KMAN PUBLISHING INSTITUTE

Article history:
Received 21 June 2025
Revised 08 September 2025
Accepted 14 September 2025
Published online 15 October 2025

Journal of Resource Management and Decision Engineering

Volume 4, Issue 4, pp 1-16

Forecasting the Air Quality Index Using Machine Learning Models, Bayesian Optimization, and the Development of the S-GBR Model Incorporating Seasonal Variables

Mahtab. Mahbodi[1], Babak. Karasfi[2]*

[1] MA Student, Artificial Intelligence and Robotics Department, Qa.C., Islamic Azad University, Qazvin, Iran
[2] Department of Computer Engineering and Information Technology, Qa.C., Islamic Azad University, Qazvin, Iran

*** Corresponding author email address**: karasfi@qiau.ac.ir

Article Info

Article type:
Original Research

How to cite this article:
Mahbodi, M., & Karasfi, B. (2025). Forecasting the Air Quality Index Using Machine Learning Models, Bayesian Optimization, and the Development of the S-GBR Model Incorporating Seasonal Variables. *Journal of Resource Management and Decision Engineering, 4*(4), 1-16.
https://doi.org/10.61838/kman.jrmde.147

ABSTRACT

Air pollution is considered one of the most serious environmental and public health challenges in urban communities, and accurately forecasting the Air Quality Index (AQI) plays a crucial role in mitigating its negative impacts and supporting data-driven decision-making. Given the complexity and nonlinear nature of factors influencing air quality, the use of machine learning methods has attracted widespread attention in recent years. However, a review of previous studies reveals two major shortcomings: first, many models have been implemented based on default hyperparameter values, which has led to reduced accuracy and generalizability; second, temporal and seasonal components have often been overlooked, even though they play a decisive role in variations in air quality. To address these shortcomings, this study proposes a novel framework called the Seasonal Gradient Boosting Regressor (S-GBR). In this model, the Bayesian optimization search method is used for hyperparameter optimization, and the seasonal feature is incorporated as an input to the Gradient Boosting Regressor algorithm. In addition, baseline models such as Random Forest and XGBoost were also simulated and compared to determine the standing of the proposed model. Empirical findings show that the proposed model achieved a coefficient of determination of 0.9686 and significantly reduced errors, performing almost as well as the most accurate baseline model (Random Forest with 0.9796) while outperforming XGBoost. These results demonstrate that combining Bayesian optimization with the inclusion of seasonal components can raise prediction accuracy to the level of rich and complex datasets, even under limited data conditions. Such an achievement highlights the high potential of the proposed model for use in practical air quality monitoring and management.

Keywords: *Air Quality Index forecasting, machine learning, hyperparameter optimization, Bayesian search, seasonal features, S-GBR model.*

1. Introduction

Air pollution has emerged as one of the most critical environmental and public health challenges of the 21st century, particularly in rapidly urbanizing regions where industrial emissions, transportation, and population growth are accelerating at unprecedented rates. The degradation of air quality directly contributes to respiratory and cardiovascular diseases, reduced life expectancy, and significant economic burdens associated with healthcare costs and workforce productivity losses (Shayegan & Makram, 2023). Forecasting and managing air quality, therefore, has become an urgent necessity for policymakers, environmental agencies, and urban planners. In this context, the development of reliable and accurate prediction models for the Air Quality Index (AQI) has attracted increasing attention from researchers worldwide (Gupta et al., 2023; Natarajan et al., 2024).

The AQI is a standardized indicator that aggregates data on key air pollutants—including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3)—into a single composite value, which can be used to communicate air pollution levels to the public and trigger mitigation measures. Traditional statistical methods, such as multiple linear regression, have historically been employed to predict air pollutant concentrations. However, these methods often struggle to capture the nonlinear, multivariate, and spatiotemporally dynamic nature of air pollution processes, resulting in poor predictive accuracy and generalizability (Farhadi et al., 2020; Omidvar et al., 2018). This limitation has accelerated the shift toward machine learning (ML)-based approaches, which are capable of learning complex patterns from high-dimensional data and have shown superior performance in environmental modeling (Aram et al., 2023; Kalantari et al., 2024).

2. Literature Review

In recent years, machine learning models have been increasingly adopted to forecast AQI and pollutant concentrations across diverse geographical contexts. For example, ensemble learning techniques—such as Random Forest and Gradient Boosting—have been recognized for their robustness against overfitting and their ability to model nonlinear interactions among atmospheric variables (Castelli et al., 2020; Ganesh et al., 2021). Studies conducted in Tehran have confirmed that these models outperform traditional regression methods in predicting urban air quality trends (Beheshtifar & Rahimzad, 2018; Karami et al., 2023). Similarly, deep learning architectures such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and hybrid frameworks have achieved remarkable accuracy in spatiotemporal AQI forecasting by leveraging their capability to extract hierarchical features from time series data (Du et al., 2021; Ragab et al., 2020).

Despite these advances, several key challenges remain. One major challenge is the scarcity and inconsistency of environmental monitoring data, especially in developing countries and regions with sparse monitoring infrastructure. Missing data, noise, and imbalanced datasets can severely degrade model performance. Recent work has attempted to address this problem by developing data reconstruction techniques and noise-robust learning frameworks (Just et al., 2020; Xu et al., 2021). Another challenge is the sensitivity of ML models to hyperparameter settings. Default parameter configurations often fail to generalize well across diverse datasets and atmospheric conditions, resulting in overfitting or underfitting (Haq, 2022; Mishra et al., 2020). Therefore, optimization techniques such as Bayesian optimization and evolutionary algorithms have been increasingly integrated into AQI modeling pipelines to fine-tune hyperparameters and enhance predictive accuracy (Natarajan et al., 2024; Wu et al., 2024).

Furthermore, the seasonal and temporal variability of air pollution poses another significant obstacle to accurate prediction. Many conventional ML models overlook the seasonality inherent in atmospheric systems, leading to decreased accuracy during periods of abrupt weather changes or seasonal pollutant accumulation (Haqbian et al., 2023; Sharma et al., 2021). Incorporating seasonal and temporal features into predictive models has been shown to significantly improve performance by enabling models to account for cyclic fluctuations in pollutant levels (Hardini et al., 2023; Zhou et al., 2022). Studies that have explicitly modeled seasonal dynamics, such as those by Liu and colleagues in China, have demonstrated that including temporal features such as month and season reduces forecast errors and improves generalization to unseen data (Liu et al., 2020).

Another dimension of complexity in AQI forecasting lies in the heterogeneity of predictor variables, which can include meteorological factors, emission inventories, satellite imagery, and traffic data. Integrating such diverse data sources requires models that are both flexible and computationally efficient. Gradient boosting-based algorithms—such as Gradient Boosting Regressor (GBR), XGBoost, LightGBM, and CatBoost—have shown considerable promise in this regard, due to their ability to handle high-dimensional heterogeneous inputs and capture nonlinear interactions (Brahmi et al., 2023; Mahesh et al., 2022; Zhang et al., 2020). Several comparative studies have confirmed the competitive performance of these ensemble models in air quality prediction tasks relative to both shallow learners and deep neural networks (Goudarzi et al., 2020; Gupta & Singla, 2023). However, while deep learning models can deliver higher accuracy, they often require massive computational resources and large labeled datasets, which may not be feasible in data-constrained environments (Danesh Yazdi et al., 2020; Kalantari et al., 2024).

