Journal of Applied Data Sciences
Vol. 6, No. 4, December 2025, pp. 2461-2476


# Air Pollution Forecasting in Almaty using Ensemble Machine Learning Models

Lyazat Naizabayeva[1], Gulbakyt Sembina[2,*], Alibek Aliman[3], Maxatbek Satymbekov[4], Nazym Barlykbay[5], Nurgul Seilova[6]

[1,2,3,4,5,6]*Information Systems Department, International Information Technology University, Manas St., 34/1, Almaty, 050000, Kazakhstan*

(Received: February 05, 2025; Revised: May 07, 2025; Accepted: August 01, 2025; Available online: September 11, 2025)

**Abstract**

This study develops an advanced forecasting methodology for air pollution levels in Almaty, Kazakhstan, focusing on fine Particulate Matter (PM2.5) and carbon monoxide concentrations. Air pollution poses significant risks to public health, and Almaty's basin location exacerbates the problem. Addressing the limitations of traditional statistical forecasting methods, we propose an ensemble machine learning approach that integrates Seasonal-Trend decomposition with gradient boosting algorithms to capture complex temporal and nonlinear patterns. The objective is to develop and validate an effective methodology for forecasting atmospheric air pollution in Almaty using machine learning methods, in particular STL decomposition, XGBoost, LightGBM models, and their ensemble combination. The novelty lies in the integration of STL decomposition with an ensemble of gradient boosting models for high-accuracy air pollution forecasting in the complex urban environment of Almaty. The dataset includes hourly measurements from over 20 monitoring stations, enabling seasonal and spatial analysis. Rigorous preprocessing techniques were applied, including outlier removal, normalization, and time series decomposition into seasonal, trend, and residual components. Two gradient boosting models, XGBoost and LightGBM, were trained separately and combined into a weighted ensemble, with optimal weights determined through cross-validation. Figures and tables illustrate data preprocessing flow, model architectures, feature importance analysis, and evaluation of predictive performance. The ensemble outperformed individual models, achieving high accuracy with coefficient of determination values exceeding 0.98 for PM2.5 and 0.83 for carbon monoxide. The findings demonstrate that integrating Seasonal-Trend decomposition with ensemble learning provides a robust and effective approach to forecasting air pollution in complex urban environments. The methodology shows strong potential for practical application in real-time air quality monitoring and warning systems, aiding policymakers and public health authorities. Future research will expand the dataset by incorporating additional factors such as traffic flow, industrial emissions, and satellite remote sensing data to enhance predictive accuracy and model interpretability.

*Keywords:* Air Quality, PM2.5, Pollution, Carbon Monoxide (CO), STL Decomposition, Xgboost, Lightgbm, Machine Learning, Forecasting, Ensemble Model

## 1. Introduction

Air pollution is recognized as one of the leading environmental threats globally, affecting human health, ecosystems, and the climate. According to the World Health Organization, millions of premature deaths annually are attributed to exposure to air pollution [1]. Fine PM2.5 and CO are among the most dangerous pollutants. Long-term exposure to these pollutants is associated with an increased risk of respiratory and cardiovascular diseases.

The city of Almaty, Kazakhstan, is particularly prone to air pollution due to its unique geographical location in a basin surrounded by mountains, which restricts air circulation and leads to the accumulation of pollutants. According to official reports, Almaty remains one of the most polluted cities in Kazakhstan [2]. Seasonal temperature inversions, increased motor vehicle emissions, and industrial activities further exacerbate the situation. Recent studies have identified causal relationships in Almaty's air pollution dynamics, emphasizing the complexity of its environmental conditions [3].

In recent decades, various traditional methods have been developed to forecast air pollution levels. These include numerical weather prediction models [4], deterministic chemical transport models [5], and classical statistical approaches such as autoregressive models and multiple linear regression [6]. Although these methods have provided

***Corresponding author: Gulbakyt Sembina (g.sembina@iitu.edu.kz)**
**DOI: https://doi.org/10.47738/jads.v6i4.821**

ISSN 2723-6471


valuable insights, they often struggle to accurately capture the nonlinear and dynamic behavior of atmospheric pollutants. Their performance heavily depends on high-quality input data and considerable computational resources, which limits their applicability in real-time forecasting systems [7]. Moreover, traditional models are generally less adaptable to sudden changes in emission patterns and meteorological conditions, leading to increased forecast uncertainty.

To address these challenges, Machine Learning (ML) techniques [8] have emerged as powerful tools for air quality forecasting. ML models can effectively uncover complex nonlinear relationships in large datasets, providing high predictive accuracy. In particular, ensemble learning methods, which combine the strengths of multiple base learners, have gained popularity due to their robustness and improved generalization capabilities. Gradient boosting algorithms, such as Extreme Gradient Boosting (XGBoost) [9] and Light Gradient Boosting Machine (LightGBM) [10], have shown outstanding performance in handling structured data and have been widely adopted for various forecasting tasks. These models are capable of managing missing data, capturing intricate feature interactions, and delivering fast and accurate predictions.

However, most machine learning models, including gradient boosting algorithms, do not explicitly account for the strong seasonal components inherent in air pollution time series. In cities like Almaty, where air quality is significantly influenced by seasonal heating, meteorological inversions, and topographical features, the ability to model seasonality is crucial for accurate forecasting. Traditional time series decomposition methods have been used to address seasonality, but many lack the flexibility to handle complex and evolving seasonal patterns.

Seasonal-Trend decomposition using Loess [11] offers a more robust alternative, allowing the separation of time series data into seasonal, trend, and residual components. STL is flexible, can accommodate nonconstant seasonality, and is resistant to outliers. By applying STL, it becomes possible to isolate the deterministic seasonal and trend behaviors, enabling machine learning models to focus on the residual, less predictable components of air quality data. This hybrid approach can enhance the performance and stability of forecasting models, especially in environments characterized by strong seasonal variations.

Although hybrid models combining decomposition techniques with machine learning have been explored, there is still a noticeable gap in integrating STL decomposition with ensemble learning algorithms like XGBoost and LightGBM for air pollution forecasting, particularly in complex urban environments such as Almaty. Existing research has mostly focused on standalone machine learning models or conventional decomposition methods, often neglecting the benefits of combining decomposition with powerful ensemble learners.

The primary goal of this study is to develop an air quality forecasting model for PM2.5 and CO concentrations in Almaty by integrating STL decomposition with ensemble machine learning techniques. To achieve this, key steps include data preprocessing and cleaning, extraction of seasonal components via STL decomposition, training of XGBoost and LightGBM models, construction of a weighted ensemble, and evaluation of model performance using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), And the Coefficient of Determination ($R^2$).

The model leverages real-world environmental and meteorological data collected from 2020 to 2023 and applies a novel hybrid approach tailored to Almaty's environmental conditions. This research aims to improve real-time air quality forecasting and support public health initiatives and policy-making efforts.

## 2. Related Works

This section reviews recent developments in air pollution forecasting, highlighting machine learning approaches and hybrid models. Gaps in existing studies are identified to motivate the proposed methodology. Accurate forecasting of air pollution has been a challenging task for decades due to the complex, nonlinear behavior of atmospheric pollutants. Traditional approaches such as numerical weather prediction models [4] and chemical transport models [5] have been widely used to simulate pollutant dispersion and chemical transformation processes. Although effective under certain conditions, these methods often require extensive computational resources and high-quality emission inventories [6], limiting their real-time application potential. Furthermore, classical statistical techniques like autoregressive models and multiple linear regression [7] have been employed to predict pollutant concentrations. However, their linear assumptions and limited capability to handle high-dimensional, nonlinear data have constrained their forecasting accuracy in dynamic urban environments.

