scientific reports

Check for updates

OPEN

$PM_{2.5}$ concentration prediction using machine learning algorithms: an approach to virtual monitoring stations

Ahmad Makhdoomi, Maryam Sarkhosh✉ & Somayyeh Ziaei

One of the most important pollutants is $PM_{2.5}$, which is particularly important to monitor pollutant levels to keep the pollutant concentration under control. In this research, an attempt has been made to predict the concentrations of $PM_{2.5}$ using four Machine Learning (ML) models. The ML methods include Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting Regressor (XGBR), Random Forest (RF) and Gradient Boosting Regressor (GBR). The mean and maximum concentration of $PM_{2.5}$ were recorded 32.84 µg/m^3 and 160.25 µg/m^2, respectively, indicating the occurrence of occasional episodes of high pollution levels from 2016 to 2022. The PM2.5 concentrations dropped below 30 µg/m^2 in 2018 due to reduced human activities during COVID-19 lockdowns but $PM_{2.5}$ levels were significantly increased because of the ongoing operation of heavy industries post-COVID-19 lockdowns during 2021. The ML models performed very well in predicting the concentrations of $PM_{2.5}$ with around 95% of their predictions falling within the factor of the observed concentration. The results presented that among the four ML algorithms, GBR confirmed good model performance compared to the other models, with the lowest MSE (5.33) and RMSE (2.31), as well as high accuracy measures. This suggests that GBR is the best model for reducing large errors, making it more robust in capturing variations in PM2.5 levels. In conclusion, the study proposed a method to obtain high-accuracy $PM_{2.5}$ prediction results using ML which are useful for air quality monitoring on a global scale and improving acute exposure assessment in epidemiological research.

Keywords Particulate matter, Machine learning models, Meteorological data, $PM_{2.5}$ prediction

Metropolitan air pollution is increasingly becoming the most awareness environmental concern because of diverse array of health complications[1]. The fine particulate matter with aerodynamic diameter less than 2.5 µm (i.e., $PM_{2.5}$) has been reported strongly associated with human diseases, such as lung cancer and premature deaths[2]. $PM_{2.5}$ has now become the chief pollutant which affects the atmospheric quality in the world, especially in the urban agglomerations. Most epidemiological studies on the effect of air pollutants show an elevated risk of cardiovascular diseases linked to $PM_{2.5}$ exposures Nevertheless, the sparse and limited air quality monitoring stations still make it difficult to acquire the accurate spatial distribution of $PM_{2.5}$ concentrations in a region. Accurate prediction of PM2.5 levels is critical to reduce these negative effects. Limitations of traditional monitoring methods, such as sparse spatial coverage, high costs, and time delays in data availability are the use of machine learning (ML) as a powerful method to address these challenges. ML methods can leverage complex, non-linear relationships in environmental data to provide accurate, real-time predictions of PM2.5 levels, enabling proactive measures to protect public health and also for efficient decision making[3,4].

Machine Learning methods because of their powerful capacity have been emerged in air quality modelling and Predicting $PM_{2.5}$ concentrations in the recent decade[5,6]. A search for more viable models than the operational air quality models lead to many studies on the use of various intelligent ML approaches that can accurately forecast in air quality indexes[7]. Several ML models are commonly employed to build air quality models, forecasting of wind speed[8,9], energy generation[10,11], quantity of solar energy[12] with comparable or better accuracy[1,13–15]. These algorithms can be achieved at a lower computational cost and with no assumptions on the atmospheric processes involved[16]. On the contrary, use of ML models provides high prediction performance with nonlinear variables and flexible modeling for PM prediction with numerical prediction models, satellite, and ground observation data as input data[17,18]. To date, more studies have leveraged ML for $PM_{2.5}$ estimation with using satellite aerosol

Department of Environmental Health Engineering, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran. ✉email: Marya.sarkhosh@yahoo.com

optical depth (AOD), the absorbing aerosol index, chemical transport model outputs, land-use data, and meteorological parameters, which have accounted for the complex atmospheric mechanisms[19,20].

Despite the use of ML models for predicting PM2.5 concentrations, several limitations remain. Some studies focus on individual algorithms without comprehensively comparing their performance. The aim of this research is to evaluate the performance of four ML algorithms in predicting PM2.5 levels, with identifying the most vigorous model for reducing large errors and providing a clear understanding of the best model for prediction PM2.5 concentrations.

Light Gradient Boosting Machine (LGBM), Gradient Boosting Regressor (GBR), Extreme Gradient Boosting Regressor (XGBR) and Random Forest (RF) were employed for high accuracy prediction and utilization of $PM_{2.5}$ in Mashhad city, Iran. These algorithms were developed in an environment running Python 3.10.12, Jupyter Lab, pandas 2.1.4, NumPy 1.26.4, Matplotlib (3.7.1), Seaborn (0.13.1), scikit-learn (sklearn) 1.3.2, SciPy and XGBoost 2.1.1.

Materials and methods

Study area

The study area is Mashhad, the second-largest city in Iran and serves as the capital of Khorasan Razavi Province. Mashhad is situated between 36° 12′ and 36° 19′ North latitude and 59° 34′ and 59° 39′ East longitude. Mashhad is ~ 280 square kilometers and its population is 3,372,660 people in 2016[21]. As the holy shrine of Imam Reza, it is estimated that over 20 million pilgrims visit this city each year. In recent years, Mashhad has undergone rapid urbanization and industrial expansion, significantly affecting air quality. The metropolitan area hosts a range of industries, including textile production, food processing, and chemical industries, all of which contribute to emissions of particulate matter (PM), nitrogen oxides (NO_x), sulfur dioxide (SO_2), and volatile organic compounds (VOCs). Additionally, heavy vehicular traffic, fueled by the city's large population and seasonal influx of visitors, exacerbates air pollution[22].

Public transportation hubs, including bus terminals, railway stations, and the airport, operate above normal capacity due to high travel demand, further deteriorating air quality. In 2017, road transport was responsible for 45% of PM10, 45% of PM2.5, 24% of NOx, and 96% of CO emissions in Mashhad. Among pollutants released from these hubs, bus stations accounted for 91% of PM10, 92% of PM2.5, 25.4% of CO, and 20% of CO2 emissions, highlighting their significant contribution to air pollution[23,24]. The study period spans from January 2016 to December 2022, covering trends in air quality amidst these ongoing developments. The study time period is from January 2016 to December 2022.