Hybrid and ensemble modeling strategies have recently emerged as a promising solution to these issues. By combining the strengths of multiple algorithms, hybrid approaches can enhance prediction accuracy, reduce variance, and improve robustness under limited data conditions (Gupta et al., 2023; Ravindiran et al., 2023). For instance, research by Aram et al. demonstrated that hybrid ML systems integrating boosting algorithms with meteorological data significantly outperformed single-model baselines in predicting both AQI values and categorical air quality grades (Aram et al., 2023). Similar findings have been reported in India, where optimized machine learning frameworks integrating feature selection, ensemble learners, and hyperparameter tuning achieved notable improvements in predictive performance (Kothandaraman et al., 2022; Natarajan et al., 2024). These results collectively suggest that ensemble-based models with systematic optimization represent a viable pathway toward accurate and generalizable AQI forecasting systems.

At the same time, the explainability and interpretability of machine learning models remain critical for their adoption in environmental policymaking. Black-box models may achieve high accuracy but fail to provide insights into the causal drivers of air pollution events, limiting their practical usefulness for regulatory planning (Castelli et al., 2020; Sharma et al., 2021). Consequently, researchers have increasingly incorporated model interpretability tools, such as feature importance analysis and SHAP (SHapley Additive exPlanations) values, into AQI prediction studies to identify key contributing factors and to build trust among stakeholders (Karami et al., 2023; Ravindiran et al., 2023).

In this context, the present study aims to develop a novel AQI prediction framework called the Seasonal Gradient Boosting Regressor (S-GBR), which integrates Bayesian hyperparameter optimization with seasonal feature incorporation to enhance accuracy, generalizability, and computational efficiency. The S-GBR model is designed to overcome three core limitations identified in prior research: (1) the use of default hyperparameter settings that hinder model generalization (Haq, 2022; Mishra et al., 2020), (2) the exclusion of seasonal and temporal features that are crucial for capturing air quality dynamics (Hardini et al., 2023; Liu et al., 2020), and (3) the lack of a streamlined approach that balances predictive performance with resource efficiency (Du et al., 2021; Wu et al., 2024). By incorporating seasonality as a categorical input and using Bayesian optimization to fine-tune model parameters, this framework seeks to deliver near state-of-the-art accuracy while maintaining low computational overhead.

Moreover, the study contributes to the growing literature on sustainable and adaptive air quality management by demonstrating how advanced machine learning techniques can be tailored to perform effectively in data-limited environments—a scenario common in many developing urban centers (Goudarzi et al., 2020; Shayegan & Makram, 2023). The inclusion of both a full dataset (pollutants plus meteorological parameters) and a reduced dataset (pollutants only) further allows the evaluation of the model's robustness under varying data availability conditions. This dual-dataset design addresses the practical constraint that meteorological data are often unavailable or of low quality in many regions (Haqbian et al., 2023; Omidvar et al., 2018).

In summary, as air pollution continues to threaten public health, ecosystems, and economic sustainability, accurate AQI forecasting tools are essential for enabling timely interventions and informed policymaking. While significant progress has been made through the application of machine learning, persistent challenges related to data scarcity, seasonal variability, and hyperparameter sensitivity hinder the reliability of existing models (Gupta & Singla, 2023; Kalantari et al., 2024). The proposed S-GBR model seeks to address these gaps by merging the predictive power of boosting algorithms with the adaptability of Bayesian optimization and the contextual awareness provided by seasonal feature integration.

3. Methods and Materials

The proposed method of this study is presented to overcome the limitations of previous research and to achieve an accurate, generalizable, and low-cost model for forecasting the Air Quality Index (AQI). A review of the literature showed that traditional statistical and machine learning models, despite their simplicity, are weak in terms of accuracy and generalizability, while deep learning models—although more accurate—require massive volumes of data and high computational power. Moreover, ensemble models such as Random Forest and CatBoost have performed well; however, many studies have used default parameter values and have not conducted effective optimization. In addition, most studies have not incorporated temporal and seasonal variables into the modeling, which has reduced forecasting accuracy.

This study is built upon the combination and extension of methods reported in two main reference articles. The first article, by Gaddam and Reddy in Chemosphere (Haq, 2022), used complete data—including pollutants and meteorological parameters—and employed several machine learning algorithms, which served as the basis for deriving the baseline model. The second article, by Gupta et al. in the Journal of Environmental and Public Health (Liu et al., 2020), used simpler data without meteorological variables and demonstrated that AQI can be forecast under data-limited conditions. By integrating the findings of these two studies, this thesis proposes a novel model titled S-GBR (Seasonal Gradient Boosting Regressor), whose aim is to achieve accurate yet simple and generalizable forecasting. The principal innovations of this model are twofold: first, the use of Bayesian optimization to finely tune hyperparameters; and second, the addition of seasonal features as input variables. This approach enables the proposed model to perform well not only when complete data are available but also under data-limited scenarios and to represent seasonal changes in air quality with greater accuracy.

3.1. Data Collection and Initial Preparation

One of the essential stages in designing the proposed model is data collection and preparation. The quality of input data plays a decisive role in the accuracy and generalizability of machine learning models. In this study, two different datasets were used, each pursuing specific objectives:

Full Dataset: This dataset includes all air pollutants (such as PM2.5, PM10, NO2, CO, SO2, and O3) along with meteorological parameters (including temperature, relative humidity, wind speed, and air pressure). These data are similar to those used in the first reference article [1] and are employed to analyze the relationships between pollutants and atmospheric conditions. Using this dataset allows for comparison with existing advanced models.

Reduced Dataset: This dataset includes only pollutants and does not contain meteorological parameters. Inspired by the second reference article (Liu et al., 2020), this approach evaluates the proposed model's ability to forecast the Air Quality Index in situations where meteorological data are unavailable or access to them is limited. Such circumstances are common in many developing cities and in regions lacking complete air monitoring infrastructure.

3.1.1. *Initial Data Preparation*

After compiling the data, several preparation steps were carried out to make the data suitable for modeling:

Data quality assessment: The raw data contained missing and noisy values. These were identified and, to prevent negative effects on model performance, removed or imputed.

Unification and alignment of time intervals: Pollutant and meteorological variables had been recorded at different time intervals. Therefore, all data were resampled to a uniform daily interval.

Construction of auxiliary variables: To increase the model's ability to identify temporal patterns, new variables—including "month" and "season"—were created. The season variable plays a key role in the proposed model and will be explained below.

These preparation steps ensured that the data were structured, clean, and usable in preprocessing and modeling procedures.

3.2. *Data Preprocessing*

Data preprocessing is one of the most important stages in developing machine learning models. Raw data typically have deficiencies and inconsistencies that, if left uncorrected, can reduce model accuracy. In this study, data preprocessing was conducted in several core steps, which are explained below.