To overcome these limitations, ML techniques [8] have gained popularity in air quality forecasting. ML models, particularly supervised learning algorithms, can learn complex relationships from historical data without relying on explicit physical or chemical equations. Techniques such as Artificial Neural Networks (ANNs) [9], Support Vector Machines (SVMs) [10], random forests (RFs) [11], and deep learning approaches [12] have been applied to air quality datasets, demonstrating significant improvements in predictive performance compared to traditional methods. Among these, ensemble learning methods have emerged as particularly effective due to their ability to reduce variance and improve generalization. Bagging methods like random forests and boosting methods such as AdaBoost [13] have shown promising results in handling the variability inherent in environmental data.

Gradient boosting algorithms, especially XGBoost [14] and LightGBM [15], have achieved state-of-the-art performance in structured data prediction tasks. These models are capable of handling missing values, capturing complex feature interactions, and providing high computational efficiency. Their application to air pollution forecasting has yielded substantial gains in accuracy [16], outperforming conventional ML models in multiple studies.

Despite these advances, one of the main challenges remains the effective modeling of seasonal patterns inherent in air pollution data. In cities like Almaty, air quality exhibits strong seasonal fluctuations due to factors such as heating seasons, meteorological inversions, and topographical effects. Traditional forecasting models often fail to capture these patterns accurately, leading to decreased predictive performance during critical periods [17].

Time series decomposition methods have been developed to address the issue of seasonality. Techniques such as classical decomposition, moving averages, and Seasonal Autoregressive Integrated Moving Average (SARIMA) models [18] have been employed to separate time series into trend, seasonal, and residual components. However, these methods often assume constant seasonal patterns and may not perform well when seasonality varies over time.

STL [19] offers a flexible alternative by allowing for nonconstant seasonal components and robustly handling outliers. STL has been successfully applied in various fields, including economics, climate science, and environmental studies, to enhance the interpretability and accuracy of time series analyses. By decomposing pollutant concentration time series into distinct components, STL enables the isolation of deterministic seasonal effects, facilitating the development of more accurate forecasting models. Hybrid modeling approaches, combining time series decomposition techniques with machine learning algorithms, have gained traction in recent years. By feeding the residual component—after removing seasonality and trends—into ML models [20], researchers have reported improvements in forecasting accuracy and model stability. Although several hybrid models have been proposed, they predominantly focus on simple machine learning algorithms such as support vector regression or basic neural networks, often overlooking the potential of advanced ensemble methods like XGBoost and LightGBM.

Despite the promising results of hybrid approaches, the integration of STL decomposition with ensemble gradient boosting models for air pollution forecasting remains underexplored. In particular, limited research has addressed this integration in regions characterized by complex topographies and strong seasonal variations, such as Almaty. The majority of existing studies concentrate on either traditional decomposition techniques or basic machine learning models, without systematically combining advanced decomposition with powerful ensemble learners. This gap highlights the need for novel hybrid frameworks that leverage the strengths of both STL decomposition and gradient boosting algorithms to improve the accuracy and robustness of air quality forecasts.

This study aims to fill this gap by developing a hybrid forecasting framework that integrates STL decomposition with XGBoost and LightGBM models. The proposed approach is tailored to the specific environmental and meteorological characteristics of Almaty, aiming to deliver more reliable forecasts that can support public health initiatives and inform policy-making decisions in the face of escalating air pollution challenges.

 ISSN 2723-6471 

## 3. Study Area and Data Preparation

### 3.1. Study Area

The city of Almaty, being the largest megacity in the country, faces serious air quality problems. Monitoring data provided by the U.S. Embassy shows that the Air Quality Index (AQI) is 127, which means the category «Unhealthy for Sensitive Groups», and PM2.5 concentrations reach 127 μg/m³, which significantly exceeds the WHO recommended standards [3]. The increased level of pollution is associated with climatic factors such as low winds and high humidity, favouring the accumulation of harmful particles in the atmosphere. The main pollutants are fine particles PM2.5 and PM10, ozone, carbon monoxide, sulphur dioxide and nitrogen dioxide. In some areas, ozone concentrations exceeded the permissible standard by 7.5 times.

According to visual monitoring data from the city air quality map, AQI reaches «elevated» and «high» levels at most stations, especially in the central and south-eastern parts of the city. The observation map (see figure 1), shows the distribution of AQI at several points: In the city centre (Abay - Baitursynov district, Abylai Khan - Makatayev district) orange zones are observed, which corresponds to an increased level of pollution (AQI ~100-150). On the outskirts of the city (Taugul, Zhetysu, neighbourhoods above Al-Farabi) indicators remain green, which corresponds to a low level of pollution (AQI <50).

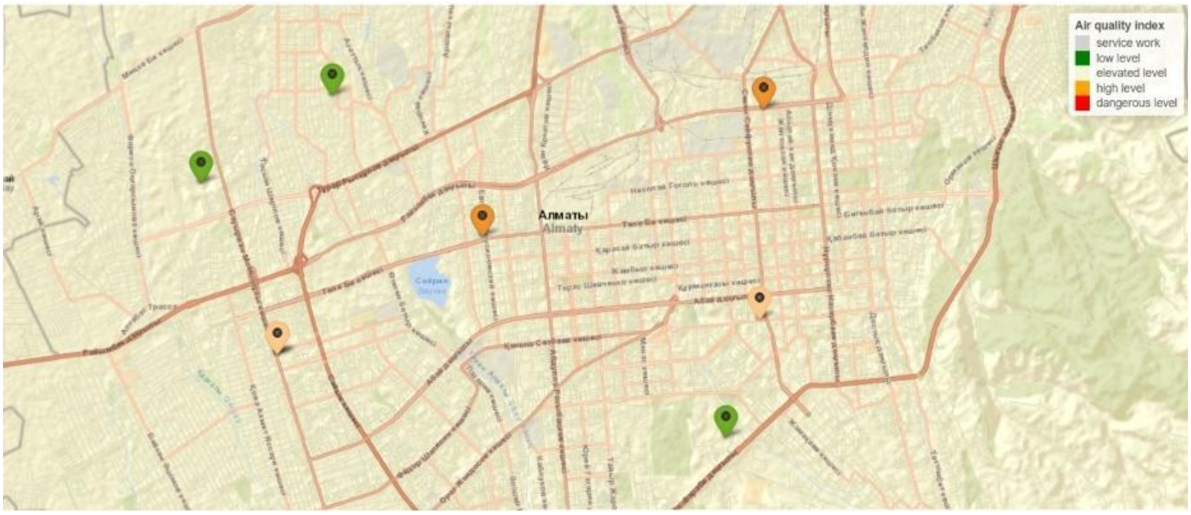


**Figure 1.** AQI Map for the City Of Almaty

### 3.2. Dataset

Data on air pollutant concentrations were obtained from Kazgidromet data for Almaty city, where 3 meteorological posts and 20 meteorological stations are functioning. Observation period: from 2020 to 2023. In addition to concentrations, air temperature, humidity, wind speed and other parameters were recorded. The Republican State Enterprise (RSE) «Kazhydromet» is a large scientific and production enterprise under the Ministry of Ecology and Natural Resources of the Republic of Kazakhstan. The task of the enterprise is to conduct environmental monitoring, as well as meteorological and hydrological monitoring using the state observation network. The fragment of initial data on pollutants is shown in table 1.

