EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2025, 26(8), 97–114
https://doi.org/10.12912/27197050/207283
ISSN 2719–7050, License CC-BY 4.0

Received: 2025.05.29
Accepted: 2025.07.17
Published: 2025.08.01

Multi-horizon air pollution prediction using interpretable machine learning techniques in a growing urban area

Endrit Fetahi[1], Prosper Eguono Ovuoraye[2], Ercan Canhasi[1], Arsim Susuri[1*], Arta Misini[1]

[1] Faculty of Computer Science, University of Prizren "Ukshin Hoti", Str. Shkronjave, Prizren, Kosovo
[2] Department of Chemical Engineering, Federal University of Petroleum Resources, P.M.B 1221, Effurun, Nigeria
* Corresponding author's email: arsim.susuri@uni-prizren.com

ABSTRACT

Air pollution continues to be a critical public health and environmental challenge, particularly in fast-growing urban areas. This study presents an interpretable, multi-horizon forecasting framework for PM2.5 concentrations in Prishtina, the capital of Kosovo. Using hourly observations from 2018 to 2024, the study evaluates the predictive performance of five machine learning models: XGBoost, LightGBM, Random Forest, Support Vector Machine, and Linear Regression. Feature engineering, incorporating pollutant lags, rolling statistics, and cyclical time encoding on model performance, was investigated. The results show that among the selected ML models, XGBoost achieves the best one-hour forecast with R^2 of 0.862, MAE of 3.524, and RMSE of 6.513, while maintaining reasonable accuracy, with R^2 of 0.50 even at 24-hour horizons. To promote transparency, the study employs SHAP (SHapley Additive exPlanations) to quantify feature importance across different forecast horizons. Key drivers include recent PM2.5 lags, wind speed, and meteorological indicators. The proposed framework offers a robust, scalable, and interpretable approach for predicting air pollution, thereby supporting efforts to reduce emissions in Prishtina and similarly affected urban environments, enabling real-time alerts and data-informed environmental policy planning. Scientifically, this study uniquely integrates multi-horizon forecasting using advanced ML models with detailed temporal feature engineering and SHAP interpretability to reveal temporal shifts in feature importance, previously unaddressed systematically in air pollution modeling literature. These insights significantly enhance the understanding of dynamic air pollution interactions and are broadly applicable to urban environments globally with analogous pollution and meteorological dynamics.

Keywords: air pollution prediction, multi-horizon, interpretable machine learning (ML), Prishtina, urban area, regression.

INTRODUCTION

Air pollution is a major challenge to environmental health. It contributes to climate change and causes serious public health problems, increasing illness and possibly death rates (Halaktionov et al., 2025; Manisalidis et al., 2020). According to the World Health Organization (WHO), the presence of gaseous pollutants, including particulate matter (PM2.5, PM10), CO, NO_2, SO_2, and O_3, among other toxic gases, raises serious toxicological concerns (Ghorani-Azam et al., 2016). Harmful effects include severe respiratory and cardiovascular diseases in humans, resulting from both long-term and short-term exposure. There is a need to understand the role of climate and meteorological factors, including temperature, atmospheric pressure, wind speed, and humidity in the dispersal of pollutants (Nakyai et al., 2025). Consequently, environmental stakeholders propose a multidisciplinary approach to provide sustainable solutions to the menace of air pollution (Manisalidis et al., 2020). Additionally, air pollution is linked to respiratory and cardiovascular diseases, and predictive models can mitigate exposure risks, improving public health and

lowering healthcare costs (Chen et al., 2024). Although several conventional approaches have been applied to air-pollution control and management, the integration of artificial intelligence and machine-learning models represents a significant technological advancement for safeguarding environmental health (Huang et al., 2025; Olawade et al., 2024). Machine-learning models, especially those for pollution detection, enhance environmental protection and foster sustainable urban development (Olawade et al., 2024). Furthermore, real-time monitoring through these models enables timely interventions to prevent pollution, effectively balancing air-quality monitoring with data privacy (Quang et al., 2025).

According to the literature, XGBoost has effectively optimized Beijing pollution data (Li et al., 2022). The study's scope was limited to PM2.5 and O_3 modeling using XGBoost and WRF-Chem. XGBoost was also implemented in (Pan, 2018). That study, though limited to historical PM2.5 data, also reported meteorological impacts. The literature shows that XGBoost simulates pollutant concentrations better by capturing spatial, temporal, and non-linear patterns (Li et al., 2022; Pan, 2018). Air-quality prediction has also been performed with neuro-fuzzy methods (Anu Priya and Khanaa, 2023) and other machine-learning models, including LightGBM, XGBoost, and Random Forest, using historical data (Ravindiran et al., 2023). These outcomes suggest that machine learning shows promise in improving air-quality prediction. However, the current study concentrates on ML models with interpretability rather than "black-box" models such as artificial neural networks, which extract features automatically and lack transparency. Prishtina presents a particular scenario from previously studied urban environments, due to its pronounced winter temperature inversions, limited industrial regulation, and rapid urbanization without corresponding infrastructure developments. These local specifics yield unique PM2.5 dynamics characterized by seasonal variability and frequent pollution peaks, differentiating it markedly from other urban settings studied extensively in prior literature. These recurring pollution episodes, if unaddressed, pose long-term health risks and environmental degradation. Governments and regulatory bodies have historically used empirical data and expert analyses to implement environmental policies. However, regulatory challenges often stem from enforcement issues, political interests, and limited funding, indicating that regulatory limitations are not solely due to the absence of modeling data. Machine learning and real-time pollution data modelling can significantly enhance decision-making by providing predictive insights, improving forecasts, identifying pollution trends, and optimizing interventions for environmental quality control.

To effectively address these unique air pollution dynamics, this study rigorously evaluates several interpretable machine learning (ML) models widely recognized in literature but rarely compared systematically across multiple forecasting horizons with detailed interpretability analysis. Specifically, we comprehensively examine the predictive performance of XGBoost, LightGBM, Random Forest, support vector machine (SVM), and Linear Regression, all extensively applied individually in prior research but rarely contrasted under unified multi-horizon scenarios and explicit SHAP-based interpretability frameworks. This comparative approach provides novel insights into model suitability and robustness, explicitly quantifying predictive skill and interpretability trade-offs across both short-term (1–3 hours) and long-term (up to 24 hours) forecasting windows, thereby significantly advancing the methodological rigor in air quality predictive modelling literature.

Scientifically, this research aims to establish a novel predictive framework explicitly addressing three significant gaps in current air pollution modelling literature:

1) the absence of a systematic, integrative temporal feature-engineering approach tailored specifically for multi-horizon forecasting scenarios,
2) the lack of comprehensive comparative evaluations of robust and interpretable machine learning models in the context of multi-horizon air quality prediction, and
3) insufficient rigorous exploration of how feature importance dynamically shifts across multiple forecasting horizons, explicitly quantified using SHAP-based interpretability methods.

Our primary hypothesis is that integrating detailed temporal and meteorological feature engineering with multi-horizon interpretability via SHAP will significantly outperform conventional methods, thus scientifically advancing both predictive accuracy and understanding of temporal pollution dynamics.

According to the existing literature, limited research has reported the integration of feature engineering, including explainable AI (XAI) techniques, such as SHAP, which were employed in real-time air pollution analysis. Integrating machine-learning techniques offers insights for informed decision-making based on pollution data (Shetty et al., 2024). Exploring AI-driven interventions to reduce air pollution can guide sustainable urban development and minimize environmental impact (Rahaman et al., 2025). Additionally, accessible forecasts empower residents to take precautions, such as avoiding outdoor activities during peak pollution hours (Ramírez et al., 2019). This present study evaluates multiple ML models to provide accurate multi-horizon forecasts for PM2.5 concentrations. By integrating a sensitivity analysis with XAI techniques, it offers real-time insights for policymakers to guide strategies like real-time alerts, targeted emission reductions, and policy planning on key air pollution drivers in Prishtina, Kosovo. The specific contributions of this study are summarized as follows:

- a comprehensive comparison of interpretable machine-learning models for multi-horizon PM2.5 forecasting across five lead times.
- the design of an extensive feature-engineering pipeline using XAI techniques that significantly improves prediction accuracy and analyzes feature importance for model transparency.
- a sensitivity analysis to identify pollution trends and adapt to urban environments, enabling proactive environmental planning with real-time pollution alerts.

The scalable pollution modeling approach is adaptable to various urban settings, enabling proactive environmental planning with real-time pollution alerts for communities and regulators. The study's findings will enhance understanding of air quality and its prediction in the area studied, forming a basis for future research.

RELATED WORK

According to the existing literature, air-pollution research has witnessed remarkable methodological advancements, largely propelled by technological innovation (Chen et al., 2024). Among these developments is the integration of machine learning (ML) models into air-quality-index control and pollution assessment, with case studies on explainable AI (XAI) and SHAP for feature-importance analysis (Kedar, 2024; Reddy and Kumar, 2023). Recent studies have applied various ML techniques to air-quality forecasting (Reddy and Kumar, 2023).

