氣汙染的建模與預測，如決策樹(Decision Tree, DT)[18, 19]、隨機森林(Random Forest, RF)[20, 21]、高斯過程回歸(Gaussian Process Regression, GPR)[22, 23]、支持向量機(Support Vector Machine, SVM)[24, 25]、人工神經網路(Artificial Neural Network, ANN) [26, 27]、卷積神經網路(Convolution Neural Network, CNN)[28, 29]、循環神經網路(Recurrent Neural Network, RNN)[30, 31]、長短期記憶(Long Short-Term Memory, LSTM)[32, 33]和雙向長短期記憶(Bidirectional Long Short-Term Memory, Bi-LSTM)[34]等。

顯而易見，機器學習因高預測能力與建模效能，已成為當今進行數據建模的主流方法，雖然已有眾多文章基此技術進行 PM2.5 濃度建模，然而，眾所周知，PM2.5 濃度時空分布極為複雜，影響 PM2.5 濃度的因素眾多，再者，台灣空氣品質監測站的幾何分布並非每個地區皆相同，前述這些因素都是建模過程中所必須考量的條件。另因各文獻中所研究區域、採用建模因子及方法比較方式等存在一些差異，因此，建模條件的不同，恐造成研究結果的變異。若就資料來源、處理方式及模型訓練等觀點來看，同一個建模方法未必適用所有地區，針對特定區域選擇適用的建模方式，可以提升空氣污染防治的效用，爰此本研究參酌文獻後選用 6 種在 PM2.5 濃度建模常見主流方法，同時也分析引入時空因子對模型精度的效益，在相同實驗基礎之下，分析台中地區較適建模之方式。

二、研究方法

2.1 資料蒐整與預處理

台中市為台灣人口第二大城市，亦為臺灣第二大都會區「臺中彰化都會區」的核心都市，空氣汙染防制為人民所關切議題並為政府施政重點項目之一，因此，本研究以台中市為研究區域。機器學習方法通常使用各種因素來預測 PM2.5 濃度，包括氣象、空氣污染、土地利用和衛星遙感數據[35-38]。所選因素取決於建模目標和可用數據。在本研究中，重點是開發一種多站 PM2.5 濃度預測模型的方法。為確

保數據的可獲得性、多樣性及穩定性，及更能區別不同模型的建模效能，在台中市內 16 個空氣品質監測站中(5 個環保署測站、6 個台中市環保局測站及 5 個台電測站)，選用 5 個環保署地面空氣品質監測站的氣象和空氣污染因子，以及相應的時空因子進行建模。

建模時所採用數據為台灣環境保護署空氣品質監測網所下載之台中市大里(Dali)、忠明(Zhongming)、豐原(Fengyuan)、沙鹿(Shalu)及西屯(Xitun)5 個監測站 2020 年 1 月 1 日至 12 月 31 日計 366 天之空氣汙染與氣象資料。資料的時間解析度為每小時 1 筆，每 1 觀測站原則上 1 天計有 24 筆觀測數據，惟當中有些時刻觀測量無法被記錄，為了讓資料在時間序列上不間斷，本研究採用最鄰近法，也就是最接近該時刻的有效觀測值作為補值。本研究選取 4 個散布於台中市的觀測站資料為建模時之訓練及驗證資料，另外為避免模型因過度擬合而影響成果評估，另採用位於建模區域內的西屯監測站觀測量實施獨立測試，以驗證不同方法所建構之 PM2.5 濃度模型之效能，本研究所採用訓練及測試監測站分布如圖 1 所示。

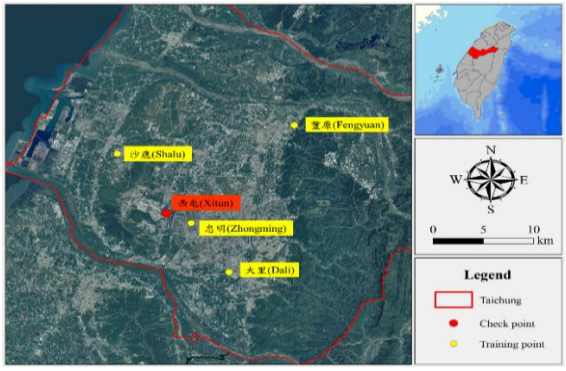

圖 1. 研究區域監測站分布圖

本研究中除採用空氣汙染及氣象因子外，考量空間與時間因子，應有助於模型的精度。例如，根據地理位置的不同，可能存在不同地區之間的 PM2.5 濃度差異。同樣，根據時間的不同，PM2.5 濃度可能會呈現出週期性的變動。因此，引入時空因子，期望可以增加模型在預測時的彈性與可靠性。實驗中以年積日(day of year)、日積時(hour of day)、經度(longitude)與緯度(latitude)為時空因子，並設計包含與不包含 4 個時空因子及包含 2 個時間因子等 3 種情境，因此 3 種情境模型的自變數分為 16、14 及 12 個。7 個空氣汙染、5 個氣象及 4 個時空模型為自變數(independent variable)，相對應的 PM2.5 濃度為應變數(dependent variable)，各變數的資料特性如表 1 所示。

表 1. 變數資料特性

變數	因子	資料類型	描述特徵
應變數	PM2.5	連續數值	[1~92];mean:15.3413; std:10.7610
自變數	CO		[0.05~9.44];mean:0.3350; std:0.2361
	NO_2		[0~87.60];mean:11.8343; std:7.6964
	NO		[0~103.70];mean:2.5893; std:4.2609
	NO_X		[0.30~147.10];mean:14.4163; std:10.2696
	SO_2		[0~18.10];mean:2.0637; std:1.1181
	O_3		[0~128.20];mean:29.7089; std:18.0736
	PM10		[0~211];mean:28.4646; std:17.9718
	Wind speed		[0~9.40];mean:166.0228; std:123.6245
	Wind direction		[0~360];mean:0.3350; std:0.2361
	relative humidity		[14~100];mean:72.4501; std:12.6652
	rainfall		[0~93.2];mean:0.0980; std:1.0931
	ambient temperature		[0.1~38.1];mean:24.6161; std:5.3295
	day of year		[1~366];mean:183.5; std:105.6559
	hour of day		[1~24];mean:12.5000; std:6.9223
	latitude		[120.5688~120.7417]; mean:120.6494;std:0.0583
	longitude		[24.0996~24.2566]; mean:24.1792;std:0.0557