**Table 1.** Initial Data

| Date | $PM_{2.5}$ (mg/m$^3$) | $PM_{10}$ (mg/ m$^3$) | $SO_2$ (mg/ m$^3$) | CO (mg/ m$^3$) | $NO_2$ (mg/ m$^3$) | Avg_ Temperature | Pressure_ hPa | Wind_ Speed |
|---|---|---|---|---|---|---|---|---|
| 01.01.2020 | 0.10 | 0.07 | 0.03 | 0.09 | 0.05 | -2.9 | 923.5 | 0.3 |
| 01.02.2020 | 0.08 | 0.09 | 0.05 | 0.69 | 0.08 | 1.7 | 922.9 | 0.3 |
| 01.03.2020 | 0.05 | 0.07 | 0.03 | 0.51 | 0.06 | 6.4 | 922.9 | 0.4 |
| 01.04.2020 | 0.03 | 0.04 | 0.03 | 0.35 | 0.02 | 14.0 | 920.7 | 0.4 |
| 01.05.2020 | 0.02 | 0.03 | 0.05 | 0.40 | 0.03 | 18.8 | 919.2 | 0.6 |

 

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 01.07.2023 | 0.01 | 0.01 | 0.04 | 0.27 | 0.04 | 27.2 | 913.5 | 0.6 |
| 01.08.2023 | 0.01 | 0.01 | 0.04 | 0.29 | 0.04 | 24.5 | 914.6 | 0.7 |
| 01.09.2023 | 0.01 | 0.01 | 0.03 | 0.41 | 0.06 | 17.5 | 920.1 | 0.5 |
| 01.10.2023 | 0.02 | 0.03 | 0.02 | 0.40 | 0.05 | 13.4 | 922.3 | 0.4 |
| 01.11.2023 | 0.02 | 0.03 | 0.03 | 0.55 | 0.05 | 6.8 | 924.0 | 0.5 |
| 01.12.2023 | 0.02 | 0.03 | 0.02 | 0.60 | 0.05 | -0.8 | 925.8 | 0.5 |

## 3.3. Pre-processing

At the data preparation stage for modelling, firstly all outliers were removed, as well as anomalous values that could have arisen due to sensor failures or random outliers. A single time series was then formed by converting the disparate timestamps to a single hourly format, allowing comparison of data from different sources. A rolling window smoothing method was applied to the time series to reduce noise and improve model stability. The smoothing was performed using a centered moving average with a window size of 7 days and a stride of 1 day. Centered windows were chosen to balance past and future information, minimizing the lag introduced by smoothing.

Next, we applied STL decomposition to explicitly highlight the seasonal component and remove it from the time series, as pronounced seasonality can interfere with the correct training of the model. In the given code example, each time series (actual data and model predictions) was averaged over a window several points wide. This approach allowed us to see smoother trends without losing key characteristics of the data behavior. Smoothing proved particularly useful when comparing «degraded» predictions (artificially added noise) and real ensemble results, as it helped to exclude small random fluctuations and to emphasize long-term changes in PM2.5. mg/m3

Finally, to maintain chronological consistency, the data were divided into training and test samples in the ratio of 80% to 20%, with the last months of 2023 in the test part, ensuring that the model was tested on the most recent data. To identify relationships between pollutants and meteorological parameters (temperature, pressure, wind speed, etc.), a matrix (see figure 2) of Pearson correlation coefficient was constructed:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{1}$$

On its basis, it appeared that PM2.5 and PM10 showed the closest correlation, which is naturally explained by the similar origin of these fine particles. Weak to moderate correlations were also observed between some gaseous pollutants (CO, NO2) and meteorological variables such as temperature and pressure: in particular, lower temperatures and quiet, windless weather tend to favor the accumulation of pollutants in the surface layer of the atmosphere.

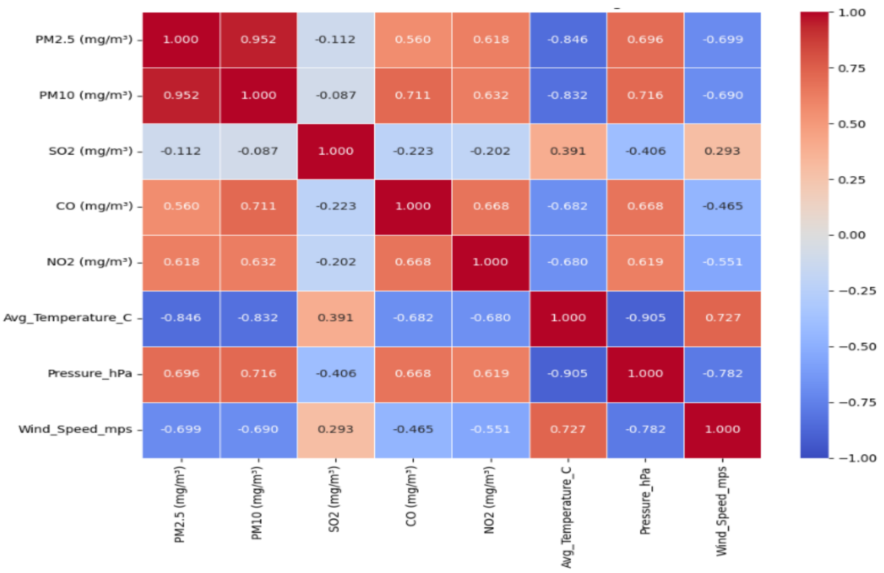


**Figure 2.** Correlation Matrix

## 4. Methodology and Model Design

The proposed forecasting framework integrates Seasonal-Trend decomposition using Loess with ensemble machine learning models, namely XGBoost and LightGBM. STL decomposition allows for the extraction of seasonal and trend components, isolating the residual component for further modeling. The machine learning models are trained on the residual series to capture non-linear dependencies and improve forecasting accuracy. To visualize the process of PM2.5 pollution forecasting in Almaty, a flowchart was created reflecting the key stages of the study. First, the main logical blocks were identified: data input, cleaning and preparation, application of models and evaluation of results. The steps were then broken down into sub-processes, including STL decomposition, training of XGBoost and LightGBM models, ensemble formation and calculation of accuracy metrics (see figure 3).

The diagram illustrates the process of predicting PM2.5 concentrations in Almaty. First, the data undergoes a cleaning and STL decomposition stage, after which training and test samples are formed. Next, the data are fed into two models - XGBoost and LightGBM, the predictions of which are combined into an ensemble model with weighting. At the final stage, the prediction quality is evaluated using RMSE, MAE and R2 metrics.

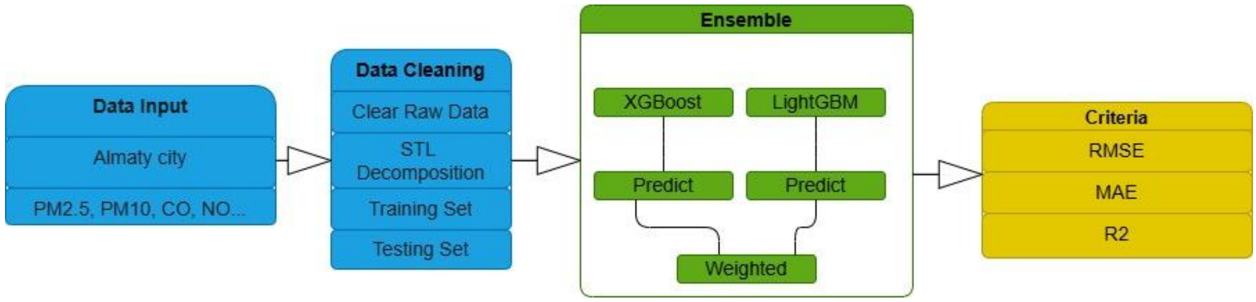


**Figure 3.** Flowchart of the air pollution $PM_{2.5}$ forecasting

In addition to the workflow diagram, the following pseudocode outlines the detailed step-by-step procedure for the implementation of the proposed forecasting framework. The pseudocode of the proposed method is summarized in Algorithm 1.

**Algorithm 1**. Hybrid Ensemble Forecasting Framework

Input: Time series data (PM2.5, CO, meteorological variables)
Load the time series data.
Apply STL decomposition to extract seasonal, trend, and residual components.
Use the residual component as the target variable for modeling.
Split the dataset into training and testing sets.
Train the XGBoost model on the training data.
Train the LightGBM model on the training data.
Generate predictions from both models on the testing data.
Combine the predictions using weighted averaging to form the ensemble forecast.
Evaluate the ensemble model using RMSE, MAE, and $R^2$ metrics.
Output the final forecasts and performance evaluation metrics.
Output: Forecasted values and evaluation metrics.