The authors in the paper (Raviteja and Reddy, 2024) employed k-nearest neighbours (KNN), Random Forest, and other ML algorithms to analyse data and forecast the air quality index (AQI), highlighting particularly strong performance from KNN and Random Forest. The authors noted that Python and the Scikit-Learn library provide convenient modules for implementing these models. Another research (Raviteja and Reddy, 2024; Chen et al., 2024; Reddy and Kumar, 2023) examined support-vector regression (SVR), XGBoost, and artificial neural networks (ANNs) for forecasting air quality in Visakhapatnam, India, evaluating performance with metrics such as mean absolute error (MAE) and root-mean-squared error (RMSE) (Chen et al., 2024). XGBoost exhibited the best predictive capability (Chen et al., 2024; Zhou et al., 2024).

Although SVR, Linear regression, XGBoost, and Random Forest remain conventional choices for air-quality assessment (Chen et al., 2024; Persis and Ben Amar, 2023; Suárez Sánchez et al., 2011; Zhou et al., 2024), more recent work has adopted statistical and deep-learning methods such as ARIMA, long short-term memory (LSTM) networks, and convolutional neural networks (CNNs) for air-quality prediction (Luo and Gong, 2023; Tsokov et al., 2022). Multi-step (multi-horizon) forecasting techniques are acknowledged in this context, though details of their implementation are often omitted.

Another work shown (Suárez Sánchez et al., 2011) used experimental data on nitrogen oxides (NOx), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), and particulate matter (PM_{10}) collected from 2006 to 2008 to develop a non-linear air-quality model for the Avilés urban area (Spain) using support-vector machines.

In contrast, the paper (Luo and Gong, 2023) proposed an ARIMA-WOA-LSTM model that combines ARIMA for linear components with LSTM networks for nonlinear patterns, with Whale Optimization Algorithm (WOA) tuning the LSTM hyperparameters. Their results showed that the hybrid model improved prediction accuracy and stability. ARIMA effectively extracts linear patterns, while LSTM captures nonlinear

relationships; systematic feature selection further enhances model performance.

Although previous studies extensively applied various machine learning and interpretability techniques, they haven't really explored dynamic shifts in feature importance across multiple forecasting horizons. Moreover, existing research predominantly treats interpretability superficially or applies it to single-horizon models, neglecting how pollutant driver importance evolves temporally. Thus, the current study provides a novel scientific advancement by explicitly quantifying and interpreting feature dynamics across multiple horizons, an approach previously unaddressed and essential for deeper insights into urban air quality forecasting.

METHODOLOGY

This section outlines the complete methodology used for developing a multi-horizon air pollution prediction framework.

Data collection

This study examines the urban environment of Prishtina, Kosovo (GPS coordinates: 42.6596 N and 21.1573 E) in the capital city of Kosovo. The geographical location of the study area is presented in Figure 1. In this study, open-source datasets comprising hourly air pollution measurements of major pollutants (PM2.5, PM10, NO_2, O3, CO, and SO_2) were collected from government-operated monitoring stations from the webpage of the Hydrometeorological Institute of Kosovo. Meteorological data, including temperature, humidity, and other weather variables, were provided by the public API of the Meteostat tool, which adheres to international WMO standards. The dataset was recorded hourly over multiple years, from 2018 to 2024, totaling 50,929 observations.

Procedure for machine learning

The overall workflow of this study is illustrated in Figure 2.

Data preprocessing

The dataset was further preprocessed to handle missing data using a forward-fill strategy to preserve the structure of the series. This assured us the continuity without introducing artificial variant into the pollutant or meteorological values.

Model selection and hyperparameter tuning

To comprehensively evaluate model performance, a suite of interpretable ML models was chosen to balance prediction performance with interpretability. These models, including XGBoost, LightGBM, SVM, Random Forest, and Linear Regression, respectively, were implemented on the data collected for the modelling and prediction of the air quality control. While extensive hyperparameter tuning was conducted to optimize model performance. The XGBoost is a powerful ensemble learning algorithm based on gradient boosting decision trees, which builds models sequentially to learn complex patterns and interactions in data. It is suitable for structured data tasks like time-series prediction, where temporal and environmental variables often interact in non-linear and high-dimensional

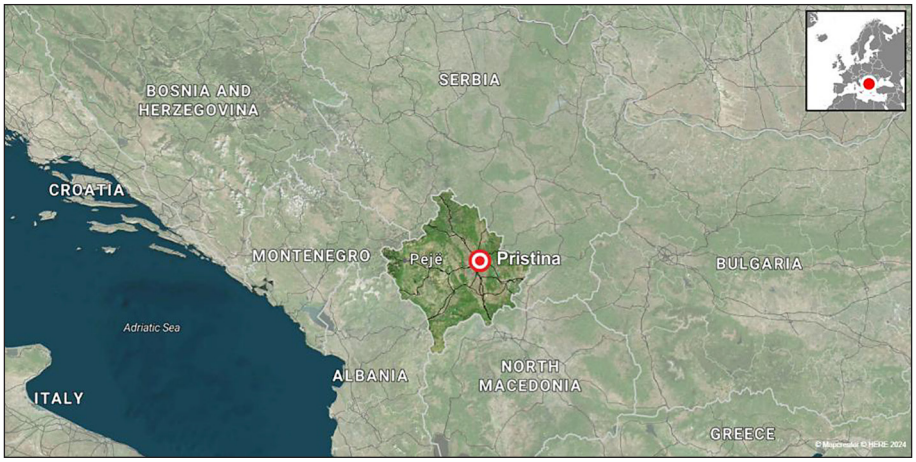

Figure 1. Location of the area studied

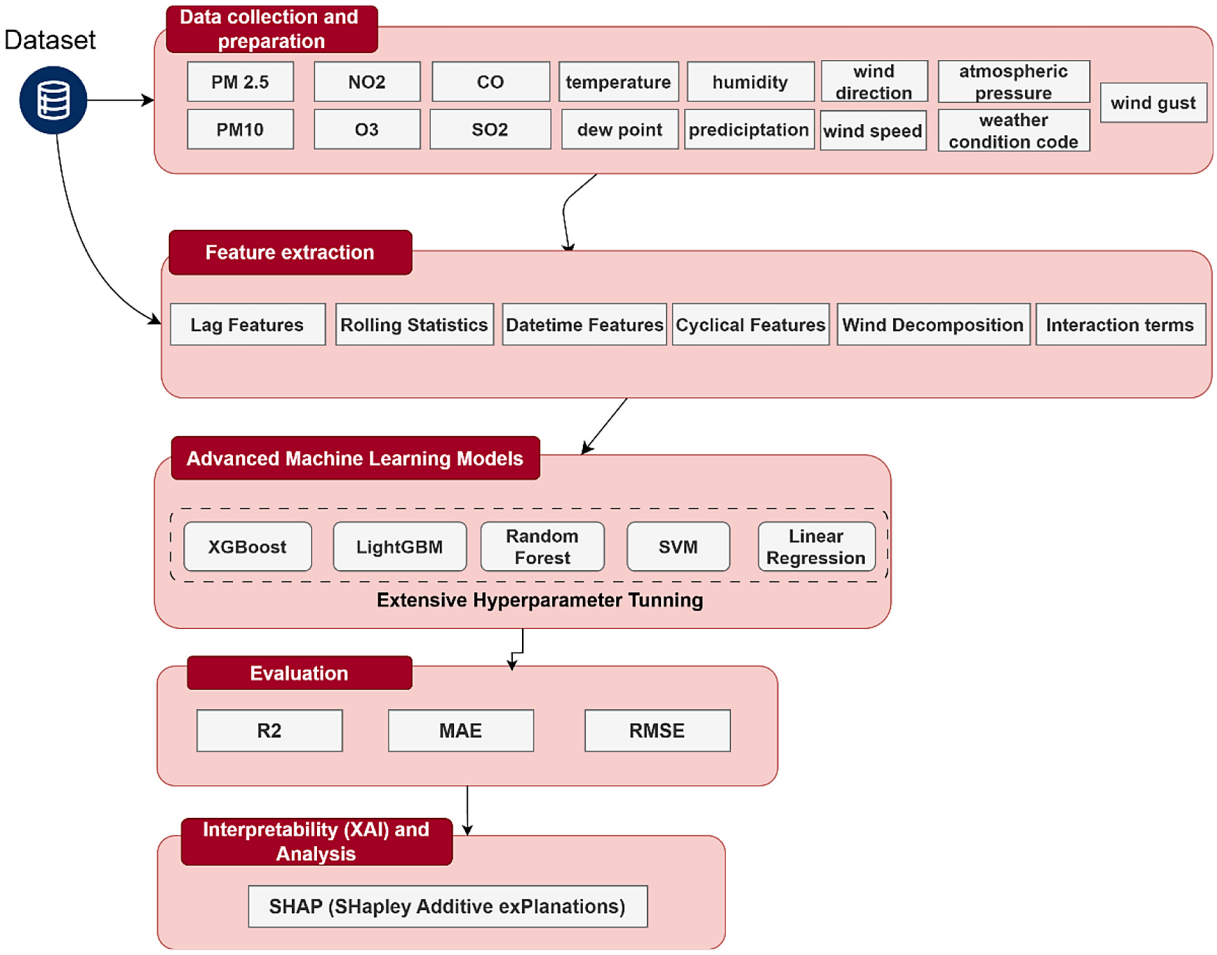

Figure 2. High-level diagram of methodology

ways. LightGBM is another gradient boosting algorithm optimized for efficiency and scalability, for the air quality and prediction datasets, using a histogram-based method to bucket continuous features and growing trees leaf-wise. It performs well on large datasets with many features and favors fast inference, making it advantageous for real-time or near-real-time applications. Random Forest is a classic ensemble method based on aggregating multiple decision trees, promoting diversity and reducing variance in predictions. It uses majority voting for decisions (McClarren, 2021).Linear Regression is a fundamental predictive model that assumes a linear relationship between input features and target variables, serving as a strong baseline in many applications (Wilson and Lorenz, 2015; Zhou et al., 2024). After selecting the machine learning models, we proceeded with hyperparameter tuning to optimize their predictive performance and ensure fair comparisons across configurations, parameters, which are shown in Table 1. For this, we run a Grid Search for each model.

