Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD38812

Indoor Air Quality Prediction Using Artificial Neural Networks

Yuchen Yang

School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
25531274@bjtu.edu.cn

Abstract. To investigate variations in indoor NO_2 concentrations under a subtropical climate, their relationships with indoor and outdoor environment parameters, and ways to improve prediction accuracy, this study monitored indoor and outdoor air quality in a teaching building from May 18 to May 31, 2025. Measurements included NO_2, PM2.5, PM10, O_3, temperature, and relative humidity. Pearson correlation coefficients were calculated, and variables with $|r| \geq 0.3$ were selected as model inputs. Multiple linear regression (MLR) and artificial neural network (ANN) models were developed and evaluated using R^2 and RMSE. Based on validation results, the preferred ANN used two hidden layers, with 150 and 75 neurons, and ReLU activation. On the test set, MLR achieved an R^2 of 0.5975 and an RMSE of 4.8597, whereas ANN achieved an R^2 of 0.8528 and an RMSE of 2.9389. Compared with MLR, ANN increased R^2 by approximately 25.53 percentage points and reduced RMSE by approximately 39.52%. ANN also reduced the underestimation observed at high concentrations and improved accuracy across concentration ranges. These findings show that indoor NO_2 depends on multiple indoor and outdoor factors and that ANN captures their complex relationships more effectively than MLR. The results support ANN-based indoor NO_2 prediction under subtropical conditions and may inform air-quality monitoring and forecasting in similar buildings. However, the study was limited by its short monitoring period, incomplete coverage of relevant factors, and limited neural-network optimization. Future studies should address these limitations.

Keywords: Artificial neural networks, indoor air quality, multiple linear regression

1. Introduction

With accelerating urbanization and continued growth in transportation and industrial activity, nitrogen oxide pollution has become an important factor affecting urban air quality. Nitrogen dioxide (NO_2) is emitted mainly by fuel combustion and motor vehicles, and long-term exposure to elevated NO_2 concentrations may adversely affect the human respiratory system [1]. Because people spend most of their time indoors, indoor NO_2 concentrations and their temporal variations are closely related to actual human exposure. Indoor NO_2 may arise from activities such as indoor combustion and may also be influenced by outdoor pollutants entering buildings through mechanical ventilation and air infiltration [2, 3]. In hot and humid tropical climates, building ventilation operation and indoor-outdoor humidity conditions have distinctive characteristics, making the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.LD38812

relationship between indoor and outdoor pollutants more complex. Research on indoor NO_2 variation and prediction under such climatic conditions therefore has practical significance [4].

Researchers in China and elsewhere have extensively studied the sources of indoor NO_2, indoor-outdoor relationships, and influencing factors. Previous studies show that outdoor NO_2 concentrations, building ventilation, air-exchange processes, and indoor emission sources can all affect indoor NO_2 levels [5-7]. Temperature, relative humidity, and other air pollutants may also be interrelated, causing indoor NO_2 concentrations to reflect the combined effects of multiple factors. A single environmental parameter therefore cannot fully characterize indoor NO_2 variation; several indoor and outdoor variables must be analyzed together.

For air-pollutant prediction, multiple linear regression (MLR) is widely used to analyze relationships between environmental variables and pollutant concentrations because of its simple structure, computational efficiency, and interpretability [8-10]. However, MLR is based primarily on an assumption of linear relationships, and its predictive ability may be limited when the input variables and target pollutant have complex nonlinear relationships. Artificial neural networks (ANNs) can learn complex input-output mappings through hidden-layer neurons and nonlinear activation functions. In recent years, they have increasingly been applied to the prediction of PM2.5, NO_2, and other air pollutants, demonstrating strong nonlinear fitting capability [11-13]. ANN performance, however, depends on the hidden-layer architecture, number of neurons, activation function, and other parameters; the network architecture must therefore be optimized for the specific monitoring data [14, 15].