此外，由於自變數的單位與計量大小不一致，研究中針對自變數採用最小值最大值正規化(Min-Max Normalization)，將其轉換介於 0 至 1 的數值，避免突出數值較高的指標在綜合分析中的影響程度，相對減弱數值水平較低指標的作用，以提升模型的收斂速度及提高模型的精準度，其轉換方式如式(1)所示；

$$X_{nom} = \frac{X - X_{min}}{X_{max} - X_{min}} \in [0,1] \quad (1)$$

式中，X_{nom}為經正規化後的自變量值；X

為正規化前自變量值; X_{min}為正規化前自變量數據集之最小值； X_{max}為正規化前自變量數據集之最大值。

2.2 理論基礎

機器學習在 PM2.5 濃度建模研究中具有廣泛的應用性。由於 PM2.5 濃度受到多個因素的影響，包括氣象條件、排放源、地理特徵等，因此建立一個準確的預測模型是具有挑戰性的。機器學習方法通常能夠處理多變量、非線性和非參數化的問題，考慮多個特徵和相互作用，可以進行模型優化和特徵選擇，並且能夠根據大量的數據進行學習和預測。基此，本研究經回顧 PM2.5 濃度建模相關文獻後，以該方法為研究對象，探討隨機森林(RF)、高斯過程回歸(GPR)、人工神經網路(ANN)、支持向量機(SVM)、決策樹(DT)及卷積神經網路(CNN)等 6 種出色的機器學習演算法於建構台中地區 PM2.5 濃度模型之效能。

2.2.1 隨機森林

隨機森林(RF)為決策樹(DT)的進化，由大量決策樹(DT)組成如圖 2 所示，通過聚合每個決策樹(DT)的貢獻來預測目標值，使得模型更穩健，更不容易過擬合，在變量數量遠大於觀察數量的條件中顯示出出色的性能。此外，它的通用性足以應用於大規模問題，容易適應各種學習任務[39]。其預測模型如式(2)所示：

$$Y_P = \frac{\sum_{i=1}^{n} T_i(X)}{n} \quad (2)$$

式中，Y_P是模型對於自變數(因子)X 的 PM2.5 濃度預測值，n 是森林中決策樹(DT)的數量，$T_i(X)$是第 i 棵決策樹(DT)對 X 的 PM2.5 濃度預測值。

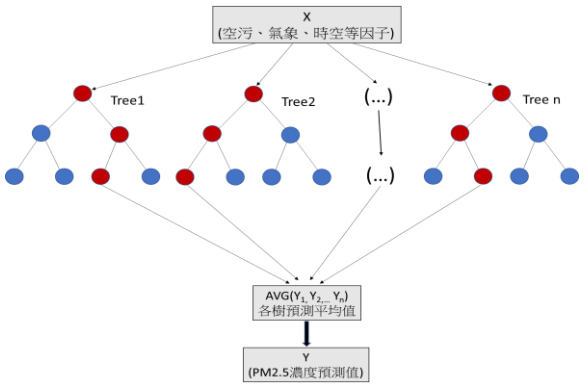

圖 2. 隨機森林示意圖

2.2.2 高斯過程回歸

高斯過程回歸(GPR)是一種用於推理的非參數貝葉斯方法，高斯過程定義一個先驗函數，在從先驗分布中觀察到一些值後將其轉換為後驗函數如圖 3 所示。其可用於直接推斷感興趣函數的分布，每個高斯過程都可以看做是多元高斯分布的無限維推廣，在小型數據集上運行良好，並且能夠提供預測的不確定性測量，被應用於解決多種形態的問題，包括材料科學、化學、物理學和生物學領域[40]。其預測模型如式(3)所示：

$$Y_P = K(X,x)K(x,x)^{-1}y \quad (3)$$

式中，Y_P是模型對於自變數(因子)X 的 PM2.5 濃度預測值，x 與 y 分別為訓練集的自變量與應變量，K為協變方函數。

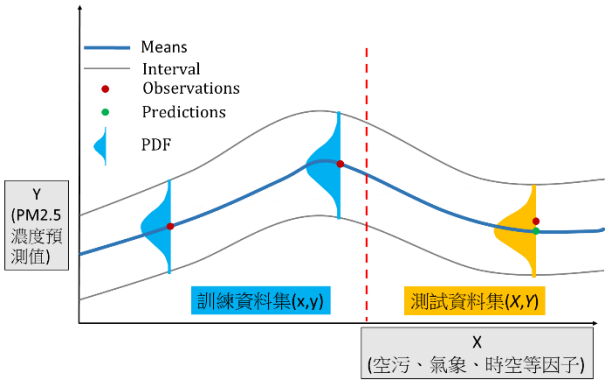

圖 3. 高斯過程回歸示意圖

2.2.3 人工神經網路

人工神經網路(ANN)受大腦神經元工作的啟發，架構由輸入層、隱藏層和輸出層組成如圖 4 所示，可在輸入層和輸出層之間堆疊多個隱藏層，以增加模型複雜度，提升建模效能。一般來說，神經網路的力量在於其神經元的互連以及與每個互連相關聯的一組權重，並自動更新網路權重以最小化誤差函數[41]。其預測模型如式(4)所示：

$$Y_P = f(X,\omega) = \omega_{L+1}\Phi_L(\omega_L\Phi_{L-1}(\cdots \Phi_2(\ \omega_2\Phi_1(\omega_1 X + b_1) + b_2)\cdots) + b_L) \quad (4)$$