Feature engineering

To thoroughly capture the dependencies and environmental complexity in air pollution dynamics, as well as to increase the accuracy of the prediction models and interpretability, we implemented extensive feature engineering for the dataset, resulting in a considerable set of features. First, lag features were generated for pollutant concentrations (PM2.5, PM10, NO_2, CO, SO_2, O_3) and meteorological variables (e.g., temperature, humidity, wind speed) at multiple hourly offsets to capture short-term dependencies. Rolling window statistics, such as mean and standard deviation, were computed for both pollutants and weather variables over 3, 6, 12, and 24-hour intervals to capture local trends and volatility. Temporal markers, including hour, day, month, and year, were extracted and transformed into cyclical components using sine and cosine functions to model seasonality and diurnal cycles. Additional derived features include wind vector decomposition into u and v components and a temperature-humidity

Table 1. Hyperparameter tuning and best parameters found for each model

Model	Hyperparameters tuned	Number of tests run	Best parameter found
XGBoost	'n_estimators': [100, 200], 'max_depth': [3, 5, 7], 'learning_rate': [0.01, 0.1, 0.2], 'subsample': [0.8, 1.0], 'colsample_bytree': [0.8, 1.0]	216 fits	'colsample_bytree': 1.0, 'learning_rate':0.1, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8
LightGBM	'n_estimators': [100, 200], 'num_leaves': [31, 63, 127], 'learning_rate': [0.01, 0.1, 0.2], 'subsample': [0.8, 1.0], 'colsample_bytree': [0.8, 1.0], 'min_child_samples': [5, 20, 50]	648 fits	'colsample_bytree': 0.8, 'learning_rate': 0.1, 'min_child_ samples': 50, 'n_estimators':100, 'num_leaves':31, 'subsample': 0.8
Random Forest	'n_estimators': [100, 200], 'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['sqrt', 'log2']	324 fits	max_depth':None, 'max_ features':'sqrt', 'min_samples_ leaf':2, 'min_samples_split':10, 'n_estimators': 200
SVM	'model__C': [0.1, 1, 10], 'model__epsilon': [0.01, 0.1, 0.2], 'model__kernel': ['rbf', 'linear'], 'model__gamma': ['scale', 'auto', 0.1, 0.01]	216 fits	'model__C': 0.1, 'model__epsilon': 0.1, 'model__kernel': 'linear', 'model__gamma': 'auto'
Linear regression	N/A	N/A	N/A

interaction term, with its lags. The detailed features are shown in Table 2.

Our feature engineering strategy explicitly differs from conventional methods through its structured temporal stratification. Integrating cyclical temporal encoding, multiple pollutant and meteorological lags, and rolling statistics explicitly provides a structured and novel approach tailored specifically to multi-horizon forecasting. Coupled systematically with SHAP interpretability, this detailed temporal feature-engineering framework uniquely clarifies how predictors dynamically shift across forecasting horizons, enhancing both predictive accuracy and interpretative depth.

Prediction strategy

To evaluate the forecasting performance of our models across varying lead times, we adopted a multi-horizon prediction strategy targeting PM2.5 concentrations at future intervals. Specifically, we trained separate models for forecasting at T+1, T+3, T+6, T+12, and T+24 hours ahead, where T represents the current time step. For each horizon, the dataset was adjusted so that the target variable corresponded to the PM2.5 level at the respective future time point. The input features, consisting only of past and current observations, ensured a strictly causal modeling approach. This design enables the models to learn distinct temporal patterns and dependencies relevant to each forecast window, supporting both short-term and long-term air quality predictions.

Interpretability analysis

To enhance transparency and support actionable insights, we employed SHAP (SHapley Additive exPlanations) in this study to interpret the contribution of each feature across different forecast horizons. SHAP values provide a consistent,

Table 2. Feature engineering list

Category	Description
Lag features	Lagged values for PM2.5, PM10, NO_2, CO, SO_2, O_3, temp, dew point, humidity, precipitation, wind direction, wind speed, pressure, and temp×rhum (1, 2, 3 lags; extended to 6, 12, 24 for PM2.5).
Rolling statistics	Rolling mean and standard deviation for PM2.5, temperature, and humidity over 3, 6, 12, and 24-hour windows.
Datetime features	Hour, day of the week, day of the month, month, and year extracted from the timestamp.
Cyclical features	Sine and cosine transformations of hour, day of the week, and month to capture periodic patterns.
Wind decomposition	Decomposed wind speed and direction into wind_u (east-west) and wind_v (north-south) components.
Interaction terms	Interaction feature combining temperature and humidity (temp × rhum), with 1, 2, and 3-hour lags.

model-agnostic measure of feature impact by estimating each variable's marginal contribution to a prediction. For each forecast model and horizon (T+1, T+3, T+6, T+12, T+24), we computed SHAP values on the test set and generated summary plots to visualize feature importance rankings.

Our approach employs SHAP analysis explicitly across multiple forecasting horizons, systematically revealing how feature contributions dynamically evolve with increasing forecast lead times. This multi-horizon interpretability analysis scientifically advances understanding beyond single-horizon interpretability, thus providing deeper insights into temporal dependencies and atmospheric-pollutant interactions.

Sensitivity analysis and optimization

In this study, sensitivity analysis involves systematically varying influential features based on SHAP insights, running predictive models under these variations (meteorological variables, pollutant, and time features), and assessing the impact on pollution levels relative to established air pollutant PM2.5 regulatory standards in Prishtina, Kosovo. This optimization approach will help identify effective and targeted pollution mitigation strategies aligned with forecast uncertainties.

Model statistics and evaluation

All experiments were conducted in a Python environment using well-known ML libraries such as Scikit-learn and SHAP. To ensure a robust evaluation of the prediction models, we employed a time-based split: the most recent 365 days were reserved for the test set, while the remaining data was used for training. This split simulates a realistic forecasting scenario in which future data is predicted based on historical observations. We compared two experimental setups: one using only basic features, and another incorporating the full set of engineered features to assess the performance gains from feature engineering. Model evaluation was based on three standard regression metrics, R^2, RMSE, and MAE: R^2 (coefficient of determination) - Indicates how much of the variation in the dependent variable can be explained by the independent variables.

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

MAE (mean absolute error) - evaluates the mean absolute disparity between predicted and actual outcomes, showing lower sensitivity to outliers than RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

RMSE gives greater weight to larger errors by squaring them, making it valuable when minimizing significant prediction mistakes is important.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

In addition to numerical results, we presented a series of visualizations and important charts to provide further insight into model behavior across different forecast horizons

RESULTS AND DISCUSSION

Model performance with basic features

The result in Table 3 showcases the performance of various machine learning models in predicting air quality using the original dataset with the basic features. This feature set includes the pollutant concentrations, meteorological variables, and timestamp-based indicators. In this setup, LightGBM outperforms other models, achieving R^2 values ≥ 0.8, the least MAE of 4.374, and an RMSE of 7.687, reflecting minimal predicting error. This outcome shows its effectiveness in capturing complex patterns without feature extraction as well. While RF and XGBoost perform well, particularly on smaller samples, they are inferior to LightGBM overall. Linear regression offers stable but lower predictive power, serving primarily as a baseline. SVM shows the weakest performance, especially as sample size increases, revealing difficulties with the data's complexity. Additionally, the performance decline with larger sample sizes in some models emphasizes the need to validate models using diverse datasets, reflecting the complexity of real-world air quality data.

