ACCSCIENCE PUBLISHING

Explora: Environment and Resource

ORIGINAL RESEARCH ARTICLE

An explainable hybrid stacked deep learning framework for forecasting PM10 concentrations in urban air

Syed Azeem Inam*, Haider Rajput, and Saddam Umer

Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Sindh, Pakistan

Abstract

Accurate and explainable forecasting of particulate matter (PM10) is increasingly essential for managing urban air quality and protecting public health. This study proposed and evaluated a hybrid stacked deep learning architecture designed to enhance PM10 and urban air quality forecasting accuracy and to provide transparent explanations for its predictions. Using a self-designed neural network and Ridge regression (the meta-learner), PM10 prediction was accomplished based on LightGBM integration. Analysis was performed on the World Air Quality Index dataset, consisting of 1.8 million observations from 380 cities globally. To demonstrate its effectiveness, the hybrid model was benchmarked against traditional time series models (Autoregressive Integrated Moving Average [ARIMA] and Seasonal ARIMA) and machine learning models, including decision tree, extreme gradient boosting, random forest, and neural network, using the mean squared error (MSE), root MSE (RMSE), mean absolute error (MAE), and R^2 metrics as evaluation metrics. Model explainability was accomplished using Shapley Additive Explanations and Local Interpretable Model-Agnostic Explanations analyses. The hybrid model achieved an R^2 of 0.9916, MSE of 4.90, RMSE of 2.21, and MAE of 0.992, surpassing the other models' performances and demonstrating strong reliability. The analysis determined the seven-day PM10 lag as the most important influential predictor, while other spatial parameters contributed minimally. The model's ability to run efficiently on general-purpose computers further ensures accessibility for resource-constrained agencies. Overall, this study demonstrates the high predictive accuracy and interpretability of the proposed hybrid framework, offering a practical and informative tool for policymakers to improve air quality and public health outcomes.

Keywords: Hybrid stacked model; Air quality; LightGBM; PM10; Shapley Additive Explanations; Local Interpretable Model-Agnostic Explanations

***Corresponding author:**
Syed Azeem Inam
(syed.azeem@smiu.edu.pk)

Citation: Inam SA, Rajput H, Umer S. An explainable hybrid stacked deep learning framework for forecasting PM10 concentrations in urban air. *Explora Environ Resour.* 2025;2(4):025380069. doi: 10.36922/EER025380069

Received: September 14, 2025

Revised: October 16, 2025

Accepted: November 10, 2025

Published online: November 27, 2025

1. Introduction

Addressing public health, policy development, and air quality monitoring, particulate matter (PM) prediction has gained prominence and captured the interest of academic and applied scientists.[1,2] Growing concerns stem from the ease with which PM can penetrate the human respiratory system, leading to long-term health issues. The inhalation of particulates is strongly correlated with increasing rates of respiratory

failure, cardiovascular decline, asthma, lung cancer, and chronic bronchitis.[3] Rising urbanization and unchecked industrial growth exacerbate the problem, contributing to elevated PM concentrations across metropolitan regions. This trend affirms the necessity of sophisticated analytical prediction tools capable of forecasting PM levels and guiding effective mitigation strategies.

Predictive analytics help recognize and forecast pollution events, enabling the targeted allocation of resources to minimize public health risks. This has to be done with specific predictive models that inform regulations, establish international standards, identify sensitive monitoring regions, and support the implementation of cost-effective public health schemes at the appropriate time to reduce pollution.[4]

Traditional air quality forecasting approaches have illuminated the value of statistical and deterministic models, particularly through linear regression and time series methods, alongside remaining frameworks deployed to provide baseline insights and real-time alerts.[5] However, these approaches disallow the forecasting of linear, static, and homogenous growth, while presuming the existence of boundaries within error variance, creating boundaries to the intricacy and randomness of the shifting atmospheric conditions.[6] Consequently, they often fail to accurately estimate the effects of complex and fluctuating emissions from various human activities, varying forecasting intervals, uneven human-made weather conditions, and the spatiotemporal conjunction of intervals. This limitation is particularly pronounced in sprawling hot zones over cities.

Therefore, there is increasing demand for more flexible and robust models that better capture the drivers of air pollution episodes and their non-linear behavior.[5,6] In recent years, many scholars and researchers have leaned toward the prowess of modern developments in machine learning (ML), and more recently, deep learning, which excel at discovering complex non-linear facets in massive datasets.[7] These approaches are in sharp contrast to those deployed in the past, which worked best with changes to the system and parameters. The modern models are more likely to provide sharper forecasts by adapting to evolving environmental factors that typically define domain boundaries.

Despite the continuing importance of classic algorithms, such as support vector machines (SVM) and random forest (RF), the gradient-boosted family of algorithms, in particular the light gradient boosting machine (LightGBM) and tailor-made neural networks (NNs), has emerged as perhaps the most profound innovations in the field.[8]

Many custom NNs have employed deep learning to improve analysis on complex time-series datasets, especially when paired with complex structures. Such models excel at untangling the deep-seated non-linear and sequential patterns that characterize long-term environmental records.[1] By separating spatiotemporal features into individual units, these models, such as Recurrent NN (RNN) and attention-based models, effectively process time-based data. The resulting dataset captures daily fluctuations, seasonal patterns, and the combined influences of meteorological conditions, emissions, and geographical factors on PM10 concentration variations.

Hence, combining LightGBM with a custom-built NN forms an unprecedented hybrid approach that synergistically leverages the advantages of both models. The LightGBM model accelerates the interaction among mixed data types, while the neural layers grasp more profound and complex spatiotemporal patterns.[1,6,8] This combined architecture monitors persistent patterns and momentary surges in pollutant concentrations, protects against the overfitting that one or the other component usually suffers from, reduces prediction errors, and enhances stability amid environmental changes or added noise.[5,7] More importantly, the integrated system provides richer, more accurate, and more robust predictions, fulfilling the high standards of contemporary PM10 monitoring.[6,7]

Given the impact that PM10 predictions may have on public health and air quality policies, PM10 modeling is no longer optional. To explain the complex behavior of advanced ML and deep learning models, it is now common for researchers to use Shapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP, grounded in cooperative game theory, helps identify which features (e.g., wind speed and temperature) contribute most to PM10 predictions, revealing the dominant factors controlling atmospheric PM10 levels and their interactions.[9] In contrast, LIME provides explanations focused on individual predictions, highlighting the specific features driving the model's output for a single observation. These approaches enhance prompt responses on a case-by-case basis.[8] Together, they not only enhance overall model transparency and build user trust but also enable routine debugging by revealing hidden biases, informing iterative improvements by flagging blame misallocations, and directing researchers to parameters that need adjustment.[7]

This work develops and validates a hybrid model for short-term forecasting of airborne PM10 levels, combining LightGBM with a custom NN specifically focused on urban areas. During this process, specific attention is

paid to areas with heavy traffic, high urbanization, and PM10 concentrations exceeding health-based guidelines, enabling precise and timely predictions vital for planning and mitigation. The model is also interpretable because it is integrated with *post hoc* tools, such as SHAP and LIME, which improve explanation-based reasoning, enabling stakeholders to access actionable recommendations. These reasons, alongside improving model ease, accuracy, and transparency, motivate this work to protect public health in urban centers globally.

