Article

Machine learning-based approaches to forecast particulate matter in the Istanbul metropolitan area

Journal of Ambient Intelligence and Smart Environments
2026, Vol. 18(2) 214–233

DOI: 10.1177/18761364261422912
journals.sagepub.com/home/ais

Atilla Mutlu[1]

Abstract

Air pollution poses a major global health challenge, with particulate matter (PM) linked to millions of premature deaths each year. This study forecasts PM concentrations in Istanbul's Kartal district using bi-daily observations collected throughout 2022. Four supervised machine learning (ML) models, including support vector machines, random forests (RFs), artificial neural networks, and K-nearest neighbors, were applied using surface meteorological variables and radiosonde-derived inversion parameters. The RF model achieved the highest predictive accuracy, with R^2 values of 0.64 for PM_{10} and 0.70 for $PM_{2.5}$, along with the lowest mean squared error. The study incorporates key enhancements, including the integration of vertical inversion metrics with surface pollutant data, the use of autocorrelation analysis to justify lagged features, and statistical evaluation of model differences using paired *t*-tests. Feature importance analysis showed that inversion thickness and lagged PM levels improved forecasts, highlighting the value of upper-air dynamics and temporal persistence. The aim of this study is to systematically evaluate multiple ML algorithms for PM forecasting at a single urban site. The findings provide transparent, site-specific methodological insights that highlight the role of upper-air dynamics and temporal persistence, offering practical implications for similar urban environments and guiding future multi-site applications.

Keywords

machine learning, SVM, RF, ANN, air quality

Received: 9 May 2025; accepted: 26 January 2026

1 Introduction

Air pollution remains one of the most critical global environmental and public health challenges, intensified by rapid urbanization, industrial activity, and evolving climatic conditions. Particulate matter (PM), particularly PM_{10} (particles $\leq$10 µm) and $PM_{2.5}$ (particles $\leq$2.5 µm), has been extensively studied due to its severe health impacts. These fine particles penetrate deeply into the respiratory system, contributing to cardiovascular and respiratory diseases, asthma, and premature mortality (Bayraktar and Mutlu, 2024; Di et al., 2016; Goudarzi et al., 2021; Hu et al., 2017; Zhang et al., 2017). According to the United Nations International Children's Emergency Fund (UNICEF), citing data from the Health Effects Institute (HEI), air pollution caused ~8.1 million deaths globally in 2021, with over 90% (around 7.8 million) attributed to $PM_{2.5}$ exposure (UNICEF, 2024). These figures underline the urgent need for accurate forecasting tools to support public health interventions and guide environmental policies, particularly in densely populated metropolitan regions.

In urban environments, PM concentrations are shaped not only by emission sources such as vehicle traffic, industrial activities, and residential heating, but also by complex meteorological dynamics. Among these, atmospheric temperature inversion (TI) plays a pivotal role. TI occurs when a warmer air mass overlays cooler surface air, suppressing vertical dispersion and trapping pollutants near the ground, especially during stable winter conditions (Koo et al., 2015; Mbululo

[1]Department of Environmental Engineering, College of Engineering, Balikesir University, Balikesir, Turkey

Corresponding author:
Atilla Mutlu, Department of Environmental Engineering, College of Engineering, Balikesir University, Balikesir 10145, Turkey.
Email: amutlu@balikesir.edu.tr

et al., 2018; Qian and Huang, 2019; Xu et al., 2019). Studies indicate that more than 90% of severe haze episodes in major cities like Beijing are sources of PM_{10} (Xu et al., 2019; Yao et al., 2019), and that the intensity and thickness of inversion layers exert an even stronger influence on pollutant accumulation than inversion frequency (Shao et al., 2023; Yang et al., 2021).

The Kartal district of Istanbul, located on the Asian side of the metropolitan area, represents a relevant case study due to its dense urban structure, mixed industrial and residential land use, and susceptibility to inversion events. Kartal hosts one of Türkiye's eight radiosonde stations, providing valuable high-resolution vertical atmospheric profiles, and also contains a long-established air quality monitoring station (AQMS). These features, combined with frequent pollution events driven by meteorological and topographical constraints, make the area well-suited for investigating the role of inversion dynamics in shaping PM levels (Baltaci, 2017; Karaca et al., 2009; Özdemir et al., 2024; Unal et al., 2011). Despite these characteristics, few studies have integrated inversion parameters into machine learning (ML)-based forecasting frameworks for Istanbul, highlighting the need for localized, data-driven approaches.

Deterministic atmospheric models such as the Weather Research and Forecasting (WRF-Chem) and the Community Multiscale Air Quality (CMAQ) are widely applied for air quality simulations; however, they require detailed emission inventories and substantial computational resources. Alternatively, ML methods offer flexible and efficient data-driven approaches capable of capturing nonlinear and multivariate interactions among meteorological variables, inversion characteristics, and pollution levels, particularly in data-rich yet resource-constrained urban environments. Recent studies have applied artificial neural networks (ANNs), random forests (RFs), support vector machines (SVMs), and K-nearest neighbors (KNNs) for PM forecasting across various urban regions (Alsaber et al., 2023; Biancofiore et al., 2017; Bozdağ et al., 2020; Gao et al., 2024), demonstrating the capability of ML to model complex atmospheric processes. Nevertheless, most of these applications rely primarily on surface meteorological variables and often overlook vertical atmospheric structure, despite its critical role in pollution dynamics.

To address this gap, the present study integrates radiosonde-derived inversion metrics with near-surface meteorological observations to improve PM_{10} and $PM_{2.5}$ forecasting performance. Specifically, this study aims to:

(1) evaluate the predictive performance of SVM, RF, ANN, and KNN models;

(2) quantify the influence of inversion characteristics (e.g. inversion thickness and inversion base/finish heights) on PM variability; and

(3) assess temporal persistence in PM levels using autocorrelation-based lag features.

By combining vertical atmospheric structure with advanced ML techniques, this study provides a scientifically robust and operationally relevant framework for improving particulate matter forecasting in urban environments.

2 Materials and methods

This study utilizes air quality and meteorological data collected in Istanbul's Kartal district from 1 January to 31 December 2022. The dataset includes concentrations of PM_{10} and $PM_{2.5}$, alongside meteorological variables such as temperature, wind speed, relative humidity, and pressure. In addition, inversion parameters, including inversion start and end heights and layer thickness, were derived from twice-daily radiosonde observations. These variables were integrated to train and evaluate four ML models, as outlined in the following subsections.

2.1 Study area and data sources

Kartal is one of Istanbul's oldest industrial regions, located ~20 km east of the city center. It spans an area of about 38 km^2 and has a population exceeding half a million. The district borders Maltepe to the west, Pendik to the east, and the Marmara Sea to the south. Major anthropogenic sources of PM_{10} include vehicular traffic, industrial facilities, and residential heating during winter months. Due to the concentration of industrial establishments in close proximity to residential zones, Kartal experiences emissions from numerous point sources (Unal et al., 2011).

Kartal's proximity to the Marmara Sea affects its local climate and weather patterns. Coastal winds and varying temperature gradients influence pollutant dispersion, while temperature inversions, particularly during winter, trap pollutants close to the ground, exacerbating air quality issues. Seasonal variations further impact pollutant levels; for example, colder months experience higher concentrations of PM_{10} and $PM_{2.5}$ due to increased heating and inversion events.

Considering these factors, Kartal is an optimal study region for investigating the influence of climatic variables and inversion layers on particulate matter concentrations. Understanding the local air pollution dynamics in Kartal is crucial for designing targeted policies and mitigation strategies that address public health risks and improve air quality management.

Moreover, Kartal is one of only eight districts in Türkiye equipped with radiosonde instrumentation, offering high-resolution vertical atmospheric profiles critical for capturing inversion dynamics. This combination of real-world relevance and high-quality data makes Kartal an ideal location for developing and testing air pollution prediction models.

The ambient air quality data for the study period were sourced from the regional air quality monitoring station, AQMS (40.91° N, 29.18° E), and the data were obtained from the network of the Turkish Ministry of Environment and Urbanization (NAQMN, 2023). Meteorological surface observations and radiosonde data were acquired from the Turkish State Meteorological Service (WMO# 17064, as a representative station of Marmara, 40.91° N, 29.15° E). The Kartal Meteorological Station is one of eight stations in the country that operates radiosonde measurements. Radiosondes, systematically deployed with helium balloons bi-daily, furnish a vertical temperature profile and are adept at estimating low-level temperature inversions (Aguilera et al., 2023). Radiosonde measurements were obtained at both 0:00 UTC and 12:00 UTC, and accessible via the Turkish State Meteorological Service Radiosonde Database (TSMS, 2023). Ground observations and radiosonde data were obtained at the Kartal met-station. Kartal was selected for this investigation, as well as in previous studies (Baltaci, 2017; Baykara et al., 2019; Flores et al., 2020; Karaca et al., 2009; Özdemir et al., 2024; Unal et al., 2011; Yavuz, 2023), due to its status as one of the eight authorized meteorological facilities for radiosonde measurement in the country and the presence of the official air quality monitoring station in a suitable location. All datasets were synchronized to bi-daily resolution, matching the radiosonde observations. Spatial coordinates of the stations were used solely for site referencing and map generation, not for geospatial modeling. Both datasets are publicly available and were accessed through the national air quality and meteorological databases.

The data for this study were obtained from two main sources: atmospheric monitoring stations and meteorological factors (highlighted in red), as shown in Figure 1.

The study utilized air quality, meteorological, and upper-atmospheric sounding data collected in the Kartal district of Istanbul over a monitoring period from January 1 to December 31, 2022. Measurements were recorded bi-daily at 00:00 and 12:00 UTC, yielding ~730 observations prior to quality control; after data cleaning and temporal alignment, a final dataset comprising 325 valid observations was retained for model development and evaluation. While the bi-daily resolution limited detailed analysis of diurnal variation, the dataset spans multiple seasons, allowing the models to implicitly account for seasonal trends in pollution dynamics. Each record includes pollutant concentrations such as PM_{10}, $PM_{2.5}$, SO_2, CO, NO, NO_2, NO_x, and O_3 (in $\mu g/m^3$), along with meteorological variables including ambient temperature (t, °C), wind speed (ws, m/s), relative humidity (rh, %), and atmospheric pressure (p, hPa). Additionally, temperature inversion characteristics were derived from radiosonde data and include the inversion start height (inv_start), finish height (inv_finish), and inversion thickness (inv_thic), all in meters.

2.2 Data preprocessing and feature engineering

Data preprocessing is critical to ensure model accuracy and reliability. Several steps were followed as shown in Figure 2.

Data preprocessing followed a systematic workflow involving data cleaning, normalization, synchronization, and feature selection (Figure 2). The process began with formatting the datetime field and removing anomalies such as missing values and duplicates. Less than 3% of values were missing and were imputed using the median for robustness. Outliers were detected using the interquartile range (IQR) method and were either removed or smoothed.

Continuous variables, such as temperature, wind speed, and inversion thickness, were normalized using the Min-Max scaling technique, which transforms all variables into a [0, 1] range, ensuring no single variable dominates due to scale differences. Prior to modeling, the dataset underwent several preprocessing steps. Since radiosonde observations were conducted twice daily (00:00 and 12:00 UTC), all data were aligned to this bi-daily temporal resolution. For normalization, Min-Max scaling was applied to each continuous input variable using the following formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X is the original value of a given feature (e.g. temperature, PM_{10}, and inversion thickness) before scaling; X_{min} is the minimum value of that feature within the training dataset; X_{max} is the maximum value of that feature within the training dataset; X_{scaled} is the normalized value of the feature, scaled to lie within the range [0, 1].

X_{min} and X_{max} are computed over the training dataset only, to prevent data leakage. This ensured that all model inputs fell within a comparable range of [0, 1]. Categorical or discrete variables (e.g. wind direction) were excluded from Min-Max scaling and retained in their original form. All transformations were applied after splitting into training and testing sets to preserve the integrity of model evaluation.