Enhanced model performance with engineered features

Table 4 presents the ML models' performance metrics, showing that engineered features significantly improve results compared with basic features. For the 1-hour horizon, LightGBM and XGBoost exhibit similarly high performance, with R^2 values of 0.860 and 0.862, respectively, and correspondingly low MAE and RMSE: 3.520

Table 3. Results of the evaluation of performance with base features

Horizon	Model	R2	MAE	RMSE
t + 1	LightGBM	0.807	4.374	7.687
	XGBoost	0.736	5.284	9.003
	RandomForest	0.795	4.611	7.94
	LinearRegression	0.805	4.764	7.729
	SVM	0.665	5.064	10.135
t + 3	LightGBM	0.516	7.207	12.186
	XGBoost	0.364	8.507	13.97
	RandomForest	0.48	7.811	12.633
	LinearRegression	0.558	7.055	11.641
	SVM	0.459	6.938	12.88
t + 6	LightGBM	0.305	9.128	14.603
	XGBoost	0.065	11.767	18.074
	RandomForest	0.185	10.415	15.814
	LinearRegression	0.357	9.011	14.05
	SVM	0.299	8.351	14.668
t + 12	LightGBM	0.331	9.003	14.33
	XGBoost	0.165	10.31	16.007
	RandomForest	0.225	10.369	15.424
	LinearRegression	0.308	9.354	14.577
	SVM	0.274	8.658	14.923
t + 24	LightGBM	0.351	8.758	14.117
	XGBoost	0.266	9.347	15.007
	RandomForest	0.319	9.268	14.456
	LinearRegression	0.418	8.571	13.359
	SVM	0.410	7.864	13.451

and 6.546 for LightGBM, and 3.524 and 6.513 for XGBoost. This outcome indicates strong model reliability in explaining variance in the air-pollution datasets (Chen et al., 2024) (B.Raviteja and Reddy, 2024) (Zhou et al., 2024) and demonstrates both models' effectiveness in making short-term predictions.

For the 3-hour horizon, performance drops for all models, suggesting greater complexity in capturing air-quality dynamics over longer periods. Nevertheless, LightGBM remains the leading model, with an R^2 of 0.733, while the performance of XGBoost and Linear Regression declines notably.

At the medium horizon (6 hours ahead), performance declines further across all models. LightGBM still leads, but its R^2 falls to 0.668; the other models trail, with RF dropping to 0.602. Longer horizons of 8, 12, and 24 hours show substantial decreases in predictive accuracy, with LightGBM falling below 0.500 in R^2 at 24 hours, underscoring the challenge of maintaining reliability over extended forecasts. Linear Regression records the lowest performance across these horizons, with R^2 values of 0.432 and 0.499 at 24 hours. Comparatively, LightGBM remains the most reliable model across all horizons.

In Figure 3 we show the comparison of the hourly XGBoost predictions with the actual PM 2.5 concentrations for all of 2024 testing set. For most of the year, the two curves sit almost on top of each other, showing that the model captures both quick day-to-day changes and slower seasonal shifts. The dashed lines plot the 24-hour rolling averages, which also match closely. This means the model reproduces longer-term patterns, such as the winter smog in January and the late-autumn pollution build-up with almost no delay. The algorithm slightly underestimates a few extreme peaks above 250 $\mu g\ m^{-3}$ during the coldest weeks, but it still detects when these spikes begin and end. Most of the remaining errors come from the height of these peaks, not their timing. Overall, the plot supports the low error values

Table 4. Results of the evaluation of performance with engineered features

Horizon	Model	R2	MAE	RMSE
t + 1	LightGBM	0.860	3.520	6.546
	XGBoost	0.862	3.524	6.513
	RandomForest	0.833	3.890	7.160
	LinearRegression	0.846	4.108	6.871
	SVM	0.839	3.693	7.022
t + 3	LightGBM	0.733	5.169	9.060
	XGBoost	0.729	5.229	9.114
	RandomForest	0.706	5.526	9.499
	LinearRegression	0.642	6.766	10.482
	SVM	0.629	5.847	10.666
t + 6	LightGBM	0.668	6.080	10.093
	XGBoost	0.647	6.300	10.414
	RandomForest	0.602	6.739	11.049
	LinearRegression	0.540	7.963	11.884
	SVM	0.533	6.829	11.975
t + 12	LightGBM	0.623	6.651	10.761
	XGBoost	0.601	6.753	11.069
	RandomForest	0.578	7.139	11.374
	LinearRegression	0.500	7.977	12.391
	SVM	0.493	7.134	12.477
t + 24	LightGBM	0.499	7.683	12.406
	XGBoost	0.503	7.673	12.353
	RandomForest	0.498	7.738	12.407
	LinearRegression	0.432	8.581	13.205
	SVM	0.451	7.612	12.980

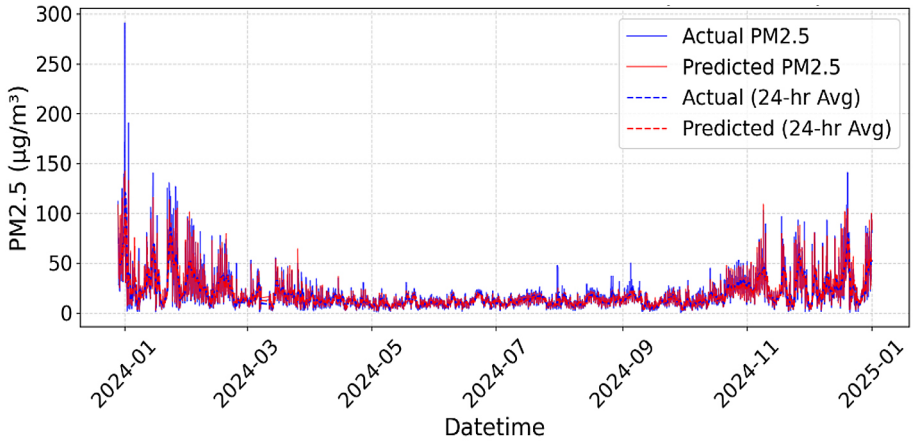

Figure 3. Actual versus predicted PM 2.5 24-hour ahead for the whole testing year 2024

in Table 3 and highlights XGBoost's ability to perform reliably across an entire year without noticeable drift or bias. The close alignment between actual and predicted values at the 24-hour horizon highlights the model's ability to deliver timely and accurate forecasts crucial for health advisories and pollution mitigation. Its effectiveness in capturing short-term variability makes it a valuable tool for real-time air quality monitoring, enabling swift responses to pollution spikes. Additionally, the smoothing effect of 24-hour averages emphasizes the need to consider various

temporal scales for thorough air quality assessments. Figure 4 presents a detailed weekly comparison plot comparing predictions made 1, 3, 6, 12, and 24 hours ahead. With only a 1-hour lead, the prediction line almost perfectly overlaps the real readings, while there is virtually no time shift or difference in height. At 3- and 6-hour leads, the peaks are only slightly lower (about 10–15%), and the timing shifts by at most one data point, so the main ups and downs still match.

The 12 and 24 hours predictions ahead smooth out the sharp swings and push the highest peaks about 3–4 hours later, yet they still capture the overall daily pattern: cleaner air in the early morning and worse air in the late afternoon. This steady loss of detail matches the gradual rise in RMSE listed in Table 3. Most of the added error comes from smaller peak heights and minor time shifts, not from a constant bias. In practice, this is helpful because the model can still warn the city of upcoming pollution several hours in advance, giving the environmental stakeholders time to respond.

Overall, the model's strong alignment of predicted and actual PM2.5 levels across different lead times shows its effectiveness in forecasting pollution trends. This reliability is essential for issuing health advisories and guiding mitigation efforts. Short-term predictions (1–3 hours) aid immediate health alerts, while longer-term forecasts (24–72 hours) assist in policy planning and resource allocation. The findings reflect the selected model's ability to replicate PM2.5 fluctuations, indicating it successfully captures key pollution drivers, meteorological conditions, emission patterns, and potentially episodic events. This understanding can guide targeted interventions (Luo and Gong, 2023); (Sanchez and Zhoual, 2024).

