toxics

MDPI

Article

Towards Cleaner Cities: Estimating Vehicle-Induced $PM_{2.5}$ with Hybrid EBM-CMA-ES Modeling

Saleh Alotaibi [1,*], Hamad Almujibah [2], Khalaf Alla Adam Mohamed [3], Adil A. M. Elhassan [2], Badr T. Alsulami [4], Abdullah Alsaluli [2] and Afaq Khattak [5,*]

1 Civil and Environmental Engineering Department, Faculty of Engineering—Rabigh Branch, King Abdulaziz University, Jeddah 21589, Saudi Arabia
2 Department of Civil Engineering, College of Engineering, Taif University, Taif 21944, Saudi Arabia; hmujibah@tu.edu.sa (H.A.); aahassan@tu.edu.sa (A.A.M.E.); amalsaluli@tu.edu.sa (A.A.)
3 Department of Civil Engineering, College of Engineering, Bisha University, Bisha 61361, Saudi Arabia; kaamohamed@ub.edu.sa
4 Department of Civil Engineering, College of Engineering and Architecture, Umm Al-Qura University, Makkah 24382, Saudi Arabia; btsulami@uqu.edu.sa
5 Department of Civil, Structural and Environmental Engineering, Trinity College Dublin, D02 PN40 Dublin, Ireland
* Correspondence: salnufiae@kau.edu.sa (S.A.); akhattak@tcd.ie (A.K.)

Abstract: In developing countries, vehicle emissions are a major source of atmospheric pollution, worsened by aging vehicle fleets and less stringent emissions regulations. This results in elevated levels of particulate matter, contributing to the degradation of urban air quality and increasing concerns over the broader effects of atmospheric emissions on human health. This study proposes a Hybrid Explainable Boosting Machine (EBM) framework, optimized using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), to predict vehicle-related $PM_{2.5}$ concentrations and analyze contributing factors. Air quality data were collected from Open-Seneca sensors installed along the Nairobi Expressway, alongside meteorological and traffic data. The CMA-ES-tuned EBM model achieved a Mean Absolute Error (MAE) of 2.033 and an R^2 of 0.843, outperforming other models. A key strength of the EBM is its interpretability, revealing that the location was the most critical factor influencing $PM_{2.5}$ concentrations, followed by humidity and temperature. Elevated $PM_{2.5}$ levels were observed near the Westlands roundabout, and medium to high humidity correlated with higher $PM_{2.5}$ levels. Furthermore, the interaction between humidity and traffic volume played a significant role in determining $PM_{2.5}$ concentrations. By combining CMA-ES for hyperparameter optimization and EBM for prediction and interpretation, this study provides both high predictive accuracy and valuable insights into the environmental drivers of urban air pollution, providing practical guidance for air quality management.

Keywords: air quality; $PM_{2.5}$; explainable boosting machine; covariance matrix adaptation evolution strategy

check for updates

Citation: Alotaibi, S.; Almujibah, H.; Mohamed, K.A.A.; Elhassan, A.A.M.; Alsulami, B.T.; Alsaluli, A.; Khattak, A. Towards Cleaner Cities: Estimating Vehicle-Induced $PM_{2.5}$ with Hybrid EBM-CMA-ES Modeling. *Toxics* **2024**, *12*, 827. https://doi.org/toxics12110827

Academic Editor: Matthias Karl

Received: 8 October 2024
Revised: 15 November 2024
Accepted: 17 November 2024
Published: 19 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Air pollution has become a pressing issue for both environmental sustainability and and public health in developing nations, intensifying in recent decades due to rapid industrial growth and urbanization [1,2]. Estimates show that nearly 7 million deaths each year can be traced to exposure to fine particulate matter with a diameter of less than 2.5 μm ($PM_{2.5}$). Moreover, approximately 91% of the world's population lives in areas where $PM_{2.5}$ concentrations surpass the acceptable limits of 10–20 $\mu g/m^3$ [3,4]. Both toxicological and epidemiological studies have established a strong association between $PM_{2.5}$ exposure and heightened risks of cardiovascular and respiratory ailments, alongside an increased incidence of premature mortality linked to prolonged exposure [5–8]. The

Global Burden of Disease (GBD) study ranks $PM_{2.5}$ as the fifth leading risk factor for global mortality, accounting for approximately 4.2 million premature deaths annually [9].

Many rapidly growing nations, such as Saudi Arabia, India, and China, face significant challenges in managing air quality [10]. With high levels of transportation-related pollution, industrial emissions, and energy use, these countries are taking steps to combat rising air pollution as part of broader sustainability initiatives. Urban centers like Riyadh, Delhi, and Beijing see elevated $PM_{2.5}$ levels due to population growth and a heavy reliance on gasoline-powered vehicles. To address these challenges, governments are promoting electric vehicles (EVs) and improving public transportation infrastructure. For example, Saudi Arabia has partnered with Lucid Motors to produce EVs domestically [11], while India and China are expanding charging infrastructure and offering EV subsidies. Investment in public transit, like Saudi Arabia's Riyadh Metro and metro expansions in India and China, aims to reduce vehicle emissions by decreasing private car reliance. These initiatives are expected to significantly lower air pollutants such as $PM_{2.5}$ and NO_x, contributing to healthier urban environments.

Rapid development across many African countries has led to significant urbanization and a sharp rise in vehicle usage, thereby intensifying energy demand [12]. This growth has profoundly impacted air quality, particularly concerning $PM_{2.5}$ levels resulting from vehicular emissions. Meteorological conditions are pivotal in intensifying $PM_{2.5}$ concentrations, as they influence the dispersion, dilution, and deposition of these fine particles [13–15]. Research indicates that unfavorable meteorological conditions can lead to elevated $PM_{2.5}$ levels, even when emissions are reduced, compared with scenarios with more favorable weather and higher emissions. Key factors such as fluctuations in humidity, wind speed, atmospheric pressure, and temperature play a critical role in shaping the spatial and temporal distribution of $PM_{2.5}$ [16,17]. Furthermore, earlier studies have demonstrated that vehicular emissions are a major contributor to $PM_{2.5}$ pollution in the environment. Specifically, vehicle-related $PM_{2.5}$ accounts for 39.8% of the total $PM_{2.5}$ concentration in Shanghai, 16% in New York, and 26% in Beijing [18]. However, there is a scarcity of research on vehicle-related $PM_{2.5}$ emissions in developing countries. Therefore, this study seeks to estimate the concentration of $PM_{2.5}$ emissions from vehicles and assess the impact of various traffic-related and environmental conditions on these emissions. We propose a new approach using the Explainable Boosting Machine (EBM) framework [19], with hyperparameter fine-tuning achieved via the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [20]. This is motivated by the following objectives:

- The EBM model was selected due to its strong predictive capabilities in forecasting $PM_{2.5.}$ Unlike black-box models, EBM maintains transparency and offers inherent interpretability, allowing stakeholders to understand the contributing factors [21–24].
- EBM is a Generalized Additive Model (GAM) that provides high interpretability by modeling feature effects independently. This aspect is crucial when assessing environmental risks such as $PM_{2.5}$, as it enables clear identification of how variables like location, humidity, and temperature contribute to $PM_{2.5}$ levels, providing actionable insights for policymakers and planners.
- To ensure optimal performance, the hyperparameters of EBM are fine-tuned using CMA-ES. CMA-ES is a robust evolutionary optimization algorithm known for efficiently navigating complex, high-dimensional search spaces [25]. Compared with traditional methods such as Grid Search or Random Search [26], CMA-ES is more adaptive and capable of handling non-linearities and interactions in the model, ensuring that EBM achieves its best possible performance on the $PM_{2.5}$ dataset.

This EBM-CMA-ES framework not only improves predictive accuracy but also preserves model transparency, making it well-suited for forecasting $PM_{2.5}$. It empowers both prediction and interpretability, ensuring a comprehensive understanding of $PM_{2.5}$ levels and their contributing factors. Figure 1 illustrates the proposed EBM-CMA-ES framework. The structure of this paper is organized as follows: Section 2 reviews the existing literature on statistical and machine learning models used for predicting $PM_{2.5}$ levels, focusing on

their strengths and limitations. Section 3 provides a detailed description of the study location and outlines the theoretical background of the methods employed, including the EBM framework and the CMA-ES for hyperparameter optimization. Section 4 evaluates and explains the model performance, conducts uncertainty analysis, and interprets the results with a focus on the inherent interpretability of EBM. Finally, Section 5 presents the conclusions and provides recommendations for future research and practical applications.

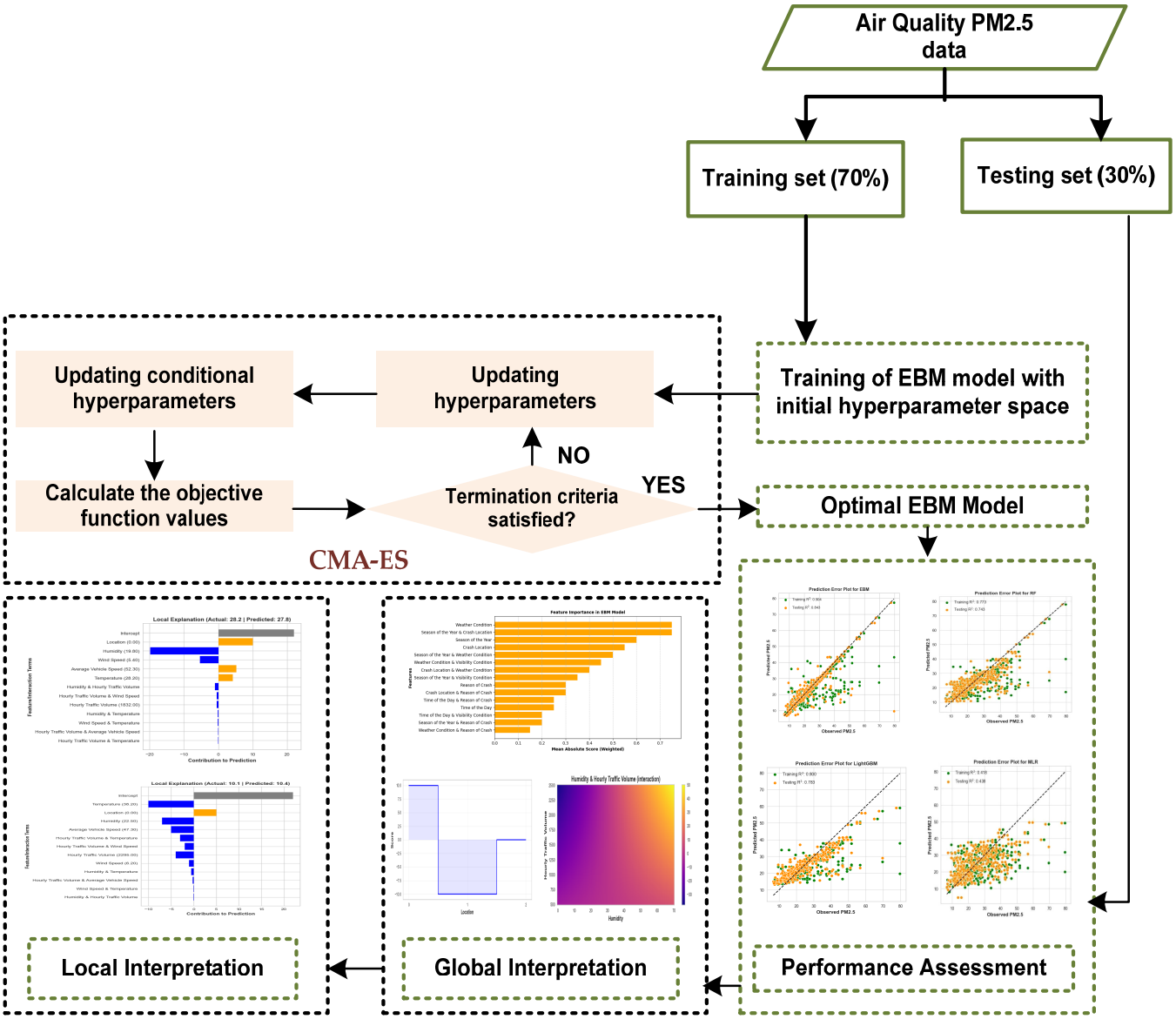


Figure 1. Proposed EBM-CMA-ES framework for the prediction and assessment of $PM_{2.5}$.