PM concentration and metrological data

This study utilized meteorological prediction data from the local data assimilation and prediction system (LDAPS). In addition, the $PM_{2.5}$ estimation selected as the final value was employed to indicate the most severe air quality condition observed. The data being monitored at the Air Quality monitoring (AQM) sites include both particulate matter (PM_{10}, $PM_{2.5}$), due to their significant impact on air quality. For this study, meteorological variables at each monitoring unit to develop the ML models including vvmin (minimum horizontal visibility), ffm (average wind speed), rrr24 (24-h rain), um (average relative humidity), dd (wind direction) and nhz (number of dust reports) were used. These datasets were obtained from the Iranian Meteorological Organization and play a noteworthy role in influencing AQI in a given area. The descriptive statistics of the parameters used in this study from 2203 matched observations, covering for the time period 2016–2022. The main focus of this study was the $PM_{2.5}$, which serves as an indicator of the overall air quality in Mashhad.

To ensure the reliability of the meteorological and air quality data, a rigorous data cleaning process was applied. Missing values were entirely removed from the dataset to prevent potential biases in the analysis. For handling noisy data, outlier detection was performed using the interquartile range (IQR) method. Specifically, outliers in PM2.5 concentrations were identified using an extended threshold of 1.5 times the IQR. Data points falling beyond this range were considered extreme values and were removed to improve data quality.

The relationship between the measured pollutants was calculated using the Spearman correlation. In this correlation, the relationships between the numerical variables are measured. Thus, it essentially provides a measure of the monotonic relationship between those two variables. The correlation ranges from − 1 to + 1. If the correlation is near to + 1 then the features are positively correlated, where − 1 means negatively correlated[25,26].

Machine learning models

Light gradient boosting machine (LGBM)

LGBM algorithm is a powerful and efficient tool for machine learning that can be used for both regression and classification tasks, especially for large-scale datasets which improve model efficiency and reduce memory usage[27]. The equation for LGBM as follow:

$$\widehat{y_i} = argmin_f E_{y,x} L_{(y,\ f(x))} \tag{1}$$

where $\widehat{y_i}$ is the predicted value, $f(x)$ represents the model's function mapping input x to an output, and $L(y, f(x))$ is the loss function measuring the error between predicted and actual values.

Extreme gradient boosting regressor (XGBR)

XGBR algorithm is an ensemble learning method that attempt the predictions from a set of simpler and weaker models to produce a stronger prediction. It performs to handle large datasets because of its robust handling of a variety of data types, relationships, distributions, and the variety of hyper parameters to achieve state-of-the-art performance in many machine learning tasks[28,29]. The following formula is for XGBR:

$$\widehat{Y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \widehat{Y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

where $\widehat{Y}_i^t$ is the prediction at iteration t, $\widehat{Y}_i^{(t-1)}$ is the prediction from the previous iteration, $f_t(x_i)$ is the new weak learner added at step t, and x_i is the input variable.

Random forest (RF)

RF algorithm is a variant of bagging that fits a multitude of decision trees on different sub-samples to find the output. Sampling features are termed column sampling and data points as row sampling. Trees are built with row and column samples. The advantage of building the model in such a way is that it is robust in estimating new data points[30]. The formula for random forest can be expressed as:

$$RF = \frac{1}{k} \sum_{k=1}^{k} h_k(x) \tag{3}$$

where K is the number of trees in the ensemble,
and $h_k(x)$ is the prediction from the k-th decision tree.

Gradient boosting regressor (GBR)

The GBR approach can be seen as a functional gradient algorithm that aims at finding an additive model that minimizes the loss function. Thus, the GBR algorithm iteratively adds at each step a new decision tree that best reduces the loss function[31]. A GBR with (n) number of trees can be stated as;

$$f_N(x_i) = \sum_{n}^{N} y_n h_n(x_i) \tag{4}$$

where h_n is a weak learner that performs poorly individually,
y_n is a scaling factor adding the contribution of a tree to the model.

Justification for model selection

The selected models—LGBM, XGBR, RF, and GBR—were chosen based on their strong performance in prior studies on air quality prediction. These models are well-suited for handling nonlinear relationships in meteorological and pollutant data. Additionally:

LGBM and XGBR: Known for their efficiency in large datasets and ability to capture complex interactions.

RF: Robust against overfitting and effective in managing noisy data.

GBR: Provides strong predictive performance by iteratively reducing errors.

These models were selected to compare different ensemble learning approaches and evaluate their predictive power for air quality forecasting.

Performance metrics and optimization

The ML-based PM prediction accuracy of this study was evaluated. In order to assess the efficiency of the model, four statistical indicators of the match, including the mean squared error (MSE), coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and mean Absolute Percentage Error (MAPE) were used to evaluate the correlation of estimated and observed PM concentrations as per Eq. (5) to (9)[32,33].

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y_i}| \tag{5}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2} \tag{7}$$

$$\%MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{y_i} \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2} \tag{9}$$

where n, y_i, and $\widehat{y_i}$ refers to the number of data points, measured value and predicted value respectively.

Statistical comparison of model performance

To assess whether the performance differences between the models were statistically significant, non-parametric statistical tests were employed. Initially, the normality of the performance metrics was examined, revealing a non-normal distribution; therefore, the Friedman test was employed as a non-parametric alternative to repeated-measures ANOVA to determine whether significant differences existed among the models. The evaluation was conducted using 10-fold cross-validation, generating ten performance scores per model for each metric, ensuring a robust estimate of predictive capability. A p-value below 0.05 in the Friedman test indicated

significant differences among models, prompting further pairwise comparisons using the Wilcoxon signed-rank test, which is designed for dependent samples. This test was applied to compare model performances across different metrics, with a p-value below 0.05 interpreted as a statistically significant difference between the compared models. The findings from these statistical analyses were summarized in a comparative table, highlighting model pairs with significant performance differences.

Result and discussion

PM (PM2.5 and PM10)

The monthly and annual trends mean time series of the data collected in this study (2016–2022) is shown in Fig. 1 which shows the $PM_{2.5}$ and PM_{10} concentration. The results reveal that PM_{10} concentrations exceeded the threshold of 60 μg/m^3 consistently during the months of September to January, as well as in June. This suggests a period of elevated pollution levels during these months. Additionally, in 2018 to 2020, low PM concentrations of both PM_{10} and $PM_{2.5}$ (falling below 65 μg/m^3 and 30 μg/m^3 respectively), are maintained for approximately five months due to the impact of the COVID-19 caused by reduced human activities and emissions. From January to May, the high magnitude and frequency of precipitation especially during March–May, lead to improvement of air quality and low $PM_{2.5}$ concentrations below 30 μg/m^3 due to wet deposition in the studied area. Base on Fig. 1 the annual mean of $PM_{2.5}$ and PM_{10} concentrations in Mashhad were higher than the World Health Organization (WHO) air quality guidelines. This implies that the prediction of $PM_{2.5}$ is significantly effects on the levels of AQI. Additionally, the mean concentrations of $PM_{2.5}$ and PM_{10} over a seven-year period from 2016 to 2022 were calculated as 32.84 μg/m^3 and 61.88 μg/m^3, respectively. These average values provide an overall understanding of the pollution levels during this time frame. Moreover, the maximum concentrations recorded for $PM_{2.5}$ and PM_{10} were 160.25 μg/m^3 and 315.55 μg/m^3, respectively, indicating the occurrence of occasional episodes of high pollution levels. Overall, these findings highlight the temporal patterns and impact of both natural and anthropogenic factors on the concentrations of $PM_{2.5}$ and PM_{10} [34]. The relationship between the AQI and $PM_{2.5}$ concentrations emphasizes the importance of monitoring and considering particulate matter in air quality assessments and predictions.