Feature importance and analysis

Figure 5 shows feature evolution, measured by mean absolute SHAP value changes across forecast horizons from 1 hour to 24 hours. At the shortest horizon (t+1h), the model's predictions

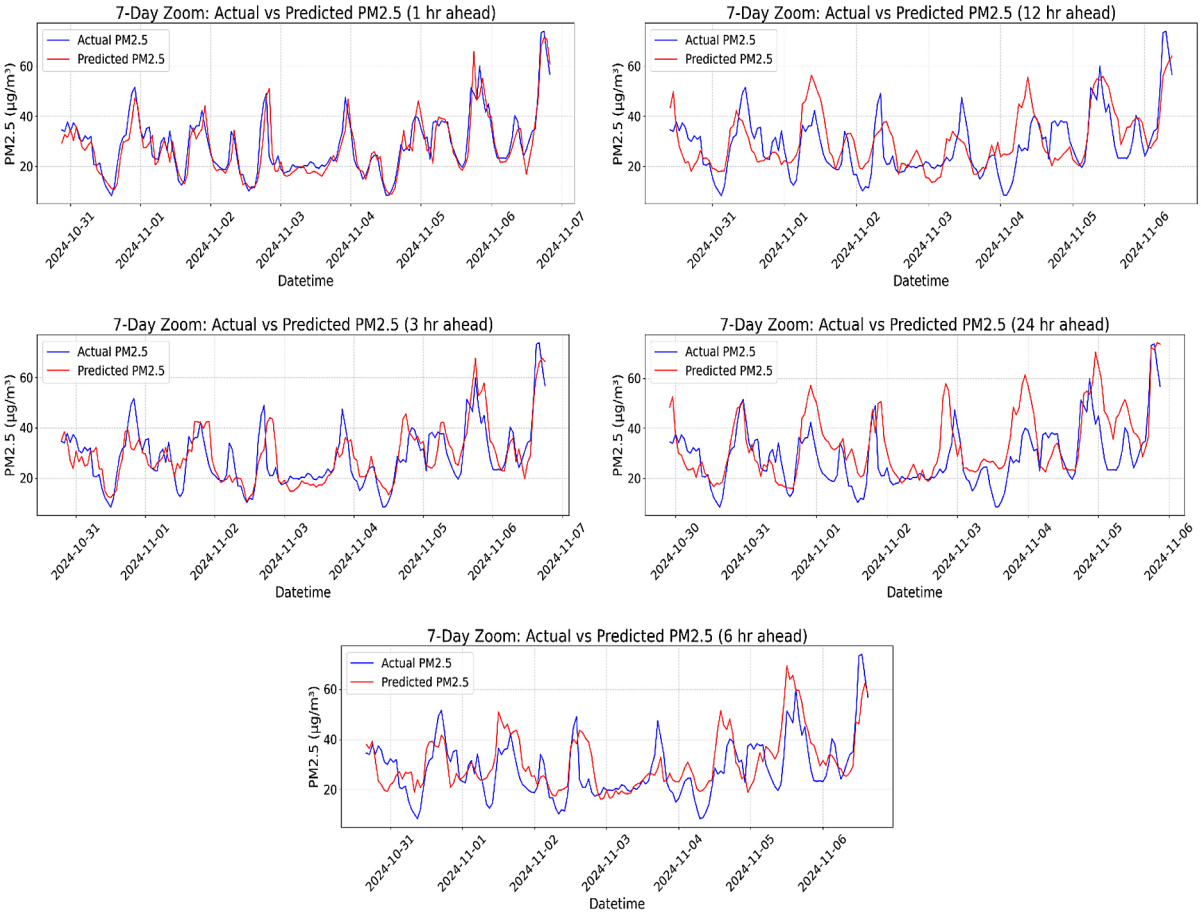

Figure 4. Actual versus predicted PM2.5 zoomed for a random 7 days in the testing set across different horizons setup

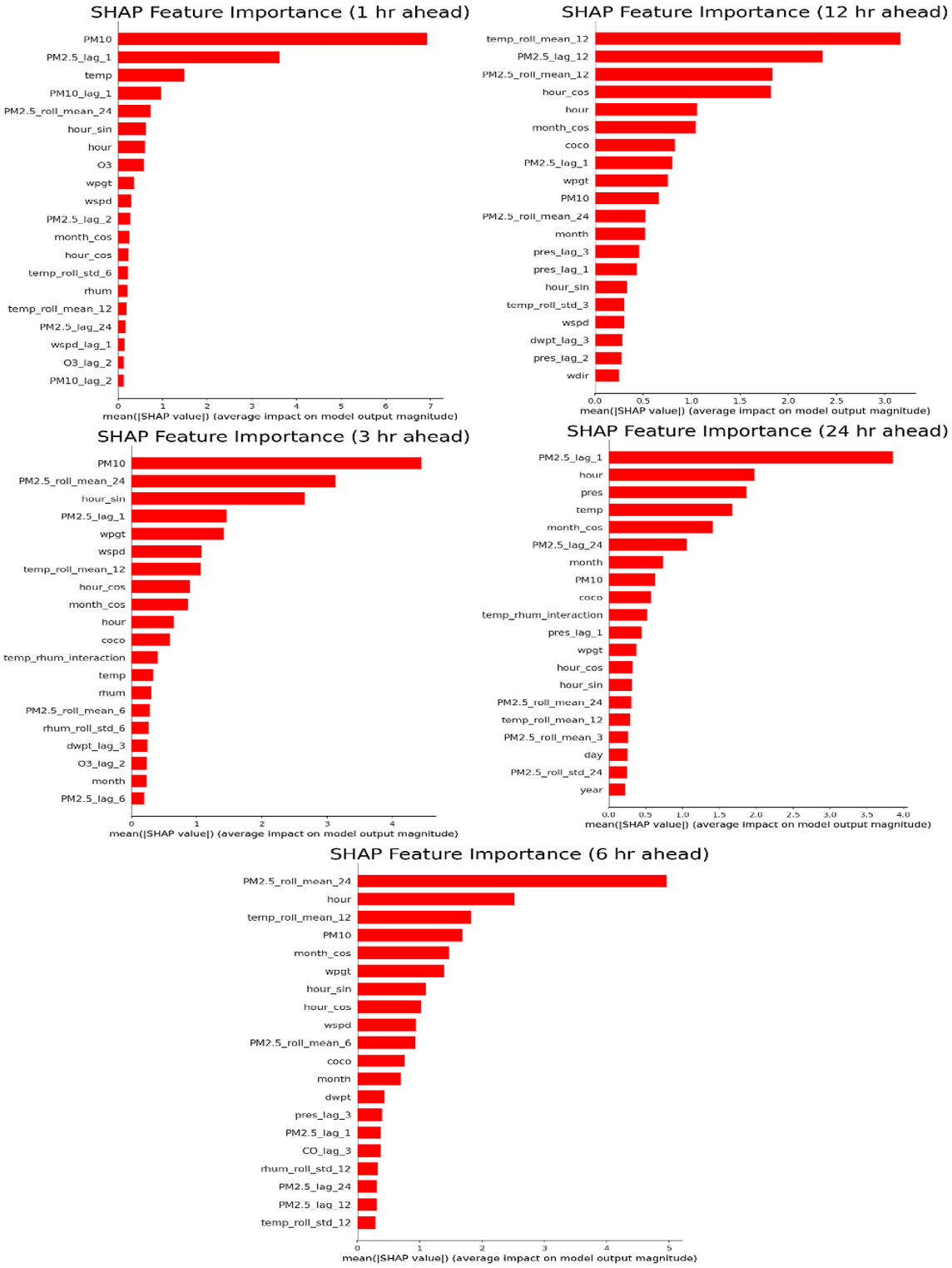

Figure 5. SHAP Feature importance for each horizon

are dominated by pollutant persistence. Specifically, the current value of PM10 and the one-hour lag of PM2.5 together explain more than half of the model's output variance. Temperature and PM10_lag_1 also contribute, but to a lesser extent, while other meteorological and cyclical features have minimal influence. This is expected, as fine particulate matter (PM2.5) levels tend to be highly autocorrelated over short periods, and PM10 often varies in parallel due to shared sources like traffic and heating.

As the forecast window increases to t+3h and t+6h, the influence of these short-term lags begins to decline. Instead, the 24-hour rolling mean of

PM2.5 becomes more important, and the hour-of-day (captured by its sinusoidal encoding) rises significantly in influence. Wind-related features, such as wind_gust and wind_speed, also gain relevance at this stage. By t+6h, the 24-hour PM2.5 rolling mean becomes the most influential feature overall. This outcome indicates that recent pollutant concentrations and immediate environmental conditions strongly influence near-future pollution levels (Cardito et al., 2023). At the same time, the 12-hour rolling mean of temperature overtakes the current temperature, and month_cos (a seasonal feature) begins to show impact. For longer forecasts such as t+12h, rolling statistics and cyclic patterns become even more prominent. The 12-hour rolling means of both temperature and PM2.5 top the list, followed by cyclic features and an interaction term between temperature and humidity (temp × rhum). The trend confirmed that metrological factors contribute more to longer-term variability in air pollution indices (Rahman and Meng, 2024; Wu et al., 2025), emphasizing the continued importance of recent pollutant levels and time-of-day effects.

PM10's importance drops to mid-range, indicating that it is less useful when the model can no longer rely on short-term persistence. At the 24-hour horizon, PM2.5_lag_1 once again becomes the top feature, likely due to its role in capturing broader temporal trends. It is followed by hour, pressure, and temperature, all of which are key variables for understanding daily air-quality cycles, including nighttime pollutant buildup and morning dispersion. Wind-related features become much less important at this horizon, as wind patterns are harder to predict a full day.

Overall, the model's decision-making shifts from relying heavily on immediate pollutant levels (68% of SHAP value weight at t+1h) to emphasizing rolling averages and cyclic descriptors. Meanwhile, the contribution of meteorological variables gradually increases to about 25%. This transition reflects established atmospheric dynamics and highlights two key operational takeaways: (i) near-term alerts require high-frequency PM data, and (ii) accurate weather forecasts and well-curated historical data are critical for effective day-ahead planning.

