Results in Engineering 26 (2025) 105557

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Prediction of BTEX concentrations in the air of Southern East Azerbaijan province, Iran using ensemble machine learning and feature analysis

Mansour Baziar [a], Negar Jafari [b], Ali Oghazyan [c], Amir Mohammadi [d,e], Ali Abdolahnejad [b,1,*], Ali Behnami [b,f,1,*]

[a] Department of Environmental Health Engineering, Ferdows Faculty of Medical Sciences, Birjand University of Medical Sciences, Birjand, Iran
[b] Department of Environmental Health Engineering, Maragheh University of Medical Sciences, Maragheh, Iran
[c] Department of Environmental Health, School of Public Health, Sabzevar University of Medical Sciences, Sabzevar, Iran
[d] Social Determinants of Health Research Center, Clinical Research Institute, Urmia University of Medical Sciences, Urmia, Iran
[e] Department of Environmental Health Engineering, School of Public Health, Urmia University of Medical Sciences, Urmia, Iran
[f] Department of Environmental Health Engineering, School of Public Health, Iran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
BTEX
Air quality
Machine learning
Stacking
Feature importance
Public health

ABSTRACT

BTEX (Benzene, Toluene, Ethylbenzene, and Xylene) compounds are prominent air pollutants with severe implications for human health. Prolonged exposure to these volatile organic compounds (VOCs) has been associated with respiratory problems, cancer, and neurological disorders. Consequently, accurate prediction of their concentrations is vital for safeguarding public health and ensuring environmental safety. In this study, we introduce the MLs (XGBRegressor, AdaBoostRegressor, ExtraTreesRegressor, and CatBoost)-Stacked-Extra Trees ensemble, an innovative machine learning approach to predict BTEX concentrations. The initial model selection process employed the LazyRegressor library, which efficiently evaluates a wide array of regression models and provides essential performance metrics. Based on R-squared values, the top-performing models identified were XGBRegressor, AdaBoostRegressor, and ExtraTreesRegressor. To further optimize the stacking ensemble, CatBoost, a high-performing model not included in LazyRegressor, was incorporated. A thorough feature analysis identified key predictors influencing BTEX concentrations, including PM_{10}, $PM_{2.5}$, humidity, temperature, wind speed, and UV index. Additionally, the contributions of each model within the ensemble were assessed, highlighting the advantages of integrating predictions from multiple models to enhance accuracy. Our findings indicate that the MLs-Stacked-Extra Trees ensemble significantly outperforms individual models, achieving R^2 values of 1.0 and 0.998 for training and testing datasets, respectively. This research underscores the potential of advanced machine learning techniques to monitor air quality and guide policy decisions aimed at mitigating health risks associated with VOCs exposure.

1. Introduction

Air pollution has emerged as a critical global issue due to its detrimental impacts on human health and the environment [1–3]. Ambient air pollution comprises particulate matter (PM), various gases, and organic and inorganic compounds [4]. Among these pollutants, volatile organic compounds (VOCs) represent a significant category prevalent in urban and industrial areas [5]. Notably, BTEX compounds—Benzene, Toluene, Ethylbenzene, and Xylene—are of particular concern due to their high volatility and toxicity [6,7]. BTEX compounds are routinely released into the atmosphere from sources such as fossil fuel combustion, motor vehicle emissions, industrial activities, and the use of organic solvents [8,9]. The US Environmental Protection Agency (USEPA) classifies BTEX chemicals as hazardous air pollutants [10,11]. Exposure to BTEX via inhalation is linked to an elevated risk of cancer and adverse effects on the central nervous system (CNS), respiratory system, kidneys, liver, and reproductive system [8]. Given the critical importance of monitoring and managing air quality, accurately predicting BTEX concentrations is essential for protecting public health and advancing environmental management efforts.

* Corresponding authors.
E-mail addresses: baziar.ehe@gmail.com (M. Baziar), n64jafari@gmail.com (N. Jafari), aoghazyan@yahoo.com (A. Oghazyan), amahammadi@gmail.com (A. Mohammadi), abdolahnejad.a@gmail.com (A. Abdolahnejad), ali.behnami64@gmail.com (A. Behnami).
[1] The corresponding authors are contributed equally to this work.

https://doi.org/10.1016/j.rineng.2025.105557
Received 13 February 2025; Received in revised form 10 May 2025; Accepted 29 May 2025
Available online 30 May 2025

Predicting air pollution is essential for maintaining environmental health, protecting public health, ensuring regulatory compliance, and supporting informed decision-making at individual, community, and governmental levels. It serves as a proactive measure to mitigate the detrimental effects of air pollution on both the environment and human health [13]. Environmental contamination processes are inherently complex, making direct quantification impractical. Additionally, identifying pollution sources, understanding their propagation and distribution patterns, and forecasting their behavior over time pose significant challenges. Consequently, developing techniques to model these processes and predict their evolution is essential [12].

Recent advancements in machine learning (ML) techniques have revolutionized environmental monitoring [14,15]. These algorithms are highly effective at predicting complex, nonlinear relationships within large datasets, which traditional statistical methods often struggle to overcome. In recent years, various ML models have been employed to forecast air pollutant concentrations using data from air quality monitoring stations and meteorological conditions [14,16,17]. The application of ML extends beyond predictive modeling to include the identification of key factors influencing air pollutant emissions [18,19]. In addition to emissions from diverse sources, meteorological parameters and their interactions significantly affect pollutant levels. Variables such as temperature, humidity, wind speed, atmospheric stability, and solar radiation play a critical role in the dispersion and dilution of air pollutants [10,16].

Traditional forecasting methods often struggle with inconsistencies due to the complex and nonlinear nature of air pollutants. In contrast, artificial intelligence (AI) and ML techniques have recently emerged as powerful tools for improving predictive accuracy in this field. These methods are notable for their adaptive learning capabilities, high precision, and ability to handle high-dimensional datasets effectively. ML techniques are particularly well-suited for uncovering intricate relationships and resolving multicollinearity issues among variables. Additionally, ML enables the quantitative assessment of pollution source impacts and facilitates monitoring within complex relational frameworks [16,20,21]. ML techniques generally outperform traditional statistical models in addressing nonlinear problems. As data-driven approaches, ML models are highly effective at uncovering underlying relationships between inputs and outputs [22,23]. They have been extensively applied to predict atmospheric pollutant concentrations at both regional and global scales [20,24,25].

In recent years, numerous algorithms have been employed for predicting air pollutant concentrations, including Artificial neural network (ANN) [26,27], Support Vector Machine (SVM) [28,29], gradient-based optimizer (GBO) [30] convolutional neural network (CNN) [31], MLR [32,33], adaptive teaching-learning-based optimization and differential evolution (ATLDE) [34], Adaptive neuro fuzzy inference system (ANFIS) [32,35], Random forest (RF) [36–39], Decision Tree (DT) [40,41], Category Boosting (Catboost) [42,43], eXtreme Gradient Boosting (XGboost) [43–45], Adaptive Boosting (Adaboost) [46–48], Long Short-Term Memory (LSTM) [49,50] and hybrid models [51–53]. One study employed a CNN-based machine learning model integrated with absorption spectroscopic gas sensing technology to simultaneously measure BTEX concentrations. The results demonstrated an R-squared value greater than 0.96 for benzene and over 0.99 for toluene, ethylbenzene, and xylene, highlighting the model's strong predictive capability for BTEX levels [31].

However, the proliferation of modeling approaches presents a significant methodological challenge: the selection of optimal models often remains arbitrary, frequently based on researcher preference or limited comparative analyses. This study addresses this gap by employing LazyRegressor, a tool that facilitates statistically grounded model selection through the automated evaluation of numerous algorithms under standardized conditions. Unlike conventional ad hoc comparisons, LazyRegressor: (1) systematically assesses predictive consistency across multiple validation folds, (2) objectively ranks models based on their ability to capture relationships between pollutants and predictors, and (3) mitigates selection bias by exhaustively testing over 40 regression algorithms. This data-driven approach is especially valuable in environmental applications, where different algorithms may capture distinct facets of atmospheric behavior—for example, ANNs for modeling nonlinearities and tree-based methods for capturing complex feature interactions.

Due to the complicated nonlinear relationships between predicted variables and inputs, a single ML model may face challenges in achieving high predictive accuracy [54]. Ensemble models are created by combining multiple individual models to produce more accurate predictions than any single model can achieve on its own. Stacked models build on this concept by employing a meta-model that optimally integrates the predictions of the base models. For instance, a study conducted in Kaohsiung, Taiwan, utilized geographically weighted regression, hybrid Kriging–land-use regression (LUR) models, and two machine learning algorithms—RF and XGBoost—to estimate BTEX concentrations. Initially, the hybrid Kriging-LUR models explained 37–52 % of the variance in BTEX concentrations. However, when XGBoost was applied, the models' explanatory power increased significantly, accounting for 61–79 % of the variance [55]. In another study conducted in Kuwait using air quality data from 2022 to 2024, a novel hybrid model was developed to enhance the prediction of benzene concentrations across three industrial zones. The results demonstrated the model's strong predictive performance, offering valuable insights for air quality management and pollution mitigation in industrial environments [56]. This approach is increasingly favored for regression and classification tasks and has demonstrated notable success in predicting contamination events by analyzing multiple quality parameters [54,57].

This study presents a novel approach to predicting BTEX air pollution levels by integrating environmental factors and particulate matter (PM) with advanced ML techniques. The research involves a comprehensive comparison of various machine learning models, including XGBoost, AdaBoost, Extra Trees, and CatBoost, and introduces a stacked ensemble approach where Extra Trees serves as the meta-learner (MLs-Stacked-Extra Trees Ensemble). The key innovation of this study is the implementation of the MLs-Stacked-Extra Trees Ensemble learning framework, with LazyRegressor employed as an efficient and straightforward tool for model selection. This framework enhances the predictive accuracy of BTEX concentration models, which is critical for effective air quality monitoring and public health management. Furthermore, the study introduces a dual-layer feature importance analysis to elucidate the contributions of environmental factors and particulate matter to BTEX levels. It also evaluates the individual contributions of each model within the ensemble to the overall prediction. By leveraging the strengths of diverse models, the MLs-Stacked-Extra Trees Ensemble approach significantly outperforms standalone models, demonstrating its potential for tackling complex environmental prediction challenges.

The results demonstrate that the ensemble approach significantly enhances predictive accuracy, effectively addressing the challenges associated with environmental prediction tasks [58,59]. This study pursues three key objectives: (1) comprehensive data characterization through Shapiro-Wilk normality testing and Spearman correlation analysis of environmental factors (PM_{10}, $PM_{2.5}$, humidity, temperature, wind speed, UV index) and BTEX concentrations, coupled with dual-layer feature importance evaluation; (2) development of an advanced MLs-Stacked-Extra Trees ensemble model utilizing LazyRegressor-selected base algorithms (XGBoost, AdaBoost, Extra Trees, CatBoost) with Extra Trees meta-learner integration; and (3) rigorous performance validation against conventional machine learning models using seven evaluation metrics (MAE, MSE, MAPE, MedAE, NSE, IA, R^2) to demonstrate predictive superiority in BTEX concentration estimation.

2. Material and methods

2.1. Study area

Maragheh, the second-largest city in East Azerbaijan Province, is located 135 km south of the provincial capital, situated on the southern slope of Sahand Mountain. The city lies adjacent to the Sufi-Chay River and spans an area of 26 square kilometers, with a population of approximately 185,000 residents. Positioned at an elevation of 1,477 meters above sea level, Maragheh is located between latitudes 37°1′–37°45′N and longitudes 46°9′–46°44′E [10,60]. Fig. 1 illustrates a map of the study area, including the locations of BTEX sampling points.

2.2. Sampling and analysis methods

Fifteen sampling locations were selected for ambient BTEX sampling, taking into account the city's traffic volume. These included three stations in low-traffic areas, six stations in medium-traffic areas, and six stations in high-traffic areas. At a certain distance from the main streets and in the urban environment, in order to quantify the actual concentration to which citizens are exposed. To capture diurnal variations in BTEX concentrations, samples were collected in two timeframes: morning (09:00–12:00) and evening (17:00–20:00). A total of 60 samples were collected during the study period, which spanned from 3 February 2021 to 6 November 2021, covering all four seasons. Sampling was conducted under stable atmospheric conditions, avoiding intense wind or precipitation.

