Proceedings of CONF-APMM 2025 Symposium: Multi-Qubit Quantum Communication for Image Transmission over Error Prone Channels
DOI: 10.54254/2753-8818/2025.GL25429

Pollutant Significance Assessment and AQI Forecasting Framework: Based on Linear Regression and xLSTM

Yiwei Wang

School of Economics and Finance, Shanghai International Studies University, Shanghai, China
0221136003@shisu.edu.cn

Abstract. Air pollution has become a global environmental problem worth considering, so accurate prediction and targeted plans for air pollution are growing increasingly important. This study aims to predict the Air Quality Index (AQI) using extended Long Short-Term Memory (xLSTM) on an air quality dataset spanning 5 years, leveraging its enhanced sequence modeling capabilities., in order to provide a scientific basis for environmental policy formulation and public health intervention. Also, linear regression was used to study which specific air pollutants have the greatest impact on AQI. This enables the environment-related institutions to develop targeted plans to control air pollution. The result of the study suggests that CO and PM2.5 are the pollutants worth most concern. Moreover, the accuracy of xLSTM in predicting AQI values that belong to the same category as the actual AQI values is 79.29%. Collectively, this integrated predictive approach provides a valuable scientific foundation for evidence-based environmental policy formulation and public health intervention.

Keywords: AQI Prediction, Linear Regression, xLSTM, Air pollution

1. Introduction

With the rapid development of industrialization and urbanization, air pollution has become a global environmental problem. In recent years, air pollution has caused a serious impact on human health and the ecosystem, which is worth considering. Air pollutants, including particulate matter (PM2.5 and PM10), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3), have not only aggravated the air quality but have also increased the incidence rate of respiratory diseases and cardiovascular diseases. Hence, accurate prediction of air pollution levels plays a pivotal role in developing effective environmental policies and public health intervention measures. Air Quality Index (AQI) is an effective index to showcase the air pollution level and reflect the air quality.

The importance of predicting air pollution lies in its profound influence in multiple fields, to name a few, public health, environmental policy formulation, establishment of early warning systems, scientific research, city planning, climate change response, and international collaboration. First of all, air pollution is one of the main health risks of urban life. Accurate prediction of air pollution enables governments to release health warnings to the public timely manner and reduce the outdoor activity of citizens, thus relieving the negative influence of severe air pollution. Besides, 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: 10.54254/2753-8818/2025.GL25429

prediction can provide the policy makers with a basis for evaluating the current environmental policies, and make more effective air pollution control methods. What's more, the city planning based on air quality prediction helps optimize resource distribution, such as adjusting traffic flow, optimizing natural resource usage, and controlling industrial emissions.

Numbers of researches have been done about AQI, focusing on its variations, forecasting methodologies, and the influence of natural and anthropogenic factors. Early investigations, such as Mohan, analyzed the annual and seasonal trends of AQI in Delhi over a nine-year period, highlighting significant seasonal shifts in pollution levels [1]. This underscores the importance of seasonal dynamics in AQI assessment.

Forecasting models have also been a prominent focus. Kumar employed principal component regression to predict short-term daily AQI across different seasons, emphasizing the role of meteorological variables in enhancing forecast accuracy [2]. Similarly, Wang developed a hybrid model integrating two-phase decomposition techniques with an optimized extreme learning machine (ELM) to improve AQI prediction precision, demonstrating the utility of advanced decomposition and machine learning methods in capturing complex AQI series patterns [3].

In this study, the micro meteorological factors influencing AQI are analyzed by Linear regression. After that, an AQI prediction based on an xLSTM model is discussed and analyzed.

2. Methods

2.1. Linear regression

Linear regression is indeed an excellent method for measuring the relationship between multiple independent variables and a single dependent variable. When building a multiple linear regression, the first thing to do is to ensure that the independent variables do not have strong multicollinearity. When the multicollinearity is high, it means that the independent variables are highly linearly dependent, and the regression coefficients in the model may become very unstable. A slight disturbance of data will cause significant changes in the estimated values of regression coefficients. Also, strong multicollinearity will increase the error of the regression coefficient. In this study, the variance inflation factor (VIF) of the covariance matrix of the independent variables is calculated to determine the multicollinearity of the independent variables. VIF of and the multicollinearity of the independent variables are positively correlated.

