



OPEN Application of the Lasso regularisation technique in mitigating overfitting in air quality prediction models

Abbas Pak¹, Abdullah Kaviani Rad², Mohammad Javad Nematollahi³✉ & Mohammadreza Mahmoudi⁴✉

As a significant global concern, air pollution triggers enormous challenges in public health and ecological sustainability, necessitating the development of precise algorithms to forecast and mitigate its impacts, which has led to the development of many machine learning (ML)-based models for predicting air quality. Meanwhile, overfitting is a prevalent issue with ML algorithms that decreases their efficacy and generalizability. The present investigation, using an extensive collection of data from 16 sensors in Tehran, Iran, from 2013 to 2023, focuses on applying the Least Absolute Shrinkage and Selection Operator (Lasso) regularisation technique to enhance the forecasting precision of ambient air pollutants concentration models, including particulate matter (PM_{2.5} and PM₁₀), CO, NO₂, SO₂, and O₃ while decreasing overfitting. The outputs were compared using the R-squared (R²), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and normalised mean square error (NMSE) indices. Despite the preliminary findings revealing that Lasso dramatically enhances model reliability by decreasing overfitting and determining key attributes, the model's performance in predicting gaseous pollutants against PM remained unsatisfactory (R²_{PM2.5} = 0.80, R²_{PM10} = 0.75, R²_{CO} = 0.45, R²_{NO2} = 0.55, R²_{SO2} = 0.65, and R²_{O3} = 0.35). The minimal degree of missing data presumably explained the strong performance of the PM model, while the high dynamism of gases and their chemical interactions, in conjunction with the inherent characteristics of the model, were the primary factors contributing to the poor performance of the model. Simultaneously, the successful implementation of the Lasso regularisation approach in mitigating overfitting and selecting more important features makes it highly suggested for application in air quality forecasting models.

Keywords Air pollution, Air quality prediction, Overfitting, Lasso regularisation, Machine learning

Air quality forecasting is an important analytical method and aims to raise a warning when pollution concentrations surpass a certain level^{1,2}. Precisely predicting air pollution levels is also essential for enacting efficient restrictions and protecting public health^{3,4}. Thus, examining and forecasting air pollution has garnered significant attention from scholars⁵⁻⁷. Machine learning (ML) techniques such as Random Forest, Extra Trees, XGBoost, and LightGBM have been attractive for forecasting applications over the past decade⁸⁻¹⁰ since they effectively perform a range of duties; thereby, many researchers have utilised ML techniques to predict air quality around the world. For instance, Castelli et al.¹¹ applied Support Vector Regression (SVR), a widely used ML technique, to estimate the concentration of pollutants and anticipate the air quality index (AQI) in California, USA. In another study, Liang et al.¹² predicted air quality index levels in several regions of Taiwan using AdaBoost, random forests, stacking ensembles, and support vector machines (SVMs) applying an 11-year dataset. He et al.¹³ and Guo et al.^{14,15} demonstrated that the ANN is effective in predicting monthly and daily aerosol concentrations in Liaocheng, Shanghai, and Chongqing, China, by identifying nonlinear relationships between the input and output variables. Furthermore, ML-based atmospheric transport models are commonly used to predict air pollution levels with high accuracy in terms of time and location. These models are beneficial

¹Department of Computer Sciences, Shahrekord University, Shahrekord, Iran. ²Department of Environmental Engineering and Natural Resources, College of Agriculture, Shiraz University, Shiraz 71946-85111, Iran.

³Department of Geology, Faculty of Sciences, Urmia University, Urmia 57561-51818, Iran. ⁴Department of Statistics, Faculty of Science, Fasa University, Fasa 74616-86131, Iran. ✉email: mj.nematollahi@urmia.ac.ir; mahmoudi.m.r@fasau.ac.ir

for regular air quality forecasts, typically predicting pollutant levels 1–3 days in advance¹⁶. In this context, Wang et al.¹⁷ developed an ML model that combines TROPOMI level-2 satellite observations with detailed meteorological data to forecast the levels of ground-level ozone (O₃) in California. ML techniques, when combined with spatiotemporal modelling, can offer more adaptable measures related to exposure. This approach has been investigated using various model architectures¹⁸. Wong et al.¹⁹ employed a Land Use Regression (LUR) model integrated with ML algorithms to evaluate the spatial-temporal fluctuations of particles that are 2.5 microns or less in diameter (i.e., PM_{2.5}). Their findings showed that the standard LUR model and the hybrid kriging-LUR model were able to recognise 58% and 89% of the fluctuations in PM_{2.5}, respectively. Therefore, the geographic pattern of air pollution has been comprehensively captured using LUR. Nevertheless, linear methods may prove difficult to implement when dealing with regional contexts and non-linear relationships²⁰. Simultaneously, enhancing the precision of conventional ML models, given the dynamic nature of pollutants and limited data availability, might pose challenges^{21,22}.

The lack of long-term data is a significant constraint for numerous studies, considering it will be essential for addressing seasonal fluctuations as well as additional variables. ML models based on short-term data might have limited generalisation capabilities when applied to various timeframes or regions^{12,23}. Consequently, forecasting air quality is a challenging endeavour because of the intricate characteristics, instability, and significant fluctuations in pollutants over time and location¹⁶. The effectiveness of the mathematical models is constrained by flaws in the emission inventory and biases in the beginning and boundary situations, along with shortcomings in the present physical and chemical schemes. The extent of disparity between the anticipated exposures produced by multiple models and one model that yields trustworthy projections is uncertain²⁴. Prior investigations employed ML and statistical models to categorise and predict air pollution. Nevertheless, the intricate nature of the air pollution dataset makes these algorithms inefficient for classifying and predicting. ML-based models encounter problems such as poor data preprocessing, class inequality concerns, data splitting, and hyperparameter tuning²⁵.

An important issue that affects the methods mentioned above is overfitting, which occurs when models achieve favourable results on training data but perform poorly in generalising their findings onto new and unseen data. The overfitting issue occurs when a model learns from noise and unrelated patterns of the training data with poor predictive performance. Lopez et al.²⁶ described overfitting as an issue where the statistical ML model learns much about noise as well as signal, which is present in the training data. Overfitting also remains an issue even in contexts involving a few dimensions, especially when there is a failure to make a correlation between the result and predictor variables robust²⁷.

Implementing regularisation techniques is crucial in improving model performance by decreasing overfitting. The performance of different regularisation techniques, such as the Frobenius norm, nuclear norm, and Lasso, has been explored to enhance the accuracy of air quality prediction^{23,28}. Lasso regularisation applies a penalty to the absolute value of regression coefficients, which reduces less important feature coefficients to zero²⁹. This process contributes to feature selection³⁰, reduction of overfitting, and enhancement of the interpretability of the model. Several studies in Iran have shown that air pollution harmfully impacts the physical and mental health of citizens, reducing labour productivity and student academic performance³¹. Prolonged exposure to ambient PM_{2.5} and O₃ significantly increased mortality in Tehran, with ischaemic heart disease being the most responsible cause³², highlighting the necessity of air pollution modelling to demonstrate its behaviour. The intention is to utilise Lasso's capacity to improve the simplicity and reliability of models to create reliable prediction models that can estimate pollution concentrations under various scenarios. Hence, the main goal of this study is to examine the utilisation of Lasso regularisation in the context of forecasting air pollution factors in the Tehran megacity, which is the most polluted city in Iran.

The rationale for employing Lasso regression in this study is rooted in its ability to handle high-dimensional datasets and perform effective feature selection. Here, we used an extensive collection of features, including concentrations of key pollutants as well as meteorological variables from 16 sensors in Tehran, spanning 10 years (2013–2023). Given the complexity of the dataset, which includes variables with potential multicollinearity and varying degrees of importance, Lasso regression was particularly suitable due to its ability to identify the key factors that influence air pollutant concentrations enhances our understanding of the underlying processes. While Lasso is inherently a linear method, it serves as an essential baseline for evaluating linear relationships within the data. In the context of air pollution forecasting for Tehran, where some relationships—such as the influence of meteorological variables on pollutant dispersion—can often be approximated as linear, Lasso regression provides a robust and interpretable modelling approach.

Materials and methods

Study area

Tehran, located at coordinates 35°41' N and 51°26' E, serves as the political centre and largest metropolis of Iran. Tehran, with an approximate area of 730 km², has an approximate population density of 10,555 individuals/km². The region's altitude varies between 900 and 1800 m above sea level. The northern area experiences a cold and arid climate, while the southern portion is characterised by a hot and dry climate. The city faces a yearly temperature range of 15° to 18 °C, with a variance of around 3° in different Sectors^{33,34}.

Data acquisition

The study dataset includes the air pollutants CO, O₃, NO₂, SO₂, PM₁₀, and PM_{2.5}. The concentrations of atmospheric PM_{2.5} and PM₁₀ (µg/m⁻³), O₃ (ppbv), NO₂ (ppbv), SO₂ (ppbv), and CO (ppmv) were measured using beta-attenuation (Met One BAM-1020, USA; Environment SA, MP 101 M, France), UV-spectrophotometry (Ecotech Serinus 10 Ozone Analyser, Australia), chemiluminescence (Ecotech Serinus 40 Oxides of Nitrogen

Analysers, Australia), and ultraviolet fluorescence (Ecotech Serinus 40 Ox), respectively. The air pollutants' real-time data from sensors is transmitted to the quality control unit. The weather sensors' location is based on a number of effective factors on air pollution, such as elevation from the ground, dominant wind currents, interval to polluting sources (e.g., industrial units and high-traffic areas), and land use. Preferably, weather sensors are situated in places where precisely indicate the properties of the immediate medium. The weather sensors in the city mainly operate in areas distinguished by high traffic¹⁰. A reference laboratory at the Iran's Sharif University of Technology is responsible for periodically checking the operational efficiency, precision, and accuracy of sensors. The sensors' data is ceaselessly received by Tehran Air Quality Control Company (TAQCC) via optical fibre. The outputs are daily publicised, complying with the validation process. Among the 22 weather sensors in Tehran, only 16 cases have been active, and data has been collected since 2013. The dataset covers ten years (Jan 2013–Dec 2023), ensuring sufficient temporal coverage for training and testing the predictive models. A complete set of 40,172 data values pertaining to 11 parameters was gathered.

Meteorological data was obtained from the Weather Underground archive, which can be explored at <https://www.wunderground.com>. There are almost 6,000 automated weather gauges located at airports worldwide. These sensors report their measurements every 1, 3, and 6 h. The weather indicators included in the present research are temperature (T in °C), relative humidity (RH in %), wind speed (WS in km/h⁻¹), dew point (DW in °C), and air pressure (AP in hPa (hectopascal)). The weather indicators are collected to capture the environmental conditions that influence the dispersion and concentration of air pollutants. The international airport linked to this system is Mehrabad Airport in Tehran, located at coordinates 35°41' N, 51°19' E. Figure 1 represents the studied geographical region, the air quality and meteorological stations, and several gadgets used in this study.