Research on indoor NO_2 prediction in tropical climates remains limited. Existing studies often examine outdoor pollution or indoor–outdoor pollutant relationships. Few studies comprehensively assess indoor NO_2 and multiple indoor and outdoor environmental factors. Building type, ventilation, and climate also affect these relationships. Therefore, existing models may not transfer directly across real-world buildings. Differences between linear and nonlinear models also require evaluation using the same dataset. Synchronized indoor and outdoor monitoring is needed to identify key factors and compare prediction methods.

This study investigated a teaching building in a tropical city. The building mainly contains offices and uses mechanical ventilation. Doors and windows remained closed during monitoring. Two AQ-SPEC-validated Air Quality Egg devices from 2018 were installed indoors and outdoors. Synchronized measurements were collected continuously from May 18 to May 31, 2025. Indoor and outdoor NO_2, PM2.5, PM10, O_3, temperature, and relative humidity were recorded every 10 minutes. The dataset contained 2,016 paired observations. Pearson correlation coefficients assessed relationships between environmental variables and indoor NO_2. Variables with $|r| \geq 0.3$ were selected as model inputs. MLR and ANN models were then developed. Data were randomly divided into training and test sets using an 80:20 ratio. ANN training data were further divided into training and validation subsets. Different hidden-layer structures and activation functions were compared. The preferred network was selected from the validation results. Both models were evaluated using R^2 and RMSE.

This study makes three main contributions. First, synchronized indoor and outdoor observations were used under tropical conditions. These data revealed relationships between indoor NO_2 and multiple environmental parameters. They also provided evidence of indoor NO_2 variation. Second, Pearson analysis identified variables closely associated with indoor NO_2. This reduced redundant inputs and supported model development. Third, MLR and ANN models were developed and compared. The ANN architecture was optimized using a validation set. Both models were tested on

DOI: 10.54254/2754-1169/2026.LD38812

the same dataset. This design evaluated linear and nonlinear approaches consistently. The findings may support indoor NO_2 forecasting and air-quality management in tropical buildings.

2. Related concepts and key technologies

2.1. Pearson correlation coefficient

In indoor-air-quality research, pollutant concentrations are generally not determined by a single environmental variable. Instead, they are jointly affected by the transport of multiple indoor and outdoor pollutants, indoor and outdoor temperature, relative humidity, and other factors. Before developing a prediction model, it is therefore necessary to analyze the relationships among the monitored variables, determine how strongly each environmental variable is associated with the target pollutant, and provide a basis for selecting model inputs. The Pearson correlation coefficient is a statistic that describes the degree of linear association between two continuous variables and ranges from -1 to 1. A positive coefficient indicates an overall positive linear trend, whereas a negative coefficient indicates an inverse trend. The closer the absolute value is to 1, the stronger the linear association; the closer it is to 0, the weaker the association.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(-\bar{y})^2}} \quad (1)$$

Equation (1) was used to calculate the Pearson correlation coefficient. Here, x_i and y_i are the paired observations of the two variables, while $\bar{x}$ and $\bar{y}$ represent their mean values. The sample size is given by n, and r represents the resulting correlation coefficient. In this study, Pearson's r was calculated between indoor NO_2 concentration and each indoor or outdoor environmental variable. The correlation results were then considered when selecting the input variables for subsequent model development.

2.2. Multiple linear regression

Multiple linear regression (MLR) was employed to describe the relationship between indoor NO_2 concentration and the selected environmental variables. Model coefficients were estimated using the least-squares criterion, which seeks to reduce the difference between the observed and predicted values. The regression model can be written as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (2)$$

In Equation (2), y is the indoor NO_2 concentration, and $x_1, x_2, \ldots, x_k$ are the input variables used in the model. The constant term is the intercept, while the other coefficients indicate how the input variables are related to the predicted NO_2 concentration. The final term represents the error that cannot be explained by the regression model. The coefficients were determined by minimizing the squared differences between measured and predicted values. MLR was then used to evaluate the linear relationship between the selected variables and indoor NO_2 concentration [8, 9].