式中，Y_P是模型對於自變數(因子)X 的

PM2.5 濃度預測值，L是網路的層數，ω_L和b_L分別是第 L 層的權重和誤差項，Φ_L是第 L 層的激勵函數。

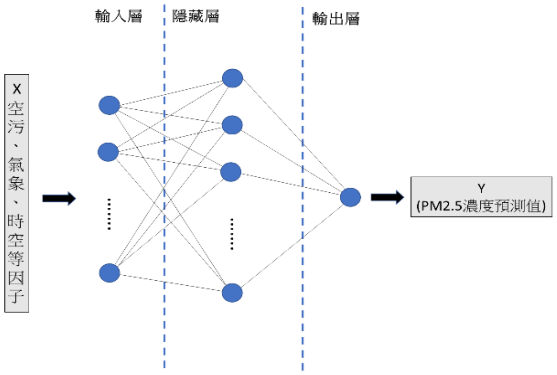

圖 4. 人工神經網路示意圖

2.2.4 支持向量機

支持向量機(SVM)的回歸問題，在於選擇一個超平面，也就是利用一條固定寬度的條帶覆蓋儘可能多的樣本點，在條帶內的點其誤差視爲零，僅計算超出條帶點的誤差，從而使得總誤差儘可能的小如圖 5 所示。此外，通過使用內核技巧，輸入數據被映射到一個新的空間中[42]，在高維特徵空間中尋找超平面，來擬合複雜的非線性問題。其預測模型如式(5)-(7)所示：

$$Y_P = f(X, \omega) = \omega^T X + b \tag{5}$$

with

$$minimize \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{6}$$

$$s.t. \begin{cases} y_i - (w \cdot x_i) - b \le \varepsilon + \xi_i \\ (w \cdot x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, m \end{cases} \tag{7}$$

式中，Y_P是模型對於自變數(因子)X 的 PM2.5 濃度預測值，ω是權重向量，b 是誤差項，i 表示觀測量序數，ξ, ξ^*分別為正向與負向的鬆弛變數(slack variables)。

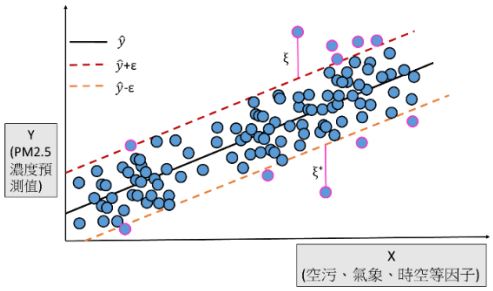

圖 5. 支持向量機示意圖

2.2.5 決策樹

決策樹(DT)以樹幹與葉子分枝概念演算法如圖 6 所示，每個節點都代表一個自變量和一個門檻值，用於將數據集分成兩個子集。這個分割過程會一直持續，直到達到某個停止條件，例如達到最大深度、節點中的樣本數少於一個預定數量等。在訓練階段學習每個決策節點的值和自變數，據此組成多個自變數間的最佳組合數據子集並以其應變數平均值作為該子集的預測值[43]。其預測模型如圖 6 所示，根據實際的回歸樹結構和訓練過程，通過遞歸地在樹的分支上選擇自變數來進行預測，最終到達葉子節點並使用葉子節點的預測值作為最終預測結果。

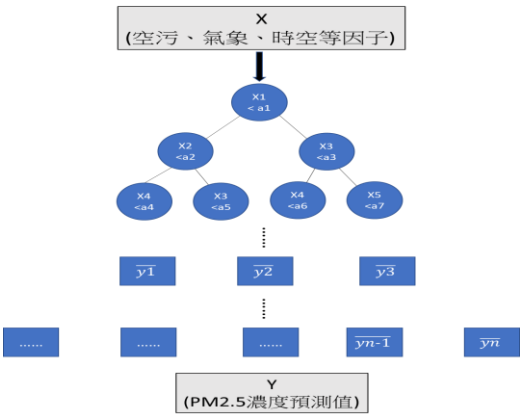

圖 6. 決策樹示意圖

2.2.6 卷積神經網路

卷積神經網路(CNN)為代表性的深度學習(Deep Learning)演算法之一，已應用於多個領域的分類和回歸問題中，尤其在影像辨別上具有相當出色的表現，基本上由三種類型的層組成，分別為卷積層、池化層和全連接層如圖

7 所示。卷積層通過輸入數據和過濾器的用戶定義矩陣之間的卷積自動從輸入數據中提取特徵;池化層用以減少數據大小;全連接層執行分類與回歸最後階段的任務[44]。CNN 回歸模型通常使用類似於 ANN 回歸模型的基本公式，但其中加入卷積層和池化層的操作，其預測模型如式(8)所示：

$$Y_P = f(\omega_L * \sigma(P(\omega_{L-1} * \cdots \sigma(P(\omega_1 * X + b_1) \cdots) \cdots) + b_{L-1}) + b_L) \quad (8)$$

式中，Y_P是模型對於自變數(因子) X 的 PM2.5 濃度預測值，ω_L和b_L分別是第 L 層的權重和誤差項，σ是激勵函數，*表示卷積操作，P 表示池化操作。

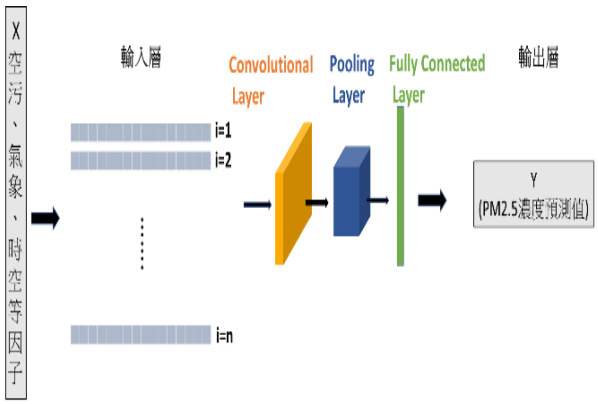

圖 7. 卷積神經網路示意圖

2.3 研究流程

本研究所採用計算軟體為 MATLAB R2022b 版本，電腦設備為宏碁(ASUSTeK COMPUTER INC.) EXPERTCENTER D700SC_M700MC 型號，作業系統為 Microsoft Windows 10 專業版。成果驗證區分為二個部分，首先於模型訓練階段，針對不同方法的訓練模型進行效能評估，訓練模型時採用 K 折交叉驗證(K-Fold Cross-Validation)，K 設定為 10，以 20%訓練資料進行模型驗證，在模型評估方面，採用本研究領域中常用的幾何驗證指標，評估模型的精度，分別為均方差(Mean Square Error, MSE)、均方根誤差(Root Mean Square Error, RMSE)、平均絕對誤差(Mean Absolute Error, MAE)、決定係數(Coefficient of determination,記為 R^2)及模型訓練時間(Time)；其次於模型測式階段，利用經過訓練好的 PM2.5 濃度模型推估西屯監測站於相應時刻的 PM2.5 濃度估計值，並與實際觀測值進行比較，以評估機器學習模型並驗證模型對獨立測試數據集的泛化能力。泛化能力是評估機器學習模型性能的重要指標之一，是機器學習模型在面對未曾見過數據時的表現能力，一個具有良好泛化能力的模型能夠在訓練數據之外的新數據上同樣有出色表現。研究流程如圖 8 所示。相關驗證指標計算方式說明如式(9)-(12)所示：

