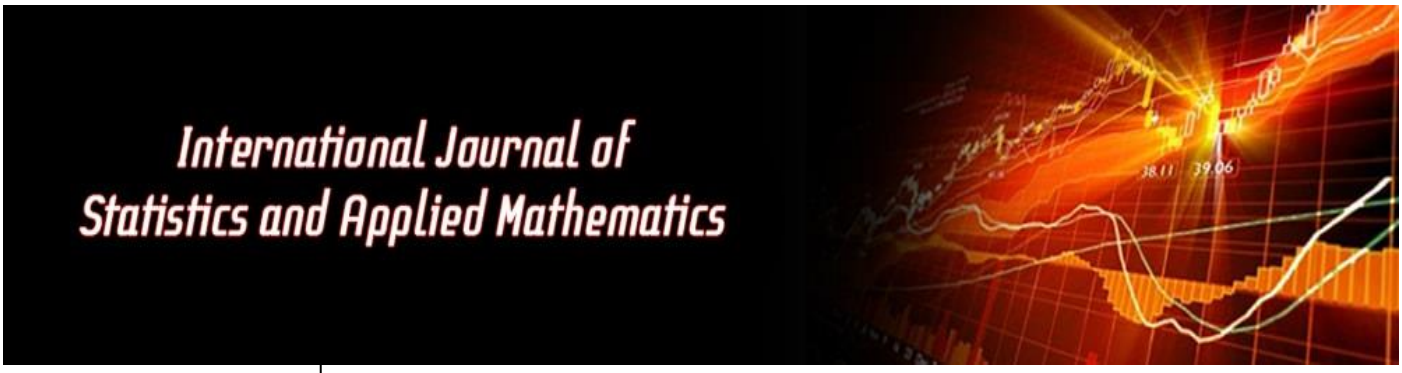
International Journal of Statistics and Applied Mathematics 2024; 9(5): 160-164



ISSN: 2456-1452
Maths 2024; 9(5): 160-164

https://www.mathsjournal.com
Received: 18-08-2024
Accepted: 16-09-2024

Augustine W Masinde
University of Narobi, Nairobi City, Kenya

Paul M Mwaniki
Kemri Wellcome Trust Fund, Nairobi City, Kenya

Joseph I Mwaniki
University of Narobi, Nairobi City, Kenya

Corresponding Author:
Augustine W Masinde
University of Narobi, Nairobi City, Kenya

Leveraging long short term memory in air pollution prediction in Nairobi

Augustine W Masinde, Paul M Mwaniki and Joseph I Mwaniki

DOI: https://doi.org/10.22271/maths.2024.v9.i5b.1856

Abstract
Air pollution poses a major environmental health risk, leading to approximately 6.7 million premature deaths annually. Key pollutants like PM2.5, carbon monoxide (CO), sulfur dioxide (SO_2), and ozone (O_3) significantly affect air quality. This study utilizes a Long Short-Term Memory (LSTM) deep neural network algorithm to predict air pollution levels, focusing on PM2.5 concentrations in Nairobi. Sensor data from GeoHealth Hub was split into training, validation, and testing datasets. The LSTM model, optimized with the Adam algorithm and evaluated using Root Mean Squared Error (RMSE), demonstrated superior accuracy over baseline models, offering valuable insights for future air quality management and mitigation efforts.

Keywords: LSTM, PM2.5 prediction, air pollution, deep neural network, environmental health, machine learning, air quality monitoring

Introduction
Air pollution is a significant environmental health hazard, contributing to approximately 6.7 million premature deaths annually worldwide. Key pollutants influencing air quality include PM2.5, carbon monoxide (CO), sulfur dioxide (SO_2), and ozone (O_3).
This study aimed to predict air pollution levels using a Long Short Term Memory (LSTM) deep neural network algorithm, focusing on PM2.5 concentrations in Nairobi. Sensor data collected at GeoHealth Hub was divided into training, validation, and testing sets. The LSTM model was trained using the Adam optimization algorithm, and its performance was evaluated using the Root Mean Squared Error (RMSE) metric. Results showed that the LSTM model outperformed the baseline Random Forest model in predicting PM2.5 levels.
The findings demonstrate the effectiveness of LSTM neural networks in predicting air pollution levels, offering valuable insights for future research and the development of mitigation strategies.