2. Related Work

$PM_{2.5}$ has emerged as a significant public health concern due to its adverse effects on human health [27,28]. Extensive research highlights the crucial role of environmental and atmospheric factors, such as temperature, humidity, and wind patterns, in shaping ambient $PM_{2.5}$ concentrations. To predict $PM_{2.5}$ levels, various statistical models have been employed, drawing on these meteorological variables. A notable example is a multiple linear regression (MLR) model designed to estimate daily $PM_{2.5}$ concentrations across different monitoring stations in the western United States. This model incorporates factors like prior day $PM_{2.5}$ levels, fire radiative power, and aerosol optical depth from satellite data, providing an effective tool for forecasting $PM_{2.5}$ fluctuations in response to changing

environmental conditions [29]. Another study utilized a Bayesian ensemble approach to create a method that integrates aerosol optical depth (AOD) data from satellites with chemical transport model (CTM) simulations, thereby improving $PM_{2.5}$ estimation [30]. Another study also applied MLR to forecast $PM_{2.5}$ levels by utilizing various risk factors, including maximum and minimum noise, temperature, and humidity [31]. Statistical models, particularly MLR, are widely utilized due to their interpretability and straightforwardness. However, they possess significant drawbacks. Traditional statistical models are based on stringent assumptions, such as linearity, normality, homoscedasticity, and independence of residuals. Violation of these assumptions can result in biased or inaccurate outcomes. Additionally, these models often struggle with complex nonlinear relationships and interactions that frequently occur in real-world datasets. This limitation arises from their inherent structure, which is constrained by predefined assumptions and lacks the flexibility needed to adapt to the multidimensional characteristics of practical data scenarios where variables may not conform neatly to theoretical expectations. As a result, when data exhibit intricate behaviors or interdependencies, conventional approaches may fail to yield reliable or robust insights. Such limitations hinder their effectiveness in capturing complex patterns compared with more adaptable methodologies. Furthermore, statistical models may encounter challenges with large datasets or high-dimensional data due to computational inefficiencies or risks of overfitting, particularly when not managed with precision [32,33].

In contrast, Artificial Intelligence (AI) and, in particular, machine learning models, is increasingly preferred for several reasons. Machine learning models are adept at managing nonlinear and complex relationships more effectively than traditional statistical approaches [34]. They can automatically detect interactions between variables without explicit specification. The machine learning models are designed to handle large-scale data and can be easily automated to learn from new data continuously, improving their predictions over time [35]. Researchers around the world are increasingly turning to machine learning models to predict $PM_{2.5}$ concentrations. For instance, a study conducted in northern Taiwan made use of the self-organizing map (SOM) technique. This approach clusters high-dimensional data into a comprehensible two-dimensional topological map, effectively highlighting the spatial and temporal distribution of $PM_{2.5}$ levels. This method enhanced the visualization and understanding of how $PM_{2.5}$ concentrations fluctuate over different locations and periods [36]. Another study employed a Random Forest (RF) model to estimate the $PM_{2.5}$ levels across China from 2005 to 2016. The proposed model significantly outperformed traditional statistical regression models in capturing spatial variability and reducing prediction errors at daily, monthly, and annual time scales [37]. Another similar study conducted in various regions of China employed an ensemble machine learning approach that combines Random Forest (RF), generalized additive models, and extreme Gradient Boosting (XGBoost) and demonstrated a strong $PM_{2.5}$ prediction accuracy [38]. In order to consider both machine learning and deep learning models, a study conducted in the Hunan province of China utilized XGBoost and a fully connected neural network (FCNN) to predict $PM_{2.5}$ concentrations using data from meteorological parameters and $PM_{2.5}$ measurements. It was observed that the XGBoost model outperformed the neural network in predicting $PM_{2.5}$ [39]. A study conducted in Malaysia employed RF and Support Vector Machine (SVM) to estimate $PM_{2.5}$ concentrations by combining satellite data, ground-measured pollutants, and meteorological factors. The RF model performed better than SVM in predicting $PM_{2.5}$ [40]. In addition, deep learning models have also been used to predict $PM_{2.5}$ concentration. A study employed a weighted long short-term memory extended model (WLSTME) to improve $PM_{2.5}$ prediction accuracy by considering site density and wind conditions. The WLSTME integrated neighbor site data, historical $PM_{2.5}$ concentrations, and meteorological data, outperforming previous methods [41]. Similarly, some other researchers employed deep convolutional neural networks (CNNs) to estimate $PM_{2.5}$ levels. This research involved generating a hallucinated reference image, computing discrepancy maps, and predicting $PM_{2.5}$ concentrations using extracted features [42]. An-

other study employed deep learning-based recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM), Bi-LSTM (Bidirectional LSTM), and Bidirectional Gated Recurrent Unit (Bi-GRU) models, alongside a CNN to predict $PM_{2.5}$ concentration using meteorological data from 2017 to 2019 from Taiwan [43].

To the best of our knowledge, no researcher has previously utilized the EBM in combination with the CMA-ES for predicting $PM_{2.5}$ concentrations. Therefore, this study adopts the EBM model to take advantage of its inherent interpretability and predictive capabilities. By employing CMA-ES for hyperparameter tuning, the EBM model performance is further enhanced, allowing it to achieve higher accuracy in $PM_{2.5}$ prediction. In addition, the inherent transparency of the EBM framework ensures that the model not only delivers accurate predictions but also provides clear, interpretable insights into the influence of various input features on $PM_{2.5}$ levels, providing a deeper understanding of the contributing risk factors.

3. Materials and Methods

3.1. Study Location and Data

The Nairobi Expressway serves as a critical transportation corridor, connecting Nairobi's urban center to Jomo Kenyatta International Airport (JKIA). This 27 km (17 miles) six-lane dual carriageway runs along the central reservations of Mombasa Road, beginning at Mlolongo, extending through Uhuru Highway, and concluding at James Gichuru Road, as illustrated in Figure 2. It is a vital route for commuters, particularly those traveling to and from the airport.

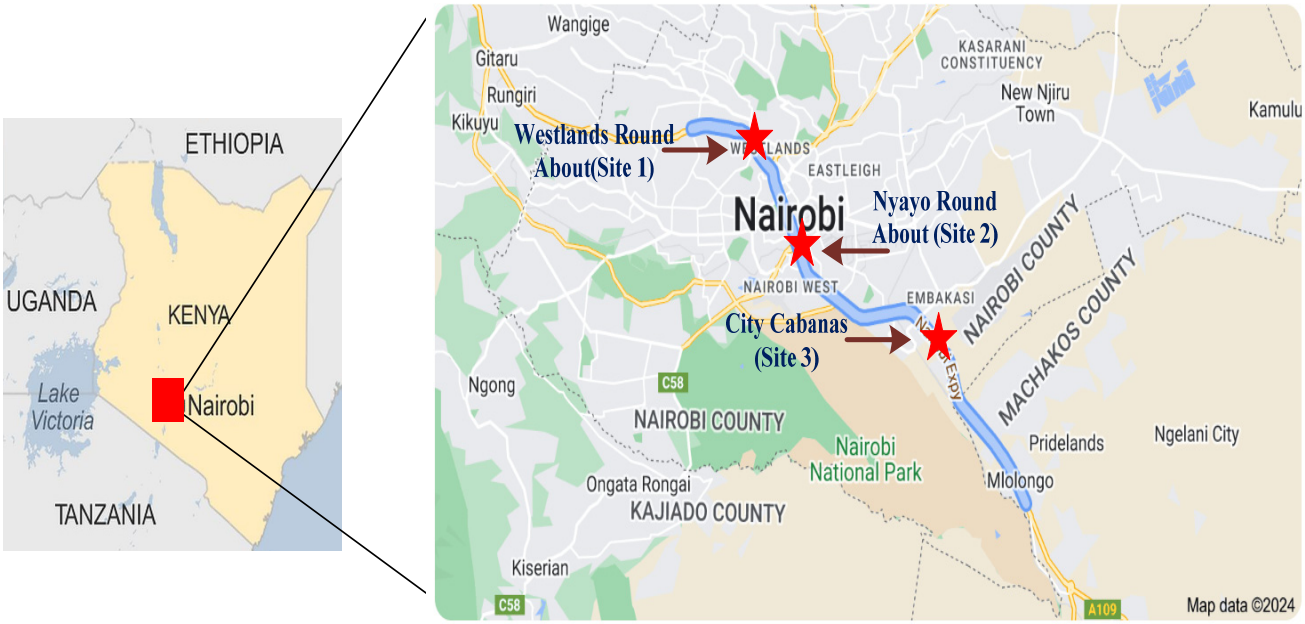


Figure 2. Sites for the data collection along Nairobi expressway.

For this study, data collection was conducted at three strategically chosen locations along the Nairobi Expressway corridor. Monitoring occurred for 12 h per day over seven consecutive days, focusing on peak hours when traffic flow and related emissions are most significant. To capture potential seasonal variations in $PM_{2.5}$ concentrations, data were collected during three distinct periods: 23–29 August 2021 (representing the dry season), 13–18 December 2021 (peak holiday season), and 21–27 March 2022 (post-holiday, also dry). August represents the dry season in Kenya, where $PM_{2.5}$ levels may be elevated due to reduced rainfall and increased dust resuspension. December coincides with the peak holiday season, likely increasing traffic volumes and vehicle-related emissions. March, as a dry month post holiday, allows for observations of typical daily traffic patterns and

ambient air quality outside peak travel periods. Table 1 provides detailed descriptions of these monitoring sites.

Table 1. Description and locations of sampling sites in Nairobi.

Sites	Description	Latitude	Longitude
Westlands roundabout (Site 1)	Located on Waiyaki Way, this is a three-lane highway in each direction adjacent to the Westlands roundabout. The area experiences a high proportion of personal vehicles and buses due to its central location and proximity to residential neighborhoods	−1.26551	36.80268
Nyayo roundabout (Site 2)	Situated in Bellevue, this is a three-lane highway in each direction. It is a busy urban route with a balanced mix of personal and commercial vehicles. Although congestion levels here are generally lower than at Westlands, it experiences similar types of traffic.	−1.31940	36.83854
City Cabanas (Site 3)	Positioned near the Airport North Road and Mombasa Road interchange, this site has a three-lane highway in each direction. Given its proximity to the airport and industrial areas, it sees a high volume of heavy vehicles, including goods transport and delivery trucks.	−1.33573	36.89217

Traffic volumes were systematically recorded across various vehicle categories, including motorcycles, passenger vehicles, buses, and goods vehicles of differing capacities (light, medium, heavy, and articulated trucks). The comprehensive datasets were compiled on ambient air pollutant concentrations using calibrated Open-Seneca sensors. The sensors were calibrated before deployment by cross-referencing with a reference-grade air quality monitoring station, involving both laboratory testing to establish baseline accuracy and field calibration to account for environmental variables such as temperature and humidity. This calibration process ensured that $PM_{2.5}$ measurements were accurate and consistent across the three monitoring locations. Along with pollutant data, hourly traffic volume, as well as average vehicle speeds and meteorological parameters, including humidity, wind speed, and temperature, were recorded to provide a holistic view of factors influencing air quality along the Nairobi Expressway.

