Tropical Environment, Biology, and Technology

Research Article
Volume 4(2), 2026, 120−135
https://doi.org/10.53623/tebt.v4i2.1386

Machine Learning-Based Air Quality Index Prediction in Hanoi: A Comparative Evaluation of Conventional and Deep Learning Models

Si Duc Truong*, Nguyen Van Thanh

Institute of Science and Technology for Energy and Environment, VAST, Nghia Do Ward, Hanoi, Vietnam

Correspondence: si.truong@miuc.edu

SUBMITTED: 28 August 2026; REVISED: 31 October 2027; ACCEPTED: 2 November 2026

ABSTRACT: Accurate air quality prediction is essential for air quality assessment, early warning, and effective air pollution management in rapidly urbanizing cities. This study evaluated the performance of conventional machine-learning and deep-learning models for Air Quality Index (AQI) prediction in Hanoi, Vietnam. A dataset comprising 8,784 observations collected during 2024–2025 was used, with 12 input variables including six meteorological parameters (temperature, relative humidity, atmospheric pressure, wind speed, wind direction, and rainfall) and six air pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3). Six models—Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), and CNN–LSTM—were developed and evaluated using RMSE, MAE, and R^2. Pearson correlation analysis showed that AQI was most strongly associated with $PM_{2.5}$ ($r = 0.97$) and PM_{10} ($r = 0.93$), followed by CO ($r = 0.75$) and NO_2 ($r = 0.61$). Among the evaluated models, LR achieved the best predictive performance (RMSE = 5.274, MAE = 4.122, $R^2 = 0.963$), followed by XGBoost (RMSE = 5.665, MAE = 4.484, $R^2 = 0.957$) and RF (RMSE = 5.809, MAE = 4.582, $R^2 = 0.955$). SVR showed moderate performance (RMSE = 11.101, MAE = 7.505, $R^2 = 0.836$), whereas LSTM (RMSE = 20.630, MAE = 16.314, $R^2 = 0.439$) and CNN–LSTM (RMSE = 21.739, MAE = 17.248, $R^2 = 0.377$) performed substantially worse. These findings demonstrate that increasing model complexity does not necessarily improve predictive accuracy and highlight the importance of selecting forecasting algorithms according to the characteristics of the dataset and prediction framework. For the dataset examined, LR, XGBoost, and RF provided the most effective approaches for AQI prediction in Hanoi.

KEYWORDS: Air Quality Index; air pollution; machine learning; deep learning; Hanoi.

1. Introduction

Air pollution has become one of the major environmental and public health challenges in large urban areas. Air pollutants such as fine particulate matter ($PM_{2.5}$), PM_{10}, NO_2, SO_2, CO, and O_3 can cause a wide range of adverse health effects, particularly under prolonged exposure. The World Health Organization (WHO) has emphasized the need to control ambient

concentrations of air pollutants to minimize risks to public health [1]. In this context, the Air Quality Index (AQI) is widely used to characterize the level of air pollution and communicate air quality information to the public [2]. Accurate AQI prediction is therefore important for providing early warnings, supporting environmental management, and mitigating the impacts of severe air pollution episodes.

Hanoi is a densely populated city with high traffic intensity and is affected by multiple emission sources, including transportation, industrial activities, construction, and regional emissions. Previous studies conducted in Hanoi have identified multiple sources contributing to particulate matter and atmospheric pollution [3, 4]. In addition to emission sources, meteorological conditions play an important role in the accumulation and dispersion of air pollutants. Tran et al. (2020) reported pronounced seasonal, monthly, and diurnal variations in $PM_{2.5}$ concentrations in Hanoi [5]. Therefore, prediction AQI using simple statistical approaches may be limited in their ability to simultaneously capture the complex relationships between historical air pollution data and environmental factors.

The rapid development of Machine Learning (ML) and Deep Learning (DL) has introduced new approaches to air quality prediction [6]. Among conventional machine learning models, Linear Regression (LR) is characterized by its simple structure, computational efficiency, and high interpretability [7]. Support Vector Regression (SVR) can capture nonlinear relationships through kernel functions [8], whereas Random Forest (RF) employs an ensemble of decision trees to improve generalization and reduce overfitting [9]. Extreme Gradient Boosting (XGBoost) further enhances the boosting framework through efficient optimization and regularization, and has consequently been widely applied to prediction tasks involving structured data [10]. Recent studies have also reported that RF and XGBoost often perform effectively when applied to multivariate tabular air quality datasets [11].

However, these conventional ML models may have limitations in directly capturing long-term dependencies in time-series data. The development of DL has therefore provided alternative approaches for air quality time-series forecasting. To improve data reliability, various unsupervised learning techniques have also been explored to detect and correct anomalies prior to subsequent data processing. Long Short-Term Memory (LSTM) networks are specifically designed to retain relevant information across sequential observations and capture long-term temporal dependencies, thereby addressing some of the limitations associated with conventional recurrent neural networks [12]. LSTM has been applied in numerous air quality forecasting studies and has demonstrated considerable potential for capturing the temporal dynamics of environmental data [13]. The hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model represents a more advanced approach by combining the feature extraction capability of CNNs with the temporal learning capability of LSTM networks, thereby improving the identification of complex patterns in time-series data [14]. Nevertheless, greater model complexity does not necessarily guarantee higher predictive accuracy; actual performance depends on several factors, including dataset size, input features, temporal characteristics, and model configuration.

Air quality prediction is particularly important for Hanoi in the context of rapid urbanization, high traffic density, and increasing construction, industrial, and residential activities. Reliable AQI estimation can support air quality assessment and environmental management, while future AQI prediction can provide timely warnings that enable authorities and the public to respond to deteriorating air quality. However, relative to practical needs,

studies systematically selecting and evaluating AQI forecasting models that are appropriate for the data characteristics and environmental conditions of Hanoi remain limited. In particular, there is a lack of examining the performance of conventional ML algorithms, such as LR, SVR, RF, and XGBoost, alongside models capable of capturing temporal dependencies, such as LSTM and hybrid CNN–LSTM. This research gap highlights the need to evaluate these models using the same dataset while explicitly considering differences in their input structures and prediction horizons, thereby assessing their predictive performance and practical applicability under different prediction settings. The novelty of this study lies in its comparative assessment of six models spanning linear regression, kernel-based learning, ensemble methods, and deep-learning architectures using a common air quality dataset from Hanoi. By examining differences in predictive performance across model families and input structures, while acknowledging that conventional models perform contemporaneous AQI estimation and deep-learning models perform one-step-ahead forecasting, the study provides evidence on the suitability and limitations of increasingly complex approaches under Hanoi's urban air pollution conditions.