Pearson correlation analysis was used to evaluate the linear relationships among variables, including meteorological parameters, inversion metrics, and pollutant concentrations. Variables with very weak correlations ($|r| < 0.1$) to target

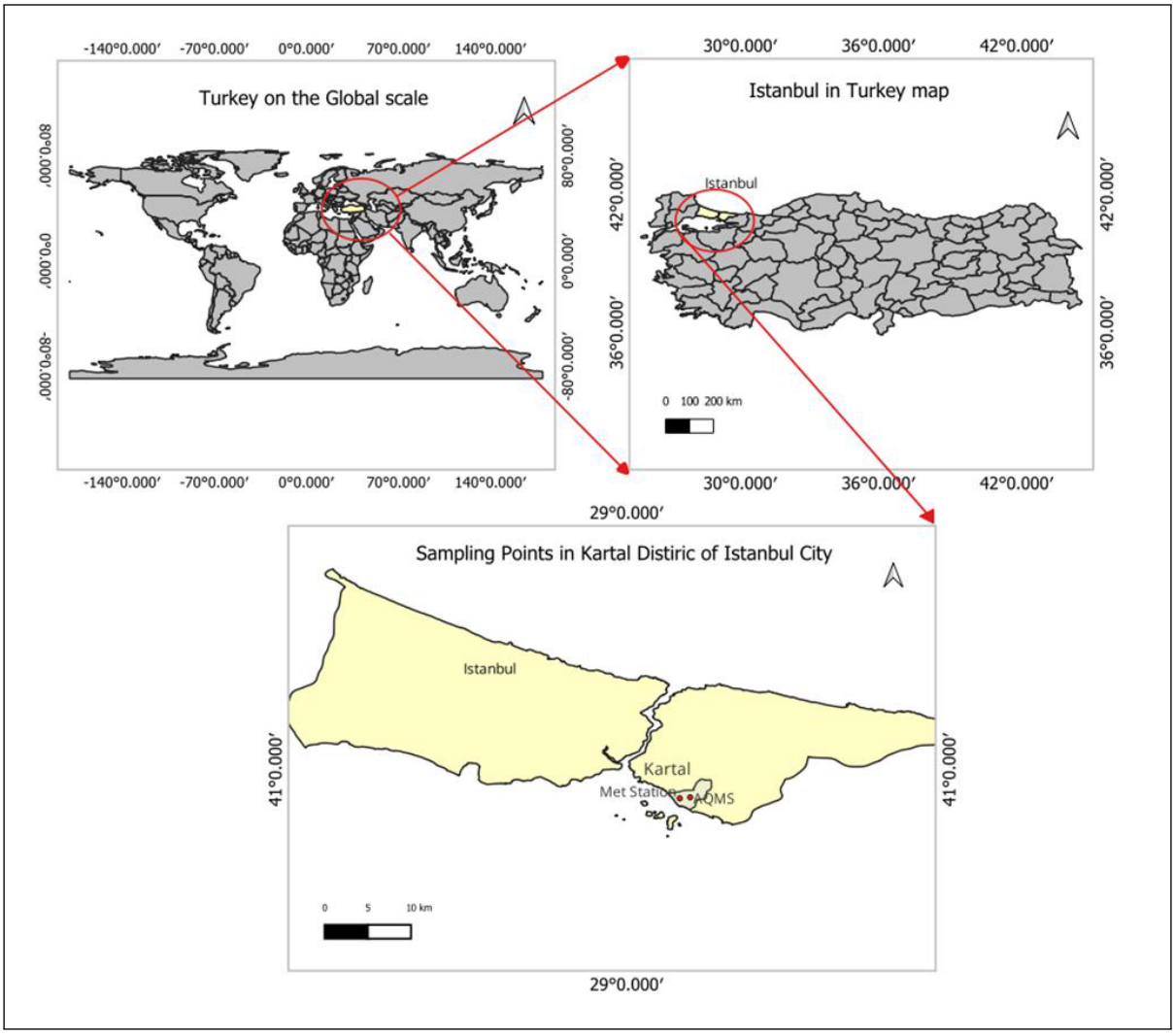

Figure 1. The study area is represented at three unique scales: global, national, and urban.

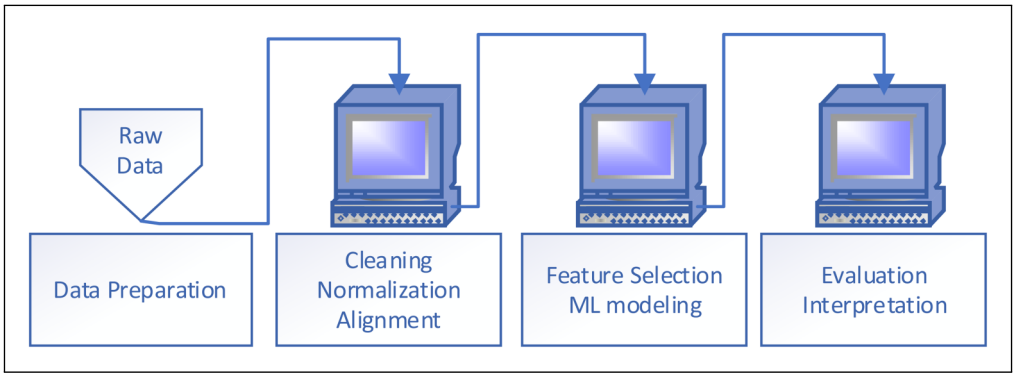

Figure 2. Overview of the initial data preparation workflow, including cleaning, normalization, and temporal alignment steps prior to modeling.

pollutants were considered less informative and excluded from modeling to reduce noise and improve model efficiency. To identify temporal dependencies, the autocorrelation function (ACF) was applied to PM_{10} and $PM_{2.5}$ concentrations, revealing short-term persistence patterns. Lags with strong positive autocorrelation (above 0.3 and statistically significant at the 95% confidence level) were selected as lagged input features. Specifically, PM_{10} exhibited notable autocorrelation up to lag 4, and $PM_{2.5}$ up to lag 3, supporting the inclusion of these lagged variables in model training. These ACF-based selections, combined with Pearson correlation analysis, ensured that only relevant and non-redundant predictors were used, enhancing the model's ability to capture the temporal dynamics of air pollution.

The final step involves temporal alignment. Since meteorological and air pollution data are recorded at different intervals, synchronization was performed to align the data uniformly. To ensure temporal alignment across all inputs, pollutant and meteorological variables were matched to the bi-daily radiosonde observation times at 00:00 and 12:00 UTC. Hourly records were either averaged or sampled at these timestamps, depending on data availability, to produce a consistent bi-daily time series. As a result, all features used in the model, including lagged values, reflect the same synchronized resolution, enabling accurate supervised learning and consistent temporal interpretation. All analyses and ML models in this study were implemented using Python version 3.11.7 (Python Software Foundation, 2023) within an Anaconda environment. The study utilized key libraries such as Scikit-learn, Pandas, NumPy, and Matplotlib to perform data processing, numerical computations, and visualization. This combination of libraries facilitated comprehensive data analysis, feature engineering, model training, and result visualization.

2.3 Lag selection and temporal dependency analysis

Feature selection plays a pivotal role in enhancing both the predictive accuracy and computational efficiency of ML models, particularly when working with multivariate environmental datasets. In this study, a two-stage feature screening approach was adopted to identify the most relevant predictors from among meteorological variables, inversion metrics, and pollutant concentrations.

First, Pearson's correlation coefficient was calculated to evaluate the strength and direction of linear associations between candidate input variables and the target outputs (PM_{10} and $PM_{2.5}$). Variables with weak correlations ($|r| < 0.1$) were excluded to reduce noise and improve model interpretability (Kalantari et al., 2024; Li et al., 2016; Zhang and Srinivasan, 2021).

Second, the temporal structure of the target variables was examined using the ACF, which quantifies the correlation of a variable with its own past values over various lag intervals. In the context of air quality, such temporal persistence often reflects recurring meteorological conditions or emission cycles (Wu et al., 2022).

Given the bi-daily sampling resolution, 00:00 and 12:00 UTC, ACF plots were generated for both PM_{10} and $PM_{2.5}$. Lag intervals with statistically significant autocorrelation ($r > 0.3$ at the 95% confidence level) were selected as additional features. This analysis revealed that PM_{10} exhibited significant autocorrelation up to lag 4 ($\sim$2 days), while $PM_{2.5}$ displayed autocorrelation up to lag 3.

These lagged values were incorporated into the model inputs to capture recent pollutant trends and temporal dependencies. By combining Pearson correlation with ACF-guided lag selection, a robust and non-redundant feature set was constructed that integrates both instantaneous and temporally persistent drivers of air pollution.

2.4 Model selection and justification

The ML regression models are commonly employed to predict continuous target variables by capturing the relationships between dependent and independent variables (Tsvetanova et al., 2018). In this study, we implemented four supervised ML models, namely SVMs, RFs, ANNs, and KNNs. These models were selected based on their established success in air quality forecasting and their complementary strengths in handling nonlinear and high-dimensional datasets.

Each algorithm employs a distinct learning mechanism: SVM uses kernel functions to model complex decision boundaries; RF leverages ensemble learning through decision trees; ANN captures nonlinear interactions via layered network structures; and KNN identifies local patterns based on proximity in feature space. This diversity facilitates a comprehensive evaluation of model performance across different learning paradigms.

Prior studies support the suitability of these models for PM forecasting. For example, Wu et al. compared ML and deep learning methods and found ensemble approaches such as RF to be highly effective when combined with lagged pollutant features (Wu et al., 2022). Araujo et al. applied ANN ensembles to estimate hospital admissions due to PM exposure, emphasizing their relevance to public health forecasting (Araujo et al., 2020). Dutta and Jinsart demonstrated that ANN outperformed both linear and tree-based models for PM_{10} prediction in Indian tier-II cities (Dutta and Jinsart, 2021). Imhanze et al. found LSTM to outperform SARIMA and Holt-Winters models in Abuja (Imhanze and Awe, 2024), while Gulati et al. showed that optimized ANN variants achieved high accuracy for $PM_{2.5}$ prediction (Gulati et al., 2023).

Despite the growing use of ML in air pollution studies, many existing models rely solely on surface-level meteorological variables. This study expands upon previous work by incorporating vertical inversion parameters, such as inversion thickness and height, into the input feature set, enabling a more physically informed and accurate representation of pollutant dynamics.

Each model was trained on a feature set that included surface meteorological variables, radiosonde-derived inversion parameters, and lagged pollutant concentrations based on ACF analysis. All continuous input variables were normalized using Min-Max scaling to ensure numerical consistency and to prevent dominance by features with larger scales.

- SVM: Implemented with an RBF kernel. Hyperparameter tuning was conducted using grid search over $C = \{1, 100, 1000\}$, $\gamma = \{0.01, 0.1, 1\}$, and $\in = \{0.1, 0.2\}$ with five-fold cross-validation.
- ANN: A feedforward network with two hidden layers (64 and 32 neurons) was trained using the Adam optimizer and ReLU activation. The learning rate was 0.001, batch size 16, and training stopped after 100 epochs or convergence.
- RF: The model used 100 decision trees with a maximum depth of 10. Feature importance scores were extracted to rank variable influence and assess interpretability.
- KNN: The number of neighbors k was optimized between 3 and 10, with the best results at $k = 5$. Euclidean distance was the metric, and all inputs were scaled to a unit range.

The hyperparameter search space was defined a priori and evaluated using GridSearchCV with five-fold cross-validation. For the RF model, the tested ranges included n_estimators = {100–500}, max_depth = {5–30}, min_samples_split = {2–10}, and min_samples_leaf = {1–5}. For SVM, we evaluated $C = \{0.1–100\}$, gamma = {0.001–1}, and kernel = {linear, RBF}. For ANN, tested architectures included 1–3 hidden layers with 16–64 neurons per layer, two activation functions (ReLU, tanh), and learning rates from 0.001 to 0.01. For KNN, the search space included k = {3–15}, weighting = {uniform, distance}, and distance metrics {Euclidean, Minkowski}. The optimal hyperparameters obtained from this search were used for training the final models.