2.3. Artificial neural networks

Artificial neural networks (ANNs) use interconnected neurons arranged in different layers to process input data and produce model outputs. The ANN developed in this study contains an input layer,

DOI: 10.54254/2754-1169/2026.LD38812

hidden layer(s), and an output layer. (1) For the input layer, each neuron corresponds to one environmental variable selected from the correlation analysis.(2) The initial design uses one hidden layer with a sigmoid activation function (range [0, 1]) for nonlinear mapping. (3) The output layer has one neuron corresponding to the predicted indoor NO_2 concentration and uses the Purelin linear activation function, which is suitable for continuous outputs. (4) The model uses backpropagation to adjust weights and thresholds and minimize error between predicted and observed values [11-13].

3. Study area and data sources

Field measurements were conducted in a teaching building located in a tropical city from May 18 to May 31, 2025. The monitored office space had an area of approximately 35 m^2 and was mechanically ventilated throughout the year, with the doors and windows generally kept closed. During the monitoring period, indoor and outdoor environmental parameters were recorded continuously to characterize their temporal variations under the local climatic conditions.Two 2018 Air Quality Egg devices were installed indoors and outdoors. The U.S. EPA AQ-SPEC had validated both devices. The instruments recorded data automatically every 10 minutes. They operated continuously throughout the monitoring period. The indoor device was positioned 1 m above the floor. It was also 0.5 m from an interior wall. The outdoor device was installed 1.8 m above an exterior platform. Wind and rain protection ensured stable outdoor measurements. Monitored variables included PM2.5, PM10, NO_2, temperature, and relative humidity. The study collected 2,016 paired indoor–outdoor observations. These observations characterized environmental changes from late spring to early summer.

4. Research methods

4.1. Correlation analysis of air pollutants

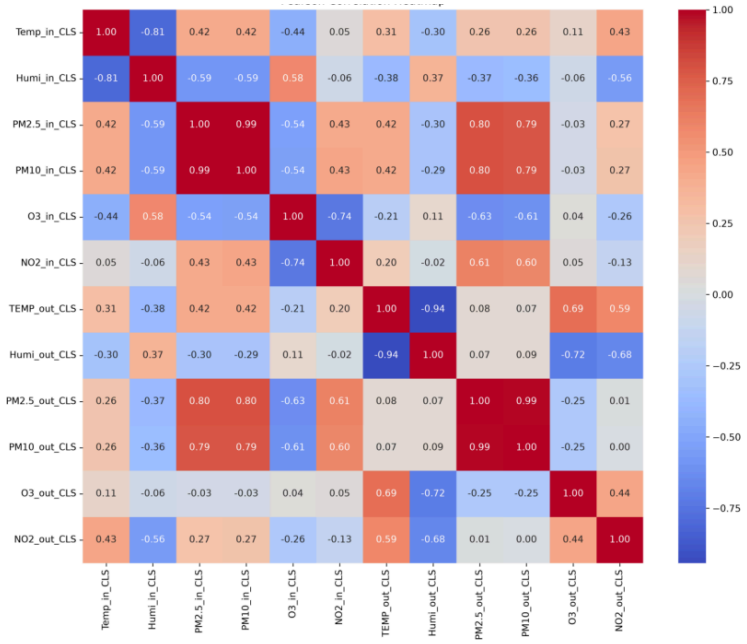


Figure 1. Heatmap of indoor and outdoor environmental parameters

DOI: 10.54254/2754-1169/2026.LD38812

Using Python and the indoor and outdoor environmental dataset, we first applied the Pearson correlation coefficient to generate a correlation matrix and heatmap for selecting model inputs. The results are shown in Figure 1.

We then analyzed the correlation results for indoor NO_2 and selected the principal independent variables by excluding variables with $|r| < 0.3$. The main influencing factors are shown in Table 1.