3.2. Hybrid EBM-CMA-ES Framework

3.2.1. Theoretical Overview of EBM

The EBM is an advanced machine learning model designed to balance high predictive accuracy with interpretability. It combines the principles of the Generalized Additive Model (GAM) and boosting algorithms to create a model that is both powerful and interpretable. This makes EBM particularly useful in domains where understanding the rationale behind predictions is crucial. The GAM represents the outcome as an additive function of the predictors, allowing for straightforward interpretation. For any m^{th} data point within a data, the general form of a GAM is given in Equation (1).

$$y = \varphi_0 + \Sigma\varphi_m(x_m) + \varepsilon \quad (1)$$

where the following hold:

- y is the predicted outcome.
- φ_0 is the intercept.
- $\varphi_m(x_m)$ are the shape functions for each feature x_m.
- ε is the error term.

Each shape function $\varphi_m(x_m)$ captures the relationship between the feature x_m and the outcome y. This allows us to visualize and understand the effect of individual features on the prediction. The boosting is an ensemble technique that enhances model performance by combining multiple weak learners. The process involves training weak learners sequentially, where each new learner focuses on correcting the errors of the previous ones. The objective is to minimize a loss function, typically using gradient descent. The process EBM

modeling begins with an initial prediction, often the mean of the target variable as given by Equation (2)

$$\hat{y}^{(0)} = \overline{y} \tag{2}$$

where the following hold:

- $\hat{y}^{(0)}$ is the initial prediction.
- $\overline{y}$ is the mean of the target variable.

The model iteratively refines the predictions by learning shape functions for each individual feature. Each iteration t involves computing residuals, fitting weak learners, updating shape function, and updating predictions. The residuals are calculated as the difference between the actual target values and the current predictions as shown in Equation (3).

$$r^{(t)} = y - \hat{y}^{(t-1)} \tag{3}$$

where the following hold:

- $r^{(t)}$ is the residual at iteration t.
- y is the actual target value.
- $\hat{y}^{(t-1)}$ is the prediction from the previous iteration.

For each feature x_i, fit a weak learner (e.g., a decision tree) to the residuals. This step focuses on learning the shape function that $\varphi_m(x_m)$ best explains the residuals for that feature as given by Equation (4).

$$\varphi_m^{(t)}(x_m) \leftarrow \text{FitWeakLearner}\left(r^{(t)}, x_m\right) \tag{4}$$

It is then followed by updating the shape functions for each individual feature by adding the contribution from the current iteration, scaled by a learning rate η as given by Equation (5).

$$\varphi_m(x_m) \leftarrow \varphi_m(x_m) + \eta \cdot \varphi_m^{(t)}(x_m) \tag{5}$$

where the following hold:

- $\varphi_m(x_m)$ is the updated shape function for feature x_m.
- η is the learning rate, controlling the step size of the update.

It also includes two-dimensional interactions between the features. The two-dimensional interactions can be rendered as heat-maps on a two-dimensional plane, the model that includes two-dimensional interaction is also interpretable. Thus, the overall prediction may be updated by adding the contributions from the updated shape functions and two-dimensional interaction, as given by Equation (6).

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \sum_{m=`}^{(M)} f_m^{(t)}(x_m) + \sum_{m=1}^{(M)} \sum_{n=1`}^{(N)} f_{m,n}^{(t)}(x_m, x_n) \tag{6}$$

where the following hold:

- $\hat{y}^{(t)}$ is the updated prediction at iteration t.
- $\sum_{m=`}^{(M)} f_m^{(t)}(x_m)$ captures the contributions from each individual feature.
- $\sum_{m=1}^{(M)} \sum_{n=1`}^{(N)} f_{m,n}^{(t)}(x_m, x_n)$ captures the contributions from interactions between pairs of features.

The iterative process continues until a stopping criterion is met. Common stopping criteria include (1) maximum number of iterations T; (2) convergence of the loss function, i.e., when changes in the loss function fall below a certain threshold. The iteration process stops when $t \geq T$ or $\Delta\text{Loss} < \varepsilon$, where $t \geq T$ is the maximum number of iteration, and ε is the threshold for convergence.

3.2.2. Interpretation of EBM

One of the key interpretability features of EBM is the shape functions $\varphi_m(x_m)$. Each shape function represents the relationship between a feature and the target variable. These functions can be visualized to understand how changes in a feature affect the prediction. For instance, if $\varphi_m(x_m)$ is a straight line, it indicates a linear relationship between x_m and y. In the case where $\varphi_m(x_m)$ is a curve, it indicates a nonlinear relationship, capturing more complex interactions between x_m and y. The additive nature of EBM allows for a clear interpretation of each feature's contribution to the final prediction. By examining the shape functions, one can see how each feature influences the outcome. A positive slope in $\varphi_m(x_m)$ indicates that higher values of x_m lead to higher predicted values of y. A negative slope in $\varphi_m(x_m)$ indicates that higher values of x_m lead to lower predicted values of y. The pairwise interaction involve heatmaps or contour plots, where x_m and x_n are on the axes, and the calculated $f(x_m, x_n)$ values fill the plot. This visualization helps in understanding which combinations of m and n contribute most to the outcome.

3.2.3. Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

CMA-ES is an advanced evolutionary algorithm specifically developed for optimization in continuous parameter spaces. It is part of the evolution strategies family, inspired by the principles of natural evolution. It adapts the covariance matrix of the search distribution to effectively explore the search space, concentrating on the most promising regions. The algorithm iteratively updates a population of candidate solutions, utilizing a multivariate normal distribution whose mean and covariance matrix are dynamically adjusted based on the performance of the selected solutions.

The core principle of CMA-ES is to represent the search distribution as a multivariate Gaussian, with its mean and covariance matrix being iteratively refined. It starts with an initial mean vector m_o and covariance matrix C_o and set initial step-size σ_o and population size λ_o. Generate offspring by sampling from a multivariate normal distribution as given by Equation (7)

$$x_k \sim \mathbf{N}\left(m_t, \sigma_t^2 C_t\right) \tag{7}$$

where m_t is the mean vector, σ_t is the step-size, and C_t is the covariance matrix at iteration t. Evaluate the objective function $f(x_k)$ for each offspring x_k. Select the top μ solutions based on their fitness values to form a new mean as shown by Equation (8)

$$m_{t+1} = \sum_{i=1}^{\mu} \omega_i x_i \tag{8}$$

where ω_i denotes the weights allocated to each selected solution. The covariance matrix C_{t+1} is updated and step-sizes σ_{t+1} are represented by Equations (9) and (10), respectively.

$$C_{t+1} = (1 - c_c)C_t + c_c \sum_{i=1}^{\mu} \omega_i (x_i - m_t)(x_i - m_t)^T \tag{9}$$

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\| p_{c,t+1} \|}{E \| N(0,1) \|} - 1\right)\right) \tag{10}$$

where c_c, c_σ, and d_σ represent the learning rates, and $p_{c,t+1}$ denote the evolution path. Continue the process until the convergence criteria are met, such as attaining the maximum number of iterations or achieving the desired fitness level.

For the EBM model, the key hyperparameters requiring careful optimization include the learning rate, the maximum number of bins, the maximum number of interaction bins, and the number of boosting iterations [44]. The learning rate determines the step size at each boosting iteration. A lower learning rate allows the model to learn more gradually, reducing the risk of overfitting and often improving accuracy, while a higher learning rate speeds up training but may compromise EBM model accuracy. The maximum number of

bins defines the number of discrete bins used to partition continuous features. A higher number of bins allows for more detailed feature representation but increases EBM model complexity. This parameter affects how well the model captures variable interactions and feature effects. The maximum number of interaction bins controls the number of bins used specifically for pairwise feature interactions, allowing the model to capture important dependencies between features. Optimizing this parameter enhances the EBM model's ability to capture complex feature relationships without overcomplicating the structure. The number of boosting iterations refers to the total number of boosting rounds. Too few iterations can lead to underfitting, as the model may not fully capture data patterns, while too many can cause overfitting by making the model overly complex.

3.3. Competitive Machine Learning Models

In addition to EBM, several other machine learning models were used to analyze $PM_{2.5}$ concentration, including XGBoost, RF, LightGBM, and AdaBoost. Each model has distinct strengths. XGBoost and LightGBM are both gradient-boosting techniques optimized for efficiency, while Random Forest is known for its robustness in various tasks. AdaBoost, on the other hand, is useful for improving weak learners through iterative weighting. Table 2 shows the summary table of different machine learning models.

Table 2. Summary Table of different machine learning models.

Model	Acronym	Key Features	Primary Applications
Extreme Gradient Boosting	XGBoost	– Utilizes gradient boosting with regularization – Highly efficient and optimized for large datasets – Handles missing values automatically	Regression and classification tasks, especially with structured/tabular data
Random Forest	RF	– Ensemble of decision trees with bootstrapping (bagging) – Reduces overfitting by averaging predictions – Robust to noise and outliers	Broad use in classification and regression, especially for feature importance analysis
Light Gradient Boosting Machine	LightGBM	– Optimized for speed and efficiency, handles large datasets – Uses leaf-wise tree growth – Highly effective for sparse data	High-dimensional data in classification and regression, effective with large datasets
Adaptive Boosting	AdaBoost	– Sequentially combines weak learners to form a strong model – Focuses on errors of previous models – Works well with simple base estimators	Binary classification and situations requiring model interpretability

3.4. Performance Measures

For evaluating the performance of regression models in machine learning, several metrics are commonly used. These include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and the coefficient of determination (R^2). Each metric provides different insights into the accuracy and performance of a model. MAE quantifies the mean magnitude of prediction errors within a dataset, disregarding the directionality of these errors. Specifically, it computes the average across a test sample of the absolute variances between each predicted value and its corresponding actual observation, treating all individual variances uniformly, as given in Equation (11).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (11)$$

MSE calculates the average of the squared differences between estimated values and the actual values. This metric represents the mean square error across the dataset, as specified in Equation (12).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (12)$$

RMSE represents the square root of the average of squared errors, as shown in Equation (13). It serves as a measure of the accuracy with which the model predicts the response.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (13)$$

R^2, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is predictable from the independent variables. It quantifies how closely data points align with the fitted regression line, as shown in Equation (14).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \quad (14)$$

where the following hold:

- n is the number of data points.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- $\overline{y}_i$ is the mean of the actual values y.

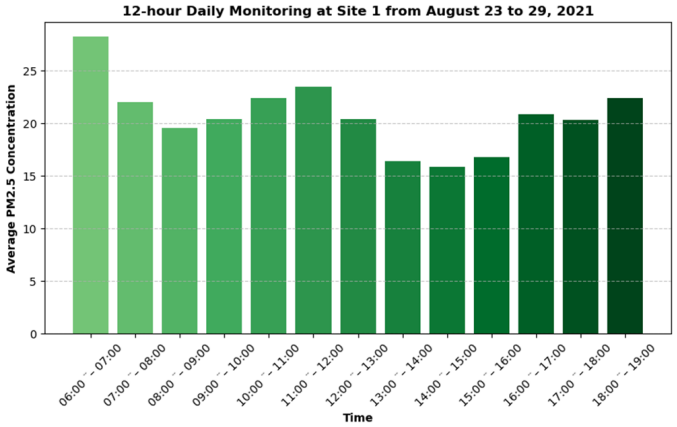
4. Results and Discussion

This study utilizes air quality data collected from sensors positioned along the Nairobi Expressway, supplemented by meteorological and traffic data, including humidity, hourly temperature, average traffic volume, average vehicle speed, wind speed, and site location. To handle missing values in the dataset, the K-Nearest Neighbors (KNNs) method was applied, imputing the missing data points based on the similarity of neighboring values [45]. This approach preserves the dataset's overall quality, making it more robust for predictive modeling. The dataset was divided into two subsets: 70% of the data was allocated for training and validation, while the remaining 30% was reserved for testing. This split follows a widely accepted machine learning practice, enabling model development and hyperparameter tuning on the training–validation set, while the test set remains unseen for performance evaluation.