Meteorological parameters, including temperature, air pressure, wind speed, wind direction, UV index, and relative humidity, were recorded during the sampling period. The NIOSH 1501 method was employed for BTEX sampling and analysis [61]. Air samples were collected at 1.5 m above ground level using charcoal sorbent tubes (SKC Inc., England, 226-01) and a vacuum pump (SKC Inc., England) operating at a flow rate of 0.2 μL/min for two hours. PM_{10} and $PM_{2.5}$ concentrations were measured using the GRIMM Model EDM180 at 1.5 m above ground level. After sampling, the activated carbon-filled glass tubes were transported to the laboratory for analysis using gas chromatography with a flame ionization detector (GC-FID). Additional details regarding sample analysis are provided in our previous study [10].

2.3. Model selection, hyperparameter tuning, and stacking ensemble approach

The LazyRegressor library offers an efficient and automated approach for evaluating a wide range of regression models. It includes 42 algorithms, spanning from basic linear models to advanced ensemble techniques. The library quickly fits multiple models to the dataset using default hyperparameters and generates essential performance metrics, such as R-Squared, RMSE, and Time Taken, for each model. Based on R-Squared values, the top-performing models identified in this study are XGBRegressor, AdaBoostRegressor, and ExtraTreesRegressor. To further enhance the stacking ensemble, CatBoost—a model not included in the LazyRegressor library—was incorporated. CatBoost excels in handling categorical variables and delivers robust predictive performance. By integrating CatBoost with the top-performing models identified by LazyRegressor, the ensemble harnesses the unique strengths of each model. This approach improves generalization, reduces overfitting, and achieves higher predictive accuracy. Literature evidence supports the efficacy of these models in managing complex relationships and high-dimensional data across both small and large datasets [42,62]. XGBoost and CatBoost, both gradient boosting models, are highly effective in regression tasks due to their superior performance and capability to handle large datasets efficiently. Additionally, they exhibit strong adaptability when applied to smaller datasets [62,63]. AdaBoost,

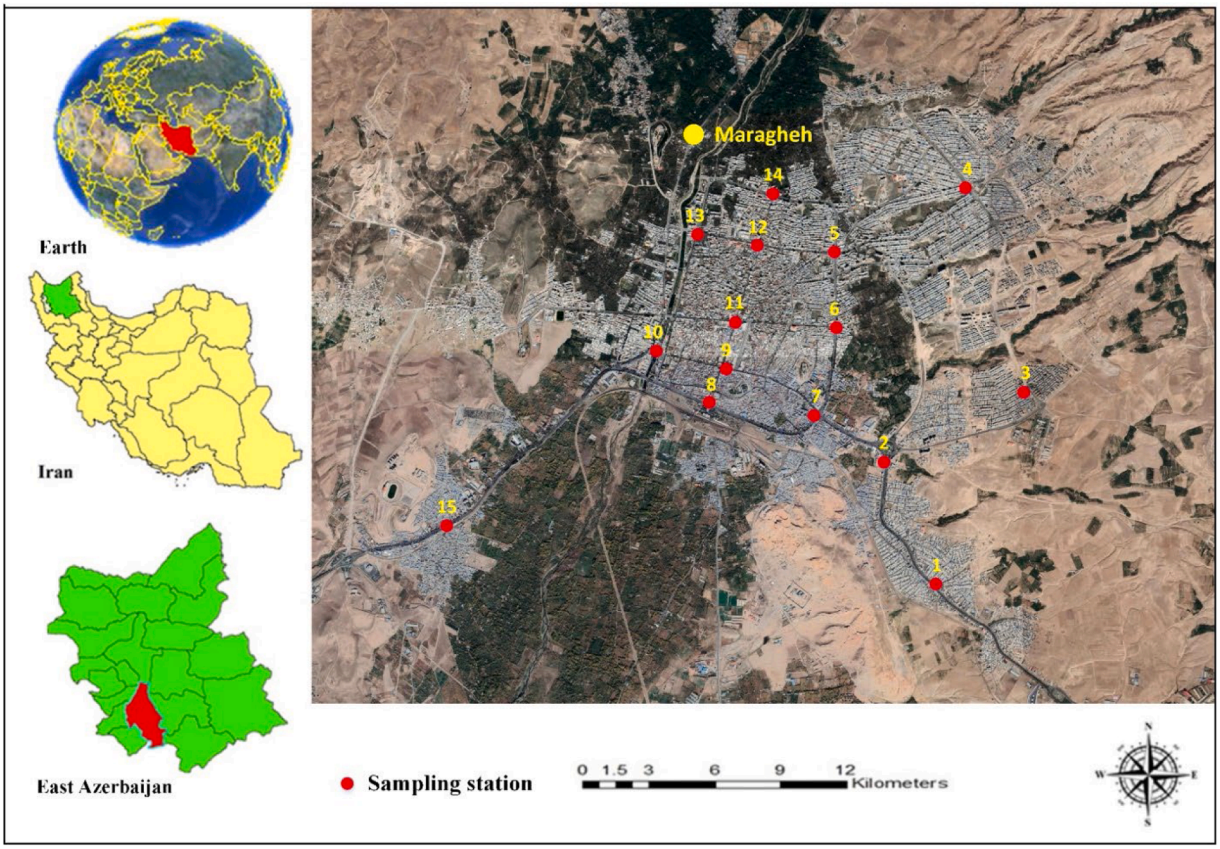

Fig. 1. BTEX sampling points of the study area.

an ensemble method, is utilized for its ability to enhance weak models by iteratively adjusting the weights of misclassified data points, effectively reducing bias regardless of dataset size [64,65]. Extra Trees, an ensemble of decision trees, is incorporated for its robustness, flexibility, and reduced susceptibility to overfitting, especially when dealing with noisy data, making it highly suitable for both small and large datasets [66,67]. The MLs-Stacked-Extra Trees ensemble combines the outputs of first-level models, with Extra Trees serving as the meta-learner to harness the strengths of individual models and improve predictive accuracy. This stacking approach enables the refinement of predictions through the meta-learner's learning process. To optimize each model's performance, hyperparameters are fine-tuned using a trial-and-error method, ensuring optimal configurations. For XGBoost, CatBoost, and AdaBoost, key hyperparameters such as the number of estimators, learning rate, and tree depth are adjusted to achieve a balance between model complexity and performance [59,62,63,67]. For Extra Trees, the tuning process emphasizes optimizing the number of trees, maximum depth, and the number of features considered for splitting [66,67]. Additionally, the stacking ensemble model undergoes optimization of its meta-learner to ensure the most effective combination of predictions from the first-level models [63]. Hyperparameter tuning is conducted iteratively, exploring various configurations to determine the optimal set that enhances predictive accuracy and robustness [68,69]. Fig. 2 presents the methodology employed in this study for developing predictive models, providing a detailed overview of the key steps involved in model training, evaluation, and optimization. This approach is designed to ensure the production of accurate, reliable, and robust predictions for BTEX air pollution levels based on environmental factors and particulate matter data. In this study, the dataset was randomly split into training (80 %) and testing (20 %) subsets, ensuring statistically adequate sample sizes for both model development and independent validation.

2.4. Performance evaluation metrics for regression models

This study utilizes a comprehensive set of performance metrics to rigorously evaluate the regression models (Table 1). The Mean Absolute Error (MAE) offers a straightforward measure of prediction accuracy by calculating the average of absolute differences between observed and predicted values. The Mean Squared Error (MSE), which emphasizes larger discrepancies, provides insights into the scale of prediction errors. To express model accuracy in relative terms, the Mean Absolute Percentage Error (MAPE) calculates the average error as a percentage, making it particularly useful for comparative analyses. The Median Absolute Error (MedAE) serves as a robust metric by focusing on the median of absolute errors, thereby reducing sensitivity to outliers. The Nash-Sutcliffe Efficiency (NSE) evaluates the model's explanatory power by comparing the variance captured by the model to the total variance, with higher values indicating superior performance. Additionally, the Index of Agreement (IA) assesses the alignment between observed and predicted values, with values closer to 1 reflecting stronger agreement [70,71].

Table 1
Summary of performance metrics and their formulas used in the analysis.

Metric	Mathematical Formula
Mean Absolute Error	$MAE = (1/n) * \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
Mean Squared Error	$MSE = (1/n) * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Mean Absolute Percentage Error	$MAPE = (1/n) * \sum_{i=1}^{n} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert \times 100$
Median Absolute Error	$MedAE = \text{median}(\lvert y_i - \hat{y}_i \rvert)$
Nash-Sutcliffe Efficiency	$NSE = 1 - \left(\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \right)$
Index of Agreement	$IA = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (\lvert y_i - \bar{y} \rvert + \lvert \hat{y}_i - \bar{y} \rvert)^2}$
R-Squared	$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

2.5. Feature analysis

Feature analysis is a crucial aspect of machine learning, as it allows researchers to assess the impact and importance of input variables on model predictions [51,67,72]. By quantifying the contribution of each feature, this analysis offers valuable insights into the relationships between predictors and the target variable, enhancing the understanding of the underlying data dynamics [73,74]. This study conducts feature analysis in two stages. The first stage focuses on the primary predictors—PM_{10}, $PM_{2.5}$, humidity, temperature, wind speed, UV index, and BTEX—evaluating their individual contributions to the predictive

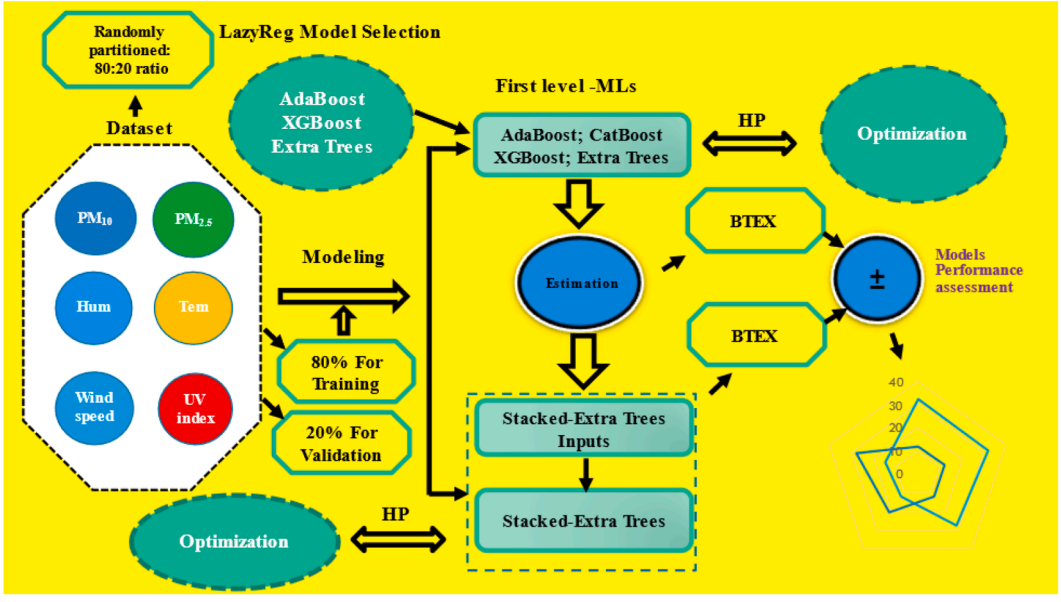

Fig. 2. Flowchart illustrating the methodology for model development, highlighting the key steps in training, evaluation, and optimization to achieve accurate and robust predictions.

performance of the base models. The second stage analyzes the MLs-Stacked-Extra Trees model to assess the importance of each base model's predictions, demonstrating how the ensemble approach enhances overall accuracy by leveraging the strengths of these models. This dual approach offers a comprehensive understanding of feature importance at both the input and model-output levels [58,59].

