International Journal of Environmental Science and Technology (2025) 22:459–476
https://doi.org/10.1007/s13762-024-05671-z

ORIGINAL PAPER

A novel ensemble machine learning method for accurate air quality prediction

M. Emeç[1] · **M. Yurtsever**[2]

Received: 13 June 2023 / Revised: 26 December 2023 / Accepted: 21 April 2024 / Published online: 6 May 2024

Abstract

Air pollution continues to be an important problem that causes health issues worldwide. Factors such as industrial development, increased vehicle traffic, and energy production have a negative impact on air quality by releasing harmful gases and particles into the atmosphere. Consequently, this can lead to respiratory diseases, cardiovascular problems, and other health complications. Predicting air quality is a crucial step in safeguarding human health and informing environmental policies. Many cities employ measurement instruments and data collection systems to monitor and forecast air quality. This data can be analyzed using machine learning models to predict future air pollution levels. This article examines the performance of a new stacking ensemble model for estimating $PM_{2.5}$, based on air quality datasets from major cities such as Beijing and Istanbul. The model combines predictions from various machine learning models. In the initial stage of the study, the performance of commonly used models in the literature, such as multi-layer perceptron, support vector regression, and random forest, were evaluated. These models were assessed for their ability to predict $PM_{2.5}$ using metrics such as mean absolute error (MAE), root mean squared error (RMSE) and R-squared (R^2). This evaluation determines the proximity of the model predictions to the actual data. The stacking ensemble model examined in this study yielded the best results for $PM_{2.5}$ predictions, with MAE of 6.67, RMSE of 8.80 and R^2 of 0.91. In conclusion, the stacking ensemble model for air pollution prediction offers a promising approach for achieving superior results compared to traditional machine learning models.

Keywords Air quality · Ensemble model · Machine learning · Regression

Introduction

Air quality is a significant concern that has a direct impact on people's health and well-being worldwide. It involves assessing the purity, clarity, and healthiness of the air in a given area. The quality of air depends on the composition, density, and distribution of gases, dust, and other particles present in the atmosphere. Many of these components are influenced by human activities. Air pollution arises from harmful substances released during activities such as energy production, transportation, industrial operations, and agriculture (Zhang et al. 2017).

$PM_{2.5}$ refers to the measurement of small particulate matter pollution in the air. Even at very low levels, $PM_{2.5}$ has been found to have severe effects on human health (Pui et al. 2014; Feng et al. 2016). It can lead to respiratory problems, cardiovascular diseases, certain types of cancer, and other severe health issues (Wang et al. 2020). The World Health Organization (WHO) reports that both indoor and outdoor air pollution is linked to approximately 7 million premature deaths annually (WHO 2022). Monitoring the $PM_{2.5}$ levels is crucial for authorities to implement preventive measures and control pollution. As a result, governments are increasingly prioritizing the estimation of $PM_{2.5}$ concentrations.

Machine learning, a sub-branch of artificial intelligence, allows computer systems to enhance their performance by leveraging sample data (Janiesch et al. 2021). Machine learning algorithms construct models based on extensive data, enabling them to process new data and make predictions. Due to its versatility, machine learning finds applications

Editorial responsibility: Mohamed F. Yassin.

✉ M. Emeç
murat.emec@istanbul.edu.tr

[1] IT Department, Istanbul University, 34116 Fatih, Istanbul, Turkey

[2] Faculty of economics and administrative sciences, Department of management information systems, Izmir Democracy University, 35140 Karabaglar, Izmir, Turkey

across various sectors, making it widely utilized today. For instance, in the financial sector, machine learning algorithms are employed to forecast financial risks, while in healthcare, they aid in diagnosis and treatment planning. In the realm of air quality forecasting, machine learning techniques have proven successful. These techniques involve the creation of statistical and mathematical models that utilize various data analysis methods to predict future values of air quality parameters. In recent years, the utilization of deep learning techniques for air quality estimation has witnessed a significant increase.

Ensemble learning is a technique that involves combining multiple machine learning algorithms to achieve improved results. This approach combines several models that employ different learning methods or utilize distinct data properties (Dong et al. 2020). By leveraging ensemble learning, prediction accuracy and decision-making can be enhanced through the aggregation of predictions from multiple models. Various ensemble learning methods exist, including Bagging, Boosting, and Stacking. Bagging, or Bootstrap Aggregating, involves generating multiple models trained on different subsets of the training data (bootstrap samples) and then aggregating their predictions to produce the final result. This technique aims to reduce variance and improve generalization by incorporating diverse models.

Boosting, on the other hand, focuses on combining weak learners, such as simple decision trees, to create a strong learner. In Boosting, models are trained iteratively, with each subsequent model emphasizing the misclassified instances from previous models. The final prediction is obtained by combining the predictions of all the weak learners. Stacking takes a different approach by creating a meta-model that combines predictions from various types of models. Instead of relying on a single model, Stacking considers the outputs of different models as additional features and trains a meta-model on this augmented dataset. The meta-model then generates the final prediction by utilizing the combined knowledge of the individual models (Wang and Yue 2019). Ensemble learning techniques like Bagging, Boosting, and Stacking provide valuable strategies for improving the performance and robustness of machine learning models by leveraging the strengths of multiple models.

When estimating $PM_{2.5}$ levels, ensemble learning can improve prediction accuracy by combining data from various sensors, meteorological sources, and air quality measurements. This approach has the potential to achieve superior results compared to individual models. The use of ensemble learning in predicting $PM_{2.5}$ levels offers a more reliable and effective approach for air quality management and informing policy decisions.

In current studies, various methods with distinct advantages, such as classical machine learning and deep learning, have been utilized (Harishkumar et al. 2020; Liang et al. 2020; Juarez and Petersen 2022). However, a crucial concern lies in how to combine these diverse methods to enhance the accuracy of air quality forecasting. Stacking Ensemble, an ensemble learning method, aims to achieve more robust and accurate predictions by integrating predictions from different machine learning algorithms. This approach allows for the collaboration of different algorithms to improve the accuracy of air quality forecasts.

For instance, by combining predictions from a Support Vector Machine (SVM) model and a Random Forest (RF) model, a more reliable air quality forecast can be obtained. The Stacking Ensemble method harnesses the complementary strengths of diverse algorithms, enabling a more comprehensive prediction. Thus, in the field of air quality forecasting, ensemble methods, particularly Stacking Ensemble, have the potential to enhance forecast accuracy and reliability.

This paper presents an important innovation in air pollution forecasting by combining different machine learning models with the 'Stack Model'. It highlights the effectiveness of non-traditional and non-linear models in air pollution forecasting by combining them with a linear fusion method such as linear regression. This study's main innovation is combining basic nonlinear models such as MLP, SVR, and RF with a linear regression model in the meta-stage. Unlike traditional ensemble methods, this aims to achieve higher precision and reliability in air quality forecasts by combining models with different structures to improve their accuracy. Hence, this method provides a significant innovation in air pollution forecasting by offering a level of accuracy that has not been achieved before by harmoniously combining non-linear models.

The main contributions of this study are as follows:

1. Proposal of a new ensemble model that integrates various machine learning models to enhance the accuracy of air pollution forecasting.
2. Providing a valuable resource for researchers and decision-makers seeking to improve the precision and reliability of air pollution forecasts for effective management and control.
3. Evaluation of the effectiveness of stacking ensemble methods in the context of time series forecasting.

The remainder of the paper is structured as follows. The related works section presents a critical literature review that

discusses previous and related work in the field. Material and methods section, the methodologies used in the ensemble learning method are explained in detail. It also provides a comprehensive description of the data used in this study. The proposed stacking ensemble model section includes the model's architecture and implementation steps. The results and discussion section covers the experimental work with a description of the experimental settings and a discussion of the results obtained. Finally, the Conclusion section concludes the article, highlighting important findings and suggesting avenues for future research.

Related work

Air quality forecasting is a crucial aspect of environmental health and air pollution control. Research conducted in this field has shown the effectiveness of various machine learning and deep learning techniques in predicting air quality. This section provides a comprehensive review of significant studies conducted in the domain of air quality forecasting.

In their study, Castelli et al. (2020) utilized support vector regression (SVR), a widely adopted machine learning technique, to forecast levels of pollutant particles and predict the air quality index (AQI). The authors found that SVR with a radial basis function (RBF) kernel yielded the most precise outcomes, successfully predicting hourly pollutant levels and AQI in California.

The estimation of the air quality index (AQI) in India has relied on the National Air Monitoring Program (NAMP), which was established by the Central and State Pollution Control Boards. In a study by Janarthanan et al. (2021), a deep learning model combining SVR and long short-term memory (LSTM) was employed to classify AQI values. The proposed model demonstrated enhanced accuracy in estimating the AQI for a specific location within the city.

In their study, Liu et al. (2019) developed regression models for the air quality index (AQI) in Beijing and the concentration of nitrogen oxide (NOX) in an Italian city. Using the SVR and random forest regression (RFR) methods, the authors demonstrated that the SVR-based model outperformed in estimating the AQI, while the RFR-based model exhibited superior performance in estimating the NOX concentration. The study emphasized the effectiveness and suitability of combining machine learning techniques with air quality prediction to address environmental issues.