2. Literature review

Recent advancements in artificial intelligence (AI) techniques have redefined air quality modeling by shifting the field from traditional rule-based approaches to more adaptive, data-centric systems.[10-13] Such a paradigm shift is aptly demonstrated by the ever-increasing accuracy of PM10 forecasts, which is a vital metric for urban health assessment. In this section, we review a spectrum of PM10 forecasting methods, comparing classical regressors with contemporary ML models. We delineate the advantages of competing approaches, such as LightGBM, deep NNs (DNNs), and advocate for a transparent stacked model that integrates these strengths while employing SHAP and LIME to elucidate the opacity of prediction.[14]

The use of linear regression, decision trees (DT), and RFs, which have traditionally been utilized to predict PM10 and other air quality metrics, is outdated.[15] Such static methods that are encapsulated by posited equations are prone to failures when used with non-linear and highly interrelated time-series data, as well as spatiotemporally overlapping time-series data. By contrast, more recent ML techniques are more accurate and more robust in handling these complexities. However, these gains come with additional hurdles, such as the requirement to process model reasoning behind forecasts and the massive computational resources needed to train deep networks on extensive datasets.[16]

Recent studies have focused on implementing various ML models for PM10 forecasting. Kujawska *et al*.[17] concluded that SVM, artificial NNs, and Kriging models effectively capture non-linear relationships between meteorological parameters and pollution concentrations. Similarly, Lei *et al*.[18] utilized ensemble methods based on DTs and reported comparable predictive improvements. Furthermore, hybrid deep learning architectures that usually amalgamate RNNs and convolutional NNs have proven effective in separating temporal lags from spatial gradients in convolutively structured air quality time series.[23] While SVMs excel at resolving complex boundaries and RFs mitigate overfitting, each method has limitations, such as sensitivity to kernel tuning and long learning times typical of NNs.[17-20]

Within the realm of environmental science, LightGBM is one of the fastest-growing algorithms due to its high speed, ease of training on large datasets, and scalability.[24] It consistently outperforms predecessors, such as extreme gradient boosting (XGB) and RFs by effectively capturing sophisticated non-linear relationships between environmental and socio-economic factors, while achieving superior speed and accuracy.[18] LightGBM's computational efficiency is one of the main reasons it is used for near-real-time air quality forecasting and management applications.[16] NNs, such as DNNs, convolutional NNs, and RNNs, are also used in environmental forecasting to address the extreme non-linearity and time-variance of PM10 concentrations.[19] Long short-term memory (LSTM), an example of RNNs, is extremely valuable for time series applications due to its ability to process sequential data. Although custom-built models often incorporate spatial and meteorological information, they need large training datasets and, without sufficient regularization, remain vulnerable to overfitting.[21]

New developments in ensemble learning have made it possible to develop hybrid models that combine LightGBM's fast, gradient-boosted tree mechanics with the deep non-linear feature extraction power of NNs.[19] LightGBM is typically the first-level learner, identifying basic patterns and interaction effects in data, while the NN layer further refines predictions by learning complex, time-variable patterns.[22] This architecture enhances the speed and scalability of LightGBM on structured data and leverages the NN's ability to trend and model non-stationary time series.

Stacked models applied across finance, energy, and transportation domains have demonstrated greater generalizability, robustness to out-of-sample data, and an improved balance between performance and explainability—an important but often neglected aspect in model evaluation.[19] Explainability becomes increasingly paramount given the growing complexity of modern ML systems used in environmental monitoring and other critical domains. For example, SHAP and LIME provide valuable insights by quantifying the contribution of individual features to predictions. SHAP allocates consistent, mathematically justified scores indicating how much each input contributes to a given output, while LIME builds ephemeral, simpler predictive models to justify a given decision. Using these techniques in conjunction with leading models, such as LightGBM and NNs, enhances trust in model decisions, supporting more transparent and responsible environmental policy-making, from

researchers to policymakers, and ultimately fostering better-informed, sustainable decisions.[23]

Most literature these days examines novel individual or ensemble models; however, few adopt a bottom-up hybrid that integrates LightGBM with a custom NN architecture. In addition, lattice models, which offer crucial in-practice transparency, have received limited attention. This research aims to address this gap by constructing a stacked hybrid architecture that combines LightGBM's computational speed with the temporal awareness of a custom NN. The model incorporates SHAP and LIME from the explainable AI (XAI) toolbox to rationalize predictions, thereby improving model fidelity and trust.

3. Methodology

3.1. Data description

The present study utilized the World Air Quality Index (WAQI) dataset to forecast the concentration of PM10 using ML and deep learning techniques. This dataset is authentic and updated thrice daily, covering approximately 380 cities worldwide. It comprises 1,798,600 records and features nine attributes, that is, country, city, date, species, count, minimum, maximum, median, and variance. The feature of distinct pollutants further extends to specific pollutants. Also, the species attribute contains the Air Quality Index (AQI), dew, humidity, precipitation, pressure, PM1, PM2.5, and PM10 (Figure 1).

In this study, values of PM10 were considered for prediction, whereas the median attribute was selected as the target variable. To further demonstrate the characteristics of the WAQI dataset, Figures 2 and 3 present the geographical coverage of PM10 concentration and its average across countries and cities.

3.2. Data preparation and preprocessing

To ensure quality and consistency, the study considered only the relevant features of dates, location, and pollutant values. The date strings were converted to date–time objects, and the dataset was further restricted to PM10. Lag features captured trends and temporal dependencies, providing the PM10 concentration values from previous days. Seven lag features were generated, representing PM10 levels from the past seven days. In addition, two rolling averages were computed, that is, a 7-day rolling average and a 30-day rolling average. The rolling averages helped smooth out short-term fluctuations and highlighted long-term trends. This enabled the model to identify patterns in PM10 concentrations over weekly and monthly periods.

Categorical features, namely, country and city, were encoded using the target encoding technique by replacing each categorical value with the mean of the target variable, that is, the value of PM10, in our case, for that category. This encoding method, being memory efficient, enabled the model to capture the relationship between categorical features and target variables effectively. The dataset was split into training and testing subsets in an 80:20 ratio, with 80% of the data utilized for training the models and 20% for testing the models under different evaluation metrics.

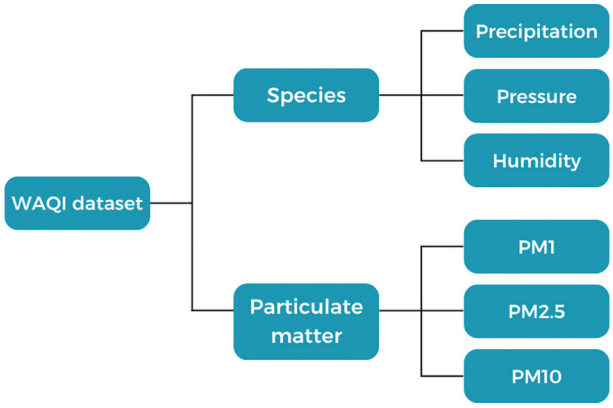

Figure 1. Features of species in the World Air Quality Index dataset

The selection of features was conducted in collaboration with domain experts and validated with algorithms. Redundant features were removed through initial correlation analysis and subsequent variance inflation factor analysis. Temporal and meteorological predictors were then validated with LightGBM's feature importance ranking. All numeric features were normalized through Min–Max scaling to prepare the data for the NN and enhance convergence and stability during training. This preprocessing made the model's parameters accurate and relevant to the environment, ensuring the approach was transparent and replicable.