2. Methods

2.1. Study Area and Data Collection.

The study was conducted in Hanoi, Vietnam, using a dataset comprising 8,784 observations collected during the 2024–2025 period. The study area included the automatic air quality monitoring station at the U.S. Embassy in Hanoi, located at No. 7 Lang Ha Street, O Cho Dua Ward, Hanoi, Vietnam (21.02167° N, 105.81889° E). The dataset included 12 input variables, consisting of six meteorological parameters: air temperature (°C), relative humidity (%), atmospheric pressure (hPa), wind speed (m/s), wind direction (°), and rainfall (mm); and six air quality parameters: $PM_{2.5}$ (µg/m^3), PM_{10} (µg/m^3), NO_2 (µg/m^3), SO_2 (µg/m^3), CO (mg/m^3), and O_3 (µg/m^3).

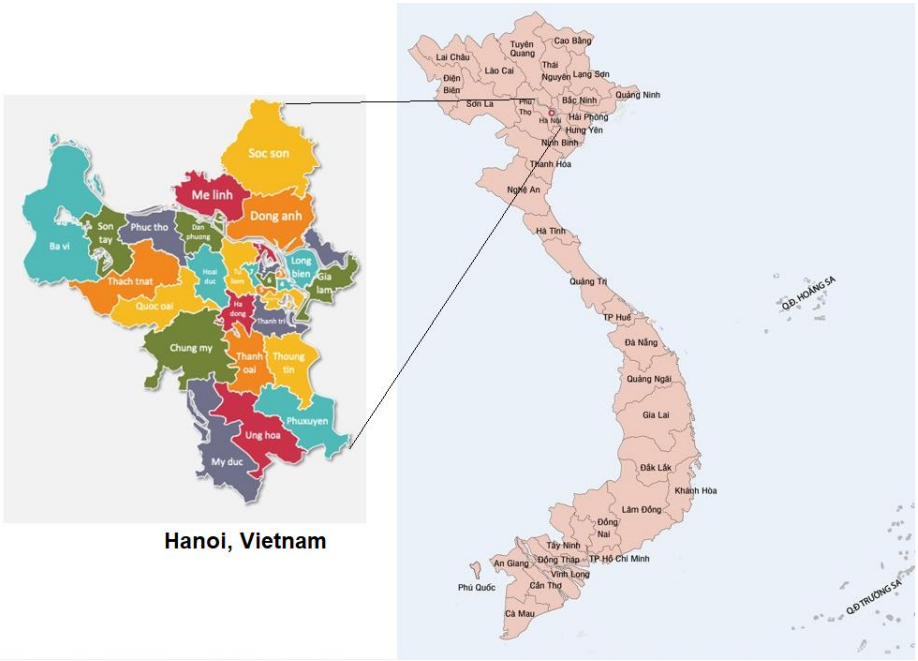

Figure 1. Location of The Study Area in Hanoi, Vietnam.

2.2. Data Preprocessing.

Prior to model development, the dataset was examined and preprocessed to ensure data consistency and reliability. Missing values, anomalous data points, and invalid observations were identified and appropriately handled before subsequent analysis. The input variables were standardized to minimize the effects of differences in measurement units and scales, particularly for the SVR, LSTM, and CNN–LSTM models. Given the time-series structure of the dataset, the chronological order of observations was preserved during data splitting to prevent information leakage from future observations. The dataset, comprising 8,784 observations, was divided into training and testing sets for model development and performance evaluation. Meteorological variables and the concentrations of $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 were used as input variables, while the AQI at the forecast time was used as the target variable. For the LSTM and CNN–LSTM models, the data were organized into sequences of consecutive observations to enable the models to capture temporal dependencies in AQI variations.

2.3. Model Development.

Linear Regression (LR) was employed as the baseline model. AQI was represented as a linear combination of the input variables, including meteorological parameters and air pollutant concentrations. The model was estimated using the ordinary least squares method, in which the regression coefficients were determined by minimizing the sum of squared differences between observed and predicted AQI values. LR offers the advantages of a simple structure, low computational cost, and straightforward interpretation of the effects of input variables.

Support Vector Regression (SVR) was implemented using a Radial Basis Function (RBF) kernel to capture nonlinear relationships between the input variables and AQI. Prior to model training, the input variables were standardized to minimize the effects of differences in scale. The main model parameters included the penalty parameter C, the ε-insensitive loss parameter, and the kernel coefficient γ, which jointly control the trade-off between prediction accuracy and model generalization.

Random Forest (RF) was developed by combining multiple regression trees. Each tree was trained using a bootstrap sample of the training dataset, while a randomly selected subset of input variables was considered at each split. The final AQI prediction was obtained by averaging the predictions generated by all individual trees. This ensemble approach enables RF to capture nonlinear relationships and interactions among input variables while reducing the risk of overfitting compared with a single decision tree.

XGBoost was implemented based on the gradient boosting framework, in which regression trees are sequentially added to correct the prediction errors generated by preceding trees. Model development considered key hyperparameters such as the number of boosting rounds, learning rate, maximum tree depth, subsampling ratio, and proportion of input features used for tree construction. The regularization mechanisms incorporated into XGBoost help control model complexity and reduce overfitting.

For the LSTM model, the dataset was transformed into sequences using a 24-time-step lag window, in which information from the preceding 24 observations was used to predict AQI at the subsequent time step. The network consisted of two stacked LSTM layers with 64 and 32 units, respectively, followed by a fully connected layer with 16 units and a single output

neuron. A dropout rate of 0.2 was applied to reduce overfitting. This architecture enables the model to capture temporal dependencies in air quality variations.

The CNN–LSTM model combined a one-dimensional convolutional neural network (1D-CNN) with LSTM layers. Two Conv1D layers with 64 and 32 filters, respectively, and a kernel size of 3 were used to extract local features from the input sequences. A max-pooling layer was subsequently applied to reduce the dimensionality of the extracted features before they were passed to two LSTM layers containing 64 and 32 units, respectively. Finally, a fully connected layer with 16 units and a single output neuron was used to generate the AQI prediction. Similar to the LSTM model, the CNN–LSTM architecture employed a 24-time-step lag window and a dropout rate of 0.2.

All models were developed using the same set of input variables and evaluated on an independent testing dataset. The simultaneous evaluation of models with different architectures enabled a systematic comparison of their ability to capture relationships among meteorological parameters, air pollutant concentrations, and AQI, thereby facilitating the identification of the most suitable model for the study dataset.