Moreover, the shown feature-importance evolution across different horizons in Figure 6 reveals how the model progressively re-weights its evidence as the forecast window expands. Rapid-decay variables as PM10 and, to a lesser extent, PM2.5_lag_1, start as the most influential signals but lose most of their explanatory weight within the first 6 h, reflecting the quick erosion of short-term persistence. Bridge variables such as PM2.5_roll_mean_24 and temp_roll_mean_12 follow an inverted-U pattern, whereas their relevance is modest at 1 h, peaks around the 6- to 12-h window when multi-hour memory is most valuable, and then tapers off as the horizon lengthens. Finally, slow-cycle descriptors as the clock index hour and its seasonal counterpart month_cos exhibit monotonic growth, moving from peripheral roles at 1 h to primary drivers in the 24-h forecast, where diurnal stability and seasonal background dominate pollutant behavior. These gradual shifts between the horizons mirror how air pollution unfolds over longer periods.

Our novel findings from the systematic SHAP multi-horizon analysis demonstrates progressive

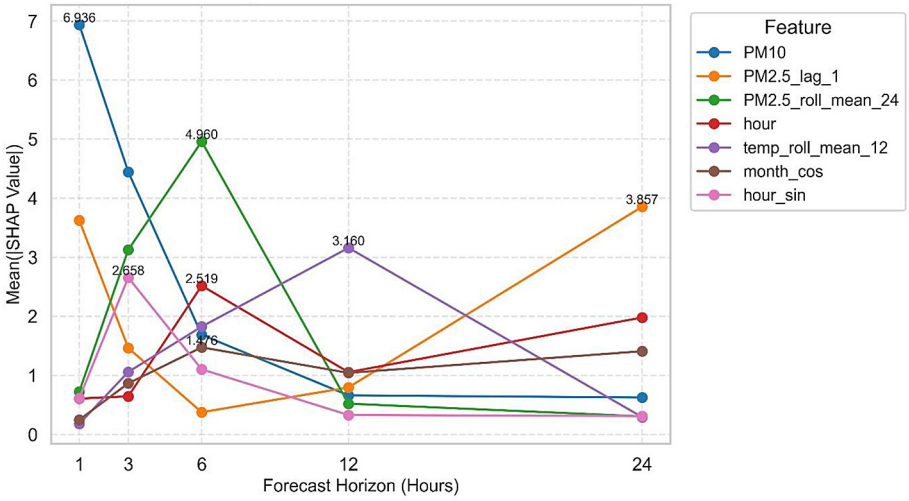

Figure 6. SHAP Feature importance evolution across different horizons

temporal shifts in feature importance, from immediate pollutant persistence (short-term horizons), toward rolling averages and cyclical meteorological conditions (mid- and long-term horizons). Explicitly characterizing this temporal evolution of features significantly advances scientific understanding of air pollution dynamics and provides insights broadly transferable to other urban environments experiencing similar meteorological and pollution dynamics.

Sensitivity analysis on metrological variables on PM2.5 pollution

In this study, a sensitivity analysis was implemented based on the outputs from the selected XGBoost and feature engineering (SHAP) analysis to assess how metrological factors affect variation in concentrations of PM2.5 changes from January to December in 2024. For the case study in Prishtina, the analysis of prediction results from the XGBoost and SHAP analysis reveals that PM2.5 was the most influential air pollutant in this case study compared to CO, SO_2, NO_2, and PM10. Additionally, the findings from SHAP analysis suggest that the dynamics of metrological variables, including wind speed, temperature, wind direction, and time variations, are key drivers of PM2.5 concentration (Rahman & Meng, 2024; Wu et al., 2025),accounting for possible air pollution in this case study. The meteorological variables temperature, wind speed, wind direction, and time were varied to 2 levels representing maximum and minimum extremes of temperature, wind speed, wind direction, and time. The targeted output is the corresponding concentration values for PM2.5 were recorded at these points extremes were recorded.

4FI model statistics

To investigate the sensitivity analysis on metrological variables on PM2.5 pollution levels, the 4FI factorial model was selected. The 4LI was adopted to predict possible combined effects or antagonistic and synergetic interaction of the four meteorological variables that can account for the PM2.5 concentration at a 98% power level, with a model lack-of-fit (LoF = 3), pure error output of 7, and degree of freedom (d.f = 4). A minimal Lof ≥ 3 relative to pure error of 7 confirmed that the selected 4FI model is valid with minimal signal-to-noise ratio.

The Pareto charts in Figures 7(a-f) analyze the ranking effect of meteorological variables temperature (a), wind speed (b), time (c), and wind direction (D) on the concentration of PM2.5 emitted during quarterly short-term January, April, August, December, and long term prediction for the year 2024. Understanding these interactions helps forecast pollution episodes and design effective air quality management plans during high pollution peaks.

In Figure 7a, temperature exhibits the highest t-value (9.65), far exceeding the Bonferroni and t-value cutoff lines, while a lower t-value (2.4), just above the t-value limit (2.13) but below the Bonferroni threshold (3.48). The statistical analysis reveals a moderately significant negative effect of wind speed and wind direction on PM2.5 levels. This finding confirms that variations in time during January are statistically associated with changes in PM2.5 levels. Time also shows a strong positive effect on PM2.5 levels during January, indicating that pollution tends to rise as the month progresses. The model analysis suggests that wind direction is less influential (Liu et al., 2020); thus, PM2.5 pollution levels tend to rise as the month advances, possibly due to weather patterns, reduced wind, and temperature inversions leading to the accumulation of pollutants over the month 30 (Ilenič et al., 2024).

A similar outcome was recorded in April, as confirmed in the Pareto chart recorded in the month April. Figure 7(b) shows that temperature (A) has the highest, statistically significant t-value (10), exceeding the Bonferroni threshold and the t-test limit. The outcome suggests a highly significant effect. This strongly indicates that temperature significantly impacts PM2.5 levels in Prishtina; higher temperatures correlate with increased PM2.5, while wind disperses pollutants. Time and wind direction have statistically insignificant or minimal effects on April's PM2.5 levels (Rahman and Meng, 2024). This outcome is confirmed by the probability output of wind speed, time, and wind direction cluster near zero scale, indicating smaller deviations and thus lesser significance.

In Figure 7(c), temperature stands out significantly with a t-value of 4.14, surpassing both the Bonferroni limit (3.4102) and the t-value threshold (2.10982). This outcome suggests that temperature has a statistically significant and dominant effect on PM2.5 levels, likely due to

its influence on atmospheric stability and pollutant dispersion.

Time shows a moderate impact with a t-value just above 1.04, suggesting a potential secondary role related to diurnal changes affecting emissions. Wind speed and wind direction have t-values near zero, indicating negligible effects on PM2.5, likely due to consistently low wind speeds and minimal directional variation in August. Thus, temperature is the primary driver of PM2.5 variability in Prishtina, and the other meteorological variables, including wind speed and direction, have minimal effects in August.

The Pareto chart in Figure 7(d) confirmed that temperature has the strongest influence on PM2.5 changes, showing a negative effect with a t-value near 9.00, well above the Bonferroni limit (3.48). This suggests that higher temperatures

