Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33736

Analysis and Prediction of Indoor NO_2 Concentration Based on Artificial Neural Network

Hanwen Zhang

School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu, China
42312058@smail.swufe.edu.cn

Abstract. Indoor air pollution poses a serious threat to public health. As a typical pollutant, the accurate prediction of nitrogen dioxide (NO_2) concentration is of great significance for health risk assessment and environmental control. Although the traditional multiple linear regression (MLR) model has a simple structure and strong interpretability, it is difficult to effectively describe the complex nonlinear relationship between indoor and outdoor multi-source environmental variables and pollutant concentration. Based on three months of high-resolution monitoring data (18,144 samples) under subtropical climate conditions, this study first uses the Pearson correlation coefficient method to screen the main independent variables affecting indoor NO_2 concentration (indoor humidity, indoor temperature, outdoor humidity, outdoor NO_2). Then it constructs MLR and BP neural network prediction models respectively, and systematically evaluates the influence of different numbers of hidden layer neurons on prediction performance. The experimental results show that the BP neural network model (R^2=0.7855, RMSE=3.6597 μg/m³) is significantly better than the MLR model (R^2=0.6742, RMSE=4.5102 μg/m³) in prediction accuracy, which verifies the effectiveness and superiority of the artificial neural network method in dealing with nonlinear prediction problems of indoor air quality.

Keywords: Artificial Neural Network, Multiple Linear Regression, Indoor Air Pollution, Prediction

1. Introduction

Air pollution has been listed by the World Health Organization as one of the most severe environmental threats to global public health. People usually pay attention to outdoor air pollution, but studies show that due to the increased air tightness of buildings and people spending more than 80% of their time indoors, indoor pollutant concentrations are often 2 to 5 times higher than those outdoors, posing a serious threat to human health [1]. Among them, nitrogen dioxide (NO_2), as an important indoor air pollutant, mainly comes from combustion processes such as indoor gas stoves and heating equipment, as well as infiltration from outdoor traffic sources [2]. Long-term or high-concentration exposure to NO_2 will significantly increase the risk of respiratory diseases, cardiovascular diseases and even asthma attacks, especially for children, the elderly and people with underlying diseases [3].

© 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: 10.54254/2753-8818/2026.33736

Accurate prediction of indoor NO_2 concentration is crucial for assessing population exposure levels and formulating effective pollution control strategies. At present, studies on indoor NO_2 concentration are mainly divided into two categories: one is mechanism models based on physical and chemical principles, such as the single-zone mass balance model (SMB). Although such models can explain the transport and transformation process of pollutants indoors, their accuracy highly depends on the accurate acquisition of key parameters such as indoor source strength, ventilation rate and deposition rate [4]. The other category is data-driven statistical or machine learning models. Existing studies show that indoor NO_2 concentration is closely related to many factors such as outdoor NO_2 level, ventilation conditions, meteorological conditions (such as temperature and humidity) and indoor activities (such as cooking) [2]. However, there are complex nonlinear relationships among these factors, which limits the prediction accuracy of traditional statistical methods such as multiple linear regression (MLR) when dealing with such nonlinear problems.

In response to the above challenges, this study proposes an indoor NO_2 concentration prediction method based on artificial neural network (ANN). Compared with traditional statistical models, ANN can better capture the complex dynamic relationship between indoor and outdoor environmental parameters and pollutant concentrations by virtue of its strong nonlinear mapping ability and self-learning characteristics [5]. Firstly, this study uses three months of continuous monitoring data under subtropical climate conditions to obtain key parameters such as indoor and outdoor NO_2 concentrations, particulate matter concentrations, temperature and humidity. Then, the Pearson correlation coefficient method is used to analyze the correlation between various factors and indoor NO_2 concentration to screen out key prediction variables. On this basis, prediction models based on multiple linear regression (MLR) and back propagation (BP) neural network are constructed and evaluated respectively, and the performance of the two models is compared.

The main contributions of this paper are as follows: (1) Based on long-term, high-frequency measured data, it systematically analyzes the coupling relationship between indoor NO_2 concentration and multiple variables in mechanically ventilated buildings under subtropical climate; (2) It constructs and optimizes an indoor NO_2 concentration prediction model based on ANN, providing a new technical path for refined prediction of indoor air quality; (3) Through comparative experiments, it quantitatively analyzes the performance advantages of the ANN model compared with the traditional MLR model, and verifies its effectiveness in dealing with complex nonlinear problems. The research results can provide a scientific basis for indoor air quality management, health risk assessment and intelligent regulation of ventilation systems.