2.4. Model Training and Validation.

Following preprocessing, the dataset was divided into training and testing sets at an 80:20 ratio. Given the time-series nature of the data, the split was performed chronologically, with the first 80% of the observations used for model training and the remaining 20% used for testing. No random shuffling was applied in order to preserve temporal continuity and minimize information leakage from the testing set into the training process. The same splitting strategy was applied to all models to ensure consistency and comparability. This strategy provides an appropriate initial assessment of temporal generalization because the models are trained on earlier observations and evaluated on unseen data from a subsequent period. The 20% holdout set also provides a substantial number of observations for assessing predictive performance beyond the training period.

For models sensitive to feature scaling, including SVR, LSTM, and CNN–LSTM, the input variables were standardized using the mean and standard deviation calculated exclusively from the training set. The resulting scaling parameters were subsequently applied to the testing set. For LR, RF, and XGBoost, the data were processed according to the specific requirements of each algorithm; in particular, RF and XGBoost did not require feature scaling because of their tree-based structures. The prediction horizon differed between the conventional machine-learning and deep-learning models. LR, SVR, RF, and XGBoost used pollutant concentrations and meteorological variables measured at time t to estimate AQI at the same observation time, representing contemporaneous AQI estimation. In contrast, LSTM and CNN–LSTM used observations from the preceding 24 time steps (t−24 to t−1) to predict AQI at time t, representing one-step-ahead forecasting.

The hyperparameters of each model were selected and tuned using the training data. For SVR, the main hyperparameters included the penalty parameter C, the ε-insensitive loss parameter, and the kernel coefficient γ. For RF, the number of trees and the number of randomly selected predictors considered at each split were examined. For XGBoost, hyperparameters including the number of boosting rounds, learning rate, maximum tree depth, and subsampling ratio were adjusted to improve predictive performance while reducing the risk of overfitting.

For LSTM and CNN–LSTM, the data were transformed into sequences using a 24-time-step lag window, in which information from the preceding 24 observations was used to predict AQI at the subsequent time step. Both models were trained using the Adam optimizer, with mean squared error (MSE) as the loss function and a batch size of 32. The maximum number of training epochs was set to 150 for LSTM and 200 for CNN–LSTM. An early-stopping strategy based on validation loss was employed to reduce overfitting, with the model weights corresponding to the best validation performance retained for subsequent evaluation.

After training and optimization, all models were evaluated using the independent 20% testing set. Predicted AQI values were compared with the corresponding observed values using the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) to assess and compare predictive performance across the models. The input structures, major hyperparameters, and training configurations of all six models are summarized in Table 1.

Table 1. Model configurations, input structures, and training settings for the six AQI prediction models.

Model	Input structure	Lag window	Major hyperparameters	Optimization and computational settings
LR	12 input variables; tabular data	None	Ordinary Least Squares (OLS)	Analytical coefficient estimation; no iterative hyperparameter tuning
SVR	12 standardized input variables; tabular data	None	RBF kernel; C = 1; ε = 0.1; γ = 0.0833	Fixed hyperparameters; Z-score normalization; R package e1071; seed = 123
RF	12 input variables; tabular data	None	500 trees; mtry = 4; nodesize = 5; unrestricted maximum depth	Fixed hyperparameters; bootstrap sampling; R package randomForest; seed = 123
XGBoost	12 input variables; tabular data	None	nrounds = 500; η = 0.05; max_depth = 6; min_child_weight = 1; subsample = 0.8; colsample_bytree = 0.8; γ = 0	Fixed hyperparameters; squared-error objective; RMSE evaluation metric; R package xgboost; seed = 123
LSTM	Sequential input (24 × 12)	24 time steps	LSTM units = 64, 32; Dense = 16; dropout = 0.2; learning rate = 0.001	Adam optimizer; MSE loss; batch size = 32; maximum 150 epochs; validation split = 20%; early stopping patience = 15; R keras3
CNN–LSTM	Sequential input (24 × 12)	24 time steps	Conv1D filters = 64, 32; kernel size = 3; pooling size = 2; LSTM units = 64, 32; Dense = 16; dropout = 0.2	Adam optimizer; MSE loss; batch size = 32; maximum 200 epochs; validation split = 20%; early stopping patience = 15; learning-rate reduction factor = 0.5 (patience = 7); R keras3

2.5. Model Performance Evaluation Metrics

The predictive performance of the models was evaluated using three metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics were calculated using the following equations:

$$R^2 = 1 - \frac{\sum (y_{pi} - y_{oi})^2}{\sum y_{pi}^2} \quad (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_{pi} - y_{oi} | \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{pi} - y_{oi})^2}{n}} \quad (3)$$

3. Results

3.1. Descriptive analysis of AQI and input variables.

The descriptive statistics showed that the mean AQI was 89.70 ± 31.50, indicating substantial variability in air quality during the study period. The mean concentrations of $PM_{2.5}$ and PM_{10} were 45.20 ± 17.40 μg/m³ and 78.10 ± 28.50 μg/m³, respectively, demonstrating considerable variation in particulate matter concentrations. The mean concentrations of gaseous pollutants, including NO_2, SO_2, CO, and O_3, were 24.20 ± 6.38 $\mu g/m^3$, 6.82 ± 2.11 $\mu g/m^3$, 0.75 ± 0.19 mg/m^3, and 24.90 ± 15.10 $\mu g/m^3$, respectively. Meteorological variables also exhibited considerable variability, with mean air temperature of 25.00 ± 6.41 °C, relative humidity of 73.90 ± 7.37%, atmospheric pressure of 1009.00 ± 4.59 hPa, wind speed of 1.52 ± 1.06 m/s, and rainfall of 0.64 ± 2.05 mm. Overall, the observed variability in air pollutant concentrations and meteorological conditions provided a sufficiently heterogeneous dataset for developing and evaluating AQI forecasting models.

Table 2. Descriptive statistics of meteorological variables, air pollutant concentrations, and AQI in Hanoi during the study period

Variable	Unit	Mean	SD	Min	Median	Max
AQI	–	89.70	31.50	15.00	89.00	200.00
$PM_{2.5}$	$\mu g/m^3$	45.20	17.40	3.00	44.70	105.00
PM_{10}	$\mu g/m^3$	78.10	28.50	8.00	77.70	174.00
NO_2	$\mu g/m^3$	24.20	6.38	3.00	24.00	47.20
SO_2	$\mu g/m^3$	6.82	2.11	1.00	6.81	13.80
CO	mg/m^3	0.75	0.19	0.14	0.75	1.47
O_3	$\mu g/m^3$	24.90	15.10	2.00	22.20	72.40
Temperature	°C	25.00	6.41	8.39	25.00	40.00
Relative humidity	%	73.90	7.37	49.20	73.90	100.00
Atmospheric pressure	hPa	1009.00	4.59	997.00	1009.00	1020.00
Wind speed	m/s	1.52	1.06	0.10	1.29	7.50
Rainfall	mm	0.64	2.05	0.00	0.00	24.80