$$MSE = \frac{\sum_{i=1}^{n}(y_{tru,i} - y_{p,i})^2}{n} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{tru,i} - y_{p,i})^2}{n}} \quad (10)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_{p,i} - y_{tru,i}|}{n} \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{tru,i} - \bar{y})^2}{\sum_{i=1}^{n}(y_{tru,i} - y_{p,i})^2} \quad (12)$$

式中，$y_{tru,i}$為 PM2.5 濃度第 i 筆實際觀測值，$y_{p,i}$為 PM2.5 濃度第 i 筆估算值，n為觀測量數，$\bar{y}$為 PM2.5 濃度實際觀測值之平均值。

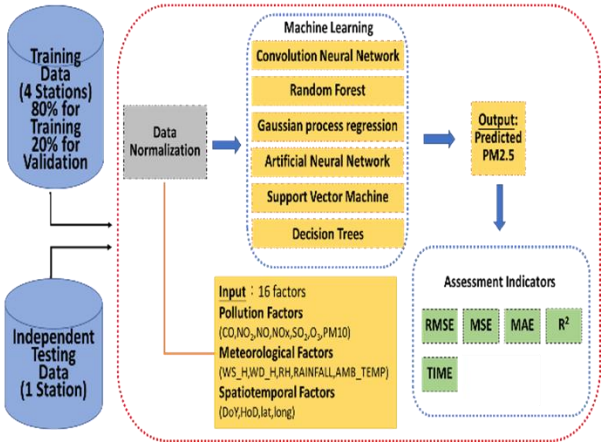

圖 8. 研究流程圖

三、成果分析

3.1 模型訓練

本章節比較了 6 種方法在 M16(包含時空因子)、M14(包含時間因子)及 M12(不包含時空因子)等 3 種不同自變數情況下的 PM2.5 濃度建模效能，研究成果分為二個部分，首先為

模型訓練，以 35,136 筆 PM2.5 濃度預測值和實際觀測值計算評估指標，相關指標成果如表 2 與圖 9 所示。結果表明，高斯過程回歸在 3 種不同自變數情境下於 RMSE、MSE、MAE 及 R^2 皆獲得最佳的訓練成果，各指標分別介於 1.67～2.01 μg/m^3、2.78～4.05 (μg/m^3)^2、0.97～1.33 μg/m^3 及 0.96～0.97，其次為隨機森林(RF)，而卷積神經網路(CNN)的成果相對較差;在訓練時間方面，決策樹(DT)的訓練時間最快，介於 14~19 秒，其次為隨機森林(RF)，介於 59~90 秒，高斯過程回歸速度最慢，介於 1,867~2,460 秒。

在此階段高斯過程回歸(GPR)的模型訓練精度最高，但相對地所耗費的時間同時也最長；決策樹(DT)與隨機森林(RF)二者模型訓練精度相近，惟隨機森林(RF)精度略優，訓練速度則較慢。就 M16、M14 及 M12 自變數模型分析如表 3 與圖 10 所示，6 種機器學習方法大都呈現一個趨勢，即加入時空因子後模型精度隨之提升，其中模型加入時間與空間因子(M16)及僅加入時間因子(M14)二者差異不大，而與不加入時空因子模型(M12)相較差異較大，由此可知時空因子對於模型訊精度的提升有顯著的效果，而當已加入時間因子後，加入空間因子對於精度的提升效果降低。時空因子對於高斯過程回歸的精度影響最為顯著。

3.2 模型測試

獨立測試以西屯站的 8,784 筆觀測量與由模型估算所得預測值進一步評估建模能力，相關指標成果如表 4 與圖 11 所示，表 4 顯示各模型與不同自變數模型情境在西屯測試站的測試結果，測試結果與訓練結果不同，隨機森林(RF)在 3 種不同自變數情境下於 RMSE、MSE、MAE 及 R^2 均獲得最佳的測試成果，各指標分別介於 4.10～4.25 μg/m^3、16.81～18.09 (μg/m^3)^2、3.11～3.24 μg/m^3 及 0.83～0.85，次之為高斯過程回歸(GPR)，而訓練階段建模效能與隨機森林(RF)相當的決策樹(DT)，其測試結果相對最差。

就測試面向觀察 M16、M14 及 M12 自變數模型的效能，6 種機器學習方法除卷積神經網路(CNN)外大都呈現一個結果，即模型在時空因子中僅加入時間因子的 M14 模型顯示出有較佳的測試成果，接續為加入時空因子的 M16 模型，而不加入時空因子的 M12 模型表現相對較差，此結果與模型訓練階段不同，亦及模型精度並非隨自變數增加而提高，顯然，時間因子相對於空間因子對於模型的影響程度較高。源自於深度學習的卷積神經網路(CNN)成果明顯不同於傳統機器學習的方法，該神經網路在圖像的特徵萃取與識別領域有很好的效能，在本研究中雖亦能有效的完成數據建模，惟表現不如其他傳統機器學習方法來的好。

為了解引入時空因子後對於成果的影響程度，分別計算 M16 模型、M14 模型及 M12 模型間的精度提升率，亦即當加入時間因子及再加入空間因子後的模型測試效果如表 5 與圖 12 所示。顯而易見，與不加入時空因子的 M12 模型相較，當加入時空因子或加入時間因子時均能提升建模的精度，然而對比前述兩種模型(M16,M14)發現，當將空間因子移除時模型的精度隨之提升，此一現像反映出，時間因子在進行 PM2.5 濃度建模時扮演相當重要角色，而空間因子在已有氣象、空污及時間因子下，對於模型的貢獻度不大，甚至有降低模型精度的情形發生。