3. Results and discussion

3.1. Analysis of descriptive statistics and The Shapiro-Wilk test

Table 2 presents a detailed summary of the descriptive statistics for key environmental variables in the dataset, including PM_{10}, $PM_{2.5}$, humidity, temperature, wind speed, UV index, and BTEX, measured across 60 observations. PM_{10} has a mean of 41.49 and a standard deviation of 11.6, reflecting notable variability around its average. Its slightly positive skewness (1.07) suggests a longer tail on the right side of the distribution, while a kurtosis of 2.27 indicates a relatively peaked distribution compared to the normal curve. Similarly, $PM_{2.5}$ exhibits a mean of 18.77, with lower skewness (0.6) indicating a less pronounced right tail. Humidity and temperature both display negative skewness (-0.18 and -0.17, respectively), indicating a concentration of higher values within their ranges. Their kurtosis values suggest platykurtic distributions, characterized by flatter shapes compared to the normal distribution. Wind speed and UV index approximate normal distributions, as evidenced by low skewness and moderate kurtosis, highlighting the relative consistency of their values. In contrast, BTEX shows the highest mean (13.21) and a substantial standard deviation (12.73), coupled with pronounced positive skewness (2.33), pointing to a right-skewed distribution with potential high-value outliers. These statistical insights provide a deeper understanding of the variability and distribution patterns of these environmental variables.

The Shapiro-Wilk test, a statistical method used to evaluate the normality of a dataset, produces a statistic ranging from 0 to 1, where values closer to 1 indicate a higher likelihood of normality. This statistic, along with the p-value, determines whether the data significantly deviates from a normal distribution. The test results reveal that most variables in the dataset do not conform to a normal distribution. For PM_{10}, the statistic is 0.932 with a p-value of 0.0024, indicating non-normality. Conversely, $PM_{2.5}$ has a statistic of 0.973 and a p-value of 0.1966, suggesting normality. The humidity variable shows a statistic of 0.850 and a p-value of 2.98×10^{-6}, confirming non-normality, while temperature has a statistic of 0.814 and an extremely low p-value of 2.99×10^{-7}, also indicating non-normality. Similarly, wind Speed (statistic: 0.904, p-value: 0.00019) and UV Index (statistic: 0.815, p-value: 3.2×10^{-7}) fail the normality test. Finally, BTEX exhibits the most significant deviation from normality, with a statistic of 0.763 and a p-value of 1.82×10^{-8}. In summary, except for $PM_{2.5}$, which aligns with a normal distribution, all other variables—PM_{10}, humidity, temperature, wind Speed, UV Index, and BTEX—deviate significantly from normality. These findings highlight the need to employ non-parametric methods for subsequent analyses.

3.2. Spearman correlation analysis (non-parametric method)

Given that most variables in the dataset were not normally distributed, Spearman correlation analysis was employed (Fig. 3). Spearman's rank correlation is a non-parametric method that evaluates the strength and direction of monotonic relationships between two variables. Unlike Pearson's correlation, which assumes normality, Spearman's correlation assesses associations based on data ranks, making it more robust for non-normally distributed variables [75,76]. The Spearman correlation matrix reveals a complex network of relationships among the variables PM_{10}, $PM_{2.5}$, humidity, temperature, wind speed, UV index, and BTEX. Notably, strong positive correlations were observed between BTEX and particulate matter, with correlation coefficients of 0.77 for PM_{10} and 0.75 for $PM_{2.5}$. These results suggest co-emission from shared sources, such as combustion processes, vehicular emissions, and industrial activities. These findings are consistent with those reported in similar studies [77,78].

These findings indicate that both fine and coarse particulate matter significantly influence BTEX concentrations [10,79–81]. Humidity shows a moderate positive correlation with BTEX (0.19), suggesting that higher humidity levels may elevate BTEX concentrations through mechanisms such as aerosol adsorption or reduced dispersion under stagnant air conditions [82,83]. Conversely, temperature exhibits a weak negative correlation with BTEX (-0.18), which may be attributed to increased solar radiation, the production of hydroxyl (OH) radicals, and the photochemical breakdown of VOCs during warmer seasons [10, 84–86]. During colder seasons, factors such as low wind speeds, atmospheric inversions, emissions from home heating systems, and reduced mixing heights contribute to air stability and hinder the dispersion of pollutants, potentially leading to higher BTEX concentrations [85,87]. Wind speed shows a weak positive correlation with BTEX (0.10), indicating that increased air movement has minimal impact on localized concentrations [78,80,82,83]. In contrast, the weak negative correlation with the UV index (-0.29) underscores the role of photochemical reactions, where greater solar radiation promotes the breakdown of BTEX compounds, thereby reducing their atmospheric concentrations [10,64].

3.3. Model selection and evaluation of machine learning models

Fig. 4 illustrates the models selected using LazyRegressor, including XGBoost, AdaBoost, and ExtraTrees, which were chosen based on their R-squared values. LazyRegressor streamlines the model selection process by automatically evaluating multiple regression models and identifying the top performers, reducing the need for manual hyperparameter tuning. Additionally, CatBoost was included to further improve model performance. The evaluation of these machine learning models reveals notable differences in their predictive capabilities during the training and testing phases [70,71] (Table 3). XGBoost demonstrates exceptional training performance, with minimal errors (MAE: 0.163, MSE: 0.116, MAPE: 1.66) and near-perfect NSE and IA values (0.999). It also maintains strong generalization during testing, achieving MAE: 2.40, MSE: 8.27, MAPE: 32.78, with high NSE (0.904) and IA (0.977). CatBoost performs robustly during training, showing slightly higher errors compared to XGBoost (MAE: 0.570, MSE: 0.532, MAPE: 11.20) and strong NSE (0.997) and IA (0.999). However, it's testing

Table 2
Descriptive statistics of input variables and BTEX concentrations in the dataset (n = sample size).

Variable	n	Mean	Median	Std	Min	Max	Skewness	Kurtosis
PM_{10}	60	41.49	40.82	11.6	19.5	82.24	1.07	2.27
$PM_{2.5}$	60	18.77	18.46	6.42	4.17	39.5	0.6	0.86
Humidity	60	29.1	30.5	13.75	8	46	-0.18	-1.62
Temperature	60	15.87	18.5	9.78	1	28	-0.17	-1.73
Wind Speed	60	3.28	3	0.89	2	5	0.13	-0.81
UV Index	60	2.4	2.5	1.74	0	5	-0.06	-1.66
BTEX	60	13.21	9.62	12.73	0.56	71.94	2.33	6.71

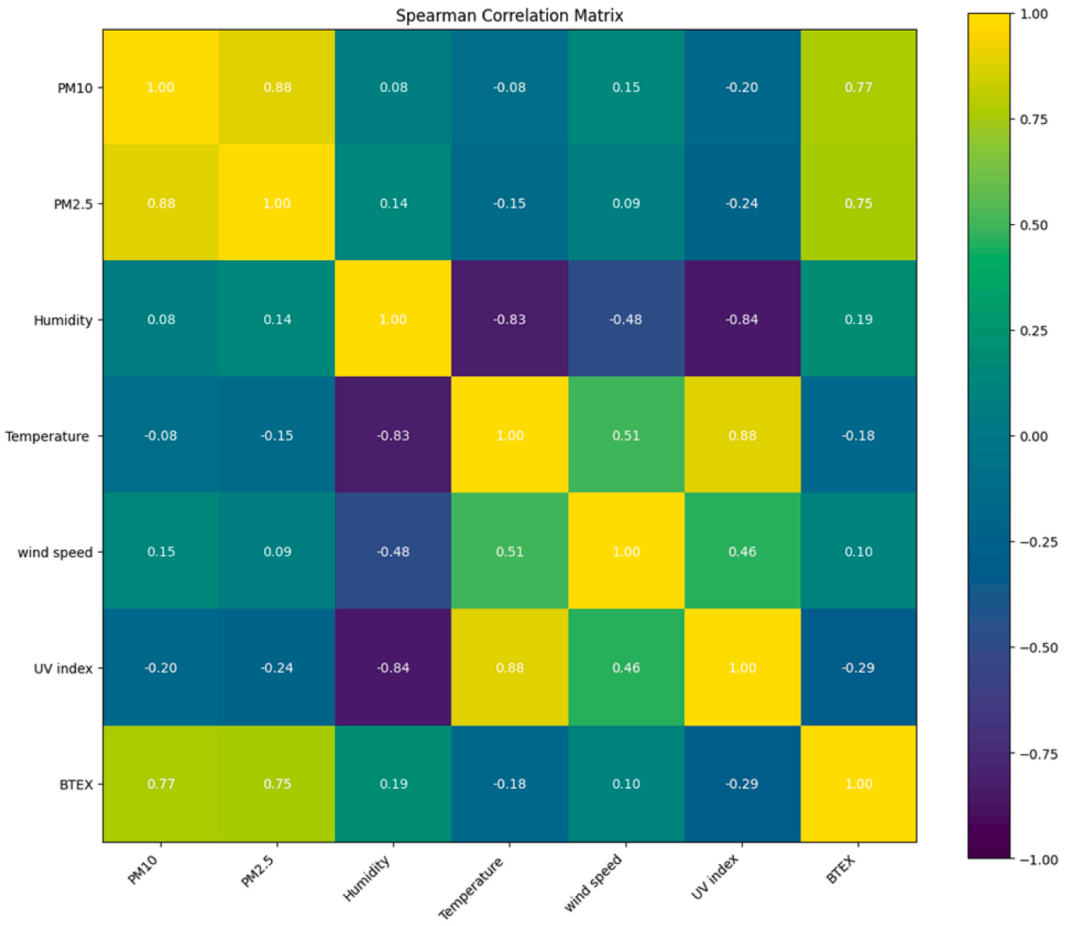

Fig. 3. Spearman correlation matrix illustrating the relationships among variables in the dataset.

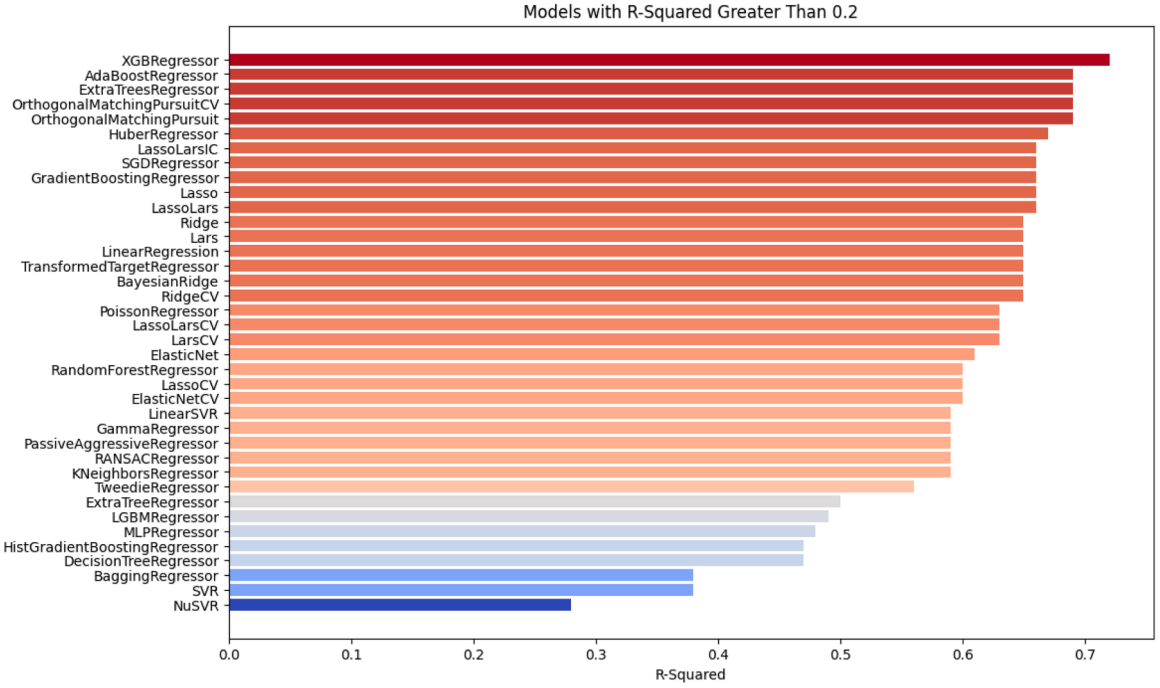

Fig. 4. Performance evaluation of machine learning models for BTEX concentration prediction, ranked by R^2 scores (testing set). Models were screened using the LazyRegressor library, with default hyperparameters.