3.2. Correlation Analysis.

Pearson correlation analysis was performed to evaluate the strength and direction of linear relationships between AQI and the input variables (Figure 2). The results showed that AQI was very strongly positively correlated with $PM_{2.5}$ ($r = 0.97$), followed by PM_{10} ($r = 0.93$), and was also strongly positively correlated with CO ($r = 0.75$) and NO_2 ($r = 0.61$). These findings indicate that AQI variability in the study dataset was particularly closely associated with particulate pollutants. The prominent role of $PM_{2.5}$ in determining air quality in Hanoi has also been reported in previous studies, in which variations in $PM_{2.5}$ concentrations were associated with local emission sources, meteorological conditions, and long-range pollutant transport [15]. These findings are also consistent with reviews of machine-learning applications in air quality forecasting, in which $PM_{2.5}$, PM_{10}, and gaseous pollutants are commonly identified as important predictors of AQI and other air quality indicators [16].

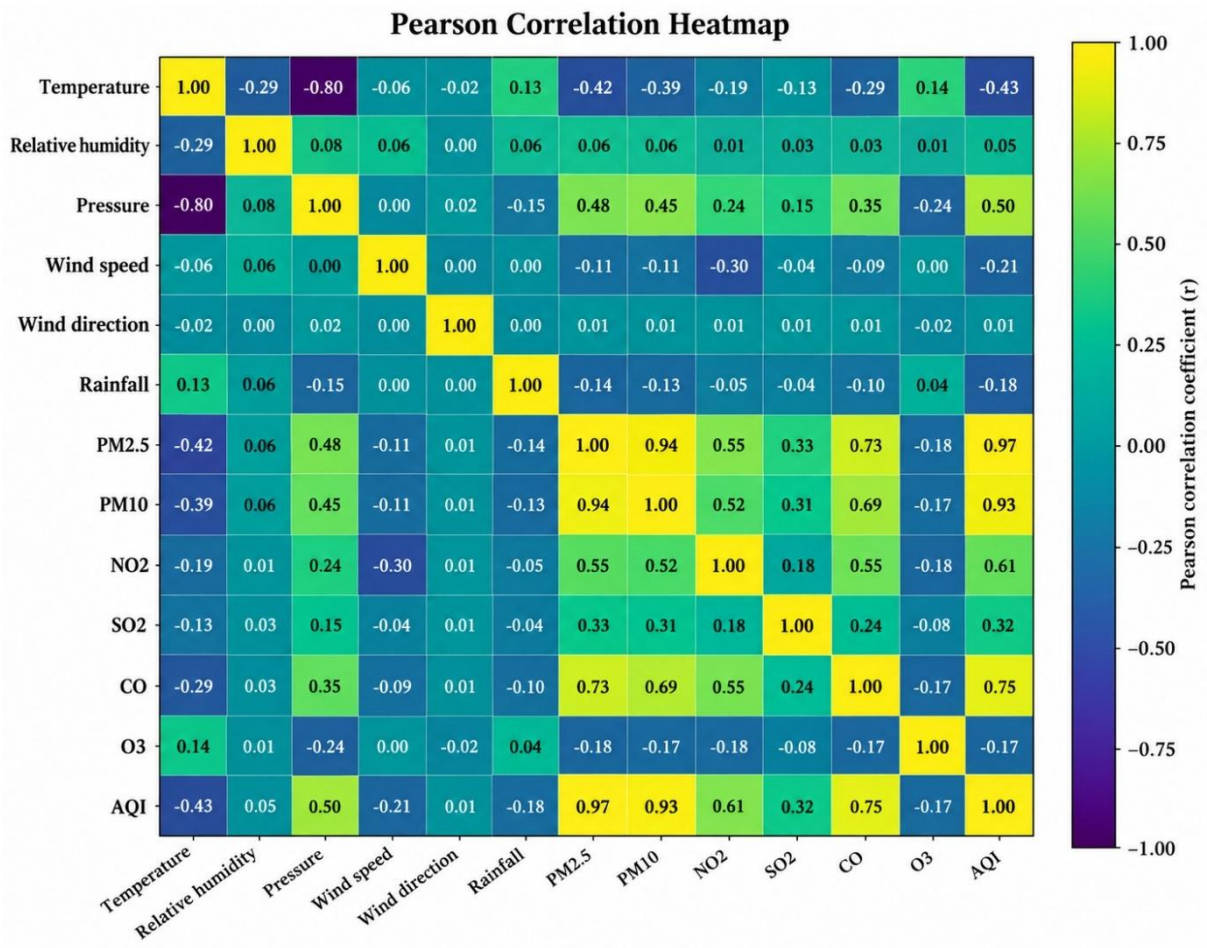

Figure 2. Pearson correlation matrix of AQI, air pollutants, and meteorological variables.

Notably, substantial correlations were also observed among the air pollutants, particularly between $PM_{2.5}$ and PM_{10} (r = 0.94), $PM_{2.5}$ and CO (r = 0.73), PM_{10} and CO (r = 0.69), $PM_{2.5}$ and NO_2 (r = 0.55), and NO_2 and CO (r = 0.55). These relationships may reflect the co-variation of pollutants associated with common emission sources and/or shared meteorological conditions. The strong correlations among $PM_{2.5}$, PM_{10}, CO, and NO_2 may reflect contributions from common combustion-related sources, particularly traffic emissions, while the association between $PM_{2.5}$ and PM_{10} may also be influenced by their overlapping particle-size fractions. The positive correlations of SO_2 with $PM_{2.5}$ (r = 0.33) and PM_{10} (r = 0.31) may be associated with sulfur-containing fuel combustion and secondary aerosol formation. In contrast, the weak negative correlations between O_3 and NO_2 (r = −0.18) and between O_3 and $PM_{2.5}$ (r = −0.18) may reflect differences in photochemical processes and atmospheric conditions, although these relationships do not establish causality. Furthermore, the very high correlation between $PM_{2.5}$ and PM_{10} indicates the potential for multicollinearity when applying linear models. Therefore, the variance inflation factor (VIF) should also be examined when the interpretation of regression coefficients is an objective of the analysis. Overall, particulate pollutants, particularly $PM_{2.5}$ and PM_{10}, exhibited the strongest associations with AQI, whereas meteorological variables generally showed weak to moderate correlations. The differences in both the magnitude and direction of correlations between the input variables and AQI further support the evaluation of nonlinear machine-learning models capable of capturing more complex relationships in AQI forecasting.