To preserve temporal integrity and avoid data leakage, we applied a chronological 70:30 split to the bi-daily dataset. The training set consisted of data from 1 January to 13 September 2022, while the test set included data from 14 September to 31 December 2022. This forward-in-time partition simulates a realistic air quality forecasting scenario. Additionally, five-fold cross-validation was performed on the training set to optimize hyperparameters and assess model robustness prior to final evaluation on the hold-out test set.

2.5 Model training and validation strategy

All ML models were optimized to accommodate the size and structure of the bi-daily dataset. A grid search strategy combined with five-fold cross-validation was applied to the training subset (70% of the full dataset) to identify the optimal hyperparameters for each algorithm.

For SVM, the best performance was achieved using a radial basis function (RBF) kernel with $C = 10$, $\gamma = 0.1$, and $\epsilon = 0.1$. The ANN model used a two-layer feedforward structure with [32, 16] neurons and ReLU activation, trained for 100 epochs with a learning rate of 0.001 and batch size of 16. For RF, the number of trees was set to 100 and the maximum depth to 10, providing a balance between model complexity and interpretability. KNN performance was optimized by testing different values of K; $K = 5$ was selected based on validation performance. All models were implemented using Scikit-learn and TensorFlow frameworks in Python.

To enhance model interpretability, feature importance scores were extracted from the RF model based on the mean decrease in impurity. Additionally, SHAP (SHapley Additive exPlanations) values were computed to quantify the marginal contribution of each feature to individual predictions. This provided a more nuanced understanding of the model's decision process and helped validate the inclusion of key meteorological and inversion-related variables.

2.6 Performance evaluation metrics

To comprehensively evaluate model performance, multiple statistical metrics were employed, including the coefficient of determination (R^2), mean squared error (MSE), paired t-tests, and residual distribution analysis. These metrics collectively assess both the predictive accuracy and statistical robustness of the ML models.

The coefficient of determination (R^2) measures the proportion of variance in the observed data that is explained by the model. It ranges from 0 to 1, with values closer to 1 indicating stronger predictive performance. R^2 is especially useful for

Table 1. Paired *t*-test results comparing residuals for PM_{10} and $PM_{2.5}$ forecasting.

Comparison	Pollutant	*p*-values	Significance[a]
RF versus SVM	PM_{10}	0.00	Significant
RF versus ANN	PM_{10}	0.3455	–
RF versus KNN	PM_{10}	0.035	Significant
RF versus SVM	$PM_{2.5}$	0.1765	–
RF versus ANN	$PM_{2.5}$	0.2078	–
RF versus KNN	$PM_{2.5}$	0.0274	Significant

PM: particulate matter; SVM: support vector machine; RF: random forest; ANN: artificial neural network; KNN: K-nearest neighbor.
[a]All statistical significance tests were performed using paired two-tailed *t*-tests at the $\alpha = 0.05$ level.

interpreting how well the model captures variability in pollution levels. The representative formula of R^2 is provided in the following equation:

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

The mean squared error (MSE) quantifies the average squared difference between observed and predicted values. Lower MSE values indicate higher model accuracy and penalize larger prediction errors more heavily. Equation (3) provides the representation of MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

where y_i is the observed values; $\hat{y}_i$is the predicted values; and $\bar{y}$is the mean of observed values.

To mitigate overfitting and ensure generalizability, five-fold cross-validation was conducted on the training set for hyperparameter tuning.

To statistically compare model performance, paired *t*-tests were conducted on the prediction residuals. These tests evaluated whether the performance of the RF model differed significantly from that of the SVM, ANN, and KNN models, for both PM_{10} and $PM_{2.5}$. To assess the statistical significance of the performance differences among models, paired two-tailed t-tests were applied to the error distributions (MAE and RMSE) of each model pair. Statistical significance was evaluated at $\alpha = 0.05$. The results, including *p*-values, are presented in Table 1.

In addition, residual distribution analysis was conducted using boxplots to visually assess model bias, spread, and the presence of outliers. These plots were derived from the prediction errors on the 30% holdout test set. Together, these evaluation tools provide a rigorous and multi-dimensional assessment of model accuracy, stability, and statistical significance.

3 Results and discussion

This section presents the results in a structured manner to evaluate the predictive performance of the models and to explore pollutant dynamics under different meteorological and inversion scenarios. The analysis begins with descriptive statistics and correlation analysis of input variables, followed by autocorrelation insights to support lagged feature selection. Model evaluation is then performed using R^2 and MSE metrics across four algorithms (SVM, RF, ANN, and KNN). Finally, residual error distributions and feature importance rankings are analyzed to assess model robustness and interpretability. This stepwise structure ensures both scientific rigor and clarity in assessing the effectiveness of the ML framework.

The results are presented in three main stages: Figures 3 to 7 illustrate the predictive performance of the ML models for PM_{10} and $PM_{2.5}$; Figures 8 and 9 compare the residual error distributions across models; and Figures 10 and 11 display the temporal characteristics of pollutant concentrations based on autocorrelation analysis.

The provided heatmap in Figure 3 illustrates the Pearson correlation coefficients between various parameters, including inversion metrics, air pollutants, and meteorological variables. The color intensity reflects the strength and direction of the correlation, ranging from −1 (strong negative) to 1 (strong positive).

As depicted in Figure 3, the correlation heatmap reveals several notable interdependencies among pollutants, meteorological variables, and inversion parameters. PM_{10} and $PM_{2.5}$ exhibit a strong positive correlation, indicating shared

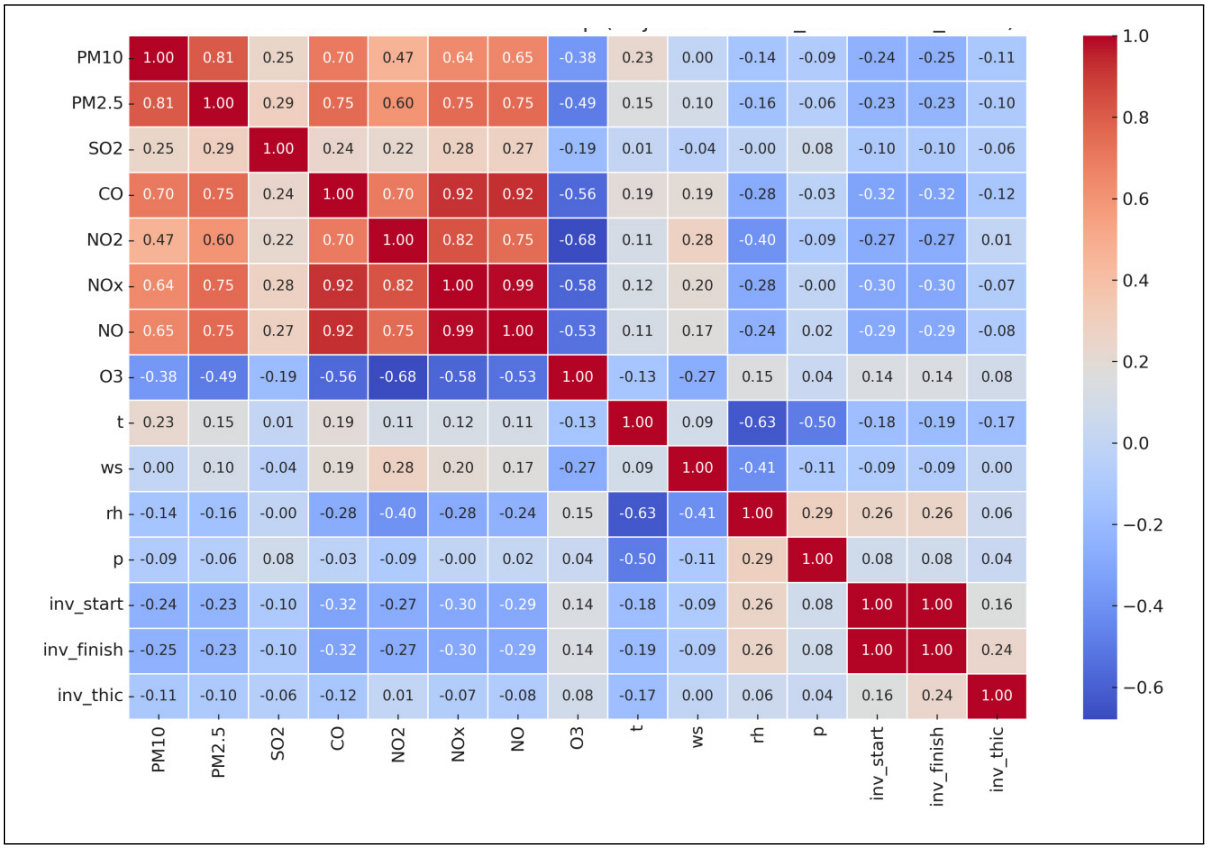

Figure 3. Heatmap showing Pearson correlation coefficients among pollutants (PM_{10}, $PM_{2.5}$, NO_x, SO_2, CO, and O_3), meteorological variables (temperature (*t*), wind speed (ws), relative humidity (rh), and pressure (*p*)), and inversion parameters (inversion thickness (inv_thic), inversion height start (inv_start), inversion height finish (inv_finish).

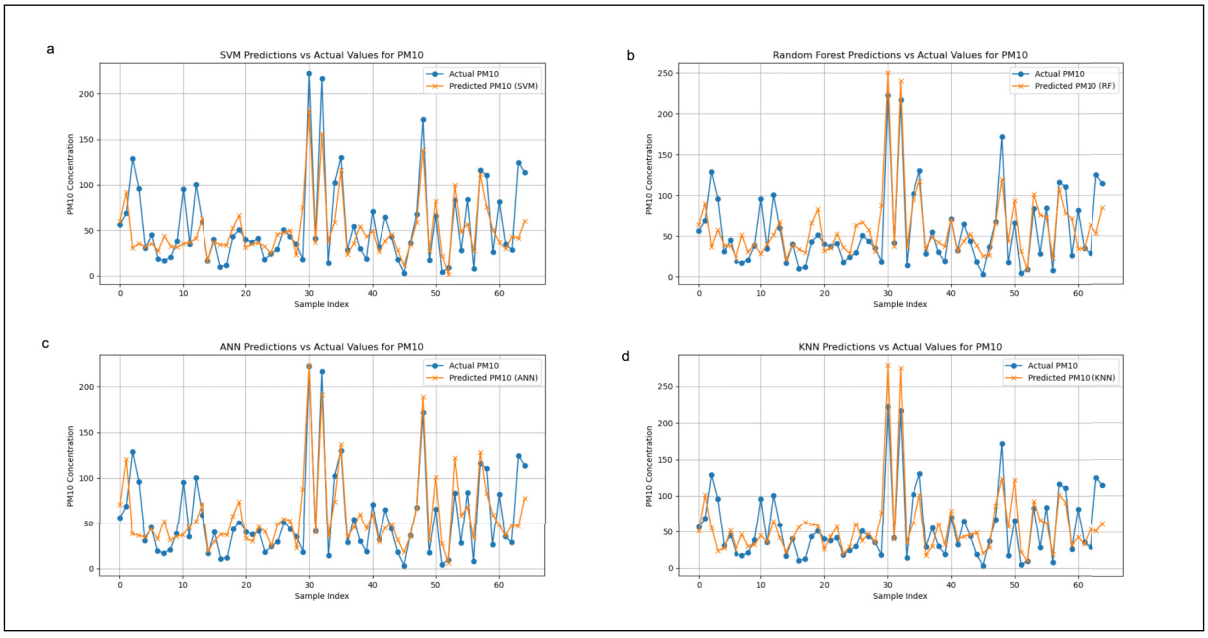

Figure 4. Comparison of predicted versus actual PM_{10} concentrations ($\mu g/m^3$) using ML models.
PM: particulate matter; ML: machine learning.