Table 3
Comparative performance evaluation of machine learning models for BTEX concentration prediction.

Model/Metric	MAE	MSE	MAPE	MedAE	NSE	IA
Training						
XGBoost	0.163	0.116	1.66	0.064	0.999	0.999
CatBoost	0.570	0.532	11.20	0.460	0.997	0.999
AdaBoost	0.833	1.66	15.55	0.424	0.990	0.810
Extra Trees	~ 0	~ 0	~ 0	~ 0	1	1
MLs-Stacked-Extra Trees	~ 0	~ 0	~ 0	~ 0	1	1
Testing						
XGBoost	2.40	8.27	32.78	2.59	0.904	0.977
CatBoost	3.04	15.74	36.75	2.410	0.817	0.932
AdaBoost	3.262	16.34	35.21	2.415	0.810	0.938
Extra Trees	3.013	12.862	42.57	2.642	0.850	0.957
MLs-Stacked-Extra Trees	0.420	0.248	3.89	0.425	0.995	0.999

performance declines, with increased errors (MAE: 3.04, MSE: 15.74, MAPE: 36.75), indicating reduced generalization. AdaBoost delivers acceptable training results, with moderate errors (MAE: 0.833, MSE: 1.66, MAPE: 15.55), but its testing performance reflects relatively weaker generalization (MAE: 3.26, MSE: 16.34, MAPE: 35.21). Extra Trees, a traditional model, and the MLs-Stacked-Extra Trees ensemble (with Extra Trees serving as the meta-learner) achieve near-perfect training metrics, with errors approaching zero and the highest possible NSE and IA values, suggesting a tendency toward overfitting [62,70,88].

During testing, Extra Trees demonstrates satisfactory but not exceptional performance, with MAE: 3.013, MSE: 12.862, MAPE: 42.57, NSE: 0.850, and IA: 0.957. In contrast, the MLs-Stacked-Extra Trees ensemble significantly outperforms all other models, achieving exceptionally low errors (MAE: 0.420, MSE: 0.248, MAPE: 3.89) and near-perfect NSE (0.995) and IA (0.999). This performance underscores the ensemble's remarkable generalization capability, despite potential overfitting concerns during training. Among individual models, XGBoost strikes the best balance between training and testing performance, establishing itself as a reliable option for practical applications. Meanwhile, the superior accuracy of the stacked ensemble highlights its suitability for high-stakes predictive tasks, provided that overfitting is carefully addressed. The exceptionally high training scores of the base models—XGBoost, AdaBoost, Extra Trees, and CatBoost—initially indicated potential overfitting, particularly for the more complex algorithms. However, the stacked ensemble model (MLs-Stacked-Extra Trees) demonstrated significantly improved generalization, as reflected in its superior performance on the test set. It achieved an R^2 of 0.998, compared to a range of 0.894–0.927 for the base models, and exhibited lower error metrics—for instance, a MAPE that was 3.89 % lower than that of the best-performing base model. This improvement likely results from the ensemble's ability to mitigate individual model biases while harnessing their combined predictive strengths. Notably, the stacked model consistently maintained this strong performance across validation sets, confirming its practical reliability despite the overfitting tendencies of the base models.

Figs. 5 and 6 present radar plots that illustrate the performance of the models in terms of R^2 for the training and test datasets, respectively. For the training data, the R^2 values for XGBoost, AdaBoost, CatBoost, Extra Trees, and MLs-Stack-Extra Trees are 0.9995, 0.991, 0.998, 1, and 1, respectively, highlighting the exceptional performance of models like Extra Trees and MLs-Stack-Extra Trees, both of which achieve perfect R^2 values. For the test data, the R^2 values for XGBoost, AdaBoost, CatBoost, Extra Trees, and MLs-Stack-Extra Trees are 0.927, 0.894, 0.920, 0.894, and 0.998, respectively. These results underscore the robustness of MLs-

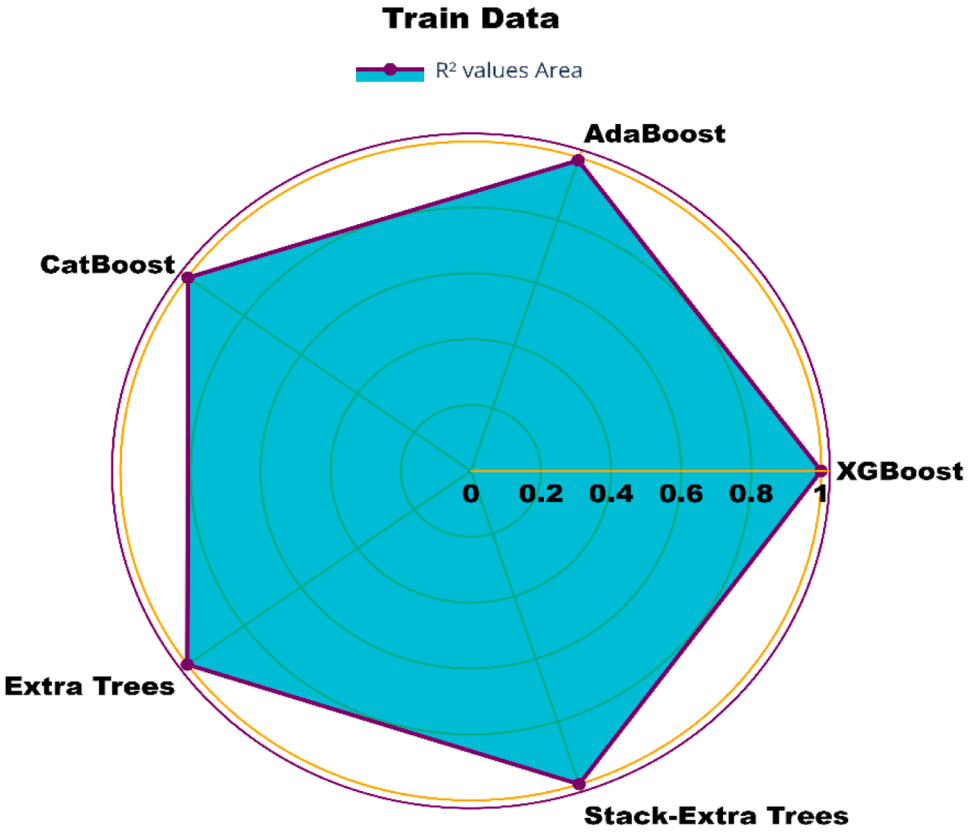

Fig. 5. Radar chart comparing R^2 scores of machine learning models for BTEX concentration prediction (training data). Models with data points closer to the outer circle (R^2=1.0) demonstrate better predictive performance. The best-performing models will have their vertices nearest to the circumference, while weaker models appear closer to the chart center.

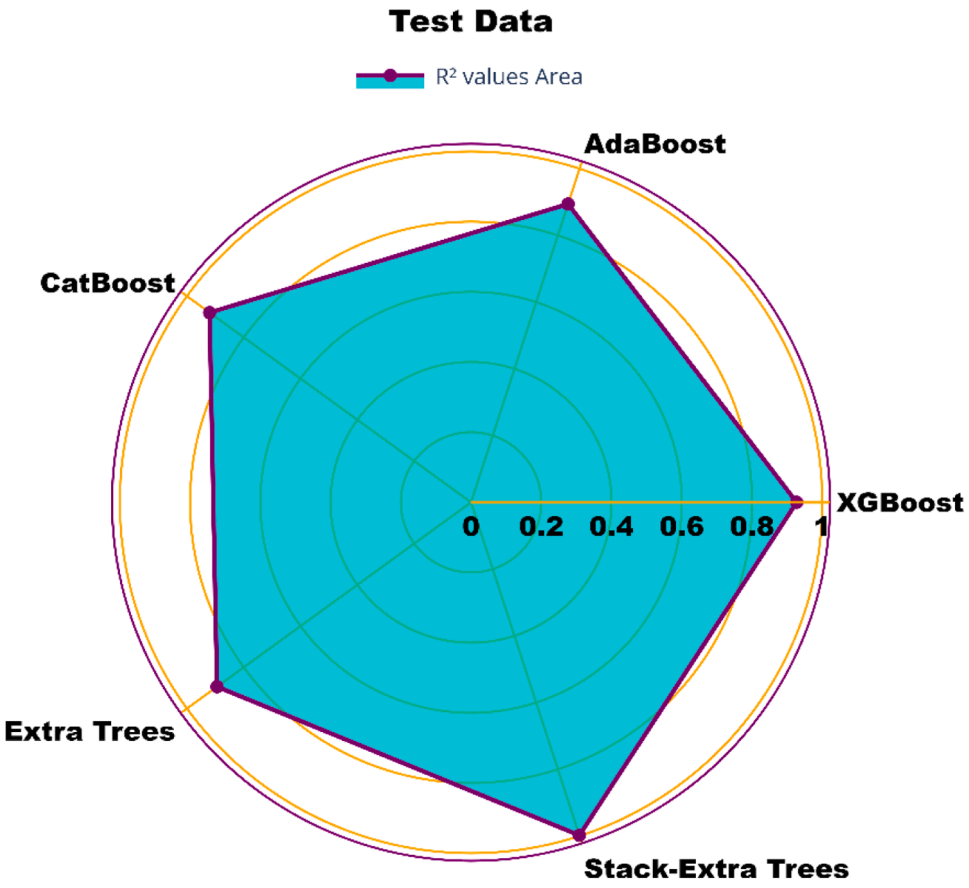

Fig. 6. Radar chart displaying R^2 scores for various machine learning models on the testing data, emphasizing the comparative performance of each model.

Stack-Extra Trees, which maintains a high R^2 of 0.998, while other models show a reduction in performance on unseen data. The radar plots offer a clear visual comparison of each model's performance across both the training and test datasets. In the radar plot, each axis represents a different model, and the distance from the center corresponds to the R^2 value. A greater distance from the center indicates a higher R^2, signifying better performance [89]. Models that maintain a consistent and large distance across both the training and test datasets, such as MLs-Stack-Extra Trees, are considered to exhibit strong generalization capabilities. Therefore, the radar plots serve as an intuitive tool for comparing the performance of multiple models in a visually compact format, effectively highlighting their strengths and weaknesses in terms of generalization [90].