The high concentration of airborne particles in Mashhad can be attributed to various factors, with heavy traffic and transportation irregularities being the main factors[35]. The excessive use of vehicles and insufficient infrastructure can lead to the emission of particulate matter into the air. Vehicles produce airborne particles that can range from regular particulate matter to toxic particles[36]. Industrial activities also play a significant role in elevating the particle concentration in Mashhad. Emissions from machinery, factories, and other industrial units can contribute to air pollution by releasing particulate matter into the atmosphere[37]. The use of fossil fuels, such as oil and natural gas, is another factor behind the high particle concentration. Fossil fuels are commonly used as sources of energy and combustion, resulting in the release of particulate matter into the air[38].

A popular explanation is that vegetation can effectively reduce the number of $PM_{2.5}$ sources by fixing the soil. At the same time, new findings verify that larger leaf area, branch, and stem surface enhance the efficiency of intercepting or capturing $PM_{2.5}$ in the subtropical broad-leaved or coniferous and broad-leaved mixed forest, thereby inhibiting effectively the concentrations of $PM_{2.5}$ in the air. As a result, winter air pollution becomes

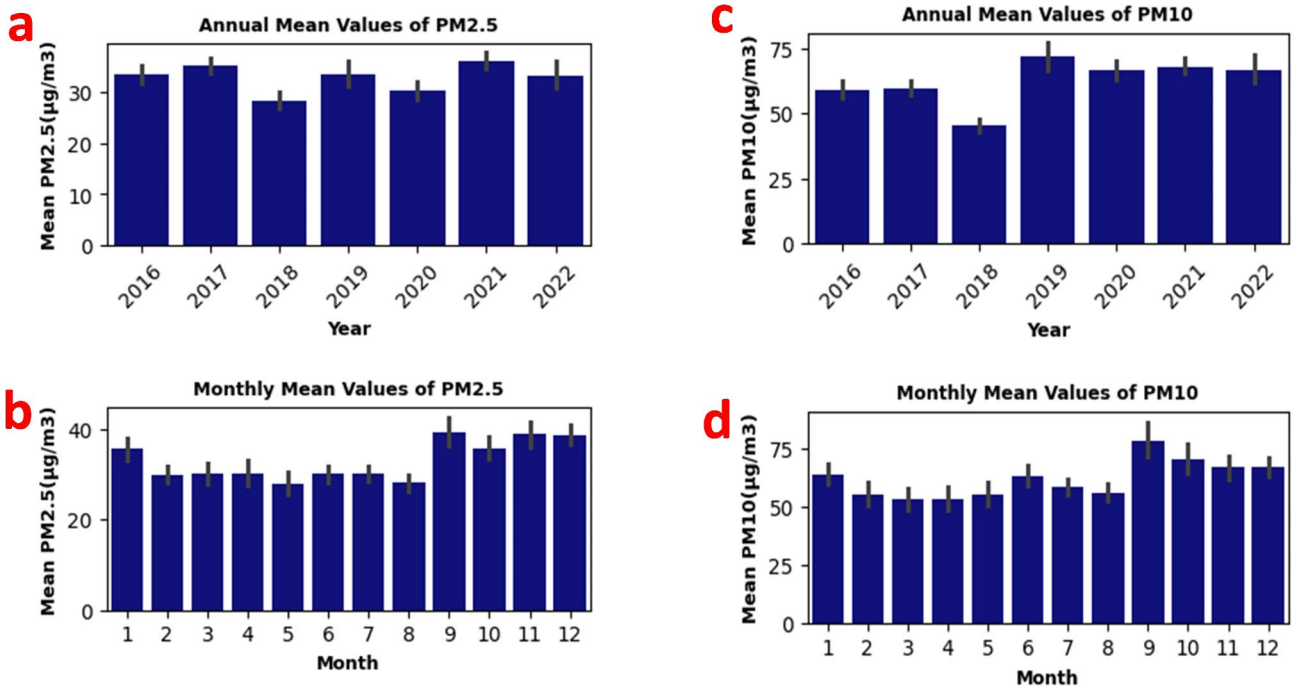

Fig. 1. Annual and monthly average values of $PM_{2.5}$ (**a**,**b**), PM_{10} (**c**,**d**).

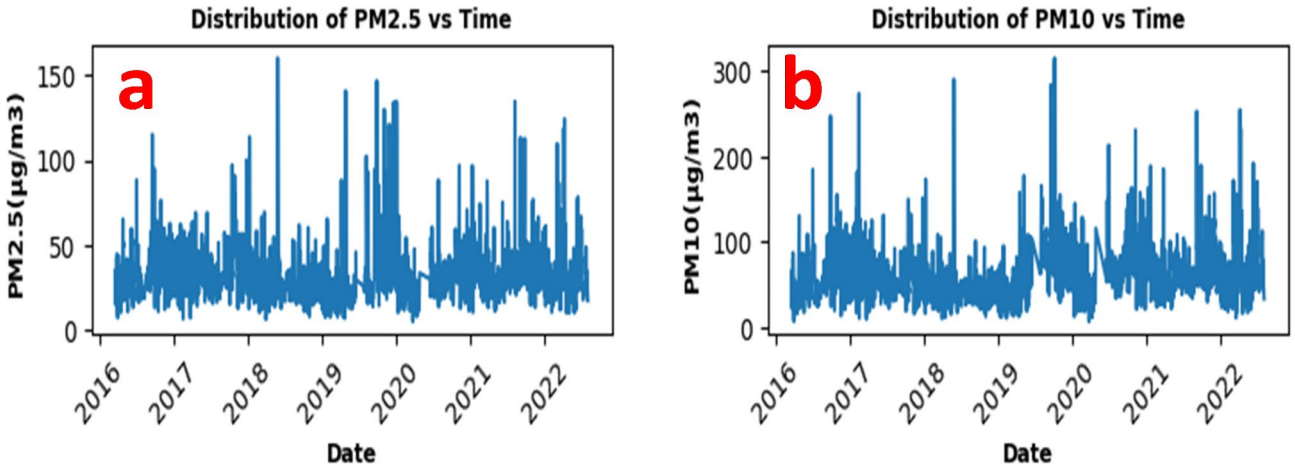

Fig. 2. Daily variations of $PM_{2.5}$ (**a**), and PM_{10} (**b**) from 2016 to 2022.