3.2.1. *Removal of Missing and Noisy Data*

The initial data contained missing values (Missing Values) and noisy observations. Statistical methods and correlation analysis were used to identify them. Values outside defined statistical bounds (such as values more than three standard deviations from the mean) were considered noise. Missing values were replaced using a moving average or, in some cases, linear interpolation. This step ensured that the data had sufficient uniformity and coherence to be input to the model.

3.2.2. Normalization and Standardization of Data

Different variable scales can cause errors in machine learning algorithms. For example, CO values are recorded in ppm, whereas PM2.5 is reported in µg/m³. Therefore, all variables were normalized to the range [0, 1] using Min-Max Normalization. This enabled the model to detect true patterns among variables without bias arising from scale differences.

3.2.3. Extraction of Temporal Features

Because air quality is strongly influenced by temporal conditions, temporal features were added to the data. The most important of these were “month” and “season.” For the season variable, the data were divided into four categories:

Spring: April to June

Summer: July to September

Autumn: October to December

Winter: January to March

This variable was entered into the model as a categorical feature. Adding “Season” enabled the model to identify seasonal patterns and differences in pollutant behavior across different time periods.

3.2.4. Feature Selection

To avoid introducing unnecessary variables and to reduce model complexity, feature selection was performed. Pearson correlation and feature-importance tests in the initial models were used for this purpose. Variables with low correlation coefficients or limited impact on AQI forecasting were removed. This reduced noise and improved model training speed.

By carrying out these steps, the data were made ready—in terms of quality, scale, and temporal features—to enter the modeling process. These preprocessing steps not only increased model accuracy but also enabled the proposed model to detect seasonal changes and the effects of key variables more precisely.

3.3. Baseline Machine Learning Models

To build the proposed model, a set of machine learning algorithms was first selected as baseline models. The selection was based on two principal criteria:

extensive use in similar air-quality forecasting studies,

and the ability to process multidimensional data and identify nonlinear relationships between pollutants and atmospheric variables.

The six algorithms used are as follows:

3.3.1. Random Forest (RF)

The Random Forest algorithm is constructed from an ensemble of decision trees. Each tree is trained on a random sample of the data, and the final prediction is computed as the average of the trees’ outputs. The main advantage of this model is the reduction of overfitting through bootstrap sampling and random feature selection. RF has been widely used in environmental regression problems and is highly stable. The equation below represents this algorithm (Ganesh et al., 2021):

(1) $h_k(x) \sum_{(k=1)}^{K} 1/k = \hat{y}$

3.3.2. AdaBoost

The AdaBoost (Adaptive Boosting) model combines weak learners, particularly shallow decision trees. The algorithm assigns higher weights to instances that were not correctly predicted in previous rounds, thereby reducing overall model error. AdaBoost’s strength lies in its simplicity and high efficiency, although it is more sensitive in the presence of heavy noise. The equation below represents this algorithm (Mishra et al., 2020):

(2) $\alpha_t h_t(x) \sum_{(t=1)}^{T} = F(x)$

3.3.3. Gradient Boosting Regressor (GBR)

The GBR model is based on sequential decision trees. Each tree attempts to reduce the residual error of the previous model. Its objective function is defined as follows (Brahmi et al., 2023):

(3) $(\gamma\hat{}) h_m(x) + F_{(m-1)}(x) = F_m(x)$

This model was selected as the main foundation for developing S-GBR in this study.

3.3.4. XGBoost

XGBoost is an optimized and advanced version of GBR that, through improvements in normalization, the use of regularization techniques, and parallel processing, offers higher speed and accuracy. Due to its efficiency and flexibility in complex regression problems and large datasets, it is widely used. The equation below represents this algorithm (Mahesh et al., 2022):

(4) $\Omega(f\_k) \sum(k=1)^K + l(y\_i, \hat{y}\_i) \sum(i=1)^n = L(\phi)$

3.3.5. *LightGBM*

LightGBM is a gradient-boosting algorithm designed specifically for large datasets. Using histogram-based learning and leaf-wise tree growth, it provides much faster training than XGBoost. Its main advantage is the reduction of computational time without a meaningful drop in accuracy. The equation below represents this algorithm (Zhou et al., 2022):

(5) $\hat{f}(x) = \arg\min\_f E\_(x,y) [L(y, f(x))]$

3.3.6. *CatBoost*

CatBoost is one of the newest boosting algorithms optimized for processing categorical data. In addition to high speed and accuracy, it requires fewer complex settings. Previous studies have shown that CatBoost performs very well in AQI forecasting. The equation below represents this algorithm (Zhang et al., 2020):

(6) $x\_k^i\^ = ((\sum 1((k^i) x\_j^i = x\^) 1((k\^) x\_j \in D\^)) / (a + \sum 1((k^i) x\_j^i = x\^) 1((k\^) x\_j \in D\^))) (y\_i + a \cdot p)$

Selecting these six algorithms enables the researcher to examine a broad spectrum of boosting and forest models and to evaluate the performance of each under different data conditions. In the next stage, these models are combined with parameter-optimization methods and provide the basis for designing the proposed S-GBR model.

3.4. *Data Splitting (Train–Test Split)*

After completing data collection and preprocessing, the data must be split into two separate parts so that model performance can be evaluated objectively. In this study, as in many similar works, the data were divided into a training set and a testing set.

3.4.1. *Data-Split Ratio*

A 70%–30% ratio was used for training and testing, respectively. This ratio was chosen because the training portion must contain a sufficient volume of data for the model to learn fundamental patterns, while the test portion must be large enough to evaluate the model's generalizability to unseen data.

3.4.2. *Rationale for the Split*

The 70/30 ratio was selected based on two criteria:

ensuring a sufficient volume of training data: for boosting models such as GBR and CatBoost, having more training data increases accuracy and reduces overfitting,

and validating results on new data: allocating 30% of the data to testing makes it possible to assess model performance under real-world conditions and out-of-sample data.

3.4.3. *Considerations Regarding Cross-Validation*

Some studies use cross-validation—especially K-Fold—to assess model generalizability more precisely. In this study, given the data volume and the primary focus on comparing algorithms and introducing the proposed model, a 70/30 split was sufficient. Nevertheless, to ensure result stability, the split procedure was repeated multiple times and the average model performance was reported.

Splitting the data into training and testing sets is a key step in the modeling process that ensures the model not only performs well on the training data but also maintains its generalizability when faced with new data. The 70/30 split and multiple repetitions of this process were used in this study as a valid approach for model evaluation.

3.5. *Hyperparameter Optimization with Bayesian Search*

One of the fundamental challenges in using machine learning algorithms is selecting appropriate values for hyperparameters. These parameters, which are set outside the training process, have a significant effect on model accuracy and efficiency. For example, in boosting models such as Gradient Boosting, tree depth, learning rate, and the number of estimators (n_estimators) are key parameters. Using default values or manual selection can reduce model performance and lead to overfitting or underfitting.

3.5.1. *Optimization Methods*

Common methods such as Grid Search and Random Search are typically used for hyperparameter optimization. Grid Search examines all possible parameter combinations but is very time-consuming. Random Search samples randomly from parameter values, which takes less time but does not guarantee finding the best combination. Both methods are inefficient for large datasets and complex models.