Among the meteorological variables, atmospheric pressure showed a moderate positive correlation with AQI ($r = 0.50$), whereas temperature exhibited a moderate negative correlation ($r = -0.43$). Wind speed ($r = -0.21$) and rainfall ($r = -0.18$) showed weak negative correlations, while relative humidity ($r = 0.05$) and wind direction ($r = 0.01$) exhibited little or no linear association with AQI. O_3 also showed a weak negative correlation with AQI ($r = -0.17$). Overall, these results are consistent with previous studies conducted in Hanoi, which have demonstrated that meteorological conditions influence particulate matter concentrations through atmospheric dispersion, transport, accumulation, and removal processes [5, 15]. However, a low correlation coefficient for an individual meteorological variable does not necessarily imply that the variable has a negligible influence on air quality, as meteorological effects may be nonlinear, season-dependent, and interact with multiple environmental factors.

3.3. Comparison of Model Performance.

The results revealed clear differences in AQI predictive performance among the evaluated models (Table 3 and Figure 3). LR achieved the best overall performance, with an RMSE of 5.274, MAE of 4.122, and R^2 of 0.963, followed by XGBoost (RMSE = 5.665; MAE = 4.484; R^2 = 0.957) and Random Forest (RMSE = 5.809; MAE = 4.582; R^2 = 0.955). This pattern is also clearly illustrated in Figure 3, where the predictions generated by LR, XGBoost, and Random Forest are closely distributed around the 1:1 reference line, indicating strong agreement between observed and predicted AQI values.

Table 3. Performance evaluation of machine learning and deep learning models for AQI prediction in Hanoi.

Model	RMSE	MAE	R^2
LR	5.274	4.122	0.963
XGBoost	5.665	4.484	0.957
Random Forest	5.809	4.582	0.955
SVR	11.101	7.505	0.836
LSTM	20.630	16.314	0.439
CNN–LSTM	21.739	17.248	0.377

SVR exhibited lower predictive performance (RMSE = 11.101; MAE = 7.505; R^2 = 0.836), as reflected by the greater dispersion of predicted values around the 1:1 reference line, particularly at higher AQI levels (Figure 3). In contrast, LSTM and CNN–LSTM showed the greatest prediction dispersion, with R^2 values of only 0.439 and 0.377, respectively. For these two models, high AQI values tended to be underestimated, whereas some low AQI values were overestimated, indicating a tendency for predictions to regress toward the central range of AQI values. Moreover, CNN–LSTM did not improve predictive performance compared with LSTM, suggesting that increasing network architectural complexity did not provide additional predictive benefits for the present dataset.

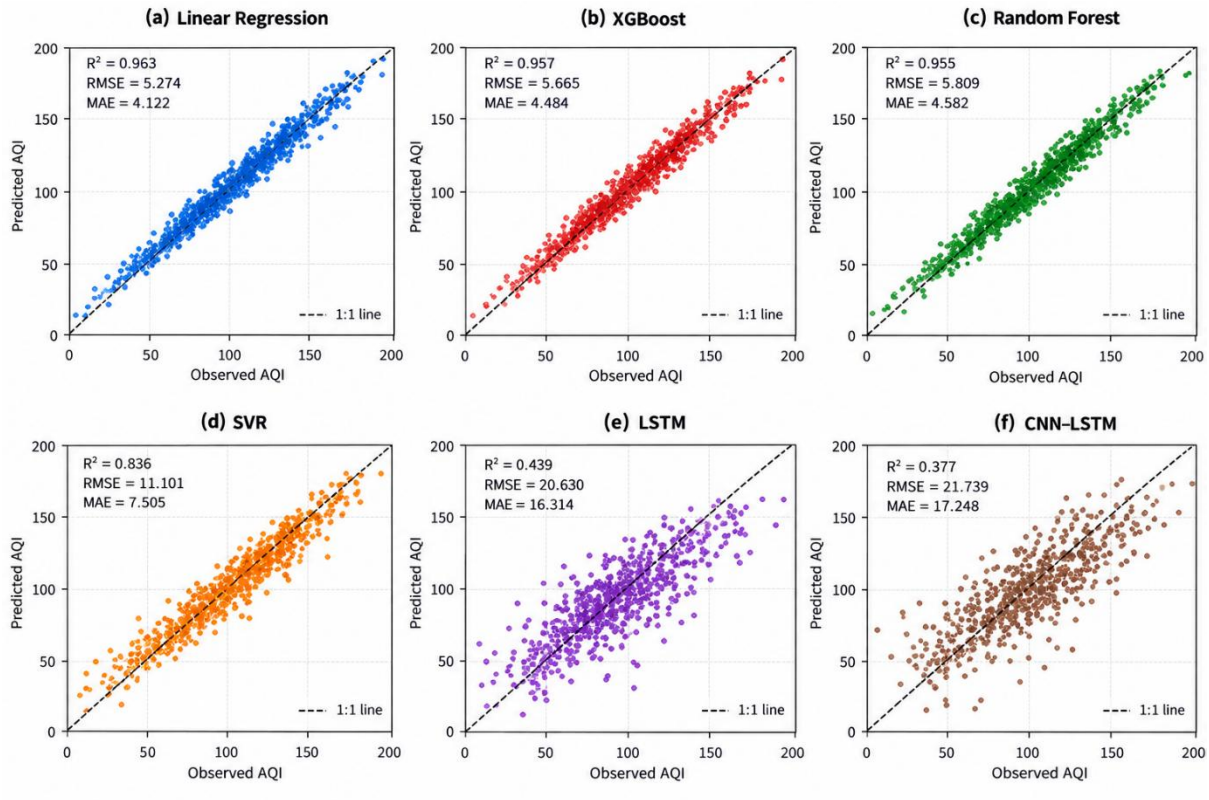

Figure 3. Observed versus predicted AQI values for the six forecasting models.

Figure 4 compares the observed and estimated AQI values obtained from LR, XGBoost, and RF over a selected period of the testing dataset. All three models closely reproduced the observed AQI fluctuations, including rapid increases and decreases in pollution levels. However, differences became more apparent during high-AQI episodes. In particular, XGBoost and RF tended to underestimate the highest pollution peak (approximately AQI 200), whereas LR provided a closer estimate. These findings are consistent with the lower RMSE and MAE values obtained for LR.

Previous studies have applied various machine learning and deep learning approaches to AQI prediction across different urban environments, with considerable differences in datasets, input variables, forecasting horizons, and validation strategies (Table 4). For example, Liu and Guo (2022) identified LSTM–SSA as the best-performing model for AQI prediction in Shanghai, whereas Pande et al. (2024) reported superior performance of Bi-RNN for daily AQI forecasting in Delhi [17] [18]. In contrast, the present study found that LR achieved the highest accuracy for contemporaneous AQI estimation in Hanoi.