The fluctuations in actual BTEX concentrations and their predicted values are illustrated in Fig. 7. This figure showcases the model's ability to replicate variations in BTEX levels, highlighting how closely the

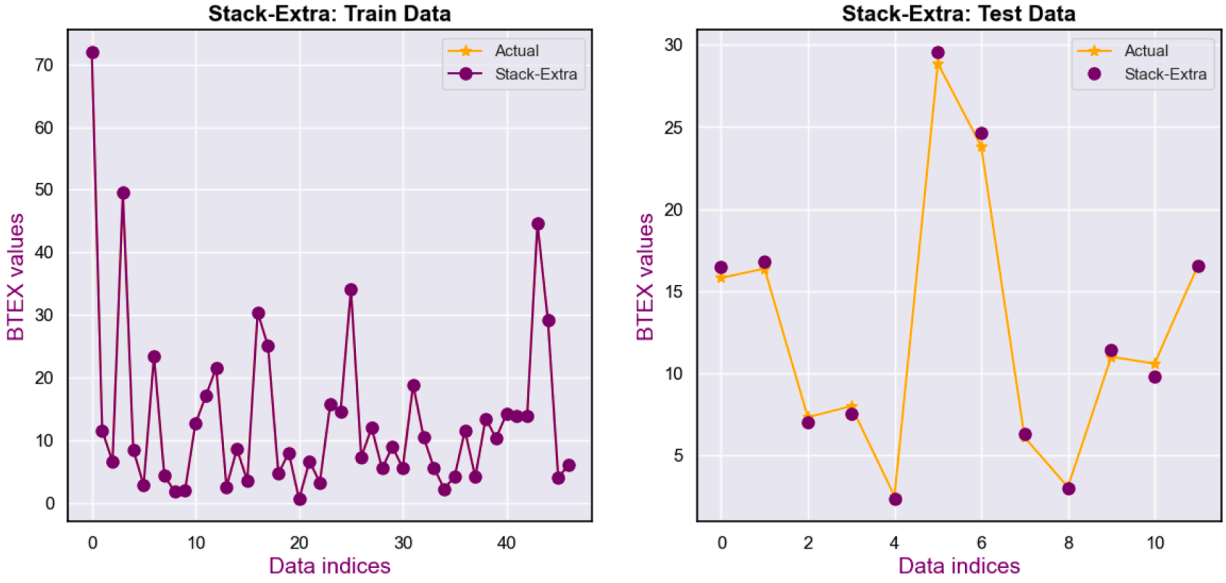

Fig. 7. Stacked ensemble model performance: Comparison of observed and predicted BTEX concentrations in training and test datasets.

predictions align with the observed data. The comparison emphasizes the model's effectiveness in capturing the trends and patterns of BTEX concentration fluctuations. Fig. 8 presents a scatter plot for the MLs-Stacked-Extra Trees ensemble, displaying the model's performance on both the training and test datasets. The scatter plot provides a visual representation of the model's predictive accuracy [58,59,88,91], illustrating how closely the predicted values align with the actual values for both datasets. It is important to note that the optimal hyperparameters for the MLs-Stacked-Extra Trees ensemble, which incorporates Extra Trees as a meta-learner, are carefully selected to enhance predictive performance. Specifically, the ExtraTreesRegressor is configured with 200 estimators, ensuring sufficient model complexity and diversity for robust predictions. The random_state is set to 230 to guarantee the reproducibility of results. To promote effective feature selection while preventing overfitting, a maximum of four features are considered at each split (max_features=4. Additionally, the model is set with no maximum depth (max_depth=None), allowing the trees to grow until they are pure, thus enabling the model to capture complex relationships within the data. These hyperparameters collectively contribute to the model's strong generalization ability and high accuracy during both training and testing phases [59,65,66]. The model tuning process followed a two-stage approach. First, random_state values ranging from 1 to 500 were systematically evaluated while keeping other parameters constant to identify the most reproducible configuration (random_state = 230). In the second stage, the remaining hyperparameters—n_estimators, max_features, and max_depth—were optimized using a trial-and-error strategy, adjusting each parameter sequentially and retaining changes only when they improved validation performance. To date, the Stacked-Extra Trees Ensemble model has not been applied to predict BTEX concentrations. However, we have compared the performance of our models with several related studies using evaluation metrics such as MAE, MSE, MAPE, RMSE, and R-squared, as shown in Table 4.

3.4. Contribution of base learners in MLs-stacked-extra trees

Fig. 9 illustrates the importance of different machine learning model predictions as inputs to the MLs-Stacked-Extra Trees model. In this ensemble approach, the predictions of individual models, including XGBoost, AdaBoost, CatBoost, and Extra Trees, are leveraged to enhance overall predictive accuracy. The percentage importance of each model's predictions in the stacking process is as follows: XGBoost at 34 %, Extra Trees at 25 %, CatBoost at 23 %, and AdaBoost at 18 %. These values demonstrate the contribution of each model to the final predictions of the ensemble. The performance scores highlight a notable improvement in accuracy due to the combination of these models, underscoring the effectiveness of the stacking technique in boosting predictive performance by capitalizing on the strengths of each individual model [58,59,91].

3.5. Contribution of input features in the first-level machine learning models

The contribution of input features in the first-level machine learning models is essential for understanding how individual features influence the model's predictions and overall performance [51,59,67]. In a stacked ensemble model, the first-level models are responsible for transforming raw input data into intermediate predictions, which are then used as inputs for the meta-model [93–95]—in this case, the Extra Trees model in the MLs-Stacked-Extra Trees ensemble. For each base model (XGBoost, AdaBoost, CatBoost, and Extra Trees), the contribution of input features varies based on the model's specific algorithm and training procedure (Fig. 10). For example, in decision tree-based models like Extra Trees, feature importance is determined by how much each feature helps reduce impurity at each node in the decision tree [66,67]. In gradient boosting methods like XGBoost, feature importance is assessed by the average reduction in the loss function across all trees that utilize a given feature [59,62,67].

By analyzing feature importance in each of these first-level models, we can gain deeper insights into the variables influencing the model's predictions [91,94]. Feature importance reflects the contribution of each feature to the model's output, offering valuable information about the factors the model deems most relevant when making predictions [59,62,67]. For instance, if PM_{10}, $PM_{2.5}$, and temperature consistently emerge as the most important features across all models (Fig. 10), it suggests that these particulate matter concentrations, along with

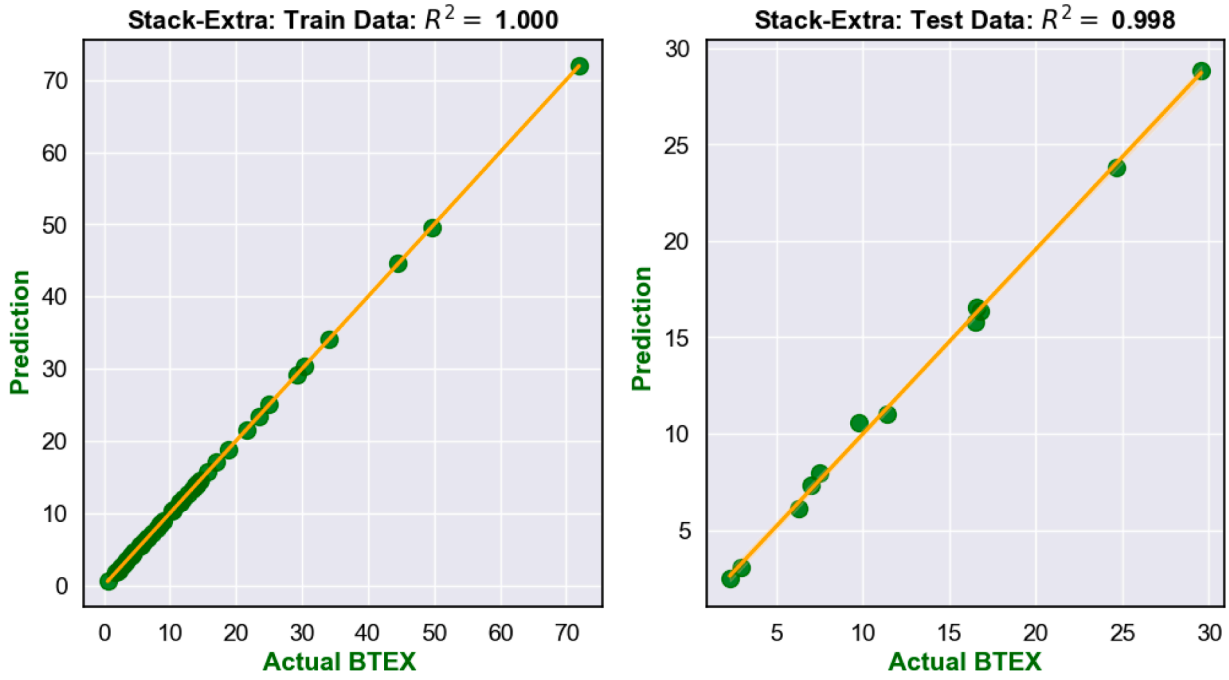

Fig. 8. Performance evaluation of the stacked ensemble model: Actual vs. predicted BTEX concentrations for training and test datasets. The brown line indicates perfect prediction (y = x).

Table 4
Comparison of the proposed model's results with those reported in related studies from the literature.

Parameters		Models	Train					Test					Ref
			R^2	MSE	RMSE	MAE	MAPE	R^2	MSE	RMSE	MAE	MAPE	
$PM_{2.5}$		ST-BPNN	0.78	-	0.0071	0.0041	-	0.78	-	0.0072	0.0041	-	[92]
		ST-KNN	0.94	-	0.0038	0.0081	-	0.85	-	0.0059	0.0030	-	
		ST-XGBOOST	0.92	-	0.0041	0.0026	-	0.87	-	0.0054	0.0031	-	
		ST-Stacking1	0.87	-	0.0054	0.0030	-	0.89	-	0.0051	0.0028	-	
		ST-Stacking2	0.88	-	0.0053	0.0031	-	0.88	-	0.0054	0.0031	-	
		ST-Stacking3	0.90	-	0.0049	0.0027	-	0.90	-	0.0047	0.0027	-	
		ST-Stacking	0.91	-	0.0046	0.0025	-	0.91	-	0.0044	0.0024	-	
NO_2	1-hr	Ensemble	0.91	-	7.23	4.52	-	0.90	-	6.10	3.77	-	[54]
	3-hr	Ensemble	0.86	-	8.69	5.77	-	0.85	-	7.51	4.92	-	
	24-hr	Ensemble	0.84	-	7.55	5.43	-	0.84	-	7.38	5.27	-	
$PM_{2.5}$		LASSO	0.87	-	22.68	-	22.68		-	28.37	-	16.67	[59]
		Adaboost	0.91	-	19.21	-	19.07		-	34	-	21.58	
		XGBoost	0.90	-	20.444	-	12.75		-	26.81	-	16.96	
		GA-MLP	0.88	-	22.04	-	28.50		-	25.45	-	17.81	
		SVR	0.91	-	19.29	-	25.96		-	27.95	-	18.92	
		Ensemble	0.90	-	20.72	-	30.06		-	23.69	-	14.43	
BTEX		XGBoost	0.999	0.116	0.340	0.163	1.66	0.927	2.4	1.54	2.40	32.78	This Study
		CatBoost	0.998	0.532	0.729	0.570	11.20	0.92	3.0	1.74	3.04	36.75	
		AdaBoost	0.991	1.66	1.28	0.833	15.55	0.894	3.2	1.8	3.262	35.21	
		Extra Trees	1	~ 0	~ 0	~ 0	~ 0	0.894	3.0	1.73	3.013	42.57	
		MLs-Stacked-Extra Trees	1	~ 0	~ 0	~ 0	~ 0	0.998	0.42	0.64	0.420	3.89	

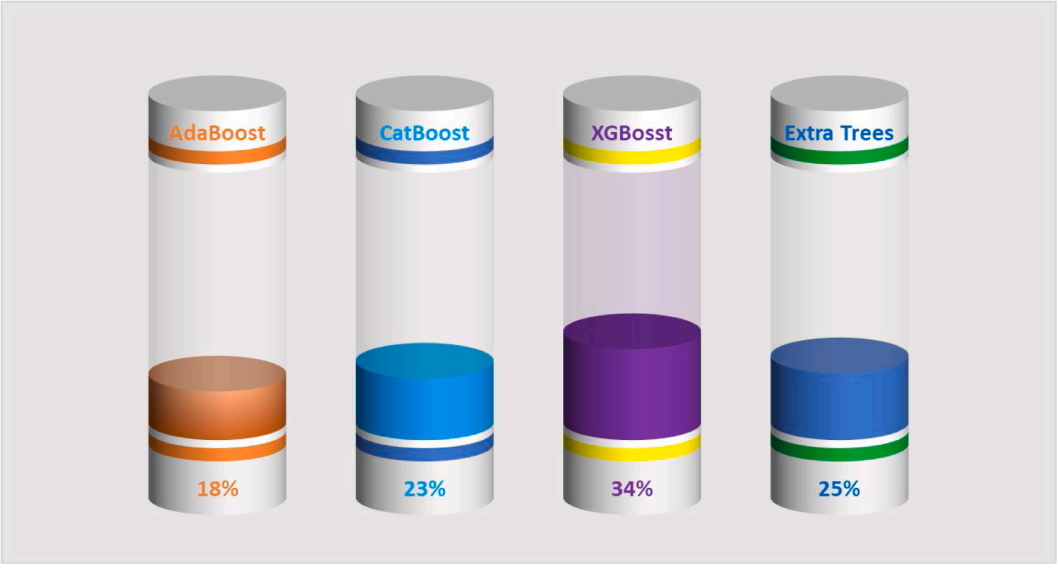

Fig. 9. Contribution of individual model predictions (XGBoost, AdaBoost, CatBoost, Extra Trees) to the MLs-Stacked-Extra Trees model and their impact on predictive accuracy.