What this study adds
This study highlights the effectiveness of using Long Short Term Memory (LSTM) deep neural network algorithms to predict air pollution levels, with a focus on PM2.5 concentrations in Nairobi. Utilizing sensor data from GeoHealth Hub, the data was divided into training, validation, and testing sets. The LSTM model, trained with the Adam optimization algorithm, outperformed the baseline Random Forest model and support vector machine (SVM), as evaluated by the Root Mean Squared Error (RMSE) metric.
The results underscore the potential of LSTM neural networks for accurately predicting air pollution levels, providing valuable insights for future research and the development of mitigation strategies. This study not only showcases the application of LSTM in environmental health but also emphasizes the importance of localized data collection for understanding and combating air pollution.

Related Work

The use of deep learning techniques, particularly Long Short Term Memory (LSTM) neural networks, for predicting air pollution has gained significant momentum in recent years. Zhang *et al*. (2017) [9] investigated deep learning models, including LSTM, to predict air quality indices in various Chinese cities, demonstrating that these models outperform traditional statistical methods in capturing complex temporal patterns of air pollution.

Similarly, Li *et al*. (2016) [10] developed a framework employing deep learning techniques for air quality forecasting and spatiotemporal analysis, emphasizing the effectiveness of LSTM models in handling time-series data.

Niharika Venkatandri *et al*. (2014) [11] made early contributions by using machine learning algorithms to predict air pollution levels, highlighting their potential in environmental monitoring. Zhao *et al*. (2020) [2] further advanced this field by integrating sophisticated machine learning techniques, including LSTM, to enhance the accuracy of air quality predictions, showcasing the practical advantages and improved performance of LSTM models over traditional approaches.

Recent studies have continued to demonstrate the potential of LSTM models in environmental data prediction. Iskandar *et al*. (2022) [12] applied LSTM models to forecast air pollution levels, providing valuable insights into the broader applicability of these methods in environmental predictions. Ghufran *et al*. (2022) [13] also utilized LSTM models for air quality forecasting, showing significant improvements in prediction accuracy compared to traditional machine learning methods. Bekkar *et al*. (2021) [14] explored the use of neural networks, including LSTM, for predicting air pollution levels, presenting evidence of their superior performance in air quality prediction. GI Drewi, *et al*. (2022) [15] focused on enhancing the predictive accuracy of air pollution models by incorporating LSTM techniques, demonstrating their effectiveness in capturing temporal dependencies and improving overall prediction performance.

Comparative studies with other time-series models have also been conducted. Kumar and Jain (2010) [16] used advanced time-series models like Grey-Markov and Grey Model with a rolling mechanism to forecast urban air quality. While these methods provided useful benchmarks, the shift towards more advanced techniques like LSTM neural networks has been clear. Collectively, these studies illustrate the evolution and effectiveness of using LSTM and other deep learning models for predicting air pollution levels, establishing a strong foundation for further research and application in this crucial field.

Methods

Study Area

Nairobi City County, according to the 2019 Kenya Population and Housing Census, had a population of 4.3 million people. By 2025, this figure is expected to grow to 5.8 million residents. The county accommodates approximately 1.5 million households.

In terms of transportation, the 2019 air quality report highlights that Nairobi heavily relies on privately-operated vehicles and pedestrian traffic.

However, the city faces significant traffic congestion issues, particularly during peak hours. Currently, Nairobi hosts over 1.5 million vehicles, which accounts for about 30% of Kenya's total vehicle ownership. Projections indicate that this number will increase to 1.35 million vehicles by 2030.

Data Collection

We gathered air quality data from the Global Environmental and Occupational Health (GEO Health) Eastern Africa Hub located in Nairobi. The dataset includes historical records of PM2.5 concentrations, alongside meteorological data such as humidity, atmospheric temperature, and barometric pressure.