3.3. Model architecture

As illustrated in Figure 4, our hybrid stacked modeling strategy combines three complementary elements to enhance predictive accuracy and protect the system from overfitting.

LightGBM is used as the primary regressor due to its efficiency on structured data and its ability to expose complex feature interactions when boosted using gradient methods. After the model was constructed, a broad grid search was performed to calculate appropriate configuration benchmarks by testing different values for the learning rate, tree depth, and levels of regularization. This base learner can identify the most prominent patterns in structured data; however, more intricate patterns, particularly those involving non-linear data in real-

Figure 2. Geographical coverage of PM10 across (A) countries and (B) cities according to the World Air Quality Index dataset accessed on [December, 2024]. Map reprinted from Microsoft PowerBI[Microsoft Corporation.]

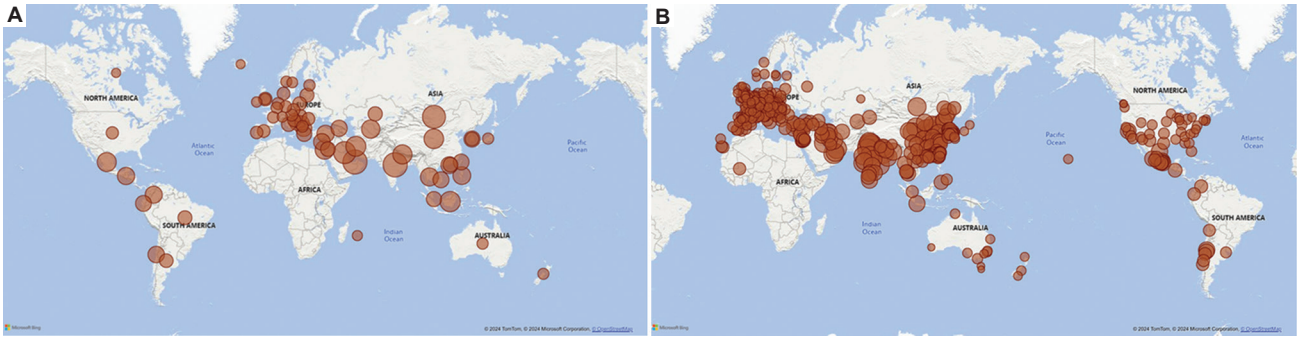

Figure 3. Average of PM10 by (A) country and (B) city according to the World Air Quality Index dataset accessed on [December, 2024]. Map reprinted from Microsoft PowerBI[Microsoft Corporation.]

world scenarios, are often left uncovered. This problem was addressed with the introduction of a framework for residual learning.

In this framework, a DNN is used for learning from the prediction mistakes resulting from the LightGBM regressor's residuals. This network, which acts as the bottom half of the model, is complemented by the base model and receives normalized feature vectors and features compatible with the base model input layer. Its skeleton consists of three densely connected hidden layers. Each hidden layer is succeeded by a fixed rate of batch normalization to moderate internal covariate shift and dropout regularization, set at 25% to avoid overfitting. At the same time, the layers are also interspersed with rectified linear unit (ReLU) activations for non-linear control. These controls assist the network in capturing and forming intricate patterns of residuals that conventional tree learners otherwise ignore.

A single linear neuron in the output layer is used for final residual corrections. The whole pipeline is then optimized in a single pass using the Adam optimizer, number of epochs learning rate policies, and mean squared error (MSE) loss. This architecture relies on the NN's universality to shift and correct all systemic biases beyond the initial LightGBM forecasts. The final step in the pipeline merges predictions from both modeling branches using a stacking ensemble. For example, in Level-1, the raw LightGBM forecasts and the residual-corrected values, defined as LightGBM predictions with corrections by the NN, are combined to form a richer feature set that captures complementary signals. These enhanced predictions are then forwarded to a Ridge regression meta-model, which reduces the multicollinearity among the base predictors and variances, as its L2 regularization is believed to achieve this goal. Using matrix algebra, the meta-learner uses the ridge regression method to compute optimal weights for each base predictor in a standard closed form. The weights are adjusted automatically with respect to validation performance. Table 1 details the parameters of the hybrid stacked model proposed in this study.

The NN was deliberately shaped for PM10 forecasting rather than adapted from a generic regression template. Its residual design corrects for the non-linear errors left by LightGBM, and each dense layer is followed by batch normalization and dropout to stabilize learning and avoid overfitting on noisy atmospheric data. ReLU activations help the model generalize across seasonal shifts, while the Adam optimizer ensures adaptive convergence. The