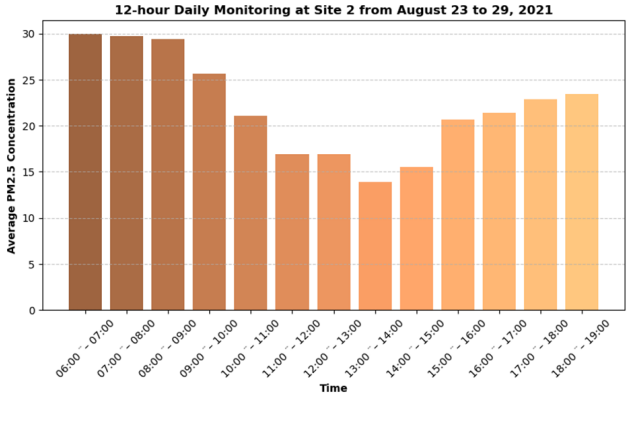
For the model development, the EBM was employed with hyperparameter optimization performed through the CMA-ES. The EBM's inherent interpretability was a key factor in its selection, as it allows for transparent analysis of feature contributions to the predictions. The performance of the EBM model was evaluated on the test set and compared with several alternative machine learning models, including RF [46], XGBoost [47], LightGBM [48], AdaBoost [49], and the MLR [50]. Among the features used in this study, site location was treated as a categorical variable, where Westlands Roundabout (Site 1) was encoded as 0, Nyayo Roundabout (Site 2) as 1, and City Cabanas (Site 3) as 2. All models were implemented in Python 3.7.1, with Table 3 presenting the descriptive statistics for the input factors used in the analysis. As discussed previously, a 12 h daily monitoring was conducted for each of the three sites. Consequently, Figure 3 illustrates the average $PM_{2.5}$ concentration at different times of the day.

Table 3. Summary statistics for different input factors.

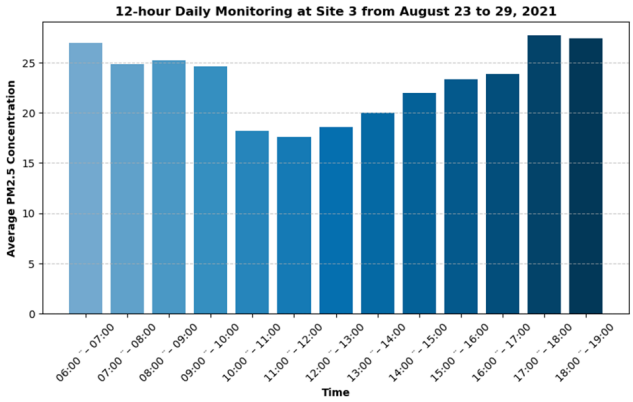
Factors	Descriptive Statistics			
	Mean	Standard Deviation	Min	Max
Humidity (%)	37.52	14.95	15.33	70.16
Temperature (°C)	28.75	6.53	18.86	44.07
Average Traffic Volume (veh/hr)	1379.05	655.16	342	3213
Average Vehicle Speed (km/hr)	46.41	9.71	24.7	62.18
Wind Speed (m/s)	6.61	3.83	2.95	11.75
Location	0.94	0.79	0	2



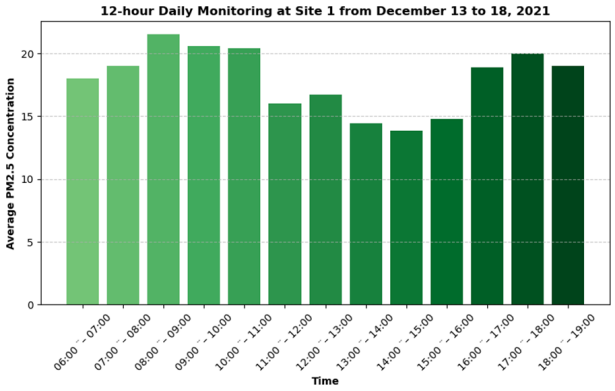
(**a**)



(**b**)



(**c**)



(**d**)

Figure 3. *Cont.*

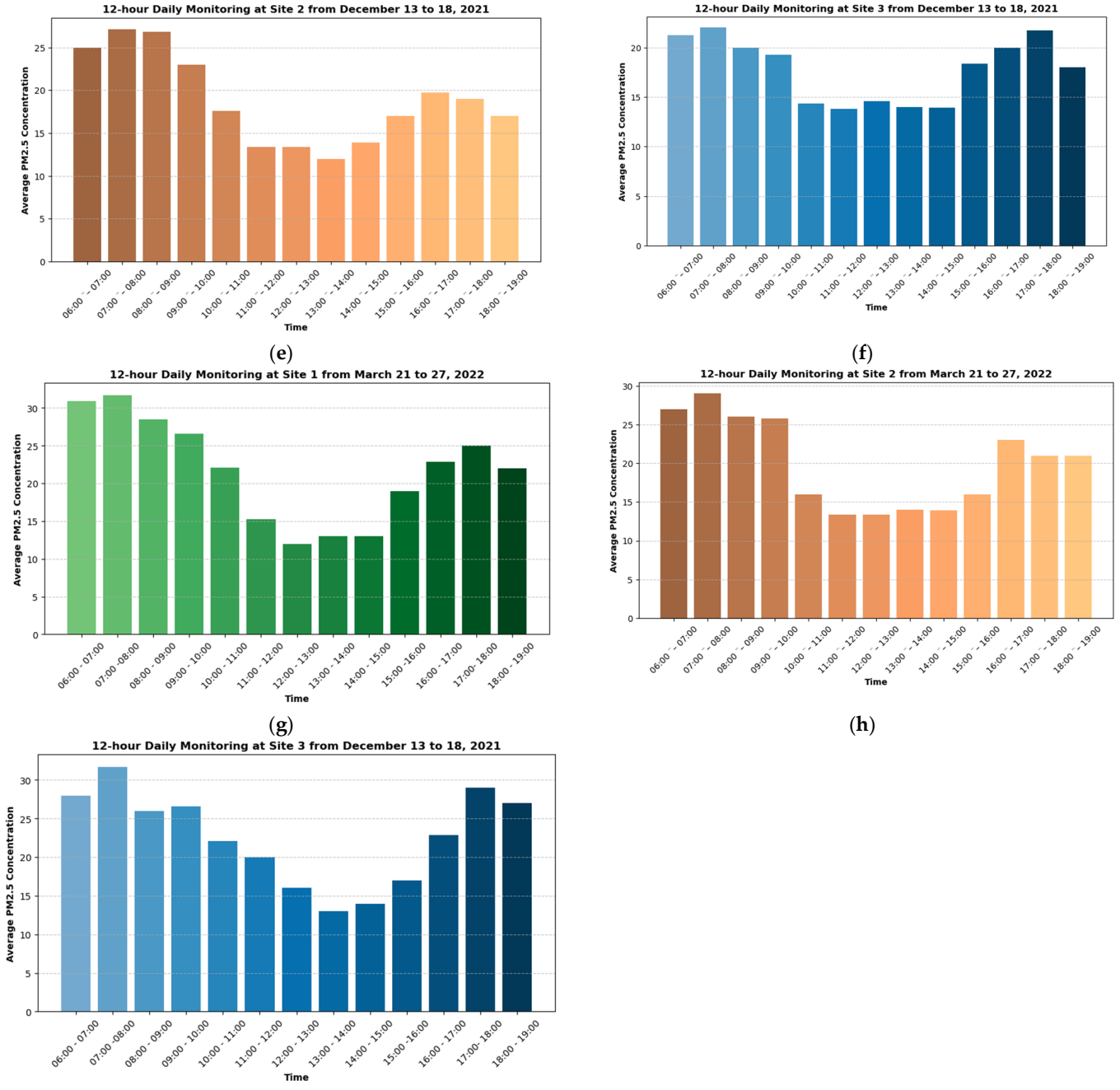


Figure 3. Twelve-hour daily variation in average $PM_{2.5}$ at different sites along Nairobi expressway.

4.1. Fine-Tuning of Hyperparameters via CMA-ES

To optimize the performance of the EBM model, hyperparameters were fine-tuned using the CMA-ES [51]. The primary objective of the tuning process was to maximize the R^2 value, a metric that indicates the predictive performance of the model by measuring the proportion of variance explained by the model predictions for comparative purposes, and the hyperparameters of other models, including XGBoost, RF, LightGBM, and AdaBoost, were also optimized. The best hyperparameters identified for the EBM model and other comparative models are summarized in Table 4.

Table 4. Hyperparameters of different machine learning models.

Models	Hyperparameters	Range	Optimal Values
EBM	n_estimators	[100, 500]	140
	max_bins	[120, 250]	185
	max_interaction_bins	[30, 120]	70
	learning_rate	[0.01, 0.1]	0.08
XGBoost	learning_rate	[0.01, 0.15]	0.08
	n_estimators	[50, 1000]	600.0
RF	n_estimators	[50, 1000]	420.0
	max_depth	[2, 12]	7.0
LightGBM	learning_rate	[0.01, 0.15]	0.13
	n_estimators	[50, 1000]	800.0
AdaBoost	learning_rate	[0.01, 0.15]	0.06
	n_estimators	[50, 1000]	180.0

4.2. Prediction Results and Comparative Analysis

The comprehensive performance evaluation across multiple machine learning models, as shown in Table 5, presents the performance metrics for predicting $PM_{2.5}$ concentrations. Among all models, the EBM, fine-tuned using the CMA-ES, demonstrated the most robust and accurate predictions. EBM outperformed the other models in both the training and testing datasets, with lower error rates and higher R^2 values, highlighting its predictive superiority and robustness. The EBM-CMA-ES model achieved an MAE of 1.615 on the training set and 2.033 on the testing set, an MSE of 15.539 on the training set and 28.134 on the testing set, an RMSE of 3.942 on the training set and 5.304 on the testing set, and an R^2 of 0.904 on the training set and 0.843 on the testing set. These results confirm that the EBM model not only performs well on the training data but also generalizes effectively to unseen data, making it highly suitable for $PM_{2.5}$ concentration forecasting.

Table 5. Performance evaluation of EBM, other competitive machine learning models, and a statistical MLR.

Models	Training Dataset				Testing Dataset			
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
EBM	1.61	15.53	3.94	0.90	2.03	28.13	5.30	0.84
XGBoost	3.58	31.56	5.62	0.81	3.84	34.58	5.88	0.78
RF	4.26	38.29	6.19	0.77	4.52	40.79	6.39	0.74
LightGBM	4.13	33.68	5.8	0.80	4.27	34.4	5.87	0.78
AdaBoost	7.01	74.39	8.63	0.55	6.75	68.18	8.26	0.57
MLR	7.55	97.98	9.95	0.41	7.23	89.12	9.44	0.43

In comparison, the second-best-performing model was XGBoost, which delivered an MAE of 3.58 on the training set and 3.84 on the testing set, an MSE of 31.56 on the training set and 34.58 on the testing set, an RMSE of 5.62 on the training set and 5.88 on the testing set, and an R^2 of 0.813 on the training set and 0.782 on the testing set. While XGBoost showed relatively strong performance, its error rates were higher, and its R^2 values were lower than those of the EBM model, indicating that it was less effective at capturing the underlying patterns in the data. At the other end of the spectrum, the MLR model displayed the weakest performance. It yielded an MAE of 7.55 on the training set and 7.23 on the testing set, an MSE of 97.98 on the training set and 89.12 on the testing set, an RMSE of 9.95 on the training set and 9.44 on the testing set, and an R^2 of 0.418 on the training set and 0.438 on the testing set. These results indicate that MLR struggled to capture the complex relationships within the dataset, making it the least suitable model for predicting $PM_{2.5}$ concentrations.

The prediction error plots in Figure 4 provide a clear visualization of the performance differences among the various models. These plots compare predicted values with actual data points, with a 45-degree reference line representing perfect predictions. The hybrid

EBM-CMA-ES model stands out due to its close alignment with the reference line, both in the training and testing datasets. The dense clustering of data points along the line highlights the model's high accuracy and low error, reinforcing its superior performance in predicting $PM_{2.5}$ levels. In contrast, the prediction error plots for alternative models, including XGBoost, Random Forest, LightGBM, AdaBoost, and MLR, show greater dispersion of points around the 45-degree line. This scattering reflects their comparatively lower predictive accuracy, with more variability in their ability to match actual values. The broader distribution of data points away from the ideal line demonstrates that these models, while still effective to varying degrees, are less reliable than the EBM-CMA-ES model in accurately forecasting $PM_{2.5}$ levels.