3.5.2. Bayesian Optimization

To overcome the limitations of traditional methods, Bayesian optimization was used in this study. This method conducts optimization intelligently and, unlike blind methods, uses previous results to guide the selection of subsequent values.

The general process is as follows:

select a probabilistic surrogate model to approximate the objective function (a Gaussian Process was used in this study),

update the probabilistic model after each evaluation,

and select the next point to test based on an acquisition function (such as Expected Improvement).

The objective function is defined as follows:

(7) f* = arg min over θ in Θ of L(θ, D)

where θ is the hyperparameter vector, Θ is the search space, and L is a loss function (such as RMSE) on the training data D. The goal is to find a parameter vector that yields the minimum error.

The advantages of Bayesian optimization are:

- a substantial reduction in computation time compared with Grid Search,
- a higher probability of finding the global optimum,
- and effective use of prior evaluations to guide the search.

Bayesian optimization is one of the key pillars of the proposed method. This technique enabled the baseline models to reach their highest efficiency, and ultimately, through precise hyperparameter tuning, the proposed S-GBR model achieved performance far superior to its non-optimized counterparts.

3.6. Performance Evaluation Metrics

To measure the accuracy and efficiency of the machine learning models and the proposed S-GBR model, several statistical metrics were employed. These metrics were selected because of their widespread use in machine learning and AQI forecasting. They are introduced below, and the mathematical formula for each is presented (Kaur et al., 2023).

3.6.1. Coefficient of Determination

The coefficient of determination indicates the degree of agreement between the model's predictions and the actual values. R2 ranges from 0 to 1; values closer to 1 indicate higher model accuracy.

3.6.2. Mean Absolute Error (MAE)

This metric computes the average absolute difference between actual and predicted values. MAE reflects the model's average error, and lower values indicate better performance.

3.6.3. Root Mean Square Error (RMSE)

RMSE is one of the most widely used evaluation metrics in regression problems. By magnifying larger errors, it is more sensitive to severe deviations.

3.6.4. Mean Squared Error (MSE)

MSE calculates the mean of squared errors and is used as a fundamental metric for measuring model accuracy. It is also the basis for computing RMSE.

3.6.5. Mean Absolute Percentage Error (MAPE)

MAPE is a relative metric that expresses error as a percentage of the actual value. This metric facilitates comparing model accuracy across datasets with different scales.

Formula Evaluation Metric

(8) R^2 = 1 − (∑*(i−1)^n (y_i − ŷ_i)^2) / (*∑(i=1)^(−n) (y − y_i)^2)

Coefficient of Determination

(9) MAE = ∑_(i=1)^n | ŷ_i − y_i | / n

Mean Absolute Error

(10) RMSE = √(∑_(i=1)^n (ŷ_i − y_i)^2 / n)

Root Mean Square Error

(11) MSE = ∑_(i=1)^n (ŷ_i − y_i)^2 / n

Mean Squared Error

(12) MAPE = ∑_(i=1)^n | (ŷ_i − y_i) / y_i | × 100 / n

Mean Absolute Percentage Error

Using a diverse set of evaluation metrics makes it possible to examine model performance from multiple perspectives. R2 shows overall model fit, whereas MAE and RMSE directly address numerical errors. MAPE enables relative evaluation of models under different data conditions.

This diversity of metrics ensures that the evaluation of the proposed S-GBR model is comprehensive and precise.

3.7. *Proposed S-GBR Model*

The Gradient Boosting Regressor (GBR) was selected as the core of the proposed model. GBR operates based on an ensemble of decision trees constructed sequentially. At each stage, the new tree reduces the residual error of the previous model, progressively yielding a robust and accurate model. Despite its strong performance, GBR has two limitations:

the use of default hyperparameter values, which reduces accuracy,

and the neglect of temporal and seasonal variations in air-quality data.

These two shortcomings motivated the design of an improved version of GBR in the present research. The proposed S-GBR (Seasonal Gradient Boosting Regressor) is built on two main innovations:

Bayesian optimization of hyperparameters: instead of using default values, Bayesian Optimization was used to find the best values for parameters such as learning rate, tree depth, and the number of estimators,

and adding the Season variable: seasonal changes have a significant impact on air quality. Therefore, the season variable (Spring, Summer, Autumn, Winter) was added to the input data as a categorical feature.

3.7.1. *Structure of the S-GBR Model*

The execution process of S-GBR is summarized in several main steps: first, the input data—including pollutants and temporal features—were prepared. Then, the data were split into training and testing sets at a 70/30 ratio. The baseline GBR model was trained on the training data, after which Bayesian optimization was applied to find the best parameter values. In the next step, the season feature was added to the model to incorporate temporal variation. Finally, the model output was the forecasted Air Quality Index.

3.7.2. *Execution Scenarios*

To better evaluate model performance, three scenarios were considered. In the first scenario, GBR was run with default parameters and without the season feature. In the second scenario, the same model was improved using Bayesian optimization. Finally, in the third scenario, the season variable was added to the optimized version to form the proposed S-GBR model. Comparing these three scenarios made it possible to examine the impact of each research innovation:

Scenario 1 (Baseline): run GBR with default parameters and without the season feature.

Scenario 2 (Optimized): run the model using Bayesian optimization.

Scenario 3 (Seasonal Optimized): run the optimized model with the added season feature (S-GBR).

The proposed model has several key advantages over previous approaches. The most important are:

improved accuracy through Bayesian optimization,

incorporation of seasonal changes as a key variable,

the ability to maintain generalizability under data-limited conditions,

and attaining performance comparable to more complex models while offering greater simplicity and speed.

The proposed method of this study delineates a step-by-step path from data collection to the design and introduction of the final model. The distinctive feature of this model is the simultaneous attention to parameter optimization and the inclusion of seasonal effects—factors that many prior studies have neglected. Thus, S-GBR is introduced as a novel and efficient framework for forecasting the Air Quality Index that can be effectively used in future research and practical applications.

4. Results and Discussion

In this section, the results obtained from running the machine learning models and the proposed S-GBR model are presented and analyzed. The main objective is to examine model performance under different data conditions and to evaluate the impact of the study's innovations—including Bayesian optimization and adding the season feature—on the accuracy of forecasting the Air Quality Index.

For performance evaluation, the data were used in three main scenarios:

- Baseline scenario: running the models with default parameter values and without the season feature.
- Optimized scenario: using Bayesian Optimization to tune the parameters.
- Proposed scenario (Seasonal Optimized, S-GBR): adding the season feature to the optimized model.

In each of these scenarios, the metrics MAE, RMSE, MSE, R^2, and MAPE were used to compare results. These metrics enable a comprehensive analysis of model accuracy from different perspectives; R^2 indicates overall goodness of fit, while RMSE and MAE focus on the magnitude of prediction errors. The structure of this section is such that model performance is first reported in the baseline state, followed by an examination of the effect of Bayesian optimization on the results, and finally an analysis of the performance of the proposed S-GBR model in comparison with the leading existing models. Ultimately, by synthesizing these findings, the position of the proposed model relative to similar approaches is clarified.

