atmosphere

MDPI

Article

Ensemble Artificial Intelligence Fusing Satellite, Reanalysis, and Ground Observations for Improved $PM_{2.5}$ Prediction

Muhammad Haseeb [1,*], Zainab Tahir [1], Syed Amer Mehmood [1], Hania Arif [2], Sumaira Kousar [3], Sundas Ghafoor [4] and Khalid Mehmood [5,*]

[1] Institute of Space Science, University of the Punjab, Lahore 54780, Pakistan; zainabtahir244@gmail.com (Z.T.); amer.spsc@pu.edu.pk (S.A.M.)
[2] Department of City and Regional Planning, Lahore College for Women University, Lahore 54000, Pakistan; hania.arif@lcwu.edu.pk
[3] Institute of Geography, University of the Punjab, Lahore 54780, Pakistan; sumaira.geog@pu.edu.pk
[4] Centre for Geographical Information System, University of the Punjab, Lahore 54780, Pakistan; sundusghafoor275@gmail.com
[5] Aix-Marseille Univ, Avignon Univ, CNRS, IRD, IMBE, ITEM, 13545 Aix-en-Provence CEDEX 4, France
* Correspondence: haseeb.res.spsc@pu.edu.pk or mrhaseeb223@gmail.com (M.H.); khalid.mehmood@imbe.fr or sipra.khalid@yahoo.com (K.M.)

Abstract

Air pollution caused by fine particulate matter ($PM_{2.5}$) poses a serious public health threat in many South Asian megacities where monitoring networks remain limited. Lahore, Pakistan—frequently ranked among the world's most polluted cities—still lacks reliable short-term $PM_{2.5}$ forecasting systems. This study develops a performance-weighted ensemble machine learning framework that integrates satellite observations, meteorological reanalysis data, and ground monitoring measurements to improve daily $PM_{2.5}$ prediction. Eleven predictor variables were processed using a unified Google Earth Engine pipeline, including MODIS aerosol optical depth, Sentinel-5P trace gases (CO, NO_2, SO_2), and ERA5 meteorological parameters. Four tree-based machine learning algorithms—Random Forest, XGBoost, LightGBM, and CatBoost—were trained using daily observations from 2019 to 2023. Model evaluation using an independent 2024 dataset showed strong predictive capability, with Random Forest achieving R^2 = 0.77 (RMSE = 24.75 μg m^{-3}), XGBoost R^2 = 0.76 (RMSE = 26.32 μg m^{-3}), CatBoost R^2 = 0.73 (RMSE = 30.39 μg m^{-3}), and LightGBM R^2 = 0.70 (RMSE = 32.75 μg m^{-3}). To further enhance performance, the best models were combined into a weighted ensemble (RF 0.5, XGBoost 0.3, and CatBoost 0.2), which produced the highest validation accuracy (R^2 = 0.77; RMSE = 23.37 μg m^{-3}). Statistical testing using paired *t*-tests and Diebold–Mariano tests confirmed that the ensemble significantly reduced forecast errors compared with individual models. Feature importance analysis revealed that surface pressure, temperature, CO, and NO_2 were the most influential predictors of $PM_{2.5}$ variability. The proposed framework demonstrates that combining satellite data, reanalysis meteorology, and ground observations through ensemble learning can provide accurate and scalable air quality forecasting for data-limited urban environments.

Keywords: $PM_{2.5}$ prediction; air quality forecasting; machine learning models; ensemble forecasting; multisource data fusion techniques; Lahore; Pakistan

Academic Editors: Antonio Donateo and Xinghua Li

Received: 2 February 2026
Revised: 1 April 2026
Accepted: 15 April 2026
Published: 18 April 2026

1. Introduction

Fine particulate matter ($PM_{2.5}$), pollution is one of the biggest environmental health problems worldwide [1]. Fine particulate matter can be extremely small, so small that

it can travel down your airway and even into your bloodstream. Exposure to $PM_{2.5}$ has been widely linked to severe health outcomes, including cardiovascular diseases, respiratory illnesses, strokes, and premature mortality, as reported in several epidemiological studies [2–5]. According to the World Health Organization, exposure to fine particulate matter results in over seven million premature deaths each year. Therefore, it is one of the biggest environmental health risks [6]. Because of urbanization and the rapidly growing industries in developing countries, the need for accurate and timely forecasting of $PM_{2.5}$ has become critical for both protecting public health and developing effective environmental policies [6,7].

South Asia is one of the most polluted areas on Earth, and Lahore, Pakistan's second largest city, is one of the dirtiest places for air pollution [8]. Based on IQAir information, Lahore has been included in the list of the most polluted cities in the world many times, particularly in the winter months when the $PM_{2.5}$ concentration levels are over 300 $\mu g/m^3$, which is significantly higher than the WHO recommended levels (https://www.iqair.com/world-most-polluted-cities, accessed on 1 February 2026). The reasons for this level of pollution in Lahore include both local causes (e.g., traffic, industrial emissions, brick kiln smoke, and open burning) and regional causes (i.e., crop residue burning in the Punjab region), in addition to weather conditions (low wind speeds and temperature inversions) that allow the pollutants to remain close to the surface [9,10]. However, real-time air quality forecasts are still limited in Lahore because there are very few monitors and/or monitoring stations, poor data management systems, and a need for the development of locally based forecast models. Therefore, it becomes an important question: can $PM_{2.5}$ levels be accurately forecasted in a heavily polluted city with limited ground-based data?

Many researchers rely on machine learning (ML) and deep learning (DL) to forecast particulate matter concentrations due to its ability to capture complicated relationships between PM, weather, and emission patterns [11,12]. Tree-based and machine learning approaches have been increasingly applied for predicting $PM_{2.5}$ concentrations because they can capture complex nonlinear relationships between meteorological variables and atmospheric pollutants. Li et al. developed a satellite-based machine learning framework to retrieve ground-level $PM_{2.5}$ concentrations across China and showed that integrating meteorological and aerosol parameters can significantly improve spatial prediction accuracy [13]. Similarly, Masood et al. proposed a hybrid extreme learning machine model combined with a snake optimization algorithm for forecasting $PM_{2.5}$ concentrations in New Delhi and reported strong prediction performance with an R^2 value of 0.928 [14]. Alawi et al. conducted a comparative analysis of several artificial intelligence models for predicting air pollutants, including $PM_{2.5}$, and found that deep learning architectures such as LSTM and GRU achieved high prediction accuracy with R^2 values exceeding 0.96 [15]. However, most of these studies have focused on large metropolitan regions with dense monitoring networks. Consequently, the applicability of such approaches in data-scarce regions, particularly in smaller cities in South Asia such as Pakistan, remains insufficiently explored. However, the use of satellite remote sensing and reanalysis data now also supplies researchers with useful information regarding air quality and weather conditions. For example, sensors such as MODIS and Sentinel-5P supply researchers with data regarding aerosol optical depth (AOD) and gaseous compounds including NO_2, SO_2, CO, O_3, and HCHO; ERA5 provides researchers with detailed meteorological parameters such as temperature, wind, and surface pressure [16–19]. The availability of sensor data allows researchers to integrate satellite observations with machine learning methods for spatial prediction and environmental monitoring [20].

Most $PM_{2.5}$ studies conducted in Pakistan were carried out by applying basic statistical techniques or short-term modeling methodologies using relatively coarse data [21,22]. Most

of these studies did not verify their model predictions with data collected in a subsequent calendar year; therefore, it is uncertain as to whether those models will accurately predict $PM_{2.5}$ concentrations outside of their initial training time frame. Consequently, current methodologies are not optimal for real-time air quality forecasting applications in data-poor cities such as Lahore, Pakistan. While several regional studies have employed machine learning techniques for estimating $PM_{2.5}$ levels, many of those studies either employed a singular methodology, limited their analysis to a short time, or employed fixed data integration techniques, and provided little emphasis on replicating results or validating across years. Moreover, no previous research conducted in Pakistan has utilized a weighted ensemble framework combining satellite-derived data (MODIS and Sentinel-5P), reanalysis data (ERA5), and ground-based measurements for predicting $PM_{2.5}$ levels.

Therefore, this study developed a performance-weighted ensemble machine learning framework for predicting daily $PM_{2.5}$ levels in Lahore utilizing ground-based measurements of $PM_{2.5}$ and data derived from satellites (Sentinel-5P and MODIS) and reanalysis (ERA5). $PM_{2.5}$ was selected as the primary target variable because it is widely recognized as one of the most harmful air pollutants affecting human health. In addition, ground-based monitoring stations in the study area consistently record $PM_{2.5}$ observations, while other pollutants are not uniformly available across the monitoring network. The primary objective was not to introduce a novel algorithm but to provide a practical and reliable $PM_{2.5}$ forecasting methodology applicable to a data-scarce megacity. Models were trained using data from 2019 to 2023 and validated with an independent 2024 dataset to evaluate their ability to generalize over time. This framework improved the reliability of predictions and provides a reproducible and transferable workflow that can be adapted to similar data-scarce megacities in South Asia. Although this study focuses on $PM_{2.5}$ prediction, the proposed ensemble machine learning framework can also be extended to estimate other atmospheric pollutants such as NO_2, SO_2, and O_3 using similar environmental predictors.

2. Material and Methods

2.1. Study Area Description

With an estimated population of over 11 million residents, Lahore is the second largest city in Pakistan and its capital, located at the southern end of the Indus basin floodplain at 31.5497° N and 74.3436° E. It is a key industrial hub and plays a leading role in the economy of the country. As a result of rapid urbanization, large amounts of industrial activity and crop residue burning by neighboring districts, Lahore has become one of the most polluted cities in terms of $PM_{2.5}$ concentration. The city experiences frequent episodes of smog during the fall and early spring months when temperatures are cooler and air masses are stagnant. Daily average $PM_{2.5}$ concentrations are often greater than 300 $\mu g/m^3$, exceeding World Health Organization standards. The semi-arid subtropical climate that occurs in Lahore results in very warm summer months, a brief monsoon season and cold winter months. Additionally, the low wind speeds experienced during winter and temperature inversions create conditions that allow pollutants to be trapped near the ground surface of the city. Due to its proximity to the international border with India, the city of Lahore is exposed to air pollution originating from this region as well. $PM_{2.5}$ data collected by the United States Consulate, at the Jail Road monitoring station, are used in this research. The station's location in the middle of the city, in a heavily populated area of high traffic density, along with its relative distance to large-scale industrial emissions, creates a representative sampling environment of typical urban conditions (Figure 1). Satellite data from MODIS and Sentinel-5P are combined with ERA5 reanalysis to provide detailed information regarding the weather conditions and pollutant source regions influencing $PM_{2.5}$ concentrations in the city of Lahore.