2. Related concepts and key technologies

2.1. Pearson Correlation Coefficient

Pearson Correlation Coefficient (PCC) is a statistical index used to measure the degree of linear correlation between two continuous variables [6]. For two zero-mean random variables a and b, its PCC is defined as: $\rho(a,b)=\frac{E(a,b)}{\sigma_a\sigma_b}$, where $E(ab)$ represents the covariance of a and b, and σ_a and σ_b represent the standard deviations of a and b respectively. The value of ρ ranges from -1 to 1. When ρ is closer to 1, the linear correlation between the two variables is stronger. A positive ρ indicates positive correlation, and a negative ρ indicates negative correlation. When $|\rho|$, there is no linear correlation between the two variables. In this study, PCC is used to analyze the correlation between indoor NO_2 concentration and various potential influencing factors (such as outdoor NO_2,

DOI: 10.54254/2753-8818/2026.33736

temperature, humidity, other pollutants, etc.), so as to provide a basis for input feature selection of the subsequent prediction model.

2.2. Multiple Linear Regression

Multiple Linear Regression (MLR) is a classic statistical modeling method used to analyze the linear relationship between a dependent variable and multiple independent variables [7]. Its mathematical model can be expressed as: $y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+\epsilon$, where y is the dependent variable (such as indoor NO_2 concentration), $x_1,x_2,\ldots,x_p$ are independent variables (such as outdoor NO_2, temperature, etc.), β_0 is the intercept term, $\beta_1,\beta_2,\ldots,\beta_p$ are the regression coefficients of each independent variable, representing the change of the dependent variable per unit change of the independent variable when other variables remain unchanged, and ϵ is the random error term. The estimation of regression coefficients usually adopts the least square method, whose goal is to minimize the sum of squared differences between predicted values and actual values of all observations. The MLR model has a simple structure and is easy to understand and interpret, but it assumes a linear relationship between variables, and its prediction ability is often limited when dealing with complex nonlinear problems.

2.3. Artificial Neural Network

Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the human brain, which can learn and simulate complex nonlinear relationships between inputs and outputs. A typical ANN consists of multiple neuron layers, including an input layer, one or more hidden layers, and an output layer. Information is transmitted through connections in the network, and each connection is attached with a weight to adjust the signal strength. During the training process, the network continuously adjusts the weights to minimize the error between predicted values and actual values [5].

The Multilayer Perceptron (MLP) network adopted in this study uses the Backpropagation (BP) algorithm for supervised learning. The core idea of the BP algorithm is "error backpropagation". The process is as follows: first, the input data is propagated forward, and the predicted output is obtained through the calculation of activation functions of each layer; then, the error of the output layer is calculated and propagated back from the output layer to the input layer layer by layer. In this process, the weights of connections in each layer are updated according to the gradient descent principle. This process (each iteration cycle is called an "epoch") is repeated until the overall error of the network drops below a preset threshold. Hidden layer neurons usually use nonlinear activation functions, such as the Sigmoid function, which maps input values to between 0 and 1, thus endowing the network with nonlinear processing ability. It is the stacking of such multi-layer nonlinear structures that enables ANN to approximate any complex function, making it very suitable for solving complex nonlinear modeling problems among indoor environmental parameters.

3. Overview of research area and data sources

This study takes an experimental building in a tropical city as the research object. The building is mainly used for office purposes, with a total construction area of about 35 m^2. It adopts mechanical ventilation all year round, and doors and windows are closed for a long time. The monitoring period is from May 17, 2020 (Sunday) to August 18, 2020 (Friday), lasting three months. The dynamic changes of indoor and outdoor environmental parameters under the background of subtropical

DOI: 10.54254/2753-8818/2026.33736

climate were systematically collected and analyzed. In this study, two sets of Air Quality Egg (2018 version) equipment verified by the Air Quality Sensor Performance Evaluation Center (AQ-SPEC) of the US Environmental Protection Agency were selected and deployed indoors and outdoors respectively for simultaneous monitoring. The instrument automatically collects data at 10-minute intervals to achieve all-weather continuous operation. The indoor sampling point is set about 1 m above the ground and 0.5 m away from the inner wall; the outdoor equipment is installed 1.8 m above the external platform of the building, with wind and rain protection devices to ensure data stability. Monitoring indicators include PM2.5, PM10, NO_2 concentration, air temperature and relative humidity. A total of 18,144 groups of indoor and outdoor paired data recorded at 10-minute intervals were obtained during the study, which systematically reflects the variation characteristics of environmental parameters under subtropical climate from late spring to early autumn.