Data preprocessing

To ensure optimal results, it is required to preprocess the data before applying the Lasso method. This involves handling missing values and standardising variables if necessary³⁵. Initially, the dataset contained a limited number of missing values. We adopted an interpolation technique that predicts the missing values by estimating them based on the surrounding data points. This adjustment ensures that the data integrity is maintained

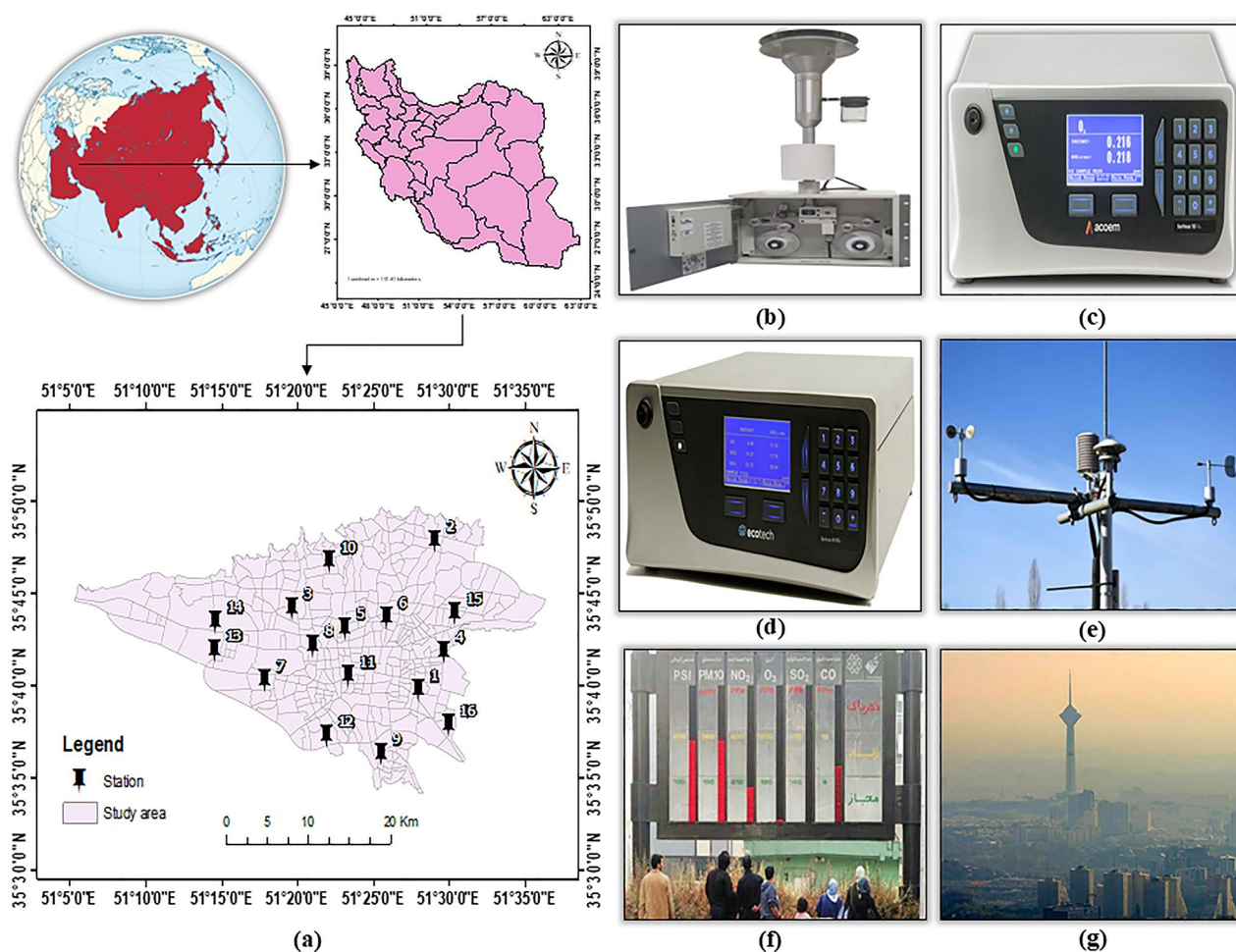


Fig. 1. (a): Map showing the position of the research region and monitoring locations; (b): Met One BAM-1020, USA (for PM); (c): Ozone Analyser (Ecotech Serinus 10, Australia); (d): NO₂ analyser (Ecotech Serinus 40, Australia); (e): The weather station; (f): Public air quality monitors; (g): Severe air pollution in Tehran.

without introducing significant biases while avoiding the information loss associated with the direct elimination of missing values. Then, due to the sensitivity of the Lasso method to the scale of features, we need to standardise the variables³⁶. To do this, each feature is subtracted from the mean and divided by its standard deviation so that it has a mean of zero and unit variance. So, variables with larger scales are prevented from dominating the regularisation process³⁷. Further, to accurately evaluate the model's performance, it is required to split the data into training and testing sets³⁸.

Methods

Overfitting in ML arises when a model acquires excessive knowledge about the intricacies and random fluctuations in the training data^{39,40}, leading to detrimental effects on its performance when applied to new data (Fig. 2). Essentially, the model becomes excessively intricate and seizes the “noise” or arbitrary variations within the training data instead of the fundamental pattern. When tested on the training data, this outcome yields high accuracy, though it fails to generalise well to unseen data. We first conduct a primary prediction for the CO variable to show this issue by employing various ML techniques. In our analyses, 80% of the data were used for model training and 20% for model testing, and the trace of R-squared (R^2) values was plotted for different models (Fig. 3). To ensure robust evaluation and consistency, the results in Fig. 3 were generated using a k-fold cross-validation approach—high differences between R^2 in the train and test models hallmark overestimation in the prediction of CO analyses. Additionally, to illustrate the phenomenon of overfitting, we have included several evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Normalised Mean Squared Error (NMSE) in Table 1. The results show lower error metrics on the training set compared to the test set, indicating the presence of overfitting to varying degrees across all the evaluated models. In this paper, the Lasso regularisation technique will be employed to counteract overfitting by incorporating a penalty equivalent to the absolute value of coefficient magnitudes into the loss function. This penalty term encourages the model to favour simpler solutions, effectively shrinking certain coefficients to zero and facilitating feature selection. By reducing model complexity, Lasso regularisation assists in mitigating overfitting and performing automatic feature selection.

Working principle of Lasso regularisation

The aim of linear regression is to minimise the loss function, generally represented by the sum of squared errors (SSE) (Eq. 1):

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

which can be stated as Eq. 2:

$$SSE = \sum_{i=1}^n \left(y_i - \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) \right)^2 \quad (2)$$

where n represents the number of observations, p denotes the number of variables that are available in the dataset, β_0 is the intercept or constant term, which represents the value of the dependent variable y when all the independent variables x_{ij} are zero; β_j for $j=1, 2, \dots, p$ are the regression coefficients for each independent variable x_{ij} . These coefficients indicate the magnitude and direction of the relationship between the independent variable x_{ij} and the dependent variable y , and x_{ij} represent the value of the j -th variable for the i -th observation ($i=1, 2, \dots, n$ and $j=1, 2, \dots, p$). Lasso regularisation introduces an additional penalty term to the loss function as follows (Eq. 3):

$$\sum_{i=1}^n \left(y_i - \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (3)$$

where λ (lambda) can take various values as follows:

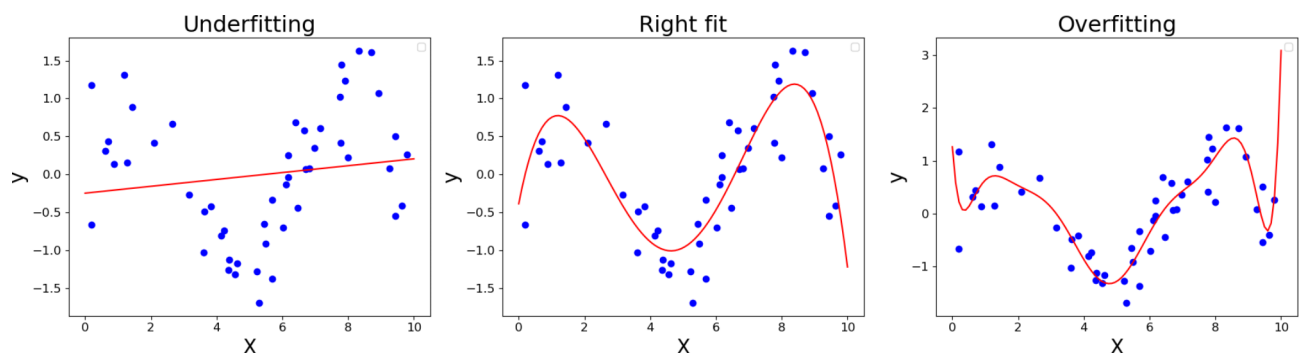


Fig. 2. Plot of fitting different models.

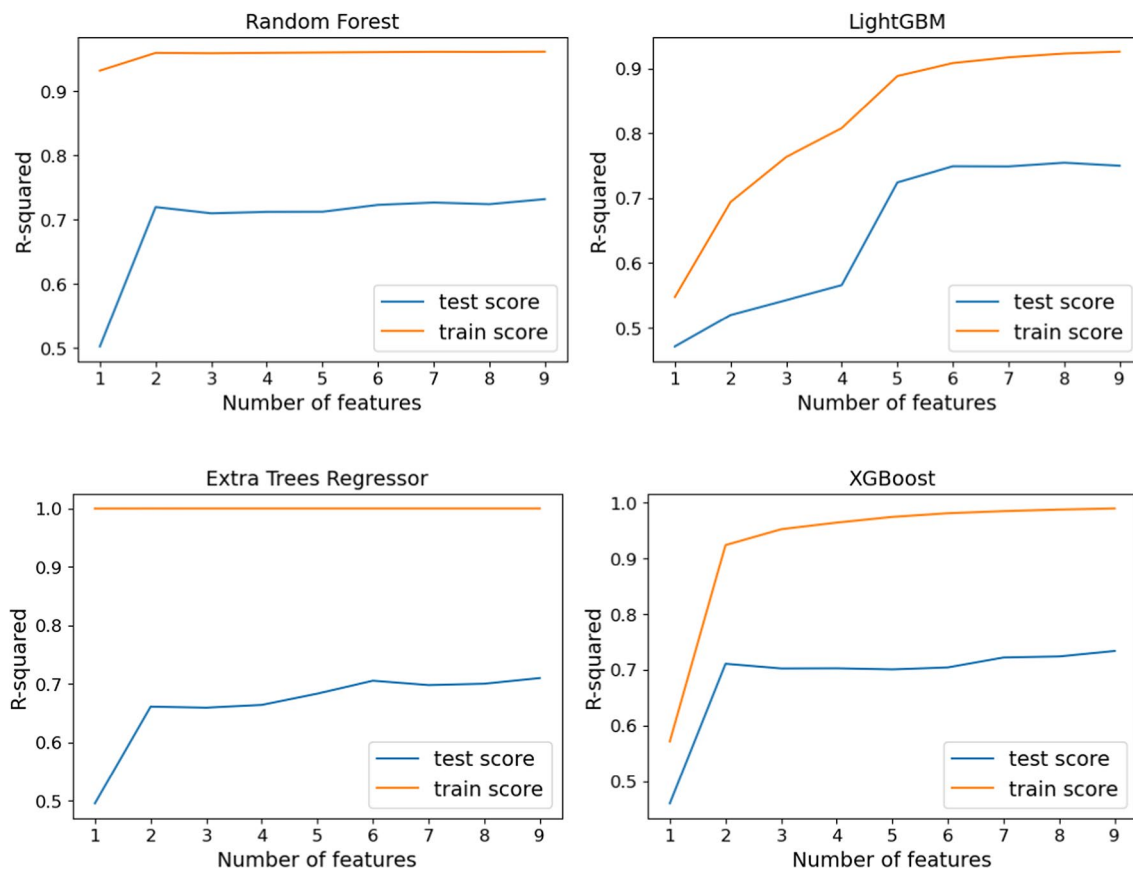


Fig. 3. Accuracy of different ML techniques, including Random Forest, LightGBM, Extra Trees Regressor, and XGBoost, in predicting CO.