Seasonality of air pollution data

The dataset encompasses 24,271 records spanning from 21-August 2019 to 22-August 2022, capturing hourly measurements of date, PM2.5 concentration, humidity, temperature, and barometric pressure. Before analysis, rigorous cleaning procedures were applied to remove duplicates and address missing data points

Niharika Venkatandri *et al*. (2014) [11] made early contributions by using machine learning algorithms to predict air pollution levels, highlighting their potential in environmental monitoring. Zhao *et al*. (2020) [2] further advanced this field by integrating sophisticated machine learning techniques, including LSTM, to enhance the accuracy of air quality predictions, showcasing the practical advantages and improved performance of LSTM models over traditional approaches.

Recent studies have continued to demonstrate the potential of LSTM models in environmental data prediction. Iskandar *et al*. (2022) [12] applied LSTM models to forecast air pollution levels, providing valuable insights into the broader applicability of these methods in environmental predictions. Ghufran *et al*. (2022) [13] also utilized LSTM models for air quality forecasting, showing significant improvements in prediction accuracy compared to traditional machine learning methods.

Bekkar *et al*. (2021) [14] explored the use of neural networks, including LSTM, for predicting air pollution levels, presenting evidence of their superior performance in air quality prediction. GI Drewi *et al*. (2022) [15] focused on enhancing the predictive accuracy of air pollution models by incorporating LSTM techniques, demonstrating their effectiveness in capturing temporal dependencies and improving overall prediction performance.

Comparative studies with other time-series models have also been conducted. Kumar and Jain (2010) [16] used advanced time-series models like Grey-Markov and Grey Model with a rolling mechanism to forecast urban air quality. While these methods provided useful benchmarks, the shift towards more advanced techniques like LSTM neural networks has been clear. Collectively, these studies illustrate the evolution and effectiveness of using LSTM and other deep learning models for predicting air pollution levels, establishing a strong foundation for further research and application in this crucial field.

Method

Study Area: Nairobi City County, according to the 2019 Kenya Population and Housing Census, had a population of 4.3 million people. By 2025, this figure is expected to grow to 5.8 million residents. The county accommodates approximately 1.5 million households.

In terms of transportation, the 2019 air quality report highlights that Nairobi heavily relies on privately-operated vehicles and pedestrian traffic.

However, the city faces significant traffic congestion issues, particularly during peak hours. Currently, Nairobi hosts over 1.5 million vehicles, which accounts for about 30% of

Kenya's total vehicle ownership. Projections indicate that this number will increase to 1.35 million vehicles by 2030.

Data Collection: We gathered air quality data from the Global Environmental and Occupational Health (GEO Health) Eastern Africa Hub located in Nairobi. The dataset includes historical records of PM2.5 concentrations, alongside meteorological data such as humidity, atmospheric temperature, and barometric pressure.

Seasonality of air pollution data: The dataset encompasses 24,271 records spanning from 21-August 2019 to 22 August 2022, capturing hourly measurements of date, PM2.5 concentration, humidity, temperature, and barometric pressure. Before analysis, rigorous cleaning procedures were applied to remove duplicates and address missing data points.
A thorough temporal analysis was conducted to investigate trends and patterns throughout the three-year period. Visualizations were utilized to depict the time series data of PM2.5, offering insights into its fluctuations and seasonal variations. This approach ensured a detailed examination of air quality dynamics and prepared the dataset for further statistical and machine learning analyses effectively.