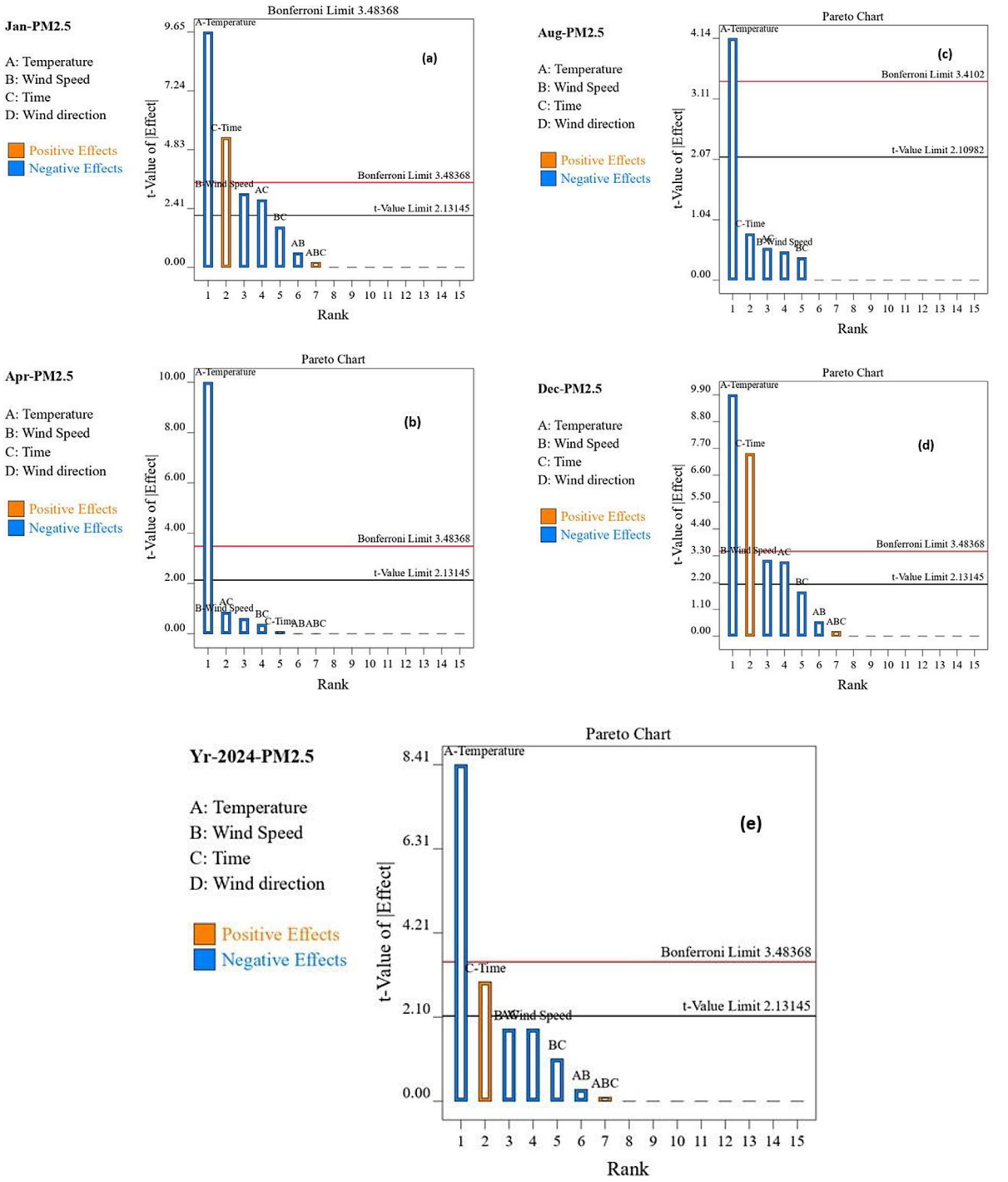

Figure 7. Pareto chart distribution of the ranking of metrological variables on PM2.5

may reduce PM2.5 due to improved particulate mixing in the atmosphere. Time also has a significant positive effect and a t-value close to 6.00, indicating peak traffic hours, and daily industrial activity influencing PM2.5 concentrations may impact PM2.5 levels. However, wind speed has a moderate positive effect (t = 3.50), suggesting that higher wind speeds may promote the mixing of particulates during this extensive period, contributing to PM2.5 levels. The Pareto ranking indicated likely synergetic effects of the meteorological variables temperature-time (A*B) and wind speed-time (B*C), both showing negative impacts, with t-values of 3.00 and 2.00. This suggests that temperature fluctuations at certain times may reduce PM2.5 pollution in Prishtina during December. We reasoned that policymakers and environmental agencies should consider implementing air quality mitigation strategies in early December and late January to reduce health risks from high PM2.5 levels, focusing on hot periods and pollution peaks.

The Pareto chart (Figure 7e) for long-term model prediction shows that temperature has the highest t-value of 8.41 and a strong negative effect, confirming that higher temperatures significantly reduce PM2.5 levels due to enhanced atmospheric mixing. Time follows with a t-value of around 3.5 and a positive effect, suggesting that daily or seasonal changes increase PM2.5 levels. The outcome possibly reflects human activities, including commuting, heating, or industrial output. Findings confirmed wind speed has a borderline significant t-value of approximately 2.1 and also has a positive effect, possibly related to particulate resuspension, and a possible built environment contributing to leading to pockets of stagnation where PM2.5 pollutants accumulate, regardless of general wind speed (Ilenič et al., 2024; Wu et al., 2025). Wind direction did not significantly impact PM2.5 levels in Prishtina. The combined effects of wind speed-time (B*C) and temperature-wind speed (A*B) are minimal, with t-values below significance thresholds, indicating that variables rarely operate in isolation. The three-way interaction of temperature-wind speed-time (A*B*C) has a negligible impact, suggesting that the combination of all four variables is merely an alias and does not create a meaningful influence on PM2.5 concentration beyond their individual or two-way effects. This indicates a lack of a compound meteorological scenario.

The final prediction of the air quality following the ML, SHAP, and sensitivity analysis is presented in Table 5. These outputs were recorded following model equations 4–8.

$$Jan - PM2.5 = 26.8 - 2.1Temp - 0.6Time - \\ - 0.6Speed - 0.12Temp \cdot Speed - 0.6 - \\ - 0.59Temp \cdot Time - 0.35Speed \cdot Time + \\ + 0.04Temp \cdot Speed \cdot Time \quad (4)$$

$$Apr - PM2.5 = 15 - 2.1Temp - \\ - 0.13Time - 0.02Speed - 0.18Temp \cdot \\ \cdot Time - 0.08Speed \cdot Time - 1.2Temp \cdot \\ \cdot Speed \cdot Time \quad (5)$$

$$Aug - PM2.5 = 12 + 0.87Temp + \\ + 0.1Time + 1.6Speed - 0.16Temp \cdot \\ \cdot Speed - 0.1Speed \cdot Time - \\ - 0.08Speed \cdot Time \quad (6)$$

$$Dec - PM2.5 = 26.8 - 2.1Temp - \\ - 0.6Time - 0.6Speed - 0.12Temp \cdot \\ \cdot Speed - 0.6 - 0.59Temp \cdot Time - \\ - 0.35Speed \cdot Time + 0.04Temp \cdot \\ \cdot Speed \cdot Time \quad (7)$$

$$YrPM2.5 = 19.8 - 1.8Temp + 0.39Time - \\ -0.63Speed - 0.063Temp \cdot Speed - \\ -0.23Speed \cdot Time + 0.02Temp \cdot \\ \cdot Speed \cdot Time \quad (8)$$

From Table 5, the short-term predicted PM2.5 outputs for January, April, August, and December correspond to 26, 14, 12, and 24, respectively. However, the long-term prediction for the year 2024 translates to a PM2.5 concentration of 19. These predicted outcomes, when compared

Table 5. Prediction analysis output for PM2.5

Response	Predicted Mean	Std Dev	SE Mean	95% CI low for Mean	95% CI high for Mean	95% TI low for 99% Pop	95% TI high for 99% Pop
Jan-PM2.5*	26	0.1	20.39	25.7594	25.7594	25.7594	25.7594
Apr-PM2.5*	14	1.0	103.58	13.7964	13.7964	13.7964	13.7964
Aug-PM2.5*	11.4528	2.4	18.20	11.4528	11.4528	11.4528	11.4528
Dec-PM2.5*	24.2984	0.05	181.97	24.2984	24.2984	24.2984	24.2984
Yr-2024-PM2.5*	18.8268	1	88.46	18.8268	18.8268	18.8268	18.8268

with the Environmental regulatory standard, established that the predicted PM2.5 outcomes for January (26 μg/m³) fall within the range of 25–50 μg/m³ of the Environmental Monitoring Standards, suggesting poor air quality at the beginning of the year. The PM2.5 output for Apr (14 μg/m³) falls within the range of 10–20 μg/m³ categorized as fair air quality. A similar outcome can be observed for the predicted output PM2.5 output for August (11.45 μg/m) also falls within the range of 10–20 μg/m³ categorized as a fair standard for air quality.

The prediction output for December confirmed PM2.5 value of 24 μg/m³. This value lies within the range of 20–25 μg/m³, rated as moderate air quality with regards to PM2.5. Overall, the long-term output for the year 2024 corresponds to 19 μg/m³ which lies within tolerable range of 20–25 μg/m³.

Implications for air quality policy

The multi-horizon modeling framework presented here gives Prishtina a ready-to-use ML pipeline that fits directly into its air-quality management system. Predictions one hour ahead achieve a mean absolute error below 4 μgm^{-3} and an R^2 of about 0.86, and they still keep reasonable skill ($R^2 = 0.50$) a full day ahead.

This accuracy at different lead times lets the city target its actions. Near-real-time alerts can trigger quick steps such as retiming traffic lights, banning heavy vehicles temporarily, or sending SMS warnings to high-risk residents. Six- to twelve-hour forecasts give utilities time to fine-tune district-heating output and public transport schedules. Forecasts 24 hours ahead support planned initiatives like "low-emission Sundays" or shifting outdoor school activities.

Each forecast comes with clear SHAP explanations. This helps policymakers explain their choices to the public and later check whether those actions reduce forecast errors. The year-long back-test on 2024 data shows stable performance with little drift and good seasonal capture, so the system can serve as an ongoing decision-support tool rather than a one-off study.

Beyond practical decision-making relevance, this study's novelty specifically lies in explicitly quantifying and interpreting the temporal evolution of feature importance across different prediction horizons using a variety of tuned ML models. This explicitly demonstrates how real-time data interpretability can clarify previously unknown temporal dependencies among meteorological and pollutant variables, significantly enhancing the understanding of air pollution dynamics in urban areas.