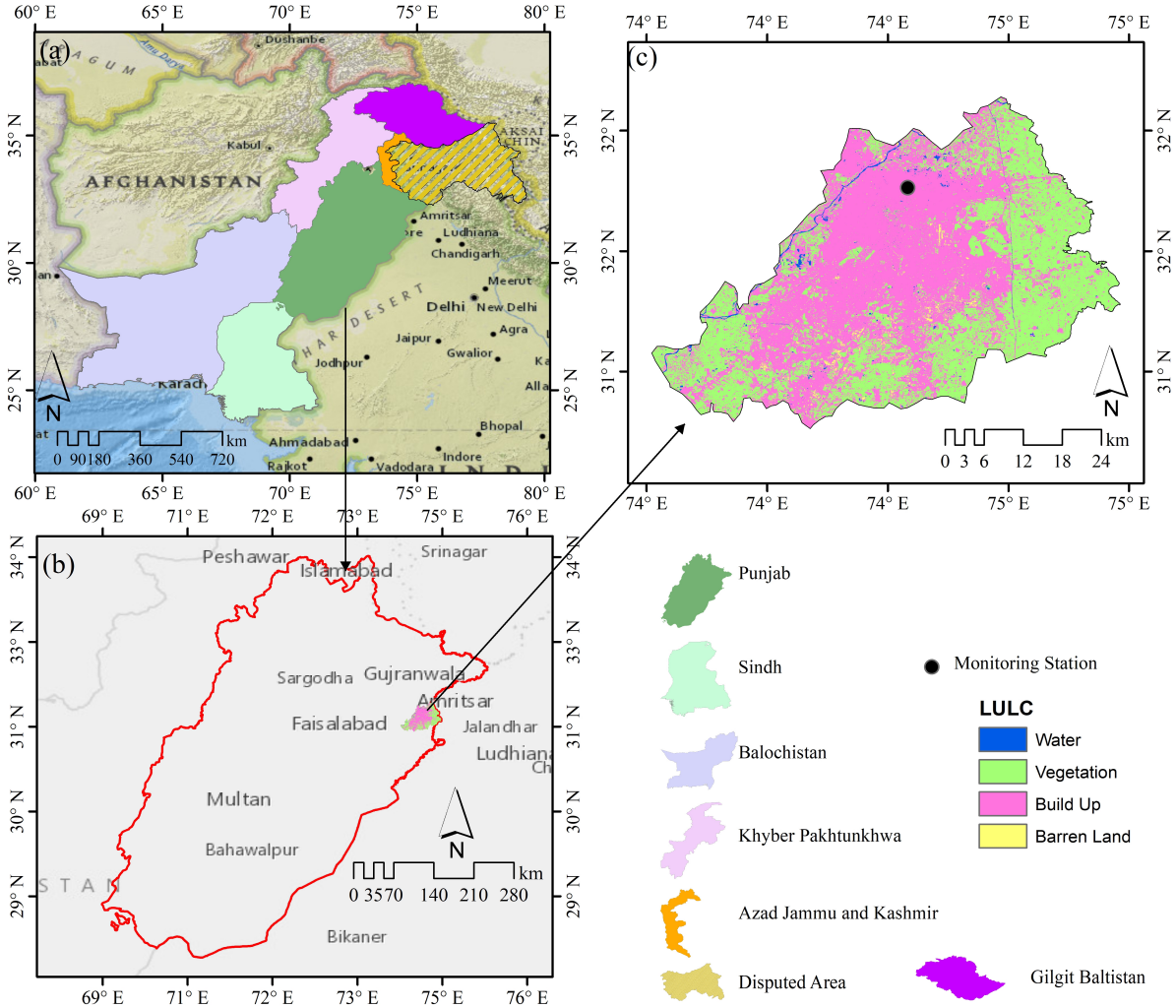

Figure 1. Geographical overview of the study area. (**a**) Map of Pakistan showing its administrative provinces, including Punjab (highlighted in green), Sindh, Baluchistan, Khyber Pakhtunkhwa, Azad Jammu and Kashmir, and the disputed area; (**b**) enlarged view of Punjab province with the study area—Lahore—indicated; (**c**) Land Use Land Cover (LULC) map of Lahore city classified into Water, Vegetation, Built-up, and Barren Land classes. The black dot represents the PM monitoring station used for air quality assessment.

2.2. Data Collection

This study develops and evaluates a robust machine learning framework to predict daily $PM_{2.5}$ concentrations over Lahore, Pakistan, using a carefully curated set of environmental, meteorological, and atmospheric variables. To ensure spatial consistency between satellite observations and the ground monitoring station, satellite-derived variables were extracted using a 10 km buffer around the station location. The average value of satellite pixels within this buffer was used to represent atmospheric conditions corresponding to the monitoring site. This approach accounts for the spatial resolution differences between satellite products and point-based ground observations and helps reduce spatial noise in satellite retrievals.

2.2.1. Dependent Variable (DV)

The core target variable in this study is the daily mean concentration of fine particulate matter ($PM_{2.5}$), expressed in micrograms per cubic meter ($\mu g/m^3$). In rapidly urbanizing regions such as Lahore, Pakistan, the accurate quantification and prediction of $PM_{2.5}$ are vital for mitigating public health risks, guiding environmental policy, and supporting

evidence-based urban planning [10]. To ensure high-fidelity ground-truth data for model training and validation, daily mean $PM_{2.5}$ concentrations were sourced from the U.S. Consulate General Air Quality Monitoring Station in Lahore, which operated a Federal Equivalent Method (FEM) sensor network from 2019 to 2024. This study has examined $PM_{2.5}$ as it varies throughout space, but it has developed a forecasting system based at one location (i.e., a single station), which does not include a spatial interpolation model for predicting the $PM_{2.5}$ patterns across the region. Single-location ML-based forecasting systems are commonly employed in atmospheric prediction. The goal is not to produce spatial maps of pollutant levels, but to provide forecasts of pollutant levels over time. Additionally, given that the U.S. Consulate is in an area of high-density urban activity and represents the "typical" or average exposure of an individual to $PM_{2.5}$ during their day, the use of a single monitoring station provides a reasonable representation of public health-related exposures. Although a single air quality monitoring station may be considered limited, the inclusion of ERA5 regional-scale meteorology and Sentinel-5P atmospheric chemistry data mitigate this limitation. In addition, other studies have demonstrated that temporal predictions of $PM_{2.5}$ can be produced using only a single air quality monitoring station as the input [23–25].

2.2.2. Independent Variables (IVs)

To forecast the amount of $PM_{2.5}$ in the atmosphere each day, we used 11 independent variables representing factors related to weather, atmospheric composition and aerosol properties that have been previously demonstrated to affect urban air quality [26,27]. The variables we chose were based on previous research, how relevant they were physically to the formation of $PM_{2.5}$, and if there was enough data available on them to be measured daily. The meteorological variables we used came from the ERA5 reanalysis database and included: air temperature at two meters; dew point temperature; surface pressure; and wind speed and direction. We derived our wind speed and direction from the U and V components of the wind vectors, as these describe how pollutants travel and disperse. The atmospheric composition was characterized using Sentinel-5P data and included NO_2, SO_2, CO, O_3, HCHO, all of which contribute to $PM_{2.5}$ through chemical reactions [28]. The variables we obtained from Sentinel-5P were scaled by 1e6 to allow us to keep the same number of decimal places during the training process of the models.

We characterized the aerosol properties using both the aerosol optical depth (AOD) from MODIS and the Aerosol Index (AI) from Sentinel-5P. Both variables give us some information about the quantity of aerosols in the atmosphere and what types of particles are present. All the variables were averaged for the district of Lahore and were matched to a daily time frame so that they could be compared to the hourly measurements of $PM_{2.5}$ made from ground-based instruments. By combining all the variables into one dataset, the machine learning models will be able to capture the interactions between the emissions, the weather and the atmospheric chemistry. Variables, abbreviations and units can be found in Table S1.

2.3. Data Preprocessing

Preprocessing of the dataset was essential for improving the performance of the machine learning algorithms. To apply robust data preprocessing to the $PM_{2.5}$ dataset used as an input to the model, a variety of preprocessing methods were employed; these included identifying missing values, identifying outliers, normalizing the data, and performing preliminary statistical evaluations. The original dataset contained eleven independent variables and one dependent variable ($PM_{2.5}$) and contained missing values due to temporary

satellite gaps and temporary sensor issues. The missing data were initially identified and documented in Table S2.

The missing values were subsequently replaced using a two-step process. First, short-term missing values were replaced by applying a linear interpolation to the existing data points based on their temporal relationship. Second, the missing values were also replaced using a correlation-based estimation procedure that utilized highly correlated variables to replace the missing data values. An example of this is the replacement of missing HCHO data values by estimating them from CO and O_3 data values. The reliability of this correlation-based estimation procedure was tested by removing 10% of the data and then replacing those data values using the correlation-based estimation procedure. This resulted in an excellent level of agreement (MAE = 2.9 and R^2 = 0.97). To evaluate if the data included any outliers, the data were evaluated using box plots (see Figure S1). It was determined that the high values of $PM_{2.5}$ and trace gases observed during the winter months were consistent with real-world air quality pollution episodes and therefore were not removed from the data. As the variables contained in the dataset have different measurement scales and ranges, the data were normalized to place all variables on the same measurement scale using a standard scaler [29,30]. A statistical summary of the preprocessed data is presented in Table S3 and feature distributions and seasonal patterns are illustrated in Figures S2 and S3. A check for multicollinearity was performed using both the Pearson correlation coefficient and the Variance Inflation Factor (VIF). The results indicated that there was no strong redundancy among the variables, and thus the preprocessing methods were applied to the independent 2024 validation dataset to maintain consistency with the 2019–2023 training data.

Wind speed and direction were not directly available as scalar or compass values but were derived from ERA5 U and V wind components. These were transformed into polar coordinates to compute wind magnitude and direction. Subsequently, annual wind rose diagrams (Figure S4) were created, illustrating the prevailing wind patterns over Lahore from 2019 to 2024. These revealed that airflow predominantly originated from the east and southeast, aligning with known meteorological behavior in the region. Wind parameters were included in the dataset to capture the role of atmospheric transport in $PM_{2.5}$ dispersion and accumulation.

2.4. Models' Development

The dataset was divided into training and testing subsets, while an independent dataset from 2024 was used for validation. To minimize the influence of temporal autocorrelation, the dataset was divided chronologically rather than randomly. $PM_{2.5}$ concentrations often exhibit temporal persistence due to meteorological conditions and emission patterns. Therefore, data from 2019 to 2023 were used for model training and testing, while the independent dataset from 2024 was reserved for validation. This temporal separation ensures that the validation dataset represents unseen time periods and reduces the risk of inflated model performance caused by temporally correlated observations.

2.4.1. Random Forest (RF)

Random Forest (RF) uses bagging and the use of a random subset of features at every split to construct multiple decision trees and then aggregate those predictions for improved accuracy and reduced overfitting. In addition, Breiman states that a single tree can be constructed from a randomly selected sample of the training set with replacement [31,32]. The focus in this study was to use RF for forecasting of daily $PM_{2.5}$ concentrations with satellite, meteorological and atmospheric chemistry variables. Hyperparameter tuning was done for three hyperparameters: (i) the number of trees to grow (n_estimators), (ii) the

maximum number of levels for the trees (max_depth), and (iii) the minimum number of examples in the child nodes for splitting (min_samples_split). A RandomizedSearchCV-based strategy was used to perform the tuning because it is much faster than a grid search [33]. A total of 100 sets of hyperparameters were tested in a 10-fold cross-validation setting with the goal of minimizing the average cross-validated root mean square error (RMSE) between predicted and actual values. This approach provided fast convergence to a solution while greatly reducing computational cost. Additionally, RF produces importance measures of the input features via two methods: Gini impurity reduction and permutation-based methods [34]. These important measures were further examined to determine the most influential predictors of $PM_{2.5}$.

2.4.2. Extreme Gradient Boosting (XGBoost)

Gradient-boosted decision trees are a family of machine learning algorithms that build upon each other to create very accurate predictions, as well as having strong regularization properties [35,36]. These models are sequential in nature, where each model is generated to correct the residuals from the previous model. The current research employed XGBoost to predict daily $PM_{2.5}$ concentrations using standardized meteorological/atmospheric datasets. Using randomized search optimization for hyperparameter tuning with 10-fold cross-validation, the hyperparameters n_estimators, max_depth, learning rate, subsample, and colsample_bytree were evaluated [33]. The L1/L2 regularization capabilities of XGBoost improve its ability to generate robust results even when the dataset contains outliers [37]. Feature importances based upon gain, cover and frequency metrics were used to evaluate the contribution of each predictor variable [38].

2.4.3. Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine (LightGBM)—created at Microsoft Research—is a fast and memory-efficient gradient boosting machine that utilizes histogram-based splitting and leaf-wise tree growth to speed up both the training of models and model predictions [39]. This makes it suitable for larger environmental and atmospheric datasets.

For this research project, we utilized LightGBM to predict daily $PM_{2.5}$ levels as they relate to meteorological and atmospheric predictor variables. Using Randomized Search Cross-Validation (RandomizedSearchCV) with a 10-fold cross-validation we optimized hyper-parameters for LightGBM (n_estimators, max_depth, learning rate, Num leaves, subsample, and colsample_bytree). As an additional measure to prevent overfitting related to the use of leaf-wise tree growth [40], we constrained the maximum depth and minimum leaf data requirements for LightGBM.