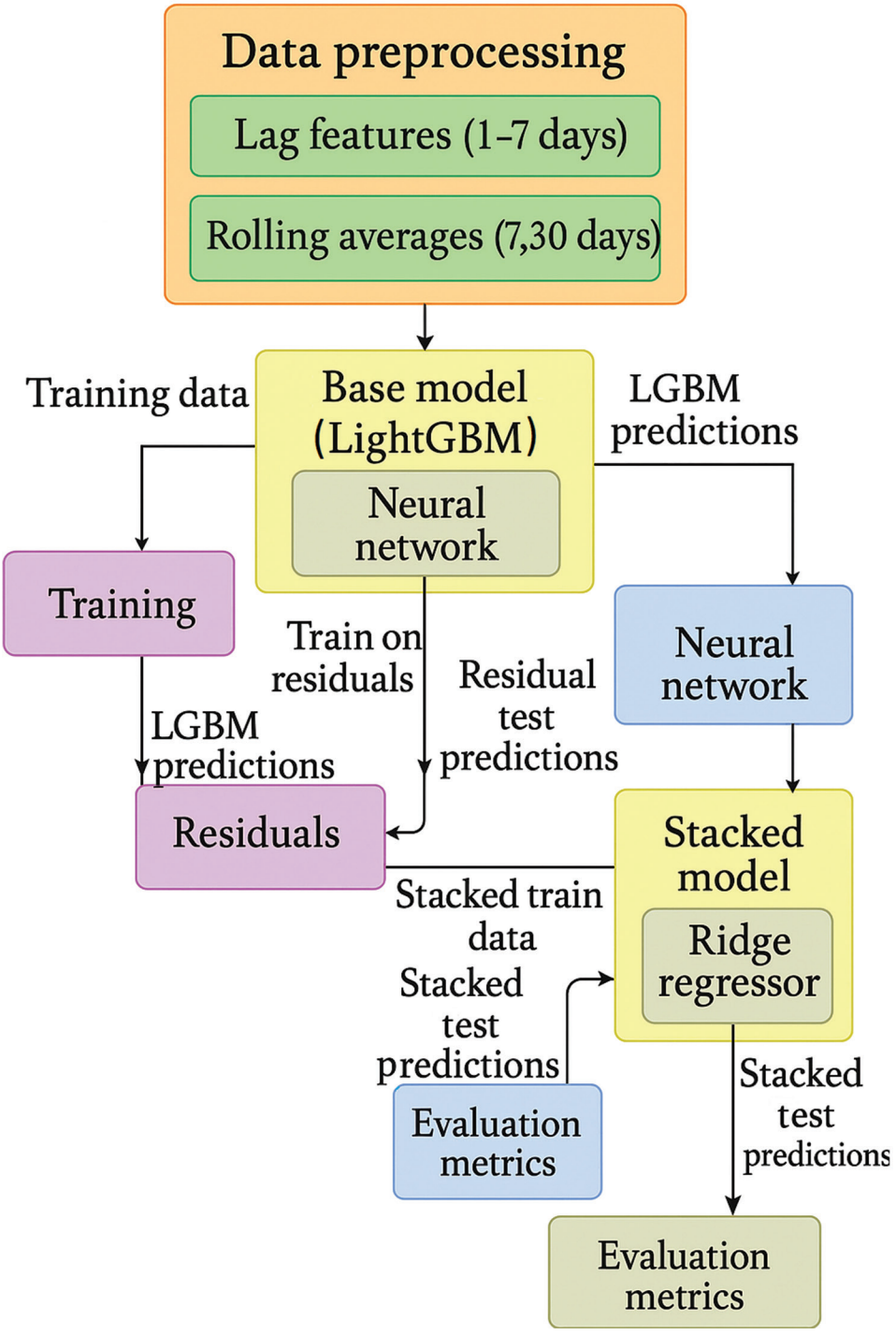

Figure 4. Proposed hybrid stacked model architecture

Ridge meta-learner integrates the outputs by assigning optimal weights through L2 regularization, reducing the effect of multicollinearity between the LightGBM and NN predictions. This configuration generates smoother and more generalizable forecasts without the instability often seen in kernel or tree-based meta-models. This stacked architecture, therefore, leverages transparent feature attribution of LightGBM, the NN's flexibility in capturing non-linear errors, and Ridges' robustness when working with multiple correlated inputs. The combined model

demonstrates superior generalization compared to any single component and consistently outperforms baseline benchmarks across a wide range of public datasets. The pseudocode of our proposed hybrid stacked model is presented in Pseudocode 1.

3.4. Model interpretability

To promote transparency and confidence in our prediction pipeline, the study conducted an in-depth interpretability assessment grounded in leading XAI techniques of SHAP and LIME. Here, SHAP applies concepts from cooperative game theory to produce a consistent global ranking of feature importance. By computing Shapley values for each observation, the method reveals how much each feature contributes to a prediction above or below the mean, thereby clarifying model behavior across the entire dataset. Meanwhile, LIME examines a single prediction at a time. It perturbs the input, fits a simple linear model around the point of interest, and uses the linear coefficients to indicate which features are most significant for that case. This local lens perfectly complements the broader overview of the SHAP analysis.

3.5. Evaluation methodology

A thorough evaluation of each aspect of the forecasting model was performed using a set of arbitrary statistical metrics concerned with the model's predictive capability and the decision model's required dimensions concerning the environment. The R^2 score, which is referred to as the coefficient of determination, for example, measures the extent to which the model identifies the total variation in the set of air-quality readings and quantifies the extent to which it offers explainable pieces of information. R^2 is a primitive of the model in the sense that all models are

Table 1. Details of the hybrid stacked model parameters

Component	Parameter	Value/Configuration
LightGBM (base model)	Learning rate	0.1
	Maximum depth	7
	No. of estimators	300
	No. of leaves	31
	Sub-sample	0.8
Residual neural network	Layer 1	Dense (256, Activation="relu")
	Layer 2	BatchNormalization()
	Layer 3	Dropout (0.2)
	Layer 4	Dense (128, Activation="relu")
	Layer 5	Dropout (0.1)
	Layer 6	Dense (64, Activation="relu")
	Output layer	Dense (1)
	Optimizer	Adam
	Loss function	Mean squared error (MSE)
	Metrics	Root MSE (RMSE), mean absolute error (MAE), R^2 score
	Batch size	256
	Epochs	100 (with early stopping)
	Early stopping	Patience=10, Monitor = "val_loss," restore_best_weights
	Validation split	0.2
Ridge (meta-model)	Alpha	1.0

Pseudocode 1. Proposed hybrid stacked neural network

Function: Residual stacking model
Input: Preprocessed dataset (features and target)
Output: Trained models and evaluation metrics

Start
1. Data preparation:
1.1. Separate features (X) and target (y)
1.2. Split data into train/test sets (70% train, 30% test)
1.3. Scale features using MinMaxScaler

2. Train base model (LightGBM):
2.1. Initialize LGBMRegressor with parameters:
learning_rate: 0.1
max_depth: 7
n_estimators: 300
num_leaves: 31
subsample: 0.8
2.2. Fit model on training data
2.3. Generate predictions for train/test sets

3. Residual learning:
3.1. Calculate residuals (true - predicted) on the training set
3.2. Build a residual neural network architecture:
Input Layer: (features_dimension,)
Dense (256, ReLU) → BatchNorm → Dropout (0.2)
Dense (128, ReLU) → Dropout (0.1)
Dense (64, ReLU)
Output: Dense (1)
3.3. Compile model with:
Optimizer: Adam
Loss: MSE
Metrics: RMSE, MAE, R²
3.4. Set early stopping (patience=10, monitor val_loss)
3.5. Train model on scaled features and residuals
3.6. Generate residual predictions for the test set
3.7. Create final residual-corrected predictions (LightGBM + NN residuals)

4. Model stacking:
4.1. Create stacking dataset:
Train: [LightGBM_pred, Residual_corrected_pred]
Test: [LightGBM_pred, Residual_corrected_pred]
4.2. Initialize Ridge regression meta-model (alpha=1.0)
4.3. Train meta-model on stacking dataset
4.4. Generate final stacked predictions

5. Evaluation:
5.1. For each model (LightGBM, Residual-corrected, Stacked):
Calculate R², MSE, RMSE, and MAE
Print evaluation metrics
End

ranked from 0 to 1, with 0 meaning that the model does not explain anything above the model's sample average, and 1 meaning that the model absolutely predicts the data; anything negative means the model is performing worse than just using the mean. The R^2 score, within the context of atmospheric pollution, suggests the effectiveness of a framework in tracking how the system copes with the complex shifts of air, emissions, and the weather. In addition, root MSE (RMSE) is the next benchmark concerned with the forecasting precision. It takes the mean of the difference squared, then takes the root. Moreover, it is a primitive of the original set. Spacing, air quality, and pollution models all rely on RMSE for decision-making, particularly when assessing the predictive range of a model. Gaps within these ranges may indicate failures to capture peak pollutant concentrations, especially those exceeding critical thresholds. Such missed or underestimated peaks can lead to unforeseen public health risks.

The values of RMSE directly address regulators and community members, as they are measured in micrograms per cubic meter, enabling straightforward comparison with legal thresholds. On the other hand, the mean absolute error (MAE) measures forecasting precision by averaging the absolute values of all prediction errors. It describes model accuracy without considering the direction of errors. Therefore, MAE provides a robust daily model performance snapshot. It is useful for assessing average overestimation, which is often critical for decision-makers. In practice, MAE can support dynamically managed resource allocation models, ensuring that responses to pollution escalations remain predictable, budget-conscious, and within an acceptable error margin.