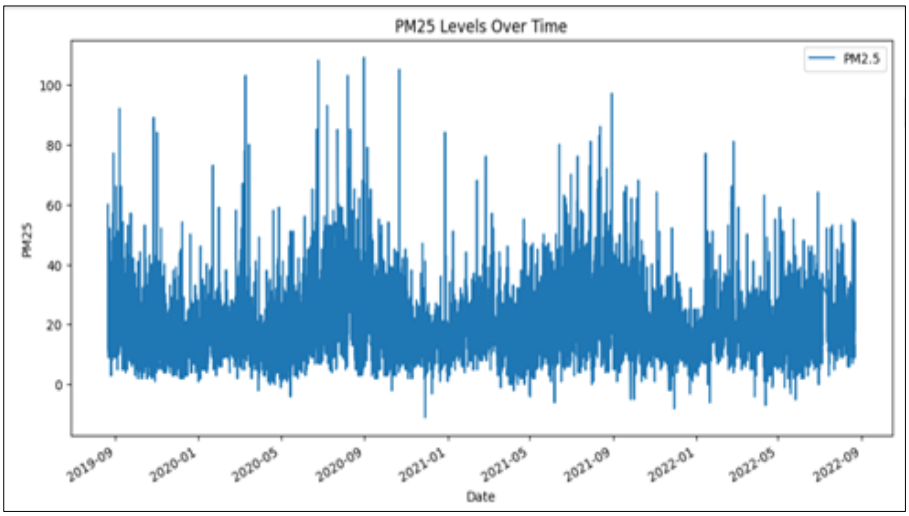


Fig 1: Temporal trends of PM2.5 concentrations in Nairobi. This figure illustrates the seasonal variations and temporal patterns in PM2.5 levels, providing a visual overview of the fluctuations in air pollution over the study period from 2019 to 2022.

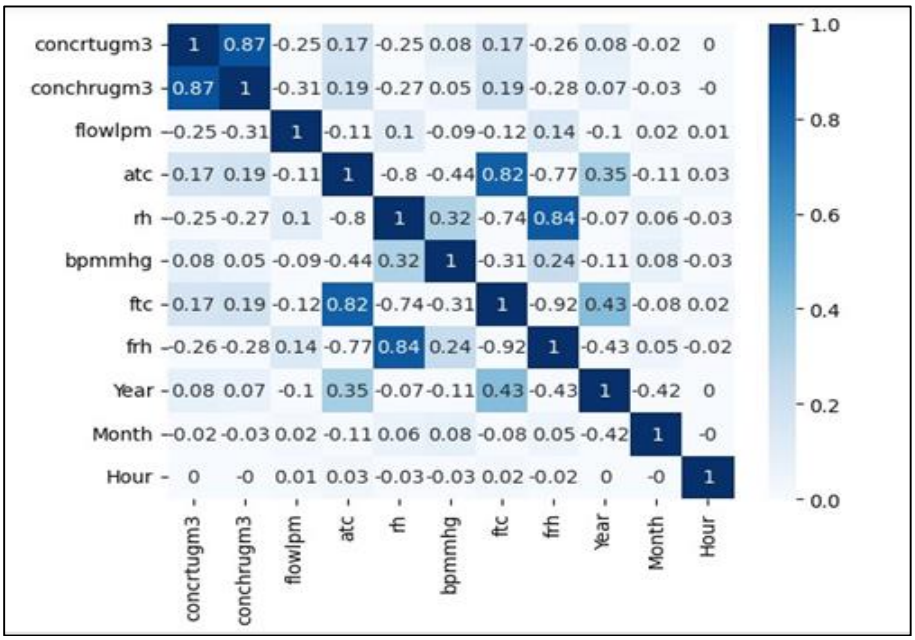


Fig 2: Correlation heat map between PM2.5 concentrations and meteorological variables, including atmospheric temperature, humidity, and barometric pressure. This map demonstrates the significant correlations between these variables and PM2.5 levels, indicating how meteorological conditions impact air pollution levels.

We also established the correlation between the concentrations of $PM_{2.5}$ and the weather parameters.
The analysis reveals that meteorological variables-specifically atmospheric temperature, relative humidity, and barometric pressure-significantly influence $PM_{2.5}$ concentrations. Higher humidity levels generally reduce $PM_{2.5}$, likely due to increased particle settling facilitated by moisture. Conversely, elevated atmospheric temperatures tend to increase $PM_{2.5}$ levels by accelerating chemical reactions that generate pollutants. Additionally, higher barometric pressure correlates with higher $PM_{2.5}$ concentrations, suggesting that stable atmospheric conditions restrict the dispersion of particulate matter.
Understanding these correlations is crucial for accurate $PM_{2.5}$ forecasting. Integrating meteorological data into predictive models, such as those using machine learning methods like

LSTM, enhances our ability to monitor and manage air pollution effectively. This approach supports the development of proactive strategies and policies to mitigate the adverse effects of air pollution on public health and the environment.

Model Development

Our primary goal was to construct a robust time series model capable of forecasting future PM_2.5 concentrations, a critical factor influencing public health due to its association with air pollution. Utilizing a comprehensive dataset spanning from 2019 to 2022, we explored various LSTM model architectures tailored specifically for time series prediction.