Table 1. Main influencing factors

Independent variable	Correlation coefficient r	Direction	Strength
PM2.5_in_CLS	0.43	Positive	Moderate
PM10_in_CLS	0.43	Positive	Moderate
O3_in_CLS	-0.74	Negative	Strong
PM2.5_out_CLS	0.61	Positive	Strong
PM10_out_CLS	0.60	Positive	Strong

The principal factors affecting indoor NO_2 and their possible mechanisms were analyzed as follows. (1) Co-variation of indoor particulate matter and NO_2: PM2.5_IN was positively correlated with NO2_IN, and PM10_IN also showed a moderate positive correlation with NO2_IN ($r = 0.43$), indicating that indoor NO_2 tends to increase as indoor particulate concentrations rise. These pollutants may be influenced by similar indoor environmental conditions. (2) Association between outdoor particulate matter and indoor NO_2: PM2.5_OUT and NO2_IN showed a strong positive correlation ($r = 0.61$), and PM10_OUT and NO2_IN were also strongly positively correlated ($r = 0.60$). These relationships were stronger than those between indoor PM2.5 or PM10 and NO2_IN. In a mechanically ventilated building whose doors and windows remain closed, outdoor particles can enter through ventilation and may affect indoor NO_2 directly or indirectly through air-exchange processes. (3) Strong negative relationship between indoor O_3 and NO_2: O3_IN and NO2_IN had a strong negative correlation ($r = -0.74$), the largest absolute correlation among all monitored variables. Thus, indoor NO_2 generally tended to be lower when indoor O_3 was high. Chemically, O_3 and NO_2 may interact: NO_2 can participate in ozone-related reactions, and O_3 may facilitate reactions between NO_2 and other reactive substances. The two species may therefore vary inversely under certain conditions. Based on this analysis, the principal variables associated with indoor NO_2 were PM2.5_IN, PM10_IN, O3_IN, PM2.5_OUT, and PM10_OUT. These variables were used as inputs to the subsequent indoor-air-pollutant prediction models.

4.2. Raw-data processing and dataset partitioning

1) Data standardization: To eliminate the effect of differences in measurement scales on regression-coefficient estimation, the selected raw features were standardized using the Z-score method so that each feature had a mean of 0 and a standard deviation of 1. The formula is:

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

In Equation (3), x is the original feature value, μ is the feature mean, σ is the feature standard deviation, and z is the standardized feature value. The first five standardized rows are shown in Table 2.

DOI: 10.54254/2754-1169/2026.LD38812

Table 2. First five rows of standardized data

PM2.5_in_CLS	PM10_in_CLS	NO2_in_CLS	PM2.5_out_CLS	PM10_out_CLS
-0.865	-0.903	0.564	-1.241	-1.146
5.299	5.063	-0.691	1.733	1.678
-0.865	-0.903	1.111	-1.21	-1.086
0.676	0.537	-0.575	0.323	-0.035
-0.865	-0.903	-1.319	-0.689	-0.395

2) Dataset partitioning: The cleaned data were randomly separated into two subsets. Eighty percent of the samples were used for model development, while the remaining 20% were reserved for testing. The same training and test samples were used for both MLR and ANN so that their prediction results were obtained under the same data conditions.

4.3. Air-quality prediction using multiple linear regression

The MLR model was developed with indoor NO_2 concentration as the response variable. Based on the correlation results presented in Section 4.1, environmental variables showing relatively strong relationships with indoor NO_2 were retained as model predictors. The MLR model was developed by estimating regression coefficients with least squares. Its basic expression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (4)$$

In Equation (4), y represents the estimated indoor NO_2 concentration. The intercept is given by the constant term, while the remaining coefficients correspond to the standardized environmental predictors. The error term accounts for the variation not explained by the regression model and is assumed to have zero mean and constant variance.

4.4. Air-quality prediction using an ANN

To account for possible nonlinear associations between indoor NO_2 and the selected environmental variables, a BP neural network was introduced as an additional prediction model. During training, prediction errors were propagated backward through the network and used to update the model parameters iteratively, allowing the network to learn the relationship between the input variables and indoor NO_2 concentration.The implementation steps were as follows.

1) BP neural-network architecture: The initial model used one input layer, one hidden layer, and one output layer. The numbers of neurons and activation functions were designed as follows. The number of input neurons equaled the number of independent variables significantly correlated with indoor NO_2 in Section 4.1, with one neuron for each feature. The output layer contained one neuron representing predicted indoor NO_2. The hidden layer initially contained five neurons; later experiments varied the number of neurons to identify the best-performing configuration. A sigmoid activation function was initially used in the hidden layer for nonlinear mapping, while a Purelin linear activation function was used in the output layer so predictions could match the continuous range of observed indoor NO_2.