In a study conducted by Fang et al. (2020), dimension reduction techniques were employed using air quality data and meteorological data in Tianjin. Missing data were filled using random forest interpolation, and feature selection was performed using locally linear embedding (LLE). A prediction model for the air quality index was then created using the LSTM. Experimental results demonstrated that the proposed method yielded positive effects in terms of dimension reduction and estimation accuracy when compared to the principal component analysis (PCA) and back propagation (BP) methods.

In a study focused on estimating $PM_{2.5}$ levels in Delhi, Sarkar et al. (2022) evaluated various data prediction methods. Among the models examined, the proposed hybrid model, LSTM-GRU, demonstrated superior performance, with an MAE value of 36.11 and an R^2 value of 0.84. Another study employed three different deep learning neural networks (LSTM, CNN, and CNN-LSTM) as well as a back propagation neural network (BPNN) for spatial–temporal estimation of $PM_{2.5}$ concentrations. The input data included past $PM_{2.5}$ concentrations, air quality data, and the most recent $PM_{2.5}$ concentrations. Yang et al. (2021) found that deep learning models outperformed in hourly $PM_{2.5}$ predictions, with the CNN-LSTM model yielding the best results when combined with LSTM. The study also highlighted that optimizing the input variables and model structure can enhance prediction performance.

Several machine learning models, such as multiple linear regression (MLR), decision tree (DT), K-nearest neighbors (KNN), random forest (RF), and XGBoost, were used in a study to forecast $PM_{2.5}$ levels for the city of Hyderabad. Long short-term memory (LSTM). A deep learning model, was also applied. The investigation showed that, with a 7.01 μg/m^3 mean absolute Error (MAE) and 0.82 R^2 value, the XGBoost regression model performed better than the other machine learning models. But when it came to $PM_{2.5}$ modeling, the LSTM deep learning model outperformed the XGBoost regression, obtaining a 0.89 R^2 value and 5.78 μg/m3 MAE (Gokul et al. 2023).

Five machine learning models were used in a study that analyzed and predicted air quality using six years' worth of air pollution data from 23 Indian cities. The XGBoost model outperformed the others and produced the maximum linearity between the predicted and real data (Kumar and Pande 2023). Three different machine learning models—SARIMA, DVM with different kernel functions, and LSTM with different hyperparameter settings—have been analyzed in order to develop an effective prediction model for the city of Ahmedabad. The findings show that DVM with the RBF kernel performed better than other models and various modifications of the DVM model in terms of multiple

evaluation criteria like the R^2 score and RMSE (Maltare and Vahora 2023).

In a study by Sethi and Mittal (2019), a feature selection method called Causality Based Linear was proposed to identify the most influential parameters on the air quality index. Experimental investigations conducted on Delhi's air quality dataset demonstrated that the proposed method successfully detected key parameters such as wind speed, carbon monoxide, and nitrogen dioxide, and yielded higher accuracy compared to other methods. Karakuş and Yıldız (2019) conducted a study revealing a significant relationship between temperature, wind speed, humidity, and the air quality index (AQI). Furthermore, there are studies that utilize sky images as input for air quality index prediction. Chen (2020) proposed a method called AQIF-DDL (Air Quality Index Forecasting based on Deep Vocabulary Learning), which combines machine vision and deep learning. The sky image is used as input to estimate the AQI value. Experimental results demonstrated that the proposed method performed well in predicting the AQI.

With the growing popularity of ensemble methods, researchers in the field of air quality prediction have started incorporating these methods into their studies. Lin et al. (2021) conducted a study where various prediction models were designed using the Gated Recurrent Unit (GRU) deep learning network, and a multiple linear regression-based ensemble learning model was proposed to integrate these models. The results showed that the MLEGRU (Multiple Linear Regression based GRU) model outperformed other ensemble methods in terms of prediction accuracy. In another study by Chang et al. (2020), a hybrid model and framework were proposed that combined ensemble learning techniques with traditional methods such as gradient boosted tree regression (GBTR), support vector machine-based regression (SVR), and LSTM. Experimental results demonstrated that the hybrid model outperformed existing models in air pollution prediction.

Furthermore, Ma et al. (2023) proposed an ensemble learning-based solution to enhance the accuracy of air pollutant predictions from the China Unified Atmospheric Chemistry Environment (CUACE) model. They utilized the random forest algorithm, XGBoost algorithm, and GBDT algorithm to improve the prediction results for $PM_{2.5}$, PM_{10}, and O_3. The model was further optimized using the grid search method. The results showed that the $PM_{2.5}$ and PM_{10} prediction accuracy increased by 60%, while the O_3 prediction accuracy increased by 70%.

The Beijing Environmental Protection Monitoring Center (BEPMC) provided an 8-year dataset for the study, which was used to create a number of prediction models. Models were created using machine learning techniques such as probabilistic voting ensemble, random forest (RF), Support vector regressor (SVR), and simple linear regression (SLR). In the study, the ensemble voting technique from several learning algorithms was applied on heterogeneous base learners. According to the experimental data, RF performed better for AQI prediction in terms of MAE and RMSE scores, whereas the maximum probabilistic voting ensemble performed better in terms of R^2 (Xiang et al. 2023).

In a separate study, a novel CNN-RF ensemble framework was presented for modeling $PM_{2.5}$ concentrations, aiming to combine the benefits of Random Forest (RF) regression capability and Convolutional Neural Network (CNN) feature extraction. The results showed that when compared to separate CNN and RF models, the suggested CNN-RF model had superior modeling capacity (Chen et al. 2023).

In another study presents a novel decomposition method that combines the benefits of two well-liked decomposition algorithms. It is based on a hybrid data preprocessing-analysis strategy and shows promise as a decomposition strategy. Additionally, a new hybrid data preprocessing-analysis technique is put forth, which is based on the successful decomposition of a promising new strategy that effectively captures the long-short patterns of complex time series. Furthermore, the study forecasts each of the decomposed components independently using long short-term memory (LSTM), a potent deep learning prediction technique. Ultimately, a novel AdaBoost-LSTM ensemble method is created to incorporate individual prediction outcomes into the ultimate forecasts, resulting in a notable enhancement in prediction efficacy. The designed hybrid model performs much better than compared models, according to experimental results (Li et al. 2023).

Overall, these studies highlight the successful application of machine learning and deep learning techniques in air quality prediction. Moreover, the utilization of ensemble methods in the field of air quality forecasting offers opportunities to enhance prediction accuracy.

Materials and methods

In this section, we will provide an overview of the datasets used in our study and the machine learning methods applied in our proposed batch ensemble learning model.

Data source

Air quality data from Beijing and Istanbul were used in this study. The dataset includes daily $PM_{2.5}$ values collected from various air quality stations in both cities. It also includes other important air parameters such as SO_2, NO_2, NOX, NO, O_3, and more. The dataset has a sufficient sample size to provide representation across different seasons, time zones, and meteorological conditions.

The air monitoring stations continuously measure $PM_{2.5}$ values along with other environmental factors that impact air quality. The air quality data used in this study were collected from air monitoring stations in Beijing (Air Quality Index Project 2022) and Istanbul (SIM 2023). These collected data represent the air quality and environmental conditions over a specific time period. For Beijing, the data range is from January 1, 2014, to April 20, 2023 (3,397 days), and for Istanbul, it is from March 1, 2013, to December 31, 2022 (3,591 days). "Daily average air quality (real-time) data for all stations (24 stations)" is presented at the address (Air Quality Index Project 2022), where we obtained the Beijing data. Also, "Daily average air quality (real-time) data for all stations (29 stations)" is presented at the address (SIM 2023) where we obtained the Istanbul data.

In major cities like Beijing and Istanbul, air quality data is collected by various air monitoring stations. For this study, a dataset was created using previously collected $PM_{2.5}$ data for training and testing purposes. Additionally, a feature vector was constructed, which includes other environmental factors (SO_2, NO_2, NOX, NO, O_3) as variables. Only Istanbul and Beijing were chosen for air quality estimation in the study due to the following reasons:

Data availability

1. Access to air quality data is crucial for the study. Istanbul and Beijing are large and densely populated cities where air quality data is more readily available compared to smaller cities. Thus, it was easier to obtain sufficient air quality data for these cities.
2. Sample Size: The dataset used for air quality estimation requires an adequate sample size to ensure reliable predictions. Major cities like Istanbul and Beijing generally have larger datasets due to the abundance of air quality monitoring stations. Increasing the sample size improves the precision of the machine learning models employed in the study.
3. Geographical Differences: Istanbul and Beijing have distinct geographical characteristics and unique factors that influence air quality. By selecting these cities, the study aimed to compare air quality forecasts in different geographical regions. This allows for the evaluation of air quality predictions in diverse locations and facilitates the comparison of results.
4. Application Orientation: The study focuses on examining air quality forecasts in a specific geographical area. Istanbul and Beijing are significant urban centers facing air pollution challenges. Estimating air quality in these cities is crucial for the development of effective environmental policies and public health measures.

While the study focused on Istanbul and Beijing, the proposed method can be adapted and extended to other cities or regions as well, providing a flexible framework for air quality estimation.