2.2. XLSTM

xLSTM (extended long short-term memory) is a newly developed, improved structure of long short-term memory networks. It performs well in long sequence tasks. xLSTM can effectively understand the relationship between time points and long-term relationships between time periods, thus making a precise prediction of AQI with historical AQI data. In xLSTM, exponential gates are introduced. Different from the sigmoid function, which traditional LSTM uses, exponential gates enable the input gates and the forget gate to control memory updates exponentially. Also, xLSTM eliminates the hidden state connections between timesteps, which enables parallelization and improves computational efficiency.

DOI: 10.54254/2753-8818/2025.GL25429

2.3. Model evaluation indicators

In this study, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 are used to weigh the predictive ability of the model. Also, since AQI data can be separated into 5 groups, the classification accuracy is also considered.

3. Results

3.1. Data

The dataset comes from Kaggle (Kaggle: Your Home for Data Science) [4]. The dataset includes daily AQI data and air pollutants in Delhi. The air pollutants are shown in Table 1. The data unit for air pollutants is μ g/m³. The dataset includes daily data from 2015/1/1 to 2020/7/1, a total of 2009 days. The AQI data is separated into five groups; the corresponding relationship of AQI and the separated group is shown in Table 2. Figure 1 shows the fluctuation of the AQI dataset. The horizontal axis represents the sequence number of the data, with a total of over 2000 data points. The vertical axis represents the AQI value. The fluctuation of AQI is quite significant.

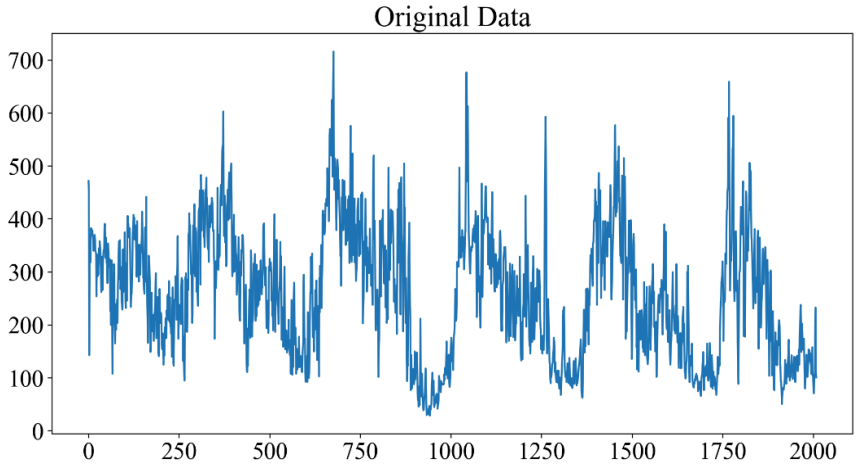

Figure 1. The variation of original data over time (photo/picture credit: original)

DOI: 10.54254/2753-8818/2025.GL25429

Table 1. Main air pollutants in Delhi

Variable Name	Description
PM2.5	Fine inhalable particles ≤2.5 micrometers in diameter
PM10	Inhalable particles ≤10 micrometers
NO	Nitrogen monoxide
NO_2	Nitrogen dioxide
NOx	Nitrogen oxides (NO + NO_2)
NH_3	Ammonia
CO	Carbon monoxide
Variable Name	Description
SO_2	Sulfur dioxide
O_3	Ground-level ozone
Benzene	Volatile organic compound (VOC)
Toluene	VOC – common solvent
Xylene	VOC – industrial solvent

Table 2. AQI classification thresholds

AQI Range	Health Impact Category
301 and above	Hazardous
201 - 300	Very unhealthy
151 - 200	Unhealthy
101 - 150	Unhealthy for sensitive groups
0 - 100	Moderate

3.2. Data preprocessing

First, delete the null values in the data. Then, the independent variables with high multicollinearity. In the study, multicollinearity between independent variables was determined by calculating VIF. The formula for calculating VIF is as follows:

$$VIF_j = \left(1 - R^2\right)^{-1} \tag{1}$$

In the formula, R^2 refers to the coefficient of determination. In the absence of multicollinearity, the variance amplification factor approaches 1. However, in reality, there is always some degree of multicollinearity between independent variables, so it is unrealistic to use the variance expansion factor equal to 1 as the criterion for evaluating multicollinearity. In this study, when VIF is less than or equal to 10, the variables are considered as having weak multicollinearity.