Metric	MAE		MSE		RMSE		NMSE	
	Train	Test	Train	Test	Train	Test	Train	Test
Random Forest	0.9993	4.0408	1.9112	39.2539	1.3824	6.2653	0.0189	0.3689
LightGBM	2.3043	4.2135	8.9180	42.2110	2.9863	6.4970	0.0882	0.3931
Extra Trees Regressor	0.0009	3.9593	0.0017	38.1330	0.0412	6.1752	0.0086	0.3598
XGBoost	1.4622	4.0853	3.7635	40.0094	1.9397	6.3253	0.0372	0.3750

Table 1. Performance metrics of the ML models.

- $\lambda = 0$: Same coefficients as ordinary least squares linear regression.
- $\lambda = \infty$: All coefficients are zero.
- $0 < \lambda < \infty$: Coefficients are between 0 and that of least squares linear regression.

The magnitude of λ determines the amount of penalty. The larger the value of λ , the more coefficients are forced to be zero to simplify the model; the smaller the value of λ , the lesser the impact, thus enabling most coefficients to remain almost as they were. An appropriate value of lambda should be chosen carefully to get the right type of sparsity. A common procedure for selecting the lambda is based on cross-validation, which is a resampling technique in which the training data are divided into multiple subsets or folds. The Lasso model is trained on a subset of the folds and evaluated on the remaining fold. This process is carried out for different values of lambda, and the lambda minimising the model's error—e.g., mean squared error or cross-validated error—is chosen. One of the most common cross-validation techniques applied to choose lambda is k-fold cross-validation⁴¹. In k-fold cross-validation, the training data is divided into k equal-sized subsets or folds as shown in Fig. 4⁴². In brief, the Lasso model is trained on k folds and evaluated on the remaining fold, repeated for each fold, and the average error across all folds is computed for each lambda value. The lambda value with the lowest average error is selected as the optimal lambda.

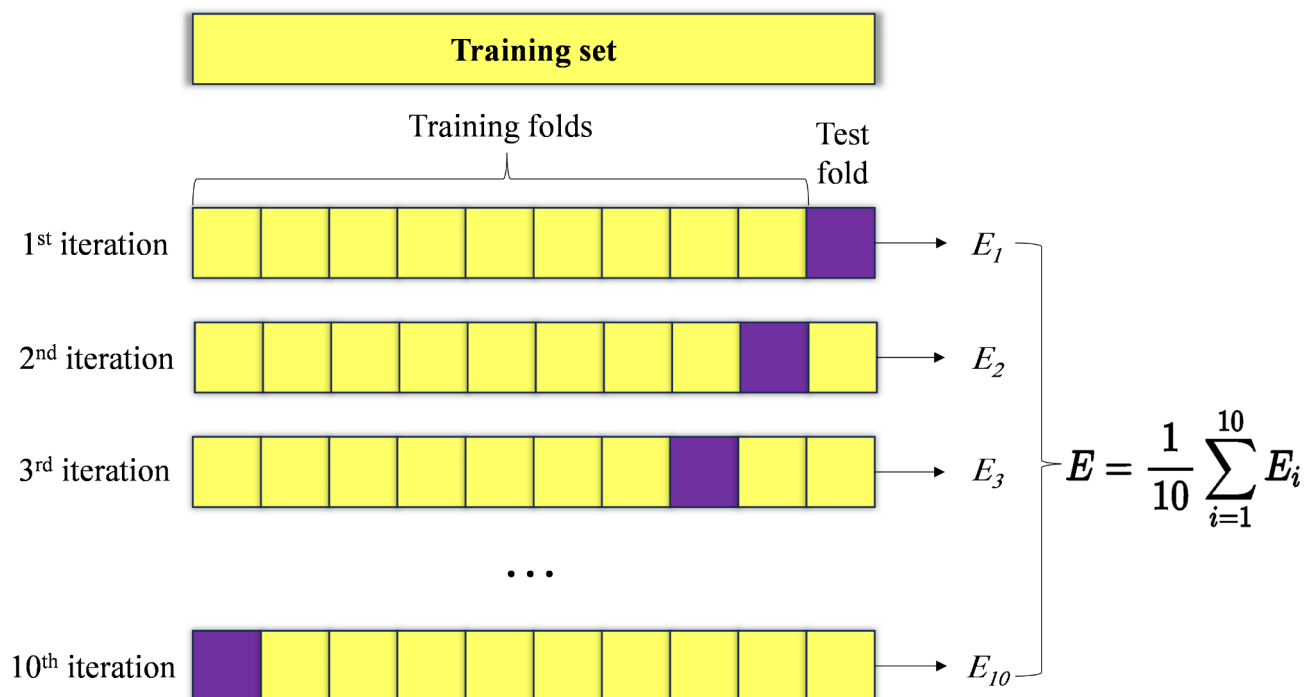


Fig. 4. The diagram of 10-fold cross-validation, adapted from⁴³.

Variable selection

Not only does the Lasso method provide predictions by model fitting of regression, but it also conducts feature selection on the data, simultaneously⁴⁴. The Lasso will automatically identify less important features and exclude these features from the final model by shrinking their coefficients towards zero. The variable selection property of Lasso has a number of practical implications. It provides better interpretability because the non-zero coefficients in this method indicate the most influential features for the target variable prediction. Secondly, Lasso variable selection may afford more efficient or parsimonious models. By excluding redundant or irrelevant features, the model's complexity is reduced, which can improve its generalising performance on unseen data. Thirdly, in Lasso, the non-zero coefficients identify the most important features w.r.t. the problem at hand and yield insight into the underlying structure of the data and relationships among the subjects. The process of researching the changes in the coefficients with the change in the level of regularisation has been referred to as tracing the lambda path or regularisation path. The lambda path is a sequence of models derived by changing the regularisation parameter, lambda (λ).

Evaluation metrics

Evaluating the model's performance to assess its predictive accuracy and generalisation capabilities is necessary while implementing the Lasso method or any other ML technique. Different evaluation metrics exist for evaluating the efficacy of the Lasso model. In this study, the most commonly used evaluation metrics, MSE, RMSE, normalised mean squared error (NMSE), and coefficient of determination (R^2) are employed, similar to Doreswamy et al.⁴⁵ and Guo et al.⁴⁶. The MSE is one of the most frequent metrics, which accounts for the average of squared differences between predicted and actual values, hence quantifying the overall quality of the model's predictions. To compute the MSE, the difference between each predicted value and its corresponding true value is squared first, and then the average of these squared differences is taken. MSE characterises the goodness of the fit of the Lasso model to the data, with lower values indicating better performance. NMSE is a variation of the MSE that provides a relative measure of prediction accuracy, considering the scale of the target variable. It's particularly useful when comparing models operating on different scales or units. It is computed as the MSE divided by the variance of the true values. R^2 is a metric that is commonly used for evaluating what percentage of variance in the target variable is predictable from the independent variables. It is computed as the quotient of explained variance over the total variance of the target variable. Given the Lasso method, R^2 conveys the overall measure of how well-selected features explain the variation in the target variable. A low value of MSE and NMSE will indicate that this Lasso model is giving rather accurate predictions, while a high R^2 will mean that the selected features describe a substantial part of the variability in the target variable.

Software and tools

Data analysis and predictive modelling were implemented using Python 3.11. The experiments were conducted on a machine equipped with an Intel Core i3-1115G4 CPU @ 3.00 GHz, 4 GB RAM, running Windows 10, and the time required to run the models ranged from 0.8397 to 1.9027 s. The Scikit-learn library is utilised for ML

procedures, Pandas is used for data manipulation, and Matplotlib is used for data visualization. Using these tools, statistical analyses and cross-validation procedures were facilitated, ensuring robust and reproducible results. The study area map and location of weather stations were provided using ArcGIS version 10.3.

Results

Finding lambda

The ideal tuning parameter λ was established by 10-fold cross-validation. The optimal λ was selected based on minimising the cross-validated mean squared error (MSE) for different target variables, as shown in Fig. 5. For the Lasso model, we used the default maximum number of iterations of 1000 to ensure sufficient iterations for convergence. The convergence tolerance was set to its default value of 1×10^{-4} , which served as the stopping criterion for the optimisation process. The optimisation was performed using a cyclic coordinate descent algorithm, which iteratively updates the coefficients by cycling through each feature, one at a time, while holding the other coefficients fixed. The final Lasso model will be trained using the indicated “1 standard error rule” (λ -1se) values. The charts reveal the optimal λ for balancing model complexity and prediction accuracy.

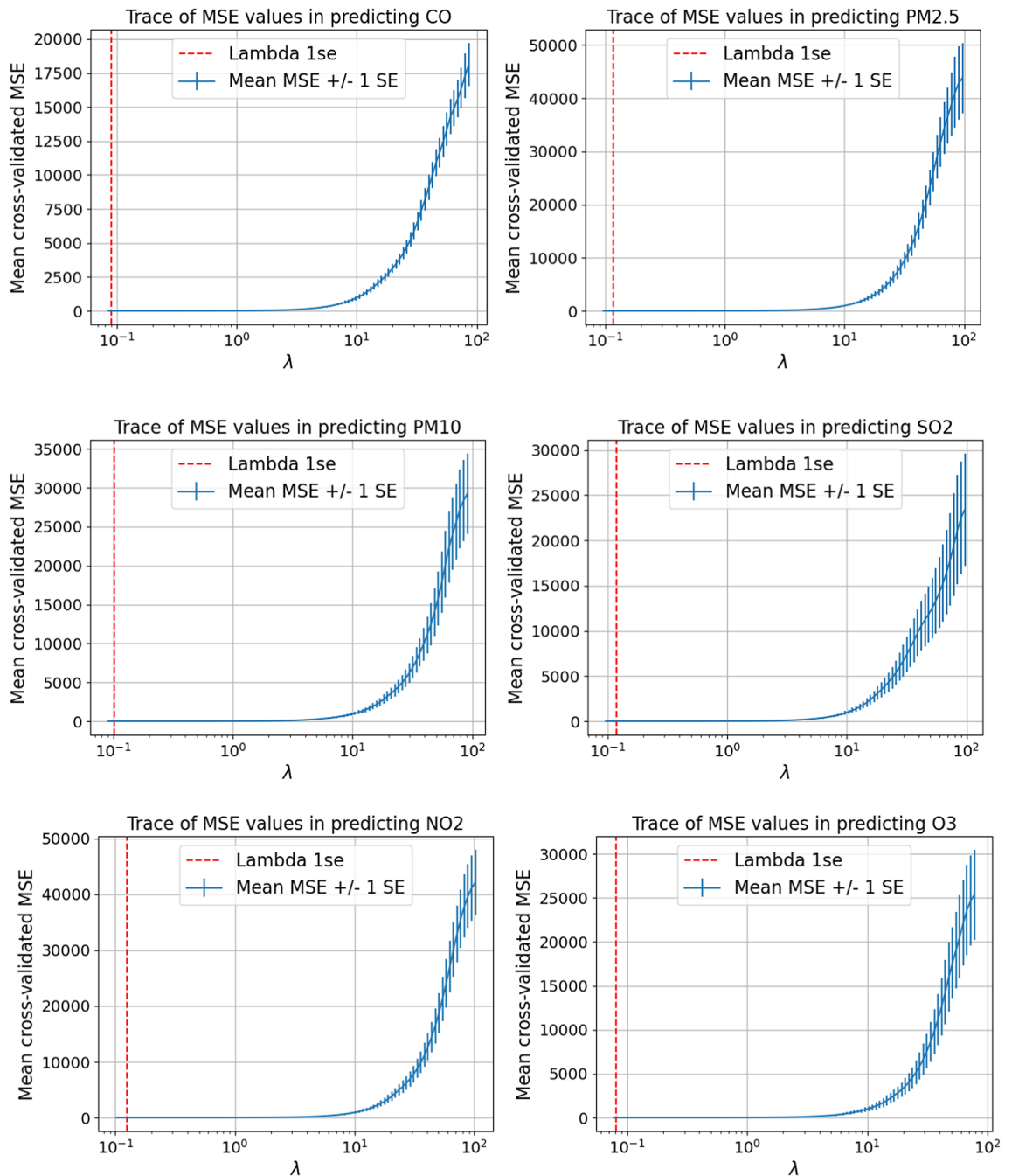


Fig. 5. Plots of cross-validation MSEs against λ values for air pollutants.