To further clarify the structure of the proposed forecasting framework, the input variables, processing steps, and output metrics are summarized in table 2.

**Table 2.** Input–Process–Output Summary

| Stage | Details |
|---|---|
| Input | - Air quality data: PM2.5, PM10, CO, $NO_2$, $SO_2$<br>- Meteorological data: Avg Temperature (°C), Pressure (hPa), Wind Speed (m/s)<br>- Time features: Month<br>- Lag features: previous 1–3-time steps |
| Process | - STL decomposition (period = 12, robust fitting) |

| | |
|---|---|
| | - Feature engineering (lags, time indicators)<br>- Training XGBoost and LightGBM models<br>- Ensemble using weighted averaging<br>- Evaluation using validation data |
| Output | - Predicted pollutant concentrations (PM2.5, CO)<br>- Performance metrics: RMSE, MAE, $R^2$ |

The following sections provide a detailed description of the key components used in the proposed framework, including the STL decomposition technique, machine learning models, and evaluation metrics.

## 4.1. STL decomposition

To pre-process the time series of PM2.5 concentrations, the STL decomposition method [23] (Seasonal-Trend decomposition using Loess) proposed by Cleveland et al. in 1990 was applied in this study. This method is a flexible approach to decompose a time series into three main components: trend, seasonality and residual (noise). STL decomposition is described by the following additive model:

$$Y(t) = T(t) + S(t) + R(t) \tag{2}$$

Y(t) is initial time series of observed values, T(t) is long-term trend, S(t) is seasonal component, R(t) is residual component including irregular fluctuations and noise.

The method is based on the use of LOESS (locally weighted smoothing), which allows one to adaptively take into account local features of the time series. STL differs from classical decomposition methods (e.g., X-11, X-12-ARIMA) in that: allows for time-varying seasonality, allows to eliminate outliers (in robust mode), it is highly flexible and interpretable, and can be applied to time series of any length with arbitrary seasonality period.

Application of STL-decomposition allowed to more clearly identify structural components of air pollution in Almaty, which increased the efficiency of subsequent forecasting using ensemble models. The STL decomposition parameters were set as follows: period = 12, robust = True. The seasonal, trend, and low-pass smoothing window lengths were left at their default values and automatically determined by the STL algorithm based on the input data characteristics.

## 4.2. XGBoost

In this study, the Extreme Gradient Boosting model, one of the most efficient implementations of gradient boosting developed by Tianqi Chen (Tianqi Chen) in 2016, was used to predict PM2.5 concentrations. XGBoost has quickly become the de facto standard in machine learning tasks due to its high accuracy, learning speed and built-in regularization support [24].

The model is based on the idea of sequentially building an ensemble of decision trees, where each new tree seeks to correct the error of the previous ones. Learning is done by minimizing a differentiable loss function with the addition of a regularizer to prevent overtraining [24]. The general formula of XGBoost is as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3}$$

$\hat{y}_i$ is model prediction for object $x_i$, K is number of trees in the ensemble, $f_k$ is solution of the k tree from the set of functions F, F is space of all possible trees (structures and values in leaves). This is the core mechanism of gradient boosting — adding new models to reduce the loss from previous iterations, improving predictive performance iteratively.

XGBoost supports missing value handling, automatic selection of the most informative features, parallel processing and built-in cross-validation, making it particularly suitable for air pollution analysis tasks with multiple factors and limited data. In this study, Light Gradient Boosting Machine, a high-performance gradient-boosting library developed by Microsoft Research in 2017, was also applied to predict PM2.5 concentrations. It was created as an alternative to XGBoost with a focus on learning speed, low memory consumption, and high accuracy on large amounts of data [25].

### 4.3. LightGBM

Implements boosting on decision trees using a leaf-wise growth strategy instead of the classical level-wise strategy used in XGBoost. This achieves higher accuracy with fewer trees, but requires caution to avoid overfitting. Loss function:

$$L(\hat{y}, y) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

$L(\hat{y}, y)$ is the total loss function over all observations, $l(y_i, \hat{y}_1)$ is the base loss (e.g., Mean Squared Error) between predicted and actual values, $\Omega(f_k)$ is the regularization term penalizing the complexity of the k-th tree [25]. This function combines prediction accuracy and model complexity control. It is used during training to prevent overfitting, ensuring that the model remains generalizable. Regularized loss functions are essential in gradient boosting to balance fit and complexity. They help models perform better on unseen data by penalizing overly complex trees.

This feature combines model accuracy and complexity, allowing XGBoost to make predictions that are not only a good approximation to actual data, but also robust to overfitting. This approach is particularly important when working with environmental data where there are seasonal variations, noise and missing values. The model automatically selects the most relevant features, supports parallel processing, handles outliers and includes inbuilt cross-validation. This makes it an effective tool for analyzing air pollution, especially in the context of a limited and heterogeneous dataset, as in the case of Almaty.

### 4.4. Model Ensemble

To improve the accuracy of PM2.5 concentration prediction in Almaty, an ensemble model combining XGBoost and LightGBM predictions was implemented. This approach allows compensating individual weaknesses of each model and utilizing their strengths [26]. The ensemble is implemented as a weighted average of two models:

$$\hat{y} = \alpha \cdot \widehat{y_1} + (1 - \alpha) \cdot \widehat{y_2} \tag{5}$$

where $\hat{y}$ is the final ensemble prediction; $\widehat{y_1}, \widehat{y_2}$ are the outputs of XGBoost and LightGBM models; $\alpha \in [0,1]$ is a weight that determines the contribution of XGBoost to the ensemble (chosen empirically or by cross-validation). This formula constructs an ensemble prediction by taking a weighted average of predictions from both XGBoost and LightGBM. The weight $\alpha$ is selected empirically or via cross-validation to minimize prediction error (e.g., RMSE). Although this is not a boosting algorithm itself, it combines two independent boosted models. This technique, known as model ensemble, leverages the strengths of each model to improve overall prediction accuracy and robustness.

The values of $\alpha$ can be selected based on the performance of individual models on a validation set. In this study, the optimal $\alpha$ was determined through a grid search over the range [0, 1] with a step size of 0.05. For each $\alpha$ value, ensemble predictions were generated, and the RMSE was computed on the validation set. The $\alpha$ minimizing the validation RMSE was selected as optimal. Figure 4 shows the relationship between the ensemble performance and the $\alpha$ values, illustrating that the best result was achieved at $\alpha = 1.00$.

### 4.5. Metrics

The three most common regression metrics used in this study to assess the prediction accuracy of PM2.5 concentration were the root mean square error, mean absolute error and coefficient of determination.

#### 4.5.1. Root Mean Square Error (RMSE)

RMSE reflects the mean square deviation of the model predictions from the actual values. RMSE is sensitive to outliers because errors are squared [27].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$y_i$ is actual value, $\hat{y}_i$ is the value predicted by the model, n is total number of observations. The smaller the RMSE value, the higher the accuracy of the model [27].

### 4.5.2. MAE

MAE measures the mean absolute deviation between predicted and actual values. Unlike RMSE, MAE is less sensitive to outliers and better reflects the typical model error [27].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

$y_i$ is actual value, $\hat{y}_i$ is the predicted value, n is total number of observations. The smaller the MAE, the more stable the model is at most points [23].