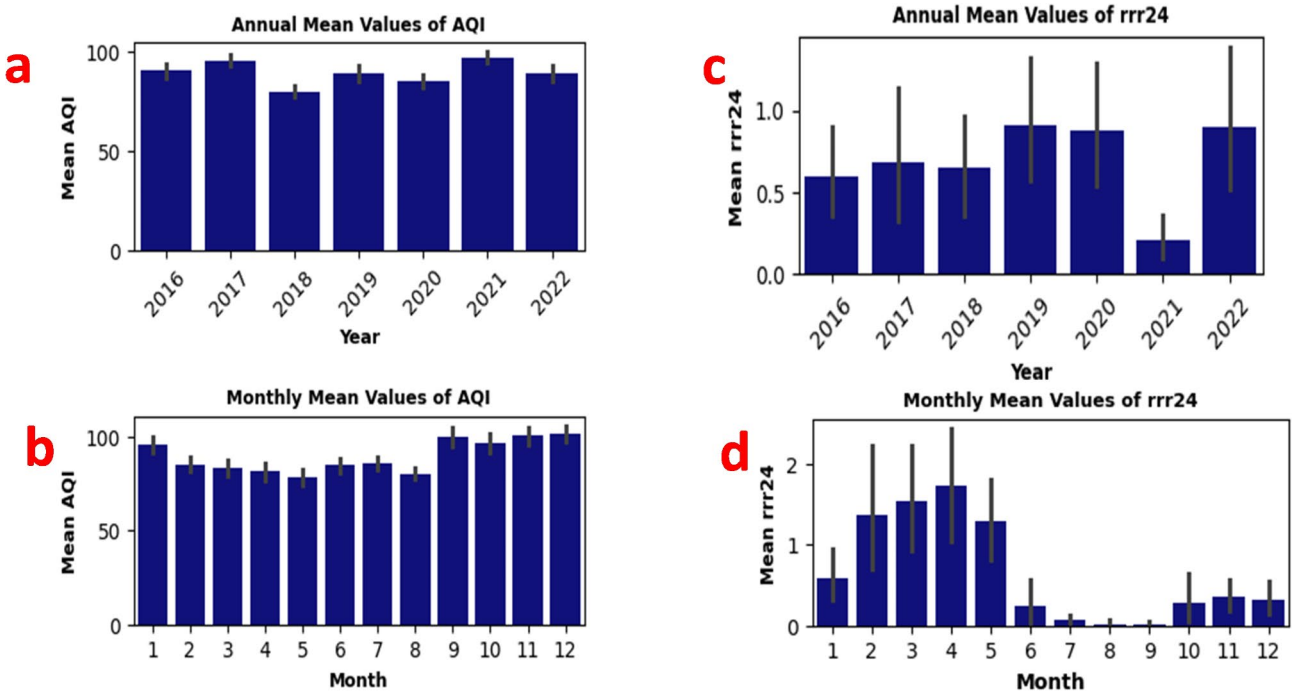

Fig. 3. Annual and monthly average values, of AQI (**a**,**b**) and Average Values of rrr24 (**c**,**d**).

a more pronounced issue in Mashhad. These atmospheric conditions, along with temperature inversions, can contribute to the formation of smog, which is characterized by a high concentration of particulate matter and pollutants. Smog can significantly degrade the air quality in the surrounding areas and contribute to higher levels of air pollution[39].

Figure 2 display the daily variations of Particulates Matter from 2016 to 2022, respectively. The data reveals a consistent Increase in PM every month and year. This pattern can be explained by a variety of factors, such as temperature inversions that lead to the buildup of PM in the winter. Equally, PM levels typically decrease in the summer. It is worth noting that in 2021, pollution levels rose compared to previous years, despite being lower in 2020. This increase may be linked to the ongoing operation of heavy industries post-Covid-19 lockdowns[40]. The resumption of industrial activities may have contributed to the rise in air pollution levels during 2021.

AQI and metrological parameters

The AQI is often closely correlated with the amount of particulate matter and pollutants in the air, and these atmospheric conditions can lead to elevated AQI levels[41]. In addition, the total amount of precipitation also plays a role in determining the AQI. According to the comparison shown in Fig. 3. and also the correlation coefficient ($r = -21$) between AQI and 24-hour rainfall (rrr-24), an increase in precipitation leads to a lower

AQI, and a decrease in precipitation leads to a higher AQI[42]. The period from February to June stands out as having significantly higher rainfall compared to the other seven months. Rainfall and AQI are closely related environmental variables. Statistical analysis clearly indicates that the AQI tends to be lower during periods of heightened precipitation, and conversely, higher during periods of decreased rainfall. Mashhad experiences the majority of its rainfall from mid-autumn to mid-spring of the following year. This seasonal pattern of rainfall might be the underlying reason for the observed changes in AQI.

The dry climate of Mashhad, marked by strong winds, plays a significant role in shaping air pollution levels. Studies indicate that raindrops can absorb airborne dust particles, which then settle due to gravity, leading to a reduction in particulate matter concentrations. Our findings further support this, showing that increased rainfall weakens AQI within a certain range (34). Additionally, wind is a key factor in the transport, dilution, and dispersion of PM2.5, contributing to seasonal variations in pollution levels. Temperature inversions, particularly during winter, further exacerbate air pollution by trapping pollutants near the surface.

Figure 4 presents the daily variations of key meteorological parameters from 2016 to 2022, providing insights into their influence on air quality dynamics. Notably, precipitation (Fig. 4e) exhibits a seasonal pattern, with increased rainfall from February to June, coinciding with improved air quality. Wind speed (Fig. 4a and c) fluctuates significantly, underscoring its role in pollutant dispersion. Temperature (Fig. 4d) follows a cyclic trend, reflecting seasonal variations, while humidity (Fig. 4f) and sunshine duration (Fig. 4g) also contribute