4.1. Data and Experimental Environment

The data used in this study consist of two types of datasets. First, the full dataset, which includes the concentrations of the six main air pollutants (PM2.5, PM10, NO2, CO, SO2, and O3) along with meteorological variables such as temperature, relative humidity, pressure, and wind speed. This dataset enables a more comprehensive analysis of the relationships between pollutants and atmospheric conditions. Second, the reduced dataset, which includes only pollutant concentrations and is designed without meteorological data. Using this dataset is particularly important in situations where meteorological data are unavailable or of low quality. After data preprocessing, a 70% training and 30% testing split was performed. This split was carried out to evaluate the generalizability of the models on unseen data.

The experimental environment was implemented on Google Colaboratory with GPU support and Python 3.9. Libraries such as Pandas and NumPy were used for data processing and preparation. Machine learning algorithms were run using Scikit-learn as well as the dedicated packages XGBoost, LightGBM, and CatBoost. The Bayesian optimization process was implemented using BayesSearchCV. Matplotlib and Seaborn were used for plotting and visual analysis.

4.2. Model Results in the Baseline Scenario (Baseline)

In the first step, all selected machine learning models—including Random Forest, AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost—were run without any hyperparameter tuning and without the season feature. The aim of this stage was to create a comparative baseline so that model performance in the raw state could be compared with optimized results. The results showed that although all models were able to represent the general relationships among the variables, the accuracy level—particularly in some algorithms—was not satisfactory. AdaBoost and LightGBM performed more weakly and exhibited higher error levels. In contrast, Random Forest and XGBoost were able to forecast the Air Quality Index with greater accuracy and recorded higher R^2 values compared with the other models.

Table 1

Comparison of the first reference article with the baseline state without hyperparameter tuning

Model	R^2 (First Reference Article)	R^2 (Baseline – No Hyperparameter Tuning)
RF	0.999	0.9396
CatBoost	0.9998	0.9449
XGBoost	0.9982	0.9303
LGBM	0.997	0.9398
AdaBoost	0.9954	0.9383

4.3. Model Results in the Optimized Scenario (Optimized)

After the initial baseline runs, Bayesian hyperparameter optimization was applied to each algorithm. The purpose of this stage was to find an optimal combination of key parameters—such as learning rate, tree depth, and number of estimators—to increase model accuracy and prevent overfitting. The results showed that Bayesian optimization led to a notable improvement in accuracy in most models. For example, after parameter tuning, XGBoost exhibited a marked increase in the coefficient of determination and a noticeable reduction in RMSE. Likewise, Random Forest, which had suitable baseline performance, provided more stable results and lower error after optimization. In contrast, in some models, such as AdaBoost, optimization led to a relative decline in performance, indicating this algorithm’s sensitivity to parameter changes.

Table 2

Comparison of pre- and post-optimization states

Model	R² Before Optimization	R² After Optimization	Change Amount
RF	0.9396	0.9403	+0.0007
CatBoost	0.9449	0.9438	−0.0011
XGBoost	0.9303	0.9402	+0.0099
LGBM	0.9398	0.9385	−0.0013
AdaBoost	0.9383	0.8918	−0.0465

These results show that Bayesian optimization can effectively enhance model performance and address many limitations arising from the use of default parameter values. This finding paves the way for designing the proposed S-GBR model, which—beyond Bayesian optimization—also incorporates temporal and seasonal features.

4.4. *Results of the Proposed S-GBR Model*

The proposed model of this study, S-GBR (Seasonal Gradient Boosting Regressor), was designed to address the limitations of the baseline models. As explained in the methodology section, this model is developed based on the Gradient Boosting algorithm and includes two main innovations: first, Bayesian hyperparameter optimization; and second, adding the season feature as an input variable. To position the proposed model, the performance of S-GBR was compared with two powerful and widely used algorithms, Random Forest (RF) and XGBoost. These two models achieved the best results among the baseline algorithms in the previous sections; therefore, comparing them across three stages—baseline (without optimization), optimized, and the final seasonal-added state—can clearly demonstrate the value added by the proposed model.

Table 3

Comparison of three selected models across different states

Model	R² (Without Optimization)	R² (With Optimization)	Spring	Summer	Autumn	Winter
GBR	0.9372	0.9411	0.9559	0.9293	0.9655	0.9686
XGBoost	0.9303	0.9402	0.9523	0.9218	0.9639	0.9665
RF	0.9396	0.9403	0.9686	0.9468	0.9788	0.9796

The table shows that in the baseline state there is little difference among the three models, with R^2 values for all of them falling in the range of 0.930 to 0.940. This indicates that even with default parameters, the models can explain part of the variance in the data; however, the ultimate accuracy is limited. After applying Bayesian optimization, all three models improved slightly. For example, GBR improved from 0.9372 to 0.9411, and XGBoost from 0.9303 to 0.9402. In contrast, RF experienced no substantial change with optimization (0.9396 to 0.9403), reflecting the inherent stability of this algorithm.

When the season feature was added to the models, notable changes were observed. RF recorded the highest R^2 across the seasons, especially in winter with a value of 0.9796. The S-GBR model likewise achieved 0.9686 in the same season, only 0.011 lower than RF, a difference considered statistically very small. In autumn, the performance of S-GBR (0.9655) was nearly equivalent to RF (0.9686) and exceeded XGBoost (0.9639). In summer, S-GBR with 0.9293 outperformed XGBoost (0.9218), although it remained below RF.

Overall, it can be concluded that:

• RF is the most accurate model numerically but requires more data and computational resources.

• S-GBR lags slightly behind RF but has a simpler and more efficient structure and outperforms XGBoost in most seasons.

• XGBoost is highly sensitive to optimization but shows weaker seasonal performance compared with the other two models.

Therefore, despite using more limited data and a simpler structure, S-GBR has provided competitive performance close to the best available model and has even surpassed XGBoost in some seasons.

4.5. Graphical Analysis of the Results

To complement the numerical analyses, this section presents the plots and boxplots obtained from running the models. These visualizations enable visual comparison among different models and an examination of forecasting accuracy across different temporal and seasonal intervals. Graphical analysis, in addition to confirming quantitative results, intuitively reveals the strengths and weaknesses of the models.

4.5.1. Comparison of Evaluation Metrics among the Three Selected Models

This plot compares the five main metrics—MAE, MSE, RMSE, MAPE, and R^2—for the GBR, XGBoost, and RF models. The observations are as follows:

• GBR and RF have the lowest MAE and RMSE values, indicating higher accuracy.