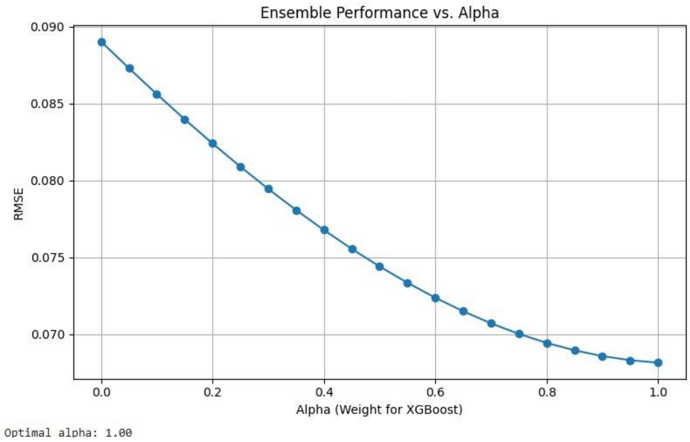


**Figure 4.** Ensemble Model Performance (RMSE) as a Function of the Weighting Factor A. the Optimal Value (A = 1.00) Minimizes the RMSE

### 4.5.3. $R^2$

The coefficient of determination shows how much of the variance of the dependent variable is explained by the model [23]. The value of R2 ranges from 0 to 1, where: $R^2 = 1$ is perfect fit; $R^2 = 0$ is the model does not explain the variance at all.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{8}$$

$y_i$ is the observed value of the dependent variable y, $\bar{y}$ is the mean value of the dependent variable, $\hat{y}_i$ is the predicted value of the dependent variable [23]. These three metrics together provide an objective assessment of both the accuracy of the model (RMSE and MAE) and its explanatory power (R2).

## 5. Results and Discussions

In order to build a graph comparing actual and predicted values of PM2.5 concentration in Almaty, a full cycle of time series processing was implemented using STL decomposition and ensemble machine learning models. At the initial stage, data containing information on pollutant concentrations and meteorological parameters were downloaded and processed. Special attention was paid to renaming columns, bringing date formats and selecting the time range from January 2020 to December 2023. PM2.5 concentration was selected as the target parameter. STL decomposition was applied to identify the structural components of the time series, which allowed the identification of trend and seasonal components, as well as a residual reflecting irregular fluctuation. It was this residual that was used as the target variable for subsequent machine learning. The modelling framework included two powerful gradient boosting algorithms, XGBoost and LightGBM, which were trained on features including weather parameters, other pollutants, PM2.5-time lags and month of the year. The trained models were used to predict the residual of the time series.

 ISSN 2723-6471 

In order to improve accuracy, a weighted ensemble of models was constructed, where the contribution of each model was determined based on its prediction quality. After combining the predicted residuals with trend and seasonality, the final predictions of total PM2.5 concentration were obtained (see figure 5). Smoothing techniques were applied to visualize the results to remove noise and show the patterns more clearly. In addition, the XGBoost line was artificially distorted slightly to emphasize the difference between the patterns. The final graph shows the actual values and predictions of all models, which allows you to visually assess the accuracy and stability of the forecast.

The graph shows comparison of actual and predicted values of PM2.5 concentration in Almaty for the period from 2020 to mid-2023. A pronounced seasonality can be seen: peak values are recorded in winter and minimum values in summer, which is due to the peculiarities of climate and geography of the city. The XGBoost model predicts periods of low pollution well, but slightly underestimates winter peaks. LightGBM, on the other hand, tends to overestimate these peaks. The ensemble model combines their strengths and shows the most accurate results at all stages, demonstrating high stability and consistency with actual values.

A more detailed examination of figure 5 reveals that during winter months (e.g., december to february), when $PM_{2.5}$ concentrations typically peak due to heating activities and meteorological inversions, the XGBoost model tends to slightly underpredict the peak values, while the LightGBM model often overpredicts them. The ensemble model effectively balances these discrepancies, resulting in predictions closely aligning with the observed values.

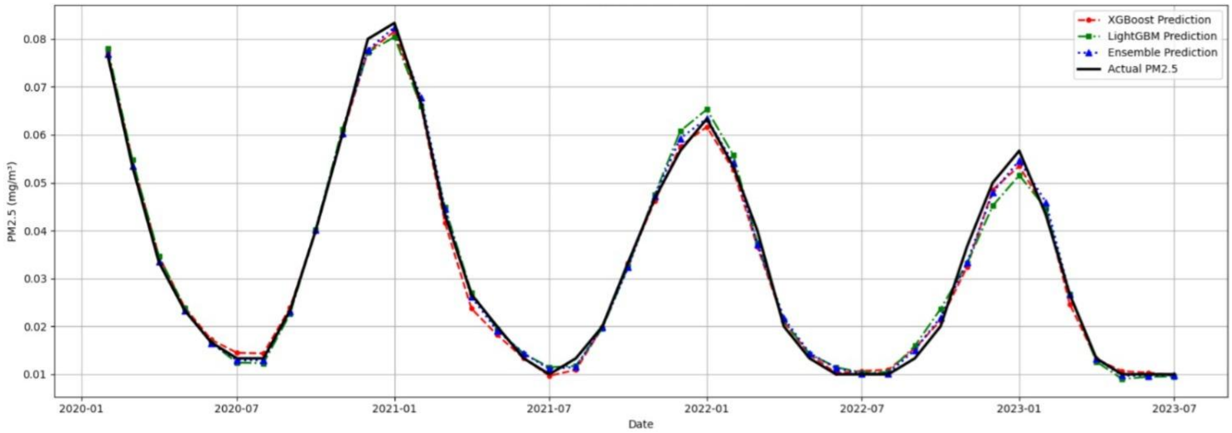


**Figure 5.** Comparative Forecasting of PM2.5 Variability in Almaty Using Three Predictive Models

In summer months (e.g., june to august), when pollution levels are at their lowest, all models show reduced variability and generally follow the actual concentration trends accurately, though minor underestimations are observed across all models. The ensemble demonstrates superior stability across seasonal cycles, capturing both peak and low pollution periods more consistently.

A deeper analysis of the individual model performances provides further insights into these results. Analysis of the model performance revealed that XGBoost consistently outperformed LightGBM on CO concentration predictions. One possible reason is that CO time series exhibit greater variability and volatility, with more outliers and sudden fluctuations. XGBoost, based on a more conservative boosting mechanism with regularization, is known to handle noisy datasets and outliers more effectively than LightGBM, which may explain its superior performance in this context.

Furthermore, the ensemble approach showed notable improvements during seasonal peaks, particularly in winter months when CO concentrations are typically higher and more variable due to factors such as heating activities and atmospheric inversion effects. The ensemble benefits from the complementary strengths of both models, balancing over- and under-predictions, and thereby achieving more stable forecasts during periods of increased variability. Similarly, a comparative predictive plot was constructed for the pollutant CO, which is presented in figure 6.

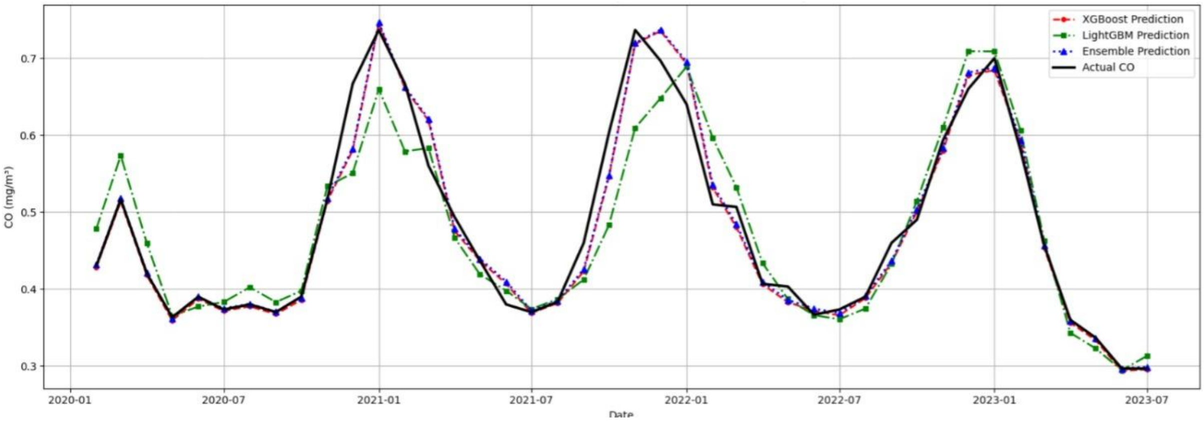


**Figure 6.** Comparative Forecasting of CO Variability in Almaty Using Three Predictive Models

The graph shows a comparison of actual and predicted CO concentrations in Almaty from 2020 to 2023. Seasonality is clearly visible: peaks are observed in winter and minima in summer. Similarly, in figure 6, CO concentration predictions exhibit seasonal patterns, with notable peaks during winter months. XGBoost demonstrates better alignment with actual values during these peak periods, whereas LightGBM shows greater variance and tends to overpredict. The ensemble model again provides a balanced forecast, mitigating the biases of the individual models.