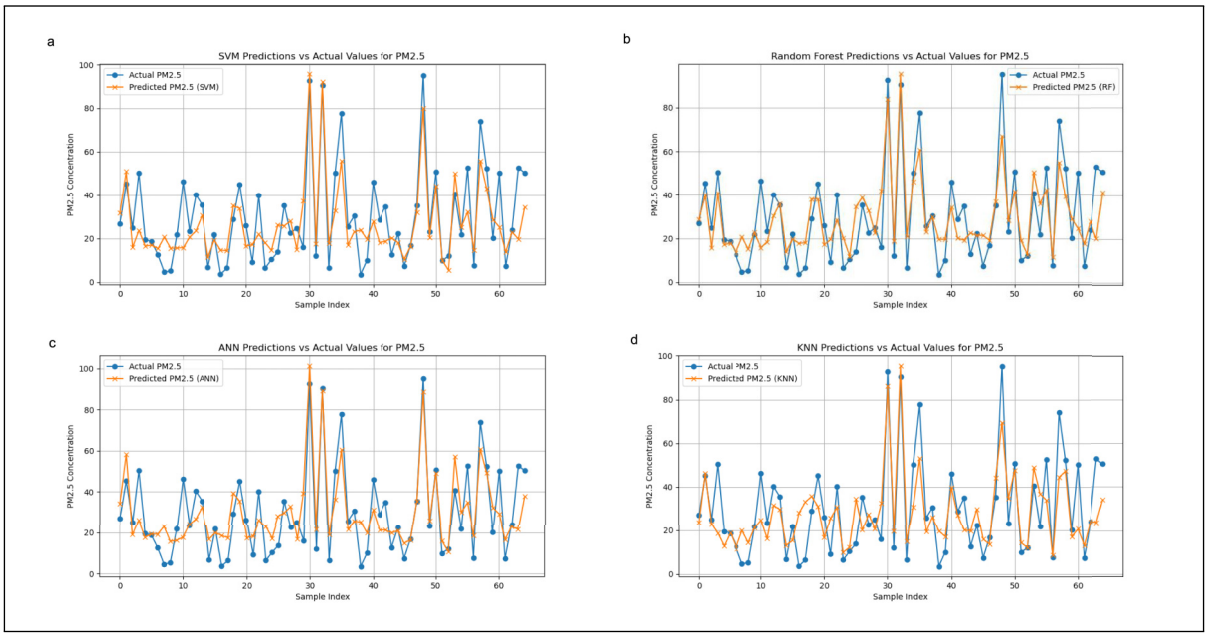

Figure 5. Comparison of predicted versus actual $PM_{2.5}$ concentrations ($\mu g/m^3$) using ML models.
PM: particulate matter; ML: machine learning.

emission sources or similar atmospheric transport dynamics. Likewise, NO, NO_2, and NO_x show strong mutual associations, consistent with their common origin from combustion-related processes. Moderate correlations between CO and SO_2 suggest overlapping emission sources, particularly from domestic heating and industrial activity.

Among the meteorological variables, wind speed (ws) and temperature (t) are moderately negatively correlated with pollutant concentrations, consistent with their roles in promoting atmospheric dispersion. Conversely, atmospheric pressure (p) shows a moderate positive correlation with both PM_{10} and $PM_{2.5}$, potentially indicating pollutant buildup under stable, high-pressure systems. Importantly, inversion thickness (inv_thic) demonstrates a positive correlation with PM levels, supporting the hypothesis that thicker inversion layers suppress vertical mixing and intensify near-surface pollution. These findings directly informed the selection of input features and emphasized the meteorological complexity influencing air quality in the Kartal district.

The SVM model was configured using a radial basis function (RBF) kernel to effectively capture complex, nonlinear relationships between features and pollutant concentrations. Hyperparameters were optimized through grid search, resulting in a regularization parameter of $C = 100$, a kernel coefficient $\gamma = 0.1$, and an epsilon margin $\varepsilon = 0.1$. These parameters balanced generalization with sensitivity to individual data patterns. Feature scaling was applied prior to training to ensure the kernel operated within a standardized input space. An epsilon-insensitive loss function was employed, which penalizes only substantial prediction errors, contributing to the robustness of the model for both PM_{10} and $PM_{2.5}$ forecasting tasks.

The RF model was designed to leverage its ensemble learning capabilities for robust regression. It consisted of 100 decision trees ($n\_estimators = 100$), with each tree having a maximum depth of 10 to prevent overfitting while capturing meaningful data patterns. The minimum number of samples required for a split ($min\_samples\_split = 2$) and the minimum number of samples required at a leaf node ($min\_samples\_leaf = 4$) were optimized to enhance the model's generalization ability. This configuration enabled the RF model to effectively capture complex relationships between features while maintaining robustness and predictive accuracy in estimating PM_{10} and $PM_{2.5}$ concentrations.

The ANN model was structured as a feedforward neural network with an input layer matching the number of predictor variables, followed by two hidden layers containing 64 and 32 neurons, respectively. Both hidden layers employed ReLU activation functions, while the output layer used a linear activation to support continuous predictions. The network was trained using the Adam optimizer with MSE as the loss function. Training was conducted over 100 epochs with a batch size of 16, and 20% of the training data was used for validation. This architecture allowed the ANN to learn complex interactions while maintaining reasonable training stability.

The KNN model relied on a distance-based approach to estimate pollutant concentrations by identifying the most similar historical observations. Hyperparameter tuning resulted in $k = 8$ for PM_{10} and $k = 6$ for $PM_{2.5}$, optimizing the balance between generalization and sensitivity. Euclidean distance was used to quantify similarity, and all input features were

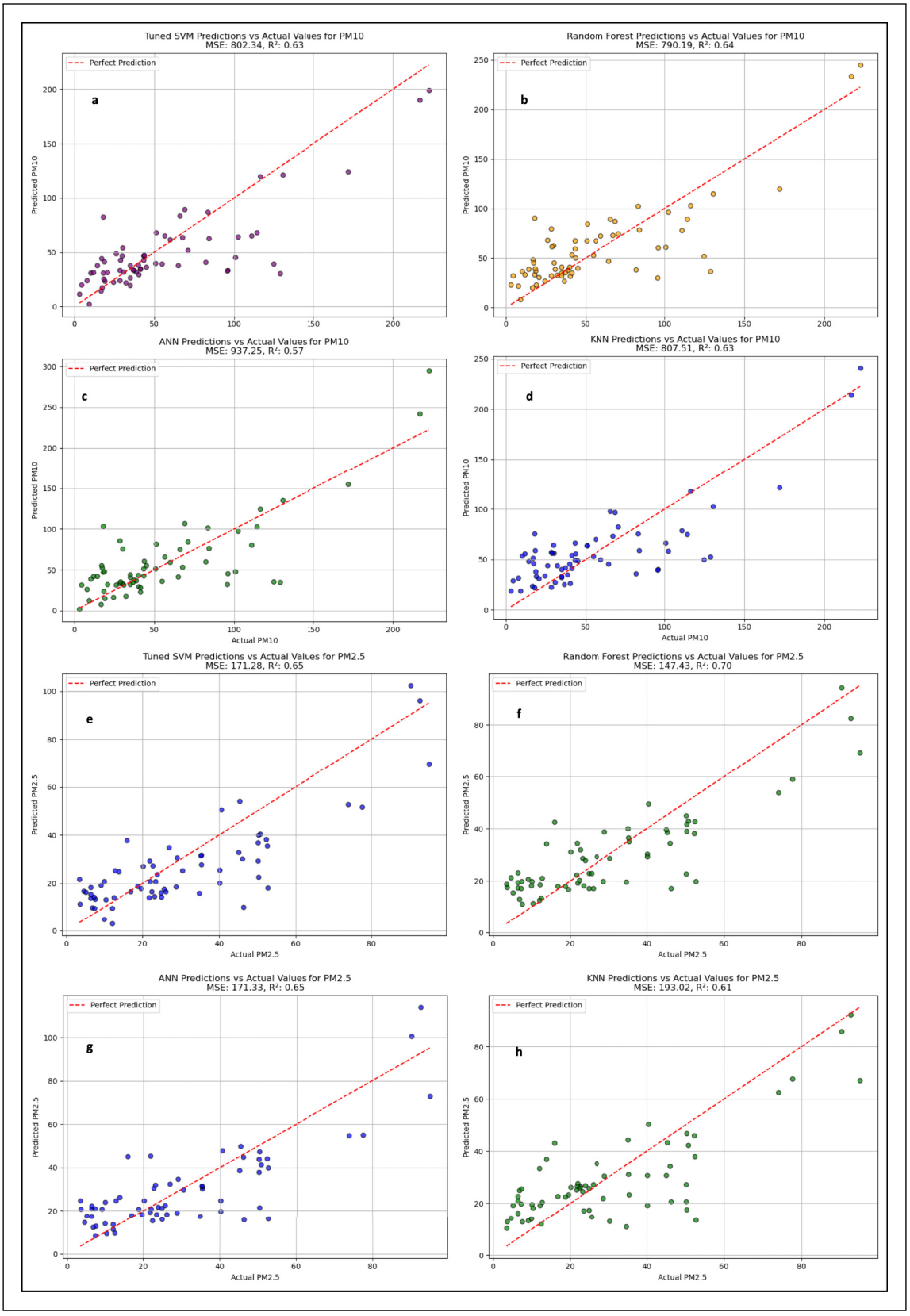

Figure 6. Performance comparison of ML models for PM_{10} (a)–(d) and $PM_{2.5}$ (e)–(h) predictions (SVM, RF, ANN, and KNN). ML: machine learning; PM: particulate matter; SVM: support vector machine; RF: random forest; ANN: artificial neural network; KNN: K-nearest neighbor.

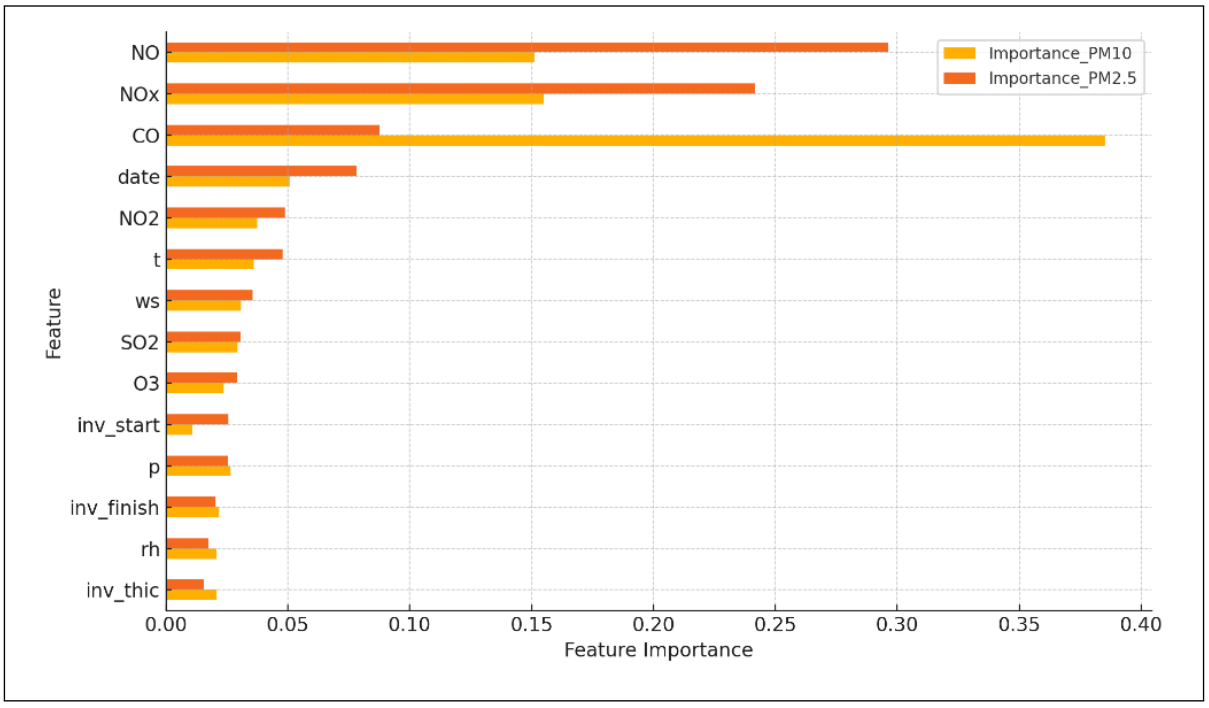

Figure 7. Feature importance rankings for PM_{10} and $PM_{2.5}$ based on the random forest model. Variables include inversion layer thickness (inv_thic, in meters), temperature (*t*, in °C), and lagged particulate matter (PM) concentrations (e.g. PM_{10} _lag1 and $PM_{2.5}$ _lag2).