Predictors	CO	PM _{2.5}	PM ₁₀	O ₃	NO ₂	SO ₂	T	DW	RH	WS	AP
CO	-	2.837	-0.799	3.632	2.917	1.801	1.463	×	-1.733	-1.108	-0.225
PM _{2.5}	2.928	-	17.405	-1.14	4.674	5.085	×	-0.976	5.283	-1.429	×
PM ₁₀	-1.246	19.422	-	×	-1.589	-2.63	×	2.507	-5.476	1.541	0.154
O ₃	-6.481	-2.308	×	-	8.031	×	15.004	0.393	-0.754	-0.100	-1.871
NO ₂	3.828	7.838	-2.750	6.204	-	-1.40	-0.502	-2.263	×	-1.556	0.866
SO ₂	1.751	5.241	-2.152	×	-1.196	-	-2.227	×	-1.661	-1.173	-0.150

Table 2. Coefficients of selected features by the Lasso model.

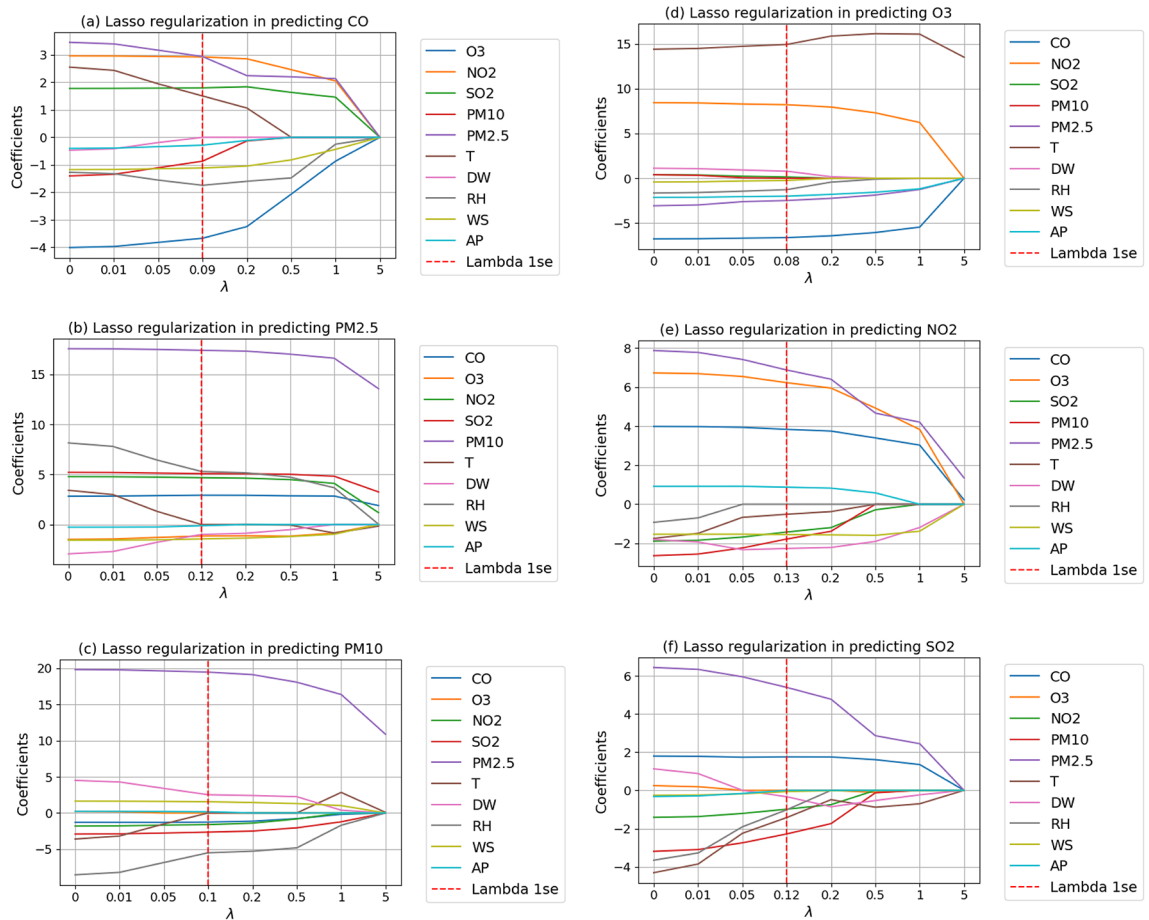


Fig. 6. Plots of coefficient paths for different λ values.

The scale of λ , shown on the x-axis, indicates that as λ grows, the adjustment impact intensifies, penalising more coefficients. The y-axis displays cross-validation average MSE values, with lower values indicating better model performance. The blue line represents the mean MSE, whereas the shaded region is the mean MSE \pm 1SE, indicating estimated variability or uncertainty. In addition, the red dashed line shows the ideal λ value using the λ -1se, which picks the model with the most consistent adjustment and an error within one standard error of the minimal error. To balance model complexity and prediction accuracy and also minimise overfitting, the ideal λ (λ -1se) for all pollutants is about 0.1. SO₂ had the highest mean cross-validated MSE values, while CO had narrower MSE ranges. Overall, mean cross-validated MSE values with narrow error bars around ideal λ are more reliable, whereas those with large error bars at extreme λ values reflect model uncertainty and variability.

Lasso regularisation and feature selection

The Lasso model coefficients of drivers affecting air pollutants, including CO, O₃, NO₂, SO₂, PM₁₀, and PM_{2.5}, are presented in Table 2. Further, the coefficient trajectory for various λ values is displayed in Fig. 6, demonstrating Lasso selective feature removal as λ grows. The x-axis (λ) displays the regularisation factor, showing how Lasso selectively eliminates features as λ increases. The coefficient routes demonstrate how changes in characteristics occur as λ rises. The coefficients of less relevant characteristics decline to 0 as λ grows, whereas critical traits

retain importance. Based on the optimal λ values (λ -1se) shown in Fig. 6 and the coefficient paths in Fig. 7, we can identify the irrelevant features in the prediction process for each dependent variable. In Fig. 6a, DW is the first feature to shrink to zero, indicating its diminishing importance in predicting CO. For $PM_{2.5}$ prediction (Fig. 6b), the features T and AP can be excluded. In Fig. 6c, the prediction of PM_{10} eliminates the features T and O_3 . When predicting O_3 (Fig. 6d), $PM_{2.5}$ and SO_2 are deemed irrelevant. In Fig. 6e, RH shrinks rapidly to zero, suggesting it has no impact on NO_2 prediction. Finally, Fig. 6f shows that DW and O_3 are removed from the prediction of SO_2 . From a physicochemical reaction perspective, the DW can affect precipitation, thereby considerably decreasing pollutant concentrations. Increased precipitation intensities correlate with decreased levels of air pollutants, such as PM_{10} , SO_2 and NO_2 , varying between 15% and 35% of the decrease relative to arid conditions⁴⁷. DW frequently corresponds with other meteorological factors, including T and RH. Such associations may result in its deletion from predictive models such as Lasso if considered superfluous^{48,49}.

The Lasso regularisation examination across the pollutants indicates major factors associated with every forecasting model. The coefficients of selected features by the Lasso model for air pollutants are presented in Table 2. O_3 was the most significant factor influencing atmospheric CO concentrations, with a strong positive coefficient (3.632). CO also had a positive coefficient with NO_2 (2.917), $PM_{2.5}$ (2.837), SO_2 (1.801), and T (1.463). $PM_{2.5}$ was most affected by PM_{10} , with a strong positive coefficient of 17.405, followed by RH (5.283), SO_2 (5.085), NO_2 (4.674), and CO (2.928), respectively. The most effective factors on atmospheric PM_{10} concentrations were

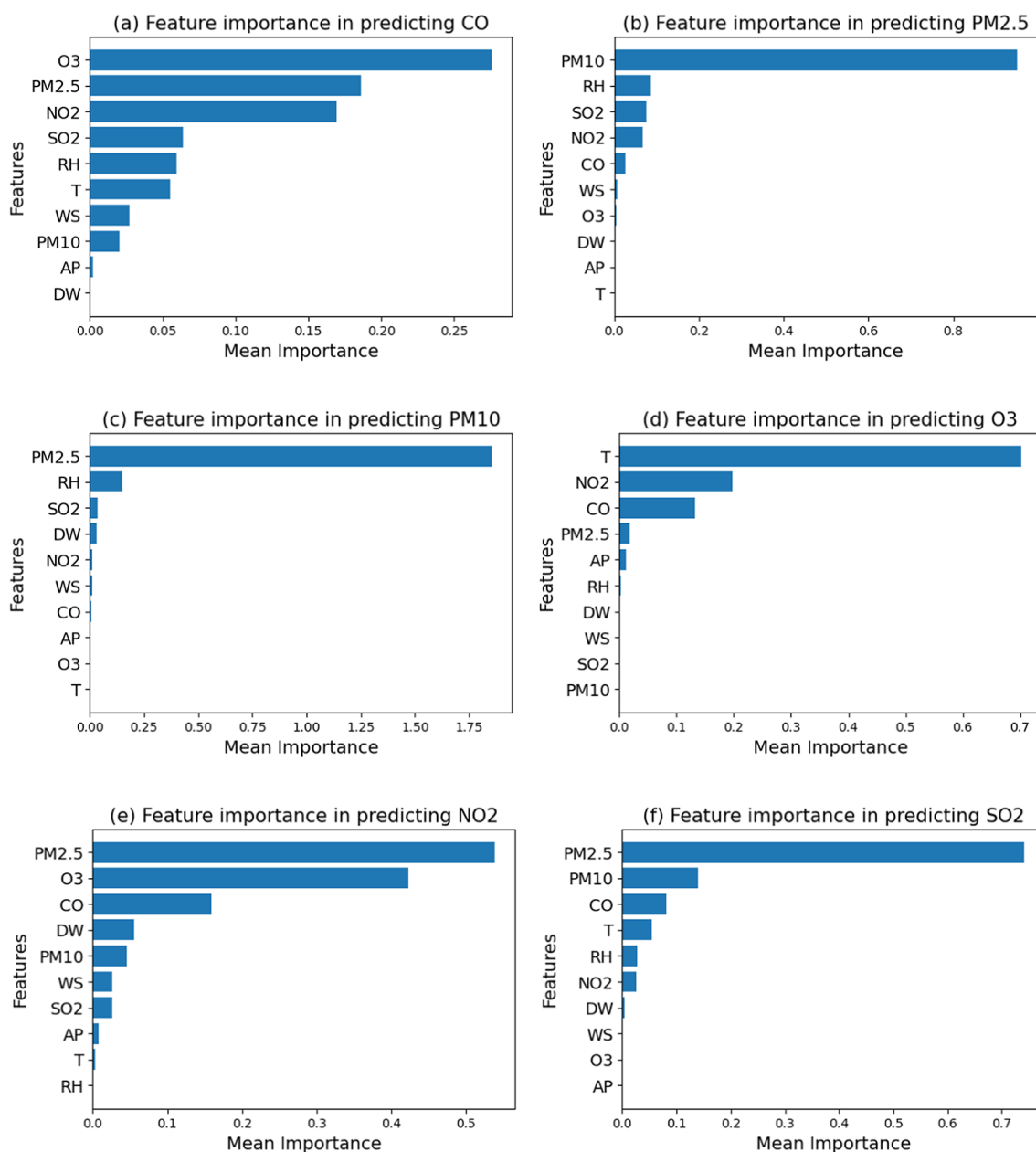


Fig. 7. Plots of permutation feature importance for air pollutants.