Data preprocessing

In the data preprocessing stage, several steps were performed to ensure accurate and reliable predictions. These steps include cleaning missing data points and abnormal values, as well as normalizing the data. The specific operations carried out on the datasets are as follows:

- Missing data handling: Missing data points were identified and filled using innovative methods. Approximately 8% of the data in both datasets were found to be missing. Rather than relying on simplistic imputation methods such as the overall mean, median, or mode, more suitable techniques were utilized to avoid potential inaccuracies in the results (Yurtsever & Emeç, 2023). In the case of estimating $PM_{2.5}$ for the next day, missing data were filled by averaging the values from the previous day and the following day.
- Outlier detection and correction: Outliers are data points that significantly deviate from the normal pattern and

Table 1 Beijing air quality dataset statistical information

Feature(s)	Count	Mean	std	Min	25%	50%	75%	Max
NOX	3397	65	51	2	36	56	78	886
O_3	3397	44	32	1	22	34	60	175
NO_2	3397	20	11	1	11	18	25	90
SO_2	3397	4	7	1	1	2	4	71
NO	3397	8	7	1	4	6	9	78
$PM_{2.5}$	3397	115	65	10	66	108	152	525

can distort predictions. However, since the estimation of $PM_{2.5}$ relied on the previous day's values, outlier elimination was not performed in this study.
- Data normalization: To ensure consistency and comparability across variables with different scales, a data normalization process was applied. This process brings the data into a standardized range, typically between 0 and 1, by scaling the values proportionally.

By conducting these preprocessing operations, the datasets were prepared for further analysis and modeling, ensuring the accuracy and reliability of the predictions.

Statistical analysis of datasets

Statistical information of the attributes in the datasets is important during model design training. The statistical information for air quality in Beijing and Istanbul is presented in the tables below.

According to Table 1, there is statistical information available for six different characteristics in the Beijing air quality dataset. Here are the descriptions of these features and the interpretations of the statistics in the table:

- NOX: Data is available for nitrous oxide (NOX) levels. There are a total of 3,397 samples. The average NOX value is 65 units. The standard deviation is 51 units, indicating a wide distribution of the data. The minimum NOX value is 2 units, while the maximum value is 886 units.
- O_3: Data on ozone (O_3) levels are provided. The average value for the 3,397 samples in the dataset is 44 units. The standard deviation is 32 units, suggesting a significant variation in O_3 levels. The minimum value is 1 unit, and the maximum value is 175 units.
- NO_2: Data is available for nitrogen dioxide (NO_2) levels. There are 3,397 examples in total. The average NO_2 value is 20 units. The standard deviation is 11 units, indicating a moderate variation in NO_2 levels. The minimum NO_2 value is 1 unit, while the maximum value is 90 units.
- SO_2: Data on sulfur dioxide (SO_2) levels are included. There are 3,397 samples in total. The average SO_2 value is 4 units. The standard deviation is 7 units, suggesting a wide distribution of SO_2 levels. The minimum SO_2 value is 1 unit, while the maximum value is 71 units.
- NO: Data is available for nitrogen monoxide (NO) levels. There are 3,397 examples in total. The average NO value is 8 units. The standard deviation is 7 units, indicating a moderate variation in NO levels. The minimum NO value is 1 unit, while the maximum value is 78 units.
- $PM_{2.5}$: Data is available for $PM_{2.5}$ particulate matter levels. There are 3,397 samples in the dataset. The average $PM_{2.5}$ value is 115 units. The standard deviation is 65 units, indicating a wide distribution of $PM_{2.5}$ levels. The minimum value is 10 units, while the maximum value is 525 units.

The provided statistics offer insights into the data distribution and central tendency for each feature. For instance, $PM_{2.5}$ values have a higher mean and wider distribution, whereas SO_2 values have a lower mean and a narrower distribution. These statistics provide an overview of the data, which can be utilized for air quality analysis and decision-making.

The information presented in Table 2 comprises crucial statistics concerning the air quality dataset in Istanbul, offering valuable insights. Subsequently, the interpretation of these statistics for each feature is provided:

NOX: There are 3,591 samples available for NOX (nitrous oxide) levels. The average NOX value is 109 units, and the standard deviation is 79 units, indicating a wide variation in NOX levels. The minimum value is 1 unit, while the maximum value is 648 units.
O_3: Data is provided for O_3 (ozone) levels, with a total of 3,591 samples. The average O_3 value is 48 units, and

Table 2 Istanbul air quality data set statistical information

Feature(s)	Count	Mean	std	Min	25%	50%	75%	Max
NOX	3.591	109	79	1	49	85	171	648
O_3	3.591	48	26	1	29	46	64	196
NO_2	3.591	40	22	0,35	23	37	53	199
SO_2	3.591	6	6	0,7	3	4	8	69
NO	3.591	38	38	0,89	16	29	41	572
$PM_{2.5}$	3.591	27	17	3	15	21	32	161

the standard deviation is 26 units, suggesting a moderate variation in O_3 levels. The minimum value is 1 unit, while the maximum value is 196 units.

NO_2: Data is available for NO_2 (nitrogen dioxide) levels, with 3,591 samples in the dataset. The average NO_2 value is 40 units, and the standard deviation is 22 units, indicating a moderate variation in NO_2 levels. The minimum value is 0.35 units, while the maximum value is 199 units.

SO_2: Data is provided for SO_2 (sulfur dioxide) levels, with a total of 3,591 samples. The average SO_2 value is 6 units, and the standard deviation is 6 units, suggesting a moderate variation in SO_2 levels.

NO: Data is available for NO (nitrogen monoxide) levels, with 3,591 samples in the dataset. The average NO value is 38 units, and the standard deviation is 38 units, indicating a wide variation in NO levels. The minimum value is 0.89 units, while the maximum value is 572 units.

$PM_{2.5}$: Data is provided for $PM_{2.5}$ (particulate matter) levels, with a total of 3,591 samples. The average $PM_{2.5}$ value is 27 units, and the standard deviation is 17 units, suggesting a moderate variation in $PM_{2.5}$ levels. The minimum value is 3 units, while the maximum value is 161 units.

The statistical table for the Istanbul air quality dataset provides insights into the average values, standard deviations, and distributions of different air pollutants. For instance, NOX levels exhibit high variability, while $PM_{2.5}$ levels show less variability. These statistics offer a general understanding of the air quality situation in Istanbul and can guide the implementation of measures to control air pollution.

Comparing the statistics in Tables 1 and 2, we can observe the following differences in air quality between Istanbul and Beijing:

1. NOX (nitrogen oxide) levels: The average NOX level in the Istanbul dataset (109 units) is higher than that in the Beijing dataset (65 units). This indicates that Istanbul has higher NOX levels compared to Beijing.
2. O_3 (ozone) levels: The average O_3 level in the Istanbul dataset (48 units) is slightly higher compared to the Beijing dataset (44 units). O_3 levels between the two cities are similar.
3. NO_2 (nitrogen dioxide) levels: The average NO_2 level in the Istanbul dataset (40 units) is higher than that in the Beijing dataset (20 units). Istanbul's NO_2 levels are higher than Beijing's.

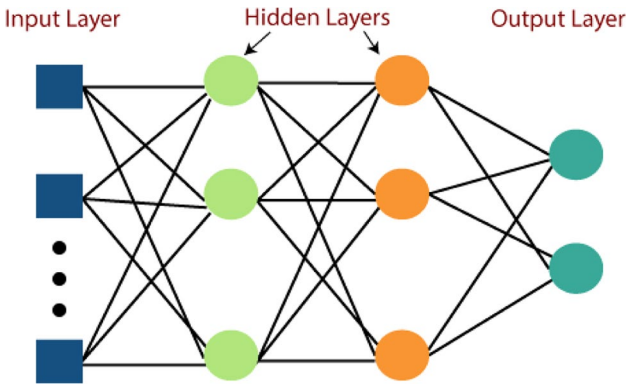

Fig. 1 Multi-layer perceptron architecture

4. SO_2 (sulfur dioxide) levels: The average SO_2 level in the Istanbul dataset (6 units) is slightly higher compared to the Beijing dataset (4 units). SO_2 levels between the two cities are similar.
5. NO (nitrogen monoxide) levels: The average NO level in the Istanbul dataset (38 units) is significantly higher than that in the Beijing dataset (8 units). Istanbul's NO levels are higher than Beijing's.
6. $PM_{2.5}$ (particulate matter) levels: The average $PM_{2.5}$ level in the Beijing dataset (115 units) is higher than that in the Istanbul dataset (27 units). Beijing's $PM_{2.5}$ levels are higher than Istanbul's.

This comparison reveals that the air quality profiles of Istanbul and Beijing differ. Istanbul tends to have higher levels of NOX, NO_2, and NO, while $PM_{2.5}$ levels are lower compared to Beijing. SO_2 and O_3 levels are similar in both cities. These differences indicate that the factors influencing air quality vary between cities, and thus, different strategies should be implemented for air pollution control accordingly.