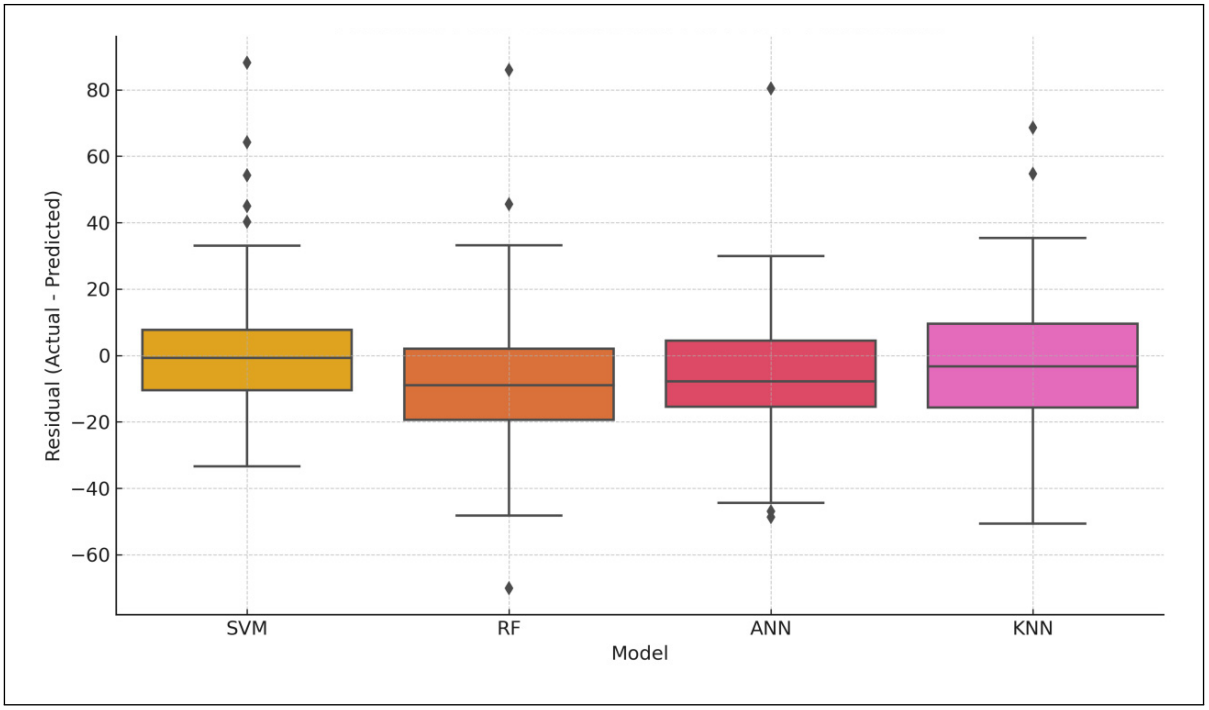

Figure 8. Boxplot of residual errors for PM_{10} predictions using four ML models.
PM: particulate matter; ML: machine learning.

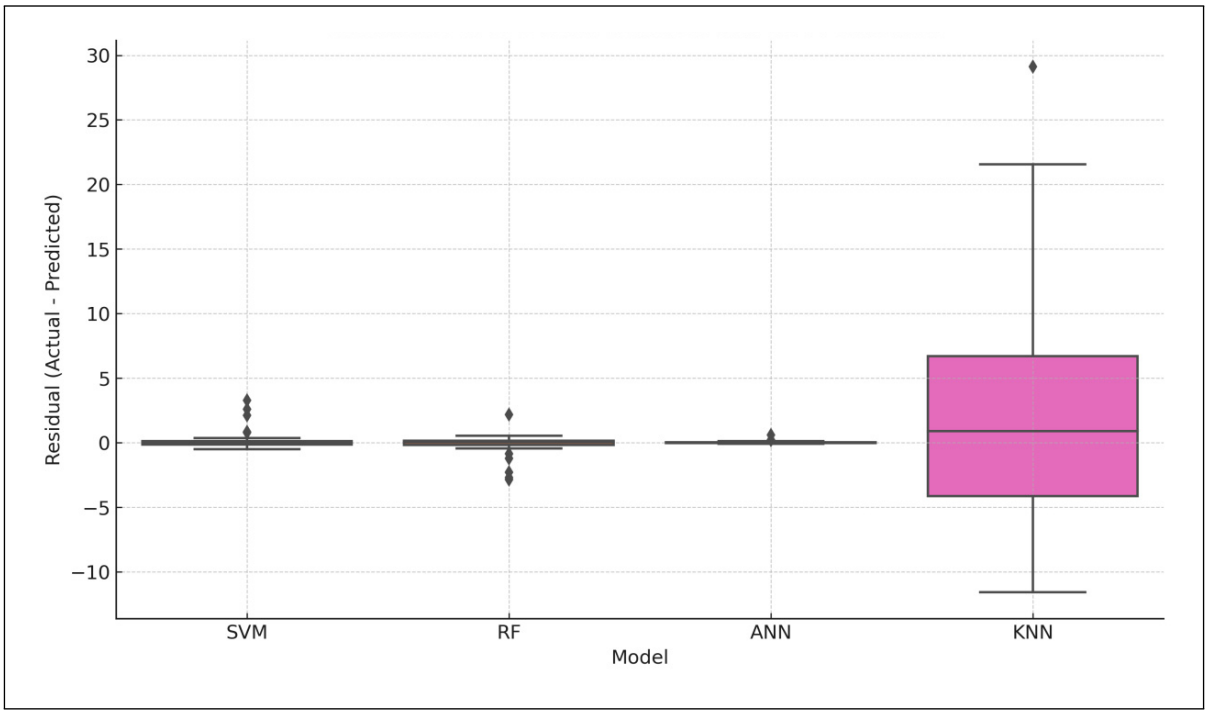

Figure 9. Boxplot of residual errors for $PM_{2.5}$ predictions using four ML models.
PM: particulate matter; ML: machine learning.

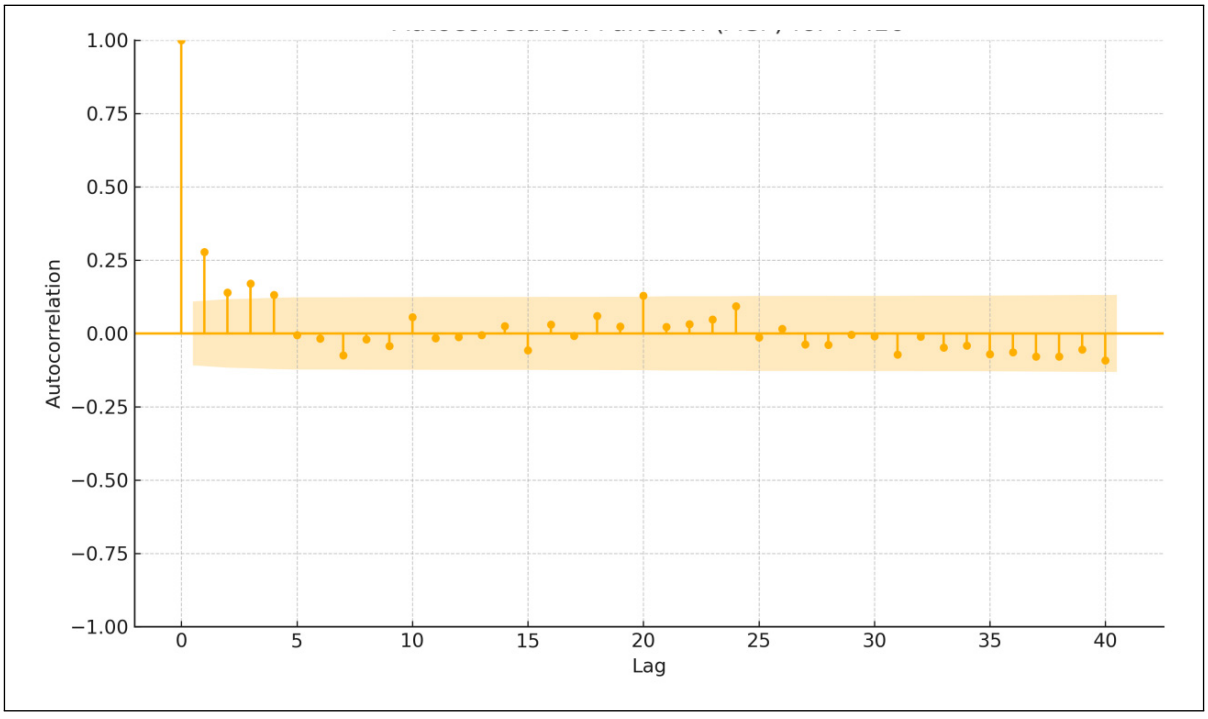

Figure 10. Autocorrelation function (ACF) for PM_{10} concentrations.

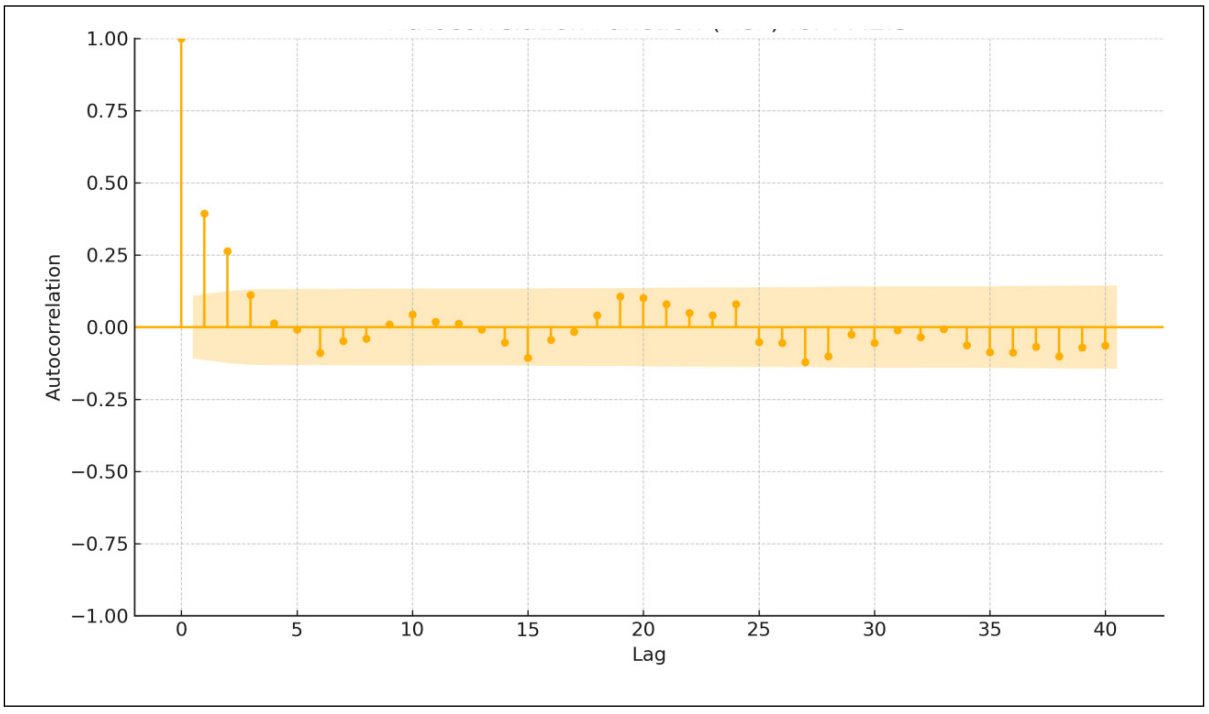

Figure 11. Autocorrelation function (ACF) for $PM_{2.5}$ concentrations.

standardized to ensure scale-invariant comparison. While conceptually simple, KNN effectively captured local variations in the dataset, particularly under stable meteorological conditions.