temperature, are strongly associated with BTEX levels and are key drivers behind the model's predictions. This trend is observed consistently across all models, highlighting the central role of PM_{10}, $PM_{2.5}$, and temperature in determining BTEX concentrations. In contrast, wind speed and UV index are consistently ranked as the least important features across all models, indicating their relatively limited influence on BTEX concentration compared to the other features. Furthermore, humidity is generally ranked second to last in importance, reinforcing the diminished significance of environmental factors such as wind speed, UV index, and humidity in predicting BTEX levels. The analysis of feature importance also reveals varying degrees of relevance for certain features across different models. This variability suggests that the models rely on different feature sets and learning patterns, which can impact their predictions. Understanding these relationships is crucial for selecting the most relevant features and optimizing the models for more accurate and interpretable predictions [59,66]. To enhance the interpretation of feature impacts on BTEX concentrations, we assessed the explainability of the XGBoost model—identified as the best-performing individual model—using SHAP values [96] (Fig. 11). The analysis identified $PM_{2.5}$ and PM_{10} as the dominant predictors of elevated BTEX levels, aligning with their shared emission sources (e.g., traffic, industrial combustion) [97–99]. Lower temperatures (in the winter season) further amplified BTEX concentrations, likely due to temperature inversions and atmospheric stability during the sampling period [10,100]. Humidity also has an approximately positive effect on increasing BTEX concentration, as the highest humidity and benzene concentration occur in the cold seasons due to reduced temperature and atmospheric stability [100]. In contrast, the UV index and wind speed showed negligible effects on BTEX concentrations—a finding likely attributable to the stable atmospheric conditions during the study period, which limited their typical roles in photochemical degradation and pollutant dispersion. A study conducted in Ahvaz, Iran, reported lower BTEX levels during summer, attributed to increased solar radiation and enhanced photochemical reactions [101]. However, in our study, the UV index

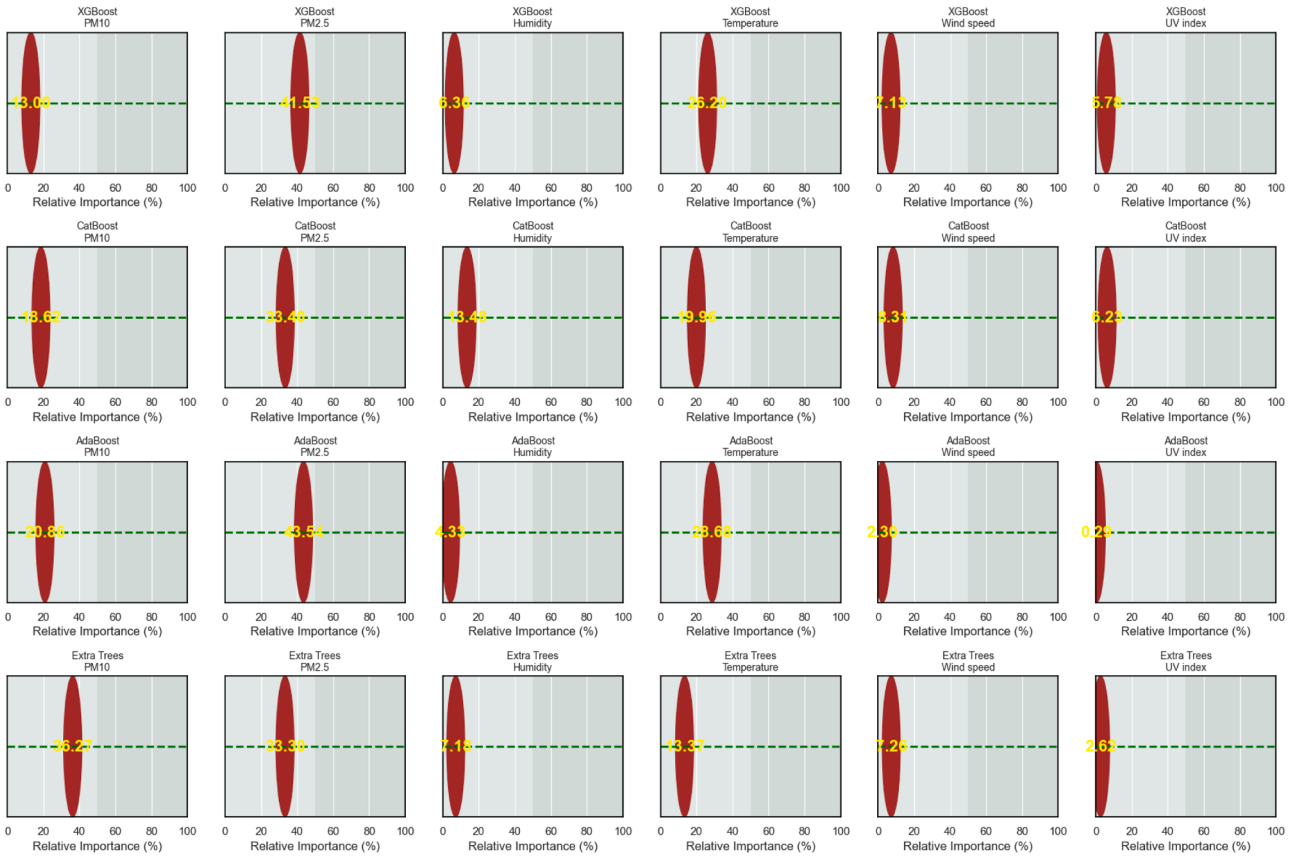

Fig. 10. Contribution of input features in the first-level machine learning models developed.

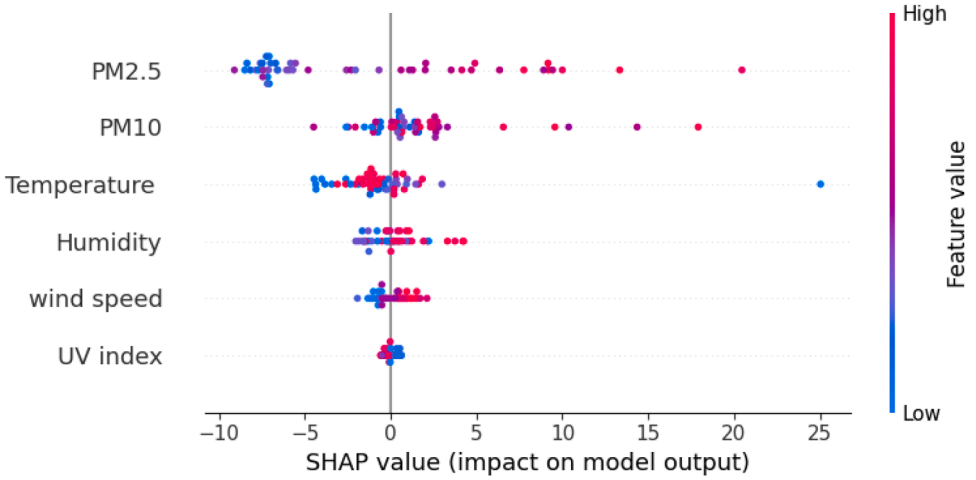

Fig. 11. SHAP value analysis of the XGBoost model for BTEX concentration prediction, highlighting $PM_{2.5}$, PM_{10}, and temperature as dominant drivers, with negligible effects from wind speed and UV index under stable air condition.

had a minimal impact on BTEX reduction, which can be explained by geographical differences. Ahvaz, located in the southernmost part of Iran near the Persian Gulf, experiences significantly higher temperatures than Maragheh, a city with a mountainous climate in the country's northwest. These contrasting conditions suggest that, in our study area, BTEX variability is primarily driven by particulate emissions and temperature, while meteorological factors such as wind and UV radiation play a secondary role under stagnant air conditions.

4. Conclusion

In conclusion, this study effectively demonstrates the capability of machine learning models, particularly the MLs-Stacked-Extra Trees ensemble, in predicting BTEX concentrations. By utilizing a diverse set of models, including XGBoost, CatBoost, AdaBoost, Extra Trees, and the stacked ensemble approach, we were able to capture complex relationships within the data, significantly enhancing predictive accuracy. Through a two-part feature importance analysis, we gained valuable insights into the contributions of individual predictors, such as PM_{10}, $PM_{2.5}$, humidity, temperature, wind speed, and UV index, while also

highlighting the importance of model predictions within the stacked ensemble. Our approach not only outperforms individual models but also offers a comprehensive framework for addressing similar environmental prediction tasks. The iterative hyperparameter tuning process ensured that each model was optimized for peak performance, bolstering the robustness of the final predictions. These findings underscore the potential of advanced ensemble learning techniques in tackling complex environmental monitoring challenges, while emphasizing the importance of careful model selection and feature analysis to optimize predictive accuracy. Our machine learning models for BTEX prediction present valuable tools for informing environmental policy and protecting public health. By accurately forecasting pollution hotspots and peak exposure periods, these models can support targeted air quality regulations, optimize the placement of monitoring stations, and guide urban planning efforts aimed at reducing community health risks. These findings are especially important for safeguarding vulnerable populations residing near industrial zones or high-traffic roadways, where chronic BTEX exposure is associated with elevated risks of cancer and respiratory diseases.

However, this study has several limitations that warrant consideration. First, financial and time constraints limited the number of sampling stations and the duration of data collection, restricting broader temporal and spatial analysis. Second, the performance of the machine learning models is inherently dependent on the quality and availability of the input data, which may affect the robustness of the results. Additionally, budgetary limitations constrained further data collection and model refinement, potentially affecting the generalizability and performance of the conclusions.

Despite these limitations, the study offers practical pathways for advancing predictive modeling in environmental health. Future research should consider: (1) integrating higher-resolution environmental and socioeconomic datasets to improve model accuracy, alongside the use of feature selection techniques such as Recursive Feature Elimination (RFE) or regularization methods like Lasso to prevent overfitting; (2) developing hybrid models that combine physical and statistical approaches to enhance both performance and interpretability; and (3) conducting rigorous cross-regional validations to ensure the applicability of findings across diverse geographic settings.

The models developed here show strong potential for integration into real-time monitoring systems. However, their implementation should strike a balance between predictive accuracy, computational efficiency, and interpretability. Future efforts should also focus on establishing guidelines for selecting context-appropriate models and exploring variable interactions through advanced techniques to produce more robust and actionable environmental insights.

CRediT authorship contribution statement

Mansour Baziar: Writing – original draft, Software, Methodology, Formal analysis. **Negar Jafari:** Writing – original draft, Resources, Methodology, Investigation. **Ali Oghazyan:** Writing – original draft, Methodology. **Amir Mohammadi:** Writing – original draft, Visualization, Methodology. **Ali Abdolahnejad:** Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Ali Behnami:** Writing – review & editing, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Ali Abdolahnejad reports financial support was provided by Maragheh University of Medical Sciences. Ali Abdolahnejad reports a relationship with Maragheh University of Medical Sciences that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to express their sincere appreciation for the financial support of Maragheh University of Medical Sciences for this research under grant number of A-10-1328-1.

Data availability

Data will be made available on request.