2.4.4. CatBoost

CatBoost is a fast gradient-boosting algorithm, built by Yandex, which provides the advantages of ordered boosting and more powerful regularization techniques to enhance performance and reduce overfitting [40,41]. CatBoost was created specifically for use with datasets having many categorical fields; however, CatBoost will still perform well on completely numerical datasets as it has shown itself to be very resistant to noise and can build complex nonlinear models.

This study applied CatBoost to forecast daily levels of $PM_{2.5}$ in relation to a variety of continuous meteorological and atmospheric predictors. The hyper-parameters of CatBoost were randomly searched through the application of 10-fold cross-validation to identify optimal settings for the number of iterations, the depth of each tree, learning rate and l2 leaf regularization. The ordered boosting capability of CatBoost showed greater generalization than traditional boosting algorithms, allowing CatBoost to be effective for environmental applications including forecasting of air quality and climate variables [42,43].

2.5. Ensemble Learning Strategy

To improve the predictive accuracy of $PM_{2.5}$, we employed an Ensemble Learning Methodology. This methodology combined three machine learning models: RF, XGBoost and CatBoost. We assigned weights to each model based on their validation performance, i.e., those models that produced lower prediction errors have larger influences on the final prediction. The weights were derived from the inverse of the validation root mean square error (RMSE) and were scaled to 0.50 for RF; 0.30 for XGBoost; and 0.20 for CatBoost. LightGBM was excluded from the ensemble as it demonstrated larger prediction errors when compared to the other three models. The ensemble model outperformed each of the individual models on the independent 2024 dataset: achieving the largest R^2 value and smallest normalized root mean square error (RMSE). The $PM_{2.5}$ forecasting framework is presented in Figure 2. The left panel provides an overview of how we integrated data from ERA5, MODIS, Sentinel-5P and ground stations into training the four tree-based models and the weighted ensemble. The right panel outlines the statistical processing of the data, including the imputation of missing values, evaluation of outliers, normalization and the calculation of the wind parameter. Additionally, we used paired *t*-tests and the Diebold–Mariano test to statistically confirm the superiority of the ensemble model over the individual models using the daily prediction errors from 2024. The paired *t*-tests and Diebold–Mariano tests evaluated if there were statistically significant differences between the predictions of the ensemble model versus each of the individual models.

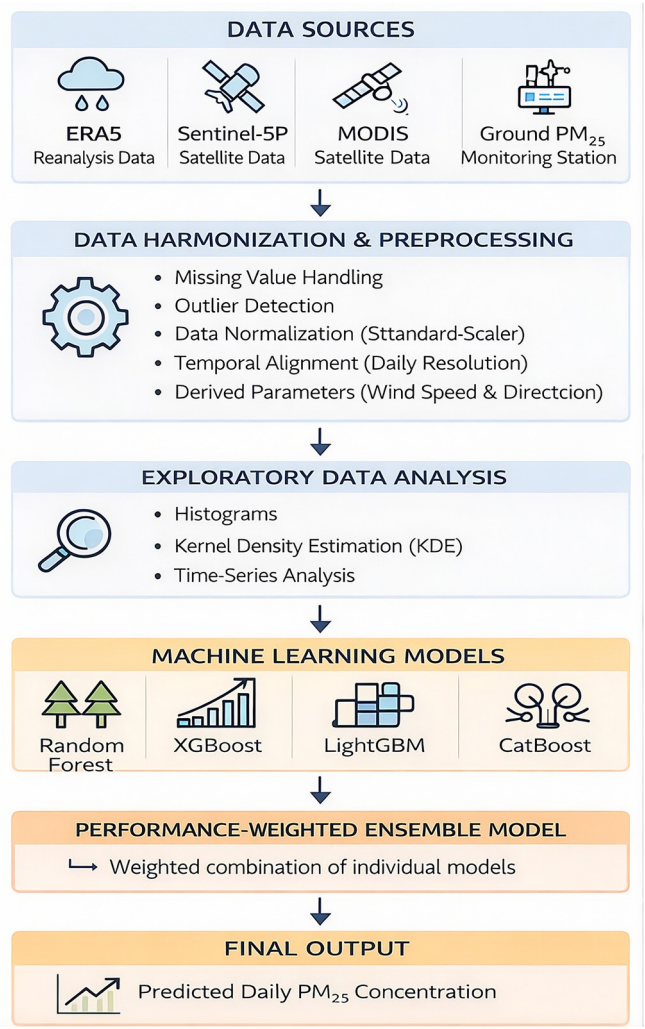

Figure 2. Machine learning workflow for daily $PM_{2.5}$ prediction over Lahore.

2.6. Evaluation of Models

To evaluate the quality of the models developed, we employed a time-based holdout method. The years 2019–2023 were used as training years, while the independent 2024 data were used as a validation dataset. We followed the same process of aggregation, interpolation, and normalization for processing the 2024 data, as was done for the training data, to ensure that the 2024 data would be consistent. Thus, this time-based split allowed us to test how well our models generalized to a new year with different meteorological and emission characteristics. To avoid the issue of overfitting, we implemented 10-fold cross-validation during the training phase by dividing the dataset into 10 parts and then running 1 validation cycle on each portion of the dataset For the boosting methods (CatBoost, XGBoost, and LightGBM), we used regularization and early stopping to limit the complexity of the models and prevent them from becoming overly complex during training.

We selected three common metrics to evaluate the quality of the models: R^2 (coefficient of determination), RMSE (root mean square error), and normalized RMSE (%) [44]. R^2 measures the amount of variability in $PM_{2.5}$ that the model can explain; RMSE (mean squared error in $\mu g/m^3$) is an indicator of the average difference in $\mu g/m^3$ between actual measurements and predictions made by the model, and it is particularly useful when evaluating performance under high pollution events [45,46]; and RMSE (%) is the percent of the average $PM_{2.5}$ concentrations in $\mu g/m^3$ for all the measured data points. Lower values of RMSE (%) and nRMSE are indicative of higher model accuracy. We computed these metrics for both the individual models and the ensemble models using the 2024 validation dataset to determine which of the $PM_{2.5}$ forecasting approaches would provide the best results.

2.7. Software and Tools

All geospatial data processing and analysis were conducted using Google Earth Engine (Google Inc., Mountain View, CA, USA). The study area map and cartographic visualizations were prepared using ArcGIS Desktop 10.8.1 (Esri, Redlands, CA, USA). Machine learning modeling and statistical analysis were performed using Python 3.12 (Python Software Foundation, Wilmington, DE, USA), with libraries including Scikit-learn, XGBoost, LightGBM, and CatBoost. Data preprocessing, visualization, and evaluation were carried out using standard Python-based scientific computing tools.

3. Results

3.1. Correlation Analysis

We used correlation analysis to evaluate the connections between the selected environmental parameters and $PM_{2.5}$, as well as to eliminate duplicate predictor variables. This step served as an exploratory analysis to understand the relationships among variables and to avoid redundancy before developing the machine learning models. We generated a Pearson correlation matrix (Figure 3) to identify the magnitude and direction of the linear relationships among the selected environmental parameters. Surface pressure showed the highest positive correlation ($r = 0.60$) to $PM_{2.5}$; thus, it indicates that during periods of stable high pressure, there is a reduction in vertical mixing of the atmosphere which results in an accumulation of pollutants. The authors found a moderate-to-strong positive correlation ($r = 0.52$) between CO and $PM_{2.5}$; therefore, we infer that CO has common sources of emissions (e.g., vehicles and industrial activities) like $PM_{2.5}$. Moderate correlations ($r = 0.36$) were observed between sulfur dioxide (SO_2) and $PM_{2.5}$, and between nitrogen dioxide (NO_2) and $PM_{2.5}$. These gas species contribute to secondary particle formation. Negative correlations ($r = -0.57$) were identified for air temperature and dew point temperature; thus, these temperature factors suggest that increased temperatures (and higher dew points)

result in greater air mass mixing and removal of $PM_{2.5}$. Negative correlations (r = −0.22) were found between wind speed and $PM_{2.5}$; thus, wind speed contributes to pollutant dispersion. Weak correlations were found between ozone (O_3), formaldehyde (HCHO), aerosol optical depth (AOD) and $PM_{2.5}$; hence, the authors inferred that O_3, HCHO, and AOD have indirect or more complicated connections to $PM_{2.5}$ concentrations.

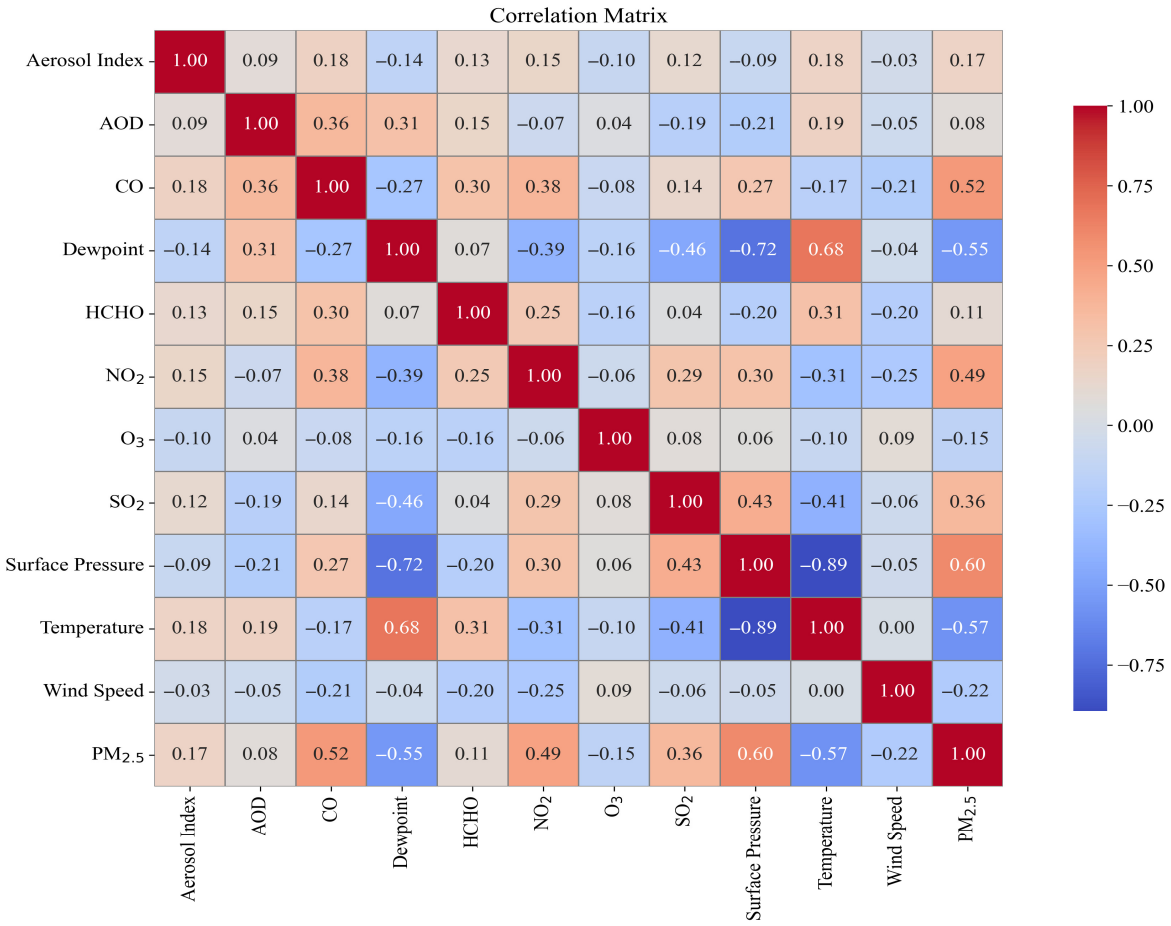

Figure 3. Pearson correlation matrix of $PM_{2.5}$ and associated predictor variables from 2019 to 2024.

To assess multicollinearity, we applied the VIF; most of the VIFs were less than five (Table S4). Thus, all the selected environmental parameters have very low redundancy. However, temperature (VIF = 6.81) and surface pressure (VIF = 6.47) slightly exceeded the threshold value of five; however, they were both included because of their important roles in controlling atmospheric instability and pollutant accumulation. Therefore, the authors conclude that the selected variables are meaningful, largely independent, and suitable for machine learning modeling.