Following the model training and evaluation procedures described in the previous section, the performance of each ML algorithm, SVM, RF, ANN, and KNN, was assessed using R^2 and MSE metrics on the test dataset. Figures 4 through 6 display the comparison of observed and predicted values for PM_{10} and $PM_{2.5,}$ enabling visual assessment of model accuracy and fit. These plots reflect the final configurations of each model after hyperparameter tuning and normalization, as outlined in the methodology.

Figure 4 illustrates the comparison between actual and predicted PM_{10} concentrations for the four ML models used in this study: SVM, RF, ANN, and KNN. These charts visually show how well the models track the actual data over the sample index. They provide an intuitive, trend-focused understanding of the models' prediction accuracy and areas where they might over- or under-predict. Subplot (a) displays the performance of the SVM model, which captures the overall trend of PM_{10} concentrations but shows slight deviations in high-concentration peaks. Subplot (b) showcases the predictions by the RF model, demonstrating its ability to closely follow the observed values with minimal deviations. Subplot (c) presents the results of the ANN model, which effectively captures both high and low concentrations, reflecting its capability to model complex relationships. Finally, subplot (d) shows the performance of the optimized KNN model ($k = 8$), which successfully captures local variations but exhibits minor inconsistencies in extreme values.

For PM_{10} prediction, the RF model achieved the best overall performance, with an R^2 of 0.64 and an MSE of 790.19, indicating its ability to capture complex feature interactions. KNN and SVM performed similarly, each with an R^2 of 0.63 and MSE values of 807.51 and 802.34, respectively. Interestingly, the ANN achieved a slightly higher R^2 of 0.57, but at the cost of a higher MSE (937.25), suggesting overfitting or higher variance in its predictions.

Figure 5 illustrates the prediction performance of the four models in estimating $PM_{2.5}$ concentrations. Subplot (a) showcases the SVM model, which demonstrates moderate accuracy but struggles slightly with large fluctuations in $PM_{2.5}$ values. Subplot (b) highlights the RF model, which captures the majority of data patterns with notable precision, particularly for mid-range concentration values. Subplot (c) features the ANN model, which performs well in identifying general trends but tends to slightly overpredict some concentration peaks. In subplot (d), the optimized KNN model ($k = 6$) effectively predicts local variations in $PM_{2.5}$, although minor mismatches are observed during periods of abrupt changes.

For $PM_{2.5}$ forecasting, the RF model again outperformed other models, achieving an R^2 of 0.70 and the lowest MSE (147.43), indicating high predictive accuracy and robustness. SVM and ANN produced similar R^2 values (0.65), with the

SVM slightly outperforming ANN in terms of MSE (171.28 vs. 171.33). The KNN model showed the weakest performance, with an R^2 of 0.61 and the highest MSE (193.02), likely due to its sensitivity to noise and lack of generalization across broader patterns.

Figure 6 presents scatter plots comparing the predicted versus observed concentrations of PM_{10} and $PM_{2.5}$ for each of the four models. These plots serve as a visual indicator of prediction accuracy, with points aligning along the 45-degree line (diagonal) indicating perfect agreement. The RF, ANN, and SVM models showed closer clustering around this line, particularly for mid-range pollutant concentrations, suggesting more accurate and consistent performance. In contrast, the KNN model exhibited greater dispersion from the diagonal, reflecting less reliable predictions and a tendency to misestimate values at the extremes.

These results align with the conclusion that the RF consistently outperforms other models in air quality prediction, particularly for datasets with combined meteorological and pollutant data. While SVM and ANN provide competitive alternatives, their slightly higher errors emphasize the importance of fine-tuning model parameters. KNN's weaker performance underscores its limitations for $PM_{2.5}$ prediction, further supporting RF's reliability as the top-performing model.

RF consistently emerged as the most effective model across both PM_{10} and $PM_{2.5}$ predictions, achieving the highest R^2 values (0.64 for PM_{10} and 0.70 for $PM_{2.5}$) and the lowest MSE values (790.19 and 147.43, respectively). This confirms RF's capability to handle complex datasets with a combination of meteorological and pollutant data, making it the most reliable model for air quality prediction. SVM and ANN demonstrated competitive and comparable performance, with SVM achieving $R^2 = 0.63$ for PM_{10} and $R^2 = 0.65$ for $PM_{2.5}$, and ANN showing $R^2 = 0.57$ for PM_{10} and $R^2 = 0.65$ for $PM_{2.5}$. While their R^2 values were close to RF, their slightly higher MSE values indicate some limitations in capturing variability as effectively as RF. However, their balanced performance highlights their utility in air quality modeling, particularly when combined with comprehensive datasets. KNN delivered the weakest results, particularly for $PM_{2.5}$ ($R^2 = 0.61$, MSE = 193.02), while showing moderate performance for PM_{10} ($R^2 = 0.63$, MSE = 807.51). These findings reflect KNN's reliance on localized patterns and sensitivity to data structure and hyperparameter selection, making it less robust for general air quality predictions.

To statistically validate differences in model performance, paired t-tests were performed to compare the residuals of the RF model against those of the other three models (SVM, ANN, and KNN), for both PM_{10} and $PM_{2.5}$ predictions. These tests assessed whether the observed differences in prediction errors were statistically significant across the same test samples. Statistical significance was evaluated at the $\alpha = 0.05$ level using paired t-tests. The resulting p-values are presented in Table 1.

Table 1 presents the results of paired t-tests comparing the prediction residuals of the RF model with those of the other models (SVM, ANN, and KNN) for both PM_{10} and $PM_{2.5}$ forecasting. The RF model showed statistically significant improvement over KNN for both pollutants ($p = 0.035$ for PM_{10} and $p = 0.0274$ for $PM_{2.5}$), suggesting that RF more effectively captured the underlying pollutant dynamics in the dataset. For PM_{10}, RF also significantly outperformed SVM ($p < 0.05$), while no statistically significant difference was found between RF and ANN. In the case of $PM_{2.5}$, differences between RF and both SVM and ANN were not significant, indicating comparable predictive accuracy among these models. These findings reinforce the robustness of the RF model, particularly in predicting PM_{10}, and support its suitability for modeling complex air quality patterns in urban environments.

To interpret model behavior, the relative importance of all input features was extracted from the RF algorithm. For both PM_{10} and $PM_{2.5}$, the top-ranking features included lagged pollutant concentrations (especially lag-1 and lag-2), inversion thickness (inv_thic), and surface temperature.

Surface meteorological variables such as temperature, humidity, and wind speed showed a consistent yet moderate influence on PM variability due to their direct role in modulating atmospheric stability and near-surface dispersion. However, inversion-derived metrics, particularly inversion thickness, exhibited a more dominant influence during stagnant winter conditions when vertical mixing is suppressed. These results indicate that inversion characteristics provide additional predictive information that is not captured by surface parameters alone, emphasizing the added value of combining both surface and upper-air dynamics in PM forecasting.

Building on this distinction, inversion height demonstrated lower importance compared to inversion thickness, which is physically expected. While inversion height denotes the altitude at which the stable layer begins, it does not fully reflect the strength or depth of the stratified layer. In contrast, inversion thickness directly represents how deeply the atmosphere is stably layered and therefore governs the volume within which pollutants are trapped. This depth-driven control on vertical dilution explains why inversion thickness emerged as a more influential predictor in the feature importance analysis.

Moreover, the analysis also suggests potential interactions between inversion dynamics and surface meteorological conditions. Inversion effects were amplified under low wind speeds and reduced boundary-layer turbulence, conditions that inhibit horizontal and vertical dispersion. High humidity and cooler temperatures frequently co-occurred with strong

inversion events, reinforcing stable stratification and promoting pollutant buildup. These interactions help explain the nonlinear patterns captured by the ML models and underscore the importance of jointly considering surface meteorology and upper-air structure in PM prediction frameworks.

These variables had the highest mean decrease in impurity, indicating their strong influence on model predictions. The prominence of lag features supports the temporal persistence observed in ACF analysis, while the role of inversion thickness highlights the atmospheric stability's critical role in pollution accumulation. A full ranking of features is shown in Figure 7.

The RF feature importance analysis reveals that lagged PM values (especially lag-1 and lag-2), inversion thickness (inv_thic), and surface temperature (t) are the most influential variables for predicting both PM_{10} and $PM_{2.5}$. This indicates that short-term pollutant persistence, combined with stable atmospheric conditions that hinder vertical dispersion, plays a critical role in particulate matter accumulation. Features like relative humidity and wind direction had lower importance, suggesting a more indirect or less consistent impact on PM concentrations in this dataset. Among the top-ranked features in Figure 7, inversion thickness (inv_thic) and surface temperature (t) emerged as the most influential predictors. A thicker inversion layer typically traps pollutants near the surface, leading to higher PM concentrations due to reduced vertical mixing. Similarly, surface temperature plays a key role in regulating atmospheric stability, as cooler conditions often correlate with stronger inversions and pollutant accumulation, while warmer conditions can enhance dispersion. These physical mechanisms explain their high predictive contribution in the RF model.

As described in the "Model selection" section, all models were trained on 70% of the dataset and tested on the remaining 30%. Figures 8 and 9 show the distribution of residual errors on the test set for PM_{10} and $PM_{2.5}$, respectively. Figure 8 presents the residual error distributions for PM_{10} predictions using SVM, RF, ANN, and KNN models. Among these, the RF model demonstrates the tightest interquartile range and least spread, indicating higher prediction consistency and lower variance. SVM and ANN models show slightly wider residual distributions, while KNN exhibits more outliers and a skewed distribution, suggesting sensitivity to data structure and local fluctuations.

Overall, the boxplot confirms the superior performance of the RF model observed in R^2 and MSE metrics and provides additional visual insight into the stability, bias, and outlier behavior of each modeling approach.

Figure 9 illustrates the distribution of residual errors for $PM_{2.5}$ predictions using SVM, RF, ANN, and KNN models. Among the models, the RF again displays the narrowest residual range and minimal outlier presence, indicating strong predictive consistency. The ANN and SVM models follow closely, showing moderately tight residual distributions with slightly more spread.

In contrast, the KNN model reveals larger residual variance and several extreme outliers, suggesting instability when predicting fine particulate concentrations. This may reflect the model's sensitivity to localized variation and data sparsity. These results reinforce the robustness of the RF model in $PM_{2.5}$ prediction and align with previously reported performance metrics. The boxplot further emphasizes model stability and accuracy, supporting the integration of ensemble methods in air quality forecasting.

Figure 10 presents the ACF plot for PM_{10}, illustrating significant temporal autocorrelation at lower lag intervals, which supports the inclusion of lagged features in the prediction models.