2) Model training and optimization: The training subset was used to fit the BP neural network, with Adam applied to update the network parameters during training. Prediction errors were

DOI: 10.54254/2754-1169/2026.LD38812

propagated backward through the network at each iteration, and parameter values were adjusted accordingly. The training process ended when the stopping condition was met or the specified maximum number of iterations was reached. RMSE was subsequently calculated to evaluate prediction error.

5. Experiments

5.1. Experimental setup

All analyses were carried out in Python under Windows 10 using PyCharm. Data preparation and numerical operations relied on pandas and NumPy, while the modeling procedures were implemented with scikit-learn. Specifically, LinearRegression was applied to construct the MLR model, MLPRegressor was used for the ANN model, and StandardScaler was used to standardize the input variables. Matplotlib and Seaborn were used to produce the prediction plots and Pearson correlation heatmap.

5.2. Evaluation metrics

Model performance was assessed using R^2 and RMSE. R^2 was calculated to indicate how well the predicted indoor NO_2 values followed the observed variation, while RMSE was used to quantify the magnitude of prediction errors. For R^2, a value approaching 1 reflects stronger consistency between the model output and the measured data. $R^2 = 1$ indicates perfect agreement, whereas $R^2 < 0$ means the model performs worse than simply predicting the mean of the target variable. R^2 is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

In Equation (5), the observed and predicted indoor NO_2 concentrations for sample i are denoted by the corresponding terms, while the mean of the observed values is represented by the averaged term. The number of samples included in the calculation is n. RMSE summarizes the typical size of the prediction error, and its value is expressed in the same unit as the measured indoor NO_2 concentration.A smaller RMSE indicates lower error and greater predictive accuracy. Because prediction errors are squared, RMSE is sensitive to large errors and effectively reflects performance on samples with substantial deviations. RMSE is calculated as follows:

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

The variables in Equation (6) are defined as in the R^2 equation.

5.3. Experimental results

5.3.1. Air-quality prediction results using MLR

The MLR model, with indoor NO_2 as the target variable, was first evaluated on the test set. The principal metrics are shown in Table 3.

DOI: 10.54254/2754-1169/2026.LD38812

Table 3. Core performance metrics of the MLR model

Evaluation metric	Value
R^2 (coefficient of determination)	0.5975
RMSE (root mean square error)	4.8597

The final R^2 was 0.5975. Thus, MLR explained approximately 59.75% of indoor NO_2 variation. This result indicates limited predictive performance. Indoor NO_2 likely depends on nonlinear processes beyond the selected linear effects. These processes include changes in mechanical-ventilation efficiency and NO_2-related chemical reactions. MLR cannot fully represent such relationships. The RMSE was 4.8597, indicating substantial differences between predictions and observations. RMSE is sensitive to large errors. Its relatively high value suggests poor accuracy for some samples. This finding reflects the model's difficulty in capturing temporal, seasonal, and ventilation-related nonlinear changes. Overall, MLR showed moderate performance. Its main advantages are simplicity and interpretability. The model identifies linear contributions from important factors, including PM2.5 and PM10. However, it poorly predicted high indoor NO_2 concentrations. This limitation reduces its suitability for applications requiring accurate forecasts.