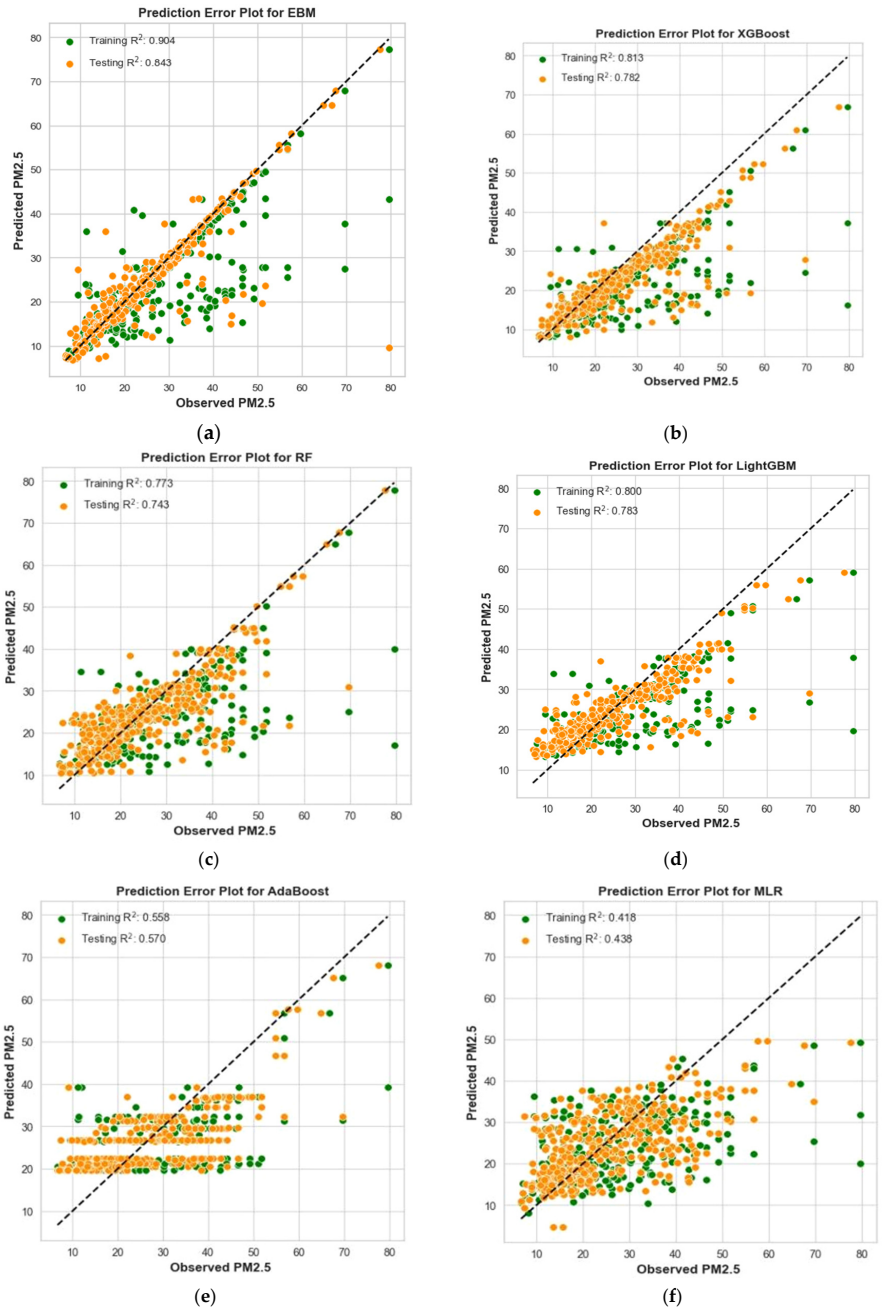


Figure 4. Prediction error plots using both training and testing datasets: (**a**) EBM; (**b**) XGBoost; (**c**) RF; (**d**) LightGBM; (**e**) AdaBoost; (**f**) MLR.

4.3. Uncertainty Analysis

The uncertainty analysis evaluates the variability and reliability of each machine learning model's predictions by comparing the ratio of predicted $PM_{2.5}$ concentrations to observed $PM_{2.5}$ concentrations against the observed values. Figure 5a–f present these ratios for each model, while Table 6 summarizes the mean ratio and standard deviation for each model.

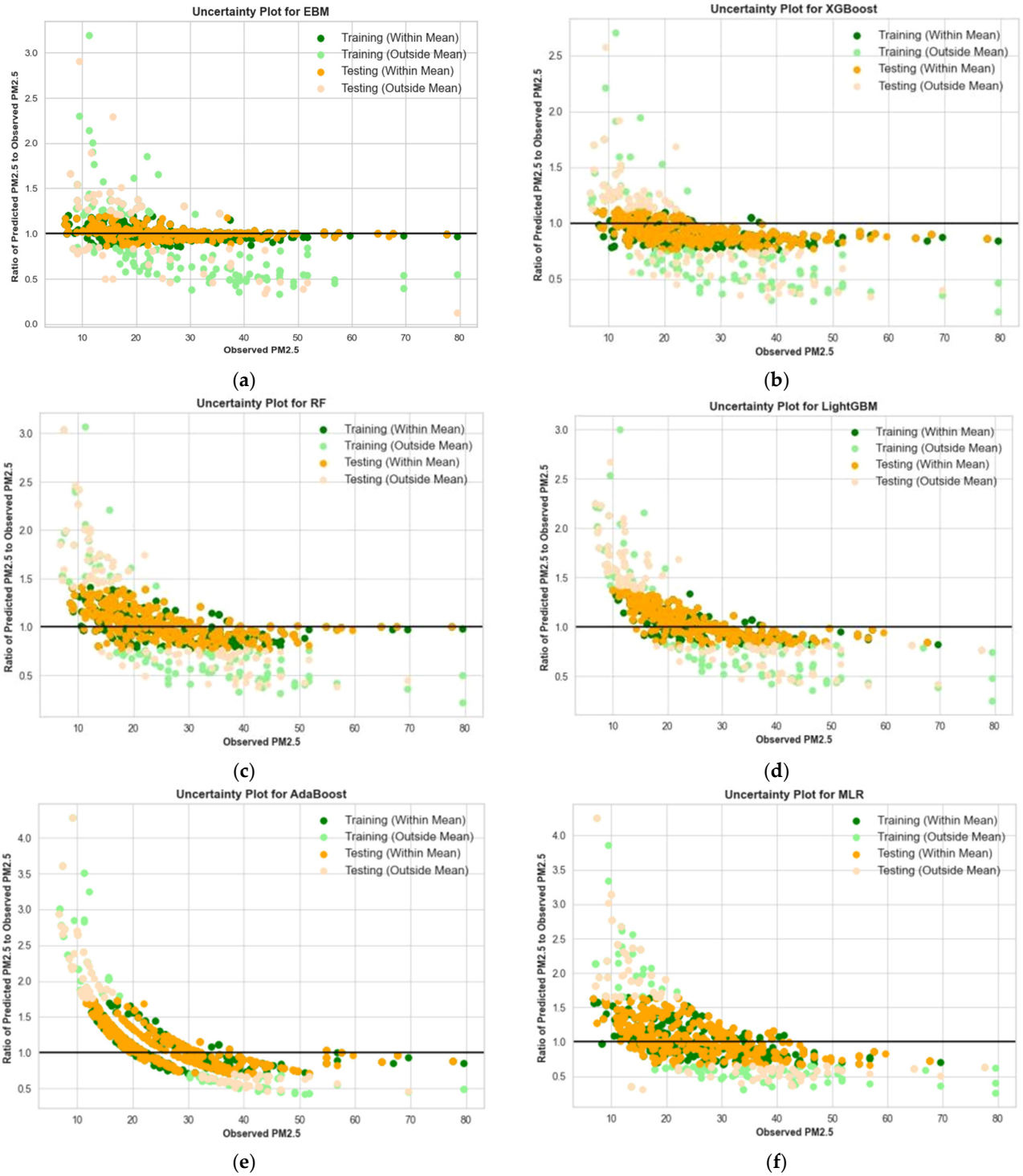


Figure 5. Uncertainty analysis of the machine learning model by plotting the ratio of predicted $PM_{2.5}$ to the observed $PM_{2.5}$ vs observed $PM_{2.5}$: (**a**) EBM model (**b**) XGBoost model (**c**) RF model; (**d**) LightGBM model; (**e**) AdaBoost model; (**f**) MLR model.

Table 6. Uncertainty analysis in terms of mean ratio and standard deviation of data points.

Models	Mean	Standard Deviation
EBM	1.0024	0.178
XGBoost	0.942	0.195
RF	0.926	0.221
LightGBM	0.934	0.199
AdaBoost	0.786	0.304
MLR	0.772	0.313

In Figure 5, the plot for each model provides a visual representation of how well the model's predictions align with the observed $PM_{2.5}$ values. For instance, Figure 5a shows that the EBM model has the most consistent performance, with the ratio of predicted to observed $PM_{2.5}$ values clustering tightly around 1.0, indicating minimal deviation from the observed values. In contrast, Figure 5e,f, which represents the AdaBoost and MLR models, respectively, shows more scattered data points, suggesting higher variability in predictions. Table 6 further supports this observation by providing the mean ratio and standard deviation for each model's predictions. The EBM model has a mean ratio close to 1.0024, indicating near-perfect accuracy and a relatively low standard deviation of 0.178, which confirms the model's high precision and reliability. Similarly, the XGBoost and LightGBM models perform reasonably well, with mean ratios of 0.942 and 0.934, respectively, and moderate standard deviations, highlighting their stability. However, the AdaBoost and MLR models show considerably higher uncertainty, with mean ratios of 0.786 and 0.772, respectively, and the highest standard deviations (0.304 and 0.313). This indicates that these models have larger deviations in predictions, suggesting lower reliability and less accuracy in forecasting $PM_{2.5}$ concentrations.

4.4. EBM Interpretation

In this section, we interpret our proposed EBM model, which emerged as the best-performing model for $PM_{2.5}$ prediction. By employing the EBM model's inherent interpretability, we are able to gain valuable insights into the factors that contribute to $PM_{2.5}$ concentrations. We begin with a global interpretation by examining the feature importance plot and then proceed to local interpretations to explore how these factors affect individual predictions.

4.4.1. EBM Global Interpretation

The feature importance plot, as shown in Figure 6, generated from the EBM model, provides a ranked list of features based on their impact on the model's predictions. The top three most significant individual factors contributing to $PM_{2.5}$ concentration predictions are location, humidity, and temperature, each of which plays a distinct role in $PM_{2.5}$ forecasting. Location stands out as the most critical factor in the model. This is likely because different locations along the Nairobi Expressway experience varying levels of pollution due to local sources such as traffic density, industrial activities, and proximity to urban centers. Location-specific factors like the presence of high-emission vehicles or specific meteorological conditions at certain sites could explain why this feature is highly influential. Humidity is the second most significant feature. Its importance likely stems from its ability to influence the behavior of airborne particles. High humidity can lead to the aggregation of particulate matter, increasing $PM_{2.5}$ concentrations. Alternatively, during low humidity conditions, dry air can enhance the resuspension of particles from surfaces, further elevating $PM_{2.5}$ levels. Therefore, fluctuations in humidity are strongly tied to changes in air pollution. Similarly, Temperature is the third most important individual feature. The relationship between temperature and $PM_{2.5}$ levels could be attributed to several factors, including the impact of temperature on atmospheric mixing and chemical reactions. Warmer temperatures may reduce the stability of the atmosphere, allowing pollutants to disperse more easily, while cooler temperatures could trap pollutants closer

to the ground. Furthermore, temperature can affect the formation of secondary particles, which contribute to $PM_{2.5}$ levels.