PM_{2.5}, DW, WS, and AP, with positive coefficients of 19.422, 2.507, 1.541, and 0.154, respectively. T had the highest positive coefficient (15.004) with O₃, followed by NO₂ (8.031) and DW (0.393), respectively. The most important factors affecting NO₂ concentration included PM_{2.5}, O₃, CO, and AP, respectively, with positive coefficients of 7.838, 6.204, 3.828, and 0.866. SO₂ had positive coefficients with PM_{2.5} (5.241) and CO (1.751).

To better understand the significance of individual features in predicting pollutant levels, we employed permutation feature importance. This method evaluates the contribution of each feature by measuring the drop in model performance after permuting the values of that feature while keeping others constant. The importance scores were calculated as the mean decrease in performance across the number of permutations. The feature importance for each pollutant is visualised using bar charts, as shown in Fig. 8, which provides a visual confirmation of the above results. This visualisation highlights the relative contributions of the predictors to the model's performance, providing insights into the factors most strongly associated with pollutant concentrations.

Model performance

The essential performance metrics of the Lasso model are presented in Table 3. The MAE quantifies the average absolute value of mistakes, irrespective of their direction; accordingly, O₃ had the greatest discrepancy, with an MAE of 11.973, compared to SO₂ with the smallest discrepancy (5.386). The average squared difference between actual and forecasted values is quantified by the MSE, making it particularly sensitive to big inaccuracies. Among the analysed variables, O₃ had the largest deviation (MSE = 277.998); on the contrary, CO had the lowest deviation (MSE = 57.574). The highest and lowest RMSE, the squared MSE, belonged to O₃ (16.673) and CO

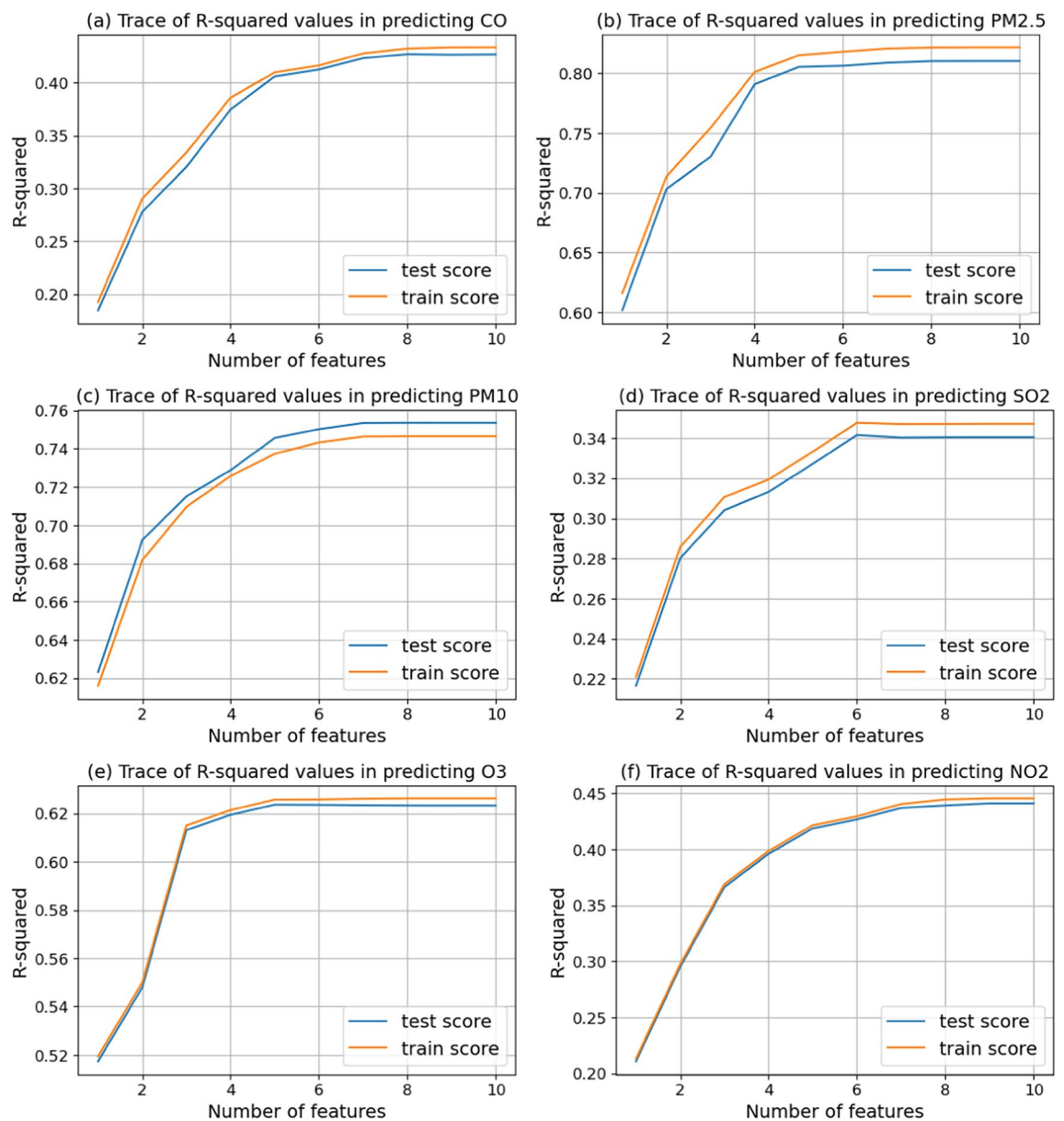


Fig. 8. Plots of accuracy scores for different numbers of features.

Metric				
Predicted Parameter	MAE	MSE	RMSE	NMSE
CO	5.979	57.574	7.589	0.559
PM _{2.5}	8.003	109.023	10.441	0.165
PM ₁₀	5.882	69.787	8.345	0.31
O ₃	11.973	277.998	16.673	0.399
NO ₂	9.083	124.909	11.176	0.565
SO ₂	5.386	65.897	8.117	0.718

Table 3. The Lasso model performance metrics.

(7.589), respectively. The NMSE, the normalised mean squared error, which indicates the level of error in the data and among the pollutants, was the highest for SO₂ (0.718) and the lowest for PM_{2.5} (0.165).

The overfitting in the Lasso regression models (Fig. 7) was analysed by comparing the R² trace to the number of features in the training and test data sets. To compare the model's performance on test and training data, the features were added to the model, and then the R² value was calculated after every addition. The model was then evaluated using experimental data. The plots show that adding features increases R² values for both training and test data, i.e., higher values in the y-axis imply an improved fit. However, a tight alignment of values demonstrates that the model is not overfitted. The training and test scores for CO (Fig. 7a) improve with feature count and maintain around the R² value of 0.45. Nevertheless, the CO model performance starts to drop significantly following seven features, and the modest discrepancy between training and test results suggests minor overfitting. The training and test scores rise quickly and remain at the R² value of 0.80 for PM_{2.5} (Fig. 7b). Additional features do not improve the model performance significantly after adding eight features in the PM_{2.5} plot, and a minimal score discrepancy implies good model generalization. The training and test scores for PM₁₀ (Fig. 7c) improve until the R² value of 0.75, and then the model efficiency remains unchanged after adding eight features. In the PM₁₀ plot, a modest score difference demonstrates a well-generalised model for PM₁₀. In the SO₂ plot (Fig. 7d), the training and test scores rise to the R² range of 0.65 to 0.7 and then subsequently stabilise. Additional features cannot increase the model's efficiency after adding six features in the SO₂ plot, and a minor score difference shows that the model does not overfit. Regarding O₃ (Fig. 7e), both training and test scores noticeably rise until the R² value of 0.65, and then the performance does not improve following the addition of five features. Small gaps pointed out strong generalisation in the prediction of O₃. The training and test R² scores for NO₂ (Fig. 7f) significantly rise to 0.45, and then the model performance improvements do not change after adding eight features. Moreover, a small distance indicates no overfitting for NO₂. Overall, the R² values for Tehran's air pollutants decrease in the order of PM_{2.5} (0.8) > PM₁₀ (0.75) > SO₂ (0.65) > NO₂ (0.55) > CO (0.45) > O₃ (0.35). Hence, CO and O₃ with lesser values reflect greater forecasting challenges. Decreased returns were identified across all pollutants upon integrating 7–8 features, defining them as the key drivers. A narrow training-test score gap pointed to strong generalisation and minimal overfitting. It is important to note that the above results were obtained using cross-validation during the implementation of the Lasso models. By repeatedly training and testing the model on different subsets of the data, the consistency of the model's predictive accuracy was confirmed. The results demonstrate that the models remain robust and perform reliably under variations in the input data.

Discussion

The model efficiency

In contrast to expectations, the findings revealed major disparities in the modelling outcomes for PM and air pollutant gases. The high R² values in the PM prediction ML models (>0.70 for PM₁₀ and >0.80 for PM_{2.5}) remarked that the model explains a large proportion of the variance in the PM concentration data (Table 3). Several factors could contribute to this high R². First, in the context of modelling, significantly correlated characteristics may predict PM_{2.5} values in the model, which is frequently accomplished by feature selection or domain understanding that detects PM concentrations. The ML algorithms have the power to accurately represent intricate and non-linear interactions⁵⁰ between input parameters and PM levels. The application of LGBM and other ML methodologies has demonstrated favourable outcomes in predicting surface concentrations of NO₂ and O₃, attaining R² values of up to 0.91 for O₃ and 0.83 for NO₂ in China^{51,52}. The Prophet forecasting model has been employed in Seoul to anticipate air pollution levels, exhibiting enhanced performance relative to conventional models⁵³. Furthermore, deep learning models, particularly Long Short-Term Memory (LSTM) networks, have been utilised to forecast air quality by amalgamating data from several sources, encompassing meteorological and pollutant data. These models have demonstrated enhanced predictive accuracy for air pollutants such as PM_{2.5}, CO, NO₂ and O₃ in Beijing⁵⁴ and Shanghai⁵⁵, with R² values of up to 0.86. The high-quality, noise-free data with a lot of observations may help the algorithm train and forecast^{56,57}. Many empirical studies have shown that the noise in the dataset dramatically leads to decreased classification accuracy and poor prediction results⁵⁸. Additional data makes correlations and patterns clearer. The evaluation of the Tehran dataset indicated that PM, particularly PM_{2.5}, was measured by a greater number of sensors. In addition, regularisation methods, such as Lasso regularisation, as applied in the present study, help prevent overfitting by penalising large coefficients and promoting simpler models and can increase generalisation to unseen data and contribute

to a high R^2 value in the test set^{59,60}. Hybrid ML models were also employed by Qiao et al.⁶¹ and Cheng et al.⁶² to develop a highly efficient $PM_{2.5}$ prediction model in China.