Overall, these strengths move the city from reactive, threshold-based warnings to a proactive, evidence-driven approach where both daily operations and long-term planning rely on the same interpretable ML pipeline.

CONCLUSIONS

This study presents a novel, interpretable multi-horizon forecasting framework that integrates advanced machine learning, systematic temporal feature engineering, and SHAP-based interpretability. A key contribution is the explicit quantification of dynamic feature shifts across forecast horizons, an area previously underexplored. This enables a deeper understanding of air pollution dynamics and feature interactions over time.

Validated through comprehensive experiments, the framework proves both effective and scalable. LightGBM and XGBoost consistently deliver strong performance across all horizons, handling high-dimensional, noisy input data with minimal feature engineering. Enhancements using simple lags and rolling means further improve accuracy. The models' ability to capture non-linear patterns surpasses linear and kernel-based baselines.

SHAP explanations provide transparent, actionable insights, supporting early-warning systems and informed policy-making. Overall, the developed interpretable multi-horizon forecasting framework explicitly balances predictive accuracy with interpretability, offering scientifically novel insights into temporal pollutant dynamics, and presenting broad applicability for urban air quality forecasting worldwide.

REFERENCES

1. Anu Priya, S., Khanaa, V. (2023). *An Intelligent Air Quality Prediction System Using Neuro-Fuzzy Temporal Classifier with Spatial Constraints* 161–175. https://doi.org/10.1007/978-3-031-23683-9_11
2. Raviteja, B. P. T., Reddy, U. S. (2024). Air quality prediction using machine learning. *International Journal For Multidisciplinary Research*, *6*(2). https://doi.org/10.36948/ijfmr.2024.v06i02.17192

3. Cardito, A., Carotenuto, M., Amoruso, A., Libralato, G., Lofrano, G. (2023). Air quality trends and implications pre and post Covid-19 restrictions. *Science of The Total Environment*, *879*, 162833. https://doi.org/10.1016/j.scitotenv.2023.162833
4. Chen, F., Zhang, W., Mfarrej, M. F. B., Saleem, M. H., Khan, K. A., Ma, J., Raposo, A., Han, H. (2024). Breathing in danger: Understanding the multifaceted impact of air pollution on health impacts. *Ecotoxicology and Environmental Safety*, *280*, 116532. https://doi.org/10.1016/j.ecoenv.2024.116532
5. Ghorani-Azam, A., Riahi-Zanjani, B., Balali-Mood, M. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. *Journal of Research in Medical Sciences*, *21*(1), 65. https://doi.org/10.4103/1735-1995.189646
6. Halaktionov, M., Bredun, V., Choudhary, R., Goroneskul, M., Kumar, A., Ouiya, F., Sydorenko, V., Markina, L. (2025). AI-Enhanced air quality assessment and prediction in industrial cities: A case study of Kryvyi Rih, Ukraine. *Ecological Engineering & Environmental Technology*, *26*(6), 45–56. https://doi.org/10.12912/27197050/203725
7. Huang, L., Duan, Q., Liu, Y., Wu, Y., Li, Z., Guo, Z., Liu, M., Lu, X., Wang, P., Liu, F., Ren, F., Li, C., Wang, J., Huang, Y., Yan, B., Kioumourtzoglou, M.-A., Kinney, P. L. (2025). Artificial intelligence: A key fulcrum for addressing complex environmental health issues. *Environment International*, *198*, 109389. https://doi.org/10.1016/j.envint.2025.109389
8. Ilenič, A., Pranjić, A. M., Zupančič, N., Milačič, R., Ščančar, J. (2024). Fine particulate matter (PM2.5) exposure assessment among active daily commuters to induce behaviour change to reduce air pollution. *Science of The Total Environment*, *912*, 169117. https://doi.org/10.1016/j.scitotenv.2023.169117
9. Kedar, M. M. M. (2024). Exploring the Effectiveness of SHAP over other Explainable AI Methods. *Interantional Journal of Scientific Research in Engineering and Management*, *8*(6), 1–5. https://doi.org/10.55041/IJSREM35556
10. Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., Geng, Y. (2022). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, *276*, 106238. https://doi.org/10.1016/j.atmosres.2022.106238
11. Liu, Z., Shen, L., Yan, C., Du, J., Li, Y., Zhao, H. (2020). Analysis of the Influence of Precipitation and Wind on PM2.5 and PM10 in the Atmosphere. *Advances in Meteorology*, *2020*, 1–13. https://doi.org/10.1155/2020/5039613
12. Luo, J., Gong, Y. (2023). Air pollutant prediction based on ARIMA-WOA-LSTM model. *Atmospheric Pollution Research*, *14*(6), 101761. https://doi.org/10.1016/j.apr.2023.101761
13. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00014
14. McClarren, R. G. (2021). Decision trees and random forests for regression and classification. In *Machine Learning for Engineers* 55–82. Springer International Publishing. https://doi.org/10.1007/978-3-030-70388-2_3
15. Nakyai, T., Santasnachok, M., Thetkathuek, A., Phatrabuddha, N. (2025). Influence of meteorological factors on air pollution and health risks: A comparative analysis of industrial and urban areas in Chonburi Province, Thailand. *Environmental Advances*, *19*, 100608. https://doi.org/10.1016/j.envadv.2024.100608
16. Olawade, D. B., Wada, O. Z., Ige, A. O., Egbewole, B. I., Olojo, A., Oladapo, B. I. (2024). Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions. *Hygiene and Environmental Health Advances*, *12*, 100114. https://doi.org/10.1016/j.heha.2024.100114
17. Pan, B. (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*, *113*, 012127. https://doi.org/10.1088/1755-1315/113/1/012127
18. Persis, J., Ben Amar, A. (2023). Predictive modeling and analysis of air quality – Visualizing before and during COVID-19 scenarios. *Journal of Environmental Management*, *327*, 116911. https://doi.org/10.1016/j.jenvman.2022.116911
19. Quang, T. Van, Doan, D. T., Ngarambe, J., Ghaffarianhoseini, A., Ghaffarianhoseini, A., Zhang, T. (2025). AI management platform for privacy-preserving indoor air quality control: Review and future directions. *Journal of Building Engineering*, *100*, 111712. https://doi.org/10.1016/j.jobe.2024.111712
20. Rahaman, M., Southworth, J., Amanambu, A. C., Tefera, B. B., Alruzuq, A. R., Safaei, M., Hasan, M. M., Smith, A. C. (2025). Combining deep learning and machine learning techniques to track air pollution in relation to vegetation cover utilizing remotely sensed data. *Journal of Environmental Management*, *376*, 124323. https://doi.org/10.1016/j.jenvman.2025.124323
21. Rahman, M., Meng, L. (2024). Examining the spatial and temporal variation of PM2.5 and its linkage with meteorological conditions in Dhaka, Bangladesh. *Atmosphere*, *15*(12), 1426. https://doi.org/10.3390/atmos15121426
22. Ramírez, A. S., Ramondt, S., Van Bogart, K., Perez-Zuniga, R. (2019). Public awareness of air pollution and health threats: challenges and opportunities for communication strategies to improve environmental health literacy. *Journal of Health Communication*,

24(1), 75–83. https://doi.org/10.1080/10810730.2019.1574320

23. Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, *338*, 139518. https://doi.org/10.1016/j.chemosphere.2023.139518
24. Reddy, G. P., Kumar, Y. V. P. (2023). Explainable AI (XAI): Explained. *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (EStream)*, 1–6. https://doi.org/10.1109/eStream59056.2023.10134984
25. Shetty, C., Seema, S., Sowmya, B. J., Nandalike, R., Supreeth, S., P., D., S., R., Y., V., Ranjan, R., Goud, V. (2024). A machine learning approach for environmental assessment on air quality and mitigation strategy. *Journal of Engineering*, *2024*(1). https://doi.org/10.1155/2024/2893021
26. Suárez Sánchez, A., García Nieto, P. J., Riesgo Fernández, P., del Coz Díaz, J. J., Iglesias-Rodríguez, F. J. (2011). Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Mathematical and Computer Modelling*, *54*(5–6), 1453–1466. https://doi.org/10.1016/j.mcm.2011.04.017
27. Tsokov, S., Lazarova, M., Aleksieva-Petrova, A. (2022). A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. *Sustainability*, *14*(9), 5104. https://doi.org/10.3390/su14095104
28. Wilson, J. R., Lorenz, K. A. (2015). *Standard Binary Logistic Regression Model* 25–54. https://doi.org/10.1007/978-3-319-23805-0_3
29. Wu, B., Zhao, S., Liu, Y., Zhang, C. (2025). Do meteorological variables impact air quality differently across urbanization gradients? A case study of Kaohsiung, Taiwan, China. *Heliyon*, *11*(2), e41694. https://doi.org/10.1016/j.heliyon.2025.e41694
30. Zhou, W., Yan, Z., Zhang, L. (2024). A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction. *Scientific Reports*, *14*(1), 5905. https://doi.org/10.1038/s41598-024-55243-x