4. Research methods

4.1. Correlation analysis of air pollutants

To explore the interaction between indoor and outdoor environmental parameters, this study uses the Pearson correlation coefficient method to analyze the correlation degree among variables. First, based on the monitoring data of indoor and outdoor air humidity, temperature, PM2.5, PM10 and NO_2 from May to August 2020, Jupyter Notebook is used to generate a correlation coefficient table (Table 1) and a heat map (Figure 1) to analyze the correlation between indoor NO_2 concentration and other variables.

Table 1. Pearson correlation coefficient table of indoor and outdoor parameters

Variable	Temp_IN	Humi_IN	PM2.5_IN	PM10_IN	NO2_IN	Temp_Out	Humi_Out	PM2.5_Out	PM10_Out	NO2_Out
Temp_IN	1.00	-0.62	0.29	0.29	-0.53	0.50	-0.43	-0.13	-0.14	-0.38
Humi_IN	-0.62	1.00	-0.34	-0.34	0.79	-0.58	0.68	0.11	0.11	0.59
PM2.5_IN	0.29	-0.34	1.00	0.99	-0.20	0.39	-0.41	0.50	0.48	-0.37
PM10_IN	0.29	-0.34	0.99	1.00	-0.20	0.39	-0.41	0.50	0.48	-0.37
NO2_IN	-0.53	0.79	-0.20	-0.20	1.00	-0.21	0.37	0.02	0.04	0.36
Temp_Out	0.50	-0.58	0.39	0.39	-0.21	1.00	-0.90	-0.12	-0.11	-0.71
Humi_Out	-0.43	0.68	-0.41	-0.41	0.37	-0.90	1.00	0.16	0.15	0.82
PM2.5_Out	-0.13	0.11	0.50	0.50	0.02	-0.12	0.16	1.00	0.99	0.14
PM10_Out	-0.14	0.11	0.48	0.48	0.04	-0.11	0.15	0.99	1.00	0.15
NO2_Out	-0.38	0.59	-0.37	-0.37	0.36	-0.71	0.82	0.14	0.15	1.00

It can be seen from Table 1 that there are varying degrees of correlation between indoor and outdoor parameters. Among them, indoor PM2.5 and indoor PM10 show an extremely strong positive correlation (r=0.9990), and outdoor PM2.5 and outdoor PM10 show an extremely strong positive correlation (r=0.9911); outdoor temperature and outdoor humidity show a strong negative correlation (r=-0.8960); indoor humidity and indoor NO_2 show a strong positive correlation (r=0.7901).

DOI: 10.54254/2753-8818/2026.33736

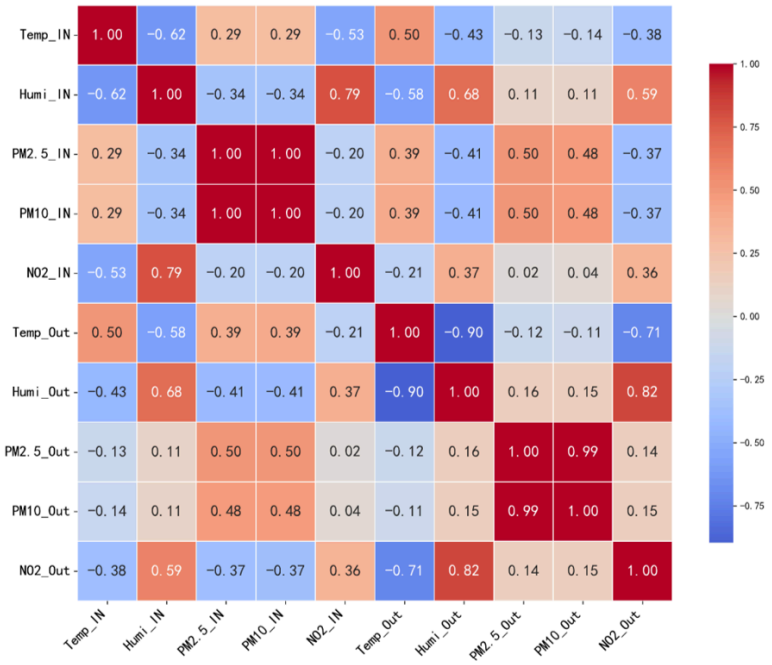


Figure 1. Heat map of indoor and outdoor parameter correlation

Then, focusing on indoor NO_2 (NO2_IN), the target air pollutant, the main independent variables affecting it are screened, and parameters with moderate or higher correlation with NO2_IN are retained as the main influencing factors for subsequent modeling. There are 4 independent variables with strong correlation with indoor NO_2 concentration, sorted by correlation intensity as follows: indoor humidity (Humi_IN, r=0.7901), indoor temperature (Temp_IN, r=-0.5340), outdoor humidity (Humi_Out, r=0.3739), outdoor NO_2 (NO2_Out, r=0.3553). Among them, indoor humidity has a strong positive correlation with indoor NO_2, indicating that higher humidity leads to higher indoor NO_2 concentration; indoor temperature has a moderate negative correlation with indoor NO_2, indicating that higher temperature leads to lower indoor NO_2 concentration. The absolute values of correlation coefficients between other variables (PM2.5_IN, PM10_IN, Temp_Out, PM2.5_Out, PM10_Out) and NO2_IN are all less than 0.3, showing weak correlation and thus being eliminated.

