

# Deep Learning Method for Urban Air Pollution Prediction: Empirical Study Based on PM<sub>2.5</sub>

Yubo Lin<sup>1\*</sup>

<sup>1</sup>School of Computer Science, Taylor's University, 47500 Subang Jaya, Malaysia

**Abstract.** A systematic comparison was conducted between traditional ARIMA models and deep learning Long Short-Term Memory (LSTM) models in predicting PM<sub>2.5</sub> in the Shanghai air quality dataset from 2014 to 2025. The first step of the research is to perform missing value imputation, outlier detection, and standardization on the original dataset. Afterwards, analyze the temporal variation characteristics and correlation of the main pollutants. Then, create and train Autoregressive Integrated Moving Average Model (ARIMA) and LSTM models and perform model assessment employing MSE, MAE, and R<sup>2</sup>. The paper will compare ARIMA with LSTM for prediction results. The results showed that LSTM reduced MSE by nearly 56% and MAE by about 31%. The paper can say that LSTM is more versatile than ARIMA in capturing nonlinear features and handling fast fluctuations. This study provides a methodology for predicting urban air pollution in meteorology and lays the foundation for building more complex prediction systems.

## 1 Introduction

Not long ago, air pollution has become a hot topic of global discussion. Among numerous air pollutants, because the PM<sub>2.5</sub> particles are very small, it can penetrate deep into the human lungs and blood vessels, posing a great threat to health. Those considered to be the most important public health risk factors in the world [1]. By developing high-precision PM<sub>2.5</sub> concentration prediction models, the government can take effective measures in a timely manner and provide scientific health protection for the people in a metropolis like Shanghai, with the deepening of urbanization and the improvement of transportation and work standards, monitoring and predicting PM<sub>2.5</sub> concentration is crucial for improving the environment and developing related policies [2].

In previous air quality prediction tasks, the use of autoregressive moving averages integrated into models and other aspects was commonly used as a benchmark method for related models. Obviously, due to the fact that this method deals with linear time series data and is not easy to capture non-linear dependencies, and is difficult to characterize over a long period of time, it is necessary to modify the manual selection of such attributes as mentioned earlier. In the process of technological development, a method called machine learning has achieved better air quality prediction results, such as support vector machine, random forest,

---

\* Corresponding author's email: 0370305@sd.taylors.edu.my

gradient boosting and other technologies [3,4]. However, the methods are grounded in feature selection for training, and they usually are limited with respect to temporal dependencies.

With the development of big data and computing power, deep learning has emerged and become an important tool for time series prediction. Recurrent neural networks, especially long short-term memory networks and gated recurrent units, perform outstandingly in sequence modeling and are particularly suitable for air pollution prediction [5]. In addition, Convolutional Neural Networks are used to extract local features of time series in order to better improve prediction accuracy and computational efficiency [6]. Hybrid models and ensemble learning have also become research hotspots in recent year. Most literature proposes the fusion of ARIMA and LSTM, combining them to limit the linear and nonlinear models and improve their prediction accuracy [7]. In recent years, cutting-edge deep learning methods like CNN and Transformer have been introduced into environmental time series prediction to make the model more interpretable and predictive [8].

However, existing research often still relies on air quality data from a single city or one of its predictive models, and systematic comparisons between traditional statistical models and deep learning models have not been fully validated [9]. Based on this, this study will select ARIMA and LSTM as two representative methods based on more than ten years of air pollution data in Shanghai. A comprehensive comparison will be made from the aspects of prediction accuracy, stability, and response to emergencies to explore the potential and possible limitations of deep learning methods in actual air quality prediction.

## **2 Method**

### **2.1 Data source and explanation**

This study utilized data from the Shanghai Environmental Monitoring Center, covering the period from 2014 to October 13, 2025. This data is recorded on a daily basis and includes six major air pollutants in this dataset: PM<sub>2.5</sub>, PM<sub>10</sub>, O<sub>3</sub>, NO<sub>2</sub>, SO<sub>2</sub> and CO. All data are sourced from public records of official monitoring stations. Some missing values are filled in through linear interpolation, and outliers are smoothed using the interquartile range (IQR) method to reduce interference from extreme values.