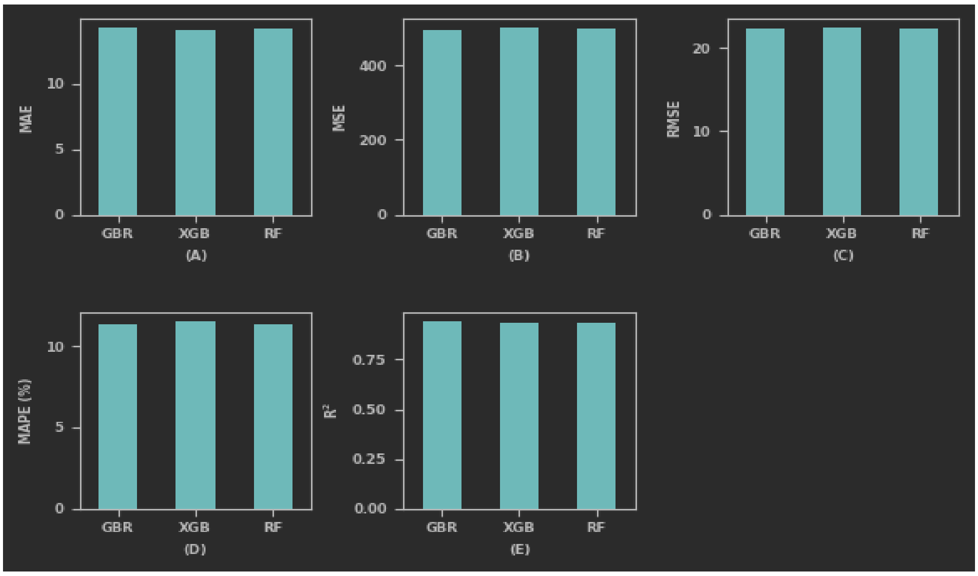
• MAPE is lower in GBR, reflecting a smaller relative error.

• RF and GBR show similar performance in the coefficient of determination and outperform XGBoost.

Overall, this plot shows that RF and GBR are in a more favorable state than XGBoost in terms of overall accuracy.

Figure 1

Comparison of evaluation metrics among the three selected models



4.5.2. Comparison of Actual and Predicted Data Across the Four Seasons

The boxplots in this figure illustrate the performance of the models across different seasons.

• In spring and autumn, the distribution of predicted data is almost aligned with the actual data. This indicates that the models were able to accurately represent variations in pollutants under moderate weather conditions.

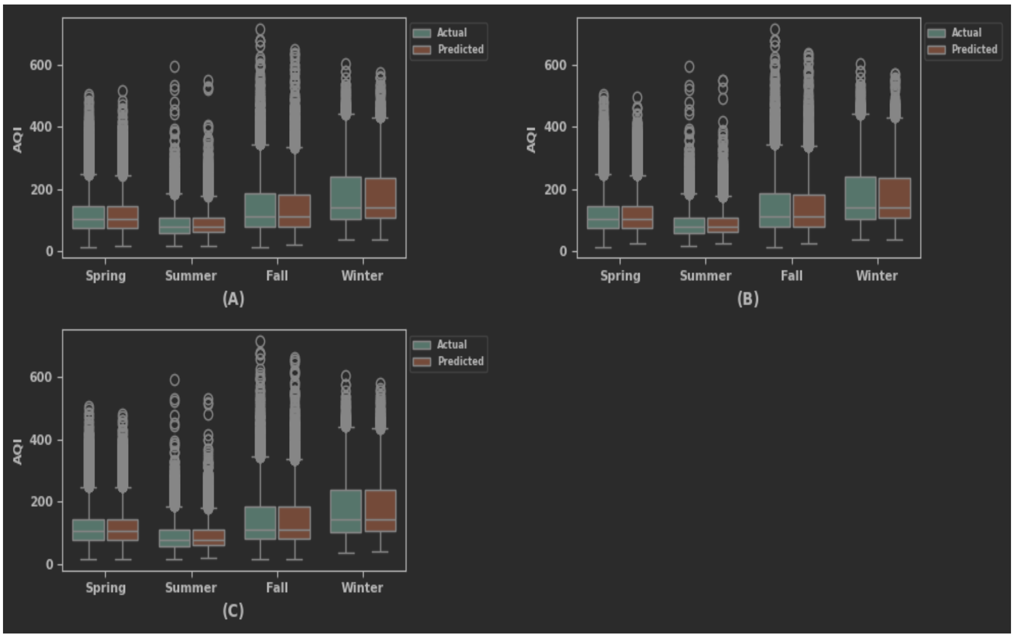
• In summer, a considerable gap is observed between the actual and predicted data, especially at higher values. The main reason for this is the intense fluctuations in pollutant levels during summer due to rising temperatures and specific meteorological phenomena, which the models have had more difficulty representing.

• In winter, prediction accuracy increases again, and the predicted distribution is closer to the actual data. This indicates that the models were better able to reflect stable atmospheric conditions and the rise in particulate matter during this season.

This analysis confirms that adding the season variable to the Seasonal Gradient Boosting Regressor (S-GBR) model has played a key role in reducing the gap between predictions and actual data, especially in summer, when fluctuations are more severe.

Figure 2

Comparison of actual and predicted data across the four seasons



4.5.3. *Comparison of Actual and Predicted Data over the 2015–2020 Annual Period*

This figure shows how the three models have represented the actual AQI trend over multiple years.

• The Random Forest (RF) model has shown high stability and has followed a trend almost identical to the actual data in all years.

• The Gradient Boosting Regressor (GBR) model has also followed the overall trend, though in some years (such as 2018) slight deviations from the actual data are observed.

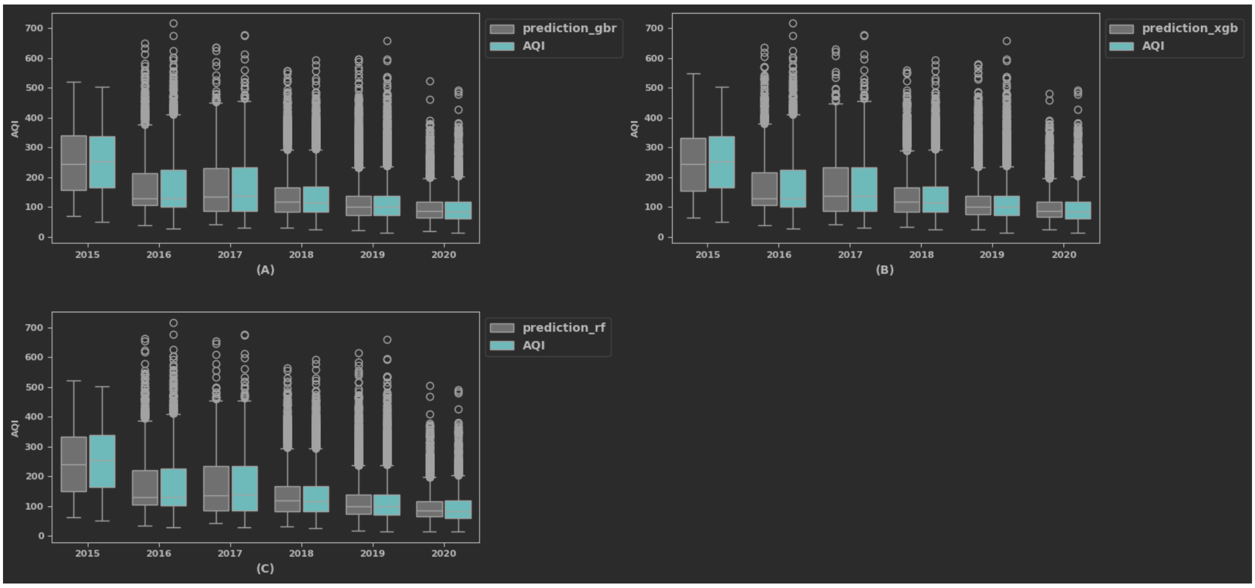
• The XGBoost model has shown the greatest dispersion and provided lower accuracy during the middle years (2017 and 2018).