4.2. Air quality prediction based on Multiple Linear Regression (MLR)

To construct a benchmark comparison model, this study adopts the multiple linear regression method to build a prediction model for indoor NO_2 concentration. As a classic statistical method, multiple linear regression achieves quantitative prediction of the target variable by establishing a linear functional relationship between multiple independent variables and the dependent variable. This method has the advantages of simple model structure, fast parameter estimation and strong interpretability of results, and is widely used in various prediction scenarios.

4.2.1. Data standardization

Given the differences in dimensions and numerical magnitudes among different environmental variables (e.g., temperature in degrees Celsius, humidity in percentage, and pollutant concentration in micrograms per cubic meter), to avoid adverse effects of numerical magnitude differences on regression coefficient estimation, the Z-score standardization transformation is first applied to the

DOI: 10.54254/2753-8818/2026.33736

original feature data. For any feature variable x, the standardization formula is: $z=\frac{x-\mu}{\sigma}$ After this transformation, each variable follows a standard normal distribution with a mean of 0 and a standard deviation of 1. This processing makes the regression coefficients of different variables comparable, and the absolute value of the regression coefficient can directly reflect the influence degree of the corresponding independent variable on the dependent variable, facilitating variable importance ranking and model interpretation.

4.2.2. Dataset division

To ensure the fairness of the subsequent comparison with the neural network model, this study divides the preprocessed dataset by random sampling. The random seed is set to 42 to ensure reproducible results, with 80% of the samples as the training set and the remaining 20% as the test set. The training set is used for model parameter estimation, and the test set is used to evaluate the model's generalization ability. The BP neural network model will use exactly the same training and test sets to ensure the two models are compared on the same data basis.

4.2.3. Model construction and parameter solution

According to the correlation analysis results in Section 4.1, there are 4 independent variables strongly correlated with indoor NO_2 concentration. This study takes these 4 variables as the input features of the multiple linear regression model and indoor NO_2 concentration as the prediction target to construct a linear prediction model for indoor NO_2 concentration.

The multiple linear regression model assumes a linear combination relationship between the dependent variable and multiple independent variables, and its basic mathematical form is expressed as: $y=\beta0+\beta1x1+\beta2x2+\beta3x3+\beta4x4+\varepsilon$ where y is the dependent variable, i.e., indoor NO_2 concentration (NO2_IN); x_1,x_2,x_3,x_4 are independent variables, corresponding to indoor humidity (Humi_IN), indoor temperature (Temp_IN), outdoor humidity (Humi_Out), and outdoor NO_2 concentration (NO2_Out) in sequence; β_0 is the intercept term, representing the basic level of the dependent variable when all independent variables are 0; $\beta_1,\beta_2,\beta_3,\beta_4$ are the regression coefficients corresponding to each independent variable, measuring the average change of indoor NO_2 concentration caused by one unit change of the corresponding independent variable while other variables remain unchanged; ε is the random error term, usually assumed to follow a normal distribution with a mean of zero.

This study uses the least squares method to estimate model parameters. The basic principle of the least squares method is to find a set of parameter estimates that minimize the sum of squared residuals between the model predicted values and the actual observed values. The objective function is: $min\sum_{i=1}^{n}(y_i-\widehat{y_i})^2=min\sum_{i=1}^{n}(y_i-(\beta0+\beta1xi1+\beta2xi2+\beta3xi3+\beta4xi4))^2$where n is the number of training samples, y_i is the true observed value of indoor NO_2 concentration for the i-th sample, and ŷ_i is the corresponding predicted value of the model. This optimization problem can be solved analytically by solving normal equations or matrix operations with high computational efficiency. By minimizing the sum of squared residuals on the training set, the optimal estimates of each regression coefficient can be obtained, thus establishing a complete multiple linear regression prediction model.