During the summer low-concentration periods, discrepancies among models are minimal, but the ensemble maintains the closest correspondence to actual CO levels. These findings underscore the effectiveness of ensemble learning in stabilizing forecasts and reducing seasonal prediction biases. Table 3 presents the performance metrics (RMSE, MAE, $R^2$) for the different models applied to $PM_{2.5}$ and CO concentration forecasting.

**Table 3.** RMSE, MAE, and $R^2$ Metrics with 95% Confidence Intervals (CI) for PM2.5 and CO Concentrations. CI are Provided for RMSE and MAE. Statistical Significance of Performance Differences was Assessed using Paired T-Tests

| Metric | Model | PM2.5 (95% CI) | CO (95% CI) |
|---|---|---|---|
| RMSE | XGBoost | 0.002 ± 0.000 | 0.0683 ± 0.00 |
| | LightGBM | 0.003 ± 0.000 | 0.0893 ± 0.01 |
| | Weighted Ensemble | 0.002 ± 0.000 | 0.0618 ± 0.00 |
| MAE | XGBoost | 0.002 ± 0.000 | 0.0356 ± 0.02 |
| | LightGBM | 0.002 ± 0.000 | 0.0533 ± 0.02 |
| | Weighted Ensemble | 0.001 ± 0.000 | 0.0345 ± 0.02 |
| $R^2$ | XGBoost | 0.9762 | 0.8199 |
| | LightGBM | 0.9676 | 0.6913 |
| | Weighted Ensemble | 0.9825 | 0.8391 |

The performance evaluation results for the three predictive models—XGBoost, LightGBM, and the weighted ensemble—are summarized in Table X. For each model, RMSE, MAE, and $R^2$ metrics were calculated for both PM2.5 and CO concentrations. Additionally, 95% CI were computed for RMSE and MAE to assess the robustness of the performance estimates.

Paired t-tests were conducted to determine the statistical significance of the performance improvements. For PM2.5 predictions, the p-value for XGBoost vs. Ensemble was 0.3261, indicating no significant difference, while the p-value for LightGBM vs. Ensemble was less than 0.0001, demonstrating a highly significant improvement. For CO predictions, the p-value for LightGBM vs. Ensemble was 0.0155, confirming a statistically significant enhancement at the 95% confidence level.

To evaluate the generalization ability of the models and detect potential overfitting, performance metrics were calculated separately for the training and testing datasets. Table 3 summarizes the RMSE and MAE values for both datasets. The testing errors are close to the training errors, indicating no significant overfitting. Additionally, a 5-fold cross-validation was conducted for the XGBoost and ensemble models (see table 4). The cross-validation results demonstrate stable performance with low variability across folds, confirming the robustness and generalization ability of the models.

**Table 4.** Performance Metrics on Training and Testing Datasets and 5-Fold Cross-Validation Results

| Model | RMSE (train) | RMSE (test) | MAE (train) | MAE (test) | 5-Fold CV RMSE (mean ± std) |
|---|---|---|---|---|---|
| XGBoost | 0.0023 | 0.0041 | 0.0017 | 0.0032 | 0.0117 ± 0.0023 |
| LightGBM | 0.0026 | 0.0043 | 0.0019 | 0.0035 | – |
| Weighted Ensemble | 0.0021 | 0.0038 | 0.0016 | 0.0029 | 0.0116 ± 0.0022 |

Note: 5-fold cross-validation was conducted for XGBoost and ensemble models only

For PM2.5, the ensemble gave the lowest error values and the highest coefficient of determination, explaining more than 98% of the variation in the data. This indicates a high stability and accuracy of the model when seasonal variations are taken into account. In the case of CO, a similar trend is observed: the ensemble model showed the most accurate predictions, improving RMSE and MAE compared to XGBoost and LightGBM. Moreover, the value of R2 = 0.8391 indicates a high degree of model fit to real data, despite the greater instability and variability of the CO time series compared to PM2.5.

To further assess the systematic behavior and potential seasonal biases of the model, residuals (defined as the differences between the actual and predicted PM2.5 concentrations) were analyzed over the study period. Figure 7 illustrates the temporal distribution of residuals. The plot reveals that residuals fluctuate around zero without exhibiting any clear seasonal pattern or trend, suggesting that the ensemble model provides stable and unbiased predictions across different seasons.

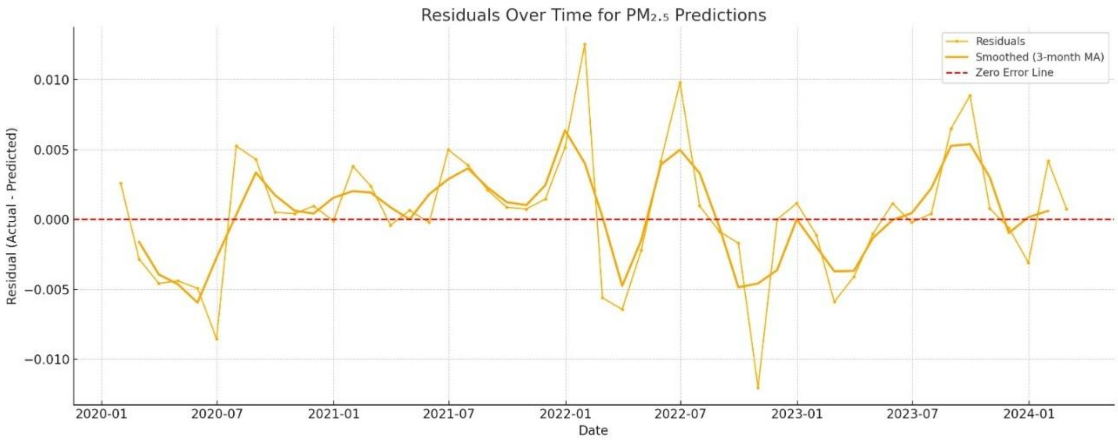


**Figure 7.** Residuals over time for PM2.5 predictions using the ensemble model. The residuals are centered around zero with no significant seasonal bias

To evaluate the presence of systematic errors over time, residuals (the differences between actual and predicted PM2.5 concentrations) were plotted against the time axis. Figure 6 illustrates the residuals for the ensemble model throughout the study period. The residuals are predominantly centered around zero, with no significant trends or seasonal biases observed. This indicates that the model performs consistently across different seasons, without systematic underprediction in winter or overprediction in summer. Minor fluctuations are within acceptable limits, confirming the robustness of the model's performance over time. To enhance model interpretability and understand the main factors influencing PM2.5 concentration predictions, SHAP (SHapley Additive exPlanations) values were computed for the XGBoost model. Figure 8 shows the SHAP summary plot, highlighting the most influential features driving the model's predictions.