3.2. Model Performance

3.2.1. Random Forest (RF)

The RF model performed well after we optimized the hyperparameters by selecting 500 trees and a Min Split Size = 2. The RF model fitted extremely well to our training data, as indicated by an R^2 value of 0.95 and an RMSE of 13.68. When applied to the testing data, the RF model performed well again with an R^2 of 0.79 and a RMSE of 21.26, demonstrating the ability to generalize the model over time. We used scatter plots (see Figure 4) to compare the predictions from the RF model to the observed $PM_{2.5}$ values. We see that the predicted $PM_{2.5}$ values are generally in close agreement with observed values and follow the 1:1 line for both the testing and training data. However, there is some slight

underestimation apparent when the pollution levels are extremely high and the model has difficulty capturing such extreme pollution levels. Time-series line plots (see Figure S5) provide additional evidence supporting our findings. For our training data, we observe a nearly perfect alignment between observed and predicted $PM_{2.5}$ values. However, for the testing data we find more variability, although the temporal patterns are captured. Overall, the RF model produces stable and accurate $PM_{2.5}$ predictions.

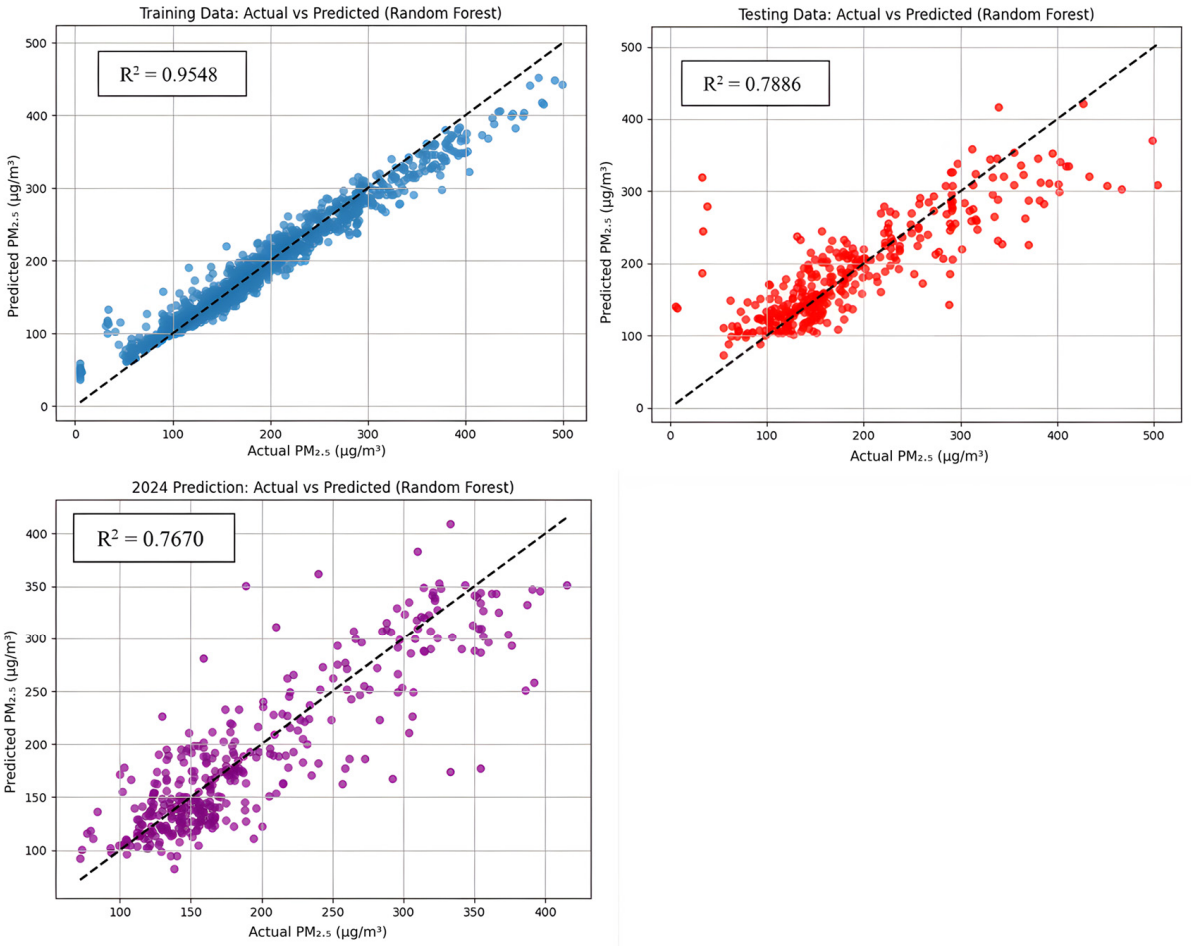

Figure 4. Actual vs. predicted $PM_{2.5}$ (scatter plots) concentrations using the Random Forest (RF) model during training, testing, and validation phases.

The above-trained RF was additionally evaluated using the daily $PM_{2.5}$ dataset from 2024, which had not been employed in the training process. The results were very positive, with an R^2 = 0.77 and RMSE = 24.75. The scatter plot depicted in Figure 4 is a good representation of how well the predicted values corresponded to the measured values, whereas the time-series depiction in Figure S5 indicates that the model can predict the seasonal variations in $PM_{2.5}$ relatively accurately. As such, this demonstrates that the RF model has the capability to make reliable predictions of $PM_{2.5}$ after the training period. The results of feature importance, as determined by the RF model, are depicted in Figure S6. Surface pressure was identified as the most important input variable, followed by temperature, CO, NO_2, and dew point temperature. Additional input variables, including SO_2, AOD, HCHO and wind speed, all contributed to the model's ability to accurately forecast $PM_{2.5}$ levels in Lahore.

3.2.2. Extreme Gradient Boosting (XGBoost)

The XGBoost model demonstrated a high level of performance with the use of 900 decision trees as a basis for building the model; a learning rate of 0.01 was used during training; the model was trained to a maximum depth of seven; and both the sample size and the number of columns sampled were set at 60 percent or less. During training, the model achieved an R-squared of 0.98 and a root mean squared error of 10.87, demonstrating a strong capability to identify the complex relationships within the data. Upon being applied to the test dataset, the model achieved an R-squared of 0.78 and an RMSE of 23.84, which demonstrates the model's ability to make accurate predictions. The scatter plots shown in Figure 5 illustrate that there is a strong correlation between observed and predicted values for $PM_{2.5}$, particularly for low-to-moderate pollution levels. The time-series plots shown in Figure S5 demonstrate that the model accurately predicts the changes in $PM_{2.5}$ on a daily and seasonal basis. When validated against the independent 2024 dataset, the model produced an R-squared of 0.76 and an RMSE of 26.32. Overall, these results indicate that the XGBoost model has a high degree of generality and can be used to predict $PM_{2.5}$ in complex urban environments over extended periods of time.

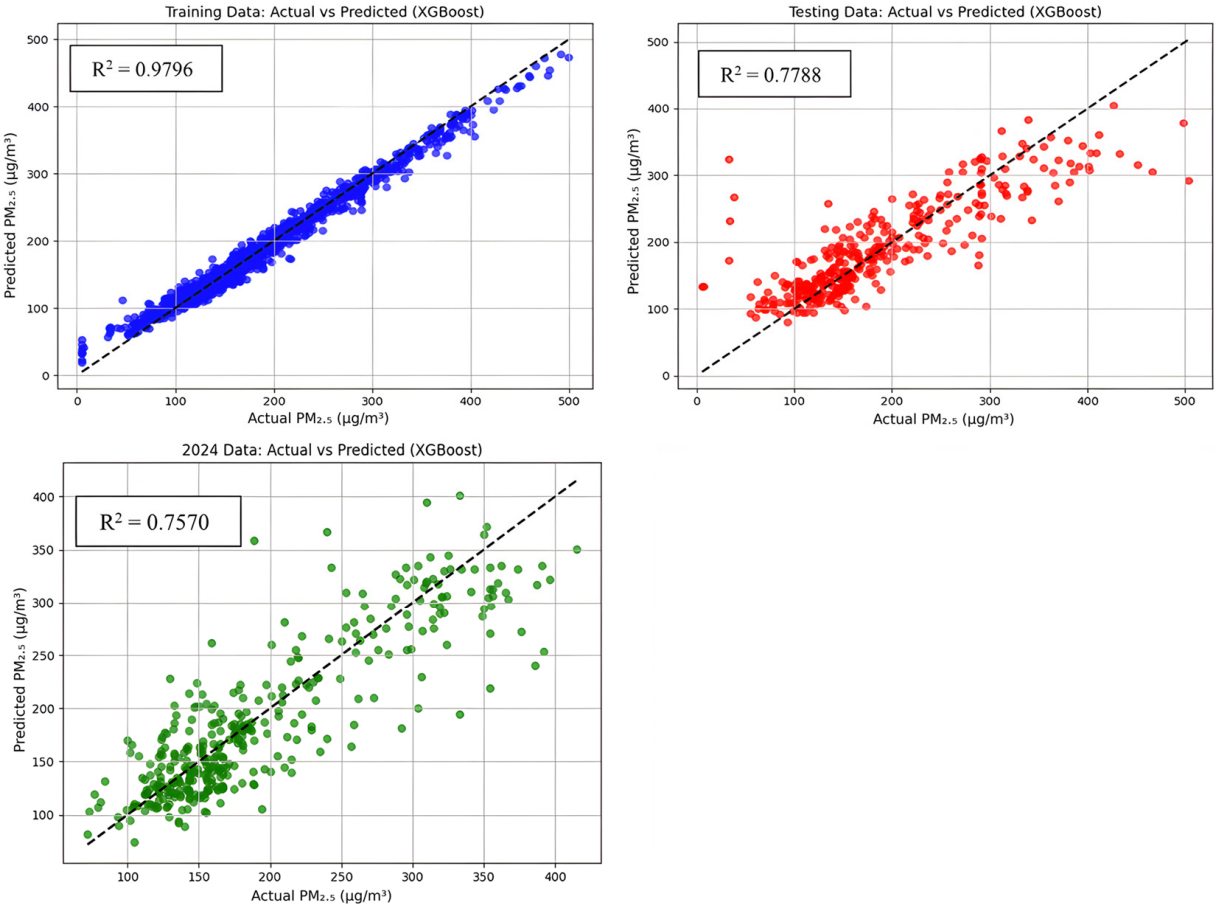

Figure 5. Performance of the XGBoost model on training, testing, and validation datasets: scatter plot of actual vs. predicted $PM_{2.5}$ concentrations.

3.2.3. Light Gradient Boosting Machine (LGBM)

The LightGBM model was successful once it had been tuned with the use of 500 trees, a 0.05 learning rate, a maximum tree depth of seven, and reasonable sampling rates. The model's performance on the training set produced an R^2 value of 0.98 and a low RMSE value of 10.35 $\mu g/m^3$. This indicates that the model produces very accurate predictions of $PM_{2.5}$ concentrations in comparison to actual values observed in the environment. Training

scatter and line plots (Figure 6) illustrate how the model's predictions closely follow the 1:1 line and also capture observed trends in $PM_{2.5}$ concentration over time. When evaluated against the test data, the model demonstrated good but lower performance than the training dataset. Specifically, the model produced an R^2 value of 0.75 and an RMSE value of 26.94 $\mu g/m^3$. Scatter plots demonstrate greater dispersion of the model's predictions at higher levels of concentration, although the general trend in $PM_{2.5}$ concentrations is still apparent (Figure S5). When tested using the 2024 dataset, the model produced an R^2 value of 0.70 and an RMSE value of 32.75 $\mu g/m^3$. Although the model's performance did decrease when applied to this previously unseen period, the model still illustrated the major seasonal trends in $PM_{2.5}$ concentrations, as well as the timing of major peak pollution events (Figure 6), demonstrating the model's utility for predicting concentrations.