### **2.2 Indicator selection and explanation**

The core prediction objective of this study is PM<sub>2.5</sub> concentration, which is one of the main components of the Air Quality Index, has a significant impact on public health. In order to improve the accuracy and stability of predictions, the study selected five items related to PM<sub>2.5</sub>. The closely related air pollution indicators as input features are PM<sub>10</sub>, O<sub>3</sub>, NO<sub>2</sub>, SO<sub>2</sub> and CO [10].

### **2.3 Methodology Introduction**

In order to compare the performance of traditional statistical models and deep learning models in air pollution prediction, this study selected ARIMA and LSTM as the main methods.

### 2.3.1 ARIMA model

ARIMA is a commonly used time series forecasting method that combines autoregression, differencing, and smoothing moving average modeling. In this study, PM<sub>2.5</sub> concentration sequence is first subjected to stationarity test, and if there is a trend or seasonal variation, it is stabilized through differential processing. Next, use autocorrelation function and partial autocorrelation function to determine the model order (p, d, q).

### 2.3.2 LSTM neural network model

LSTMs are a type of RNN designed for handling time series data. They can remember things for a long time and have input gates, output gates, and forget gates, and because of this, they no longer have the gradient vanishing problem that the traditional RNNs have.

In this study, The LSTM model takes several hours of pollutant data as input. Used to predict the PM<sub>2.5</sub> concentration at the next moment. The network structure includes two layers of LSTM hidden layers, each layer has 64 neurons and is connected to a fully connected output layer. Adam is the optimizer used, the loss function used is MSE, the learning rate set is at 0.001, and the activation function used is tanh.

### 2.3.3 Model evaluation

Both models are modeled and predicted on the same training, validation, and testing sets to ensure comparability. By calculating the MSE, MAE, and R<sup>2</sup> indicators, a comprehensive evaluation is conducted on the prediction accuracy and stability of the model. The experimental results are used to analyze the differences between ARIMA and LSTM in processing linear and nonlinear features, to compare the advantages and limitations of traditional methods and deep learning methods in air quality prediction.

## 3 Results & Discussion

### 3.1 Data preprocessing

In data preprocessing, all indicators are normalized using the Min-Max normalization method:

$$X' = \frac{X - X_{min}}{x_{max} - x_{min}} \quad (1)$$

Scale the values to the range of [0,1] to avoid the influence of different dimensions on model training. The performance evaluation indicators for prediction include mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R<sup>2</sup>), which are used to measure the prediction accuracy and stability of the model. Through statistics, it was found that there is a missing value of approximately 3.2% for O<sub>3</sub>, SO<sub>2</sub> about 1.5%, there are relatively few missing indicators. In order to preserve temporal continuity and fill data gaps, linear interpolation is employed. Using the interquartile range method to detect outliers:

$$IQR = Q3 - Q1 \quad (2)$$

If the data exceeds the range of [Q1 - 1.5IQR, Q3 + 1.5IQR], it is judged as an outlier. After testing, The maximum value of PM<sub>2.5</sub> at 409 µg/m<sup>3</sup> is considered an extreme pollution event, Retain as abnormally high value samples for model robustness testing. The test data is Table 1:

**Table 1.** Statistical Results of Data for Different Air Pollution Indicators

indicator	average	standard deviation	minimum	maximum
PM <sub>2.5</sub>	97.22	39.20	14	409
PM <sub>10</sub>	44.04	20.68	6	254
O <sub>3</sub>	43.75	21.54	1	143
NO <sub>2</sub>	17.49	9.35	1	69
SO <sub>2</sub>	3.98	3.14	1	37
CO	5.84	2.21	1	19

The data analysis results of the table 1 are as follows: PM<sub>2.5</sub> in Shanghai the average concentration is about 97.22  $\mu\text{g}/\text{m}^3$ , higher than the recommended safety limit of 25  $\mu\text{g}/\text{m}^3$  by the World Health Organization, indicating severe long-term air pollution; the standard deviation of PM<sub>2.5</sub> and PM<sub>10</sub> is relatively large, and the volatility is significant; significant volatility; the concentration of O<sub>3</sub> has obvious seasonality and the highest fluctuation amplitude; the highest fluctuation amplitude; The average value of SO<sub>2</sub> is relatively low, indicating a significant effect of coal combustion control in recent years.