Moreover, as similarly concluded in this study, it has been reported that decreasing the number of features in the model and the suggested feature optimisation contribute to higher interpretability of the model and give insights into the most crucial factor influencing air quality⁶³. However, ML-based air pollution research is associated with a gap, because of inappropriate handling and optimisation of the data²⁵. Weather conditions, different levels of pollutants (e.g., NO_2), and temporal variables (time of day, day of the week, and seasonal fluctuations) as important features might be substantially associated with PM concentrations^{64,65}. Weather factors, including T, RH, and WS, affect PM chemical reactions, production, and dispersion, respectively⁶⁶. Additionally, air pressure can reflect meteorological conditions that impact pollution levels⁶⁷. Thus, it appears that the combination of the mentioned factors has been instrumental in achieving a relatively high R^2 for PM.

At the same time, even after addressing overfitting and minimising the gap between training and testing data, the prediction models for gaseous chemicals, including CO, SO_2 , O_3 , and NO_2 still have R^2 values < 0.70 (Table 3); consequently, ML techniques still struggle to predict the majority of air pollution, as similarly reported by Mendez et al.⁶⁸ and Sharma et al.⁶⁹. Gaseous pollutants such as NO_2 , O_3 and SO_2 are influenced by intricate, nonlinear interactions among multiple components, complicating the correct prediction of their correlations by machine learning models. In contrast to PM, which exhibits more regular patterns, gaseous pollutants are acutely responsive to variations in atmospheric conditions, vehicular traffic, and industrial operations, resulting in swift oscillations in their concentrations. These abrupt alterations can lead to considerable fluctuations in the levels of air gaseous pollutants, thus complicating predictive accuracy relative to the more consistent trends observed in PM^{70,71}.

The air pollution dataset may include missing values owing to equipment malfunction or servicing, reducing the precision of models⁷². One possible explanation for the reduction in the R^2 value is that there was a higher amount of missing data in gases compared to PM in Tehran. In other words, the absence of data for PM was minimal, necessitating merely interpolation without the requirement for additional analysis. Furthermore, additional sensors for PM have been proposed as a potential factor contributing to the model's superior performance. Top-quality information needs to be gathered routinely and at multiple places to adequately reflect pollution shifts. In Tehran, a lack of previous data limits reliable model development since the gradual establishment of stations and their subsequent deployment over ten years may result in inconsistencies in the data. It has been reported that discontinuities in air pollution data, such as gaps in spatial and temporal data, add complexity to the forecast process⁷³. Moreso, fluctuations in weather, traffic, and manufacturing operations dramatically and promptly affect gaseous pollutants, thereby rendering real-time data processing difficult⁷⁴. Furthermore, shifts in political approaches and natural calamities additionally influence the air quality, which is often hard to forecast⁷⁵. Pattanayak and Kumar⁷⁶ highlighted how natural calamities can significantly affect environmental conditions, including air quality, and how these impacts are intertwined with political decisions and responses. Regarding this issue, it is noteworthy to highlight the shifts in the approach towards addressing air pollution in Tehran over the years under examination. Throughout the examined years, there have been multiple updates in the rules regarding vehicular traffic, factory operations, and the temporary shutdown of schools and offices during polluted days in Tehran. Model adaptability and dependability require a thorough real-world evaluation and continual updates⁷⁷ for application in various regions of Tehran because applying ML algorithms to current systems is difficult owing to compatibility and standards difficulties, and models need to adapt across areas and levels of pollution for widespread adoption. Atmospheric researchers may overlook some of these important issues when using ML in air pollution assessments⁷⁸.

Air quality forecasting

It is imperative to interpret pollutant models and interactions from the viewpoint of the environment to comprehend pollutant interactions and improve the modelling process. This is because numerous ML models, including SVM and multilayer perceptron (MLP), are plagued by the black box issue, which complicates the interpretation of the physical significance of the predictions. This lack of transparency can result in problems such as overfitting and local minima⁷⁹. In this context, the findings of this study demonstrated that O_3 has a significant prediction error in all metrics, though its lower NMSE signals superior data variance performance. SO_2 had the lowest MAE and greatest NMSE, reflecting minor mean errors except for unreliable model performance. CO, PM_{10} , and NO_2 exhibited moderate to significant errors, with varying NMSE values, highlighting various degrees of performance. Despite greater overall flaws, $PM_{2.5}$ provided the most successful prediction accuracy, as similarly obtained by Zhang et al.⁸⁰ and Chen et al.⁸¹, considering its lower NMSE than actual data variability.

To comprehend the interrelations between contaminants, from a chemical point of view, O_3 production occurs through photochemical interactions between NO_2 and volatile organic compounds (VOCs) under sunlight, and their nonlinearity and environmental sensitivity make prediction challenging⁸². As a secondary contaminant, O_3 is created by atmospheric processes, which must be accurately modelled⁸³. Temperature and direct sunlight affect O_3 generation and breakdown. It has been demonstrated that temperature-dependent changes in local chemistry and increased emissions of NO_2 in warmer conditions significantly contribute to higher O_3 levels⁸⁴. Moreover, O_3 forecasting is complicated by NO_2 catalytic cycles and interactions with other pollutants like CO and SO_2 ^{85,86}. As a result, O_3 regeneration and NO_2 reactivity in the atmosphere generate a dynamic system that has proved hard to describe^{87,88}. Different regions exhibit varying O_3 -temperature dynamics, influenced by local meteorological conditions and precursor emissions⁸⁹. Ren et al.¹⁸ and Eslami et al.⁹⁰ have demonstrated that nonlinear ML techniques, such as Random Forest and Extreme Gradient Boosting, have obtained superior prediction accuracy compared to linear models, particularly in the context of spatiotemporal modelling; however, precisely measuring peak O_3 levels continues to be a major obstacle. Lasso is intended for linear correlations and may have difficulty capturing intricate, non-linear interactions

commonly observed in air chemistry and pollutant dynamics. Moreover, the regularisation procedure may result in omitting potentially pertinent features, which could be especially vital for precisely predicting O_3 level due to its sensitivity to various environmental conditions. Alternatively, techniques such as Thresholded Lasso (TL), Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) may yield enhancements. These methodologies mitigate the biases and feature selection challenges associated with Lasso, rendering them more appropriate for intricate prediction tasks like O_3 -level modelling⁹¹. Consequently, traditional ML approaches have limitations in accuracy and interpretability for predicting these pollutants. The prediction of air pollution has preliminarily depended on physical and chemical models, thereby their efficiencies are influenced by precisely considering the intricate dynamics of air pollutant transport, consisting of the long-range transport and secondary formation of pollutants via atmospheric chemical reactions⁵⁶.

Air quality forecasting and evaluation systems are efficient tools for improving air quality and public health, reducing acute air pollution episodes, especially in urban regions, and decreasing the potential effects on climate, ecosystems and agriculture. Despite that, the air pollution predictive models need a higher optimisation, knowing the best-suited combining of data and algorithms for different dependent variables sounds difficult⁹². Despite the admirable efforts of researchers and administrators in Tehran, the air quality of urban areas is continuously declining, influencing the quality of air, water and land in this region. On a global scale, the issue of air pollution remains a significant concern, with detrimental effects that impact residents and the environment. According to the Lancet Commission on Pollution and Health, atmospheric pollution caused nine million premature mortalities in 2015, rendering it the leading hazardous driver of illness and early mortality worldwide⁹³. Hence, it is highly imperative to implement every potential measure to tackle air pollution.

Emerging geospatial intelligence technologies, along with big data analytics, machine learning and artificial intelligence, remarkably strengthen early warning systems for air pollution induced by climate change. Using such technologies, real-time data collection and analysis are facilitated, leading to prompt attempts at pollution incidents and thus creating more sustainability in urban media. These technologies help urban areas improve public health consequences and expand more efficient environmental policies. Moreover, integrating observations obtained from both ground sensors and satellite remote sensing instruments to air pollution is a growing necessity. Connecting low-cost sensors is significant for collecting data, although data quality obtained from these sensors is of great importance. There is still a crucial issue regarding different physical scales in air pollution modelling, in particular in cities influenced by long-range transport and localised air quality guidelines. Eventually, long-term, prospective, and interdisciplinary studies, along with international collaborations, are needed to tackle global air pollution.

Conclusions

The results of this study demonstrated how Lasso regularisation raised the accuracy and reliability of air pollution models by overcoming the challenge of overfitting. Regularisation decreased the model's complexity on account of the addition of the penalty term, thereby enhancing performance expressed as sparsity and improving generalisation performance. The very important stride in the prediction of a wide range of pollutant variants, notably reducing overfitting and selecting the most important features for the models, was accomplished by applying Lasso on a rich dataset from Tehran over a decade. The findings highlight the capability of Lasso regularisation as a promising technique in air quality prediction that could support governments in devising successful policies on air pollution management. The closeness of the training and test set performances of the model across various contaminants highlights its durability and dependability. Despite the difficulties of forecasting specific gaseous pollutants due to their complicated behaviours and interactions, Lasso regularisation has proved advantageous in increasing model interpretability and precision. Moreover, the application of Lasso allows the identification of the most important predictors out of a large number of variables, which helps in pinpointing aspects that bear the most influence on air quality. This feature selection power is very important because simplification of the models does not come with a sacrifice in predictive strength and hence makes the models more applicable in the real world. The current study opens new avenues for future research in areas like the combination of Lasso with other advanced regularisation methods and machine learning algorithms to improve model performance. In addition, the method could be applied to other ecological data and developed to forecast diverse ecological effects. Besides, the conclusions of this study underlined how necessary it is to continuously update and monitor any kind of prediction model when the environmental conditions and behaviours of the pollutants are in variable states. Air pollution is dynamic, driven by variables such as urbanisation, industrial activity, and climate fluctuations, necessitating the use of robust and adaptable modelling techniques. With an added ability to manage very large data sets and relevant feature selection, the Lasso regularisation is one convenient method for continuous air quality evaluation and control.

The ramifications of the study extend beyond scholarly activities, providing real advantages to environmental authorities and urban planners. Using the more powerful prediction capabilities of Lasso-regularised models, policymakers may adopt more focused and successful environmental protection actions, devote budgets systematically, and create urban settings that reduce exposure to hazardous pollutants. Such higher precision in these models may additionally collaborate with early warning systems for high pollution occurrences, therefore protecting public health. Overall, including Lasso regularisation in air quality prediction models is a significant step forward in environmental assessment, as well as providing a potential route for further investigation and operational uses in air pollution control.

Data availability

Due to ethical reasons data were not provided in the manuscript but will be available on request from corresponding author.