Figure 10 presents the ACF of PM_{10} values, which was used in the feature selection step to determine the optimal lag depth included in the models. It illustrates the temporal characteristics of PM_{10} concentrations over a 40-lag horizon. The results reveal a clear and gradually decreasing autocorrelation pattern, with significant positive correlations observed up to approximately lag 5. This suggests that PM_{10} levels exhibit strong short-term persistence, meaning current concentrations are influenced by values from the previous 1 to 5 days. The presence of statistically significant autocorrelations at low lags supports the inclusion of lagged PM_{10} values as input features in the predictive models. This temporal memory likely reflects the cumulative effects of persistent emission sources, atmospheric stability conditions, and slow pollutant dispersion, particularly during inversion episodes. After lag 5, the correlation values decline toward zero, indicating diminishing influence from older observations. These findings confirm that PM_{10} concentrations are not temporally independent and validate the use of dynamic (time-lagged) features to enhance model accuracy. Incorporating these autocorrelated patterns allows the models to better capture pollution episodes and the underlying temporal dynamics of urban air quality in Kartal.

While the ML models used in this study do not provide prediction intervals by default, we examined model uncertainty through boxplots of residual errors and statistical testing. These visualizations (Figures 8 and 9) highlight differences in error variability and outliers between models. Additionally, paired t-tests were conducted to assess whether observed performance differences were statistically significant. Future work could explore Bayesian ML models or ensemble uncertainty methods to quantify predictive confidence intervals more explicitly.

Figure 11 shows the temporal structures of $PM_{2.5}$ concentrations, indicating significant positive autocorrelations up to approximately lag 4. This suggests that recent $PM_{2.5}$ values have a strong influence on current levels, reflecting short-term persistence typical of fine particulate pollution.

As shown in Figure 11, the autocorrelation declines beyond lag 4, indicating that the temporal influence weakens over time; however, the early lags retain significant predictive value. These results support the inclusion of lagged $PM_{2.5}$ values in ML models to enhance forecasting accuracy by capturing short-term temporal dependencies.

The autocorrelation analysis indicates that both PM_{10} and $PM_{2.5}$ exhibit short-term temporal dependencies, with statistically significant autocorrelations observed in the initial lags. PM_{10} demonstrates strong autocorrelation up to lag 5, suggesting that current concentrations are notably influenced by values from the previous 5 days. This persistence is likely associated with consistent emission sources and limited atmospheric dispersion under certain conditions. Similarly, $PM_{2.5}$ shows significant autocorrelation up to lag 4, reflecting a comparable but slightly shorter memory effect. The difference may be attributed to the finer particles' higher reactivity and sensitivity to short-term meteorological variations. These findings confirm that both pollutants retain temporal patterns, supporting the integration of lagged concentration values into the ML models. Doing so enhances the models' ability to capture pollution trends and forecast high-concentration episodes more accurately.

Although the ACF-guided inclusion of lagged features improved model performance, the study did not evaluate alternative lag configurations or perform robustness checks across different temporal subsets (e.g. seasonal splits). As such, model rankings should be interpreted within the context of the current lag design and data scope. Future studies could incorporate rolling-window validation or lag sensitivity analysis to confirm the generalizability of these findings.

Overall, the analysis confirms that RF is the most robust and reliable model for predicting PM_{10} and $PM_{2.5}$ concentrations. SVM and ANN provide solid alternatives, particularly when tuned appropriately, while KNN may be best suited as a baseline model or for tasks emphasizing localized variability. Although both RF and ANN are nonlinear models, the superior performance of RF in this study may be attributed to its ensemble-based structure and robustness to overfitting. While the ANN was trained with dropout and tuned via grid search, its performance showed more variability, likely due to the limited dataset size and higher sensitivity to weight initialization and architecture choices. In contrast, RF is less dependent on fine-tuned parameters and tends to generalize well in datasets with moderate feature dimensionality and structured inputs. These results underscore the importance of integrating meteorological and pollutant data to maximize predictive accuracy across all models.

Table 2 provides a comparative analysis of prior studies employing ML models for PM_{10} and $PM_{2.5}$ forecasting, highlighting differences in dependent and independent variables, evaluation metrics, and predictive accuracy.

This study reports an R^2 of 0.63 for PM_{10} and 0.65 for $PM_{2.5}$ using meteorological and pollutant data, demonstrating consistent and reliable performance. Compared to prior studies, such as Plocoste and Laventure (2023) with an R^2 of 0.58 for PM_{10} and Gayen et al. (2022) with an R^2 of 0.43 for $PM_{2.5}$ using meteorological and landcover data, this study results highlight the importance of integrating pollutant data (Gayen et al., 2022; Plocoste and Laventure, 2023). While the findings fall slightly behind Yi et al. (2019) for $PM_{2.5}$ ($R^2 = 0.71$), they align with mid-range SVM performances in literature, confirming the model's versatility when enriched with combined datasets (Yi et al., 2019). RF in this study achieved R^2 values of 0.64 for PM_{10} and 0.70 for $PM_{2.5}$, placing the results at the mid-range compared to prior studies. Kalantari et al. (2024) and Bozdağ et al. (2020) reported similar performances for PM_{10} ($R^2 = 0.61$–0.62), while Gündoğdu and Elbir (2024b) observed an R^2 of 0.73 for $PM_{2.5}$ using meteorological data alone (Bozdağ et al., 2020; Gündoğdu and Elbir, 2024b; Kalantari et al., 2024). The inclusion of pollutant data emphasizes RF's ability to enhance predictions with more complex datasets. However, the study results remain lower than those of Wang et al. (2023) for $PM_{2.5}$ ($R^2 = 0.78$), suggesting additional data, such as satellite or landcover information, could further optimize performance (Wang et al., 2023). This study's ANN model achieved R^2 values of 0.57 for PM_{10} and 0.65 for $PM_{2.5}$, comparable to studies like Dutta and Jinsart (2021) for PM_{10} ($R^2 = 0.65$) (Dutta and Jinsart, 2021). Gulati et al. (2023) reported stronger ANN performance for $PM_{2.5}$ ($R^2 = 0.81$) with pollutant data alone, indicating that ANN's architecture and hyperparameter tuning can greatly influence outcomes (Gulati et al., 2023). While the results demonstrate the model's non-linear relationship capture, they underline its sensitivity to feature engineering, as highlighted by lower performances in simpler datasets, such as Gayen et al. (2022), with an R^2 of 0.56 for $PM_{2.5}$. KNN in this study achieved R^2 values of 0.63 for PM_{10} and 0.61 for $PM_{2.5}$ with meteorological and pollutant data, consistent with findings such as Alsaber et al. (2023) for PM_{10} ($R^2 = 0.71$) and Gao et al. (2024) for $PM_{2.5}$ ($R^2 = 0.68$) (Alsaber et al., 2023; Gao et al., 2024; Gayen et al., 2022). However, Narkhede et al. (2023) reported superior results ($R^2 = 0.85$ for $PM_{2.5}$), suggesting that optimized hyperparameter tuning or additional geospatial inputs could enhance your model's performance (Narkhede et al., 2023). KNN's variability in predictive power, as observed in this study, aligns with its reliance on dataset quality and neighbor selection.

Study findings align well with the mid-range performances reported in the literature. The RF and SVM consistently demonstrate reliable predictions across both PM_{10} and $PM_{2.5}$, with RF showing slightly better accuracy for $PM_{2.5}$ in this study. While the ANN and KNN results are competitive, they emphasize the importance of careful hyperparameter tuning and the inclusion of diverse datasets to maximize predictive power. Overall, the models validate the role of integrating pollutant and meteorological data in improving ML-based air quality predictions.

Table 2. Comparison of the employed prior models for PM_{10} and $PM_{2.5}$ forecasting.

Models	Predicted (*dependent variables*)	Estimator (*independent variables*)	Metrics		Reference
			R^2	MSE	
SVM	PM_{10}	Met. data	0.56		Bozdağ et al. (2020)
SVM		Met. data	0.64	0.005	Yağmur (2022)
SVM		Met. data	0.58		Plocoste and Laventure (2023)
SVM		Met. & pollutant data	0.63	802.34	This study
RF		Met. data	0.61		Bozdağ et al. (2020)
RF		Met. data	0.62	121,221	Kalantari et al. (2024)
RF		Met. data	0.68	0.004	Yağmur (2022)
RF		Met. & pollutant data	0.64	790.19	This study
ANN		Met. data	0.42	184,916	Kalantari et al. (2024)
ANN		Met. data	0.62	0.005	Yağmur (2022)
ANN		Met. & pollutant data	0.65	497.8	Dutta and Jinsart (2021)
ANN		Met. & pollutant data	0.57	937.25	This study
KNN		Met. data	0.46	174,256	Kalantari et al. (2024)
KNN		Met. data	0.46		Plocoste and Laventure (2023)
KNN		Met. & pollutant data	0.71		Alsaber et al. (2023)
KNN		Met. & pollutant data	0.63	807.51	This study
SVM	$PM_{2.5}$	Met., water body, & landcover data	0.43		Gayen et al. (2022)
SVM		Pollutant data	0.34		Cheng et al. (2019)
SVM		Met. data	0.71		Yi et al. (2019)
SVM		Met. & pollutant data	0.65	171.28	This study
RF		Met. & pollutant data	0.78		Wang et al. (2023)
RF		Met., water body, land cover data	0.68		Gayen et al. (2022)
RF		Met. data	0.73		Gündoğdu and Elbir (2024b)
RF		Met. & pollutant data	0.70	147.43	This study
ANN		Pollutant data	0.81	90.63	Gulati et al. (2023)
ANN		Met., water body, land cover data	0.56		Gayen et al. (2022)
ANN		Pollutant data	0.40		Cheng et al. (2019)
ANN		Met. & pollutant data	0.65	171.33	This study
KNN		Met., emissions, pollutant data	0.68		Gao et al. (2024)
KNN		Met. data	0.83	4.49	Yaqoob et al. (2024)
KNN		Met. & pollutant data	0.85		Narkhede et al. (2023)
KNN		Met. & pollutant data	.61	193.02	This study

MSE: mean squared error; PM: particulate matter; SVM: support vector machine; RF: random forest; ANN: artificial neural network; KNN: K-nearest neighbor.

It is important to acknowledge that the modeling framework was developed and validated using data from a single location. Although the model demonstrated consistent performance across the dataset, the lack of cross-seasonal validation and spatial replication in adjacent districts remains a limitation. Future studies incorporating seasonal segmentation and regional generalization would enhance the model's broader applicability.

4 Conclusion

This study assessed the predictive capabilities of four supervised ML models, RF, SVMs, ANNs, and KNNs, for forecasting PM_{10} and $PM_{2.5}$ concentrations in Istanbul's Kartal district. The input dataset integrated surface meteorological variables, radiosonde-derived upper-air inversion characteristics, and lagged pollutant concentrations.

Among the evaluated models, RF consistently delivered the most accurate and stable predictions, achieving the highest R^2 values (0.64 for PM_{10} and 0.70 for $PM_{2.5}$) and the lowest MSE. ANN and SVM followed with competitive performance, while KNN exhibited greater error variability and lower predictive reliability. Paired *t*-tests confirmed that RF's performance was statistically superior to KNN for both pollutants and significantly better than SVM for PM_{10}.

A key strength of this study lies in the integration of upper-atmospheric inversion features, specifically inversion thickness, into the predictive framework, which are often neglected in traditional air quality modeling. The use of ACF analysis revealed short-term temporal dependencies in pollutant levels, supporting the inclusion of lagged variables. Feature importance analysis further emphasized the relevance of prior pollutant values and inversion characteristics, particularly in the RF model, highlighting the influence of vertical atmospheric stability on pollution accumulation.

While SVM and ANN also produced competitive results, particularly for $PM_{2.5,}$ their performance was somewhat less consistent across pollutants compared to RF. KNN yielded the weakest predictive metrics, yet served as a useful baseline for comparison. Paired *t*-tests further confirmed that RF's performance was statistically superior to KNN and SVM for PM_{10}, and to KNN for $PM_{2.5}$, underscoring its robustness under varying pollutant and feature conditions.