### 3.2 Time series change analysis

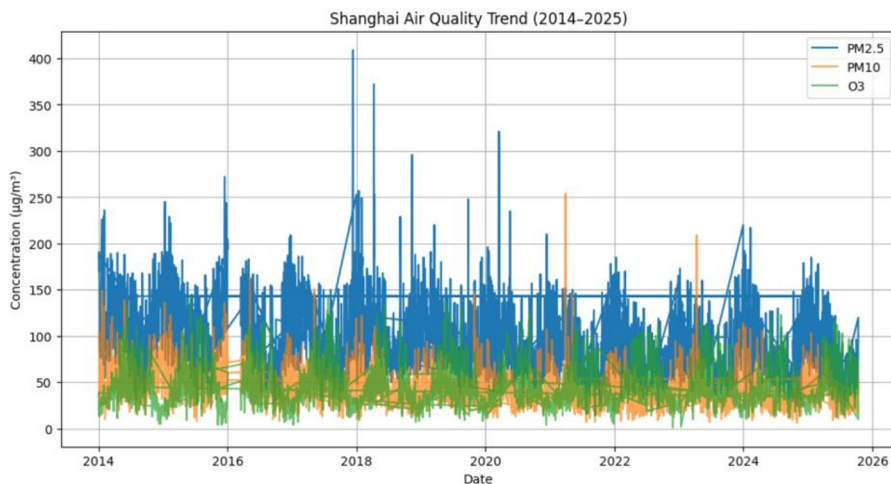
**Fig. 1.** Shanghai Air Quality Trend Chart (Picture credit: Original)

Figure 1 shows the three main air pollutants in Shanghai from 2014 to 2025 PM<sub>2.5</sub> the time series trend of daily average concentrations of PM<sub>10</sub> and O<sub>3</sub>. Through comparative analysis, it can be seen that the air quality in Shanghai has shown an overall improvement trend over the past decade.

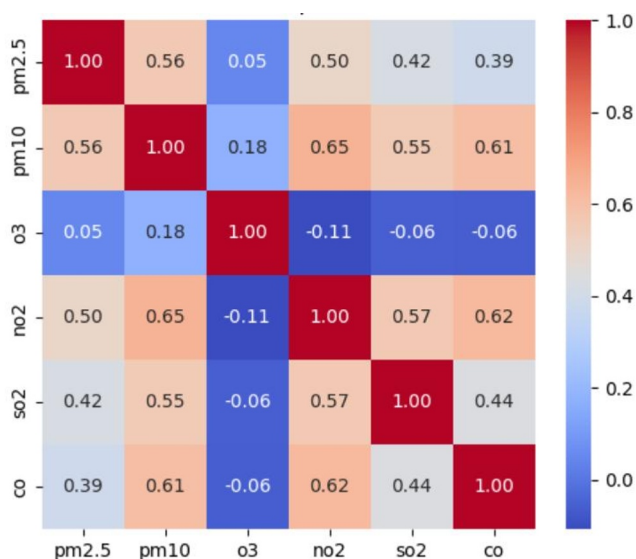
The change curves of PM<sub>2.5</sub> and PM<sub>10</sub> show a high degree of consistency, the two are basically synchronized in peak and valley positions, indicating a strong commonality in their pollution sources. This synchronicity is mainly due to the emission of inhalable particulate matter generated by human activities such as motor vehicle exhaust.

However, the trend of O<sub>3</sub> shows obvious seasonal characteristics. The concentration of O<sub>3</sub> significantly increases in summer, while it is generally lower in winter. This transformation has something to do with O<sub>3</sub>'s formation mechanisms and processes. O<sub>3</sub> is produced because of NO<sub>x</sub> and volatile organic compounds photochemical reactions with intense solar radiation. Therefore, the meteorological conditions of high temperature and

strong light in summer are conducive to the generation of ozone, while the low temperature and weak light conditions in winter are not conducive to its formation. It is worth noting that, O<sub>3</sub> and PM<sub>2.5</sub> shows a negative correlation, meaning that when particulate matter decreases, ozone concentration may increase. The characteristic of “PM<sub>2.5</sub> falling, O<sub>3</sub> rising” is more pronounced in some years.