就測試結果顯示，隨機森林(RF)在不同的自變數模型均具有最佳成果表現，為分析與其它方法的建模精度差異程度，以 RMSE 與 MAE 為指標計算其精度提升率(意旨隨機森林(RF)相較於其它方法在精度方面的提升程度)如表 6 與圖 13 所示，M16、M14 及 M12 在 RMSE 方面精度提升率分別介於 1.90%～22.41%、0.24%～22.50%與 1.39%～21.15%；而在 MAE 部分，精度提升率分別介於 4.66%～22.47%、1.29%～22.28%與 1.85%～22.06%。

經本研究綜合分析各項評估指標，顯示在演算法方面，隨機森林(RF)與其於方法相比表現出相對穩健的建模效能，而在時空因子部分，僅加入時間因子對於模型精度提升相較於同時加入時空因子較為顯著。

表 2. 模型訓練成果比較表

Method	RMSE(μg/m^3)			MSE(μg/m^3)^2			MAE(μg/m^3)			R^2			Training Time(s)		
	M16	M14	M12	M16	M14	M12	M16	M14	M12	M16	M14	M12	M16	M14	M12
RF	2.81	2.84	3.03	7.88	8.04	9.17	1.96	1.99	2.15	0.92	0.92	0.91	90	59	64

GPR	1.67	1.73	2.01	2.78	2.98	4.05	0.97	1.04	1.33	0.97	0.97	0.96	2460	2237	1867
ANN	3.43	3.57	3.73	11.74	12.76	13.93	2.58	2.68	2.80	0.89	0.88	0.86	968	293	840
SVM	3.57	3.69	3.86	12.76	13.59	14.88	2.58	2.67	2.80	0.87	0.86	0.85	465	640	404
DT	2.92	2.97	3.15	8.55	8.83	9.92	1.97	2.00	2.16	0.92	0.92	0.91	14	18	19
CNN	4.10	4.40	4.30	16.85	19.32	18.54	2.97	3.22	3.15	0.83	0.82	0.80	579	597	570

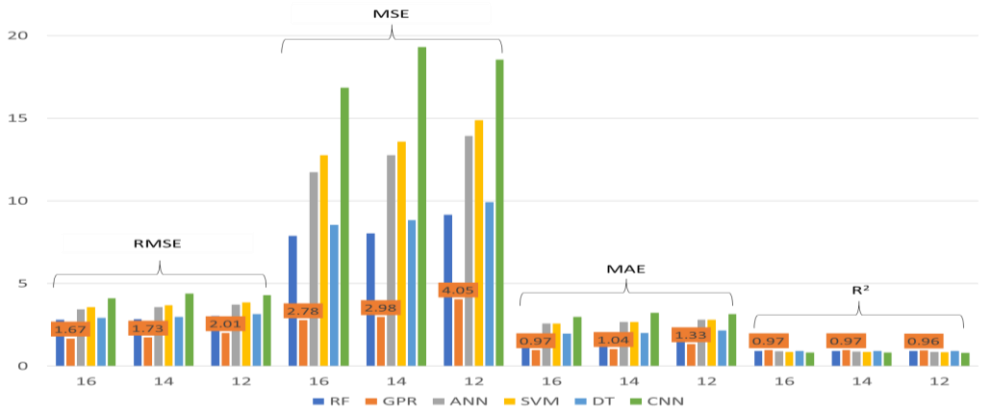

圖 9. 模型訓練成果比較圖

表 3. 不同自變數模型訓練精度提升率成果比較表

Method	**RMSE(μg/m³)**			**MAE(μg/m³)**		
	Improvement Rate (%)					
	M14-16	M12-14	M12-16	M14-16	M12-14	M12-16
RF	1.06	6.27	7.26	1.51	7.44	8.84
GPR	3.47	13.93	16.92	6.73	21.80	27.07
ANN	3.92	4.29	8.04	3.73	4.29	7.86
SVM	3.25	4.40	7.51	3.37	4.64	7.86
DT	1.68	5.71	7.30	1.50	7.41	8.80
CNN	6.82	-2.33	4.65	7.76	-2.22	5.71

Improvement Rate (%) = $(M_i - M_j)/M_i \times 100$

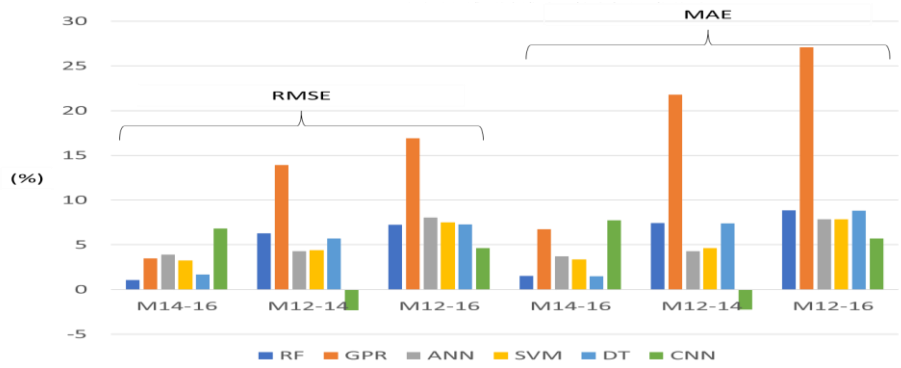

圖 10. 不同自變數模型訓練成果比較圖

表 4. 模型測試成果比較表

Method	**RMSE(μg/m³)**			**MSE(μg/m³)²**			**MAE(μg/m³)**			**R²**		
	M16	M14	M12	M16	M14	M12	M16	M14	M12	M16	M14	M12
RF	4.12	4.10	4.25	17.01	16.81	18.09	3.07	3.07	3.18	0.84	0.85	0.83