3.3. Result of linear regression

After calculation, the following variables are retained: PM2.5, NOx, NH3, CO, O3, Toluene, Xylene. The result of the linear regression is shown in Table 3.

DOI: 10.54254/2753-8818/2025.GL25429

Table 3. The result of the linear regression

Variables	Coef.	p
Constant	73.980***	<0.001
PM2.5	1.186***	<0.001
NOx	0.425***	<0.001
NH3	-0.360***	<0.001
CO	6.902***	<0.001
O3	0.542***	<0.001
Toluene	-0.361***	<0.001
Xylene	-0.210	0.729

According to the result, PM2.5, NOx, CO, and O3 showed a statistically significant positive effect on AQI, while NH3 and Toluene had a statistically significant negative effect on AQI. Xylene demonstrated no significant association with AQI, implying it may not be a relevant predictor in this model. The ecoefficiency of CO and PM2.5 is very significant, implying that the local government of Delhi should take controlling coal and motor vehicle emissions into consideration.

3.4. XLSTM prediction

The model parameters are configured as follows: a sequence length of 30, meaning 30 days of data are used to predict a single day. The model architecture consists of 2 sLSTM layers and 1 mLSTM layer, with a hidden size of 16. This medium-complexity configuration is well-suited for AQI prediction. The first 70% of the dataset is used for training, and the rest 30% is for testing. The training loss is shown in Figure 2. The blue line represents train loss while the orange line represents validation loss. The horizontal axis shows the number of epochs, and the vertical axis is the value of train loss and validation loss. After 10 epochs and before 30 epochs, the model gradually converges. After 30 epochs, the model tends towards stability.

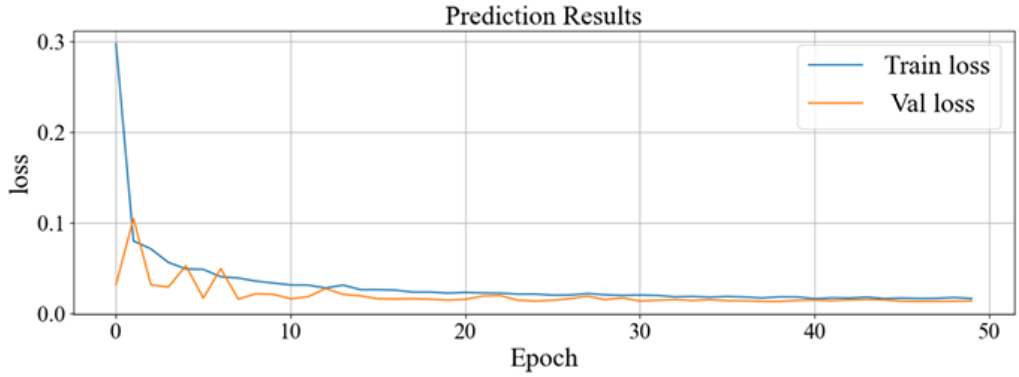

Figure 2. Training loss (photo/picture credit: original)

The evaluation metrics are shown in Table 4. The relatively high R^2 indicates that the model performs well in explaining the variation of the data and predicting the data. The 25.86MAE means that the average difference between the AQI values predicted by the model and the true AQI values is about 25.86 index points. The error might cause the prediction of AQI to be wrongly classified.

DOI: 10.54254/2753-8818/2025.GL25429

RMSE is much higher than MAE, indicating that the model does not perform well in explaining extreme values of AQI.

Table 4. Evaluation metrics

Evaluation Metric	Test Set Performance
R^2	0.727
MAE	25.86
RMSE	42.08

In order to further analyze the performance of the model in prediction, the predicted data is classified into 5 groups, and compared to the actual class of AQI data, thus calculating the classification accuracy. The classification confusion matrix is showcased in Figure 3. According to the confusion matrix, most wrongly classified data falls into the categories adjacent to it. This indicates that the reason for the model's error is that the prediction error is relatively large, rather than the model tending to classify most of the data into a certain category. The classification accuracy is 79.29%.