References

[1] J. Awewomom, F. Dzeble, Y.D. Takyi, W.B. Ashie, E.N.Y.O. Ettey, P.E. Afua, O. Akoto, Addressing global environmental pollution using environmental control techniques: a focus on environmental policy and preventive environmental management, Discov. Environ. 2 (1) (2024) 8.
[2] B. Halder, I. Ahmadianfar, S. Heddam, Z.H. Mussa, L. Goliatt, M.L. Tan, A. H. Jawad, Machine learning-based country-level annual air pollutants exploration using Sentinel-5P and Google Earth Engine, Sci. Rep. 13 (1) (2023) 7968.
[3] G. Venkatraman, N. Giribabu, P.S. Mohan, B. Muttiah, V. Govindarajan, M. Alagiri, S.A. Karsani, Environmental impact and human health effects of polycyclic aromatic hydrocarbons and remedial strategies: a detailed review, Chemosphere (2024) 141227.
[4] A.M. Parenteau, S. Hang, J.R. Swartz, A.S. Wexler, C.E. Hostinar, Clearing the air: A systematic review of studies on air pollution and childhood brain outcomes to mobilize policy change, Dev. Cogn. Neurosci. (2024) 101436.
[5] A. Zalel, D.M. Broday, Revealing source signatures in ambient BTEX concentrations, Environ. Pollut. 156 (2) (2008) 553–562.
[6] C.-T. Chang, B.-Y. Chen, Toxicity assessment of volatile organic compounds and polycyclic aromatic hydrocarbons in motorcycle exhaust, J. Hazard. Mater. 153 (3) (2008) 1262–1269.
[7] J.M.M. Mello, H.L. Brandão, A. Valério, A.A.U. de Souza, D. de Oliveira, A. da Silva, Biodegradation of BTEX compounds from petrochemical wastewater: kinetic and toxicity, J. Water. Process. Eng. 32 (2019) 100914.
[8] A.H. Khoshakhlagh, S. Yazdanirad, A. Ducatman, Climatic conditions and concentrations of BTEX compounds in atmospheric media, Environ. Res. (2024) 118553.
[9] I. Muda, M.J. Mohammadi, A. Sepahvad, A. Farhadi, R. Fadhel Obaid, M. Taherian, M. Farhadi, Associated health risk assessment due to exposure to BTEX compounds in fuel station workers, Rev. Environ. Health 39 (3) (2024) 435–446.
[10] A. Behnami, N. Jafari, K.Z. Benis, F. Fanaei, A. Abdolahnejad, Spatio-temporal variations, ozone and secondary organic aerosol formation potential, and health risk assessment of BTEX compounds in east of Azerbaijan Province, Iran, Urban. Clim. 47 (2023) 101360.
[11] USEPA, Hazardous Air Pollutants, USEPA, 2019. Editor.
[12] L.A.M. García, F.S. Lasheras, P.J.G. Nieto, L.Á. de Prado, A.B. Sánchez, Predicting benzene concentration using machine learning and time series algorithms, Mathematics 8 (12) (2020) 1–22.
[13] V. Yadav, A.K. Yadav, V. Singh, T. Singh, Artificial neural network an innovative approach in air pollutant prediction for environmental applications: a review, Res. Eng. 22 (2024) 102305.
[14] I. Gryech, C. Asaad, M. Ghogho, A. Kobbane, Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: a systematic literature review, Eng. Appl. Artif. Intell. 137 (2024) 109182.
[15] R. Yan, J. Liao, J. Yang, W. Sun, M. Nong, F. Li, Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering, Expert Syst. Appl. 169 (2021) 114513.
[16] F. Mohammadi, H. Teiri, Y. Hajizadeh, A. Abdolahnejad, A. Ebrahimi, Prediction of atmospheric PM2. 5 level by machine learning techniques in Isfahan, Iran, Sci. Rep. 14 (1) (2024) 2109.
[17] G. Suthar, N. Kaul, S. Khandelwal, S. Singh, Predicting land surface temperature and examining its relationship with air pollution and urban parameters in Bengaluru: a machine learning approach, Urban. Clim. 53 (2024) 101830.
[18] A.-L. Balogun, A. Tella, L. Baloo, N. Adebisi, 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 (2021) 100989.
[19] D.B. Olawade, O.Z. Wada, A.O. Ige, B.I. Egbewole, A. Olojo, B.I. Oladapo, Artificial intelligence in environmental monitoring: advancements, challenges, and future directions, Hygiene Environ. Health Adv. 12 (2024) 100114.
[20] Z. Peng, B. Zhang, D. Wang, X. Niu, J. Sun, H. Xu, Z. Shen, Application of machine learning in atmospheric pollution research: a state-of-art review, Sci. Total Environ. 910 (2024) 168588.
[21] L. Zhao, Z. Li, L. Qu, A novel machine learning-based artificial intelligence method for predicting the air pollution index PM2. 5, J. Clean. Prod. 468 (2024) 143042.

[22] M. Baziar, A. Behnami, N. Jafari, A. Mohammadi, A. Abdolahnejad, Machine learning-based Monte Carlo hyperparameter optimization for THMs prediction in urban water distribution networks, J. Water. Process. Eng. 73 (2025) 107683.

[23] M. Baziar, M. Yousefi, V. Oskoei, A. Makhdoomi, R. Abdollahzadeh, A. Dehghan, Machine learning-based prediction of heating values in municipal solid waste, Sci. Rep. 15 (1) (2025) 14589.

[24] S. Chadalavada, O. Faust, M. Salvi, S. Seoni, N. Raj, U. Raghavendra, R. Acharya, Application of artificial intelligence in air pollution monitoring and forecasting: a systematic review, Environ. Model. Softw. 185 (2025) 106312.

[25] A. Masood, K. Ahmad, A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: fundamentals, application and performance, J. Clean. Prod. 322 (2021) 129072.

[26] A.N. Al–Dabbous, P. Kumar, A.R. Khan, Prediction of airborne nanoparticles at roadside location using a feed–forward artificial neural network, Atmos. Pollut. Res. 8 (3) (2017) 446–454.

[27] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y.Tahmasebi Birgani, M. Rahmati, Air pollution prediction by using an artificial neural network model, Clean. Technol. Environ. Policy. 21 (6) (2019) 1341–1352.

[28] K. de Hoogh, H. Héritier, M. Stafoggia, N. Künzli, I. Kloog, Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland, Environ. Pollut. 233 (2018) 1147–1154.

[29] N.M. Eldakhly, M. Aboul-Ela, A. Abdalla, A novel approach of weighted support vector machine with applied chance theory for forecasting air pollution phenomenon in Egypt, Int. J. Comput. Intell. Appl. 17 (01) (2018) 1850001.

[30] A. Samadi-Koucheksaraee, S. Shirvani-Hosseini, I. Ahmadianfar, B. Gharabaghi, Optimization algorithms surpassing metaphor, in Computational intelligence for water and environmental sciences, Springer, 2022, pp. 3–33.

[31] M. Hashemitaheri, E. Ebrahimi, G. de Silva, H. Attariani, Optical sensor for BTEX detection: integrating machine learning for enhanced sensing, Adv. Sens. Energy Mater. 3 (3) (2024) 100114.

[32] F. Karimi, J. Amanollahi, M. Reisi, M. Darand, Prediction of air quality using vertical atmospheric condition and developing hybrid models, Adv. Space Res. 72 (4) (2023) 1172–1182.

[33] S.R. Shams, S. Kalantary, A. Jahani, S.M. Parsa Shams, B. Kalantari, D. Singh, Y. Choi, Assessing the effectiveness of artificial neural networks (ANN) and multiple linear regressions (MLR) in forcasting AQI and PM10 and evaluating health impacts through AirQ+ (case study: Tehran), Environ. Pollut. 338 (2023) 122623.

[34] H. Tao, A.O. Al-Sulttani, M.A. Saad, I. Ahmadianfar, L. Goliatt, S.S.U.H. Kazmi, Z. M. Yaseen, Optimized ensemble deep random vector functional link with nature inspired algorithm and boruta feature selection: multi-site intelligent model for air quality index forecasting, Process Saf. Environ. Protect. 191 (2024) 1737–1760.

[35] S. Shirvani-Hosseini, A. Samadi-Koucheksaraee, I. Ahmadianfar, B. Gharabaghi, Data mining methods for modeling in water science, in Computational intelligence for water and environmental sciences, Springer, 2022, pp. 157–178.

[36] A. Analitis, B. Barratt, D. Green, A. Beddows, E. Samoli, J. Schwartz, K. Katsouyanni, Prediction of PM2.5 concentrations at the locations of monitoring sites measuring PM10 and NOx, using generalized additive models and machine learning methods: a case study in London, Atmos. Environ. 240 (2020) 117757.

[37] S. Araki, M. Shima, K. Yamamoto, Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan, Sci. Total Environ. 634 (2018) 1269–1277.

[38] W. Ding, X. Qie, Prediction of air pollutant concentrations via RANDOM forest regressor coupled with uncertainty analysis-A case study in Ningxia, Atmosphere 13 (6) (2022) 960.

[39] C. Silibello, G. Carlino, M. Stafoggia, C. Gariazzo, S. Finardi, N. Pepe, G. Viegi, Spatial-temporal prediction of ambient nitrogen dioxide and ozone levels over Italy using a Random Forest model for population exposure assessment, Air Qual., Atmos. Health 14 (2021) 817–829.

[40] Y. Gao, Z. Wang, C.-Y. Li, T. Zheng, Z.-R. Peng, Assessing neighborhood variations in ozone and PM2. 5 concentrations using decision tree method, Build. Environ. 188 (2021) 107479.

[41] W.N. Shaziayani, A.Z. Ul-Saufie, S. Mutalib, N.Mohamad Noor, N.S. Zainordin, Classification prediction of PM10 concentration using a tree-based machine learning approach, Atmosphere (Basel) 13 (4) (2022) 538.

[42] L. Mampitiya, N. Rathnayake, Y. Hoshino, U. Rathnayake, Forecasting PM10 levels in Sri Lanka: a comparative analysis of machine learning models PM10, J. Hazard. Mater. Adv. 13 (2024) 100395.

[43] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, C. Sonne, Air quality prediction by machine learning models: a predictive study on the Indian coastal city of Visakhapatnam, Chemosphere 338 (2023) 139518.

[44] G. Ravindiran, S. Rajamanickam, K. Kanagarathinam, G. Hayder, G. Janardhan, P. Arunkumar, S.K. Muniasamy, Impact of air pollutants on climate change and prediction of air quality index using machine learning models, Environ. Res. 239 (2023) 117354.

[45] Z. Wang, X. Wu, Y. Wu, A spatiotemporal XGBoost model for PM2. 5 concentration prediction and its application in Shanghai, Heliyon. 9 (2023), 12.

[46] Z. Li, K. Gan, S. Sun, S. Wang, A new PM2. 5 concentration forecasting system based on AdaBoost-ensemble system with deep learning approach, J. Forecast. 42 (1) (2023) 154–175.

[47] D. Thamizhselvi, B. Kasi, K. Kamalakkannan, S. Bharath, S. Gowtham, M. Kishore, Air quality prediction using adaboost. 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), IEEE, 2023.

[48] J. Wang, D. Wang, F. Zhang, C. Yoo, H. Liu, Soft sensor for predicting indoor PM2. 5 concentration in subway with adaptive boosting deep learning model, J. Hazard. Mater. 465 (2024) 133074.

[49] B. Zhang, Z. Wang, Y. Lu, M.-Z. Li, R. Yang, J. Pan, Z. Kou, Air pollutant diffusion trend prediction based on deep learning for targeted season—North China as an example, Expert Syst Appl 232 (2023) 120718.

[50] Z. Wu, Y. Tian, M. Li, B. Wang, Y. Quan, J. Liu, Prediction of air pollutant concentrations based on the long short-term memory neural network, J. Hazard. Mater. 465 (2024) 133099.

[51] S. Karimi, M. Asghari, R. Rabie, M.E. Niri, Machine learning-based white-box prediction and correlation analysis of air pollutants in proximity to industrial zones, Process Saf. Environ. Protect. 178 (2023) 1009–1025.

[52] D. Wang, S. Wei, H. Luo, C. Yue, O. Grunder, A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine, Sci. Total Environ. 580 (2017) 719–733.

[53] B. Wu, C. Wu, Y. Ye, C. Pei, T. Deng, Y.J. Li, D. Wu, Long-term hourly air quality data bridging of neighboring sites using automated machine learning: a case study in the Greater Bay area of China, Atmos. Environ. 321 (2024) 120347.

[54] T. Peng, J. Xiong, K. Sun, S. Qian, Z. Tao, M.S. Nazir, C. Zhang, Research and application of a novel selective stacking ensemble model based on error compensation and parameter optimization for AQI prediction, Environ. Res. 247 (2024) 118176.