In summary, the main pollutant concentrations in Shanghai from 2014 to 2025 exhibit the following characteristics: Seasonal fluctuations in O<sub>3</sub> concentration show high levels in the summer and low levels in the winter. Furthermore, the trend of particulates and ozone has opposite directions for their respective changes.

### 3.3 Correlation analysis of pollutants



**Fig. 2.** Thermal map of pollutant correlation (Picture credit: Original)

Figure 2 shows the Pearson correlation coefficient heatmap between six major air pollutants (PM<sub>2.5</sub>, PM<sub>10</sub>, O<sub>3</sub>, NO<sub>2</sub>, SO<sub>2</sub> and CO) in Shanghai from 2014 to 2025. By using color depth and numerical labeling, the degree of linear correlation between various pollutants can be intuitively reflected, providing data support for subsequent pollution source analysis and multivariate prediction model construction.

As can be seen from figure 2, there is a significant positive correlation between PM<sub>10</sub> and NO<sub>2</sub> ( $r = 0.65$ ) and CO ( $r = 0.61$ ) indicating that the concentration trends of the three are relatively consistent and may come from common emission sources, such as vehicle exhaust and combustion activities. PM<sub>10</sub> is also moderately correlated with SO<sub>2</sub> ( $r = 0.55$ ) indicating that coal combustion and industrial emissions still have a certain impact on particulate matter concentration.

The correlation coefficient between PM<sub>2.5</sub> and PM<sub>10</sub> is 0.56, indicating a moderate level of correlation. Although this is lower than the high correlation reported in some literature ( $r > 0.8$ ), it still reflects a certain degree of synchronicity between the two. So, this could be from checking the data daily and how the data is influenced by weather conditions like humidity and wind speed. Overall, the changes in PM<sub>2.5</sub> are still to some extent influenced by PM<sub>10</sub> both of which can be considered representative indicators of particulate matter pollution.

The positive correlation between  $\text{NO}_2$  and  $\text{SO}_2$  ( $r = 0.57$ ) as well as  $\text{CO}$  ( $r = 0.62$ ) indicates that, these three gaseous pollutants have strong consistency in their temporal changes, reflecting their common source characteristics in urban energy consumption and transportation activities. Both  $\text{NO}_2$  and  $\text{CO}$  mainly come from motor vehicle emissions, while  $\text{SO}_2$  mainly comes from industrial combustion and coal-fired power generation, this multi-source superposition often leads to a simultaneous increase in pollutant concentrations in the core areas of cities during high emission periods.

It is worth noting that the correlation between  $\text{O}_3$  and other pollutants is weak or even negative. The correlation coefficient between  $\text{O}_3$  and  $\text{NO}_2$  is  $-0.11$ , with  $\text{SO}_2$  being  $-0.06$ , and with  $\text{PM}_{2.5}$  is  $0.05$ , indicating that the trend of ozone changes is not synchronized with the primary pollutant. This result is consistent with the formation law of photochemical pollution: under strong solar radiation and low particulate matter load conditions, the concentration of  $\text{O}_3$  is prone to increase, while when the concentration of particulate matter and gaseous precursors (such as  $\text{NO}_2$ ) is high, ozone generation is actually inhibited.

In summary, the heatmap reveals the multi-source complexity of air pollution in Shanghai: transportation and combustion emissions are the main sources, while ozone pollution has an independent formation mechanism from particulate matter.

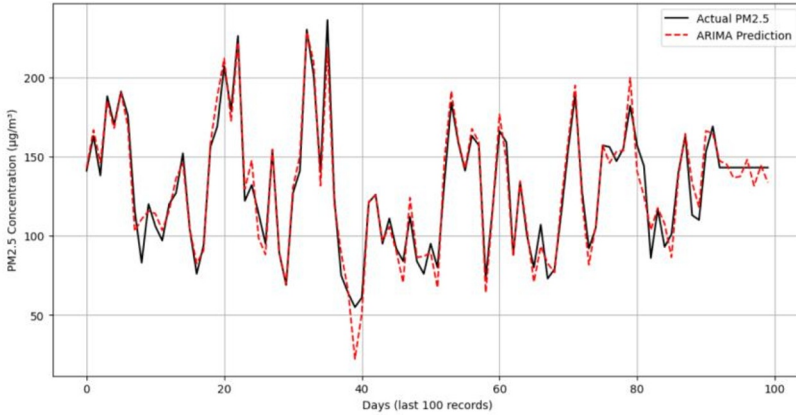
### 3.4 Comparison of Model Prediction Results