Table 2. Experimental models and results

S. No.	Algorithm	MSE	RMSE	MAE	R^2
1	LightGBM+neural network (hybrid stacked model)	4.90	2.21	0.91	0.9916
2	Multi-layer perceptron (MLP\ FCNN)	9.77	2.85	1.82	0.9862
3	XGB regressor	18.21	4.27	1.30	0.9689
4	Random forest regressor (RFR)	20.91	4.57	1.08	0.9643
5	Extra trees regressor (ETR)	35.75	5.98	1.82	0.9390
6	Decision tree regressor (DTR)	63.73	7.98	2.22	0.8913
7	ARIMA	886.11	29.77	15.13	−0.1906
8	SARIMA	887.05	29.78	15.14	−0.1918

Abbreviations: ARIMA: Autoregressive Integrated Moving Average; FCNN: Fully connected neural network; MAE: mean absolute error; MSE: Mean squared error; RMSE: Root mean squared error; SARIMA: Seasonal Autoregressive Integrated Moving Average; XGB: Extreme gradient boosting.

The MSE summarizes all deviations by employing a squared-average method, focusing more on larger errors. This property makes MSE a valuable measure of dispersion in model predictions and a standard loss function in deep learning, where optimizers minimize it to avoid expensive, large-magnitude prediction errors. During critical pollution events, such as concentration spikes, this heightened sensitivity to large deviations acts as an additional security for public health decision-making. Together, these four metrics form a self-reinforcing evaluation system. R^2 measures the model's explanatory power, RMSE acts as a guardrail against overly optimistic under-predictions, MAE conveys expected day-to-day accuracy, and MSE informs system optimization through iterative design. Using all four ensures consistent system reliability across routine assessments, intermediate validations, and high-alert preparedness evaluations.

4. Results

4.1. Comparative evaluation of regression models

To evaluate the predictive quality of the eight regression procedures, four summary statistics, that is, MSE, RMSE, MAE, and R^2, were measured. Together, all these metrics represented the model's error size, error symmetry, and total explainability. Table 2 presents the results obtained using state-of-the-art ensemble learners and conventional techniques, covering the full spectrum of modeling techniques.

The strongest overall performance was demonstrated by the LightGBM–NN hybrid stacked model, which consistently outperforms all other models across every evaluation metric. It achieved the lowest error values, with an MSE of 4.90, RMSE of 2.21, and MAE of 0.91, along with an R^2 of 0.9916, indicating that the model explains more than 99% of the total variance in the data. These results validate the effectiveness of the hybrid stacked architecture, which combines tree-guided learning with deep feature extraction to reduce bias and variance. This combination is often preferred in ML due to the better performance on complex, noisy datasets, offering broader generalization and increased stability.

The second-best performer is the multi-layer perceptron (MLP), which achieved an R^2 of 0.9862, an MSE of 9.77, and an MAE of 1.82, indicating a widening performance gap. While the MLP maintains a strong ability to capture non-linearity, its overall fit remains inferior to that of the stacked architecture. Among ensemble tree techniques, the RF regressor and XGB regressor trail behind in accuracy, as single-tree models are eclipsed by neural and hybrid models. The RF regressor demonstrated a modest but stable performance with an MAE of 1.08, whereas the XGB

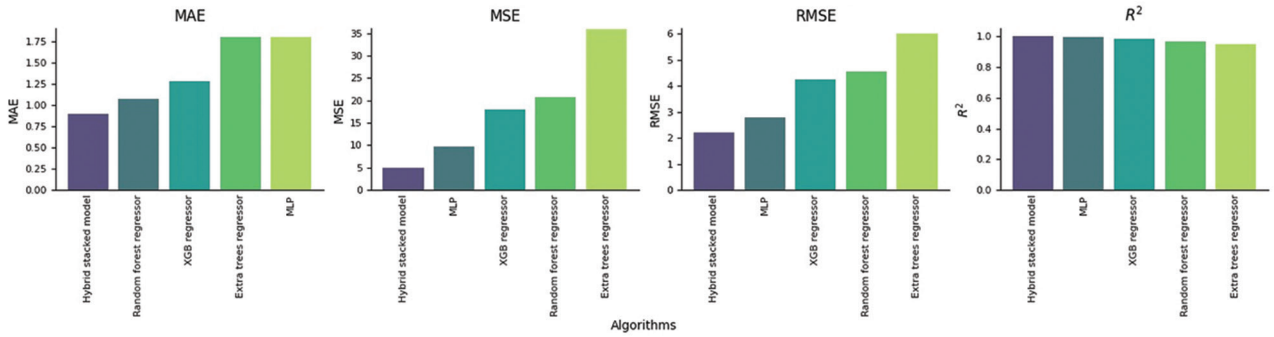

Figure 5. Comparison of evaluation metrics for the top five best-performing models
Abbreviations: MAE: mean absolute error; MLP: Multi-layer perceptron; MSE: Mean squared error; RMSE: Root mean squared error; XGB: Extreme gradient boosting.

regressor lagged with a higher MAE of 1.30.

The extra trees regressor and decision tree regressor emphasize the importance of ensembling, yet their relatively poor performance underscores the limitations of shallow learners on complex datasets. The decision tree regressor, with an MAE of 2.22 and an R^2 of 0.8913, ranked lowest among the tree-based models, reflecting the challenges shallow models face in capturing complex data patterns.

Regarding traditional time series methods, both Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models faced criticism for their underwhelming performance compared to present ML models. Their RMSE values, 29.77 and 29.78, respectively, and MSE values, 886.11 and 887.05, respectively, were generally unacceptable in practical applications. Moreover, their R^2 of −0.1906 and −0.1918, respectively, were the poorest scores in the study, indicating that they performed worse than a simple predictor that merely predicted the historical mean. These shortcomings likely stem from the rigid structure of ARIMA and SARIMA models, which cannot easily incorporate unbounded external factors typical of contemporary environmental sensor data streams.

Overall, the results demonstrate a common theme: while traditional time series benchmarks may provide a baseline, only the hybrid stacked ensemble approaches near-perfect explainability and achieve impressively low error rates. This model sets a new standard for predictive applications that demand trustworthiness and high accuracy. Figure 5 demonstrates the performance of the five best-performing algorithms.

The complete hybrid model was trained on a standard desktop system (Intel i7-11700, 16 GB RAM, no GPU) to validate computational feasibility. Despite handling 1.8 million records, training was completed in approximately

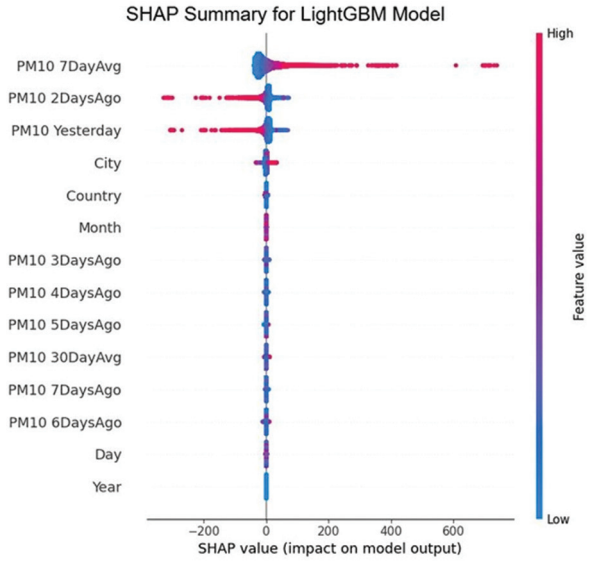

Figure 6. SHAP analysis of LightGBM
Abbreviation: SHAP: Shapley Additive Explanations.