DOI: 10.54254/2753-8818/2026.33736

4.3. Air quality prediction based on ANN Neural Network

Indoor air quality is interactively affected by various environmental factors, and the relationship between pollutant concentration and environmental parameters often presents complex nonlinear coupling characteristics. It is difficult to fully capture this complex mapping relationship by relying solely on linear models. To break through the limitation of linear models in fitting nonlinear relationships, this study introduces a Back Propagation (BP) neural network to construct an indoor NO_2 concentration prediction model. BP neural network is a multi-layer feedforward network based on the error backpropagation algorithm, which has been theoretically proven to approximate any complex nonlinear function with arbitrary precision.

4.3.1. Data normalization

Different from the Z-score standardization method adopted by the linear regression model, the neural network model is sensitive to the value range of input data. Excessive magnitude differences among input variables may cause the activation function to enter the saturation zone prematurely, reducing the network's training efficiency and learning ability. Therefore, this study uses the Min-Max normalization method to perform linear transformation on each feature variable, mapping the original data to the interval [0, 1]. This normalization method can retain the distribution pattern of the original data while ensuring all input features are at the same magnitude, facilitating stable training and fast convergence of the neural network. The transformation formula of Min-Max normalization is: $x_{norm}=\frac{x-x_{min}}{x_{max}-x_{min}}$

4.3.2. Dataset division

To ensure comparability with the multiple linear regression model, the BP neural network adopts exactly the same training and test set division. That is, using the same random seed (random_state=42), the dataset is randomly divided at a ratio of 80%/20%, ensuring the two models are trained and tested on the same samples. This design enables the subsequent performance comparison to truly reflect the advantages and disadvantages of the models themselves, rather than the impact of data division differences.

4.3.3. Network topology design

This study constructs a three-layer BP neural network, namely one input layer, one hidden layer and one output layer. The specific design of each layer is as follows:

(1) Input layer: According to the correlation screening results in Section 4.1, the number of neurons in the input layer is set to 4, corresponding to the four feature variables: indoor humidity, indoor temperature, outdoor humidity, and outdoor NO_2 concentration. (2) Hidden layer: One hidden layer is set with 5 neurons. The Sigmoid function is selected as the hidden layer activation function, and its mathematical expression is: $f(x)=\frac{1}{1+e^{-x}}$ The nonlinear characteristic of this function enables the network to fit complex mapping relationships. (3) Output layer: The output layer has 1 neuron, corresponding to the predicted value of indoor NO_2 concentration. Since this task is a regression prediction problem, the Purelin linear function is adopted as the output layer activation function, with the expression: $g(x)=x$. (4) Network training: The backpropagation algorithm is used for network training, with the mean squared error as the loss function: $MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ The loss function is converged through iterative optimization to obtain the trained BP neural network

DOI: 10.54254/2753-8818/2026.33736

prediction model. After training, the R^2 and RMSE indicators are calculated on the test set for comparative analysis with the test results of the multiple linear regression model.

5. Experiments

5.1. Experimental setup

The experimental platform was built based on the Windows 10 operating system (64-bit). Python 3.14.3 was selected as the experimental programming language. The core dependent libraries and versions are as follows: pandas 2.0.3 and numpy 1.24.3 were used for data processing and analysis to implement data reading, cleaning, merging, array calculation and other operations. Data standardization, dataset division and correlation analysis relied on the scikit-learn 1.3.0 library. Data visualization was completed with matplotlib 3.7.2 and seaborn 0.12.2 to draw charts such as correlation coefficient heat maps, comparison charts between predicted and actual values. In addition, the pearsonr function in the scipy 1.10.1 library was used to test the significance of correlation coefficients. The experimental hardware environment is: CPU is Intel Core i9-14900HX, RAM is 32GB, and hard disk is 1TB SSD. The random seed was fixed at 42 during the experiment to ensure the repeatability and reproducibility of experimental results.

5.2. Evaluation metrics

To quantitatively evaluate the prediction performance of each model, the coefficient of determination (R^2) and root mean square error (RMSE) were selected as evaluation indicators in this study. These two indicators measure the prediction effect of the model from different perspectives, where R^2 reflects the degree of interpretation of data variability by the model, and RMSE reflects the absolute deviation between predicted and actual values. The coefficient of determination R^2 measures the goodness of fit of the model to the observed data, representing the proportion of dependent variable variation that can be explained by the model. Its value ranges from 0 to 1. The closer the value is to 1, the stronger the model's ability to interpret data and the better the prediction effect; conversely, the closer the value is to 0, the weaker the model's interpretation ability. The formula for R^2 is: $R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$. The root mean square error RMSE measures the average deviation between predicted and actual values, reflecting the prediction accuracy of the model. RMSE is sensitive to large errors and can effectively distinguish the prediction performance differences of different models. The smaller the value, the lower the model prediction error and the higher the prediction accuracy. The formula for RMSE is: $RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$.