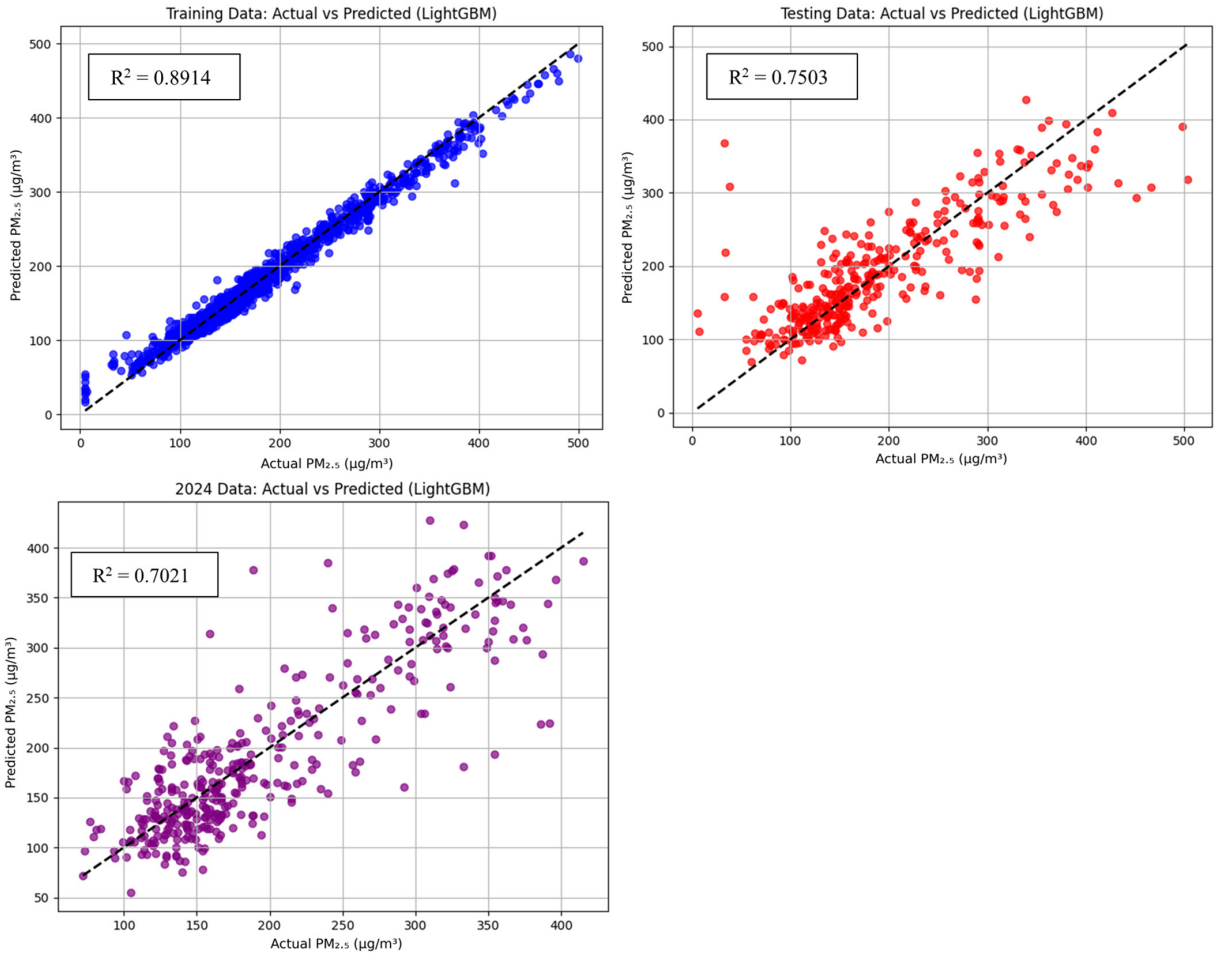

Figure 6. Scatter plots of actual vs. predicted $PM_{2.5}$ concentrations for the LGBM model during training, testing, and validation phases.

3.2.4. CatBoost

CatBoost also showed accuracy in both its training and tuning stages. During the training phase of the model, it produced a very high R^2 of 0.97 and very low RMSE of 11.82 $\mu g/m^3$ on the training set, which indicates that there was a good correlation between the actual $PM_{2.5}$ values that were measured, and the values that the model predicted (see Figure 7). It appears as though the model correctly identified how different atmospheric conditions can affect $PM_{2.5}$ concentrations. When the CatBoost model was tested against the test set, the model continued to perform at a relatively high level with an R^2 of 0.76 and an RMSE of 24.23 $\mu g/m^3$. Scatter and time-series plots are presented in Figures 7 and S5 and illustrate how the predicted $PM_{2.5}$ values generally tracked the observed values near

one another, with most points clustered along the 1:1 line. When this model was run against the independent 2024 test set, it produced an R^2 of 0.73 and an RMSE of 30.39 μg/m^3, which supports the idea that CatBoost will generalize well to new datasets, making it suitable for use in predicting future $PM_{2.5}$ concentrations.

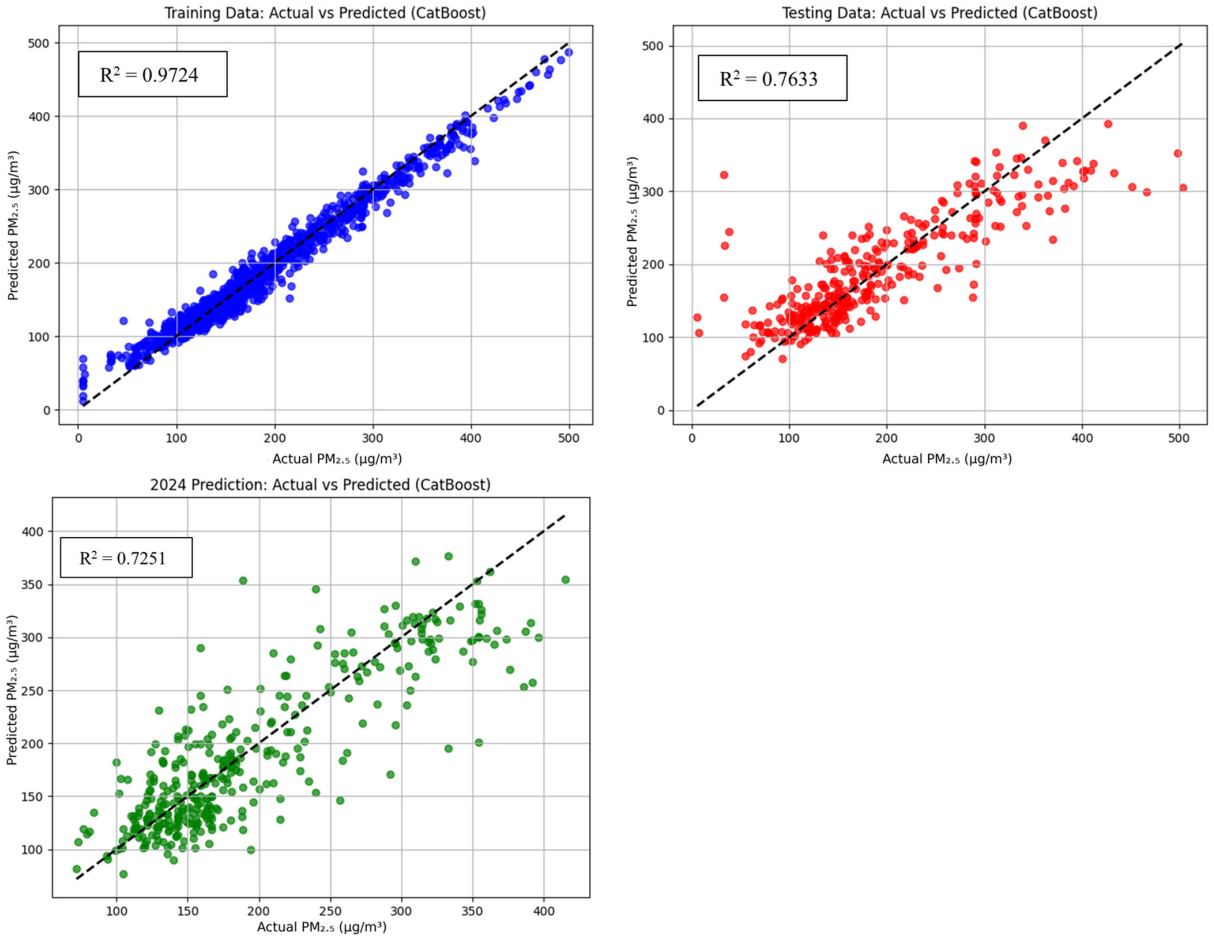

Figure 7. Scatter plots of actual vs. predicted $PM_{2.5}$ concentrations for the CatBoost model during training, testing, and validation phases.

3.2.5. Weighted Super Ensemble Model

To further enhance prediction accuracy, a weighted super ensemble model was constructed by integrating the outputs of four individually trained models: RF, XGBoost, LGBM, and CatBoost. The ensemble was strategically designed using a linear combination of predictions from these base learners, where the weights were empirically optimized based on their individual generalization strengths on test data. Specifically, the final prediction for each observation was calculated using the following weighted average:

$$\text{Ensembled} = (0.5 \times \text{RF}) + (0.3 \times \text{XGBoost}) + (0.2 \times \text{CatBoost})$$

The LGBM model was excluded from the weighted scheme due to its comparatively lower standalone performance in 2024 validation (R^2 = 0.7021), as it marginally reduced the ensemble's accuracy in initial trials. Because the 2024 dataset followed the same preprocessing and standardization pipeline as the training years, the resulting validation metrics directly reflect genuine temporal generalization rather than preprocessing differences. Upon evaluating the ensemble on the independent 2024 dataset, the model achieved a significantly improved R^2 score of 0.7728 and RMSE of 23.37 μg/m^3. This performance

surpasses the predictive capabilities of each standalone model, highlighting the effectiveness of intelligent model fusion. The notable increase in R^2 and reduction in RMSE indicate the ensemble's enhanced generalization to unseen, temporally out-of-sample data.

The scatter plot (Figure 8) reveals a tight clustering of predictions around the 1:1 reference line, affirming high agreement between predicted and actual $PM_{2.5}$ values. Additionally, the time-series line plot illustrates the ensemble's capacity to accurately capture both seasonal variations and short-term fluctuations in particulate matter concentrations throughout 2024. Overall, this weighted ensemble methodology showcases a robust and interpretable strategy for $PM_{2.5}$ forecasting. By leveraging the complementary strengths of tree-based learners and mitigating their individual weaknesses, the model provides a resilient forecasting solution tailored to dynamic urban environments such as Lahore. A comparison of the performance of each model individually to the ensemble model was conducted using paired *t*-tests and Diebold–Mariano (DM) forecast error testing on the 2024 independent test data (Table S5). The DM tests indicated that the ensemble had a statistically significantly lower mean absolute percentage error than both CatBoost (−3.19) and XGBoost (−2.63); it also outperformed Random Forest (−2.12) in terms of the DM statistics. However, the *t*-test for differences between ensemble and RF did not show significance (*p*-value = 0.15). Therefore, these results indicate that the ensemble has more consistent and lower variance output values compared to the individual models, especially through reduction in high-error spikes from day to day, which is supported by the statistical evidence of the ensemble being the most reliable and operationally robust model examined.