This research, however, has limitations. The analysis was conducted using data from a single urban monitoring station, which may constrain spatial generalizability. Although the long temporal span and bi-daily resolution captured seasonal dynamics implicitly, the lack of explicit seasonal segmentation or multi-site validation limits the broader applicability of the findings.

Future research should aim to expand this framework to multiple locations, incorporate seasonally adaptive or site-specific calibrations, and assess robustness across varying lag depths. Additional features, such as satellite-based aerosol products, traffic emissions, and real-time inversion forecasts, could further enhance model performance and practical utility. Moreover, integrating this approach into operational early-warning systems may enable timely interventions, such as traffic regulations or public health advisories during high-pollution episodes.

By emphasizing interpretability, methodological rigor, and practical relevance, this study provides a scalable, replicable, and scientifically grounded framework for urban air quality forecasting. The findings offer actionable insights for environmental planners and policymakers, particularly regarding the role of inversion dynamics and pollutant persistence in shaping air quality, while laying the groundwork for future data-driven air pollution mitigation strategies.

Despite the strong predictive performance achieved, this study has several limitations that should be acknowledged. The analysis is based on data from a single monitoring site and a single year, and model evaluation relied on *k*-fold cross-validation without explicit spatial transfer testing. Consequently, the findings primarily reflect site-specific atmospheric and emission characteristics and may not be directly transferable to other urban settings with different climatic or geographic conditions, such as coastal versus inland regions. Future research should therefore extend this framework to multi-site and multi-year datasets, incorporate explicit spatial validation strategies, and evaluate model robustness across diverse urban and regional environments.

Acknowledgements

The author extends sincere appreciation to Balıkesir University for its institutional support. Special thanks are also due to the Provincial Directorate of Environment and the Provincial Meteorological Service in Istanbul, Türkiye, for their invaluable assistance in providing essential environmental and meteorological datasets.

Author note

The author has read and understands the ethical responsibilities outlined in the journal's "Instructions for Authors" and affirms compliance with all applicable guidelines.

ORCID iD

Atilla Mutlu https://orcid.org/0000-0002-0777-0863

Ethical considerations

This research relied exclusively on publicly available or authorized datasets and did not involve human participants or animal subjects; therefore, ethical approval was not required.

Consent to participate

The author affirms their full participation in all aspects of the study, including the research and publication process.

Consent for publication

The author provides consent for the publication of this work and confirms agreement with the final version of the article.

Author contributions

The author was solely responsible for the conceptualization of the study, data acquisition and analysis, and the drafting and revision of the article.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author gratefully acknowledges the financial support provided by the Division of Scientific Research Projects at Balıkesir University under

Project No: BAP.2025/032. This funding was instrumental in enabling the collection, processing, and analysis of data, as well as the development of the models employed in the study.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

The data utilized in this study are available upon reasonable request. Air pollution and meteorological datasets were obtained from officially authorized, real-time monitoring stations and include high-resolution measurements critical to the research objectives.

References

Aguilera R, Luo N, Basu R, et al. (2023) A novel ensemble-based statistical approach to estimate daily wildfire-specific PM2.5 in California (2006–2020). *Environment International* 171: 107719.

Alsaber A, Alsahli R, Al-Sultan A, et al. (2023) Evaluation of various machine learning prediction methods for particulate matter PM10 in Kuwait. *International Journal of Information Technology* 15: 4505–4519.

Araujo LN, Belotti JT, Alves TA, et al. (2020) Ensemble method based on artificial neural networks to estimate air pollution health risks. *Environmental Modelling & Software* 123. DOI: 10.1016/j.envsoft.2019.104567.

Baltaci H (2017) Spatial and temporal variation of the Extreme Saharan Dust Event over Turkey in March 2016. *Atmosphere* 8: 41.

Baykara M, Im U and Unal A (2019) Evaluation of impact of residential heating on air quality of megacity Istanbul by CMAQ. *Science of the Total Environment* 651: 1688–1697.

Bayraktar OM and Mutlu A (2024) Analyses of industrial air pollution and long-term health risk using different dispersion models and WRF physics parameters. *Air Quality, Atmosphere & Health*. DOI: 10.1007/s11869-024-01573-8.

Biancofiore F, Busilacchio M, Verdecchia M, et al. (2017) Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmospheric Pollution Research* 8: 652–659.

Bozdağ A, Dokuz Y and Gökçek ÖB (2020) Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. *Environmental Pollution* 263: 114635.

Cheng Y, Zhang H, Liu Z, et al. (2019) Hybrid algorithm for short-term forecasting of PM2.5 in China. *Atmospheric Environment* 200: 264–279.

Di Q, Kloog I, Koutrakis P, et al. (2016) Assessing PM2.5 exposures with high spatiotemporal resolution across the Continental United States. *Environmental Science and Technology* 50: 4712–4721. 20160422.

Dutta A and Jinsart W (2021) Air pollution in Indian cities and comparison of MLR, ANN and CART models for predicting PM10 concentrations in Guwahati, India. *Asian Journal of Atmospheric Environment* 15: 2020131.

Flores RM, Mertoğlu E, Özdemir H, et al. (2020) A high-time resolution study of PM2.5, organic carbon, and elemental carbon at an urban traffic site in Istanbul. *Atmospheric Environment* 223: 117241.

Gao Z, Do K, Li Z, et al. (2024) Predicting PM2.5 levels and exceedance days using machine learning methods. *Atmospheric Environment* 323: 120396.

Gayen BK, Dutta D, Acharya P, et al. (2022) Exploring the effect of waterbodies coupled with other environmental parameters to model PM2.5 over Delhi-NCT in northwest India. *Atmospheric Pollution Research* 13: 101614.

Goudarzi G, Hopke PK and Yazdani M (2021) Forecasting PM2.5 concentration using artificial neural network and its health effects in Ahvaz, Iran. *Chemosphere* 283: 131285.

Gulati S, Bansal A, Pal A, et al. (2023) Estimating PM2.5 utilizing multiple linear regression and ANN techniques. *Scientific Reports* 13: 22578.

Gündoğdu S and Elbir T (2024b) A data-driven approach for PM2. 5 estimation in a metropolis: Random forest modeling based on ERA5 reanalysis data. *Environmental Research Communications* 6: 035029.

Hu X, Belle JH, Meng X, et al. (2017) Estimating PM(2.5) concentrations in the Conterminous United States using the random forest approach. *Environmental Science and Technology* 51: 6936–6944.

Imhanze OS and Awe OO (2024). Predicting Air Quality in an Urban African City Using Four Comparative Novel Time Series Models. *Available at SSRN 4701176*.

Kalantari E, Gholami H, Malakooti H, et al. (2024) Evaluating traditional versus ensemble machine learning methods for predicting missing data of daily PM10 concentration. *Atmospheric Pollution Research* 15: 102063.

Karaca F, Anil I and Alagha O (2009) Long-range potential source contributions of episodic aerosol events to PM10 profile of a megacity. *Atmospheric Environment* 43: 5713–5722.

Koo Y-S, Choi D-R, Kwon H-Y, et al. (2015) Improvement of PM10 prediction in East Asia using inverse modeling. *Atmospheric Environment* 106: 318–328.

Li H, Fan H and Mao F (2016) A visualization approach to air pollution data exploration—A case study of air quality index (PM2.5) in Beijing, China. *Atmosphere* 7: 35.

Mbululo Y, Qin J, Hong J, et al. (2018) Characteristics of atmospheric boundary layer structure during PM2.5 and ozone pollution events in Wuhan, China. *Atmosphere* 9: 359.

NAQMN (2023) National Air Quality Monitoring Network, https://sim.csb.gov.tr/Services/AirQuality (accessed 11.21.2023 2023).

Narkhede G, Hiwale AS, Pawar M, et al. (2023) Comparative analysis of prediction models for particulate matter (PM 2.5) prediction. In: 2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI), pp.1–6: IEEE.

Özdemir ET, Birinci E and Deniz A (2024) Multi-source observations on the effect of atmospheric blocking on air quality in İstanbul: A study case. *Environmental Monitoring and Assessment* 196: 698.

Plocoste T and Laventure S (2023) Forecasting PM 10 concentrations in the Caribbean area using machine learning models. *Atmosphere* 14: 134.

Python Software Foundation (2023) *Python. 3.11.7 ed*. Wilmington, DE: Python Software Foundation.

Qian W and Huang J (2019) Applying the anomaly-based weather analysis on Beijing severe haze episodes. *Science of the Total Environment* 647: 878–887.

Shao M, Xu X, Lu Y, et al. (2023) Spatio-temporally differentiated impacts of temperature inversion on surface PM2.5 in eastern China. *Science of the Total Environment* 855: 158785.

TSMS (2023) Turkish State Meteorological Service Radiosonde Database https://mevbis.mgm.gov.tr/mevbis/ui/index.html (accessed 21 November 2023).

Tsvetanova I, Zheleva I, Filipova M, et al. (2018) Statistical analysis of ambient air PM10 contami-nation during winter periods for Ruse region, Bulgaria. *MATEC Web Conf* 145: 01007.

Unal YS, Toros H, Deniz A, et al. (2011) Influence of meteorological factors and emission sources on spatial and temporal variations of PM10 concentrations in Istanbul metropolitan area. *Atmospheric Environment* 45: 5504–5513.

UNICEF (2024). Air pollution accounted for 8.1 million deaths globally in 2021, becoming the second leading risk factor for death, including for children under five years, https://www.unicef.org/press-releases/air-pollution-accounted-81-million-deaths-globally-2021-becoming-second-leading-risk (accessed 24 July 2025).

Wang Z, Chen P, Wang R, et al. (2023) Estimation of PM2.5 concentrations with high spatiotemporal resolution in Beijing using the ERA5 dataset and machine learning models. *Advances in Space Research* 71: 3150–3165.

Wu A, Harrou F, Dairi A, et al. (2022) Machine learning and deep learning-driven methods for predicting ambient particulate matters levels: A case study. *Concurrency and Computation: Practice and Experience* 34: e7035.

Xu T, Song Y, Liu M, et al. (2019) Temperature inversions in severe polluted days derived from radiosonde data in North China from 2011 to 2016. *Science of the Total Environment* 647: 1011–1020.

Yağmur EÇ (2022) Atmosferik Partikül Maddelerin Makine Öğrenmesi İle Tahmini: Beşiktaş, İstanbul Örneği. *Konya Journal of Engineering Sciences* 10: 807–826.

Yang Y, Ni C, Jiang M, et al. (2021) Effects of aerosols on the atmospheric boundary layer temperature inversion over the Sichuan Basin, China. *Atmospheric Environment* 262: 118647.

Yao Y, He C, Li S, et al. (2019) Properties of particulate matter and gaseous pollutants in Shandong, China: Daily fluctuation, influencing factors, and spatiotemporal distribution. *Science of The Total Environment* 660: 384–394.

Yaqoob I, Kumar V and Chaudhry SA (2024) Machine learning calibration of low-cost sensor PM 2.5 data. In: 2024 IEEE International Symposium on Systems Engineering (ISSE), pp.1–8: IEEE.

Yavuz V (2023) An analysis of atmospheric stability indices and parameters under air pollution conditions. *Environmental Monitoring and Assessment* 195: 934.

Yi L, Mengfan T, Kun Y, et al. (2019) Research on PM2.5 estimation and prediction method and changing characteristics analysis under long temporal and large spatial scale-A case study in China typical regions. *Science of the Total Environment* 696: 133983.

Zhang H and Srinivasan R (2021) A biplot-based PCA approach to study the relations between indoor and outdoor air pollutants using case study buildings. *Buildings* 11: 218.

Zhang Q, Jiang X, Tong D, et al. (2017) Transboundary health impacts of transported global air pollution and international trade. *Nature* 543: 705–709.