5.3. Experimental results

5.3.1. Air quality prediction results based on Multiple Linear Regression (MLR)

To establish a benchmark prediction model, multiple linear regression was used to predict indoor NO_2 concentration in this study. The four variables screened in Section 4.1 (indoor humidity, indoor temperature, outdoor humidity, outdoor NO_2 concentration) were taken as input features. The training set and test set were randomly divided at a ratio of 80%/20% (random seed set to 42). The

DOI: 10.54254/2753-8818/2026.33736

coefficient of determination R^2 and root mean square error RMSE were calculated on the test set, and the results are shown in Table 2.

Table 2. Prediction performance of Multiple Linear Regression (MLR) model (test set)

Evaluation Metric	Value
R^2	0.6742
RMSE	4.5102 μg/m³

It can be seen from Table 2 that the R^2 of the MLR model on the test set is 0.6742, indicating that the selected four independent variables can explain about 67.42% of the variation in indoor NO_2 concentration, and the model has certain prediction ability. The RMSE is 4.5102 μg/m³, reflecting that the average deviation between the predicted and actual values is within an acceptable range.

This study plotted the comparison curve (Figure 2) and scatter plot (Figure 3) of the predicted and actual values of the first 200 samples in the test set. It can be seen from Figure 2 that the prediction curve of the MLR model is basically consistent with the overall trend of the actual value curve, and can well follow the fluctuation direction of indoor NO_2 concentration. However, in some concentration peak areas, the predicted values are significantly lower than the actual values, showing an underestimation phenomenon; in the stable stage, the prediction curve is relatively smooth and fails to fully reproduce the local fluctuation details of the actual values. The scatter plot in Figure 3 shows that most sample points are distributed near the ideal fitting line, but the predicted values are systematically low in the high-concentration area (actual value greater than 70 μg/m³). Overall, the MLR model can grasp the overall trend, but its ability to characterize extreme values and subtle fluctuations is still insufficient.

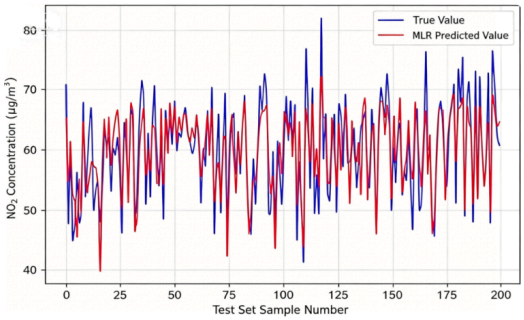


Figure 2. Comparison curve of predicted and actual values of MLR model

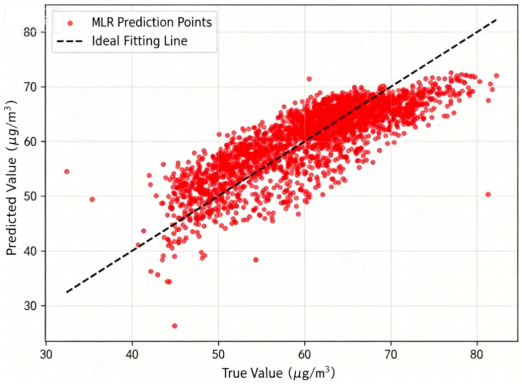


Figure 3. Scatter plot of predicted and actual values of MLR model

DOI: 10.54254/2753-8818/2026.33736

5.3.2. Air quality prediction results based on BP Neural Network

To explore the influence of network structure on prediction accuracy, this study set the number of hidden layer neurons from 2 to 8 for traversal experiments on the premise of fixing a single hidden layer (maximum number of iterations: 3300). The R^2 and RMSE of each model were calculated on the test set, and the results are shown in Table 3.

Table 3. Prediction performance of ANN model with different numbers of hidden layer neurons

Number of Hidden Layer Neurons	R^2	RMSE
2	-0.0006	7.9043
3	0.7463	3.9798
4	0.7801	3.7059
5	0.7538	3.921
6	0.7832	3.6794
7	0.7849	3.6647
8	0.7855	3.6597

It can be seen from Table 3 that when the number of neurons is 2, the model can hardly learn effectively ($R^2 \approx 0$). When the number of neurons increases to 3, the model performance improves sharply, with R^2 reaching 0.7463. As the number of neurons increases from 3 to 8, the overall prediction ability of the model remains at a high level, with R^2 stable between 0.75 and 0.79, and RMSE stable between 3.66 and 3.98 μg/m³. Among them, the model achieves the best performance when the number of neurons is 8, with R^2 of 0.7855 and RMSE of 3.6597 μg/m³. This indicates that appropriately increasing the number of hidden layer neurons can improve the fitting ability of the network, but the performance improvement tends to be gentle after reaching a certain complexity.