[55] C.-Y. Hsu, Y.-T. Zeng, Y.-C. Chen, M.-J. Chen, S.-C.C. Lung, C.-D. Wu, Kriging-based land-use regression models that use machine learning algorithms to estimate the monthly BTEX concentration, Int. J. Environ. Res. Public Health 17 (19) (2020) 6956.

[56] E.T. Al-Shammari, Hybrid model for benzene prediction in Kuwait's industrial regions, Int. J. Appl. Geospatial Res. (IJAGR) 15 (1) (2024) 1–23.

[57] Z. Ning, S. Gao, Z. Gu, C. Ni, F. Fang, Y. Nie, C. Wang, Prediction and explanation for ozone variability using cross-stacked ensemble learning model, Sci. Total Environ. 935 (2024) 173382.

[58] Y. Xie, W. Sun, M. Ren, S. Chen, Z. Huang, X. Pan, Stacking ensemble learning models for daily runoff prediction using 1D and 2D CNNs, Expert Syst. Appl. 217 (2023) 119469.

[59] B. Zhai, J. Chen, Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China, Sci. Total Environ. 635 (2018) 644–658.

[60] N. Jafari, A. Behnami, F. Ghayurdoost, A. Solimani, A. Mohammadi, M. Pourakbar, A. Abdolahnejad, Analysis of THM formation potential in drinking water networks: effects of network age, health risks, and seasonal variations in northwest of Iran, Heliyon. 10 (2024), 14.

[61] NIOSH, *Method 1501: Niosh Manual of Analytical methods (Mam)*. 2003: USA.

[62] D.H. Djarum, Z. Ahmad, J. Zhang, Reduced Bayesian optimized stacked regressor (RBOSR): a highly efficient stacked approach for improved air pollution prediction, Appl. Soft Comput. 144 (2023) 110466.

[63] C.-H. Yang, C.-H. Wu, K.-H. Luo, H.-C. Chang, S.-C. Wu, H.-Y. Chuang, Use of machine learning algorithms to determine the relationship between air pollution and cognitive impairment in Taiwan, Ecotoxicol. Environ. Saf. 284 (2024) 116885.

[64] A. Abdellatif, H. Mubarak, S. Ahmad, T. Ahmed, G. Shafiullah, A. Hammoudeh, H.M. Gheni, Forecasting photovoltaic power generation with a stacking ensemble model, Sustainability. 14 (17) (2022) 11083.

[65] H. Liu, C. Chen, Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: a case study in China, J. Clean. Prod. 265 (2020) 121777.

[66] J. Kerckhoffs, G. Hoek, L.T. Portengen, B. Brunekreef, R.C. Vermeulen, Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces, Environ. Sci. Technol. 53 (3) (2019) 1413–1421.

[67] K. Ravindra, S.S. Bahadur, V. Katoch, S. Bhardwaj, M. Kaur-Sidhu, M. Gupta, S. Mor, Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections, Sci. Total Environ. 858 (2023) 159509.

[68] F. Lautenschlager, M. Becker, K. Kobs, M. Steininger, P. Davidson, A. Krause, A. Hotho, OpenLUR: off-the-shelf air pollution modeling with open features and machine learning, Atmos. Environ. 233 (2020) 117535.

[69] J.-J. Zhu, M. Yang, Z.J. Ren, Machine learning in environmental research: common pitfalls and best practices, Environ. Sci. Technol. 57 (46) (2023) 17671–17689.

[70] A.A.M. Ahmed, S.J.J. Jui, E. Sharma, M.H. Ahmed, N. Raj, A. Bose, An advanced deep learning predictive model for air quality index forecasting with remote satellite-derived hydro-climatological variables, Sci. Total Environ. 906 (2024) 167234.

[71] D. Shakya, V. Deshpande, M.K. Goyal, M. Agarwal, PM2. 5 air pollution prediction through deep learning using meteorological, vehicular, and emission data: a case study of New Delhi, India, J. Clean. Prod. 427 (2023) 139278.

[72] A.K. Rad, S.-O. Razmi, M.J. Nematollahi, A. Naghipour, F. Golkar, M. Mahmoudi, Machine learning models for predicting interactions between air pollutants in Tehran Megacity, Iran, Alex. Eng. J. 104 (2024) 464–479.

[73] M.J. Jiménez-Navarro, M. Martínez-Ballesteros, F. Martínez-Álvarez, G. Asencio-Cortés, Explaining deep learning models for ozone pollution prediction via embedded feature selection, Appl. Soft Comput. 157 (2024) 111504.

[74] S. Masmoudi, H. Elghazel, D. Taieb, O. Yazar, A. Kallel, A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection, Sci. Total Environ. 715 (2020) 136991.

[75] S.V. Razavi-Termeh, A. Sadeghi-Niaraki, S.-M. Choi, Effects of air pollution in spatio-temporal modeling of asthma-prone areas using a machine learning model, Environ. Res. 200 (2021) 111344.
[76] H. Zhang, R. Srinivasan, X. Yang, S. Ahrentzen, E.S. Coker, A. Alwisy, Factors influencing indoor air pollution in buildings using PCA-LMBP neural network: a case study of a university campus, Build. Environ. 225 (2022) 109643.
[77] F. Abbasi, H. Pasalari, J.M. Delgado-Saborit, A. Rafiee, A. Abbasi, M. Hoseini, Characterization and risk assessment of BTEX in ambient air of a Middle Eastern City, Process Saf. Environ. Protect. 139 (2020) 98–105.
[78] Y. Hajizadeh, M. Mokhtari, M. Faraji, A. Mohammadi, S. Nemati, R. Ghanbari, M. Miri, Trends of BTEX in the central urban area of Iran: a preliminary study of photochemical ozone pollution and health risk assessment, Atmos. Pollut. Res. 9 (2) (2018) 220–229.
[79] M.T. Latif, H.H. Abd Hamid, F. Ahamad, M.F. Khan, M.S. Mohd Nadzir, M. Othman, N.M. Tahir, BTEX compositions and its potential health impacts in Malaysia, Chemosphere 237 (2019) 124451.
[80] M. Miri, M. Rostami Aghdam Shendi, H.R. Ghaffari, H. Ebrahimi Aval, E. Ahmadi, E. Taban, A. Azari, Investigation of outdoor BTEX: concentration, variations, sources, spatial distribution, and risk assessment, Chemosphere 163 (2016) 601–609.
[81] J.G. Cerón Bretón, R.M. Cerón Bretón, S. Martínez Morales, J.D. Kahl, C. Guarnaccia, R.D.C.L. Severino, M.P. Uc Chi, Health risk assessment of the levels of BTEX in ambient air of one urban site located in Leon, Guanajuato, Mexico during two climatic seasons, Atmosphere (Basel) 11 (2) (2020) 165.
[82] A.H. Khoshakhlagh, S. Yazdanirad, A. Ducatman, Climatic conditions and concentrations of BTEX compounds in atmospheric media, Environ. Res. 251 (2024) 118553.
[83] K.M. Mullaugh, J.M. Hamilton, G.B. Avery, J.D. Felix, R.N. Mead, J.D. Willey, R. J. Kieber, Temporal and spatial variability of trace volatile organic compounds in rainwater, Chemosphere 134 (2015) 203–209.
[84] M. Kermani, A. Jonidi Jafari, M. Gholami, F. Taghizadeh, K. Masroor, A. Abdolahnejad, F. Fanaei, Characterisation of PM2. 5–bound PAHs in outdoor air of Karaj megacity: the effect of meteorological factors, Int. J. Environ. Anal. Chem. 103 (14) (2023) 3290–3308.
[85] R. Maleki, Z. Asadgol, M. Kermani, A. Jonidi Jafari, H. Arfaeinia, M. Gholami, Monitoring BTEX compounds and asbestos fibers in the ambient air of Tehran, Iran: seasonal variations, spatial distribution, potential sources, and risk assessment, Int. J. Environ. Anal. Chem. 102 (16) (2022) 4220–4237.
[86] Y. Zhang, Y. Mu, J. Liu, A. Mellouki, Levels, sources and health risks of carbonyls and BTEX in the ambient air of Beijing, China, J. Environ. Sci. 24 (1) (2012) 124–130.
[87] A. Masih, A.S. Lall, A. Taneja, R. Singhvi, Inhalation exposure and related health risks of BTEX in ambient air at different microenvironments of a terai zone in north India, Atmos. Environ. 147 (2016) 55–66.
[88] E. Kalantari, H. Gholami, H. Malakooti, M. Eftekhari, P. Saneei, D. Esfandiarpour, A.R. Nafarzadegan, Evaluating traditional versus ensemble machine learning methods for predicting missing data of daily PM10 concentration, Atmos. Pollut. Res. 15 (5) (2024) 102063.
[89] S.A. Sai, S.N. Venkatesh, S. Dhanasekaran, P.A. Balaji, V. Sugumaran, N. Lakshmaiya, P. Paramasivam, Transfer learning based fault detection for suspension system using vibrational analysis and radar plots, Machines 11 (8) (2023) 778.
[90] J. Zhou, W. Huang, F. Chen, Facilitating machine learning model comparison and explanation through a radial visualisation, Energies. (Basel) 14 (21) (2021) 7049.
[91] M. Lu, Q. Hou, S. Qin, L. Zhou, D. Hua, X. Wang, L. Cheng, A stacking ensemble model of various machine learning models for daily runoff forecasting, Water (Basel) 15 (7) (2023) 1265.
[92] L. Feng, Y. Li, Y. Wang, Q. Du, Estimating hourly and continuous ground-level PM2. 5 concentrations using an ensemble learning algorithm: the ST-stacking model, Atmos. Environ. 223 (2020) 117242.
[93] M.A. Anjum and A. Alanzi, Smart Urban planning: an intelligent framework to predict traffic using stack ensembling approach, (2024).
[94] N.U. Khan, M.A. Shah, C. Maple, E. Ahmed, N. Asghar, Traffic flow prediction: an intelligent scheme for forecasting traffic flow using air pollution data in smart cities with bagging ensemble, Sustainability. 14 (7) (2022) 4164.
[95] W. Yu, S. Li, T. Ye, R. Xu, J. Song, Y. Guo, Deep ensemble machine learning framework for the estimation of PM 2.5 concentrations, Environ. Health Perspect. 130 (3) (2022) 037004.
[96] A. Dehghan, V. Oskoei, T. Khajavi, M. Baziar, M. Yousefi, Machine learning-based prediction of the C/N ratio in municipal organic waste, Environ. Technol. Innov. 37 (2025) 103977.
[97] N. Kanjanasiranont, Assessment of BTEX, PM10, and PM2. 5 concentrations in Nakhon Pathom, Thailand, and the health risks for security guards and copy shop employees, Atmosphere (Basel) 16 (2) (2025) 212.
[98] M. Kermani, H. Arfaeinia, K. Masroor, A. Abdolahnejad, F. Fanaei, A. Shahsavani, M.H. Vahidi, Health impacts and burden of disease attributed to long-term exposure to atmospheric PM10/PM2. 5 in Karaj, Iran: effect of meteorological factors, Int. J. Environ. Anal. Chem. 102 (18) (2022) 6134–6150.
[99] S.Z. Sajani, S. Marchesi, A. Trentini, D. Bacco, C. Zigola, S. Rovelli, D.M. Cavallo, Vertical variation of PM2. 5 mass and chemical composition, particle size distribution, NO2, and BTEX at a high rise building, Environ. Pollut. 235 (2018) 339–349.
[100] P. Nakhjirgan, F. Fanaei, A. Jonidi Jafari, M. Gholami, A. Shahsavani, M. Kermani, Extensive investigation of seasonal and spatial fluctuations of BTEX in an industrial city with a health risk assessment, Sci. Rep. 14 (1) (2024) 23662.
[101] H.D. Rad, A.A. Babaei, G. Goudarzi, K.A. Angali, Z. Ramezani, M.M. Mohammadi, Levels and sources of BTEX in ambient air of Ahvaz metropolitan city, Air Qual., Atmos. Health 7 (2014) 515–524.