Figures 3 and 4 only selected the last 100 records from the dataset to compare the performance of ARIMA and LSTM models in predicting  $\text{PM}_{2.5}$  concentration in Shanghai. To verify the advantages of deep learning models in time series prediction, this study used the same training and testing dataset partitioning to model and evaluate both models under the same data constraints. In the figure provided, the actual values of  $\text{PM}_{2.5}$  concentration observed are shown with a black line. The red, dashed line breakdown shows the expected values from the ARIMA model, and the blue, dashed line describes the LSTM model expected values.

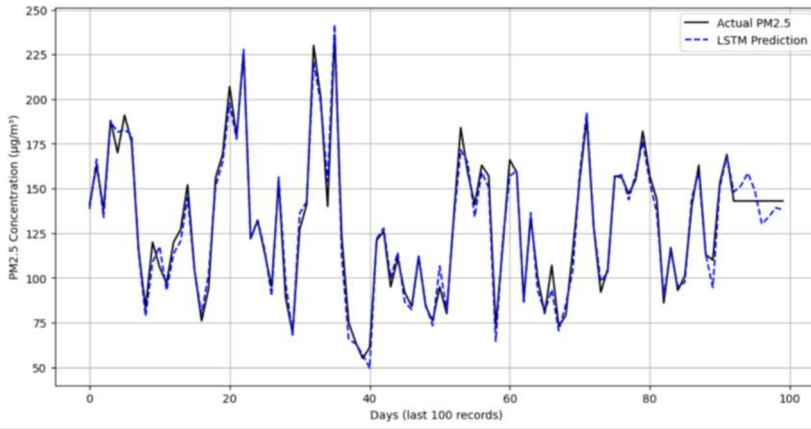
From Figures 3 and 4, it can be seen that both models can effectively capture the overall fluctuation trend of  $\text{PM}_{2.5}$  concentration, but there are differences in the predicted values of local fluctuations and peak values. The ARIMA model fits accurately in stable stages (such as lower concentration variation intervals), demonstrating its good performance in capturing linear trends and periodic fluctuations in data. However, there is a certain lag and bias in the predictions of ARIMA models during rapid rise or fall stages. This is due to the linear assumption of the model, which makes it difficult to capture the complex dynamic characteristics of air pollution.

In contrast, the prediction curve of the LSTM model is more consistent with the observed values, especially in the interval where the concentration rapidly increases or decreases. When learning from long sequences, LSTM alleviates the issues of gradient vanishing within traditional RNNs by applying input, forget, and output gates. This model is capable of not only capturing short-term fluctuations, but the long-term dependencies of pollutant concentrations, making it even more effective when handling environmental data, which consists of nonlinearity, a time lag, and seasonal fluctuations.

Based on the evaluation results, ARIMA model can capture and fit the trend of stable periods well, but there is a significant lag when rapid increases or decreases occur. During periods of high volatility, the predicted values of the LSTM model are closer to the observed values, indicating its advantage in handling nonlinear features and long-term dependencies.



**Fig. 3.** Comparison between ARIMA and Actual Values (Picture credit: Original)



**Fig. 4.** Comparison between LSTM and Actual Values (Picture credit: Original)

**Table 2.** Model Evaluation Results

Model	MSE	MAE	R <sup>2</sup>
ARIMA	91.38	7.31	0.94
LSTM	39.99	5.01	0.97

$$\text{MSE reduction rate} = \frac{\text{ARIMA MSE} - \text{LSTM MSE}}{\text{ARIMA MSE}} \times 100\% \tag{3}$$

$$\text{MAE reduction rate} = \frac{\text{ARIMA MAE} - \text{LSTM MAE}}{\text{ARIMA MAE}} \times 100\% \tag{4}$$

As shown in table 2, the mean square error decreased by about 56% and the average absolute error decreased by about 31%, indicating that its ability to capture complex nonlinear features is significantly better than traditional statistical models. Table 3 shows the comparative analysis between ARIMA model and LSTM model.