2 h and 20 min, using <5 GB of memory. Generating a 1-month forecast took less than 30 s. These figures confirm that the model can run comfortably on non-specialized hardware, making it viable for city agencies and research labs without high-performance computing resources.

4.2. Insights from SHAP analysis

In this study, SHAP was used to clarify the inner workings of the LightGBM model, the NN, and the pooled stacked ensemble, making each predictor easier for researchers to interpret. In the LightGBM analysis (Figure 6), the SHAP analysis identified three recent lagged PM10 readings—averages over the past 7 days, 2 days, and the previous day—as the key factors affecting air quality forecasts. Their dominance reflects long-standing environmental

theory, which posits that PM levels tend to persist and gradually decay rather than change abruptly. By contrast, fixed features, such as city and country showed almost zero effect, confirming that present pollution loads, rather than location or time-specific factors, consistently drive the model.

The NN (Figure 7) showed a nearly identical pattern, that is, recent PM10 levels ranking highest in importance. However, its layered structure assigned small but noticeable weight to location, suggesting that the NN captures region-specific trends or demographic patterns when such data are present. This additional responsiveness generates a slightly richer, albeit less parsimonious, prediction whenever site-level behavior varies, reflecting the NN's use of curved activation functions, producing smoother and more gradual influence distributions compared to the sharper transitions seen in tree-based models.

The stacked hybrid model exhibited a distinct pattern of interpretability when examined using SHAP (Figure 8). In this setting, the final prediction was primarily derived from the outputs of the two core models, that is, the LightGBM and the NN. Their combined output overshadows the influence of raw environmental features, which contribute almost nothing on their own. This pattern underscores the meta-learner's role as a coordinator rather than an interpreter, leaving most of the explanatory work to the base models while fine-tuning the final score by averaging their predictions.

More importantly, the SHAP analysis revealed reassuring consistency at global and local levels across all configurations. Recent PM10 readings emerge as steady and robust predictors, and the capability of the hybrid stacked architecture stems from the skillful combination of these familiar signals. The dominance of the 7-day lag feature is not arbitrary—it mirrors the atmospheric persistence of PM. PM10 concentrations evolve gradually due to meteorological stability, boundary-layer retention, and continuous urban emissions. The strong autocorrelation between present and past readings reflects this physical behavior, rather than being a mere statistical correlation. While spatial features contributed marginally, SHAP and LIME demonstrated that they interact indirectly with temporal lags, showing that geography amplifies or dampens pollutant retention under specific weather conditions. This reinforces the environmental plausibility of the model's reasoning.

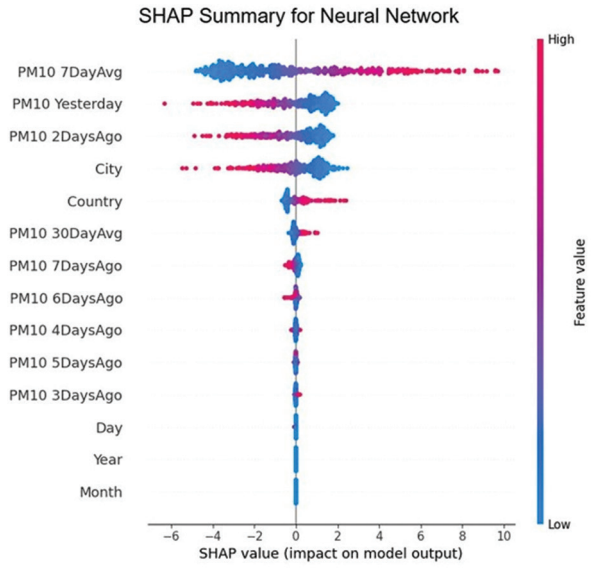

Figure 7. SHAP analysis of the NN
Abbreviations: NN: Neural network; SHAP: Shapley Additive Explanations.

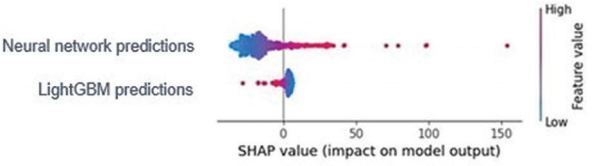

Figure 8. SHAP analysis of the proposed hybrid stacked model
Abbreviation: SHAP: Shapley Additive Explanations.

4.3. Local explanation for model decisions

Typically, LIME zooms in and highlights the factors contributing to each prediction. For the NN in this study, the LIME analysis (Figure 9) identified "PM10 2 DaysAgo > 0.03" as the most significant feature, exerting a strong

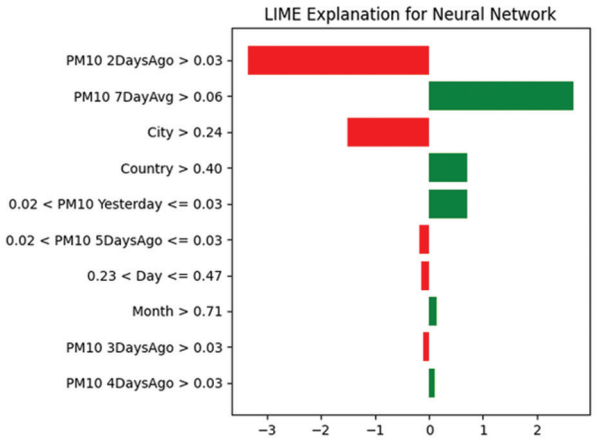

Figure 9. LIME analysis of NN
Abbreviations: LIME: Local Interpretable Model-Agnostic Explanations; NN: Neural network.

negative influence that lowers the forecast when this threshold is exceeded. In contrast, "PM10 7-Day Avg > 0.06" emerged as the most significant positive driver, enhancing the prediction accuracy. The "City > 0.24" feature exerted a negative but comparatively minor effect; in contrast, "Country > 0.40" contributed a slight positive boost. "0.02 < PM10 Yesterday ≤ 0.03," "0.02 < PM10 5 days ago ≤ 0.03," and calendar indicators, such as day or month also played a role, though their contributions remained modest. The LIME breakdown aligns with the global SHAP view, highlighting the lagged PM10 series as the primary contributor while revealing spatial features as context-sensitive nuances effectively captured by the NN.

The LIME summary for the LightGBM model presents a different balance of influences while conveying a similar overall message (Figure 10). Here, "PM10 7DayAvg > 32.00" stood out with a strong positive influence on predictions, whereas "PM10 2DaysAgo > 31.00" acted as the second-largest driver but effect in the opposite direction. Impacts from other lagged features and the city feature provide only slight fine-tuning. The key insight is that LightGBM forecasts depend almost entirely on recent air quality measurements, with the model sharply focusing on these dominant signals in a recursive manner.