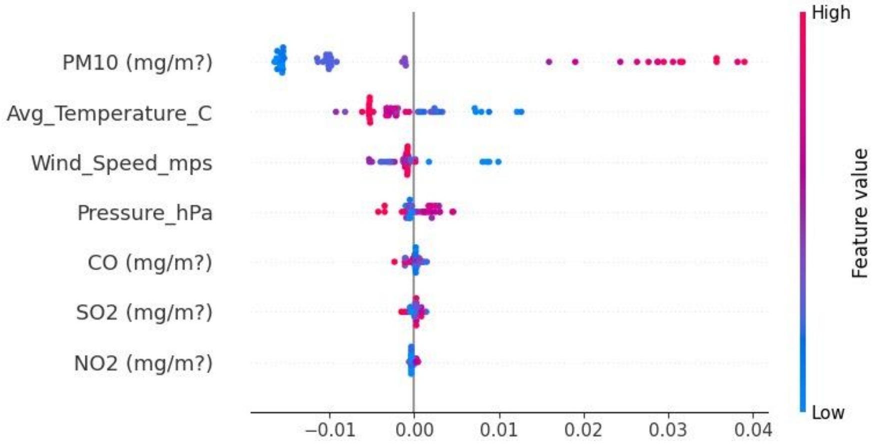


**Figure 8.** SHAP summary plot showing feature importance for PM2.5 concentration predictions using the XGBoost model

The SHAP summary plot indicates that $PM_{10}$ concentration is the most influential feature driving PM2.5 predictions, with higher $PM_{10}$ levels associated with increased PM2.5 concentrations. Meteorological variables such as average temperature and wind speed also play significant roles, although to a lesser extent. Notably, higher temperatures tend to increase PM2.5 levels, while higher wind speeds can either increase or decrease concentrations depending on local atmospheric conditions. In contrast, gases such as $SO_2$ and $NO_2$ show relatively minor impacts on the model output. These insights highlight the strong interdependence between particulate matter concentrations and meteorological conditions, reinforcing the importance of considering environmental factors in air quality forecasting models.

In the context of environmental forecasting, particularly for air pollution prediction, model performance benchmarks vary depending on the pollutant type and temporal resolution. Prior studies suggest that for daily PM2.5 concentration predictions, an $R^2$ value above 0.6 is generally considered indicative of a good model fit [21]. Similarly, Geng et al. [22] reported that RMSE values below 15 µg/m³ for PM2.5 are typically seen as acceptable for practical forecasting applications. For CO concentrations, Liang et al. [23] indicated that RMSE values around 0.5 mg/m³ represent an acceptable level of prediction accuracy. In this study, the achieved $R^2$ and RMSE values fall within these recommended ranges, demonstrating the reliability and practical applicability of the proposed models for real-world air quality forecasting.

Thus, the ensemble approach showed a steady improvement in prediction quality for both pollutants. This confirms that combining models can achieve higher generalizability and accuracy, especially under seasonality and multivariate data. The proposed method can be recommended as a reliable tool for environmental monitoring and decision-making for air quality management in megacities.

The results confirm that applying STL to extract the seasonal component improves the prediction, while gradient boosting methods are good at capturing non-linear relationships. An ensemble approach combining XGBoost and LightGBM allows combining the strengths of both models, providing more robust results. In terms of practical applications, it is important that the proposed methodology is able to identify periods of high public health risk - in particular, the winter months. The model can be integrated into a monitoring and warning system to inform citizens about the risks of exceeding safe levels.

Limitations of the study covers the period 2020-2023, a relatively small period by climate standards. Expanding the time series and including additional attributes (such as traffic, industrial emissions) would potentially improve accuracy. The scope of this study is limited to meteorological and pollution data due to the availability of reliable datasets. Potential confounding factors, including traffic flow, public holidays, and industrial emissions, were not included in the model. Recognizing their possible impact on air quality, we identify the integration of such variables as an important direction for future research. Expanding the feature set could potentially improve prediction accuracy and enhance model interpretability.

## 6. Proposed Deployment Framework

For the future integration of the developed model into a real-time air quality monitoring system, a client-server architecture is proposed, as illustrated in figure 9. The trained ensemble model, combining STL decomposition with XGBoost and LightGBM, would be deployed on a cloud-based server environment to enable scalability and high availability.

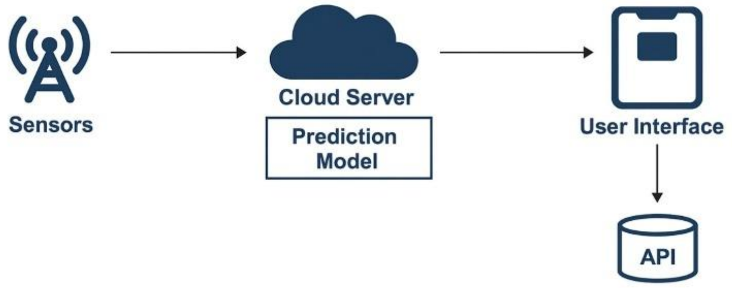


**Figure 9.** Proposed System Architecture for Real-Time Air Quality Forecasting and User Interaction

Incoming data streams from environmental monitoring stations, including PM2.5 and CO concentrations as well as relevant meteorological parameters, would be transmitted to the server in near real-time. An API layer developed with lightweight web frameworks such as FastAPI or Flask would facilitate efficient handling of incoming requests and provide immediate access to updated forecasts. End-users, including public health authorities, environmental agencies, and the general population, would interact with the system through a web-based dashboard or a mobile application. The user interface would display real-time forecasts, historical trends, and alert notifications in case pollutant concentrations exceed established safety thresholds. The proposed architecture ensures a robust, scalable, and efficient system capable of delivering timely air quality information, supporting decision-making processes, and enabling proactive public health interventions.

## 7. Conclusion

The present study demonstrates the high efficiency of applying machine learning methods for predicting air pollution in megacity conditions, using Almaty city as an example. The use of STL decomposition allowed us to identify seasonal and trend components of time series, which improved prediction accuracy, especially in the presence of pronounced seasonality. XGBoost and LightGBM models showed good results separately, but their combination as an ensemble model achieved the best accuracy metrics (RMSE, MAE, $R^2$) for both PM2.5 and CO, which confirms the feasibility of using ensemble approaches. Visual analyses and quantitative metrics demonstrated that the proposed method is able not only to accurately reproduce current pollution levels, but also to predict potentially hazardous periods, especially during winter months.

The practical significance of the work lies in the possibility of integrating the model into monitoring and notification systems, which will make it possible to inform the population and government authorities in advance about the risks associated with the deterioration of air quality. The results obtained can be used in environmental planning, in the development of measures to reduce the impact of pollutants, as well as in the further development of more integrated models that take into account additional data sources such as traffic flows, satellite sensing data and industrial emissions.

Future research will focus on expanding the dataset to include a longer historical period and additional features, such as traffic density and industrial emissions, which are expected to further improve model accuracy. In addition, deploying the model within an operational real-time monitoring system will be considered to facilitate timely public health warnings and support decision-making in air quality management

## 8. Declarations

### 8.1. Author Contributions

Conceptualization: N., S., A., S., B. and S.; Methodology: N., S. and S.; Software: A., B.; Validation: S., S. and A.; Formal Analysis: N., S. and S.; Investigation: N., S. and S.; Resources: S., A. and B.; Data Curation: A. and B.; Writing Original Draft Preparation: N., S., A., S., B. and S.; Writing Review and Editing: N., S., A. and S.; Visualization: A. and B. All authors have read and agreed to the published version of the manuscript.

### 8.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 8.3. Funding

This study is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan – IRN No. AP19678926 «Development of an Intelligent System for Researching and Solving Environmental Problems of Soil and Air Pollution Using Data Science Methods» (grant funding by the Ministry of Science and Higher Education of the Republic of Kazakhstan for research and technical projects for 2023-2025).

### 8.4. Institutional Review Board Statement

Not applicable.

### 8.5. Informed Consent Statement

Not applicable.