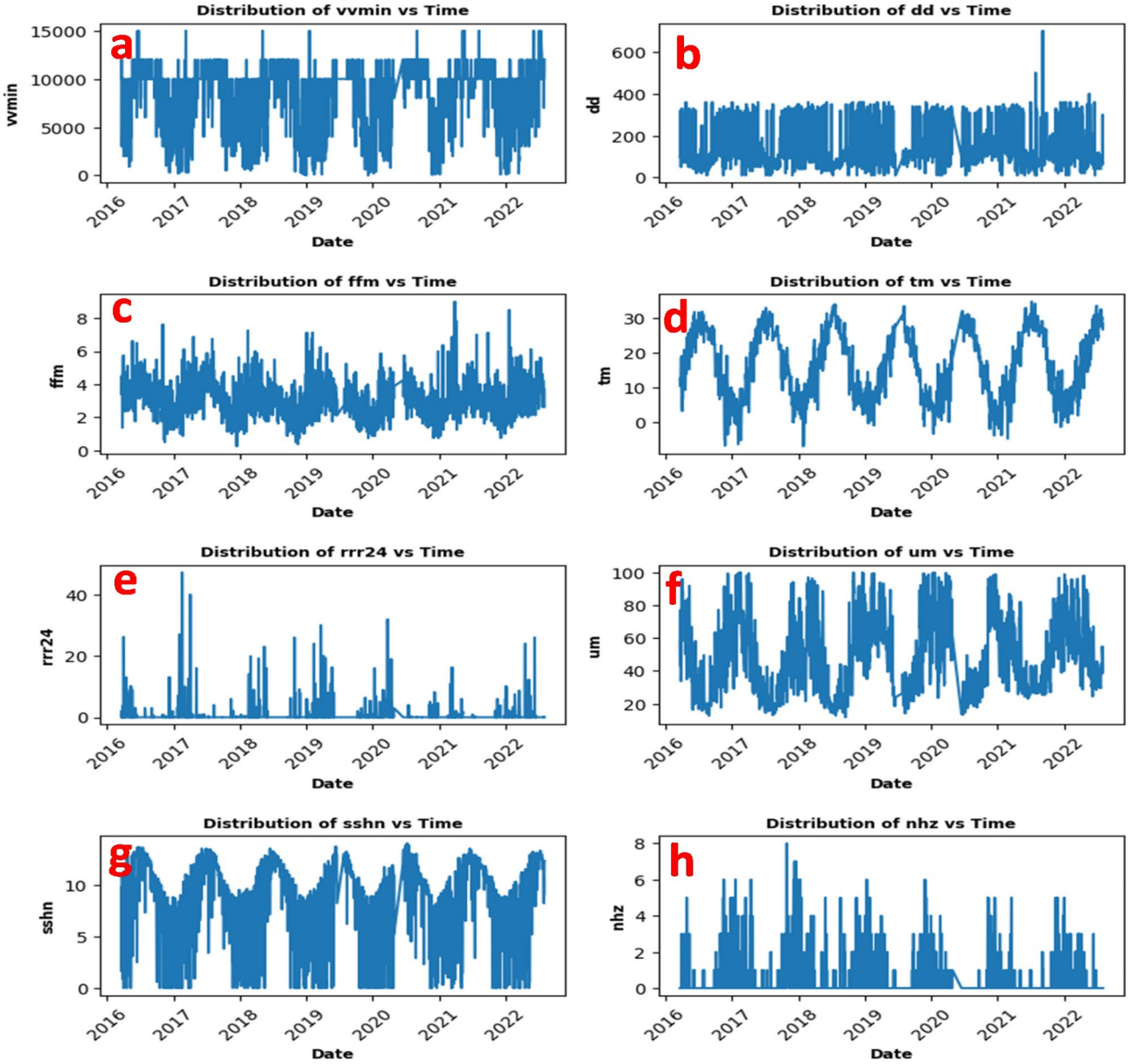

Fig. 4. Daily variations of minimum horizontal visibility (**a**), wind direction (**b**) 24-h rain average wind speed (**c**), average temperature (**d**), 24-hour rain (**e**), average relative humidity (**f**) number of sunny hours (**g**) and number of dust (**h**) from 2016 to 2022.

to atmospheric conditions that impact pollution levels. Additionally, cloud cover (Fig. 4h) varies throughout the year, influencing radiation balance and atmospheric stability. These meteorological factors collectively determine the dispersion, deposition, and resuspension of particulate matter, shaping long-term air quality trends in Mashhad.

It is widely accepted that the rising temperature increases the height of the mixing layer which helps the vertical diffusion of the atmosphere, thereby providing more space for the dilution of surface pollutants. Simultaneously, the turbulent mixing effect from the thermal and dynamic forces of the underlying surface has a direct impact on the migration and transformation of pollutants in the mixed layer[43].

Spearman correlation and evaluation results

Figure 5 illustrates the correlation among various variables. Notably, a strong positive correlation exists between the AQI and PM_{10} with the $PM_{2.5}$ (with a correlation coefficient of 0.95) variables. The correlation results indicate that a substantial portion of the prediction of $PM_{2.5}$ concentrations can be attributed to AQI and PM_{10} levels. The heatmap correlation indicated that PM2.5 concentrations are a main factor influencing the AQI in the studied area. Furthermore, this finding implies that monitoring and controlling PM2.5 concentrations could be an effective strategy for improving air quality. Conversely, other variables exhibited weak to moderate correlations with $PM_{2.5}$. Among the meteorological variables analyzed, four demonstrated a negative correlation with $PM_{2.5}$; however, these correlations were relatively weak. Overall, only those variables exhibiting correlations greater than +0.1 or less than −0.1 with $PM_{2.5}$ were retained for predicting its concentration.

In this study, four ML models were employed to estimate the $PM_{2.5}$ concentrations, consist of GBR, RF, XGBR and LGBM. For training and evaluation purposes, 80% of the available data was used to train the models, while the remaining 20% was used for evaluation. Grid search was utilized to discover the optimal parameters for the models. Table 1 presents the comparison evaluation metrics of the predicted $PM_{2.5}$ for each ML algorithm. These metrics included MAE, MSE, RMSE, MAPE and R^2. Each metric provided unique insights into different aspects of the models predictive accuracy, goodness of fit, and ability to capture the variations in the $PM_{2.5}$ concentrations. These evaluation metrics were used for both the training and testing datasets, but since these values are more important for the testing dataset, they were analyzed specifically for this set.

Table 1 presents the performance of the machine learning models in predicting PM2.5 concentrations. Each model's effectiveness was assessed using MAE, MSE, RMSE, MAPE, and R^2 across training (80%) and validation/testing (20%) datasets.