GPR	4.20	4.11	4.31	17.67	16.93	18.57	3.22	3.11	3.24	0.84	0.85	0.83
ANN	4.81	4.23	4.44	23.11	17.89	19.7	3.78	3.21	3.35	0.81	0.84	0.81
SVM	4.32	4.23	4.40	18.64	17.87	19.38	3.29	3.16	3.27	0.83	0.83	0.81
DT	5.31	5.29	5.39	28.25	28.03	29.08	3.96	3.95	4.08	0.77	0.77	0.75
CNN	4.68	4.99	4.69	21.93	24.92	22.04	3.49	3.65	3.51	0.80	0.79	0.78

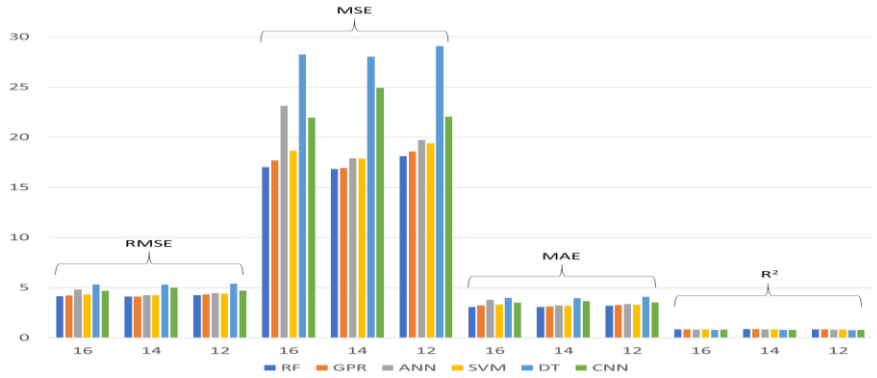

圖 11. 模型測試成果比較圖

表 5. 不同自變數模型測試精度提升率成果比較表

Method	RMSE(μg/m³)			MAE(μg/m³)		
	Improvement Rate (%)					
	M14-16	M12-14	M12-16	M14-16	M12-14	M12-16
RF	-0.49	3.53	3.06	0.00	3.46	3.46
GPR	-2.19	4.64	2.55	-3.54	4.01	0.62
ANN	-13.71	4.73	-8.33	-17.76	4.18	-12.84
SVM	-2.13	3.86	1.82	-4.11	3.36	-0.61
DT	-0.38	1.86	1.48	-0.25	3.19	2.94
CNN	6.21	-6.40	0.21	4.38	-3.99	0.57
Improvement Rate (%) = $(M_i - M_j)/M_i \times 100$						

圖 12. 不同自變數模型測試成果精度提升率比較圖

表 6. RF 模型測試成果相較於其它模型精度提升率比較表(西屯)

Method	RMSE(μg/m³)			MAE(μg/m³)		
	RF					
	Improvement Rate (%)					
	M16	M14	M12	M16	M14	M12

GPR	1.90	0.24	1.39	4.66	1.29	1.85
ANN	14.35	3.07	4.28	18.78	4.36	5.07
SVM	4.63	3.07	3.41	6.69	2.85	2.75
DT	22.41	22.50	21.15	22.47	22.28	22.06
CNN	11.97	17.84	9.38	12.03	15.89	9.40

Improvement Rate (%) = (Method − RF)/Method×100

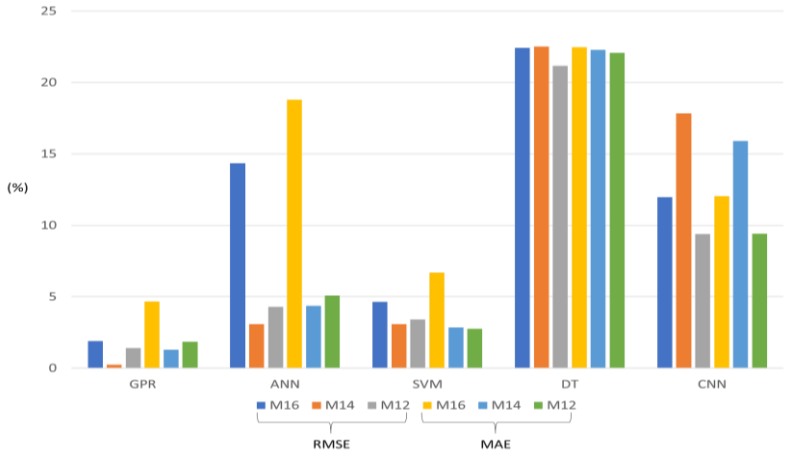

圖 13. RF 模型測試成果相較於其它模型精度提升率示意圖

四、結論

本研究目的為獲得台中地區適當之 PM2.5 濃度建模方式，並置重點於機器學習演算法之選擇，以及在已有空氣汙染與氣象因子的條件下，引入時間與空間因子對於模型精度的影響，爰此評估隨機森林(RF)、卷積神經網路(CNN)、高斯過程回歸(GPR)、人工神經網路(ANN)、支持向量機(SVM)及決策樹(DT)等 6 種廣泛應用於迴歸分析的機器學習方法對台中市 PM2.5 濃度建模的效能，同時分析在模型中引入時空因子的效果。研究中除評估模型的訓練精度外，並以西屯監測站的觀測數據做為獨立測試資料，希冀獲取台中市較適的建模方式，提供對於空氣汙染、數據分析、模型推估及機器學習等領域有興趣的人員進行參考，並可依此做更進一步的研究，依據本研究成果，得出以隨機森林(RF)為方法並導入時間因子具有最佳的建模效能，可綜整以下幾點結論：

(1)時間與空間因子的選用將影響模型的精度，本研究中發現時間因子對於模型精度的提升相當顯著，反映出能輔助氣象與空氣汙染因子對於 PM2.5 濃度隨著時間而變化(日夜與季節性)描述的不足。

(2)PM2.5 濃度在不同的地點呈現空間上的變異性，這可能是由於源頭排放、地形、建築物密度與交通流量等因素所導致。因此，本研究中引入經度與緯度作為空間因子，然而該因子對於模型精度的提升並不顯著，甚至在已加入時間因子的模型造成精度衰退的現象，顯示出空氣汙染、氣象及時間因子足以描述相應地區的 PM2.5 濃度，額外引入空間因子反而造成不良約制效果。