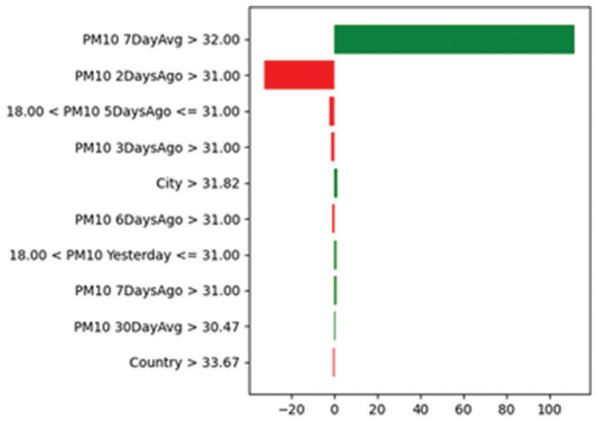

Figure 10. LIME analysis of LightGBM
Abbreviation: LIME: Local Interpretable Model-Agnostic Explanations.

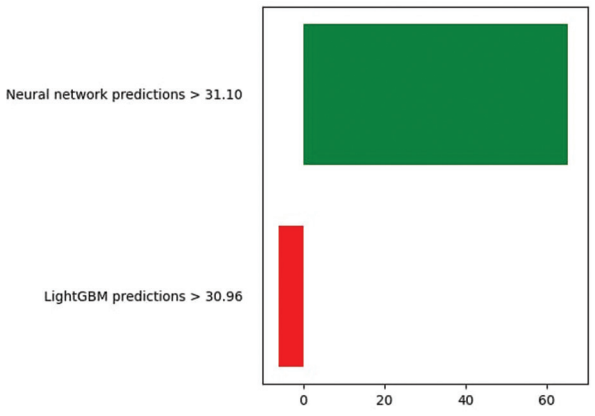

Figure 11. LIME analysis of the hybrid stacked model
Abbreviation: LIME: Local Interpretable Model-Agnostic Explanations.

By contrast, the LIME analysis of the hybrid stacked model (Figure 11) revealed a remarkably clean and intuitive interpretation. Here, the final prediction for each instance largely stemmed from the separate outputs of the underlying NN and LightGBM, with the meta-learner acting primarily as a blending mechanism. When the NN's score exceeded 31.10, the meta-learner assigned it a heavy positive weight; however, if the reading of LightGBM surpassed 30.96, it slightly reduced that weight, maintaining a small gap between them. The logic is straightforward: The composer leverages the strengths of its two bases, trusting their ensemble view more than any single input feature.

4.4. Synthesis

Combining SHAP and LIME analyses yielded a remarkably unified account of model performance. The stacked hybrid system not only outpaced every competing architecture across key evaluation metrics but does so with scientific rigor and transparency. This superior performance primarily stems from the ensemble's adept use of recent PM10 readings, a conclusion repeatedly corroborated by both the global perspective of SHAP and the localized viewpoint offered by LIME. The meta-predictive framework selectively forwards the most confident outputs from its base learners, thereby anchoring the final forecast in robust evidence. As a result, the hybrid design proves practically resilient, highly interpretable, and broadly adaptable—qualities essential for public agencies and environmental modelers alike. Altogether, these findings offer a compelling case for deploying stacked ensembles, such as LightGBM with NN, when minimizing error and maximizing clarity are critical. Backed by extensive benchmarking and layered interpretability, such a hybrid stacked framework established an ambitious benchmark for reliable, data-driven decision support in environmental,

Table 3. Comparison of the hybrid stacked model with state-of-the-art models

S. No.	Model	Year	R^2
1.	ANN[1]	2024	0.7850
2.	CNN–LSTM[16]	2023	0.8800
3.	LightGBM+NN (hybrid stacked model)	2025	0.9916

Abbreviations: ANN: Artificial neural network; CNN: Convolutional neural network; LSTM: Long short-term memory; NN: Neural network.

health, and policy arenas. To further validate the outcomes of our proposed hybrid stacked model, it was compared with state-of-the-art models (Table 3).

5. Discussion

This study contributes to the body of knowledge regarding forecasting PM10 levels and the associated hybrid stacked model by demonstrating that a hybrid stacked model significantly outperforms traditional models in explained variance and error reduction. In this context, the model's R^2 value of 0.9916 indicates a significant improvement over traditional statistical models, such as ARIMA and SARIMA, which exhibited negative explained variance values of approximately −0.19. This highlights the inability of these models to capture the complex, high-dimensional, and non-linear nature of air pollution data. The hybrid model also surpasses the standalone XGB ML model (R^2 = 0.9689) and the MLP (R^2 = 0.9862). These findings demonstrate the advantage of integrating tree-based models with NNs to capture complex relationships while maintaining computational efficiency.[24]

Additional tests were conducted by segmenting cities according to industrial activity and climate type to evaluate model robustness across different environments. In all cases, the hybrid model maintained R^2 values above 0.98, albeit with slightly greater variance in coastal and high-emission regions. Such differences correspond with documented patterns of pollution persistence rather than model errors. Visual assessments of predicted and measured PM10 concentrations confirmed the geometric center of the regions, and the residuals were evenly distributed, supporting the model's adaptability to various urban settings. This capability is essential for air-quality management systems intended for practical policy implementation.

These findings validate and extend other studies indicating the need for advanced temporal dependence and non-linear methodologies in air quality prediction. More recent studies confirm the effectiveness of hybrid approaches that integrate variational mode decomposition with attention-augmented LSTM models in improving predictive accuracy by capturing latent dynamic features—consistent with our models of reduced forecasting errors.[24] Feature importance assessments using SHAP and LIME indicate that the lagged PM10 parameters, especially the seven-day rolling mean and the one-day lagged concentration, are the most significant features. These findings support the air quality studies, emphasizing the time lag and autocorrelation of pollutants.[25,26] The developed model offers apparent practical advantages over conventional deterministic and statistical models, which often fail to adequately capture the non-linear, spatiotemporal complexities of pollution data. The Ridge meta-learner balances base learners' contributions to reduce potential overfitting and measurement noise dominance. This hybrid approach retains fast prediction capabilities while relying heavily on accurate, regularly updated lag features.

The model can be strengthened and generalized by incorporating real-time meteorological and auxiliary information. In summary, the developed hybrid stacked model delivers enhanced statistical measures and, to some extent, an interpretable framework aligning feature importance with domain knowledge. This is useful as it offers the relevant policymakers and stakeholders in environmental health actionable and pragmatic decision support. Incorporating additional real-time environmental data will be critical to streamline the model for effective air quality management in other regions. Complex frameworks built around physical air quality models should be tested in such regions.

6. Conclusion

The hybrid stacked model proposed in this study demonstrated strong accuracy in forecasting PM10 levels. Further development of such models could assist public health officials and environmental managers in optimizing health and environmental outcomes. Considering ease of use and policy relevance, this study recommends that decision-makers integrate the proposed model with other AI-based forecasting engines to enhance air quality monitoring and alerting systems, thereby improving model performance. The results further reveal that model accuracy tends to diminish when historical time-series pollutant data are used to estimate a model's parameters. The real-time measurement of exogenous data, such as vehicle counts, industrial emissions, and even weather conditions, would improve the model's operational performance and predictive efficacy. These forecasting models could support policymakers in implementing dynamic air quality management strategies, including real-time public advisory systems for urban areas that experience recurring pollution episodes. This aligns with recent bulletins from the World Meteorological Organization, which advocate for greater unification of air quality and climate initiatives, along with improved integrated monitoring and international collaboration, to reduce aerosol pollutants and public health risks. Such studies should aim to validate the model in other regions and socio-economic conditions that lack adequate air quality monitoring infrastructure, thereby promoting active and equitable air quality governance.