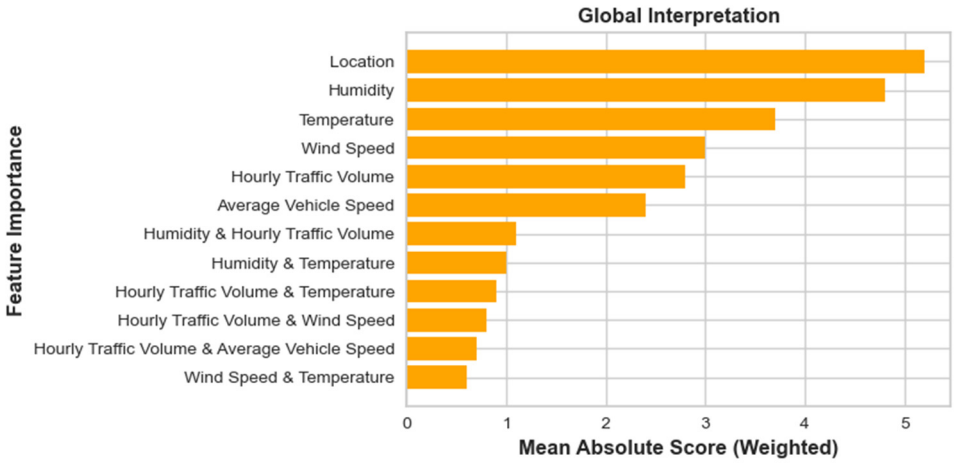


Figure 6. Global factors importance analysis via EBM.

In addition, the most significant interaction term is between humidity and hourly traffic volume, indicating that the combined effect of these two features plays an important role in $PM_{2.5}$ concentration levels. This interaction likely highlights the dual impact of traffic emissions and atmospheric conditions. Increased traffic volume results in higher emissions of particulate matter, particularly from combustion engines. When traffic levels rise, $PM_{2.5}$ concentrations are expected to increase. The interaction with humidity suggests that the effect of traffic on $PM_{2.5}$ concentrations may vary depending on the moisture content in the air. Under high humidity conditions, the particles emitted by vehicles could cluster more easily, resulting in higher recorded $PM_{2.5}$ levels. Conversely, at lower humidity levels, the particles might behave differently, potentially reducing their clustering but increasing their dispersal.

Figure 7 illustrates the impact of three key locations, including Westlands Roundabout (Location 1), Nyayo Roundabout (Location 2), and City Cabanas (Location 3), on $PM_{2.5}$ concentrations. At the Westlands Roundabout, the score is strongly positive, around 10, indicating that this location contributes significantly to higher $PM_{2.5}$ levels, likely due to heavy traffic and congestion. In contrast, Nyayo Roundabout shows a sharp drop in the score to around -10, reflecting a negative impact on $PM_{2.5}$ levels, possibly due to smoother traffic flow or more favorable environmental conditions, which help reduce pollution. Lastly, City Cabanas has a score close to 0, suggesting a neutral impact on $PM_{2.5}$ concentrations, with minimal influence on pollution levels. The filled blue areas highlight the magnitude of these effects, with Westlands being a significant contributor to pollution, Nyayo acting as a reducer, and City Cabanas having a negligible impact.

Similarly, Figure 8 illustrates the relationship between humidity and its impact on $PM_{2.5}$ concentrations. As humidity increases from 15 to 65, its influence on $PM_{2.5}$ levels fluctuates in a nonlinear pattern. Initially, at lower humidity levels around 15 to 25, the score shows a slight positive contribution, indicating a modest effect on $PM_{2.5}$ concentrations. As humidity rises to around 30, the score dips briefly, showing that moderate humidity has a reducing effect, leading to a temporary decrease in concentrations. However, as humidity continues to increase beyond 35, the score rises significantly, indicating that higher humidity levels have a more pronounced positive impact on $PM_{2.5}$ concentrations. This illustrates that when humidity reaches around 40 to 55, the concentration of $PM_{2.5}$ particles increases considerably. Mechanisms such as particle deliquescence and aqueous-phase chemical reactions play their roles at elevated humidity. Deliquescence occurs when hygroscopic particles absorb moisture at specific humidity thresholds, transitioning from solid to aqueous phases, which enhances their ability to act as reaction sites for atmospheric chemicals. This process facilitates the formation of secondary inorganic aerosols, notably ammonium

nitrate, under humid conditions. Research indicates that ammonium nitrate formation is more efficient in moist environments due to the increased solubility and reactivity of precursor gases like ammonia and nitric acid in aqueous aerosols [52]. Furthermore, studies have shown that the deliquescence and efflorescence relative humidities of aerosol particles are critical in determining their phase states and subsequent chemical reactivity, directly impacting $PM_{2.5}$ levels [53]. The score peaks at the highest humidity levels, around 55 to 60, although the upward trend diminishes slightly towards the end, implying that while very high humidity still increases $PM_{2.5}$ levels, the effect stabilizes.

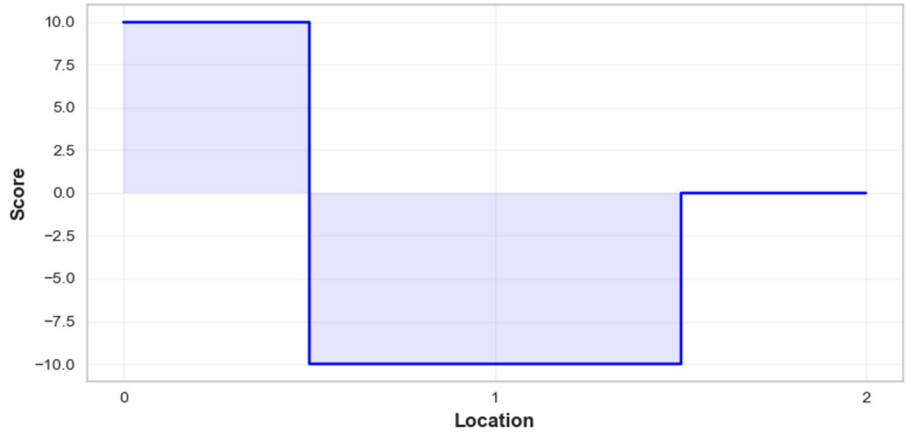


Figure 7. Influence of location on $PM_{2.5}$ concentrations.

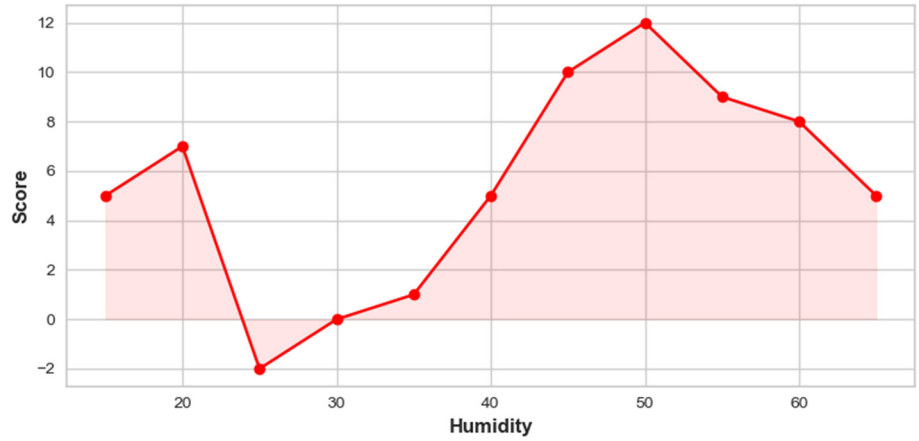


Figure 8. Influence of humidity on $PM_{2.5}$ concentrations.

Figure 9 illustrates the relationship between temperature and its impact on $PM_{2.5}$ concentrations. As the temperature increases from 15 to 40, the score shows a clear downward trend. Initially, at lower temperatures, particularly between 15 and 20, the score rises significantly, indicating a strong positive influence on $PM_{2.5}$ concentrations. However, as temperatures exceed 20, the score begins to decline and eventually drops below 0 around 30, suggesting that higher temperatures have a diminishing effect, leading to no impact on $PM_{2.5}$ levels. This trend implies that as temperatures rise, the environment's capacity to maintain or increase $PM_{2.5}$ concentrations decreases. The shaded area under the curve highlights the range of scores corresponding to specific temperature intervals, with the light green shading emphasizing the transition from a positive influence to a neutral or negative effect. Overall, the plot indicates that while lower temperatures contribute significantly to $PM_{2.5}$ concentrations, higher temperatures are associated with a reduction in influence, ultimately leading to negligible effects on pollution levels.

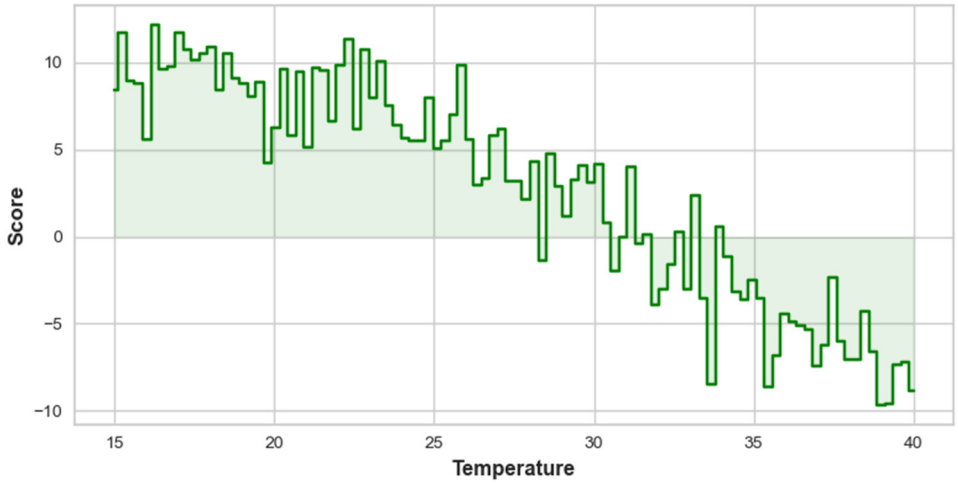


Figure 9. Influence of temperature on $PM_{2.5}$ concentrations.

In the case of feature interaction, the EBM heatmap illustrates the interaction between humidity and hourly traffic volume, with a color gradient indicating the corresponding interaction scores, as shown in Figure 10. Areas with humidity levels below 40 are predominantly shaded in purple, illustrating a minimal effect on $PM_{2.5}$ concentrations. In contrast, as humidity increases to between 40 and 60, the colors transition to yellow and orange, indicating a strong positive influence on $PM_{2.5}$ levels, particularly when combined with high traffic volumes (above 1500). This shows that optimal conditions for higher $PM_{2.5}$ concentrations occur within this humidity and traffic volume range. The heatmap effectively communicates that higher humidity and traffic volumes can exacerbate air pollution, making it essential to monitor these parameters for effective environmental management.

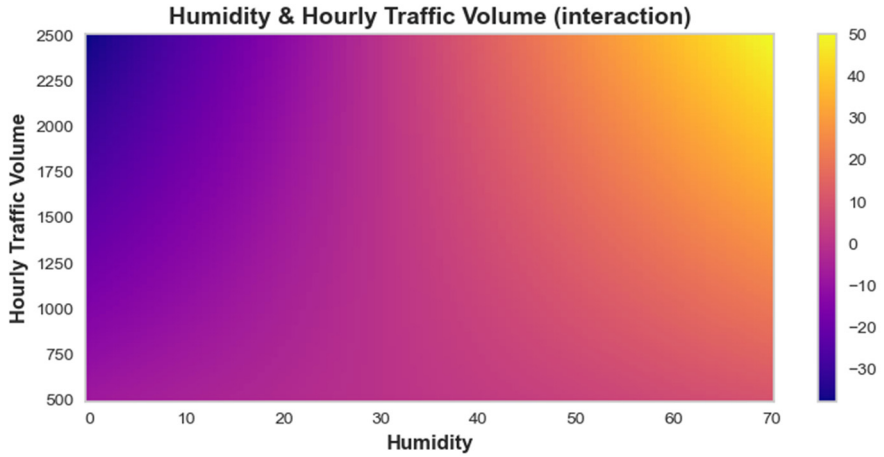


Figure 10. EBM-based heatmap for the interaction of humidity and hourly traffic volume.