Multi-layer perceptron (MLP)

The multi-layer perceptron (MLP) is a neural network architecture consisting of an input layer, one or more hidden layers, and an output layer. Illustrated in Fig. 1, each layer comprises nodes or neurons that apply weights to the inputs and pass them through an activation function (Gardner and Dorling 1998). The MLP primarily learns through the backpropagation algorithm and can be further optimized by adjusting hyperparameters, including the number of hidden

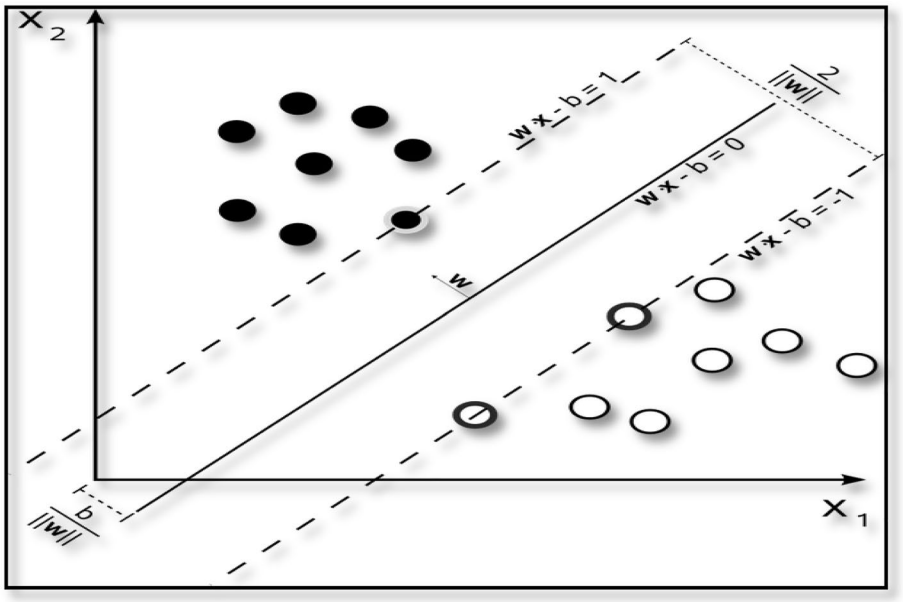

Fig. 2 Support vector regression architecture

layers, number of neurons, choice of activation functions, and more.

In the MLP model, the output of each hidden layer is obtained by using the output of the previous layer as input. The final output layer applies an activation function to produce the prediction result. For regression problems, linear activation functions are commonly used, while for classification problems, appropriate activation functions such as sigmoid or softmax are utilized.

To train the MLP model, a learning process is employed where the weight and bias values are optimized based on the training dataset. The backpropagation algorithm is a widely used method for training MLP models. The MLP model can be customized for specific applications by specifying parameters such as the input and output layer sizes, the number and sizes of hidden layers, the activation functions, and the training algorithm.

Support vector regression (SVR)

Support vector regression (SVR) is a regression-focused machine learning model derived from the support vector machines (SVM) algorithm, which is predominantly utilized for classification purposes. SVR aims to predict a regression function by constructing a hyperplane around the data points, as depicted in Fig. 2. During the learning process, SVR ensures that the data points stay within a certain constraint. The model can be optimized by adjusting various hyperparameters, such as the choice of kernel function and the tolerance parameter for low marginal error.

The goal of SVR is to provide an estimate that closely approximates the true output value. It formulates the problem as an optimization task, seeking to minimize the error on

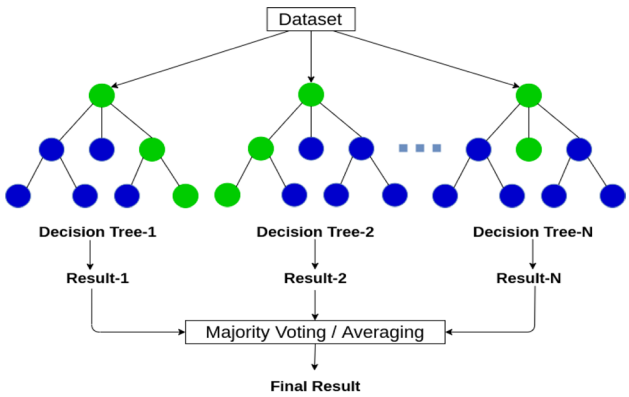

Fig. 3 Random forest architecture

the training dataset. To achieve this, a specialized optimization algorithm is used to determine the weight vector (w) and the bias term (b). In SVR, it is possible to model non-linear relationships by employing the "kernel trick". The data are transformed into a higher-dimensional feature space using this method so that linear separation is possible. As opposed to being linear in the original feature space, it enables SVR to capture complex patterns and connections in the data.

The general formula of SVR can be expressed as follows (Chen et al. 2015):

$$h(x) = w^T * x + b.$$

In this formula,

- h(x) represents the estimated output value.

- w represents the weight vector of the support vector regression.
- x represents the input feature vector.
- b represents a constant value called the bias or correction term.
- T represents vector transpose.

SVR can be customized with adjustable parameters (such as C and ε). These parameters are employed to manage the adaptability of the model and mitigate the risk of overfitting. The performance and predictive power of the model depend on the correct setting of these parameters. SVR generally performs well in regression problems compared to other models and can also deliver good results, especially when noise or outliers are present in the dataset.

Random forest (RF)

The Random Forest (RF) algorithm, which was initially introduced by Breiman (2001), is an ensemble approach that leverages the combination of multiple decision trees to construct a more powerful model. As shown in Fig. 3, each decision tree is trained on a random subsample and makes individual predictions. RF is particularly known for its ability to capture interactions and feature importance rankings. The final prediction is generated by RF through the amalgamation of predictions from multiple trees. This approach helps to reduce noise, prevent overfitting, and generate more generalized predictions. To achieve further optimization, RF can be fine-tuned by adjusting hyperparameters, including the number of trees, maximum depth, and minimum partition size.

The general formula for random forests can be expressed as (Ao et al. 2019):

$$h(x) = \Sigma(T(i)(x)).$$

In this formula,

- h(x) represents the estimated output value.
- T(i)(x), i. represents the predictive value obtained from the tree.
- Σ is the sum symbol and represents the sum of the forecast values from all trees.

RF is an ensemble model created by combining multiple decision trees. To train each tree, varying subsets of features are utilized, and the bootstrap sampling method is employed. Consequently, each tree makes independent predictions using distinct features and data points. The RF model obtains an ensemble estimate by averaging the predictions from all trees or by applying the consensus principle. This approach mitigates the risk of overfitting, as RF is formed by combining decision trees. Moreover, RF exhibits robust performance even in high-dimensional and noisy datasets. A significant advantage of RF is its ability to evaluate feature importance. The feature importance rank quantifies the information gain of the RF model based on the features utilized in each tree. In classification problems, RF is typically trained with measures such as "gini impurity" or "entropy," while regression problems employ methods that minimize the variance of the target variable (e.g., mean square error). As an ensemble model, RF produces more robust and accurate predictions by combining multiple decision trees. As a consequence, RF has emerged as a highly popular machine learning model with a wide array of applications in different fields.

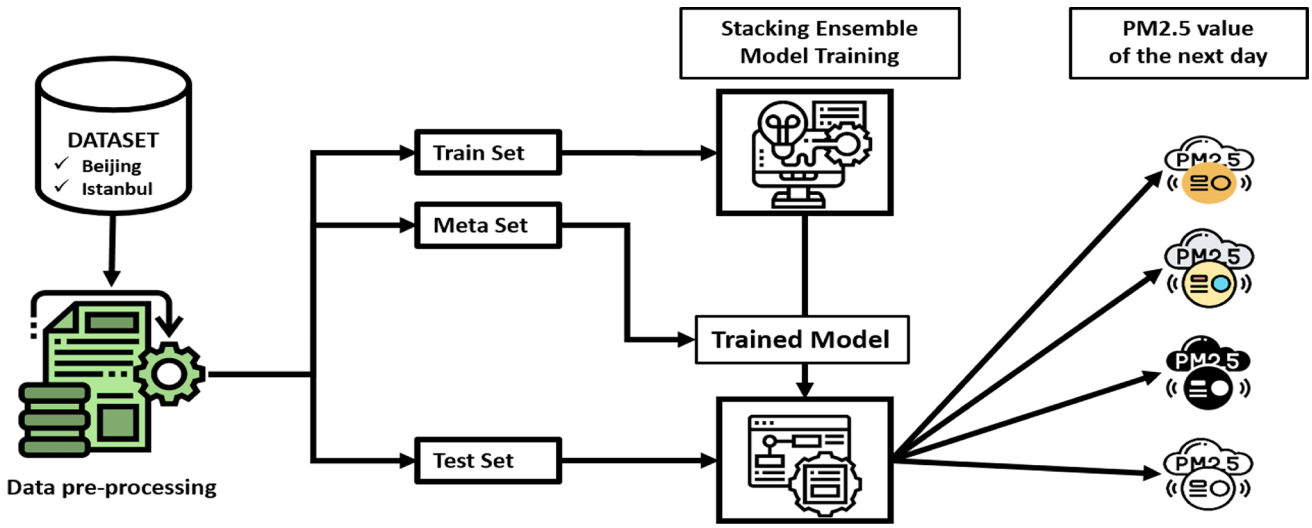

Fig. 4 Prediction method of $PM_{2.5}$ air quality value for the next day

Proposed stacking ensemble model

In this study, an ensemble of machine learning methods is utilized to address the limitations of traditional techniques employed in air quality forecasting. The stack ensemble machine learning approach involves training multiple models simultaneously and combining their outputs to generate predictions. This technique leverages the diverse features and strengths of each model, leading to more accurate and reliable predictions.

The first step involves training individual models to predict $PM_{2.5}$ values. Each model incorporates environmental factors (such as SO_2, NO_2, NOX, NO, O_3) that influence air quality. Figure 4 illustrates the proposed estimation method for $PM_{2.5}$ predictions.