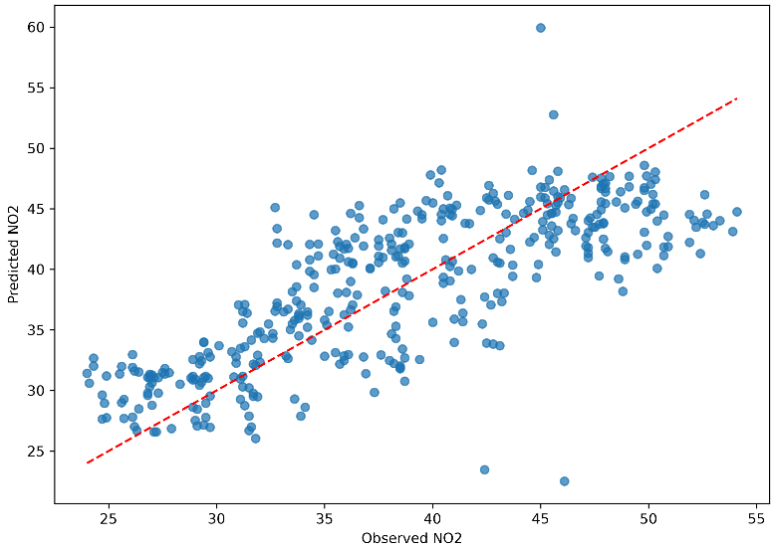


Figure 2. Comparison of observed and predicted values for the MLR model

Figure 2 compares MLR predictions of indoor NO_2 with observed concentrations. The predictions show partial agreement, but the points remain widely dispersed. An R^2 of 0.5975 indicates that MLR explained approximately 59% of the variation. About 41% remained unexplained. Most observations ranged from approximately 25 to 50. At low and moderate concentrations, several points approached the ideal line, $y = x$. This pattern indicates reasonable representation of the overall trend. However, dispersion increased as observed concentrations rose. Predictive stability therefore decreased at higher concentrations. When observed NO_2 exceeded approximately 45, predictions mostly ranged from 40 to 48. Observed values reached approximately 54. Several points consequently fell below the ideal line. This pattern indicates underestimation of high NO_2 concentrations. Combined with an RMSE of 4.8597, the plot confirms substantial absolute errors. A few distant points contributed strongly to RMSE. MLR captured the general trend but performed poorly for highly variable or

DOI: 10.54254/2754-1169/2026.LD38812

high-concentration samples. It remains useful as a baseline model. However, its linear structure limits performance under complex, nonlinear conditions.

5.3.2. Air-quality prediction results using the neural network

For indoor NO_2 prediction, different ANN architectures were first compared on the validation set. Eight configurations were tested, including one- and two-hidden-layer networks using either ReLU or Tanh activation. R^2 and RMSE were used as performance metrics, and the results are shown in Table 4.

Table 4. Performance of ANN hidden-layer configurations

Hidden-layer architecture	Activation function	R^2 (coefficient of determination)	RMSE (root mean square error)
(50,)	ReLU	0.8304	3.2006
(100,)	ReLU	0.8427	3.0830
(150,)	ReLU	0.8349	3.1580
(100,50)	ReLU	0.8576	2.9332
(150,75)	ReLU	0.8676	2.8279
(50,)	Tanh	0.8269	3.2333
(100,)	Tanh	0.8189	3.3075
(100,50)	Tanh	0.8471	3.0389

The results show that hidden-layer architecture and activation function materially affected ANN performance. Across all configurations, R^2 ranged from 0.8189 to 0.8676 and RMSE ranged from 2.8279 to 3.3075. (1) Single-hidden-layer ReLU networks: When the number of hidden neurons increased from 50 to 100, R^2 improved from 0.8304 to 0.8427 and RMSE decreased from 3.2006 to 3.0830, indicating that a moderate increase in hidden neurons improved prediction. At 150 neurons, however, R^2 fell to 0.8349 and RMSE rose to 3.1580. Thus, more neurons did not necessarily improve performance; 100 neurons gave the best result among the single-layer ReLU networks. (2) Improvement with two hidden layers: With ReLU activation, expanding the network from one to two hidden layers further improved performance. The (100, 50) architecture achieved $R^2 = 0.8576$ and RMSE = 2.9332, while the (150, 75) architecture improved R^2 to 0.8676 and reduced RMSE to 2.8279. Relative to the best single-hidden-layer architecture, (100,), the (150, 75) architecture raised R^2 by approximately 0.0249 and reduced RMSE by approximately 0.2551. Under the present data and parameter settings, the two-hidden-layer network had greater fitting capacity and learned the complex relationships between features and indoor NO_2 more fully. (3) ReLU versus Tanh: For comparable architectures, ReLU generally outperformed Tanh. For (50,), ReLU achieved $R^2 = 0.8304$ and RMSE = 3.2006, compared with 0.8269 and 3.2333 for Tanh. For (100,), ReLU achieved $R^2 = 0.8427$ versus 0.8189 for Tanh and reduced RMSE from 3.3075 to 3.0830. For (100, 50), ReLU achieved $R^2 = 0.8576$ versus 0.8471 for Tanh. ReLU was therefore better suited to this indoor NO_2 prediction task. (4) Preferred architecture: Among the eight configurations, the (150, 75) architecture with ReLU achieved the best performance, with the highest R^2 (0.8676) and lowest RMSE (2.8279). It provided the best balance of fit and prediction error among the candidates and was selected for subsequent testing.