Among all models, RF exhibited the best performance based on MAE (0.47) and MAPE (1.5%), indicating its superior ability to minimize absolute errors and relative percentage deviations. However, despite its strong performance in these metrics, RF had a slightly higher MSE (8.73) and RMSE (2.95) in the testing phase

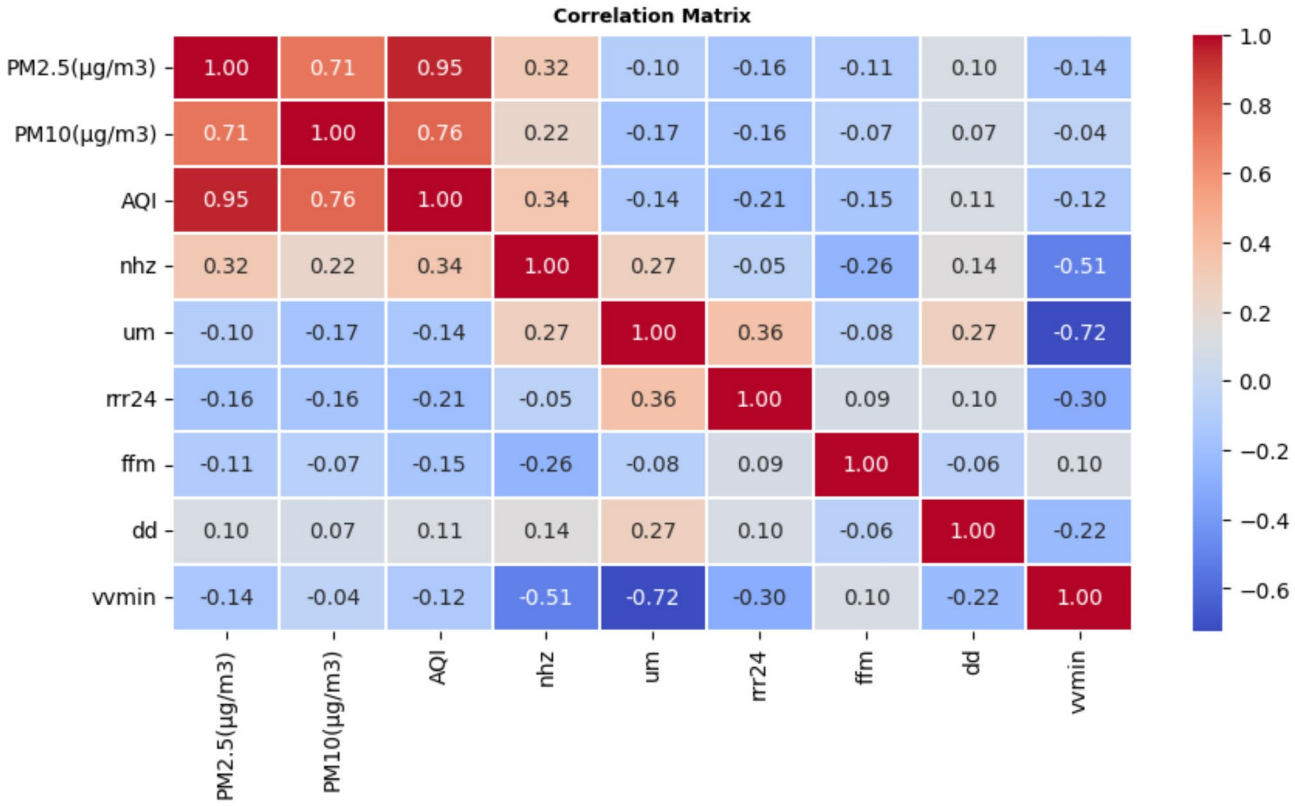

Fig. 5. Correlation Heat map between the AQI and PM and meteorological parameters.

S. no.	Model name	Training set (80%)					Validation/testing set (20%)				
		MAE	MSE	RMSE	MAPE (%)	R^2	MAE	MSE	RMSE	MAPE (%)	R^2
1	LGBM	0.75	5.74	2.4	2.3	0.9784	0.9	10.2	3.19	2.6	0.9611
2	RF	0.19	0.88	0.94	0.7	0.9967	0.47	8.73	2.95	1.5	0.9686
3	GBR	0.4	1.4	1.18	1.7	0.9947	0.54	5.33	2.31	1.9	0.9802
4	XGBR	0.41	0.53	0.73	1.5	0.9981	0.81	5.92	2.43	2.7	0.9781

Table 1. Displays the performance of the models in predicting the $PM_{2.5}$.

S. No	Comparison	*p*-value			
		MAE	MSE	RMSE	R^2
1	RF and GBR	0.0098	0.0195	0.0371	0.0488
2	RF and XGB	0.0039	0.2754	0.3750	0.3223
3	RF and LGBM	0.0019	0.0195	0.0839	0.0839
4	GBR and XGB	0.0137	0.0137	0.0137	0.0195
5	GBR and LGBM	0.0019	0.0039	0.0039	0.0039
6	XGB and LGBM	0.0839	0.6953	0.6953	1.0000

Table 2. Statistical comparison of model performance.

compared to GBR, which had the lowest MSE (5.33) and RMSE (2.31). This suggests that GBR is better at reducing large errors, making it more robust in capturing variations in PM2.5 levels.

XGBR performed well in terms of R^2 (0.9781), demonstrating high predictive power, but it had the highest MAPE (2.7%), indicating that its predictions deviated the most in percentage terms. Meanwhile, LGBM exhibited the weakest performance across multiple metrics, with the highest MSE (10.2) and the lowest R^2 (0.9611), suggesting greater residual errors and lower predictive reliability.

Despite variations in individual metrics, all models achieved an R^2 above 96%, demonstrating strong overall predictive capability. However, when considering robustness and error minimization, GBR appears to provide the most balanced performance, while RF is superior in absolute and percentage error reduction. Conversely, LGBM performed the weakest in all error-based evaluations.

The results of the statistical comparison of model performance using the Friedman test and subsequent Wilcoxon signed-rank test are presented in Table 2. The Wilcoxon signed-rank test results demonstrate that Random Forest (RF) and Gradient Boosting Regressor (GBR) exhibit significant differences across all metrics, with p-values below 0.05. Similarly, RF and XGBoost (XGB) show significant differences in MAE ($p = 0.0039$) but not in MSE, RMSE, or R^2, indicating that their predictive capabilities are comparable in these aspects. Conversely, RF and LightGBM (LGBM) differ significantly in MAE and MSE ($p < 0.05$) but not in RMSE and R^2, suggesting that while these models may vary in error magnitude, their overall variance explanations are similar. The comparison between GBR and XGB reveals significant differences across all four performance metrics ($p < 0.05$), emphasizing that these models behave differently. Additionally, GBR and LGBM exhibit significant differences across all metrics, reinforcing that GBR and LGBM yield distinct predictive performances.

Figure 6. shows actual (blue) and predicted (yellow) PM2.5 concentrations from 2016 to 2022, with GBR and RF models demonstrating better alignment with actual values. However, notable deviations occur, particularly during 2018–2019 and 2019–2020, where sharp pollution spikes were likely influenced by temperature inversions and increased winter emissions, and in late 2020 to early 2021, possibly due to industrial activity or traffic fluctuations. LGBM and XGBR struggle with high PM2.5 levels, suggesting limitations in handling extreme values. To improve predictions, incorporating meteorological and traffic data, data augmentation, and hybrid modeling approaches could enhance accuracy, especially during pollution surges.