The inputs of the trained models are listed below:

- SO_2
- NO_2
- NOX
- NO
- O_3
- $PM_{2.5}$ (Current day).

The output of the trained models is listed below:

- $PM_{2.5}$(Next day).

Upon closer examination of Fig. 4, it reveals the regression models employed in the initial step and their corresponding training process. The stacking ensemble regression method involves the utilization of multiple regression models simultaneously. In this approach, diverse regression models are trained, and predictions from each model are obtained. These predictions are then combined through a higher-level meta-model, yielding the final predictive value. Figure 4 showcases various stages, including data input, data preprocessing, model selection and training, prediction fusion, and final prediction. The fusion process is typically executed using a meta-model or meta-regression model, which incorporates the weights or importance of different predictions. The figure demonstrates the consolidation of predictions and the application of the meta-model. Moreover, it demonstrates the process of generating and evaluating the final estimate in the concluding stage. In summary, the phrase "Prediction of the $PM_{2.5}$ air quality value for the next day with the stacking ensemble regression method" denotes the estimation process utilizing a methodology that combines multiple regression models.

In this study, the estimation of $PM_{2.5}$ values has been successfully carried out using machine learning models like MLP, SVR, and RF. Each model possesses distinct advantages and disadvantages. Performance evaluation and comparative analysis are conducted to assess the successes and strengths of these models in air quality prediction. These analyses aid in identifying the most suitable model for estimating $PM_{2.5}$ values in Beijing and Istanbul. The collected dataset is split into two sets: the training dataset, used to train the stacking model, and the test dataset, utilized to assess the model's performance. Segmentation is achieved by randomly splitting the dataset, and typically an 80% training and 20% testing ratio is preferred. However, in this study, the dataset was divided into three subgroups.

The final prediction is generated by the Meta model through the combination of predictions from the trained models. This aggregation is commonly accomplished through weighting or combining the models' estimates.

In the weighting method, for instance, each model's estimate is multiplied by a specific weight based on its performance and then combined by taking the weighted average. The aggregation method is employed to enhance the accuracy of predictions and obtain more reliable outcomes. In the initial stage, each base model is separately trained and focuses on different features used for predicting air quality. The stacked ensemble model is designed to leverage several advantageous aspects of machine learning algorithms. The first stage of the stack includes base models such as SVR, RF, and multi-layer perceptron.

The choice of using MLP, SVR, and RF models in this study is based on several factors:

1. Using Different Types of Models: The aim of the study is to create a stronger prediction model by combining different machine learning models. MLP is an artificial neural network-based model capable of capturing complex relationships. SVR is an effective regression model that can be flexibly adapted through kernel functions. RF, as an ensemble learning method, combines multiple decision trees to capture dataset features and reduce noise. By combining these different types of models, the forecasting performance can be improved.
2. Model Diversity: The stacking ensemble approach relies on combining different models. In this case, incorporating MLP, SVR, and RF models with distinct characteristics can yield better results by increasing the diversity

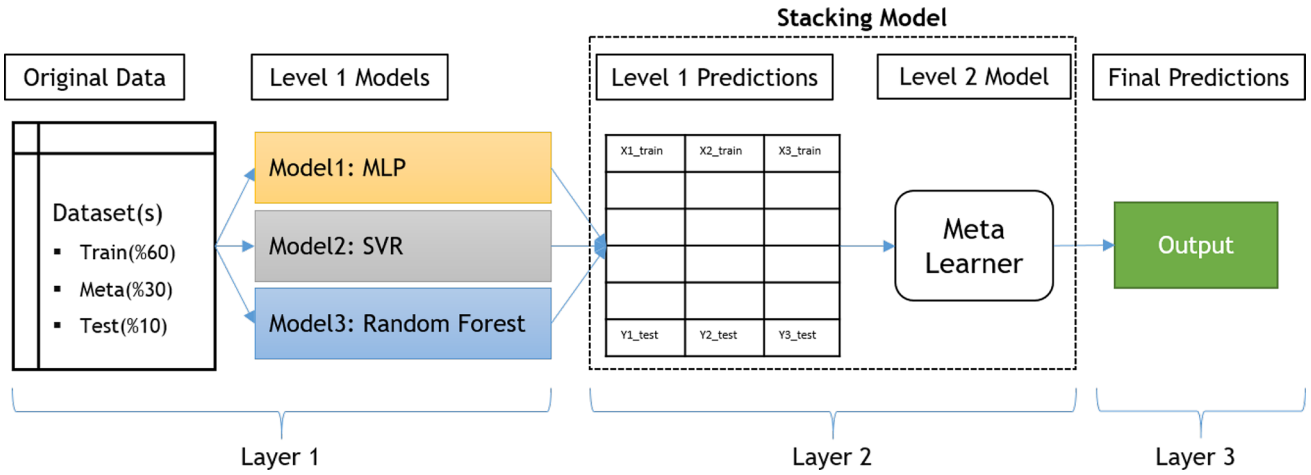

Fig. 5 3-layer stack batch learning architecture by temporal flow for air quality $PM_{2.5}$ prediction

Table 3 Distribution of datasets

Dataset(s)	Training data (%60)	Meta training data (%30)	Test data (%10)	Total
Beijing air quality	2.038	1.019	340	3.397
Istanbul air quality	2.155	1.077	359	3.591

of predictions. Model diversity enhances the stability of the ensemble structure and reduces overfitting.

3. Performance and Stability: MLP, SVR, and RF models have demonstrated successful performance in a substantial body of literature concerning regression problems, including air quality prediction. These models possess properties that enhance performance and stability, making them suitable for air quality forecasting. Combining different models of this nature can lead to more precise and reliable predictions.
4. Achievements in Current Literature: MLP, SVR, and RF models have achieved successful results in previous studies related to air quality forecasting and similar domains. Their documented success in the literature contributes to their preference and utilization in the stacking ensemble structure. For these reasons, MLP, SVR, and RF models were chosen for this study. By combining these models, the aim is to achieve improved performance and forecast accuracy in air quality prediction.

The second stage of the stack involves using the predictions from the underlying models as input for the stacking model, which serves as the meta-model. In this study, a solitary meta-model, which is commonly employed, namely the Linear Regression ("OLS") model, was selected. Figure 5 provides an architectural representation of the proposed method for air quality forecasting.

The stacking model will undergo further training to combine the predictions of the base models and estimate $PM_{2.5}$ values based on these predictions. The stacking model takes into account the weights and impacts of the underlying models when combining their predictions. At this stage, the ensemble model is trained and fine-tuned using cross-validation and optimization techniques. The goal is to create an ensemble model that can achieve higher accuracy in estimating $PM_{2.5}$ values by leveraging the combined predictions of the baseline models.

There is no universally "correct" approach to community stack modeling. It primarily relies on practical experience, combined with extensive experimentation and testing.

A popular method entails utilizing multiple machine learning algorithms with varying hyperparameters and subsequently feeding them into a metamodel. In more intricate scenarios, it is common to incorporate supplementary model layers preceding the metamodel (Cao et al. 2023).

Stacking batch modeling is guided by two fundamental principles. The first principle emphasizes the creation of a composite of algorithms capable of accurately estimating the relevant data within the specific field of study, employing diverse methodologies. The second principle is to ensure precision in separating the various layers of training, validation, and test data to prevent any leakage of information into subsequent layers (Akyol 2020). In this exercise, the target variable will be the $PM_{2.5}$ forecast for one period ahead (i.e., one day) in both the Beijing and Istanbul datasets.

The dataset is frequently divided into training and test sets when using learning models. The data in this study is split into three categories: training, meta-training, and testing for the stacking ensemble model that is being suggested.

The data's time series nature is taken into consideration throughout the training procedure.

The test set consists of the last 10% of the data in the time sequence, while the remaining portion is split into two groups. Two-thirds of the observations are designated as grid search training data, used to train the individual base models. The remaining one-third of the observations serve as meta-model training data, enabling the training of a meta-model that combines the predictions from the base models. By having separate training data for the base models and the meta-model, the model can effectively capture the patterns and relationships in the data (Kwon et al. 2019).

The distribution of the training and test data can be found in Table 3. After dividing the data into training (grid search and meta-model) and testing subsets, the data scaling phase is initiated prior to model design. StandardScaler() is applied exclusively to the grid search training data during the data scaling procedure, which is subsequently inserted and transformed accordingly.

The second set of data, known as the meta-model training set, is employed to train each base model. The purpose of this training is to generate predictions for the target variable. Subsequently, these predictions serve as explanatory variables in the meta-model.

The proposed stack ensemble model consists of the following stages:

1. Data Subset: The data is divided into three subsets:
2. 60% training set
3. 30% meta training set,
4. 10% test set.
5. Training: To determine the optimal hyperparameters, each of the three base models undergoes a GridsearchCV process. Once the hyperparameters of interest are tuned, the models are trained using the entire gridsearch training data, thus finalizing their hyperparameter settings.
6. Generating Base Model Predictions: The base models generate predictions on the meta model training set. These predictions serve as the explanatory variables to train the meta model on the target variable.
7. Base Model Predictions on Test Set: The base models make predictions on the test set. The generated predictions from the base models are subsequently inputted into the meta model.
8. Meta model Prediction and Scoring: The meta model utilizes the predictions made by the base models on the test set as inputs, generating its own predictions for the target variable. These predictions are then assessed and scored to evaluate the performance of the stacked ensemble model.