DOI: 10.54254/2754-1169/2026.LD38812

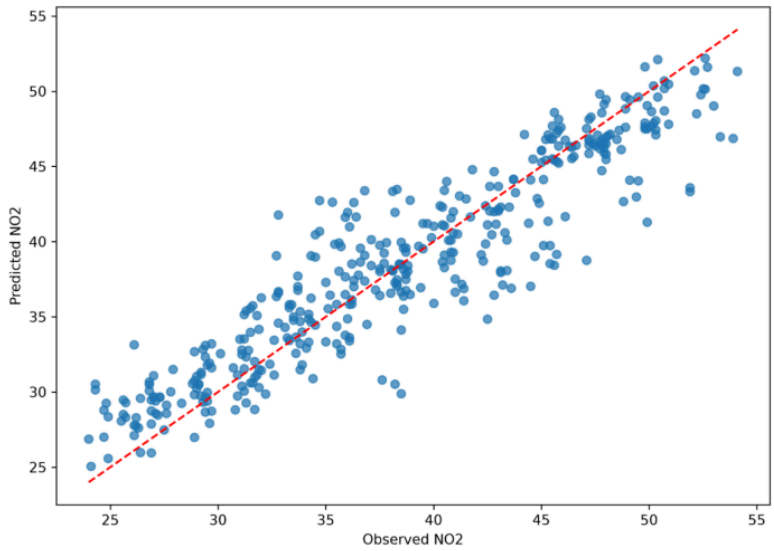


Figure 3. Comparison of observed and predicted values for the best ANN model

Figure 3 shows the test-set results obtained with the selected ANN configuration. The network contained two hidden layers with 150 and 75 neurons and used the ReLU activation function. After this configuration was identified from the validation results, the model was fitted again using the training data and then applied to the test samples. The resulting R^2 and RMSE were 0.8528 and 2.9380, respectively. As shown in the figure, the predicted concentrations are generally distributed around the $y = x$ reference line, although some deviations remain for individual samples.

5.3.3. Comparison of the models

To compare predictive performance for indoor NO_2, the MLR and ANN models were evaluated using the core metrics shown in Table 5.

Table 5. Core performance metrics of the two models

Model	R^2 (coefficient of determination)	RMSE (root mean square error)
MLR (multiple linear regression)	0.5975	4.8597
ANN (artificial neural network)	0.8528	2.9389

ANN clearly outperformed MLR on both metrics. Its R^2 was 0.8528, compared with 0.5975 for MLR, an increase of approximately 25.53 percentage points. This result shows that ANN had greater predictive capacity and used the input features more effectively. The principal reason is that hidden layers and nonlinear activation functions allow ANN to learn complex nonlinear relationships between the inputs and indoor NO_2, whereas MLR models only linear relationships and is therefore limited when relationships are complex. ANN also achieved RMSE = 2.9389, a reduction of approximately 1.9208, or 39.52%, from the MLR value of 4.8597. The overall deviation between predictions and observations was therefore substantially smaller. Consistent with the scatterplots, most ANN predictions were more tightly concentrated around the ideal line, whereas MLR predictions were more dispersed and showed larger errors for some moderate- and high-concentration samples. ANN consequently provided more stable predictions across concentration

DOI: 10.54254/2754-1169/2026.LD38812

ranges. Under the data and model settings used in this study, ANN is the more suitable model for indoor NO_2 prediction.