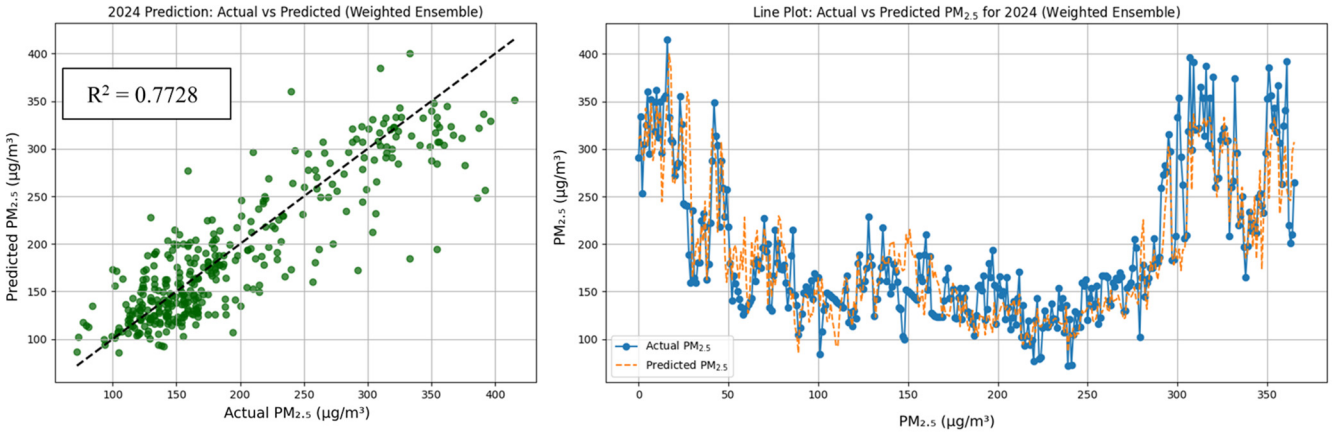

Figure 8. Scatter plot and time-series comparison of actual versus predicted daily $PM_{2.5}$ concentrations for the year 2024 using the weighted super ensemble model (RF + XGBoost + CatBoost).

Independent validations for the year 2024 were performed by evaluating the uncertainty structure of the ensemble model using residual diagnostics. The residual distribution is shown in Figure S7. X is approximately symmetrical and has a Gaussian-like shape about zero. This suggests there is little to no systematic error in the predictions made from the ensemble model. There is no funneling, nor are there any discernible patterns in the residual versus predicted plot (Figure S7). Therefore, it can be concluded that the error in the ensemble model is homogeneous, or the same, at all levels of concentration, including low, moderate and high $PM_{2.5}$ concentrations. In addition, the Q-Q plot (Figure S7) demonstrates that the residuals of the ensemble model are nearly perfectly aligned with the theoretical normal line. Together, the results of these diagnostic tests show that the prediction errors of the ensemble model have been stable over time, the ensemble model has demonstrated strong generalization in time, and the ensemble model is operating reliably.

3.3. Model Performance Comparison

The performance of each model was evaluated at three levels of analysis, training, testing and validation, to assess how accurately the models predicted and generalized to $PM_{2.5}$ levels. The evaluation of model performance at each level of analysis is illustrated in both Figure 9 and summarized in Table 1. The models performed well during the training phase, as evidenced by their high R^2 values. The model with the greatest accuracy during the training phase was LightGBM (R^2 = 0.98), followed closely by XGBoost (R^2 = 0.98) and CatBoost (R^2 = 0.97). Random Forest also had a high R^2 value (0.95). All the models had relatively low RMSE values, particularly LightGBM (RMSE = 10.35 μg/m^3) for the training phase. This indicates that each model can effectively match the training data.

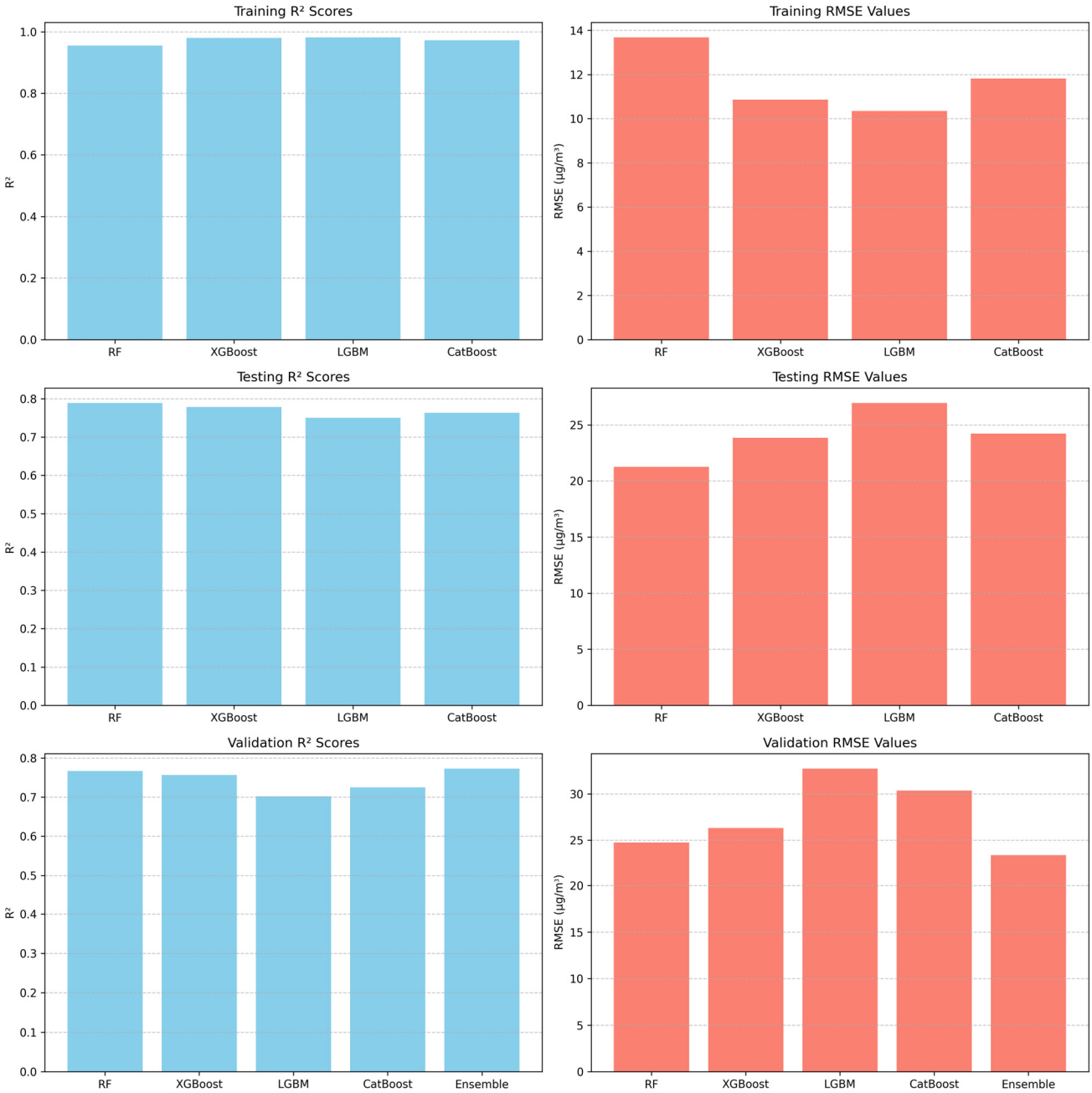

Figure 9. Comparative model performance across training, testing, and validation phases. Left column panels depict R^2 scores, while right column panels display RMSE (μg/m^3) values.

Table 1. Performance of machine learning models during training and testing phases.

Model	Phase	R^2	RMSE (μg/m^3)	nRMSE (%)
Random Forest	Training	0.9548	13.68	7.50%
	Testing	0.7886	21.26	11.66%
XGBoost	Training	0.9796	10.87	5.96%
	Testing	0.7788	23.84	13.07%
LGBM	Training	0.9814	10.35	5.67%
	Testing	0.7503	26.94	14.77%
CatBoost	Training	0.9724	11.82	6.48%
	Testing	0.7633	24.23	13.28%

Performance declined slightly in the testing phase as the models are used on unseen data, which is expected behavior. Random Forest had the highest R^2 value (0.79) and lowest RMSE value (21.26 μg/m^3) among the four models tested. Both XGBoost and CatBoost also performed well; however, LightGBM' s performance was significantly lower than the other two models as it had a higher RMSE (26.94 μg/m^3), indicating that LightGBM may have experienced some degree of overfitting.

Table 2 presents the validation performance of the models using the independent 2024 dataset. As such, among the individual models, Random Forest once again produced the highest R^2 (0.7670) and lowest RMSE (24.75 μg/m^3) for the validation phase. The weighted super ensemble achieved the highest predictive accuracy, with R^2 = 0.7728, RMSE = 23.37 μg/m^3, and nRMSE = 12.11%, outperforming the individual machine learning models. This demonstrates that the combination of Random Forest, XGBoost and CatBoost models can improve the accuracy of predictions by mitigating the deficiencies of each individual model. It is apparent that the generalization capabilities of LightGBM are the poorest (R^2 = 0.7021; RMSE = 32.75 μg/m^3) of the models tested; thus, it will be excluded from the ensemble. The ability of the ensemble model to perform better than the individual models represent a benefit of combining multiple models to provide reliable $PM_{2.5}$ forecasts.

Table 2. Independent validation performance of individual models and the final ensemble using the 2024 dataset.

Model	R^2	RMSE (μg/m^3)	nRMSE (%)	Remarks
Random Forest (RF)	0.7670	24.75	12.82%	Strong generalization; baseline performer
XGBoost	0.7570	26.32	13.63%	Robust gradient boosting learner
LightGBM (LGBM)	0.7021	32.75	16.97%	Overfitting observed; excluded from ensemble
CatBoost	0.7251	30.39	15.74%	Stable performance across years
Weighted Super Ensemble	0.7728	23.37	12.11%	Best accuracy; combines RF + XGBoost + CatBoost

4. Discussion

4.1. Model Performance Overview

The ML models were able to accurately forecast daily $PM_{2.5}$ concentrations in Lahore; however, the variation in their ability to generalize was significant. Individual learner RF had the most consistent and reliable performance in terms of testing accuracy (R^2 ~ 0.79; RMSE ~ 21 μg/m^3), as well as independent validation, in 2024 (R^2 = 0.767). The structural nature of RF as a bagging-based model seems to provide a favorable bias–variance trade-off,

allowing for stable performance and avoiding overfitting. The GBM models (XGBoost, CatBoost, and LightGBM) also had competitive performance and achieved testing R^2 values of ~0.78, ~0.76, and ~0.75, respectively. However, the boosting models, especially LGBM, exhibited larger variances in performance between training and testing and this is indicative of their propensity to overfit when there is noisy data and/or limited temporal variability in the input data. Nonetheless, the GBMs' sensitivity to complex nonlinear relationships did improve performance during extreme pollution events. The performance-weighted ensemble model ultimately outperformed each of the base learners. Through optimal weighting of RF, XGBoost, and CatBoost (0.5, 0.3, 0.2), the ensemble achieved an independent year accuracy (R^2 = 0.7728; RMSE = 23.37 $\mu g/m^3$) greater than each of the individual models. Although the 2024 dataset followed the same preprocessing pipeline as the training period, confirming temporal comparability, its single-station nature limits spatial generalization. This improvement was due to the ensemble's ability to capitalize on the stability of RF while leveraging the fine-grained pattern detection capabilities of the boosting models. Additionally, ensemble forecasts were found to be closer aligned with actual values and better captured both seasonal trends in pollution and extreme pollution peaks. In summary, the results demonstrate the potential of ensemble methods to enhance the generalizability of temporal and spatial aspects of $PM_{2.5}$ forecasting in South Asian cities like Lahore which are characterized by limited data availability.

4.2. Comparison with Existing Studies

Our study's predictive performance is consistent with other studies on $PM_{2.5}$ forecasting in South and East Asia. The R^2 values for daily $PM_{2.5}$ forecasts obtained with machine learning methods are generally around 0.7 and 0.85, depending on data quality and the complexity of the used model [47]. In fact, numerous studies in the region have been able to produce similar results using Random Forest, XGBoost or hybrid deep learning framework data [48–50] and they were able to do so thanks to the high density of the existing monitoring networks and high-resolution meteorological data available for these areas.

However, most previous studies conducted in Pakistan have focused on the analysis of correlations or short-term time-series models based on coarse inputs [21]. With the exception of very recent studies that rely on satellites (e.g., P-CNN models using Sentinel-5P data [22]), which significantly improve the spatial resolution but remain limited due to the lack of ground observation data, our ensemble model demonstrates an exceptionally high level of accuracy (R^2 = 0.7728; RMSE = 23.37 $\mu g/m^3$) considering that we use a single ground station and combine MODIS AOD, Sentinel-5P trace gases and ERA5 meteorological data.