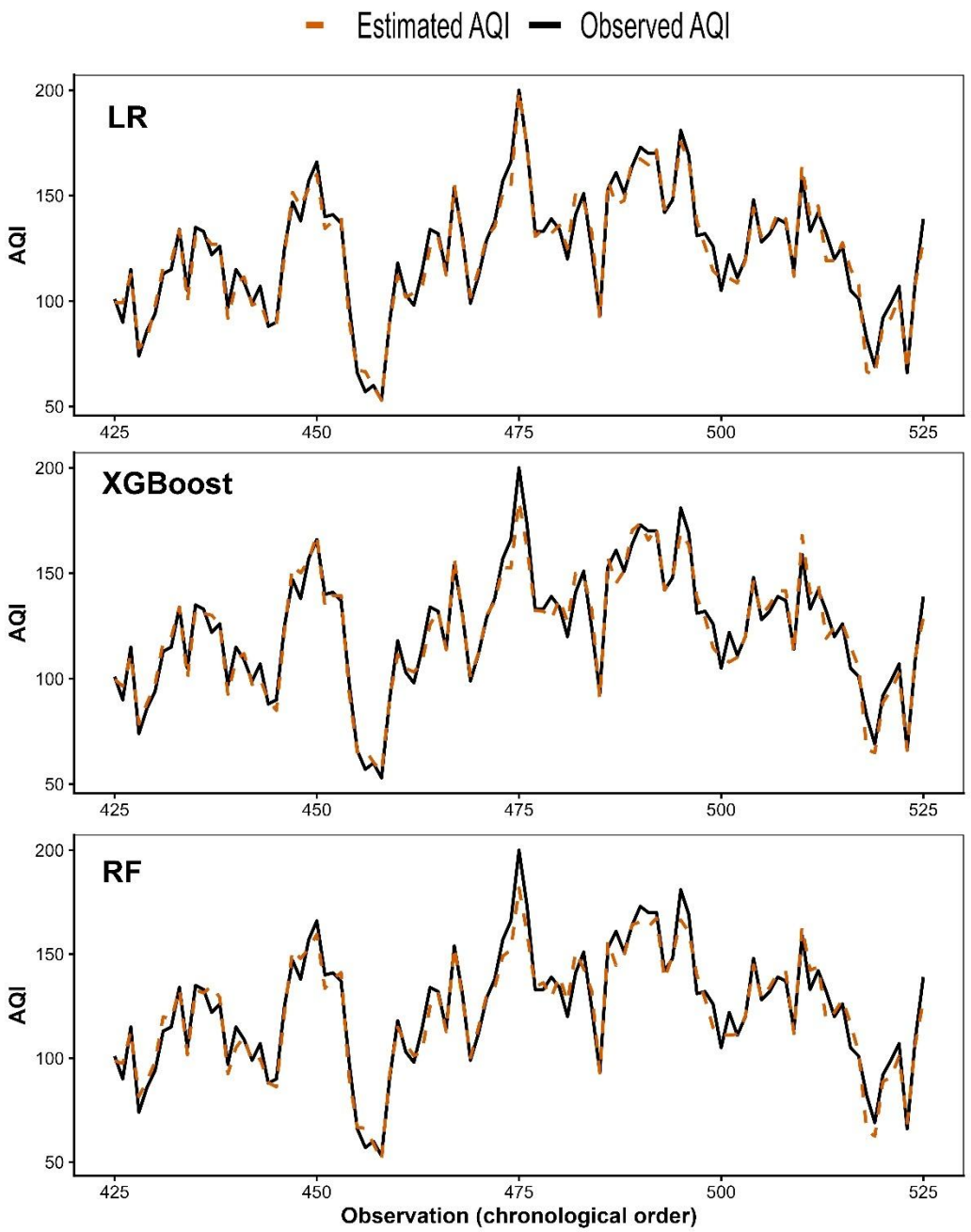

Figure 4. Comparison of observed and estimated AQI variations using Linear Regression (LR), XGBoost, and Random Forest (RF) models over a selected period of the testing dataset.

These results indicate that greater model complexity does not necessarily translate into better predictive performance for the present dataset. A large-scale review by Méndez et al. (2023) similarly demonstrated that model performance depends substantially on the characteristics of the dataset, input variables, and forecasting design [16]. More recently, Chianese and Riccio (2026) emphasized that no single model consistently performs best under all conditions when models are evaluated using leakage-controlled rolling evaluation procedures; rather, performance depends on factors such as the forecast horizon, predictor availability, monitoring regime, and validation design [19]. Therefore, the findings of the present study are more appropriately interpreted as evidence that the relative advantages of different algorithmic approaches depend on the structure and characteristics of the prediction task, rather than as

evidence that conventional machine-learning methods are inherently superior to deep-learning approaches.

Table 4. Comparison of the present study with previous machine learning-based AQI prediction studies.

Location	Dataset	Predictors	Horizon	Algorithms	Validation	Performance	Ref
Shanghai, China	Hourly, Feb 2021–Jan 2022	Six air pollutants	Hourly, daily, weekly	RF, XGBoost, MLP, LSTM, LSTM–SSA	NR	LSTM–SSA; AQI MAPE 1.53%	[17]
Delhi, India	Daily; size NR	NR	Daily AQI	LSTM, Bi-LSTM, Bi-RNN, KRR	Training/testing; details NR	Bi-RNN; R^2 = 0.954, RMSE = 25.755	[18]
Dehradun and Kashipur, India	Size NR	Six air pollutants	NR	LR, RF, SVR, XGBoost, LSTM, GRU, others	Holdout and five-fold walk-forward	RF (Dehradun holdout); R^2 = 0.995, RMSE = 4.60	[20]
Present study, Hanoi, Vietnam	8,784 observations, 2024–2025	12 pollutant and meteorological variables	Contemporaneous estimation / one-step-ahead forecasting	LR, SVR, RF, XGBoost, LSTM, CNN–LSTM	Chronological 80:20	LR (estimation): R^2 = 0.963, RMSE = 5.274; LSTM (forecasting): R^2 = 0.439, RMSE = 20.630	This study