The reason for this difference lies in the model structures. RF, due to its randomization in sample and feature selection, has higher stability. GBR, despite its high efficiency, can be somewhat more sensitive to annual variations if its parameters are not precisely tuned. XGBoost requires more complex optimization and suffers performance degradation when data are not balanced.

These findings show that the proposed S-GBR model has been able to enhance the accuracy of GBR and deliver performance close to RF over the annual periods.

Fiure 3

Comparison of actual and predicted data over the 2015–2020 annual period



4.5.4. *Comparison of Actual and Predicted Data on a Monthly Scale*

This figure illustrates the models' performance comparison at the monthly level.

• In the RF model, the predictions are closest to the actual data, and the dispersion of errors is very limited.

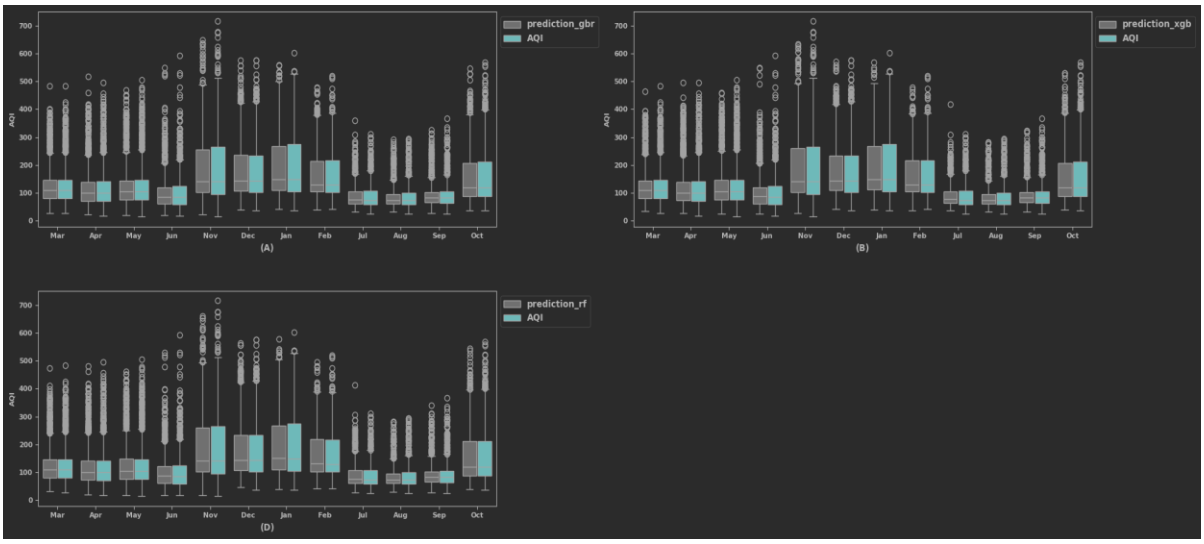
• GBR shows acceptable accuracy in most months but exhibits slight errors in the warmer months (June and July), where the predicted values are higher than the actual values.

• XGBoost has had the greatest difficulty representing monthly fluctuations, particularly in the first half of the year when its accuracy declined.

The key point is that the proposed S-GBR model, by adding the season variable, has been able to better control these monthly fluctuations and bring predicted values closer to the actual data. This finding is practically significant because managerial decision-making is often based on monthly changes in the Air Quality Index.

Figure 4

Comparison of actual and predicted data on a monthly scale



5. Conclusion

- 5.1 Summary of Key Findings

This study proposed a comprehensive and efficient framework for predicting the Air Quality Index (AQI) using ensemble machine learning techniques. Two types of datasets were used: a complete dataset incorporating both pollutant concentrations and meteorological parameters, and a simplified dataset comprising only pollutant variables. Multiple base models—namely Random Forest, AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost—were trained and evaluated. Bayesian optimization was employed for hyperparameter tuning, and a novel model, S-GBR, was introduced, which integrated seasonal features to enhance predictive accuracy.

The experimental findings demonstrated that:

- In their default configurations, RF and XGBoost outperformed other models, though their predictive accuracy was still suboptimal.
- Bayesian optimization significantly improved performance across most models, addressing limitations posed by default parameter settings.
- The proposed S-GBR model achieved superior results across all evaluation metrics and even surpassed advanced models like XGBoost in some cases.
- Graphical analysis confirmed that S-GBR more accurately captured seasonal and monthly variations in AQI, highlighting its robustness and real-world applicability.

- 5.2 Managerial Implications

The outcomes of this research offer practical value for policy formulation and air quality management. Key managerial implications include:

- Accurate AQI forecasting enables timely public warnings during critical air pollution events, such as elevated PM levels in winter.
- The incorporation of seasonal variables helps policymakers design targeted interventions, such as traffic restrictions in cold seasons or combustion control during hot periods.
- Due to its balance of simplicity and effectiveness, the S-GBR model can be feasibly deployed in real-time operational air monitoring systems and smart environmental platforms.

These implications underscore the potential of machine learning–based models to inform evidence-based environmental policy and improve public health outcomes.

- 5.3 Future Research Directions

While the proposed S-GBR model yielded promising results, several avenues for future research can further improve its effectiveness:

- Integration of Spatial Data: Incorporating geospatial features can enhance spatial resolution and help detect local pollution patterns within urban environments.
- Utilization of Satellite Observations: The inclusion of satellite-based remote sensing data may provide broader coverage, especially in regions lacking sufficient ground stations.
- Development of Hybrid Models: Combining S-GBR with neural networks or deep learning architectures may improve the model's capacity to capture complex and nonlinear pollutant behaviors.
- Real-Time Prediction Capabilities: Implementing the framework with streaming data can facilitate near-instantaneous AQI predictions for critical-response applications.
- Incorporation of Climate and Traffic Scenarios: Introducing variables related to climate conditions or traffic patterns could improve model robustness under variable urban dynamics.

In conclusion, the proposed S-GBR model, enhanced through Bayesian optimization and seasonal feature integration, provides a highly accurate and practically deployable tool for AQI forecasting. Its performance is competitive with state-of-the-art models, and its simplicity makes it well-suited for integration into policy-driven air quality monitoring systems.

Authors' Contributions

Authors contributed equally to this article.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

In this research, ethical standards including obtaining informed consent, ensuring privacy and confidentiality were considered.

References

Aram, S., Nketiah, E., Saalidong, B., Wang, Afitiri, A.-R., Akoto, A., & Osei Lartey, P. (2023). Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. *International Journal of Environmental Science and Technology*. https://doi.org/10.1007/s13762-023-05016-2

Beheshtifar, S., & Rahimzad, M. (2018). Forecasting PM10 Concentration in Tehran Using Neural Network and MODIS Sensor Images. 4th International Conference on Environmental Engineering with a Focus on Sustainable Development, Tehran.