Emphasizing the link between air quality forecasting, public policy decision support systems, and public health protection will support the more refined integration of formulated strategies to manage exposure to PM pollution for vulnerable populations. Ultimately, this study contributes a transparent and computationally efficient framework that advances the interface between science and policy, alleviating the transition toward healthier and more resilient urban areas.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Syed Azeem Inam
Formal analysis: Syed Azeem Inam
Investigation: All authors
Methodology: Syed Azeem Inam
Writing – original draft: All authors
Writing – review & editing: Syed Azeem Inam

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The datasets used and/or analyzed during the present study are available from the corresponding author upon reasonable request.

References

1. Shepelev V, Glushkov A, Slobodin I, Cherkassov Y. Measuring and modelling the concentration of vehicle-related PM2.5 and PM10 emissions based on neural networks. *Mathematics*. 2023;11:1144.

   doi: 10.3390/math11051144.

2. Manono BO, Sadiq FK, Sadiq AA, Matsika TA, Tanko F. Impacts of air quality on global crop yields and food security: An integrative review and future outlook. *Air*. 2025;3(3):24.

   doi: 10.3390/air3030024

3. Ramaiah M, Vanmathi C, Khan MZ, Noorwali A, Jain R, Agarwal P. COVID19: Forecasting air quality index and particulate matter (PM2.5). *Comput Mater Continua*. 2021;67:3363-3380.

   doi: 10.32604/cmc.2021.014991

4. Sakhrieh A, Hamdan MA, Ata MB. Air quality assessment and forecasting using neural network model. *J Ecol Eng*. 2021;22:1-11.

   doi: 10.12911/22998993/137444

5. Shams SR, Jahani A, Kalantary S, Moeinaddini M, Khorasani N. Artificial intelligence accuracy assessment in NO_2 concentration forecasting of metropolises air. *Sci Rep*. 2021;22:1-11.

   doi: 10.1038/s41598-021-81455-6

6. Li T, Hua M, Wu X. A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5). *IEEE Access*. 2020;8:26933-26940.

   doi: 10.1109/access.2020.2971348

7. Waseem KH, Mushtaq H, Abid F, *et al.* Forecasting of air quality using an optimized recurrent neural network. *Processes*. 2022;10:2117.

   doi: 10.3390/pr10102117

8. Ben Jabeur S, Khalfaoui R, Ben Arfi W. The effect of green energy, global environmental indexes, and stock markets in predicting oil price crashes: Evidence from explainable machine learning. *J Environ Manage*. 2021;298:113511.

   doi: 10.1016/j.jenvman.2021.113511

9. Ding H, Noh G. A hybrid model for spatiotemporal air quality prediction based on interpretable neural networks and a graph neural network. *Atmosphere (Basel)*. 2023;14:1807.

   doi: 10.3390/atmos14121807

10. Wu Y, Hu J, Irfan M, Hu M. Vertical decentralization, environmental regulation, and enterprise pollution: An evolutionary game analysis. *J Environ Manage*. 2024;349:119449.

    doi: 10.1016/j.jenvman.2023.119449.

11. Hu J. Synergistic effect of pollution reduction and carbon emission mitigation in the digital economy. *J Environ Manage*. 2023;337:117755.

    doi: 10.1016/j.jenvman.2023.117755

12. Feng L, Lu J, Hu J, Irfan M, Wu K. Divergent carbon emission mitigation pathways toward sustainable development: Heterogeneous effects of the digital economy in urban centers versus boundary regions. *Sustain Cities Soc*. 2025;132:106808.

    doi: 10.1016/j.scs.2025.106808

13. Ur Rahim M, Hussain M, Inam SA, Hashim H. Ignition behavior of supercritical liquid fuel in combustion system. *J Mech Continua Math Sci*. 2021;16(8):22-34.

doi: 10.26782/jmcms.2021.08.00003

14. Inam SA, Khan AA, Mazhar T, *et al.* PR-FCNN: A Data-driven hybrid approach for predicting PM2.5 concentration. *Discov Artif Intell*. 2024;4(1):75.

    doi: 10.1007/s44163-024-00184-7

15. Haupt SE, Gagne DJ, Hsieh WW, *et al.* The history and practice of AI in the environmental sciences. *Bull Am Meteorol Soc*. 2022;103(5):E1351-E1370.

    doi: 10.1175/BAMS-D-20-0234.1

16. Wang C, Chang C. Forecasting air quality index considering socio-economic indicators and meteorological factors: A data granularity perspective. *J Forecast*. 2023;42:1261-1274.

    doi: 10.1002/for.2962

17. Kujawska J, Kulisz M, Oleszczuk P, Cel W. Machine learning methods to forecast the concentration of PM10 in Lublin, Poland. *Energies (Basel)*. 2022;15:6428.

    doi: 10.3390/en15176428

18. Lei TMT, Siu SWI, Monjardino J, Mendes L, Ferreira F. Using machine learning methods to forecast air quality: A case study in Macao. *Atmosphere (Basel)*. 2022;13(9):1412.

    doi: 10.3390/atmos13091412

19. Yang G, Lee H., Lee G. A hybrid deep learning model to forecast particulate matter concentration levels in Seoul, South Korea. *Atmosphere (Basel)*. 2020;11(4):348.

    doi: 10.3390/atmos11040348.

20. Tsalikidis N, Mystakidis A, Koukaras P, *et al.* Urban traffic congestion prediction: A multi-step approach utilizing sensor data and weather information. *Smart Cities*. 2024;7:233-253.

    doi: 10.3390/smartcities7010010

21. Zukaib U, Maray M, Mustafa S, Haq NU, Rehman Khan AU, Rehman F. Impact of COVID-19 lockdown on air quality analyzed through machine learning techniques. *PeerJ Comput Sci*. 2023;9:e1270.

    doi: 10.7717/peerj-cs.1270

22. Inam SA, Khan AA, Ahmed N, *et al.* A novel deep learning approach for investigating liquid fuel injection in combustion system. *Discov Artificial Intell.* 2025;5(1):32.

    doi: 10.1007/s44163-025-00248-2

23. Arboleda-Florez M, Castro Zuluaga CA. Interpreting direct sales' demand forecasts using SHAP values. *Production*. 2023;33:e20220035.

    doi: 10.1590/0103-6513.20220035

24. Wang X, Zhang S, Chen Y, *et al*. Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD-GAT-BiLSTM. *Sci Rep.* 2024;14(1):17841.

    doi: 10.1038/s41598-024-68874-x

25. Ji Y, Zhi X, Wu Y, *et al*. Regression analysis of air pollution and pediatric respiratory diseases based on interpretable machine learning. *Front Earth Sci (Lausanne)*. 2023;11:1105140.

    doi: 10.3389/feart.2023.1105140

26. Yenkikar A, Mishra VP, Bali M, Ara T. Explainable forecasting of air quality index using a hybrid random forest and ARIMA model. *MethodsX*. 2025;15:103517.

    doi: 10.1016/j.mex.2025.103517