The clear relationship steps between the architectural initial models and the assembly results in Fig. 5 are listed below:

A. Training of each model: First of all, each basic model (MLP, SVR, RF) is ensured to make predictions on Training-data data.
B. These predictions are saved in a DataFrame named df_pred. The predictions of each model correspond to the columns of the DataFrame, respectively.
C. Then, using these predictions, the test scores of each model are calculated in a loop. MAE, RMSE, and R^2 scores are added to lists named (for example, test1_scores, test2_scores, and test3_scores).
D. After the loop completes, a DataFrame named results is created. This DataFrame contains the name of each model and their corresponding MAE, RMSE, and R^2 scores.
E. A new column named "Stack Model" is added to the df_pred DataFrame. This column contains the predictions of the "Stack Model" obtained by combining (with Linear Regression) the predictions of the base models using the trained meta data (meta-data).
F. Test scores for the "Stack Model" (MAE, RMSE, and R^2) are calculated and added to the results DataFrame.

This process generates the predictions of each model in the dictionary of trained models (reg_dic), saves them in a DataFrame, measures the performance of each model, combines them into a result DataFrame, and finally calculates the predictions of the meta model (combining model) and their performance. combines these results. This is a stacking process that makes a combined forecast using the predictions of the underlying models.

Results and discussion

The performance of the stacking ensemble machine learning model presented in the article was evaluated using various metrics. To determine the best evaluation method, two

Table 4 $PM_{2.5}$ prediction results in Istanbul dataset

Model(s)	Mean absolute error (MAE)	Root mean squared error (RMSE)	R^2
Multi-layer perceptron	7,43	9,51	0.87
Support vector regression	8,40	10,01	0.79
Random forest	7,37	9,50	0.88
Stack ensemble model	6,67	8,80	0.91

Table 5 $PM_{2.5}$ prediction results in Beijing dataset

Model(s)	Mean absolute error (MAE)	Root mean squared error (RMSE)	R^2
Multi-layer perceptron	13,44	18,91	0.66
Support vector regression	18,19	23,50	0.58
Random forest	13,13	17,70	0.65
Our stack model	12,42	15,84	0.77

model performance tests were considered (Madan et al. 2020): root mean square error (RMSE) and mean absolute error (MAE). Another metric used was R-square, a statistical measure of the fit of the regression model.

Performance metrics used in our study;

$$\mathrm{RMSE} = \sqrt{\left[\Sigma\left(P_i - X_i\right)^2/n\right]} \quad (1)$$

$$\mathrm{MAE} = \Sigma\left[P_i - X_i\right]/n \quad (2)$$

- P_i: The observed value (i_{th})
- X_i: The predicted value (i_{th})
- n: The total number of observations

$$R^2 = \mathrm{SSR}/\mathrm{SST} \quad (3)$$

- SSR is Sum of Squared Regression also known as variation explained by the model.
- SST is Total variation in the data also known as sum of squared total.

The mean absolute error (MAE) represents the average absolute difference between the predicted values and the actual values. A lower MAE value indicates that the estimates are, on average, closer to the true values. The root mean squared error (RMSE) is the square root of the average squared difference between the predicted values and the actual values. Similar to MAE, a lower RMSE value suggests that the estimates are, on average, closer to the true values. R^2 a regression model is used to explain the relationship between the dependent variable (outcome variable) and the independent variables. Its values range between 0 and 1. The higher the value of R^2, the better the regression model is considered to fit the data.

Based on the information provided in Table 4, the $PM_{2.5}$ predictions of four different models (MLP, SVR, RF, and Stack Ensemble Model) were evaluated. The MAE values were examined, and it was found that the Stack Ensemble Model achieved the lowest MAE value of 6.67. This suggests that the Stack Ensemble Model produces predictions that are closer to the true values compared to the other three models. Similarly, when analyzing the RMSE values, it was observed that the Stack Ensemble Model attained the lowest RMSE value of 8.80. Similarly, when the R^2 values were analyzed, it was observed that the Stack Ensemble Model achieved the highest R^2 score of 0.91. This indicates that the Stack Ensemble Model tends to make more precise and accurate predictions compared to the other models. Based on these findings, it can be concluded that the Stack Ensemble Model outperforms the other three models in terms of $PM_{2.5}$ predictions. It demonstrates better performance by producing more accurate predictions with lower MAE, RMSE and higher R^2 values.

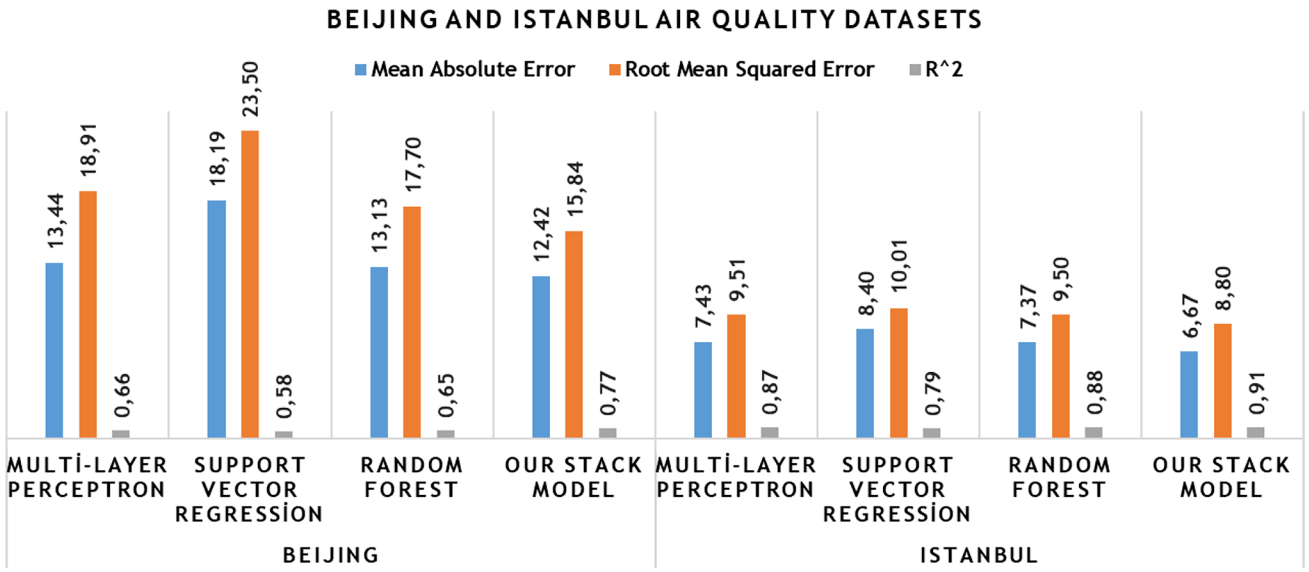

Fig. 6 $PM_{2.5}$ prediction graph of applied machine learning algorithms

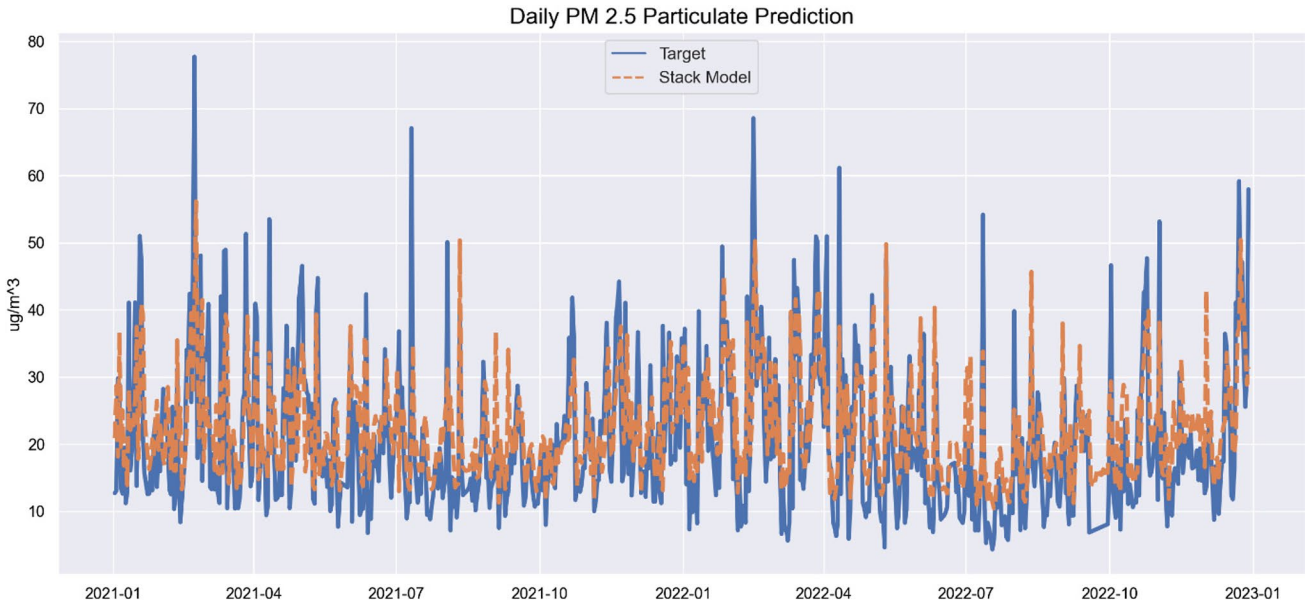

Fig. 7 $PM_{2.5}$ Istanbul daily forecast graph of stacking model

Based on the information provided in Table 5, the $PM_{2.5}$ estimations of the four models (MLP, SVR, RF, and Stack Ensemble Model) were evaluated. When examining the MAE values, it is evident that the Stack Ensemble Model achieved the lowest MAE value of 12.42. This indicates that the Stack Ensemble Model produces estimations that are closer to the true values compared to the other three models. Furthermore, considering the RMSE values, it is observed that the Stack Ensemble Model obtained the lowest RMSE value of 15.84. Moreover, when R^2 values are considered, it is seen that the Agglomeration Model has the highest R^2 score of 0.77. This suggests that the Stack Ensemble Model tends to generate more accurate and precise predictions compared to the other models. In summary, Table 5 demonstrates that the Stack Ensemble Model performs better in terms of $PM_{2.5}$ predictions and provides more accurate estimations compared to the other three models. With lower MAE, RMSE and higher R^2 values, the Stack Ensemble Model is