In this study, the BP neural network with 8 hidden layer neurons was selected as the optimal prediction model. The comparison curve between its predicted and actual values is shown in Figure 4, and the scatter plot is shown in Figure 5. It can be seen from Figure 4 that the prediction curve of the optimal ANN model is in good overall agreement with the actual value curve. Whether it is the subtle changes in the stable stage or the sharp fluctuations in the peak area, the predicted values can well follow the changing trend of the actual values. Compared with the MLR model (Figure 2), the prediction accuracy of the ANN model in the peak area is significantly improved. The scatter plot in Figure 5 shows that the predicted value scatter points are closely distributed on both sides of the ideal fitting line. Compared with the MLR model (Figure 3), the number of deviated sample points is significantly reduced, especially the prediction deviation in the high-concentration area is significantly reduced, verifying that the ANN model has better prediction stability and generalization ability.

DOI: 10.54254/2753-8818/2026.33736

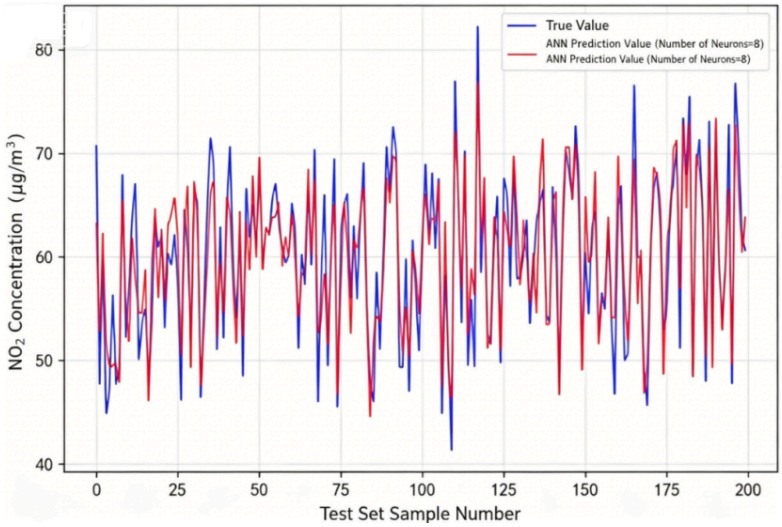


Figure 4. Comparison curve of predicted and actual values of ANN model

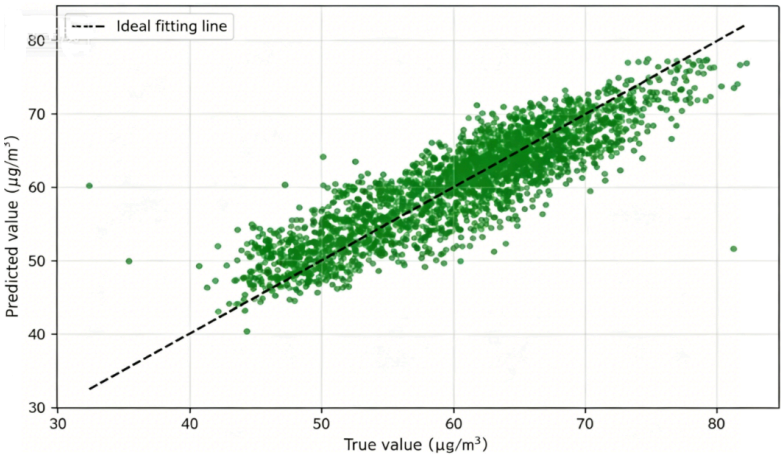


Figure 5. Scatter plot of predicted and actual values of ANN model

5.3.3. Model comparison results

To comprehensively evaluate the prediction performance of the two models for indoor NO_2 concentration, this study compared the coefficient of determination R^2 and root mean square error RMSE of multiple linear regression (MLR) and the optimal BP neural network (ANN, number of hidden layer neurons = 8) on the test set. The results are shown in Table 4.