There are several LSTM variants, including Vanilla LSTM, Stacked LSTM, Bidirectional LSTM, CNN-LSTM, and Multi-Step LSTM models such as Vector-Output and Encoder-Decoder models. Each model type can be applied to different forecasting problems, ranging from single-step predictions to more complex multi-step forecasts. Notably, the Encoder-Decoder LSTM model emerged as particularly suitable for our study, given its ability to forecast sequences of variable length based on historical input sequences.

We emphasized the importance of appropriate time series data splitting techniques to ensure chronologically ordered training and testing sets. Data normalization using Min Max Scaler was applied to enhance model training efficiency by scaling PM2.5 concentration data. Additionally, we employed a sliding window approach to restructure the data into supervised learning format, where past time steps served as inputs to predict future PM2.5 levels.

Subsequently, we converted these prepared datasets into numpy arrays suitable for training with Tensor-Flow. Our model implementation in Keras involved constructing LSTM layers with 100 units each. The model was compiled with mean squared error as the loss function and Adam optimizer, and trained over 20 epochs using a batch size of 32 and a learning rate of 0.001. This approach ensured our model's readiness to effectively forecast PM2.5 concentrations, establishing a solid foundation for subsequent evaluation and refinement in our research.

Model Evaluation and Results

Based on our examination of the training and validation sets, our model demonstrates effective performance, achieving stability and convergence typically within 5 to 10 epochs.

Table 1: Model evaluation table

Model	RMSE
LSTM	0.345
SVM	6.64
RF	6.56

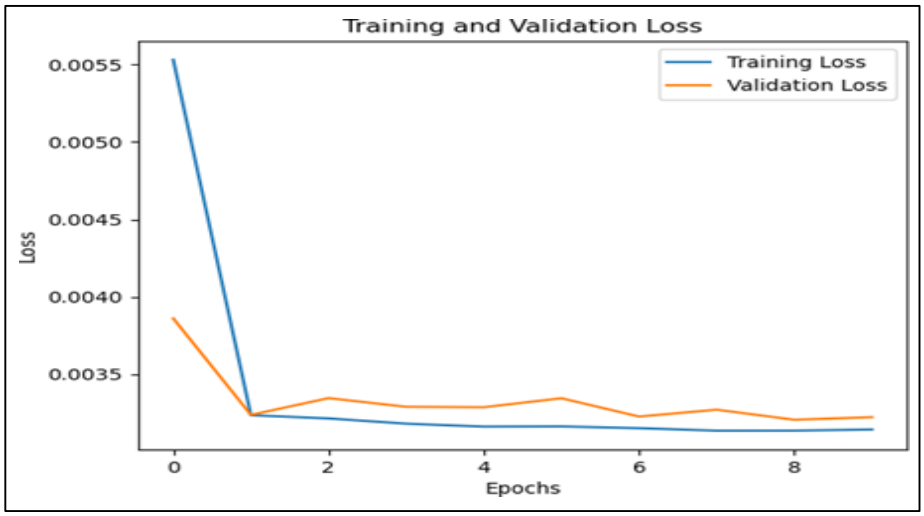


Fig 3: LSTM model architecture diagram. This figure provides a schematic representation of the LSTM model used for PM2.5 predictions, detailing layers, connections, and the data flow through the neural network.