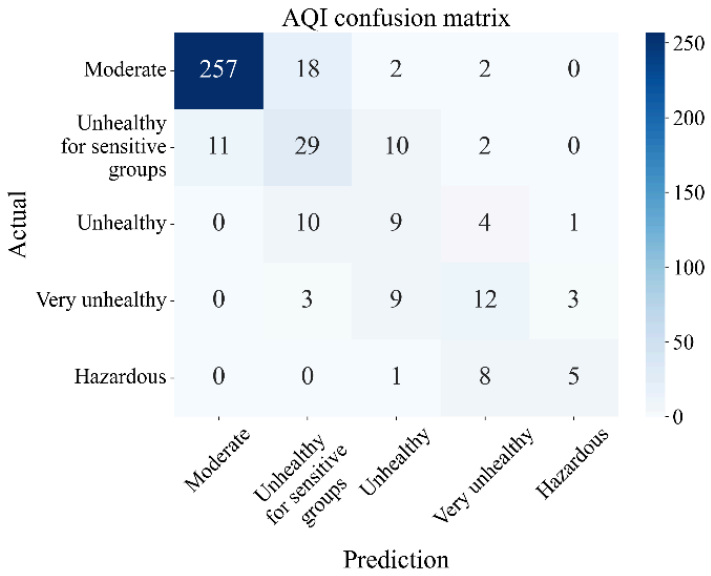

Figure 3. Classification confusion matrix (photo/picture credit: original)

4. Discussion

Many studies have proved the prediction accuracy of LSTM-based models. Siami-Namini's research demonstrates that the average reduction in error rates obtained by LSTM was between 84 - 87 percent when compared to ARIMA in time series prediction [5]. However, the time cost of training the LSTM-based model is much higher than traditional models for time series prediction.

A limitation of this study is its exclusive reliance on data from Delhi. This contrasts with approaches like Liu, who leveraged integrated data from Beijing, Tianjin, and Shijiazhuang to exemplify the enhanced predictive power achievable through multi-source integration [6]. The findings of this study might be less generalizable and could potentially benefit from future incorporation of diverse data sources.

DOI: 10.54254/2753-8818/2025.GL25429

Also, time series prediction is a method that only uses historical data to predict future data. It ignores the geography and climate factors that truly affect AQI. Study of Jumaah uses GIS-based methods to predict AQI in Kuala Lumpur [7]. Also, Guo's research established a mathematical model that considers various influencing factors, including inversion layer at night and the inter-effect of adjacent areas [8]. The omission of these factors, demonstrated as critical in the literature, likely constrains the model's explanatory depth and predictive robustness across diverse scenarios.

5. Conclusion

The study uses linear regression to analyze the AQI data of Delhi from 2015/1/1 to 2020/7/1, and finds that CO and PM2.5 are the pollutants that influence the AQI the most. Then, the xLSTM prediction model predicted the future AQI with 30 history data points. The accuracy of the prediction is 79.29%. The core findings of the study provide a practical way to predict the air quality level with historical data, expanding the application and practice of machine learning in the field of AQI prediction. Although xLSTM has demonstrated good prediction performance and accuracy in prediction tasks, since the prediction method only uses history data to predict future data, it cannot consider all factors that may affect air quality, such as meteorological conditions, industrial activities, etc. Unexpected fluctuation in meteorological conditions or sudden change in industrial policies might undermine the prediction performance of the model. For more practical air quality prediction, it may be necessary to combine other types of data to improve the accuracy of the prediction.

To conclude, the study provides a useful tool for air quality prediction, but there is still room for improvement. Future research can explore how to integrate more environmental and meteorological data to enhance the predictive ability of models. Through these efforts, it can better understand and respond to changes in air quality, providing stronger support for environmental protection, public health, and policy-making by relevant departments.

References

[1] Mohan, M., & Kandya, A. (2007). An analysis of the annual and seasonal trends of air quality index of Delhi. Environmental Monitoring and Assessment.

[2] Kumar, A., & Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research.

[3] Wang, D., Wei, S., Luo, H., Yue, C., & Grunder, O. (2016). A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. The Science of the Total Environment.

[4] Kaggle. (2020, January 1). Clean air? India's air quality □□. https: //www.kaggle.com

[5] Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1394–1401).

[6] Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K., & Tsao, C.-C. (2017). Urban air quality forecasting based on multi-dimensional collaborative support vector regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. PLOS ONE.

[7] Jumaah, H. J., Ameen, M. H., Kalantar, B., Rizeei, H. M., & Jumaah, S. J. (2019). Air quality index prediction using IDW geostatistical technique and OLS-based GIS technique in Kuala Lumpur, Malaysia. Geomatics, Natural Hazards and Risk.

[8] Guo, Z. (2016). The research of AQI index changing regularity mainly in Tianjin. Unpublished manuscript or thesis.