6. Conclusion

This study investigated indoor NO_2 prediction in a teaching building in a tropical city. Pearson correlation analysis with $|r| \geq 0.3$ as the selection criterion identified the main environmental features associated with indoor NO_2 and provided a basis for model development. MLR and ANN models were then constructed. MLR explained approximately 59.75% of the variation in NO_2 but underestimated some high-concentration observations. Validation-based optimization of the ANN architecture and activation function identified two hidden layers (150, 75) with ReLU activation as the best configuration, with validation $R^2 = 0.8676$ and RMSE = 2.8279. On the final test set, ANN achieved $R^2 = 0.8528$ and RMSE = 2.9389, improving R^2 by 25.53 percentage points and reducing RMSE by 1.9208 (39.52%) relative to MLR. Its predictions were also generally closer to the observations. Overall, ANN modeled nonlinear relationships more effectively than MLR and provided more accurate and stable indoor NO_2 predictions under the conditions of this study.

References

[1] Taha, S. S., Idoudi, S., Alhamdan, N., et al. (2025). Comprehensive review of health impacts of the exposure to nitrogen oxides (NOx), carbon dioxide (CO2), and particulate matter (PM). *Journal of Hazardous Materials Advances, 19*, 100771.

[2] Alqahtani, F. K., Deeb, Y. I., Sherif, M., et al. (2025). Assessment of traffic related air pollution effects on indoor air quality in educational buildings. *Scientific Reports, 15*(1), 41476.

[3] Kadiri, K., Turcotte, D., Gore, R., et al. (2024). Determinants of indoor NO2 and PM2.5 concentration in senior housing with gas stoves. *Toxics, 12*(12), 901.

[4] Mani, G., Viswanadhapalli, J. K., Stonier, A. A. (2022). Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models. *Journal of Engineering Research, 10*(2), 179-194.

[5] Suri, R. S., Jain, A. K., Kapoor, N. R., et al. (2023). Air quality prediction-a study using neural network based approach. *Journal of Soft Computing in Civil Engineering, 7*(1), 93-113.

[6] Chaturvedi, P. (2024). Air quality prediction system using machine learning models. *Water, Air, & Soil Pollution, 235*(9), 578.

[7] Natarajan, S. K., Shanmurthy, P., Arockiam, D., et al. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports, 14*(1), 6795.

[8] El Jihaoui, M., Abra, O. E. K., Mansouri, K. (2025). Factors affecting student academic performance: A combined factor analysis of mixed data and multiple linear regression analysis. *IEEE Access, 13*, 15946–15964.

[9] Mohammadagha, M., Najafi, M., Kaushal, V., et al. (2026). Machine learning models for reinforced concrete pipes condition prediction: The state-of-the-art using artificial neural networks and multiple linear regression in a Wisconsin case study. *Innovative Infrastructure Solutions, 11*(2), 65.

[10] Lu, X., Teh, S. Y., Tay, C. J., et al. (2025). Application of multiple linear regression model and long short-term memory with compartmental model to forecast dengue cases in Selangor, Malaysia based on climate variables. *Infectious Disease Modelling, 10*(1), 240–256.

[11] Castillo-Girones, S., Munera, S., Martinez-Sober, M., et al. (2025). Artificial neural networks in agriculture, the core of artificial intelligence: What, when, and why. *Computers and Electronics in Agriculture, 230*, 109938.

[12] Shobayo, O., & Saatchi, R. (2025). Developments in deep learning artificial neural network techniques for medical image analysis and interpretation. *Diagnostics, 15*(9), 1072.

[13] Guo, Q., He, Z., & Wang, Z. (2025). Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in Liaocheng City. *Scientific Reports, 15*(1), 6798.

[14] Ali, M., Rabehi, A., Souahlia, A., et al. (2025). Enhancing PV power forecasting through feature selection and artificial neural networks: A case study. *Scientific Reports, 15*(1), 22574.

[15] Qureshi, M., Ishaq, K., Daniyal, M., et al. (2025). Forecasting cardiovascular disease mortality using artificial neural networks in Sindh, Pakistan. *BMC Public Health, 25*(1), 34.