Conclusions

Air pollution is a global problem and researchers from all around the world are working to discover a solution. To accurately forecast the $PM_{2.5}$ concentrations, machine learning techniques were investigated. The present study assessed the performance of four machine learning models including RF, LGBM, XGBR, GBR for predicting the accurate $PM_{2.5}$ concentrations in Mashhad. The performance of the models in predicting $PM_{2.5}$ concentrations in Mashhad between 2016 and 2022 was compared. The evaluation metrics used included R^2, MAPE, RMSE, MSE, and MAE, which were calculated for both the training and testing datasets. However, the final analysis was performed on the results of the testing dataset. The GBR performed strong capability in predicting $PM_{2.5}$ concentrations based on the R^2, MSE and RMSE metrics which outperformed of the other ML models The R^2 value for the testing dataset was above 96% for all of ML models.

Given the annual and monthly average PM2.5 concentrations in Mashhad, which often exceed the EPA standard (15 $\mu g/m^2$) and WHO guideline (10 $\mu g/m^2$), proactive measures are necessary to mitigate the associated health risks. In addition to expanding green spaces and public transport, targeted strategies should be implemented, including stricter industrial emission controls, improved fuel quality regulations, and the adoption

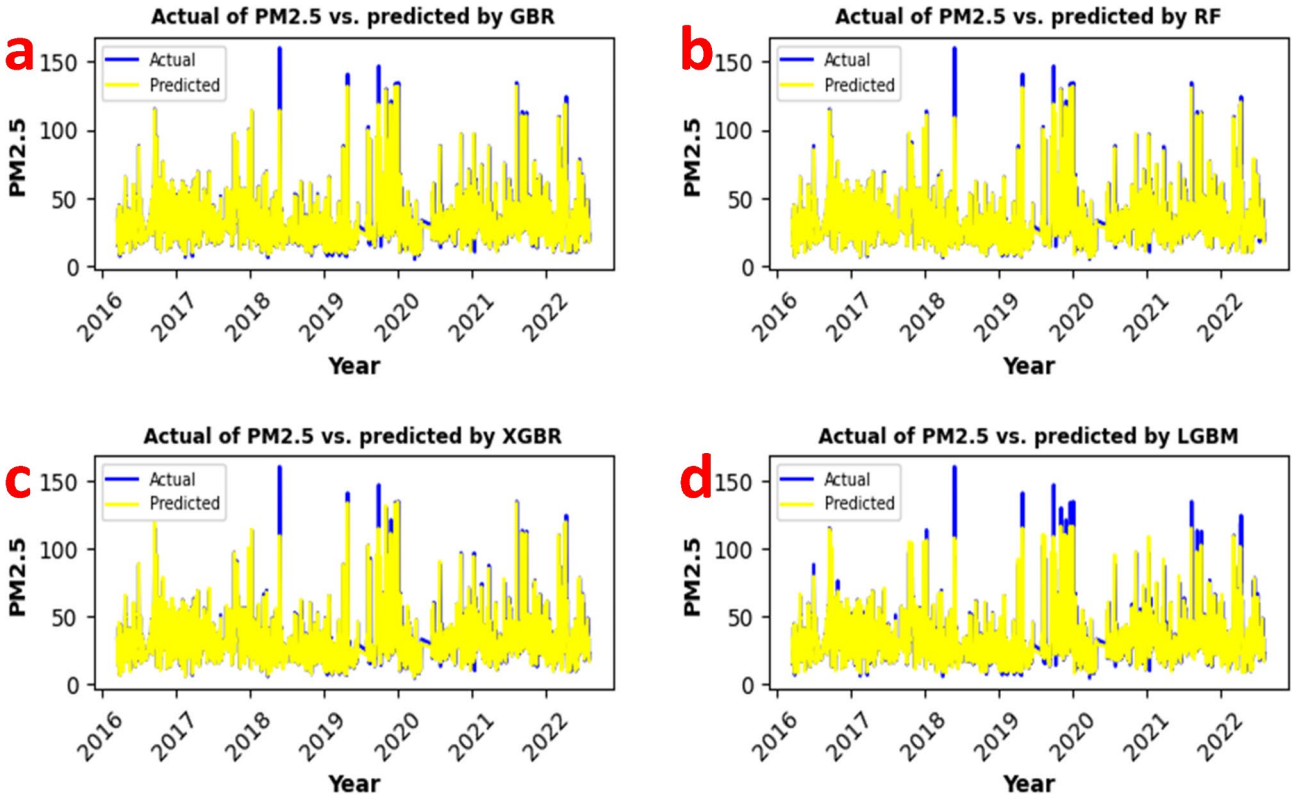

Fig. 6. Actual $PM_{2.5}$ values and the predicted $PM_{2.5}$ values by GBR (**a**), RF (**b**), XGBR (**c**) and LGBM (**d**).

of advanced air filtration systems in high-traffic areas. Furthermore, phasing out older, high-emission vehicles and enforcing stricter urban air quality monitoring policies would be crucial steps in reducing pollution levels. Given the significant impact of PM2.5 on public health, particularly in large metropolitan cities like Mashhad, high-accuracy PM forecasting is essential for informed environmental policymaking, public health monitoring, and air quality improvement initiatives.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to data confidentiality but are available from the corresponding author on reasonable request.

Received: 15 October 2024; Accepted: 25 February 2025
Published online: 08 March 2025

References

1. Suleiman, A., Tight, M. R. & Quinn, A. D. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmos. Pollut. Res.* **10**, 134–144 (2019).
2. He, B., Xu, H. M., Liu, H. W. & Zhang, Y. F. Unique regulatory roles of NcRNAs changed by PM2.5 in human diseases. *Ecotoxicol. Environ. Saf.* **255**, 114812 (2023).
3. McCarron, A. et al. Public engagement with air quality data: using health behaviour change theory to support exposure-minimising behaviours. *J. Expo Sci. Environ. Epidemiol.* **33**, 321–331 (2023).
4. Yang, H., Wang, W. & Li, G. Prediction method of PM2.5 concentration based on decomposition and integration. *Measurement* **216**, 112954 (2023).
5. Hardini, M., Sunarjo, R. A., Asfi, M. & Riza Chakim, M. H. Ayu Sanjaya, Y. P. Predicting air quality index using ensemble machine learning. *ADI J. Recent. Innov.* **5**, 78–86 (2023).
6. Wu, C. et al. A hybrid deep learning model for regional O_3 and NO_2 concentrations prediction based on Spatiotemporal dependencies in air quality monitoring network. *Environ. Pollut.* **320**, 121075 (2023).
7. Doan, Q. C., Chen, C., He, S. & Zhang, X. How urban air quality affects land values: exploring non-linear and threshold mechanism using explainable artificial intelligence. *J. Clean. Prod.* **434**, 140340 (2024).
8. Malakouti, S. M. Improving the prediction of wind speed and power production of SCADA system with ensemble method and 10-fold cross-validation. *Case Stud. Chem. Environ. Eng.* https://doi.org/10.1016/j.cscee.2023.100351 (2023).
9. Malakouti, S. M. Estimating the output power and wind speed with ML methods: A case study in Texas. *Case Stud. Chem. Environ. Eng.* **7**, 100324 (2023).
10. S266730532300073X.
11. Malakouti, S. M. et al. Advanced techniques for wind energy production forecasting: leveraging multi-layer Perceptron + Bayesian optimization, ensemble learning, and CNN-LSTM models. *Case Stud. Chem. Environ. Eng.* **10**, (2024).
12. Malakouti, S. M., Menhaj, M. B. & Suratgar, A. A. The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Clean. Eng. Technol.* **15**, 100664 (2023).