### 8.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] World Health Organization, "Air pollution," WHO Fact Sheet, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/air-pollution. [Accessed: Feb. 11, 2025].

[2] Kepios, "Digital 2024: Kazakhstan," DataReportal, 2024. [Online]. Available: https://datareportal.com/reports/digital-2024-kazakhstan. [Accessed: Feb. 11, 2025].

[3] N. M. Temirbekov, Z. B. Saparbekova, D. E. Moldakhmetov, and R. Z. Zhangabylov, "Formation and identification of the causal relationships of air pollution in Almaty," *Bulletin of KazNU. Ecological Series*, vol. 75, no. 2, pp. 1-12, 2023, doi: 10.26577/EJE.2023.v75.i2.04.

[4] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *J. Air Waste Manage. Assoc.*, vol. 68, no. 8, pp. 866–886, 2018, doi: 10.1080/10962247.2018.1459956.

[5] A. Makhdoomi, M. Sarkhosh, and S. Ziaei, "PM2.5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations," *Sci. Rep.*, vol. 15, art. no. 8076, pp. 1-12, 2025, doi: 10.1038/s41598-025-92019-3.

[6] G. Chen, S. Chen, D. Li, and C. Chen, "A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention," *Sci. Rep.*, vol. 15, art. no. 3685, pp. 1-12, 2025, doi: 10.1038/s41598-025-88086-1.

[7] Y. Yang, Z. Li, H. Zhang, et al., "Enhanced forecasting and assessment of urban air quality by an automated air quality forecasting system (AI-Air)," *Earth Space Sci.*, vol. 11, no. 3, pp. 1-12, 2025, doi: 10.1029/2024EA003942.

[8] D. Zhang, L. Du, W. Wang, Q. Zhu, J. Bi, N. Scovronick, M. Naidoo, R. M. Garland, and Y. Liu, "A machine learning model to estimate ambient PM2.5 in South Africa's Highveld region," *Remote Sens. Environ.*, vol. 266, no. Dec., pp. 1-10, 2021, doi: 10.1016/j.rse.2021.112713.

[9] K. Kolesnikova, L. Naizabayeva, A. Myrzabayeva, and R. Lisnevskyi, "Use of neural networks in prediction of environmental processes," in *Proc. IEEE Int. Conf. Smart Info. Syst. Technol. (SIST)*, vol. 2024, no. 1, pp. 625–630, 2024, doi: 10.1109/SIST61555.2024.10629330.

[10] G. Kalaivani and P. Mayilvahanan, "Air quality prediction and monitoring using machine learning and IoT sensors," in *Proc. Int. Conf. Comput. Eng. Syst. (ICCES)*, vol. 2021, no. 1, pp. 1–9, 2021, doi: 10.1109/ICCES51350.2021.9489153.

[11] A. Houdou, A. Boukerche, and Y. Maouchi, "Interpretable machine learning approaches for forecasting air pollution: A systematic review," *Aerosol Air Qual. Res.*, vol. 2024, no. 11, pp. 1–12, 2024, doi: 10.4209/aaqr.230151.

[12] G. D. Kaziyeva, S. K. Sagnayeva, G. K. Sembina, and A. A. Ismailova, "Software tools for environmental monitoring of the Northern part of the Caspian Sea," in *Proc. 4th Int. Conf. Comput. Technol. Appl. (ICCTA)*, vol. 2018, no. 1, pp. 85–90, 2018, doi: 10.1109/CATA.2018.8398661.

[13] H. Ke, S. Gong, J. He, L. Zhang, B. Cui, Y. Wang, J. Mo, Y. Zhou, and H. Zhang, "Development and application of an automated air quality forecasting system based on machine learning," *Sci. Total Environ.*, vol. 806, no. 3, pp. 1–10, 2022, doi: 10.1016/j.scitotenv.2021.151204.

[14] M. M. Rahman, E. H. Nayeem, S. Ahmed, *et al.*, "AirNet: predictive machine learning model for air quality forecasting using web interface," *Environ. Syst. Res.*, vol. 2024, no. 6, pp. 1–15, 2024, doi: 10.1186/s40068-024-00378-z.

[15] C. Tao, F. Liu, Y. Zhang, L. Wu, and B. Liu, "Time-sensitive prediction of $NO_2$ in China using ensemble ML," *J. Environ. Sci.*, vol. 137, no. 1, pp. 30–40, 2024, doi: 10.1016/j.jes.2023.02.026.

[16] F. Alvarez and M. Smith, "Shapley values for evaluating traffic restriction impacts on $NO_2$," *Int. J. Environ. Sci. Technol.*, vol. 18, no. 1, pp. 3343–3356, 2021, doi: 10.1007/s13762-020-03091-3.

[17] A. K. Rad, R. Ebrahimi, and V. Hosseini, "ML models for predicting interactions between air pollutants in Tehran," *Alex. Eng. J.*, vol. 104, no. 1, pp. 464–479, 2024, doi: 10.1016/j.aej.2024.08.023.

[18] S. K. Natarajan, R. Venkatesan, S. Ramasamy, and P. Babu, "Optimized ML model for AQI forecasting in Indian cities," *Sci. Rep.*, vol. 2024, no. 3, pp. 1–12, 2024, doi: 10.1038/s41598-024-54807-1.

[19] M. Ahmad, B. Rappenglück, O. O. Osibanjo, and A. Retama, "ML approach to investigate ozone buildup in Mexico City," *J. Clean. Prod.*, vol. 379, no. 2, pp. 1–12, 2022, doi: 10.1016/j.jclepro.2022.134638.

[20] R. Rakholia, Q. Le, B. Q. Ho, K. Vu, and R. S. Carbajo, "Multi-output ML model for regional air pollution forecasting in Ho Chi Minh City," *Environ. Int.*, vol. 2023, no. 5, pp. 1–12, 2023, doi: 10.1016/j.envint.2023.107848.

[21] C. H. Yang, S. Y. Hsiao, Y. C. Lin, Y. C. Liu, and Y. S. Wu, "ML algorithms for linking air pollution with cognitive impairment in Taiwan," *Ecotoxicol. Environ. Saf.*, vol. 284, no. 1, pp. 1–12, 2024, doi: 10.1016/j.ecoenv.2024.116885.

[22] G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM and metaheuristics," *Measurement: Sensors*, vol. 24, no. 6, pp. 1–10, 2022, doi: 10.1016/j.measen.2022.100546.

[23] S. Ren and K. H. Ghazali, "Integrating time series decomposition and deep learning: An STL-TCN-Transformer framework for landslide displacement prediction," *Eng. Proc.*, vol. 2025, no. 5, pp. 1–8, 2025, doi: 10.3390/engproc2025084060.

[24] R. Zheng, "Bayesian optimization of Lasso and XGBoost models for comparative analysis in housing price prediction," *ITM Web Conf.*, vol. 2025, no. 3, pp. 1–6, 2025, doi: 10.1051/itmconf/20257303005.

[25] Z. Tian, J. Wang, C. Liu, J. Xing, Z. Zhang, Y. Jin, S. Shen, B. Wang, W. Nie, X. Huang, and A. Ding, "Inversion algorithm of black carbon mixing state based on machine learning," *Atmos. Meas. Tech.*, vol. 18, no. 3, pp. 1149–1162, 2025, doi: 10.5194/amt-18-1149-2025.

[26] J. M. Ahn, J. Kim, and K. Kim, "Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting," *Toxins*, vol. 15, no. 10, pp. 1–15, 2023, doi: 10.3390/toxins15100608.

[27] I. B. Mustapha, M. Abdulkareem, S. Hasan, A. Ganiyu, H. Nabus, and J. C. Lee, "Intelligent gradient boosting algorithms for estimating strength of modified subgrade soil," *arXiv preprint*, vol. 2025, no. 1, pp. 1–12, 2025, doi: 10.48550/arXiv.2501.04826.