(3)本研究所採用 6 種機器學習方法應用於台中市的 PM2.5 濃度建模，均有不錯的建模能力，R^2 大多高於 0.8，RMSE 大都小於 5 $\mu g/m^3$。

(4)高斯過程回歸(GPR)對於模型訓練有較好的精度，測試精度則僅次於隨機森林(RF)，在精度表現上良好，惟所需的訓練時間與其他方法相較倍數增加，為其缺陷。

(5)隨機森林(RF)於本研究中不論於訓練或測試階段皆表現出優異的建模效能，提供穩健的建模結果，可為對空氣污染、數據分析、模型估計和機器學習感興趣的研究人員提供參考。

(6)卷積神經網路(CNN)為深度學習之一，已廣泛應用於影像分類領域並展現出優異的能力，本研究取其出色特徵萃取能力進行

PM2.5 濃度模型回歸，惟成果顯示建模效能相較其於方法不佳，因此對於此方法單獨運用於回歸問題上仍須進一步探討，亦或可採 Hybrid 方式結合其他方法。

(7)研究數據使用地面監測站所得空氣汙染與氣象資料為主，未來將加入氣溶膠光學厚度(Aerosol Optical Depth, AOD)及臺灣氣候變遷推估資訊與調適知識平台(Taiwan Climate Change Projection Information and Adaptation Knowledge Platform, TCCIP)所提供高解析氣象資料，並研究整合不同方法的 Hybrid 模型，希藉此提升 PM2.5 濃度模型的建模效能。

(8)本研究主要貢獻為基於機器學習方法，分析 6 種常見方法，在台中地區進行 PM2.5 濃度建模的精度、速度及導入時間與空間因子對模型的影響，最終得出隨機森林(RF)建模效能最佳，在時空因子中，僅額外導入時間因子，對模型精度提升最為顯著，此一建模方法與研究中對於各方法的豐富分析成果，對於區域性的建模提供實用的參考價值。

參考文獻

[1] W. C. Hinds, and Y. Zhu, *Aerosol technology: properties, behavior, and measurement of airborne particles*: John Wiley & Sons, 2022.

[2] Y.-C. Hong, J.-T. Lee, H. Kim, and H.-J. Kwon, "Air pollution: a new risk factor in ischemic stroke mortality," *Stroke,* vol. 33, no. 9, pp. 2165-2169, 2002.

[3] W. Q. Gan, M. Koehoorn, H. W. Davies, P. A. Demers, L. Tamburic, and M. Brauer, "Long-term exposure to traffic-related air pollution and the risk of coronary heart disease hospitalization and mortality," *Environmental health perspectives,* vol. 119, no. 4, pp. 501-507, 2011.

[4] Z. J. Andersen, M. Hvidberg, S. S. Jensen, M. Ketzel, S. Loft, M. Sørensen, A. Tjønneland, K. Overvad, and O. Raaschou-Nielsen, "Chronic obstructive pulmonary disease and long-term exposure to traffic-related air pollution: a cohort study," *American journal of respiratory and critical care medicine,* vol. 183, no. 4, pp. 455-461, 2011.

[5] O. Raaschou-Nielsen, Z. J. Andersen, R. Beelen, E. Samoli, M. Stafoggia, G. Weinmayr, B. Hoffmann, P. Fischer, M. J. Nieuwenhuijsen, and B. Brunekreef, "Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE)," *The lancet oncology,* vol. 14, no. 9, pp. 813-822, 2013.

[6] L. A. Darrow, M. Klein, W. D. Flanders, J. A. Mulholland, P. E. Tolbert, and M. J. Strickland, "Air pollution and acute respiratory infections among children 0–4 years of age: an 18-year time-series study," *American journal of epidemiology,* vol. 180, no. 10, pp. 968-977, 2014.

[7] P. V. T. Anh, "Control Air Pollution to The Sustainable Development Goals Vietnam Perspective," *Administrative and Environmental Law Review,* vol. 4, no. 1, pp. 49-64, 2023.

[8] I. Djalalova, L. Delle Monache, and J. Wilczak, "PM2. 5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model," *Atmospheric Environment,* vol. 108, pp. 76-87, 2015.

[9] G. Geng, Q. Zhang, R. V. Martin, A. van Donkelaar, H. Huo, H. Che, J. Lin, and K. He, "Estimating long-term PM2. 5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model," *Remote sensing of Environment,* vol. 166, pp. 262-270, 2015.

[10] P. E. Saide, G. R. Carmichael, S. N. Spak, L. Gallardo, A. E. Osses, M. A. Mena-Carrasco, and M. Pagowski, "Forecasting urban PM10 and PM2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model," *Atmospheric Environment,* vol. 45, no. 16, pp. 2769-2780, 2011.

[11] F. Mao, J. Hong, Q. Min, W. Gong, L. Zang, and J. Yin, "Estimating hourly full-coverage PM2. 5 over China based on TOA reflectance data from the Fengyun-4A satellite," *Environmental Pollution,* vol. 270, pp. 116119, 2021.

[12] J. Wei, Z. Li, A. Lyapustin, L. Sun, Y. Peng, W. Xue, T. Su, and M. Cribb, "Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications," *Remote Sensing of Environment,* vol. 252, pp. 112136, 2021.

[13] Y. Liu, K. He, S. Li, Z. Wang, D. C.

Christiani, and P. Koutrakis, “A statistical model to evaluate the effectiveness of PM2. 5 emissions control during the Beijing 2008 Olympic Games,” *Environment international,* vol. 44, pp. 100-105, 2012.

[14] C.-D. Wu, Y.-T. Zeng, and S.-C. C. Lung, “A hybrid kriging/land-use regression model to assess PM2. 5 spatial-temporal variability,” *Science of the Total Environment,* vol. 645, pp. 1456-1464, 2018.

[15] P. D. Sampson, M. Richards, A. A. Szpiro, S. Bergen, L. Sheppard, T. V. Larson, and J. D. Kaufman, “A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2. 5 concentrations in epidemiology,” *Atmospheric environment,* vol. 75, pp. 383-392, 2013.

[16] Y. Liu, G. Cao, N. Zhao, K. Mulligan, and X. Ye, “Improve ground-level PM2. 5 concentration mapping using a random forests-based geostatistical approach,” *Environmental Pollution,* vol. 235, pp. 272-282, 2018.