13. Sun, J., Gong, J. & Zhou, J. Estimating hourly PM2.5 concentrations in Beijing with satellite aerosol optical depth and a random forest approach. *Sci. Total Environ.* **762**, 144502 (2021).
14. Yang, L., Xu, H., Yu, S. & Estimating PM2.5 concentrations in Yangtze river Delta region of China using random forest model and the Top-of-Atmosphere reflectance. *J. Environ. Manag.* **272**, 111061 (2020).
15. Kim, B. Y., Lim, Y. K. & Cha, J. W. Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based machine learning algorithms. *Atmos. Pollut. Res.* **13**, (2022).
16. Gardner, M. W. & Dorling, S. R. Statistical surface Ozone models: an improved methodology to account for non-linear behaviour. *Atmos. Environ.* **34**, 21–34 (2000).
17. Berrocal, V. J. et al. A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration. *Atmos. Environ.* **222**, 117130 (2020).
18. Ghahremanloo, M. et al. Estimating daily high-resolution PM2.5 concentrations over Texas: machine learning approach. *Atmos. Environ.* **247**, 118209 (2021).
19. Tang, D. et al. Comparison of GOCI and Himawari-8 aerosol optical depth for deriving full-coverage hourly PM2.5 across the Yangtze river Delta. *Atmos. Environ.* **217**, 116973 (2019).
20. Wang, W. et al. Deriving hourly PM2.5 concentrations from Himawari-8 AODs over Beijing–Tianjin–Hebei in China. *Remote Sens.* **9**. https://doi.org/10.3390/rs9080858 (2017).
21. Heidari, E. A. et al. Assessing VOC emissions from different gas stations: impacts, variations, and modeling fluctuations of air pollutants. *Sci. Rep.* **14**, 16617 (2024).
22. Kardani-Yazd, N., Kardani-Yazd, N. & Mansouri Daneshvar, M. R. Strategic spatial analysis of urban greenbelt plans in Mashhad City, Iran. *Environ. Syst. Res.* **8**, 30 (2019).
23. Yousofpour, Y. et al. Ecosystem services and economic values provided by urban park trees in the air polluted City of Mashhad. *Sustain. Cities Soc.* **101**, 105110 (2024).
24. Miri, M. et al. Mortality and morbidity due to exposure to outdoor air pollution in Mashhad Metropolis, Iran. The AirQ model approach. *Environ. Res.* **151**, 451–457 (2016).
25. Williams, D. R. & Rast, P. Back to the basics: rethinking partial correlation network methodology. *Br. J. Math. Stat. Psychol.* **73**, 187–212 (2020).
26. Demir, E., Bilgin, M. H., Karabulut, G. & Doker, A. C. The relationship between cryptocurrencies and COVID-19 pandemic. *Eurasian Econ. Rev.* **10**, 349–360 (2020).
27. Chen, H. et al. Shield attitude prediction based on Bayesian-LGBM machine learning. *Inf. Sci. (Ny).* **632**, 105–129 (2023).
28. Asselman, A., Khaldi, M. & Aammou, S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact. Learn. Environ.* **31**, 3360–3379 (2023).
29. Tran, D. A. et al. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong delta, Vietnam. *Ecol. Indic.* **127**, 107790 (2021).
30. Prajwala, T. R. A comparative study on decision tree and random forest using R tool. *Int. J. Adv. Res. Comput. Commun. Eng.* **4**, 196–199 (2015).
31. Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N. & Taki, M. Y. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *J. Pet. Sci. Eng.* **208**, 109244 (2022).
32. Chicco, D., Warrens, M. J. & Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj Comput. Sci.* **7**, e623 (2021).
33. Althoff, D. & Rodrigues, L. N. Goodness-of-fit criteria for hydrological models: model calibration and performance assessment. *J. Hydrol.* **600**, 126674 (2021).
34. Liu, X., Zou, B., Feng, H., Liu, N. & Zhang, H. Anthropogenic factors of PM2.5 distributions in China's major urban agglomerations: A spatial-temporal analysis. *J. Clean. Prod.* **264**, 121709 (2020).
35. Mohammadi, M. et al. Relationships between ambient air pollution, meteorological parameters and respiratory mortality in Mashhad, Iran: a time series analysis. *Pollution* **8**, 1250–1265 (2022).
36. Harrison, R. M. Airborne particulate matter. *Philos. Trans. R Soc. A.* **378**, 20190319 (2020).
37. Aminiyan, M. M. et al. Occurrence and source apportionment of polycyclic aromatic hydrocarbons (PAHs) in dust of an emerging industrial City in Iran: implications for human health. *Environ. Sci. Pollut. Res.* **28**, 63359–63376 (2021).
38. Maciejczyk, P., Chen, L. C. & Thurston, G. The role of fossil fuel combustion metals in PM2. 5 air pollution health associations. *Atmos. (Basel).* **12**, 1086 (2021).
39. Bilal, M., Hassan, M., Tahir, D. B. T., Iqbal, M. S. & Shahid, I. Understanding the role of atmospheric circulations and dispersion of air pollution associated with extreme smog events over South Asian megacity. *Environ. Monit. Assess.* **194**, 82 (2022).
40. Pal, S. et al. Effects of lockdown due to COVID-19 outbreak on air quality and anthropogenic heat in an industrial belt of India. *J. Clean. Prod.* **297**, 126674 (2021).
41. Horn, S. A. & Dasgupta, P. K. The air quality index (AQI) in historical and analytical perspective a tutorial review. *Talanta* **267**, 125260 (2024).
42. Li, Y., Chen, Y., Karimian, H. & Tao, T. Spatiotemporal analysis of air quality and its relationship with meteorological factors in the Yangtze river Delta. *J. Elem.* **25**, (2020).
43. Kanawade, V. P. et al. What caused severe air pollution episode of November 2016 in new Delhi? *Atmos. Environ.* **222**, 117125 (2020).

Author contributions

M.S. and A.M. wrote the main manuscript. S.Z. prepared figures. A.M. analyzed the data and wrote the code.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.