Table 4. Comparison of prediction performance between MLR and ANN models

Model	R^2	RMSE
Multiple Linear Regression	0.6742	4.5102
BP Neural Network	0.7855	3.6597

It can be seen from Table 4 that the BP neural network model is superior to the multiple linear regression model in both evaluation indicators. Specifically, the R^2 of the ANN model is 0.7855, which is 0.1113 higher than the 0.6742 of the MLR model, with an increase of about 16.51%; the RMSE of the ANN model is 3.6597 μg/m³, which is 0.8505 μg/m³ lower than the 4.5102 μg/m³ of

DOI: 10.54254/2753-8818/2026.33736

the MLR model, with a decrease of about 18.86%. This result shows that the BP neural network can more effectively capture the complex mapping relationship between indoor and outdoor environmental variables and indoor NO_2 concentration. Among them, the improvement of the ANN model's ability to predict peak concentration is particularly significant, indicating that the neural network has obvious advantages in dealing with extreme value problems with more prominent nonlinear characteristics. In contrast, the MLR model is limited by the linear assumption. Although it can explain about 67% of the concentration variation, its ability to characterize nonlinear relationships is limited. In summary, the BP neural network model constructed in this study performs better than the multiple linear regression model in the indoor NO_2 concentration prediction task, verifying the effectiveness and superiority of the artificial neural network method in dealing with nonlinear prediction problems of indoor air quality.

6. Conclusion

Based on three months of high-resolution monitoring data (18,144 samples) from an office building under subtropical climate conditions, this study focused on the prediction of indoor NO_2 concentration. The Pearson correlation coefficient method was used to screen key influencing factors, and multiple linear regression (MLR) and BP neural network (ANN) prediction models were constructed respectively, with their prediction performance systematically compared. The main conclusions are as follows: Indoor NO_2 concentration has significant correlations with indoor humidity ($r=0.7901$), indoor temperature ($r=-0.5340$), outdoor humidity ($r=0.3739$), and outdoor NO_2 concentration ($r=0.3553$). Among them, indoor humidity shows a strong positive correlation, indoor temperature shows a moderate negative correlation, and outdoor factors show weak positive correlations. Particulate matter (PM2.5, PM10) has a weak correlation with indoor NO_2 ($|r|<0.3$), indicating differences in the sources and diffusion patterns of NO_2 and particulate matter. The prediction performance of the BP neural network model is significantly better than that of the multiple linear regression model. The ANN model achieves an R^2 of 0.7855 and an RMSE of 3.6597 $\mu g/m^3$, while the MLR model yields an R^2 of 0.6742 and an RMSE of 4.5102 $\mu g/m^3$. Compared with MLR, the ANN model increases R^2 by 16.51% and reduces RMSE by 18.86%. The number of hidden layer neurons significantly affects the prediction performance of the ANN model. When the number of neurons increases from 2 to 3, the model performance improves sharply (R^2 rises from -0.0006 to 0.7463). The model achieves the best performance when the number of neurons is 8 ($R^2=0.7855$, RMSE=3.6597 $\mu g/m^3$). Appropriately increasing the number of neurons can enhance the network's fitting ability, but the performance improvement tends to slow down after a certain complexity is reached. From the visual analysis of prediction effects, although the MLR model can capture the overall trend of indoor NO_2 concentration changes, it shows systematic underestimation in peak regions. The ANN model can better follow concentration fluctuations, especially with outstanding advantages in extreme value prediction, verifying the effectiveness of neural network methods in handling nonlinear problems. In summary, the BP neural network model constructed in this study can effectively capture the nonlinear mapping relationship between indoor and outdoor environmental parameters and indoor NO_2 concentration. It exhibits good prediction accuracy and generalization ability under the subtropical climate background, and can provide methodological support for intelligent building environment regulation and health early warning.

DOI: 10.54254/2753-8818/2026.33736

References

[1] González-Martín J, Kraakman N J R, Pérez C, et al. A state-of-the-art review on indoor air pollution and strategies for indoor air pollution control [J]. *Chemosphere, 2021, 262:* 128376.
[2] Zota A, Adamkiewicz G, Levy J I, et al. Ventilation in public housing: implications for indoor nitrogen dioxide concentrations [J]. *Indoor Air, 2005, 15*(6): 393-401.
[3] Koenig J Q. Health effects of indoor air pollution [M]//Indoor Air Quality Handbook. New York: McGraw-Hill, 2000.
[4] Hu Y, Zhao B. Relationship between indoor and outdoor NO_2: A review [J]. *Building and Environment, 2020, 180:* 106909.
[5] Kulkarni H, Bharath N, Siddesh C S, et al. An introduction to artificial neural network [J]. *International Journal of Advanced Research in Innovation and Ideas in Education, 2016, 1(5*): 27-30.
[6] Benesty J, Chen J, Huang Y, et al. Pearson correlation coefficient [M]//Noise Reduction in Speech Processing. *Berlin: Springer, 2009:* 1-4.
[7] Çelikten H. Evaluating machine learning (RF, XGBoost) and statistical model (MLR) for PM_{10} and air quality prediction: A case from Kars, Türkiye [J]. *Atmospheric Pollution Research, 2026:* 102975.