Finally, comparing our results to those of larger regional and global studies confirms that our approach is competitive. Indeed, large-scale models in China and the United States (that integrate chemical transport models, large numbers of monitoring stations and sophisticated ML ensembles) generally provide R^2 values between 0.80 and 0.90 [51,52]. These studies are obviously better equipped to collect rich observational datasets than our study, and yet our single-city ensemble achieves comparable accuracy levels, which underlines the potential of the data fusion and optimal weighting of ensemble members for $PM_{2.5}$ forecasting in South Asian megacities.

More broadly, our results support that machine learning can also be used to obtain reliable daily $PM_{2.5}$ predictions when combining satellite products and meteorological reanalysis data at high spatio-temporal resolutions, even in the case of a very limited number of monitoring stations. Future research should therefore develop and apply similar ensemble models to several cities in Pakistan and/or in South Asia to evaluate the transferability of our approach over different urban environments.

4.3. Strengths and Limitations of Models

When each of the three algorithms is used to forecast $PM_{2.5}$ concentrations in a city as geographically diverse as Lahore, Pakistan, they have differing strengths and limitations. Random Forest (RF) produced the most stable and reliable results as it has the strongest generalization characteristics on both the test data and independent validation data. RF's use of a bagging-based structure to reduce variance and prevent overfitting makes RF particularly effective in modeling daily variations in pollution. However, one well-documented weakness of RF is its tendency to smooth extreme values, which is reflected in the model's underestimation of some of the worst smog episodes.

The two boosting models—XGBoost and CatBoost—had the best performance during training and therefore demonstrate the best recognition of patterns of complex nonlinear relationships. Their ability to recognize interactions between weather and atmospheric conditions also allows them to identify peak pollution episodes more frequently than RF. However, these models are more susceptible to overfitting than RF and therefore have a slightly higher generalization error. LightGBM (LGBM), although theoretically like XGBoost, has a larger gap in performance between training and validation than either XGBoost or CatBoost. LGBM's leaf-wise tree growth strategy likely contributes to the overfitting that occurs in this relatively small dataset, thus reducing its reliability for independent prediction. As a result, LGBM made minimal contributions to the final ensemble.

The ensemble model takes advantage of the unique strengths of RF, XGBoost and CatBoost while minimizing their individual weaknesses. The weighted combination (RF = 50%, XGB = 30%, and CatBoost = 20%) produces the highest accuracy and most consistent performance in terms of different pollution levels. Although the ensemble provides a moderate increase in accuracy, it demonstrates the benefits of using multiple models to predict the same outcome by providing a measure of diversity among the model errors. Therefore, this study provides additional evidence of the benefits of multi-model fusion for operational forecasting. There are still many limitations that must be addressed before this model can be considered fully validated. For example, the model is trained on data from a single monitoring station and therefore lacks the ability to accurately model pollutant concentrations at other locations. In addition, coarse-resolution satellite and reanalysis inputs do not capture local-scale pollutant dynamics, which are often important in cities like Lahore.

Furthermore, the evaluation of satellite-derived variables was conducted using a fixed 10 km buffer around the ground monitoring station to maximize spatial coverage and ensure sufficient pixel representation. While this approach improves data availability, it may introduce spatial averaging errors and reduce the ability to capture fine-scale variability in pollutant concentrations. This limitation is particularly important in heterogeneous urban environments such as Lahore, where air pollution can vary significantly over short distances. Additionally, the analysis relies on data from a single monitoring station, which further constrains the spatial representativeness of the results. Future studies will focus on incorporating multiple ground stations and exploring adaptive or smaller buffer sizes to better capture localized pollution dynamics and improve model robustness.

Finally, key emission-related predictors (i.e., traffic volume and industrial activity) were unavailable and we could not capture short-term emission spikes. The model's ability to predict future pollutant concentrations is limited by its reliance on the input of concurrent pollutant concentrations (i.e., NO_2, SO_2, and CO). Therefore, the model cannot be used as a standalone predictor of future pollutant concentrations until it is paired with an output from a module that predicts pollutant concentrations or a chemical transport model. Nonetheless, the model passed independent validation in 2024 and demonstrated the ability to generalize to new time periods, and the ensemble approach provided the best trade-off between complexity and predictive stability.

4.4. Policy and Operational Implications

The improvements in the newly proposed models have useful implications for both policy and practice in cities like Lahore and many others. The ensemble model is reliable enough to support a daily Air Quality Early Warning System since the model has an R^2 of approximately 0.77 and its RMSE values are less than the low-20s ($\mu g/m^3$). Such a system can aid authorities in providing Smog Alerts, closing schools, etc., when there are high levels of pollution. The model also shows that it can help planners plan for critical times such as Winter Inversions, Post-Harvest Crop Burning, etc. The results from feature importance show that combustion-related pollutants (CO and NO_2) and meteorological conditions (surface pressure and temperature) have strong influences on the levels of $PM_{2.5}$. These results will allow authorities to take targeted action such as stricter control of traffic and brick kilns when there is stagnant weather. Although the model does not represent a solution to long-term air pollution, the model will help to minimize exposure and the risk of health problems associated with extreme smog events.

The model is also applicable to other data-scarce cities in South Asia. Since the model uses satellite and reanalysis data, along with limited ground monitoring, the model provides reliable $PM_{2.5}$ forecasting at low cost. The model can also be connected to public alert platforms such as Mobile Apps, to provide Easy-to-Understand air quality information. Overall, this study demonstrates how interpretable machine learning models can support improved air quality management, improve public awareness of air quality issues, and help protect public health in highly polluted urban areas.

5. Conclusions

This study has developed a framework for forecasting daily $PM_{2.5}$ concentrations in Lahore, based upon an ensemble of four different types of machine learning algorithms utilizing data collected between 2019 through 2023 for training purposes and data collected independently for testing in 2024. Each type of model used demonstrated very good predictive ability; however, Random Forest demonstrated the greatest predictive ability of all four base models (testing $R^2 = 0.79$; RMSE ≈ 21 $\mu g\ m^{-3}$; 2024 $R^2 = 0.767$). XGBoost and CatBoost also showed excellent predictive ability (testing $R^2 = 0.78$ and 0.76); however, they appeared to have slightly larger generalization errors than Random Forest. LightGBM had the greatest R^2 ($R^2 = 0.98$) of the training scores; however, it had the largest reduction in R^2 during validation, indicating possible overfitting. A performance-weighted ensemble was created using Random Forest (RF): 0.5, XGBoost (XGB): 0.3, and CatBoost (CatBoost): 0.2 and had the most overall performance. The performance-weighted ensemble had an independent validation $R^2 = 0.7728$, RMSE = 23.37 $\mu g\ m^{-3}$, and MAE = 17.29 $\mu g\ m^{-3}$, which were greater values than each base learner. The performance-weighted ensemble increased the accuracy of independent-year predictions by approximately 1–2 $\mu g\ m^{-3}$ compared to the best individual model. Statistical comparison using paired *t*-tests and Diebold–Mariano forecast accuracy tests confirmed that the ensemble significantly reduced forecast error relative to the individual models, particularly CatBoost and XGBoost. The use of multiple complementary predictors is shown to result in measurable improvements in prediction accuracy. Residuals from the ensemble were confirmed to be stable as they were centered about zero, normally distributed (QQ plot aligned), and did not correlate with the level of predicted concentration.

Analysis of feature importance revealed that surface pressure, temperature, and relative humidity, along with CO and NO_2, as trace gases, were the strongest predictors, and together accounted for more than 60% of $PM_{2.5}$ variability. Therefore, the use of ERA5 meteorology, MODIS AOD, and Sentinel-5P trace gas data, in combination, is effective in capturing both emission-driven and weather-driven pollution dynamics. Although the

modeling framework achieved performance comparable to many regional and international studies, there are limitations regarding its generalizability due to the dependence of the results solely on one monitoring station and the lack of high-resolution emission data. However, the demonstrated time period generalization to an unseen year (2024) demonstrates the robustness of the model in a data-poor environment. To expand the applicability of the framework, future research should incorporate multi-station datasets, near-real-time emission proxies, and spatial modeling techniques to allow for real-time and operationally deployable air quality forecasting across larger urban areas. In summary, this study provides quantitative evidence that the application of ensemble learning with freely available satellite imagery can provide highly accurate, scalable, and interpretable $PM_{2.5}$ forecasts in regions with limited air quality monitoring capabilities and provides a foundation for developing early warning and public health advisory systems in Pakistan.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/atmos17040411/s1, Figure S1: Boxplots depicting outliers for each variable across the full dataset. High-value peaks are retained as legitimate atmospheric phenomena rather than being treated as noise. Figure S2: Temporal variation in predictor variables and $PM_{2.5}$ concentrations (2019–2024). Clear seasonal and inter-annual patterns are visible, which were leveraged for supervised learning. Figure S3: Histograms and KDE curves representing the distribution of the eleven independent variables and $PM_{2.5}$ concentrations (2019–2024). Variables exhibit varied distributions, suggesting the need for normalization during preprocessing. Figure S4: Year-wise wind rose diagrams (2019–2024) representing wind speed and direction derived from ERA5 U and V components. Eastern and southeastern winds dominate, significantly influencing $PM_{2.5}$ transport in Lahore. Figure S5: Temporal validation plots of actual vs. predicted $PM_{2.5}$ concentrations for the RF, LightGBM, CatBoost, and XGBoost models during training, testing, and validation phases. Figure S6: Feature importance score of the RF model. Figure S7: Residual diagnostics for the ensemble $PM_{2.5}$ forecasting model during independent 2024 validation. (a) Residual distribution (histogram with KDE) (b) Residuals versus predicted $PM_{2.5}$ (c) Q–Q plot illustrating strong alignment between empirical and theoretical quantiles. Table S1. Summary of variables, abbreviations, and measurement units used in this study. Table S2: Missing value report for predictor and target variables (2019–2024). Table S3: Statistical summary of all input variables, showing the range and variability of the data used for modelling $PM_{2.5}$ across the training and validation periods. Table S4: Variance Inflation Factor (VIF) values of predictors used in the $PM_{2.5}$ modelling framework. Table S5: statistical comparison between the ensemble model and individual base learners (RF, XGBoost, and CatBoost).

Author Contributions: M.H., Z.T. and K.M. performed the methodology, conceptualization, writing, and editing. M.H., Z.T., S.A.M., S.K., H.A. and S.G. performed the validation, formal analysis, and investigation, while M.H. and S.A.M. supervised this research. All authors contributed to the review and final approval of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: All authors have read, understood, and have complied, as applicable, with the statement on the "Ethical responsibilities of Authors" as found in the Instructions for Authors and are aware that with minor exceptions, no changes can be made to authorship once the paper is submitted.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding authors upon reasonable request.

Acknowledgments: We appreciate the helpful suggestions made by the reviewers who wish to remain anonymous. Muhammad Haseeb gratefully acknowledges the Institute of Space Science,

University of the Punjab, Lahore, Pakistan for its continuous institutional support, technical guidance, and provision of software and computational resources essential for the successful completion of this study. K. Mehmood gratefully acknowledges institutional support from Aix-Marseille Univ, Avignon Univ, CNRS, IRD, IMBE, ITEM, Aix-en-Provence, France.