Received: 23 October 2024; Accepted: 23 December 2024

Published online: 02 January 2025

References

- Li, X., Hussain, S. A., Sobri, S. & Md Said, M. S. Overviewing the air quality models on air pollution in Sichuan Basin, China. *Chemosphere* **271**, 129502. <https://doi.org/10.1016/j.chemosphere.2020.129502> (2021).
- Mao, W., Wang, W., Jiao, L., Zhao, S. & Liu, A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustain. Cities Soc.* **65**, 102567. <https://doi.org/10.1016/j.scs.2020.102567> (2021).
- Mitreska Jovanovska, E. et al. Methods for urban air pollution measurement and forecasting: Challenges, opportunities, and solutions. *Atmosphere* **14**, 9. <https://doi.org/10.3390/atmos14091441> (2023).
- Guo, Q. et al. Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol Air Qual. Res.* **20**, 1429–1439. <https://doi.org/10.4209/aaqr.2020.03.0097> (2020).
- Chang, Y. S. et al. An LSTM-based aggregated model for air pollution forecasting. *Atmos Pollut Res.* **11**, 1451–1463. <https://doi.org/10.1016/j.apr.2020.05.015> (2020).
- Mehmood, K. et al. Predicting the quality of air with machine learning approaches: Current research priorities and future perspectives. *J. Clean. Prod.* **379**, 134656. <https://doi.org/10.1016/j.jclepro.2022.134656> (2022).
- Jain, S., Kaur, N., Verma, S., Kavita, Hosen, A. S. M. S. & Sehgal, S. S. Use of machine learning in air pollution research: A bibliographic perspective. *Electronics* **11**, 21. <https://doi.org/10.3390/electronics11213621> (2022).
- Bochenek, B. & Ustrnul, Z. Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere* **13**, 2. <https://doi.org/10.3390/atmos13020180> (2022).
- Rybarczyk, Y. & Zalakeviciute, R. Machine learning approaches for outdoor air quality modelling: A systematic review. *Appl. Sci.* **8**, 12. <https://doi.org/10.3390/app8122570> (2018).
- Rad, A. K. et al. Machine learning for determining interactions between air pollutants and environmental parameters in three cities of Iran. *Sustainability* **14**, 13. <https://doi.org/10.3390/su14138027> (2022).
- Castelli, M., Clemente, F. M., Popovič, A., Silva, S. & Vanneschi, L. A machine learning approach to predict air quality in California. *Complexity* **2020** (8049504). <https://doi.org/10.1155/2020/8049504> (2020).
- Liang, Y. C., Maimury, Y., Chen, A. H. L. & Juarez, J. R. C. Machine learning-based prediction of air quality. *Appl. Sci.* **10**, 9151. <https://doi.org/10.3390/app10249151> (2020).
- He, Z., Guo, Q., Wang, Z. & Li, X. Prediction of monthly PM_{2.5} concentration in Liaocheng, China employing artificial neural network. *Atmosphere* **13**, 1221. <https://doi.org/10.3390/atmos13081221> (2022).
- Guo, Q., He, Z. & Wang, Z. Predicting daily PM_{2.5} concentration employing wavelet artificial neural networks based on meteorological elements in Shanghai, China. *Toxics* **11**, 51. <https://doi.org/10.3390/toxics11010051> (2023).
- Guo, Q., He, Z. & Wang, Z. Prediction of hourly PM_{2.5} and PM₁₀ concentrations in Chongqing City, China based on artificial neural network. *Aerosol Air Qual. Res.* **23**, 220448. <https://doi.org/10.4209/aaqr.220448> (2023).
- 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.* **20**, 128–138. <https://doi.org/10.4209/aaqr.2019.08.0408> (2020).
- Wang, W., Liu, X., Bi, J. & Liu, Y. A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology. *Environ. Int.* **158**, 106917. <https://doi.org/10.1016/j.envint.2021.106917> (2022).
- Ren, X., Mi, Z. & Georgopoulos, P. G. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. *Environ. Int.* **142**, 105827. <https://doi.org/10.1016/j.envint.2020.105827> (2020).
- Wong, P. Y. et al. Using a land use regression model with machine learning to estimate ground level PM_{2.5}. *Environ. Pollut.* **277**, 116846. <https://doi.org/10.1016/j.envpol.2021.116846> (2021).
- Wang, A., Xu, J., Tu, R., Saleh, M. & Hatzopoulou, M. Potential of machine learning for prediction of traffic-related air pollution. *Transp. Res. Part. D Transp. Environ.* **88**, 102599. <https://doi.org/10.1016/j.trd.2020.102599> (2020).
- Bellinger, C., Mohamed Jabbar, M. S., Zaiane, O. & Osornio-Vargas, A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health.* **17**, 907. <https://doi.org/10.1186/s12889-017-4914-3> (2017).
- Essamlali, I., Nhaila, H. & El Khaili, M. Supervised machine learning approaches for predicting key pollutants and for the sustainable enhancement of urban air quality: A systematic review. *Sustainability* **16**, 3. <https://doi.org/10.3390/su16030976> (2024).
- Zhu, D., Cai, C., Yang, T. & Zhou, X. A machine learning approach for air quality prediction: Model regularization and optimization. *Big Data Cogn. Comput.* **2** <https://doi.org/10.3390/bdcc2040027> (2018).
- Berrolcal, V. J. et al. A comparison of statistical and machine learning methods for creating national daily maps of ambient PM_{2.5} concentration. *Atmos. Environ.* **222**, 117130. <https://doi.org/10.1016/j.atmosenv.2019.117130> (2020).
- Haq, M. A. & SMOTEDNN A novel model for air pollution forecasting and AQI classification. *Computers Mater. Continua.* **71**, 1. <https://doi.org/10.32604/cmc.2022.021968> (2022).
- Montesinos López, O. A., Montesinos López, A. & Crossa, J. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Springer International Publishing, 109–139. https://doi.org/10.1007/978-3-030-89010-0_4 (2022).
- Subramanian, J. & Simon, R. Overfitting in prediction models – is it a problem only in high dimensions? *Contemp. Clin. Trials.* **36**, 636–641. <https://doi.org/10.1016/j.cct.2013.06.011> (2013).
- Kernbach, J. M. & Staartjes, V. E. Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and overfitting. In *Machine Learning in Clinical Neuroscience* (Springer International Publishing, 15–21. https://doi.org/10.1007/978-3-030-85292-4_3 (2022).
- Zhao, P. & Yu, B. Stagewise Lasso. *J. Mach. Learn. Res.* **8**, 2701–2726 (2007). https://www.researchgate.net/publication/220319951_Stagewise_Lasso
- McNeish, D. M. Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivar. Behav. Res.* **50**, 471–484. <https://doi.org/10.1080/00273171.2015.1036965> (2015).
- Farzanegan, M. R., Gholipour, H. F. & Javadian, M. Air pollution and internal migration: Evidence from an Iranian household survey. *Empiric Econom.* **64**, 223–247. <https://doi.org/10.1007/s00181-022-02253-1> (2022).
- Faridi, S. et al. Long-term trends and health impact of PM_{2.5} and O₃ in Tehran, Iran, 2006–2015. *Environ. Int.* **114**, 37–49. <https://doi.org/10.1016/j.envint.2018.02.026> (2018).
- Rad, A. K., Shariati, M. & Zarei, M. The impact of COVID-19 on air pollution in Iran in the first and second waves with emphasis on the city of Tehran. *J. Air Pollut Health.* <https://doi.org/10.18502/japh.v5i3.5391> (2020).
- Daneshpour, K., Naghipour, A. & Rad, A. K. The relationships between air pollution, warming, and health in Tehran Metropolis, Iran, during 2015–2019. *J. Environ. Health Sustain. Dev.* <https://doi.org/10.18502/jehsd.v8i4.14436> (2023).
- Kotsiantis, S. B., Kanellopoulos, D. & Pintelas, P. E. Data preprocessing for supervised learning. *Int. J. Comput. Sci.* **1**, 111–117 (2006). https://www.researchgate.net/publication/228084519_Data_Preprocessing_for_Supervised_Learning
- Muthukrishnan, R. & Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In *IEEE International Conference on Advances in Computer Applications (ICACA)* (2016), 18–20. (2016). <https://doi.org/10.1109/ICACA.2016.7887916>