**Table 3.** ARIMA model vs. LSTM model

	ARIMA	LSTM
core principle	Describing linear trends through autoregression (AR), difference (I), and moving average (MA)	Using gate controlled loop units (input gate, forget gate, output gate) to capture

		long-term dependencies and nonlinear features
Ability to perform non-linear fitting	Weak, can only depict linear trends	Strong, able to learn multidimensional nonlinear relationships
Ability to respond to sudden fluctuations	Significant lag and poor adaptability to abnormal changes	Sensitive response, able to capture rapid fluctuation characteristics
training complexity	Low, high computational efficiency	High, requiring longer training time
Applicable scenarios	Short term stable data prediction, trend analysis, and explanatory research	Long term complex sequence modeling, air quality prediction, nonlinear dynamic analysis
summary	Simple structure, efficient computation, strong interpretability, but insufficient capture of nonlinear features	High fitting accuracy and strong generalization ability, suitable for complex pollution time series prediction tasks

## 4 Conclusion

In summary, this research evaluated the predictive capabilities of ARIMA and LSTM models on PM<sub>2.5</sub> data in Shanghai, It has been confirmed that the LSTM model can better capture the variability of nonlinear changes, seasonal changes, and sudden pollution events, And it performs better than the ARIMA model in all error indicators, and the relationship between gas oxygen and substances is weak, resulting in a negative correlation. It is necessary to incorporate various meteorological factors and photochemical mechanisms in the future to improve predictability, and researchers can also start from the complementary advantages of the two models, Thus, an ARIMA-LSTM hybrid model is constructed, which can handle both linear trends with ARIMA capability and non-linear feature learning with LSTM model to achieve better predictive performance.

## References

1. World Health Organization. Air pollution and health. <https://www.who.int/news-room/fact-sheets/detail/air-pollution> (2021)
2. F. Dominici, R. D. Peng, M. L. Bell, L. Pham, A. McDermott, S. L. Zeger, & J. M. Samet. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10), 1127–1134. (2006) <https://jamanetwork.com/journals/jama/fullarticle/202503>
3. V. I. Kontopoulou, A. D. Panagopoulos, I. Kakkos, & G. K. Matsopoulos. A review of ARIMA vs. machine learning approaches for time series forecasting in data-driven networks. *Future Internet*, 15(8), 255. (2023) <https://www.mdpi.com/1999-5903/15/8/255>
4. L. Breiman. Random forests. *Machine Learning*, 45(1), 5–32. <https://link.springer.com/article/10.1023/A:1010933404324> (2001)
5. N. Zaini, Z. Zainal, & N. Sulaiman. PM<sub>2.5</sub> forecasting for an urban area based on deep learning and decomposition method. *Scientific Reports*, 12, Article 21769. <https://www.nature.com/articles/s41598-022-21769-1> (2022)
6. A. Borovykh, S. Bohte, & C. W. Oosterlee. Conditional time series forecasting with convolutional neural networks. arXiv preprint. <https://arxiv.org/abs/1703.04691> (2017)

7. G. Zhang, B. E. Patuwo, & M. Y. Hu. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), 35–62. (1998) [https://doi.org/10.1016/S0169-2070\(97\)00044-7](https://doi.org/10.1016/S0169-2070(97)00044-7)
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, & I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. (2017) <https://arxiv.org/abs/1706.03762>
9. L. Bai, J. Xu, & Y. Chen. Air pollution forecasting: A review of methods and applications. *Journal of Environmental Sciences*, 66, 330–347. (2018) <https://www.sciencedirect.com/science/article/pii/S1001074217312299?via%3Dihub>
10. Z. Kazi, S. Filip, & L. Kazi. Predicting PM<sub>2.5</sub>, PM<sub>10</sub>, SO<sub>2</sub>, NO<sub>2</sub>, NO and CO air pollutant values with linear regression in R language. *Applied Sciences*, 13(6), 3617. (2023) <https://www.mdpi.com/2076-3417/13/6/3617>