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

References

1. Haseeb, M.; Tahir, Z.; Amer, S.; Arif, H.; Almutairi, K.F.; Soufan, W.; Tariq, A. Comparative Analysis of Machine Learning Models for Predicting $PM_{2.5}$ Concentrations Using Meteorological and Chemical Indicators. *J. Atmos. Sol. Terr. Phys.* **2024**, *263*, 106338. [CrossRef]
2. Makhdoomi, A.; Sarkhosh, M.; Ziaei, S. OPEN $PM_{2.5}$ Concentration Prediction Using Machine Learning Algorithms: An Approach to Virtual Monitoring Stations. *Sci. Rep.* **2025**, *15*, 8076. [CrossRef]
3. Bekkar, A.; Hssina, B.; Douzi, S.; Douzi, K. Air-Pollution Prediction in Smart City, Deep Learning Approach. *J. Big Data* **2021**, *8*, 161. [CrossRef]
4. Zhou, S.; Wang, W.; Zhu, L.; Qiao, Q.; Kang, Y. Environmental Science and Ecotechnology Deep-Learning Architecture for $PM_{2.5}$ Concentration Prediction: A Review. *Environ. Sci. Ecotechnol.* **2024**, *21*, 100400. [CrossRef] [PubMed]
5. Guo, Q.; He, Z.; Wang, Z. Predicting of Daily $PM_{2.5}$ Concentration Employing Wavelet Artificial Neural Networks Based on Meteorological Elements in Shanghai, China. *Toxics* **2023**, *11*, 51. [CrossRef] [PubMed]
6. Zhang, Z.; Zhao, M.; Zhang, Y. How Does Urbanization a Ect Public Health ? New Evidence from Countries Worldwide. *Front. Public Health* **2023**, *10*, 1096964.
7. Chen, H.; Deng, G.; Liu, Y. Monitoring the Influence of Industrialization and Urbanization on Spatiotemporal Variations of AQI and $PM_{2.5}$ in Three Provinces, China. *Atmosphere* **2022**, *13*, 1377. [CrossRef]
8. Fatima, M.; Butt, I.; Atta, A. Assessment Of Air Pollution And Its Association with Population Health: Geo-Statistical Evidence. *Geogr. Environ. Sustain.* **2023**, *2*, 93–101. [CrossRef]
9. Youns, I.; Bukhari, O.; Butt, I.; Shafiq, M. Air Pollution Patterns Identification in Lahore City Using GIS.; Pakistan. *Pak. Geogr. Rev.* **2009**, *64*, 40–48.
10. Malhi, H.; Ahmed, I.; Nasim, I.; Khurshid, I.; Haider, R.; Nawaz, R.; Irshad, M.; Khan, A.; Shah, S. Monitoring of Ambient Air Pollution in Lahore City. *Pak. J. Emerg. Sci. Technol. (PJEST)* **2023**, *4*, 65–73. [CrossRef]
11. Zhang, H.; Liu, Y.; Zhang, C.; Li, N. Machine Learning Methods for Weather Forecasting: A Survey. *Atmosphere* **2025**, *16*, 82. [CrossRef]
12. Waqas, M.; Chueasa, B.; Wangwongchai, A. Arti Fi Cial Intelligence and Numerical Weather Prediction Models: A Technical Survey. *Nat. Hazards Res.* **2024**, *5*, 306–320. [CrossRef]
13. Li, S.; Ding, Y.; Xing, J.; Fu, J.S. Retrieving Ground-Level $PM_{2.5}$ Concentrations in China (2013–2021) with a Numerical-Model-Informed Testbed to Mitigate Sample-Imbalance-Induced Biases. *Earth Syst. Sci. Data* **2024**, *16*, 3781–3793. [CrossRef]
14. Masood, A.; Hameed, M.M.; Srivastava, A.; Pham, Q.B.; Ahmad, K.; Fatin, S.; Razali, M.; Baowidan, S.A. Improving $PM_{2.5}$ Prediction in New Delhi Using a Hybrid Extreme Learning Machine Coupled with Snake Optimization Algorithm. *Sci. Rep.* **2023**, *13*, 21057. [CrossRef] [PubMed]
15. Alawi, O.A.; Kamar, H.M.; Alsuwaiyan, A.; Yaseen, Z.M. Temporal Trends and Predictive Modeling of Air Pollutants in Delhi: A Comparative Study of Artificial Intelligence Models. *Sci. Rep.* **2024**, *14*, 30957. [CrossRef] [PubMed]
16. Lin, H.; Li, S.; Xing, J.; Yang, J.; Wang, Q.; Dong, L.; Zeng, X. Fusing Retrievals of High Resolution Aerosol Optical Depth from Landsat-8 and Sentinel-2 Observations over Urban Areas. *Remote Sens.* **2021**, *13*, 4140. [CrossRef]
17. Zhang, Y.; Li, Z.; Bai, K.; Wei, Y.; Xie, Y.; Zhang, Y.; Hong, J.; Xu, H.; Guang, J.; Lv, Y.; et al. Satellite Remote Sensing of Atmospheric Particulate Matter Mass Concentration: Advances, Challenges, and Perspectives. *Fundam. Res.* **2021**, *1*, 240–258. [CrossRef]
18. Nagy, A.; El, A.; Mohamed, Z.; Walaa, S.; Manal, M. Monitoring Air Quality of Sharkia Governorate Using Remote Sensing. *Acta Geophys.* **2024**, *72*, 3737–3750. [CrossRef]
19. Chen, X.; Ding, J.; Wang, J.; Ge, X.; Raxidin, M.; Liang, J.; Chen, X.; Zhang, Z.; Cao, X.; Ding, Y. Retrieval of Fine-Resolution Aerosol Optical Depth (AOD) in Semiarid Urban Areas Using Landsat Data: A Case Study in Urumqi, NW China. *Remote Sens.* **2020**, *12*, 467. [CrossRef]
20. Almeida, B.; David, J.; Campos, F.S.; Cabral, P. Satellite-Based Machine Learning Modelling of Ecosystem Services Indicators: A Review and Meta-Analysis. *Appl. Geogr.* **2024**, *165*, 103249. [CrossRef]
21. Ali, M.A.; Bilal, M.; Wang, Y.; Nichol, J.E.; Mhawish, A.; Qiu, Z.; de Leeuw, G.; Zhang, Y.; Zhan, Y.; Liao, K.; et al. Accuracy Assessment of CAMS and MERRA-2 Reanalysis $PM_{2.5}$ and PM_{10} Concentrations over China. *Atmos. Environ.* **2022**, *288*, 119297. [CrossRef]

22. Ahmed, M.; Xiao, Z.; Shen, Y. Estimation of Ground $PM_{2.5}$ Concentrations in Pakistan Using Convolutional Neural Network and Multi-Pollutant Satellite Images. *Remote Sens.* **2022**, *14*, 1735. [CrossRef]
23. Adong, P.; Bainomugisha, E.; Dev, S. Evaluating Machine Learning Methods for $PM_{2.5}$ Estimation Using Satellite AOD, Low-Cost and Reference-Grade Monitors in Kampala. *Int. J. Environ. Sci. Technol.* **2025**, *22*, 15747–15756. [CrossRef]
24. Zoqi, M.J.; Aliaqa, A.D.; Rasouli, S.J.; Daneshvar, M.R.M. Prediction of Daily $PM_{2.5}$ Concentrations Using Neural Networks at 3 Urban Monitoring Stations With Diverse Land Uses. *Environ. Health Insights* **2025**, *19*, 11786302251391304. [CrossRef]
25. Safira, D.A.; Kuswanto, H.; Ahsan, M. Improving the Forecast Accuracy of $PM_{2.5}$ Using SETAR-Tree Method: Case Study in Jakarta, Indonesia. *Atmosphere* **2025**, *16*, 23. [CrossRef]
26. Niu, J.; Li, H.; Liu, X.; Lin, H.; Zhou, P.; Zhu, X. Driving Force of Meteorology and Emissions on $PM_{2.5}$ Concentration in Major Urban Agglomerations in China. *Atmosphere* **2024**, *15*, 1499. [CrossRef]
27. Islam, N.; Toha, T.R.; Islam, M.M.; Ahmed, T. Spatio-Temporal Variation of Meteorological Influence on $PM_{2.5}$ and PM_{10} over Major Urban Cities of Bangladesh. *Aerosol Air Qual. Res.* **2023**, *23*, 220082. [CrossRef]
28. Veefkind, J.P.; Aben, I.; McMullan, K.; Förster, H.; de Vries, J.; Otter, G.; Claas, J.; Eskes, H.J.; de Haan, J.F.; Kleipool, Q.; et al. TROPOMI on the ESA Sentinel-5 Precursor: A GMES Mission for Global Observations of the Atmospheric Composition for Climate, Air Quality and Ozone Layer Applications. *Remote Sens. Environ.* **2012**, *120*, 70–83. [CrossRef]
29. Liu, Z.; Fang, Z.; Hu, Y. A Deep Learning-Based Hybrid Method for $PM_{2.5}$ Prediction in Central and Western China. *Sci. Rep.* **2025**, *15*, 10080. [CrossRef]
30. Nourmohammad, E.; Rashidi, Y. Ground Data Analysis for $PM_{2.5}$ Prediction Using Predictive Modeling Techniques. *J. Air Pollut. Health* **2025**, *10*, 61–82. [CrossRef]
31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
33. Cahya, Z.; Cahya, D.; Nugroho, T.; Zuhri, A.; Agusta, W. CNN Model with Parameter Optimisation for Fine-Grained Banana Ripening Stage Classification. *ACM Int. Conf. Proc. Ser.* **2022**, *13*, 90–94. [CrossRef]
34. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef]
35. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
36. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
37. Ma, J.; Yu, Z.; Qu, Y.; Xu, J.; Cao, Y. Application of the XGBoost Machine Learning Method in $PM_{2.5}$ Prediction: A Case Study of Shanghai. *Aerosol Air Qual. Res.* **2019**, *20*, 128–138. [CrossRef]
38. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 4766–4775. [CrossRef]
39. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates Inc.: New York, NY, USA, 2017; Volume 30.
40. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Proceedings of the Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates Inc.: New York, NY, USA, 2018; Volume 31.
41. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support. *arXiv* **2018**, arXiv:1810.11363. [CrossRef]
42. Kanagarathinam, K.; Krishnan, S.; Manikandan, R. Water Quality Prediction: A Data-Driven Approach Exploiting Advanced Machine Learning Algorithms with Data Augmentation. *J. Water Clim. Change* **2023**, *15*, 431–452. [CrossRef]
43. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for Big Data: An Interdisciplinary Review. *J. Big Data* **2020**, *7*, 94. [CrossRef] [PubMed]
44. Varoquaux, G.; Cheplygina, V. Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future. *NPJ Digit. Med.* **2022**, *5*, 48. [CrossRef] [PubMed]
45. Willmott, C.J.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
46. Alexander, L.V.; Zhang, X.; Peterson, T.C.; Caesar, J.; Gleason, B.; Klein Tank, A.M.G.; Haylock, M.; Collins, D.; Trewin, B.; Rahimzadeh, F.; et al. Global Observed Changes in Daily Climate Extremes of Temperature and Precipitation. *J. Geophys. Res. Atmos.* **2006**, *111*, 0148–0227. [CrossRef]
47. Wang, Y.; Nemitz, E.; Tomlinson, S.J.; Carnell, E.J.; Yao, L.; Scheffler, J.; Liska, T.; Pearson, C.; Dragosits, U.; Venkataraman, C.; et al. Response of South Asia PM 2.5 Pollution to Ammonia Emission Changes and Associated Impacts on Human Health. *Environ. Int.* **2025**, *195*, 109207. [CrossRef]
48. Gupta, P.; Zhan, S.; Mishra, V.; Aekakkararungroj, A.; Markert, A.; Paibong, S.; Chishtie, F. Machine Learning Algorithm for Estimating Surface $PM_{2.5}$ in Thailand. *Aerosol Air Qual. Res.* **2021**, *21*, 210105. [CrossRef]

49. Song, Z.; Chen, B.; Huang, Y.; Dong, L.; Yang, T. Estimation of $PM_{2.5}$ Concentration in China Using Linear Hybrid Machine Learning Model. *Atmos. Meas. Tech.* **2021**, *14*, 5333–5347. [CrossRef]
50. Kunjir, G.M.; Tikle, S.; Das, S.; Karim, M.; Kumar, S. Assessing Particulate Matter ($PM_{2.5}$) Concentrations and Variability across Maharashtra Using Satellite Data and Machine Learning Techniques. *Discov. Sustain.* **2025**, *6*, 238. [CrossRef]
51. Wang, S.; Wang, P.; Qi, Q.; Wang, S.; Meng, X.; Kan, H.; Zhu, S.; Zhang, H. Improved Estimation of Particulate Matter in China Based on Multisource Data Fusion. *Sci. Total. Environ.* **2022**, *869*, 161552. [CrossRef]
52. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. An Ensemble-Based Model of PM 2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution. *Environ. Int.* **2019**, *130*, 104909. [CrossRef]