A particularly noteworthy finding was that LR, the simplest model evaluated, achieved the highest predictive accuracy. This result may be attributed to the very strong relationships between AQI and several pollutant predictors, particularly $PM_{2.5}$ ($r = 0.97$) and PM_{10} ($r = 0.93$), followed by CO ($r = 0.75$) and NO_2 ($r = 0.61$). The method used to calculate AQI provides an important basis for interpreting this finding. Pollutant-specific indices are derived from pollutant concentrations using predefined breakpoints and linear interpolation within each concentration interval, after which the overall AQI is determined from the pollutant-specific indices. Consequently, when the concentrations of $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 are directly included as predictors, the relationship between the predictors and AQI inherently contains a piecewise-linear structure [21]. This characteristic may partly explain why a relatively simple linear model achieved a very high R^2 value of 0.963.XGBoost and RF also achieved excellent predictive performance, with only small differences compared with LR. This finding demonstrates that tree-based ensemble models can effectively capture the underlying structure of AQI data. Méndez et al. (2023) reported that RF, boosting algorithms, and other nonlinear ML methods are widely used in air quality forecasting, particularly when complex interactions among predictors are present [16]. A recent study on $PM_{2.5}$ forecasting also demonstrated that the relative performance of statistical models, tabular ML approaches, and sequence-based DL models varies according to the forecast horizon and data conditions, highlighting the importance of evaluating different model families within a common experimental framework [22]. In the present study, the relatively small performance differences among LR, XGBoost, and RF suggest that linear relationships may account for a substantial proportion of the predictive structure in the dataset, while ensemble models may additionally capture residual nonlinear relationships and interactions.

For SVR, the model achieved an RMSE of 11.101, MAE of 7.505, and R^2 of 0.836, representing substantially lower performance than LR, XGBoost, and RF. However, this result does not necessarily indicate that SVR is inherently unsuitable for AQI prediction, as the performance of kernel-based algorithms is highly dependent on data characteristics, preprocessing procedures, and model configuration. Méndez et al. (2023) identified SVM/SVR as a widely used approach in air quality forecasting, although its relative performance varies

considerably across studies [16]. As hyperparameters were predefined rather than systematically optimized across models, the relative ranking of SVR and other algorithms should be interpreted with caution. Future studies should apply consistent hyperparameter optimization procedures to improve the robustness of model comparisons. This consideration further highlights the importance of applying consistent hyperparameter-tuning procedures across models to ensure a fair comparison.

LSTM and CNN–LSTM exhibited substantially lower predictive performance than the conventional machine-learning models. The distribution of predicted values showed that both models tended to underestimate some high AQI values while overestimating some low AQI values, resulting in a narrower prediction range concentrated around the central AQI values. Moreover, CNN–LSTM did not improve performance over LSTM, with R^2 values of 0.377 and 0.439, respectively, indicating that increasing architectural complexity did not provide additional predictive benefits under the current data conditions. Previous studies have shown that the performance of recurrent and hybrid neural networks depends substantially on data availability, temporal structure, input-window design, regularization, and validation procedures. Furthermore, insufficient high-resolution spatiotemporal data may increase the risk of overfitting in deep-learning models [23]. Therefore, the relatively poor performance of LSTM and CNN–LSTM may be attributable to suboptimal representation of temporal dependencies, an insufficient amount of data to fully exploit the advantages of deep neural networks, or the possibility that much of the predictive information was already captured by the pollutant predictors.

An important methodological consideration is the difference in temporal input configurations among the models. LR, RF, XGBoost, and SVR used pollutant concentrations and meteorological variables measured at the same time as the target AQI, whereas LSTM and CNN–LSTM used observations from the preceding 24 time steps to predict the subsequent AQI value. Therefore, the conventional models had direct access to contemporaneous pollutant concentrations involved in AQI calculation, while the deep-learning models relied on historical information. This difference may substantially contribute to the observed performance gap and limits the direct comparability of the models. This design also raises concerns about potential target leakage, as contemporaneous pollutant concentrations used to calculate AQI were included as predictors. Although this does not necessarily constitute data leakage in the conventional sense, it creates a direct mathematical relationship between the inputs and the target, potentially inflating the apparent predictive performance of LR and other conventional models. Therefore, their results should be interpreted primarily as contemporaneous AQI estimation rather than genuine future AQI forecasting. In practical applications, conventional models could support real-time AQI monitoring in Hanoi, whereas LSTM and CNN–LSTM may contribute to short-term early-warning systems after further improvement and validation. Moreover, although the dataset contained 8,784 observations, the 24-step input window may not adequately capture longer-term or seasonal variations. The absence of explicit temporal predictors, such as hour of day, day of week, and seasonal indicators, may further limit the ability of deep-learning models to represent temporal variability. Additional limitations include the single-station spatial coverage, potential delays or missing predictor data, reduced reliability during extreme pollution events, and uncertain generalizability to other monitoring locations. These explanations remain hypotheses that require further evaluation using

consistent forecast horizons and alternative input-window configurations. Future studies should incorporate multi-station datasets, improve extreme-event forecasting, and apply rolling-origin and external validation to assess operational applicability. Overall, LR, XGBoost, and RF achieved higher accuracy for contemporaneous AQI estimation, whereas LSTM and CNN–LSTM exhibited lower accuracy in the one-step-ahead forecasting task. These results should not be interpreted as evidence of the general superiority of conventional machine-learning methods over deep-learning approaches for future AQI forecasting. Instead, they emphasize the importance of aligning prediction horizons, predictor availability, and validation strategies when selecting and evaluating AQI prediction models.

4. Conclusions

This study systematically compared six conventional machine-learning and deep-learning models, including LR, SVR, RF, XGBoost, LSTM, and CNN–LSTM, for AQI forecasting in Hanoi, Vietnam, using 8,784 observations collected during 2024–2025. The results demonstrated substantial differences in predictive performance among the evaluated models. LR achieved the best performance, with an RMSE of 5.274, MAE of 4.122, and R^2 of 0.963, followed closely by XGBoost (RMSE = 5.665, MAE = 4.484, R^2 = 0.957) and RF (RMSE = 5.809, MAE = 4.582, R^2 = 0.955). In contrast, SVR showed lower accuracy, while LSTM and CNN–LSTM exhibited substantially poorer predictive performance under the investigated data and model configurations. The strong correlations of AQI with $PM_{2.5}$ (r = 0.97) and PM_{10} (r = 0.93) indicate that particulate pollutants contained substantial information relevant to AQI prediction and may partly explain the strong performance of the relatively simple LR model. Overall, the findings demonstrate that greater model complexity does not necessarily result in improved AQI predictive performance. For the present dataset, LR provided the best balance of predictive accuracy, computational simplicity, and interpretability, while XGBoost and RF also showed strong potential for practical AQI prediction. The poorer performance of LSTM and CNN–LSTM should therefore not be interpreted as a general limitation of deep learning but rather as evidence that model effectiveness depends on data characteristics, temporal structure, input configuration, and validation strategy. Future studies should evaluate these models using longer-term and higher-resolution spatiotemporal datasets, additional monitoring locations, different forecast horizons, and systematic hyperparameter optimization. Such investigations would provide a more comprehensive assessment of model generalizability and support the development of reliable early-warning systems for air quality management in Hanoi and other rapidly urbanizing cities.