[17]K. Masroor, F. Fanaei, S. Yousefi, M. Raeesi, H. Abbaslou, A. Shahsavani, and M. Hadei, “Spatial modelling of PM2. 5 concentrations in Tehran using Kriging and inverse distance weighting (IDW) methods,” *Journal of Air Pollution and Health,* vol. 5, no. 2, pp. 89-96, 2020.

[18] S. Kumar, S. Mishra, and S. K. Singh, “A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere,” *Heliyon,* vol. 6, no. 11, pp. e05618, 2020.

[19] K. Harishkumar, K. Yogesh, and I. Gad, “Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models,” *Procedia Computer Science,* vol. 171, pp. 2057-2066, 2020.

[20] Q. Di, H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, M. B. Sabath, C. Choirat, P. Koutrakis, and A. Lyapustin, “An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution,” *Environment international,* vol. 130, pp. 104909, 2019.

[21] S. Kumar, S. Mishra, and S. Singh, "A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere, Heliyon, 6, e05618," 2020.

[22] T. Zheng, M. H. Bergin, R. Sutaria, S. N. Tripathi, R. Caldow, and D. E. Carlson, “Gaussian process regression model for dynamically calibrating a wireless low-cost particulate matter sensor network in Delhi,” *Atmos. Meas. Tech. Dis*, 2019.

[23] H. Jafarian, and S. Behzadi, “Evaluation of PM2. 5 emissions in Tehran by means of remote sensing and regression models,” *Pollution,* vol. 6, no. 3, pp. 521-529, 2020.

[24] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, “Modeling PM 2.5 urban pollution using machine learning and selected meteorological parameters,” *Journal of Electrical and Computer Engineering,* vol. 2017, 2017.

[25] Y. Zhou, F.-J. Chang, L.-C. Chang, I.-F. Kao, Y.-S. Wang, and C.-C. Kang, “Multi-output support vector machine for regional multi-step-ahead PM2. 5 forecasting,” *Science of the Total Environment,* vol. 651, pp. 230-240, 2019.

[26] G. Goudarzi, P. K. Hopke, and M. Yazdani, “Forecasting PM2. 5 concentration using artificial neural network and its health effects in Ahvaz, Iran,” *Chemosphere,* vol. 283, pp. 131285, 2021.

[27] S.-M. Choi, and H. Choi, “Artificial Neural Network Modeling on PM10, PM2. 5, and NO2 Concentrations between Two Megacities without a Lockdown in Korea, for the COVID-19 Pandemic Period of 2020,” *International Journal of Environmental Research and Public Health,* vol. 19, no. 23, pp. 16338, 2022.

[28] S. Chae, J. Shin, S. Kwon, S. Lee, S. Kang, and D. Lee, “PM10 and PM2. 5 real-time prediction models using an interpolated convolutional neural network,” *Scientific Reports,* vol. 11, no. 1, pp. 11952, 2021.

[29] J. Li, M. Jin, and H. Li, “Exploring spatial influence of remotely sensed PM2. 5 concentration using a developed deep convolutional neural network model,” *International Journal of Environmental Research and Public Health,* vol. 16, no. 3, pp. 454, 2019.

[30] B. Liu, S. Yan, J. Li, Y. Li, J. Lang, and G. Qu, “A spatiotemporal recurrent neural network for prediction of atmospheric PM2. 5: a case study of Beijing,” *IEEE Transactions on Computational Social Systems,* vol. 8, no. 3, pp. 578-588, 2021.

[31] X. Dai, J. Liu, and Y. Li, “A recurrent neural network using historical data to predict time series indoor PM2. 5 concentrations for

residential buildings," *Indoor air,* vol. 31, no. 4, pp. 1228-1237, 2021.

[32] X. Gao, and W. Li, "A graph-based LSTM model for PM2. 5 forecasting," *Atmospheric Pollution Research,* vol. 12, no. 9, pp. 101150, 2021.

[33] K. Qadeer, W. U. Rehman, A. M. Sheri, I. Park, H. K. Kim, and M. Jeon, "A long short-term memory (LSTM) network for hourly estimation of PM2. 5 concentration in two cities of South Korea," *Applied Sciences,* vol. 10, no. 11, pp. 3984, 2020.

[34] B. Zhang, H. Zhang, G. Zhao, and J. Lian, "Constructing a PM2. 5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks," *Environmental Modelling & Software,* vol. 124, pp. 104600, 2020.

[35] P.-Y. Wong, H.-Y. Lee, Y.-C. Chen, Y.-T. Zeng, Y.-R. Chern, N.-T. Chen, S.-C. C. Lung, H.-J. Su, and C.-D. Wu, "Using a land use regression model with machine learning to estimate ground level PM2. 5," *Environmental Pollution,* vol. 277, pp. 116846, 2021.

[36] X. Wang, W. Sun, K. Zheng, X. Ren, and P. Han, "Estimating hourly PM2. 5 concentrations using MODIS 3 km AOD and an improved spatiotemporal model over Beijing-Tianjin-Hebei, China," *Atmospheric Environment,* vol. 222, pp. 117089, 2020.

[37] J. Shen, D. Valagolam, and S. McCalla, "Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2. 5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea," *PeerJ,* vol. 8, pp. e9961, 2020.

[38] X. Li, and X. Zhang, "Predicting ground-level PM2. 5 concentrations in the Beijing-Tianjin-Hebei region: a hybrid remote sensing and machine learning approach," *Environmental pollution,* vol. 249, pp. 735-749, 2019.

[39] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine,* vol. 47, no. 1, pp. 31-39, 2017.

[40] C. Williams, and C. Rasmussen, "Gaussian processes for regression," *Advances in neural information processing systems,* vol. 8, 1995.

[41] J. Zupan, "Introduction to artificial neural network (ANN) methods: what they are and how to use them," *Acta Chimica Slovenica,* vol. 41, pp. 327-327, 1994.

[42] C. Cortes, and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[43] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, pp. 81-106, 1986.

[44] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," *ISPRS journal of photogrammetry and remote sensing,* vol. 173, pp. 24-49, 2021