4.4.2. EBM Local Interpretation

In addition to providing a global interpretation of the overall model's behavior, the EBM can also be used for local interpretation, offering insights into how individual features influence the prediction for specific samples. In this study, we consider two randomly selected samples, i.e., sample #12 and sample #38 of the testing dataset. For sample #12, as shown in Figure 11, the model reveals the contributions of various factors to the predicted outcome of 27.8, allowing us to understand the specific impact of the most important features in this case. The feature with the strongest positive contribution is location (0.00), which adds 10 units to the prediction. This indicates that the sample's location is a key driver in increasing the predicted value.

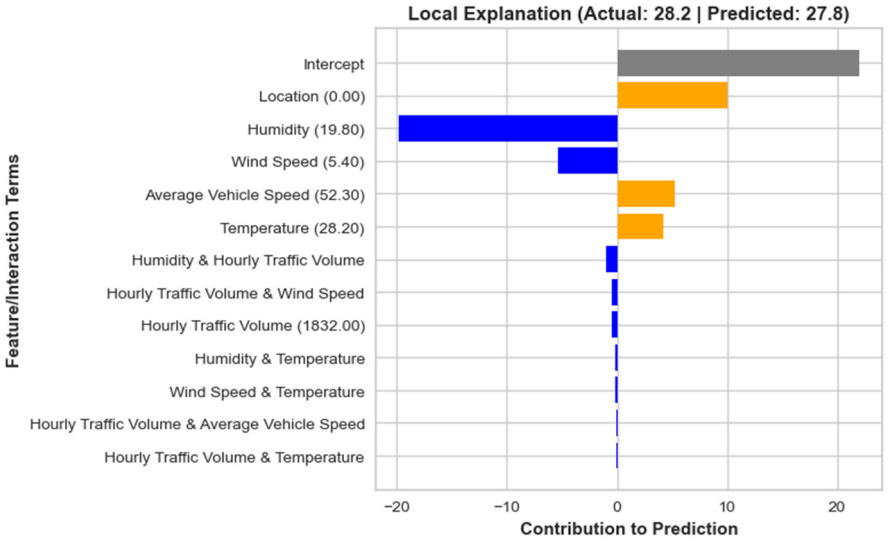


Figure 11. EBM-based local interpretation of Sample # 12 in testing dataset.

On the other hand, humidity (19.80) has a significant negative contribution of approximately −19.8 units, showing that the humidity levels, in this case, strongly reduce the prediction. Finally, wind speed (5.40) also contributes negatively, lowering the predicted value by −5.4 units, showing that higher wind speeds are associated with a reduction in the outcome for this sample. In this local interpretation, we see that while location plays a major positive role, humidity and wind speed act as substantial negative influences, shaping the final prediction together. This level of insight helps to explain why the model predicted 27.8 for this particular sample, providing transparency into the model's decision-making process.

This local interpretation plot for sample #38 highlights how the top three factors, temperature (38.20), location (0.00), and humidity (22.50), contribute to the model's predicted value of 10.4, which is close to the actual value of 10.1, as shown by Figure 12. Temperature (38.20) has the strongest negative influence on the prediction, reducing it by approximately −10 units. This suggests that higher temperatures in this specific case lead to a lower predicted outcome, making temperature one of the key drivers pushing the prediction downward.

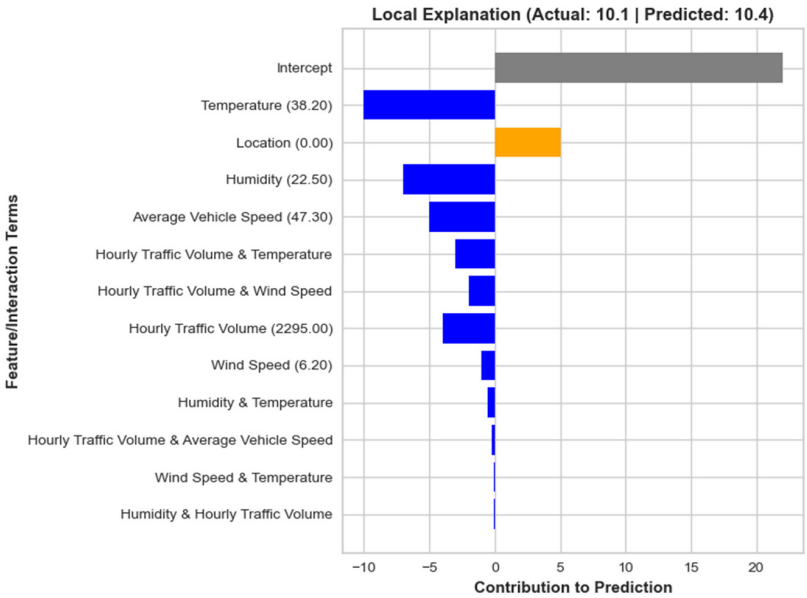


Figure 12. EBM-based local interpretation of Sample # 12 in testing dataset.

In contrast, location (0.00) has a significant positive impact, contributing about five units to the prediction. The specific location of this sample plays a major role in increasing the predicted value, offsetting some of the negative effects from other factors. On the other hand, humidity (22.50) has a substantial negative contribution, lowering the prediction by about −7 units. This indicates that higher humidity levels for this sample are associated with a decrease in the predicted value. Together, these three factors explain much of the predicted outcome, with temperature and humidity reducing the value while location increases it.

5. Conclusions and Recommendations

This study developed a Hybrid Explainable Boosting Machine (EBM) framework, optimized with the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), to predict vehicle-related $PM_{2.5}$ concentrations along the Nairobi Expressway. The model effectively captured the influence of environmental and traffic-related factors, providing both high predictive accuracy and interpretability. The findings provide valuable insights for air quality management in developing urban areas:

- The EBM-CMA-ES model was the best-performing model, achieving an MAE of 2.033 and an R^2 of 0.843 on the testing set, significantly outperforming alternative models like XGBoost, RF, and MLR.
- The MLR model performed the worst, with an MAE of 7.226 and an R^2 of 0.438, indicating its limitations in capturing the complex relationships between environmental factors and $PM_{2.5}$ levels.
- Based on the EBM global interpretation results, location was identified as the most critical factor influencing $PM_{2.5}$ concentrations, with areas near the Westlands roundabout showing the highest levels, likely due to traffic congestion.
- Humidity was found to have a strong positive effect on $PM_{2.5}$ levels, with medium to high humidity linked to increased particle concentrations. Elevated humidity promotes hygroscopic growth, enabling fine particles to absorb water, increasing their size and mass, which elevates $PM_{2.5}$ concentrations. Humidity also enhances aerosol acidity, facilitating secondary aerosol formation. These processes make humidity a critical factor in increasing PM2.5 levels [54]. Temperature showed an inverse relationship with $PM_{2.5}$ concentrations, where higher temperatures were associated with reduced $PM_{2.5}$ levels, likely due to enhanced atmospheric mixing.
- The interaction between humidity and traffic volume was significant, demonstrating that high traffic volume combined with increased humidity results in higher $PM_{2.5}$ concentrations, highlighting the need for targeted interventions in such conditions.

5.1. Limitations of Study

This study has several limitations that may affect the generalizability and scope of its findings. Firstly, data were only collected from three monitoring sites along the Nairobi Expressway. This limited spatial coverage may reduce the model's ability to generalize to other urban regions with varying traffic patterns and environmental conditions. Areas with different road configurations, traffic intensities, or urban layouts may exhibit different $PM_{2.5}$ concentrations, which the current study does not capture.

Moreover, this study considered a limited set of meteorological variables, including humidity, temperature, and wind speed, and did not incorporate other potentially significant factors, such as atmospheric pressure, precipitation, or pollutant interactions. The absence of these variables may lead to an incomplete understanding of the factors influencing $PM_{2.5}$ concentrations. Additionally, the model did not account for temporal dynamics, meaning it did not consider how traffic patterns or weather conditions change over time (e.g., during different seasons or times of the day), which could further affect $PM_{2.5}$ levels.

5.2. Future Recommendations

To enhance the accuracy and generalizability of the hybrid EBM-CMA-ES model, future studies should focus on extending data collection over longer periods to capture seasonal and long-term variations in air quality and traffic patterns. By increasing the duration of data collection, the model can be more robust in accounting for temporal dynamics, which are essential for understanding how factors like traffic volume and meteorological conditions fluctuate over time. Additionally, expanding sensor coverage by deploying more air quality monitoring stations along the expressway and in surrounding areas will provide a more comprehensive understanding of $PM_{2.5}$ distribution. This will allow for a finer spatial analysis and improve the model's ability to generalize to other urban environments.

Incorporating additional meteorological factors, such as atmospheric pressure, precipitation, and solar radiation, will further enhance the model's predictive capability by accounting for a broader range of environmental influences on $PM_{2.5}$ concentrations. Policy interventions should be targeted in high-risk areas, such as the Westlands roundabout, particularly during periods of high traffic and humidity, when $PM_{2.5}$ levels are likely to peak. Future research should also explore the use of temporal models to account for time-dependent changes in traffic and weather conditions, improving the accuracy of air pollution forecasts and aiding in more informed decision-making for urban air quality management.

Author Contributions: Conceptualization, A.K.; data curation, A.A.M.E. and A.K.; formal analysis, S.A., H.A. and A.A.M.E.; funding acquisition, A.A.; investigation, H.A.; methodology, S.A. and K.A.A.M.; project administration, B.T.A.; resources, H.A. and A.A.; supervision, S.A.; validation, B.T.A. and A.A.; visualization, H.A. and A.K.; writing—original draft, K.A.A.M.; writing—review and editing, K.A.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Taif University, Saudi Arabia, Project No. (TU-DSPP-2024-33).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-33). In addition, we extend our gratitude to Caroline Matara for providing the air quality data, which has been exclusively utilized for research purpose. We also acknowledge the use of the Grammarly tool for grammar checking in the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. McMichael, A.J. The urban environment and health in a world of increasing globalization: Issues for developing countries. *Bull. World Health Organ.* **2000**, *78*, 1117–1126. [PubMed]
2. Qing, W. Urbanization and global health: The role of air pollution. *Iran. J. Public Health* **2018**, *47*, 1644.
3. Southerland, V.A.; Brauer, M.; Mohegh, A.; Hammer, M.S.; Van Donkelaar, A.; Martin, R.V.; Apte, J.S.; Anenberg, S.C. Global urban temporal trends in fine particulate matter (PM2·5) and attributable health burdens: Estimates from global datasets. *Lancet Planet. Health* **2022**, *6*, e139–e146. [CrossRef]
4. Giannadaki, D.; Lelieveld, J.; Pozzer, A. Implementing the US air quality standard for PM 2.5 worldwide can prevent millions of premature deaths per year. *Environ. Health* **2016**, *15*, 88. [CrossRef] [PubMed]
5. Lippmann, M. Toxicological and epidemiological studies of cardiovascular effects of ambient air fine particulate matter (PM2.5) and its chemical components: Coherence and public health implications. *Crit. Rev. Toxicol.* **2014**, *44*, 299–347. [CrossRef]
6. Thangavel, P.; Park, D.; Lee, Y.-C. Recent insights into particulate matter (PM2.5)-mediated toxicity in humans: An overview. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7511. [CrossRef]
7. Meng, X.; Zhang, Y.; Yang, K.-Q.; Yang, Y.-K.; Zhou, X.-L. Potential harmful effects of PM2.5 on occurrence and progression of acute coronary syndrome: Epidemiology, mechanisms, and prevention measures. *Int. J. Environ. Res. Public Health* **2016**, *13*, 748. [CrossRef]