Acknowledgments

This research received no external funding.

Author Contribution

Si Duc Truong: Conceptualization, Methodology, Data Analysis, Writing, Nguyen Van Thanh: Writing-Review and Editing.

Competing Interest

The authors declare no competing interests.

References

[1] WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide; World Health Organization: Geneva, Switzerland. (accessed on 1 June 2026) Available online: https://www.who.int/publications/i/item/9789240034228.

[2] Horn, S.A.; Dasgupta, P.K. (2024). The Air Quality Index (AQI) in historical and analytical perspective: A tutorial review. *Talanta*, *267*, 125260. https://doi.org/10.1016/j.talanta.2023.125260.

[3] Nghiem, T.-D.; Nguyen, T.T.T.; Nguyen, T.T.H.; Ly, B.-T.; Sekiguchi, K.; Yamaguchi, R.; Pham, C.-T.; Ho, Q.B.; Nguyen, M.-T.; Duong, T.N. (2020). Chemical characterization and source apportionment of ambient nanoparticles: A case study in Hanoi, Vietnam. *Environmental Science and Pollution Research*, *27*, 30661–30672. https://doi.org/10.1007/s11356-020-09417-5.

[4] Ly, B.-T.; Kajii, Y.; Nguyen, T.-Y.-L.; Shoji, K.; Van, D.-A.; Do, T.-N.-N.; Nghiem, T.-D.; Sakamoto, Y. (2020). Characteristics of roadside volatile organic compounds in an urban area dominated by gasoline vehicles, a case study in Hanoi. *Chemosphere*, *254*, 126749. https://doi.org/10.1016/j.chemosphere.2020.126749.

[5] Tran, C.C.; Ta, T.D.; Duong, A.T.; Phan, O.T.K.; Nguyen, D.A. (2020). Analysis on temporal pattern of fine particulate matter (PM2.5) in Hanoi, Vietnam and the impact of meteorological conditions. *Journal of Environmental Protection*, *11*, 246–256. https://doi.org/10.4236/jep.2020.113014.

[6] Thanh, N.V.; Thuong Giang, P.; Thuy, N.T.T. (2026). Applications of artificial intelligence in environmental management in Vietnam: A mini review. *Sustainable Environmental Insight*, *3*. https://doi.org/10.53623/sein.v3i1.1008.

[7] Su, X.; Yan, X.; Tsai, C.L. (2012). Linear regression. *WIREs Computational Statistics*, *4*, 275–294. https://doi.org/10.1002/wics.1198.

[8] Awad, M.; Khanna, R. (2015). Support vector regression. In Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers; Apress: Berkeley, CA, USA; pp. 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4.

[9] Salman, H.A.; Kalakech, A.; Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007.

[10] Chen, T.; Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: New York, NY, USA; pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[11] Karmoude, M.; Munhungewarwa, B.; Chiraira, I.; McKenzie, R.; Kong, J.; Smith, B.; Ayana, G.; Njara, N.; Mathaha, T.; Kumar, M.; Mellado, B. (2025). Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks. *Science of the Total Environment*, *1002*, 180593. https://doi.org/10.1016/j.scitotenv.2025.180593.

[12] Graves, A. (2012). Long short-term memory. In Supervised Sequence Labelling with Recurrent Neural Networks; Studies in Computational Intelligence, Volume 385; Springer: Berlin/Heidelberg, Germany; pp. 37–45. https://doi.org/10.1007/978-3-642-24797-2_4.

[13] Zaini, N.A.; Ean, L.W.; Ahmed, A.N.; Malek, M.A. (2022). A systematic literature review of deep learning neural network for time series air quality forecasting. *Environmental Science and Pollution Research*, *29*, 4958–4990. https://doi.org/10.1007/s11356-021-17442-1.

[14] Zhang, J.; Li, S. (2022). Air quality index forecast in Beijing based on CNN-LSTM multi-model. *Chemosphere*, *308*, 136180. https://doi.org/10.1016/j.chemosphere.2022.136180.

[15] Ly, B.-T.; Matsumi, Y.; Vu, T.V.; Sekiguchi, K.; Nguyen, T.-T.; Pham, C.-T.; Nghiem, T.-D.; Ngo, I.-H.; Kurotsuchi, Y.; Nguyen, T.-H.; Nakayama, T. (2021). The effects of meteorological conditions and long-range transport on PM2.5 levels in Hanoi revealed from multi-site

measurement using compact sensors and machine learning approach. *Journal of Aerosol Science*, *152*, 105716. https://doi.org/10.1016/j.jaerosci.2020.105716.

[16] Méndez, M.; Merayo, M.G.; Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. *Artificial Intelligence Review*, *56*, 10031–10066. https://doi.org/10.1007/s10462-023-10424-4.

[17] Liu, X.; Guo, H. (2022). Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): A case study of Shanghai. *Atmospheric Pollution Research*, *13*, 101551. https://doi.org/10.1016/j.apr.2022.101551.

[18] Pande, C.B.; Kushwaha, N.L.; Alawi, O.A.; Sammen, S.S.; Sidek, L.M.; Yaseen, Z.M.; Pal, S.C.; Katipoğlu, O.M. (2024). Daily scale air quality index forecasting using bidirectional recurrent neural networks: Case study of Delhi, India. *Environmental Pollution*, *351*, 124040. https://doi.org/10.1016/j.envpol.2024.124040.

[19] Chianese, E.; Riccio, A. (2026). Air-quality forecasting across monitoring, predictor, and validation regimes: A systematic mapping review and decision framework. *Forecasting*, *8*, 72. https://doi.org/10.3390/forecast8040072.

[20] Bhatt, D.; Goswami, S.; Verma, G.; Pandey, B.K. (2026). Air quality index prediction using machine learning and deep learning: A comparative analysis of Dehradun and Kashipur in Uttarakhand, India. *Advances in Research*, *27*, 116–130. https://doi.org/10.9734/air/2026/v27i41661.

[21] How is the AQI calculated? (accessed on 1 June 2026) Available online: https://www.epa.gov/outdoor-air-quality-data/how-aqi-calculated.

[22] Chisilev, B.-I.; Pelican, E. (2026). Assessment of machine learning methods for urban air pollution forecasting. *Stochastic Environmental Research and Risk Assessment*, *40*. https://doi.org/10.1007/s00477-026-03331-x.

[23] Lawrence, S.; Bhathmanabhan, S. (2026). Harnessing deep learning for air pollution forecasting: Trends, techniques, and future prospects. *Artificial Intelligence Review*, *59*. https://doi.org/10.1007/s10462-026-11496-8.



© 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).