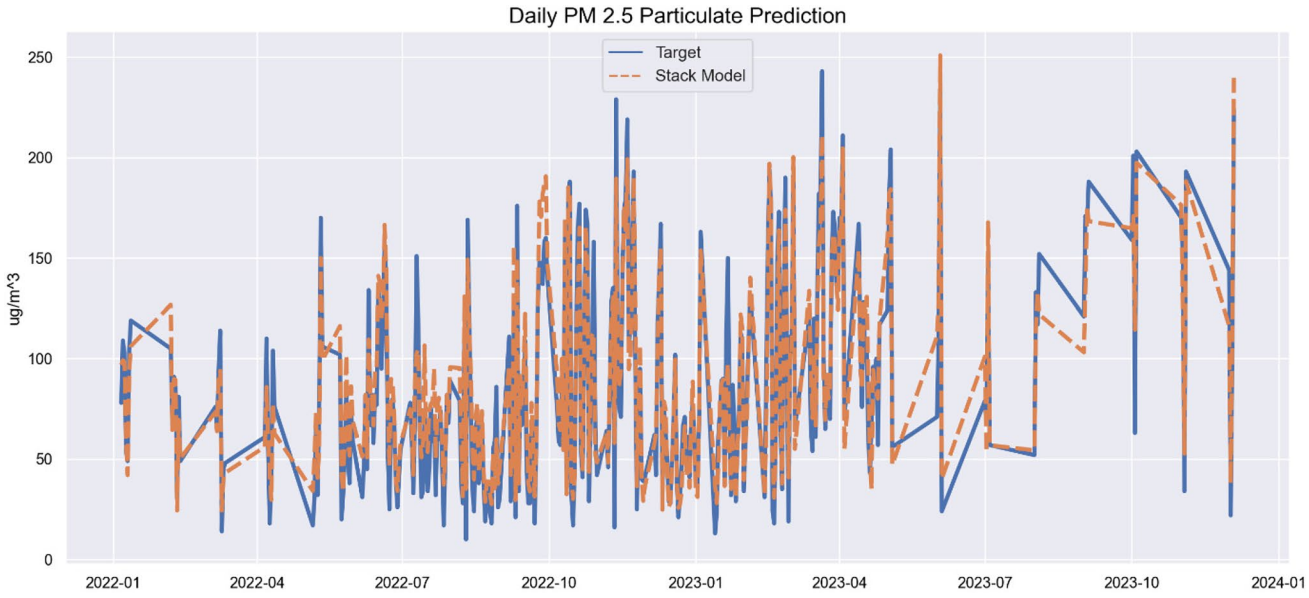

Fig. 8 $PM_{2.5}$ Beijing daily forecast graph of stacking model

capable of producing more reliable and accurate results for $PM_{2.5}$ predictions.

It is noteworthy that the values in Table 4 generally exhibit lower error values, whereas the values in Table 5 have higher error values. Upon examining the values in both tables, it is evident that the proposed model consistently achieves lower error values compared to the other three models. This suggests that the proposed model has a tendency to generate more accurate and reliable results for $PM_{2.5}$ predictions.

When comparing the results with the data presented in Fig. 6, it can be observed that the values in the Istanbul dataset generally exhibit lower errors. This suggests that the estimates in the Istanbul dataset are closer to the true values. On the other hand, the values in the Beijing dataset, the second dataset, tend to have higher errors. This indicates that the estimates in the Beijing dataset have more inaccuracies compared to the values in the Istanbul dataset, and they are not very close to the true values. In conclusion, when evaluating the results from both datasets, it can be concluded that the proposed model performs better in both cases and provides more accurate predictions. The Stack Ensemble Model emerges as a reliable and effective model for air quality prediction.

However, when conducting a comprehensive evaluation, it becomes evident that there is a significant disparity in the error values between both tables. This discrepancy can mainly be attributed to the statistical characteristics of the datasets presented in Tables 1 and 2. Upon closer examination of these tables, it can be observed that the standard deviation of $PM_{2.5}$ values in the Istanbul dataset is 17, whereas the standard deviation in the Beijing dataset is 65. With a difference of approximately 800% between the standard deviations, it is noteworthy that the disparity in the predicted $PM_{2.5}$ values amounts to approximately 80% (RMSE: 8.80/15.84).

Based on the data in Fig. 7, it is observed that the target values (actual $PM_{2.5}$) change over time and the values estimated by the Stack Model are close to these target values. In general, it is seen that the estimated values move in parallel with the target values.

Figure 8 illustrates the daily estimation of actual $PM_{2.5}$ values and depicts how these values change over time. Initially, the values start at a low level and then exhibit fluctuations. There is a period during which the values tend to increase, followed by a decrease before starting to increase again in the last period. Overall, the "Target" values demonstrate temporal variability.

The data series labeled "Stack Ensemble Model" represents the predicted values, which are displayed alongside the corresponding dates. In the figure, it is noticeable that the "Stack Ensemble Model" values are generally higher than the "Target" values. While the prediction model manages to make predictions that are quite close to the target values at certain points, it is worth noting that there are instances where the predictions deviate significantly from the target values.

Estimates appear to fit well in some cases, but deviate from target values in other cases. It can be said that the prediction model can be improved or need more data to obtain better results. It is also noteworthy that 'Target' values tend to increase over time and reach higher levels over time. This upward trend may indicate that targets have become more difficult or demand has increased over the relevant period. As a result, the graph shows that the prediction model can make accurate predictions in some cases, but there is room for improvement and the 'Target' values increase over time.

When examining past studies on air quality index estimation, it is evident that machine learning, deep learning, hybrid, and ensemble methods have been commonly employed. Castelli et al. (2020) successfully utilized SVR in their research. SVR offers the advantage of having kernel functions that can be applied to both linear and nonlinear problems, while also avoiding local minimums on the error surface. In a study by Janarthanan et al. (2021), SVR was combined with LSTM to create a model. The integration of LSTM, which has the ability to capture dependencies among historical data, with SVR yielded superior results compared to using SVR alone. Ensemble methods have also shown improvements in air quality predictions by combining various techniques such as Linear Regression-GRU (Lin et al. 2021), RF, XGBoost, and GBDT (Ma et al. 2023). Ensemble models aim to achieve enhanced performance by aggregating multiple models. In this study, the ensemble models outperformed the single methods (MLP, SVR, and RF). Additionally, Xiang et al. (2023) achieved successful results with the maximum probabilistic voting ensemble method.

In our study, we developed an air pollution prediction model using data from multiple monitoring stations to measure air quality in large cities such as Beijing and Istanbul. However, we did not have access to measurement data at different locations in a given city (e.g., different districts in Beijing or different neighborhoods in Istanbul). Therefore, the datasets presented in this study include daily average air quality parameters, which are the sum of all monitoring

stations in the respective cities. Therefore, we do not have the opportunity to make a comparison between stations in order to identify air quality variation in different locations (such as neighborhoods or districts). Instead, we focused on estimating air pollution levels at an overall city scale. The city-wide air quality estimates were based on a model combining data from all stations. In this case, it is not possible to make a specific assessment of the representative capacity of each station for a particular region, or of the different impacts in different geographical areas of the city.

The "stacking method" combines the forecasts of different models to create a more powerful forecasting model. At this stage, Linear Regression ("OLS"), a widely accepted meta-model for combining the forecasts of the leading models, was preferred. The simplicity and efficiency of the OLS model, its ability to produce statistically significant results, and its wide acceptance in the literature are the reasons behind this choice. Moreover, Linear Regression improves the model's interpretability by increasing the estimates' clarity. For these reasons, using the OLS in the Aggregation model is a strategic choice regarding reliability and performance.

Conclusion

This study showcases the effectiveness of a novel stacking ensemble machine learning approach for $PM_{2.5}$ value prediction in Beijing and Istanbul. The analyses and performance evaluations indicate that the estimation outcomes achieved through this method hold significant potential for applications in areas such as air quality management, public health, and environmental policy-making. Moreover, the insights provided regarding data preprocessing techniques and the implementation of diverse machine learning models can contribute to the progress of other researchers working on similar projects and the advancement of methodologies in this field.