8. Wan Mahiyuddin, W.R.; Ismail, R.; Mohammad Sham, N.; Ahmad, N.I.; Nik Hassan, N.M.N. Cardiovascular and respiratory health effects of fine particulate matters (PM2.5): A review on time series studies. *Atmosphere* **2023**, *14*, 856. [CrossRef]
9. Tang, Z.; Jia, J. The Association between the Burden of PM2.5-Related Neonatal Preterm Birth and Socio-Demographic Index from 1990 to 2019: A Global Burden Study. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10068. [CrossRef]
10. Anwar, M.N.; Shabbir, M.; Tahir, E.; Iftikhar, M.; Saif, H.; Tahir, A.; Murtaza, M.A.; Khokhar, M.F.; Rehan, M.; Aghbashlo, M. Emerging challenges of air pollution and particulate matter in China, India, and Pakistan and mitigating solutions. *J. Hazard. Mater.* **2021**, *416*, 125851. [CrossRef]
11. Alhakbani, A. *Battery Blueprint: Saudi Arabia's Strategic Foray into the Battery Value Chain*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2024; Available online: https://dspace.mit.edu/handle/1721.1/156985 (accessed on 7 October 2024).
12. Boadi, K.; Kuitunen, M.; Raheem, K.; Hanninen, K. Urbanisation without development: Environmental and health implications in African cities. *Environ. Dev. Sustain.* **2005**, *7*, 465–500. [CrossRef]
13. Sadiq, A.A. *Effect of Particulate Emissions from Road Transportation Vehicles on Health of Communities in Urban and Rural Areas, Kano State, Nigeria*; Université Claude Bernard-Lyon I: Villeurbanne, France, 2022; Available online: https://theses.hal.science/tel-03963351 (accessed on 7 October 2024).
14. Kinney, P.L.; Gichuru, M.G.; Volavka-Close, N.; Ngo, N.; Ndiba, P.K.; Law, A.; Gachanja, A.; Gaita, S.M.; Chillrud, S.N.; Sclar, E. Traffic impacts on PM2.5 air quality in Nairobi, Kenya. *Environ. Sci. Policy* **2011**, *14*, 369–378. [CrossRef] [PubMed]
15. Kebe, M.; Traore, A.; Manousakas, M.I.; Vasilatou, V.; Ndao, A.S.; Wague, A.; Eleftheriadis, K. Source apportionment and assessment of air quality index of PM2.5–10 and PM2.5 in at two different sites in urban background area in Senegal. *Atmosphere* **2021**, *12*, 182. [CrossRef]
16. Chen, Z.; Chen, D.; Zhao, C.; Kwan, M.-p.; Cai, J.; Zhuang, Y.; Zhao, B.; Wang, X.; Chen, B.; Yang, J. Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism. *Environ. Int.* **2020**, *139*, 105558. [CrossRef]
17. Karimian, H.; Li, Q.; Li, C.; Chen, G.; Mo, Y.; Wu, C.; Fan, J. Spatio-temporal variation of wind influence on distribution of fine particulate matter and its precursor gases. *Atmos. Pollut. Res.* **2019**, *10*, 53–64. [CrossRef]
18. Tong, R.; Liu, J.; Wang, W.; Fang, Y. Health effects of PM2.5 emissions from on-road vehicles during weekdays and weekends in Beijing, China. *Atmos. Environ.* **2020**, *223*, 117258. [CrossRef]
19. Nori, H.; Jenkins, S.; Koch, P.; Caruana, R. Interpretml: A unified framework for machine learning interpretability. *arXiv* **2019**, arXiv:1909.09223.
20. Jędrzejewski-Szmek, Z.; Abrahao, K.P.; Jędrzejewska-Szmek, J.; Lovinger, D.M.; Blackwell, K.T. Parameter optimization using covariance matrix adaptation—Evolutionary strategy (CMA-ES), an approach to investigate differences in channel properties between neuron subtypes. *Front. Neuroinform.* **2018**, *12*, 47. [CrossRef]
21. Chen, Z.; Tan, S.; Nori, H.; Inkpen, K.; Lou, Y.; Caruana, R. Using explainable boosting machines (ebms) to detect common flaws in data. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bilbao, Spain, 13–17 September 2021; pp. 534–551.
22. Liu, G.; Sun, B. Concrete compressive strength prediction using an explainable boosting machine model. *Case Stud. Constr. Mater.* **2023**, *18*, e01845. [CrossRef]
23. Maxwell, A.E.; Sharma, M.; Donaldson, K.A. Explainable boosting machines for slope failure spatial predictive modeling. *Remote Sens.* **2021**, *13*, 4991. [CrossRef]
24. Khattak, A.; Chan, P.-w.; Chen, F.; Peng, H. Assessing wind field characteristics along the airport runway glide slope: An explainable boosting machine-assisted wind tunnel study. *Sci. Rep.* **2023**, *13*, 10939. [CrossRef] [PubMed]
25. Bajer, L.; Pitra, Z.; Repický, J.; Holeňa, M. Gaussian process surrogate models for the CMA evolution strategy. *Evol. Comput.* **2019**, *27*, 665–697. [CrossRef] [PubMed]
26. Anggoro, D.A.; Mukti, S.S. Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure. *Int. J. Intell. Eng. Syst.* **2021**, *14*, 198–207.
27. Deng, Y.; Wang, J.; Sun, L.; Wang, Y.; Chen, J.; Zhao, Z.; Wang, T.; Xiang, Y.; Wang, Y.; Chen, J. Effects of ambient O3 on respiratory mortality, especially the combined effects of PM2.5 and O3. *Toxics* **2023**, *11*, 892. [CrossRef]
28. Hoy, A.; Mohan, G.; Nolan, A. An investigation of inequalities in exposure to PM2.5 air pollution across small areas in Ireland. *Int. J. Health Geogr.* **2024**, *23*, 17. [CrossRef]
29. Marsha, A.; Larkin, N.K. A statistical model for predicting PM2.5 for the western United States. *J. Air Waste Manag. Assoc.* **2019**, *69*, 1215–1229. [CrossRef]
30. Murray, N.L.; Holmes, H.A.; Liu, Y.; Chang, H.H. A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation. *Environ. Res.* **2019**, *178*, 108601. [CrossRef]
31. Ameen, M.H.; Jumaah, H.J.; Kalantar, B.; Ueda, N.; Halin, A.A.; Tais, A.S.; Jumaah, S.J. Evaluation of PM2.5 particulate matter and noise pollution in Tikrit University based on GIS and statistical modeling. *Sustainability* **2021**, *13*, 9571. [CrossRef]
32. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Wortman Vaughan, J.W.; Wallach, H. Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–52.
33. Wei, W.; Ramalho, O.; Malingre, L.; Sivanantham, S.; Little, J.C.; Mandin, C. Machine learning and statistical models for predicting indoor air quality. *Indoor Air* **2019**, *29*, 704–726. [CrossRef]

34. Matara, C.; Osano, S.; Yusuf, A.O.; Aketch, E.O. Prediction of Vehicle-induced Air Pollution based on Advanced Machine Learning Models. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 12837–12843. [CrossRef]
35. Khanzode, K.C.A.; Sarode, R.D. Advantages and disadvantages of artificial intelligence and machine learning: A literature review. *Int. J. Libr. Inf. Sci. (IJLIS)* **2020**, *9*, 3.
36. Chang, F.-J.; Chang, L.-C.; Kang, C.-C.; Wang, Y.-S.; Huang, A. Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques. *Sci. Total Environ.* **2020**, *736*, 139656. [CrossRef]
37. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ.* **2018**, *636*, 52–60. [CrossRef]
38. Xiao, Q.; Chang, H.H.; Geng, G.; Liu, Y. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environ. Sci. Technol.* **2018**, *52*, 13260–13269. [CrossRef]
39. Peng, J.; Han, H.; Yi, Y.; Huang, H.; Xie, L. Machine learning and deep learning modeling and simulation for predicting PM2.5 concentrations. *Chemosphere* **2022**, *308*, 136353. [CrossRef]
40. Zaman, N.A.F.K.; Kanniah, K.D.; Kaskaoutis, D.G.; Latif, M.T. Evaluation of machine learning models for estimating PM2.5 concentrations across Malaysia. *Appl. Sci.* **2021**, *11*, 7326. [CrossRef]
41. Xiao, F.; Yang, M.; Fan, H.; Fan, G.; Al-Qaness, M.A. An improved deep learning model for predicting daily PM2.5 concentration. *Sci. Rep.* **2020**, *10*, 20988. [CrossRef] [PubMed]
42. Sun, K.; Tang, L.; Qian, J.; Wang, G.; Lou, C. A deep learning-based PM2.5 concentration estimator. *Displays* **2021**, *69*, 102072. [CrossRef]
43. Kristiani, E.; Lin, H.; Lin, J.-R.; Chuang, Y.-H.; Huang, C.-Y.; Yang, C.-T. Short-term prediction of PM2.5 using LSTM deep learning methods. *Sustainability* **2022**, *14*, 2068. [CrossRef]
44. Khattak, A.; Zhang, J.; Chan, P.-W.; Chen, F.; Almujibah, H. Explainable Boosting Machine: A Contemporary Glass-Box Strategy for the Assessment of Wind Shear Severity in the Runway Vicinity Based on the Doppler Light Detection and Ranging Data. *Atmosphere* **2023**, *15*, 20. [CrossRef]
45. Pujianto, U.; Wibawa, A.P.; Akbar, M.I. K-nearest neighbor (k-NN) based missing data imputation. In Proceedings of the 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 23–24 October 2019; pp. 83–88.
46. Zamani Joharestani, M.; Cao, C.; Ni, X.; Bashir, B.; Talebiesfandarani, S. PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* **2019**, *10*, 373. [CrossRef]
47. Ma, J.; Yu, Z.; Qu, Y.; Xu, J.; Cao, Y. Application of the XGBoost machine learning method in PM2.5 prediction: A case study of Shanghai. *Aerosol Air Qual. Res.* **2020**, *20*, 128–138. [CrossRef]
48. Zhong, J.; Zhang, X.; Gui, K.; Wang, Y.; Che, H.; Shen, X.; Zhang, L.; Zhang, Y.; Sun, J.; Zhang, W. Robust prediction of hourly PM2.5 from meteorological data using LightGBM. *Natl. Sci. Rev.* **2021**, *8*, nwaa307. [CrossRef]
49. Zhao, Y.; Hasan, Y.A. Comparison of three classification algorithms for predicting PM2.5 in Hong Kong rural area. *J. Asian Sci. Res.* **2013**, *3*, 715–728.
50. Zhao, R.; Gu, X.; Xue, B.; Zhang, J.; Ren, W. Short period PM2.5 prediction based on multivariate linear regression model. *PLoS ONE* **2018**, *13*, e0201011. [CrossRef]
51. Loshchilov, I.; Hutter, F. CMA-ES for hyperparameter optimization of deep neural networks. *arXiv* **2016**, arXiv:1604.07269.
52. Wei, Y.; Tian, X.; Huang, J.; Wang, Z.; Huang, B.; Liu, J.; Gao, J.; Liang, D.; Yu, H.; Feng, Y. New insights into the formation of ammonium nitrate from a physical and chemical level perspective. *Front. Environ. Sci. Eng.* **2023**, *17*, 137. [CrossRef]
53. Tanatachalert, T.; Jumlongkul, A. Correlation Between Relative Humidity and Particulate Matter During the Ongoing of Pandemic: A Systematic Review. *Aerosol Sci. Eng.* **2023**, *7*, 295–302. [CrossRef]
54. Ding, J.; Zhao, P.; Su, J.; Dong, Q.; Du, X.; Zhang, Y. Aerosol pH and its driving factors in Beijing. *Atmos. Chem. Phys.* **2019**, *19*, 7939–7954. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.