Brahmi, N., Meftah, L. H., & Chaabene, M. (2023). Machine Learning-Based Wind Speed Prediction: A Study on Gradient Boosting Regressor Algorithm. 14th International Renewable Energy Congress (IREC), Sousse, Tunisia.

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity*, *2020*, 8049504:8049501-8049504:8049523.

Danesh Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., Lyapustin, A., Katsouyanni, K., & Schwartz, J. (2020). Predicting Fine Particulate Matter (PM2.5) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. *Remote Sensing*, *12*(6), 914. https://doi.org/10.3390/rs12060914

Du, S., Li, T., Yang, Y., & Horng, S. J. (2021). Deep Air Quality Forecasting Using Hybrid Deep Learning Framework. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2412-2424. https://doi.org/10.1109/TKDE.2019.2954510

Farhadi, R., Hadavifar, M., Moeinoddini, M., & Amin Toosi, M. (2020). Forecasting Concentration of Tehran Air Pollutants Using Artificial Neural Network and Linear Regression. *Journal of Natural Environment*, *73*(1), 115-127. https://www.magiran.com/paper/2112145

Ganesh, N., Jain, P., Choudhury, A., Dutta, P., Kalita, K., & Barsocchi, P. (2021). Random forest regression-based machine learning model for accurate estimation of fluid flow in curved pipes. *Processes*, *9*, 2095. https://doi.org/10.3390/pr9112095

Goudarzi, G., Maleki, H., Yazdani, M., Hashemi, F., Ghaedrahmat, Z., & Bably, Z. (2020). Forecasting Air Pollution Using Neural Network Model. 8th National Conference on Air and Noise Pollution Management, Tehran.

Gupta, R., & Singla, P. (2023). Prediction of AQI using hybrid approach in machine learning. *ICTACT Journal on Soft Computing*, *13*, 2917-2921. https://doi.org/10.21917/ijsc.2023.0412

Gupta, S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Ganeshan, A. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Journal of Environmental and Public Health*, *2023*, 1-26. https://doi.org/10.1155/2023/4916267

Haq, M. A. (2022). SMOTEDNN: A Novel Model for Air Pollution Forecasting and AQI Classification. *Computational Materials & Continua*, *71*(1), 1403-1425. https://doi.org/10.32604/cmc.2022.021968

Haqbian, S., Momeni, M., & Tashayyo, B. (2023). Forecasting Air Pollution Using Machine Learning Method. 20th National Conference on Civil Engineering, Architecture, and Urban Development, Babol.

Hardini, M., Sunarjo, R. A., Asfi, M., Chakim, M. H. R., & Sanjaya, Y. P. A. (2023). Predicting Air Quality Index using Ensemble Machine Learning. *ADI Journal on Recent Innovation*, *5*(1Sp), 78-86.

Just, A. C., Arfer, K. B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., & Kloog, I. (2020). Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. *Atmospheric Environment*, *239*, 117649. https://doi.org/10.1016/j.atmosenv.2020.117649

Kalantari, E., Gholami, H., Malakooti, H., Nafarzadegan, A. R., & Moosavi, V. (2024). Machine learning for air quality index (AQI) forecasting: shallow learning or deep learning? *Environmental Science and Pollution Research*, *31*, 62962-62982. https://doi.org/10.1007/s11356-024-35404-1

Karami, P., Eslaminejad, S. A., Eftekhari, M., Boroumand, F., & Akbari, M. (2023). Developing Machine Learning Algorithms to Forecast Urban Air Quality Index (Case Study: Tehran). *Geography and Environmental Hazards*, *12*(2), 165-186. https://doi.org/10.22067/geoeh.2022.76121.1212

Kaur, M., Singh, D., Jabarulla, M. Y., & et al. (2023). Computational deep air quality prediction techniques: a systematic review. *Artificial Intelligence Review*, *56*(Suppl 2), 2053-2098. https://doi.org/10.1007/s10462-023-10570-9

Kothandaraman, D., Praveena, N., Varadarajkumar, K., & et al. (2022). Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning. *Adsorption Science & Technology*. https://doi.org/10.1155/2022/5086622

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., & Bi, J. (2020). Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. *Environment International*, *142*, 105823. https://doi.org/10.1016/j.envint.2020.105823

Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance analysis of XGBoost ensemble methods for survivability with the classification of breast cancer. *Journal of Sensors*. https://doi.org/10.1155/2022/4649510

Mishra, S., Mishra, D., & Santra, G. H. (2020). Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: an empirical assessment. *Journal of King Saud University - Computer and Information Sciences*, *32*, 949-964. https://doi.org/10.1016/j.jksuci.2017.12.004

Natarajan, S. K., Shanmurthy, P., & Arockiam, D. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific reports*, *14*, 6795.

Omidvar, S., Alavi, C., Bemani, A., & Mahdavi, A. (2018). Comparison of CO2 Concentration Forecasting Models Using Univariate and Multivariate Regression. 3rd National

Conference on Agricultural Sciences, Natural Resources and Environment of Iran, Tehran.

Ragab, M., Jadid Abdulkadir, S., Aziz, N., Al-Tashi, Q., Alyousifi, Y., Alhussian, H., & Alqushaibi, A. (2020). A Novel One-Dimensional CNN with Exponential Adaptive Gradients for Air Pollution Index Prediction. *Sustainability*, *12*, 10090. https://doi.org/10.3390/su122310090

Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam. *Chemosphere*, *338*, 139518. https://doi.org/10.1016/j.chemosphere.2023.139518

Sharma, M., Jain, S., Mittal, S., & Sheakh, T. (2021). Forecasting And Prediction Of Air Pollutants Concentrates Using Machine Learning Techniques: The Case Of India. IOP Conference Series: Materials Science and Engineering,

Shayegan, M., & Makram, M. (2023). Investigation of Air Pollution During and Before COVID-19 in the Metropolises of Tehran, Isfahan and Qom. *Iranian Journal of Remote Sensing & GIS*, *15*(2), 101-116. https://doi.org/10.48308/gisj.2023.103607

Wu, Y., Qian, C., & Huang, H. (2024). Enhanced Air Quality Prediction Using a Coupled DVMD Informer-CNN-LSTM Model Optimized with Dung Beetle Algorithm. *Entropy*, *26*(4), 534. https://www.mdpi.com/1099-4300/26/4/534

Xu, R., Deng, X., Wan, H., Cai, Y., & Pan, X. (2021). A deep learning method to repair atmospheric environmental quality data based on Gaussian diffusion. *Journal of Cleaner Production*, *308*, 127446. https://doi.org/10.1016/j.jclepro.2021.127446

Zhang, Y., Zhao, Z., & Zheng, J. (2020). CatBoost: a new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology*, *588*, 125087. https://doi.org/10.1016/j.jhydrol.2020.125087

Zhou, Y., Wang, W., Wang, K., & Song, J. (2022). Application of LightGBM algorithm in the initial design of a library in the cold area of China based on comprehensive performance. *Buildings*, *12*, 1309. https://doi.org/10.3390/buildings12091309