In this study, the regression models used for the stacking ensemble method and the performance of these models were analyzed in detail. This analysis showed that the stacking ensemble method has lower MAE, RMSE and R^2 values compared to other methods. These results indicate that the stacking ensemble method can perform better in $PM_{2.5}$ estimation task.

This study highlights the significance of generating precise and trustworthy forecasts for effective air pollution management and control. Given the severe impact of air pollution on human health and the pressing need for environmental sustainability, the development of models capable of accurate air quality predictions becomes a crucial tool for policy makers and decision makers. By providing reliable insights and forecasts, such models can aid in formulating informed strategies and policies to mitigate the adverse effects of air pollution and ensure a healthier and more sustainable environment.

The limitations of this study should also be considered. For example, data preprocessing steps such as completing deficiencies in the data set and correcting abnormal values can be further improved. In addition, a more comprehensive study including a wider time period and more weather parameters can contribute to further improvement of forecast performance.

Future studies should further test the performance of the stacking ensemble method in estimating air pollution in different air quality datasets and different regions. In addition, a more comprehensive comparison can be made by considering different regression models and meta-model options. The performance of forecast accuracy can be tested by considering traffic and other meteorological factors. In this way, it will be possible to optimize the stacking ensemble model to provide the best performance in air pollution forecasting. This study makes an important contribution highlighting the usability of machine learning models and the potential of the stacking ensemble method in air pollution prediction. The results show that the stacking ensemble method performs better in $PM_{2.5}$ prediction and further research is needed in future studies.

Declarations

Conflict of interest There are no conflicts of interest, and all the authors are interested in publishing the manuscript.

Ethical approval This article contains no studies with human participants or animals performed by authors.

References

Air Quality Index Project, TW Beijing air pollution: real-time air quality index (2022). https://aqicn.org/city/beijing/

Akyol K (2020) Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection. Expert Syst Appl 148:113239. https://doi.org/10.1016/j.eswa.2020.113239

Ao Y, Li H, Zhu L, Ali S, Yang Z (2019) The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. J Petroleum Sci Eng 174:776–789. https://doi.org/10.1016/j.petrol.2018.11.067

Breiman L (2001) Random forests. Mach Learn 45:5–32

Cao Y, Liu G, Sun J, Bavirisetti DP, Xiao G (2023) PSO-Stacking improved ensemble model for campus building energy consumption forecasting based on priority feature selection. J Build Eng 72:106589. https://doi.org/10.1016/j.jobe.2023.106589

Castelli M, Clemente FM, Popovič A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity https://doi.org/10.1155/2020/8049504

Chang YS, Abimannan S, Chiao HT, Lin CY, Huang YP (2020) An ensemble learning based hybrid model and framework for air pollution forecasting. Env Sci Poll Res 27:38155–38168. https://doi.org/10.1007/s11356-020-09855-1

Chen B (2020) Air quality index forecasting via deep dictionary learning. IEICE Trans Inf Syst 103(5):1118–1125. https://doi.org/10.1587/transinf.2019EDP7296

Chen MH, Chen YC, Chou TY, Ning FS (2023) $PM_{2.5}$ concentration prediction model: a CNN–RF ensemble framework. Int J Environ Res Public Health 20(5):4077. https://doi.org/10.3390/ijerph20054077

Chen R, Liang CY, Hong WC, Gu DX (2015) Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. Appl Soft Comput 26:435–443. https://doi.org/10.1016/j.asoc.2014.10.022

Dong X, Yu Z, Cao W, Shi Y, Ma Q (2020) A survey on ensemble learning. Front Comput Sci 14:241–258. https://doi.org/10.1007/s11704-019-8208-z

Fang H, Feng Y, Zhang L, Su M and Yang H (2020) A long short-term memory neural network model for predicting air pollution index based on popular learning. In: Database systems for advanced applications. DASFAA 2020 International Workshops: BDMS, SeCoP, BDQM, GDMA, and AIDE, Jeju, South Korea, September 24–27, 2020, Proceedings 25. Springer International Publishing, pp 190–199

Feng S, Gao D, Liao F, Zhou F, Wang X (2016) The health effects of ambient $PM_{2.5}$ and potential mechanisms. Ecotoxicol Environ Saf 128:67–74. https://doi.org/10.1016/j.ecoenv.2016.01.030

Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32(14–15):2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0

Gokul PR, Mathew A, Bhosale A, Nair AT (2023) Spatio-temporal air quality analysis and $PM_{2.5}$ prediction over Hyderabad City, India using artificial intelligence techniques. Ecol Inf 76:102067. https://doi.org/10.1016/j.ecoinf.2023.102067

Harishkumar KS, Km Y, Gad I (2020) Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

Janarthanan R, Partheeban P, Somasundaram K, Elamparithi PN (2021) A deep learning approach for prediction of air quality index in a metropolitan city. Sustain Cities Soc 67:102720. https://doi.org/10.1016/j.scs.2021.102720

Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electron Markets 31(3):685–695. https://doi.org/10.1007/s12525-021-00475-2

Juarez EK, Petersen MR (2022) A comparison of machine learning methods to forecast tropospheric ozone levels in Delhi. Atmosphere 13(1):46. https://doi.org/10.3390/atmos13010046

Karakuş CB, Yıldız S (2019) Hava kalite indeksi ile meteorolojik parametreler arasındaki ilişkinin çoklu regresyon yöntemi ile belirlenmesi. Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi 8(2):698–711. https://doi.org/10.28948/ngumuh.598118

Kumar K, Pande BP (2023) Air pollution prediction with machine learning: a case study of Indian cities. Int J Environ Sci Technol 20(5):5333–5348. https://doi.org/10.1007/s13762-022-04241-5

Kwon H, Park J, Lee Y (2019) Stacking ensemble technique for classifying breast cancer. Healthc Inf Res 25(4):283–288. https://doi.org/10.4258/hir.2019.25.4.283

Li Z, Gan K, Sun S, Wang S (2023) A new $PM_{2.5}$ concentration forecasting system based on AdaBoost-ensemble system with deep learning approach. J Forecast 42(1):154–175. https://doi.org/10.1002/for.2883

Liang YC, Maimury Y, Chen AHL, Juarez JRC (2020) Machine learning-based prediction of air quality. Appl Sci 10:9151. https://doi.org/10.3390/app10249151

Lin CY, Chang YS, Abimannan S (2021) Ensemble multifeatured deep learning models for air quality forecasting. Atmosph Poll Res 12(5):101045. https://doi.org/10.1016/j.apr.2021.03.008

Liu H, Li Q, Yu D, Gu Y (2019) Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl Sci 9(19):4069. https://doi.org/10.3390/app9194069

Ma J, Ma X, Yang C, Xie L, Zhang W, Li X (2023) An air pollutant forecast correction model based on ensemble learning algorithm. Electronics 12(6):1463. https://doi.org/10.3390/electronics12061463

Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2020 2nd international conference on advances in computing, communication control and networking (ICACCCN). IEEE, pp 140–145

Maltare NN, Vahora S (2023) Air quality index prediction using machine learning for Ahmedabad city. Digit Chem Eng 7:100093. https://doi.org/10.1016/j.dche.2023.100093

Pui DY, Chen SC, Zuo Z (2014) $PM_{2.5}$ in China: measurements, sources, visibility and health effects, and mitigation. Particuology 13:1–26. https://doi.org/10.1016/j.partic.2013.11.001

Sarkar N, Gupta R, Keserwani PK, Govil MC (2022) Air quality index prediction using an effective hybrid deep learning model. Environ Poll 315:120404. https://doi.org/10.1016/j.envpol.2022.120404

Sethi JK, Mittal M (2019) A new feature selection method based on machine learning technique for air quality dataset. J Stat Manag Syst 22(4):697–705. https://doi.org/10.1080/09720510.2019.1609726

SIM (Sürekli izleme merkezi) | T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı (2023). https://sim.csb.gov.tr/

Wang B, Eum KD, Kazemiparkouhi F, Li C, Manjourides J, Pavlu V, Suh H (2020) The impact of long-term $PM_{2.5}$ exposure on specific causes of death: exposure-response curves and effect modification among 53 million US Medicare beneficiaries. Environ Health 19:1–12. https://doi.org/10.1186/s12940-020-00575-0

Wang D, Yue X (2019) The weighted multiple meta-models stacking method for regression problem. In: 2019 Chinese control conference (CCC). IEEE, pp 7511–7516

WHO (2022) Household air pollution. 28 Nov 2023

Xiang X, Fahad S, Han MS, Naeem MR, Room S (2023) Air quality index prediction via multi-task machine learning technique: spatial analysis for human capital and intensive air quality monitoring stations. Air Qual Atmos Health 16(1):85–97. https://doi.org/10.1007/s11869-022-01255-3

Yang J, Yan R, Nong M, Liao J, Li F, Sun W (2021) $PM_{2.5}$ concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. Atmos Poll Res 12(9):101168. https://doi.org/10.1016/j.apr.2021.101168

Yurtsever M, Emeç M (2023) Potable water quality prediction using artificial intelligence and machine learning algorithms for better sustainability. Ege Academic Rev 23(2):265–278. https://doi.org/10.21121/eab.1252167

Zhang Q, Jiang X, Tong D, Davis SJ, Zhao H, Geng G et al (2017) Transboundary health impacts of transported global air pollution and international trade. Nature 543(7647):705–709. https://doi.org/10.1038/nature21712