37. De Vito, E., Rosasco, L., Rudi, A. & Regularization From inverse problems to large-scale machine learning. In *Harmonic and Applied Analysis: From Radon Transforms to Machine Learning* (Springer International Publishing, 245–296. https://doi.org/10.1007/978-3-030-86664-8_5. (2021).
38. Dinov, I. D. Model performance assessment, validation, and improvement. In *Data Science and Predictive Analytics: Biomedical and Health Applications Using R* (Springer International Publishing, 477–531. https://doi.org/10.1007/978-3-031-17483-4_9. (2023).
39. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122. <https://doi.org/10.1177/1745691617693393> (2017).
40. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **1168**, 022022. <https://doi.org/10.1088/1742-6596/1168/2/022022> (2019).
41. Le, N. et al. K-Fold cross-validation: An effective hyperparameter tuning technique in machine learning on GNSS time series for movement forecast. In *Recent Research on Geotechnical Engineering, Remote Sensing, Geophysics and Earthquake Seismology* (Springer Nature Switzerland, 377–382. https://doi.org/10.1007/978-3-031-43218-7_88. (2024).
42. Fazekas, A. & Kovacs, G. Enumerating the k-fold configurations in multi-class classification problems. *arXiv preprint arXiv:2401.13843* (2024). <https://doi.org/10.48550/arXiv.2401.13843>
43. Ashfaq, J. & Iqbal, A. Introduction to support vector machines and kernel methods. (2019). https://www.researchgate.net/publication/332370436_Introduction_to_Support_Vector_Machines_and_Kernel_Methods
44. Kim, Y. & Kim, J. Gradient LASSO for feature selection. In *Proceedings of the Twenty-First International Conference on Machine Learning* Banff, Alberta, Canada, (2004). <https://doi.org/10.1145/1015330.1015364>
45. Doreswamy, K. S. H., Km, Y. & Gad, I. Forecasting air pollution particulate matter (PM_{2.5}) using machine learning regression models. *Proc. Comput. Sci.* **171**, 2057–2066. <https://doi.org/10.1016/j.procs.2020.04.221> (2020).
46. Guo, Q., He, Z. & Wang, Z. Monthly climate prediction using deep convolutional neural network and long short-term memory. *Sci. Rep.* **14**, 17748. <https://doi.org/10.1038/s41598-024-68906-6> (2024).
47. Yavuz, V. Variations in air pollutant concentrations on dry and wet days with varying precipitation intensity. *Atmosphere* <https://doi.org/10.3390/atmos15080896> (2024).
48. Rashed, S., Al-Taai, O. & Al-Salihi, A. Analysis study of meteorological parameters and their relationship with some concentration of tropospheric gases over Erbil City. *ZANCO J. Pure Appl. Sci.* **29**, 227–241. <https://doi.org/10.21271/zjpas.29.s4.27> (2017).
49. Zhang, Y., Sun, Q., Liu, J. & Petrosian, O. Long-term forecasting of air pollution particulate matter (PM_{2.5}) and analysis of influencing factors. *Sustainability* <https://doi.org/10.3390/su16010019> (2023).
50. Li, L., Rong, S., Wang, R. & Yu, S. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chem. Eng. J.* **405**, 126673. <https://doi.org/10.1016/j.cej.2020.126673> (2021).
51. Wang, Y., Yuan, Q., Li, T., Zhu, L. & Zhang, L. Estimating daily full-coverage near surface O₃, CO, and NO₂ concentrations at a high spatial resolution over China based on S5P-TROPOMI and GEOS-FP. *ISPRS J. Photogr. Remote Sens.* **175**, 311–325. <https://doi.org/10.1016/j.isprsjprs.2021.03.018> (2021).
52. Kang, Y. et al. Estimation of surface-level NO₂ and O₃ concentrations using TROPOMI data and machine learning over East Asia. *Environ. Pollut.* **288**, 117711. <https://doi.org/10.1016/j.envpol.2021.117711> (2021).
53. Shen, J., Valagolam, D. & McCalla, S. Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM_{2.5}, PM₁₀, O₃, NO₂, SO₂, CO) in Seoul, South Korea. *PeerJ* **8**, e9961. <https://doi.org/10.7717/peerj.9961> (2020).
54. Seng, D., Zhang, Q., Zhang, X., Chen, G. & Chen, X. Spatiotemporal prediction of air quality based on LSTM neural network. *Alex. Eng. J.* <https://doi.org/10.1016/j.aej.2020.12.009> (2021).
55. Wu, L., An, J. & Jin, D. Predictive model for O₃ in Shanghai based on the KZ filtering technique and LSTM. *Huanjing Kexue.* **45**, 5729–5739. <https://doi.org/10.13227/j.hjxx.202311150> (2024).
56. Koo, Y. S., Choi, Y. & Ho, C. H. Air quality forecasting using big data and machine learning algorithms. *Asia-Pacific J. Atmos. Sci.* **59**, 529–530. <https://doi.org/10.1007/s13143-023-00347-z> (2023).
57. Budach, L. et al. The effects of data quality on machine learning performance. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2207.14529> (2022). arXiv:2207.14529.
58. Gupta, S. & Gupta, A. Dealing with noise problem in machine learning datasets: A systematic review. *Proc. Comput. Sci.* **161**, 466–474. <https://doi.org/10.1016/j.procs.2019.11.146> (2019).
59. Trivedi, U., Bhatt, M. & Srivastava, P. Prevent overfitting problem in machine learning: A case focus on linear regression and logistic regression. In *Proceedings of the International Conference on Machine Learning* 345–349. (2021). https://doi.org/10.1007/978-3-030-66218-9_40
60. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodological)*. **58**, 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x> (1996).
61. Qiao, W. et al. The forecasting of PM_{2.5} using a hybrid model based on wavelet transform and an improved deep learning algorithm. *IEEE Access.* **7**, 142814–142825. <https://doi.org/10.1109/ACCESS.2019.2944755> (2019).
62. Cheng, Y., Zhang, H., Liu, Z., Chen, L. & Wang, P. Hybrid algorithm for short-term forecasting of PM_{2.5} in China. *Atmos. Environ.* **200**, 264–279. <https://doi.org/10.1016/j.atmosenv.2018.12.025> (2019).
63. Neo, E. X. et al. Yanto. Artificial intelligence-assisted air quality monitoring for smart city management. *PeerJ Comput. Sci.* **9**, e1306. <https://doi.org/10.7717/peerj-cs.1306> (2023).
64. Balogun, A. L., Tella, A., Baloo, L. & Adebisi, N. A. A review of the inter-correlation of climate change, air pollution and urban sustainability using novel machine learning algorithms and spatial information science. *Urban Clim.* **40**, 100989. <https://doi.org/10.1016/j.uclim.2021.100989> (2021).
65. Verma, P. et al. Assessment of human and meteorological influences on PM₁₀ concentrations: Insights from machine learning algorithms. *Atmos. Pollut. Res.* **15**, 102123. <https://doi.org/10.1016/j.apr.2024.102123> (2024).
66. Chen, Z. et al. Influence of meteorological conditions on PM_{2.5} concentrations across China: A review of methodology and mechanism. *Environ. Int.* **139**, 105558. <https://doi.org/10.1016/j.envint.2020.105558> (2020).
67. Ning, G. et al. Impact of low-pressure systems on winter heavy air pollution in the northwest Sichuan Basin, China. *Atmos. Chem. Phys.* **18**, 13601–13615. <https://doi.org/10.5194/acp-18-13601-2018> (2018).
68. Mendez, M., Merayo, M. G. & Nunez, M. Machine learning algorithms to forecast air quality: A survey. *Artif. Intell. Rev.* <https://doi.org/10.1007/s10462-023-10424-4> (2023).
69. Sharma, G., Khurana, S., Saina, N., Gupta, G. & Shivansh & Comparative analysis of machine learning techniques in air quality index (AQI) prediction in smart cities. *Int. J. Syst. Assur. Eng. Manag.* **15**, 3060–3075. <https://doi.org/10.1007/s13198-024-02315-w> (2024).
70. De Mattos Neto, P. et al. A methodology to increase the accuracy of particulate matter predictors based on time decomposition. *Sustainability* <https://doi.org/10.3390/su12187310> (2020).
71. Palanichamy, N., Haw, S., S. S., Murugan, R. & Govindasamy, K. Machine learning methods to predict particulate matter PM_{2.5}. *F1000Research* **11** <https://doi.org/10.12688/f1000research.73166.1> (2022).
72. Kim, T., Kim, J., Yang, W., Lee, H. & Choo, J. Missing value imputation of time-series air-quality data via deep neural networks. *Int. J. Environ. Res. Public Health.* **18**, 221213. <https://doi.org/10.3390/ijerph18221213> (2021).
73. Boaz, R. M., Lawson, A. B. & Pearce, J. L. Multivariate air pollution prediction modeling with partial missingness. *Environmetrics* **30**, e2592. <https://doi.org/10.1002/env.2592> (2019).

74. Nilesch, N., Narang, J., Parmar, A. & Chaudhari, S. IoT and ML-based AQI estimation using real-time traffic data. In *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)* 1–6. (2022). <https://doi.org/10.1109/WF-IoT54382.2022.10152160>
75. Smith, G. S., Anjum, E., Francis, C., Deanes, L. & Acey, C. Climate change, environmental disasters, and health inequities: The underlying role of structural inequalities. *Curr. Environ. Health Rep.* **9**, 80–89. <https://doi.org/10.1007/s40572-022-00336-w> (2022).
76. Pattanayak, A. & Kumar, K. S. K. Fiscal transfers, natural calamities and partisan politics: Evidence from India. *Econom Disasters Clim. Change.* **6**, 375–392. <https://doi.org/10.1007/s41885-022-00111-z> (2022).
77. Shaheen, K., Hanif, M. A., Hasan, O. & Shafique, M. Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks. *J. Intell. Robot Syst.* **105**, 9. <https://doi.org/10.1007/s10846-022-01603-6> (2022).
78. Tang, D., Zhan, Y. & Yang, F. A review of machine learning for modeling air quality: Overlooked but important issues. *Atmos. Res.* **300**, 107261. <https://doi.org/10.1016/j.atmosres.2024.107261> (2024).
79. Lu, W. Z. & Wang, D. Learning machines: Rationale and application in ground-level ozone prediction. *Appl. Soft Comput.* **24**, 135–141. <https://doi.org/10.1016/j.asoc.2014.07.008> (2014).
80. Zhang, Z., Zeng, Y. & Yan, K. A hybrid deep learning technology for PM_{2.5} air quality forecasting. *Environ. Sci. Pollut Res.* **28**, 39409–39422. <https://doi.org/10.1007/s11356-021-12657-8> (2021).
81. Chen, M. H., Chen, Y. C., Chou, T. Y. & Ning, F. S. PM_{2.5} concentration prediction model: A CNN-RF ensemble framework. *Int. J. Environ. Res. Public Health.* **20**, 54077. <https://doi.org/10.3390/ijerph20054077> (2023).
82. Li, X. et al. Examining the implications of photochemical indicators for O₃–NO_x–VOC sensitivity and control strategies: A case study in the Yangtze River Delta (YRD), China. *Atmos. Chem. Phys.* **22**, 14799–14811. <https://doi.org/10.5194/acp-22-14799-2022> (2022).
83. Galán, M. D. Implication of secondary atmospheric pollutants in the air quality: A case-study for ozone. In *environmental sustainability. Ch 2* <https://doi.org/10.5772/intechopen.95481> (2021). Syed Abdul Rehman, K. Ed., IntechOpen.
84. Romer, P. S. et al. Effects of temperature-dependent NO_x emissions on continental ozone production. *Atmos. Chem. Phys.* **18**, 2601–2614. <https://doi.org/10.5194/acp-18-2601-2018> (2018).
85. Yoo, J. M. et al. Spatiotemporal variations of air pollutants (O₃, NO₂, SO₂, CO, PM₁₀, and VOCs) with land-use types. *Atmos. Chem. Phys.* **15**, 10857–10885. <https://doi.org/10.5194/acp-15-10857-2015> (2015).
86. Al-Janabi, S., Alkaim, A., Al-Janabi, E., Aljeboree, A. & Mustafa, M. Intelligent forecaster of concentrations (PM_{2.5}, PM₁₀, NO₂, CO, O₃, SO₂) caused air pollution (IFCsAP). *Neural Comput. Appl.* **33**, 14199–14229. <https://doi.org/10.1007/s00521-021-06067-7> (2021).
87. Moonen, P. C., Cape, J. N., Storeton-West, R. L. & McColm, R. Measurement of the NO + O₃ reaction rate at atmospheric pressure using realistic mixing ratios. *J. Atmos. Chem.* **29**, 299–314. <https://doi.org/10.1023/A:1005936016311> (1998).
88. Fu, Y. et al. Association and interaction of O₃ and NO₂ with emergency room visits for respiratory diseases in Beijing, China: A time-series study. *BMC Public Health.* **22**, 2265. <https://doi.org/10.1186/s12889-022-14473-2> (2022).
89. Tzani, C. On the relationship between total ozone and temperature in the Troposphere and the lower stratosphere. *Int. J. Remote Sens.* **30**, 6075–6084. <https://doi.org/10.1080/01431160902798429> (2009).
90. Eslami, E., Salman, A. K., Choi, Y., Sayeed, A. & Lops, Y. A. A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. *Neural Comput. Appl.* **32**, 7563–7579. <https://doi.org/10.1007/s00521-019-04287-6> (2020).
91. Pokarowski, P., Rejchel, W., Soltys, A., Frej, M. & Mielniczuk, J. Improving Lasso for model selection and prediction. *Scand. J. Stat.* **49**, 831–863. <https://doi.org/10.1111/sjos.12546> (2019).
92. Sulaimon, I. et al. Air pollution prediction using machine learning: A review. In *EDMIC 2021 Conference Proceedings Environmental Design & Management International Conference Obafemi Awolowo University, Ile-Ife*. (2021). https://www.researchgate.net/publication/354150208_Air_Pollution_Prediction_using_Machine_Learning_-_A_Review
93. Fuller, R. et al. Pollution and health: A progress update. *Lancet Planet. Health.* **6**, e535–e547. [https://doi.org/10.1016/S2542-5196\(22\)00090-0](https://doi.org/10.1016/S2542-5196(22)00090-0) (2022).

Acknowledgements

The authors express their gratitude to Shiraz University and Urmia University for providing research facilities.

Author contributions

Abbas Pak: Writing — original draft, Conceptualization, Formal analysis, Methodology. Abdullah Kaviani Rad: Writing — original draft, Resources, Data curation. Mohammad Javad Nematollahi: Writing — review and editing, Validation, Supervision, Project administration. Mohammadreza Mahmoudi: Validation, Supervision, Writing — review and editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.J.N. or M.M.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024