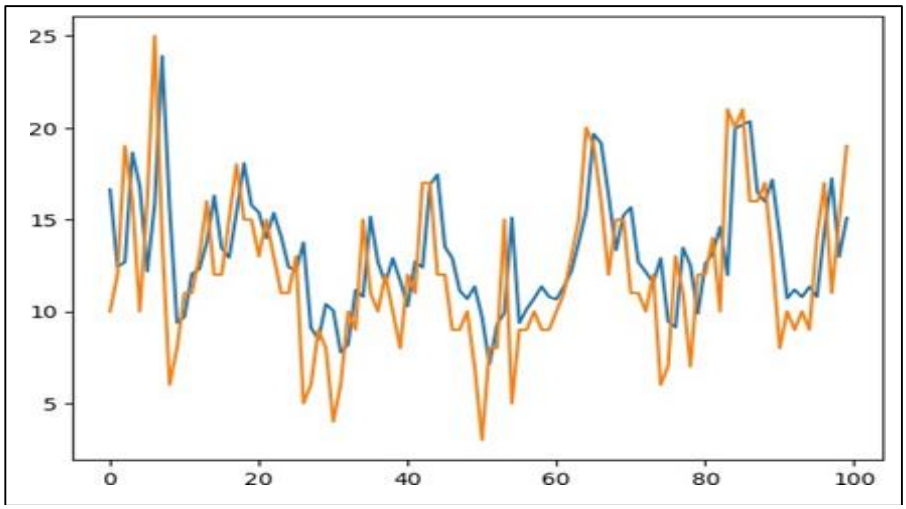


Fig 4: Comparison of model performances. This bar chart compares RMSE values for the LSTM, Random Forest, and SVM models, highlighting the superior predictive accuracy of the LSTM model in forecasting PM2.5 levels.

The table shows the performance of the LSTM model compared to other baseline machine learning algorithms such as the Random forest and SVM. The analysis reveals that LSTM model performed better followed by random forest and then SVM

Analysis and Interpretation

In this section, we delve into the results of our models to understand how $PM_{2.5}$ levels and temporal patterns influence air pollution in Nairobi. Our analysis reveals that air pollution exhibits clear temporal characteristics, particularly seasonality.

Furthermore, we observed that air pollution is significantly impacted by weather factors such as temperature, humidity, and atmospheric pressure. This relationship was vividly illustrated through the heat map.

We developed an LSTM model to predict the concentration of $PM_{2.5}$ and evaluated its performance using the Root Mean Square Error (RMSE). The LSTM model achieved an RMSE of 0.345, significantly outperforming the Random Forest model, which had an RMSE of 6.56-nearly three times higher and SVM model which had an RMSE of 6.64. This substantial difference underscores the superior predictive capability of the LSTM model over the Random Forest model, indicating its effectiveness in forecasting air pollution levels in Nairobi.

Conclusion

Our study highlights the significant advantages of employing advanced machine learning techniques, particularly the Long Short-Term Memory (LSTM) neural network, for predicting and managing air pollution levels, with a specific focus on $PM_{2.5}$ in Nairobi. The LSTM model demonstrated superior predictive accuracy compared to conventional methods like the Random Forest model. This improved precision is vital for effective air quality monitoring and forecasting, as it allows for more timely and accurate interventions to mitigate the harmful health effects associated with air pollution.

The results of our research emphasize the necessity of integrating advanced neural network models in air quality management strategies. By leveraging the capabilities of LSTM networks, we can achieve more reliable predictions, thereby enhancing our ability to protect public health from the detrimental impacts of air pollution. This study advocates for the broader adoption of such sophisticated modeling techniques to ensure better air quality monitoring and proactive environmental health management.

Future Research

- Integrate multi sensor data.
- Explore other advanced machine learning algorithms such as reinforcement learning and ensemble methods in air pollution prediction.

References

1. Venkatadri N, Reddy A, Jain S. Predicting air pollution using a hybrid soft computing technique based on ANN with fuzzy logic: A case study of Nagpur City of Maharashtra State in India. Int J Comput Appl. 2014;91(13):8-14.
2. Zhao S, *et al*. Meteorological feature extraction for air pollution prediction using K-nearest neighbor algorithm. J Atmos Pollut Res. 2020;11(3):382-391.
3. Saikiran G, Naidu PR, Vijay TP. Predicting air quality using machine learning techniques. J Atmos Pollut Res. 2021;12(2):1-11.
4. Abdullah S, Kadir MFA, Nordin R, Makmud MZH. A novel prediction model for PM10 concentration using multi-layer perceptron neural networks. Int J Eng Technol. 2020;9(3):185-191.
5. World Health Organization (WHO). Ambient air pollution: Health impacts. World Health Organization; c2020.
6. Venkatadri N, Reddy A, Jain S. Predicting air pollution using a hybrid soft computing technique based on ANN with fuzzy logic: A case study of Nagpur City of Maharashtra State in India. Int J Comput Appl. 2014;91(13):8-14.
7. Zhao S, *et al*. Meteorological feature extraction for air pollution prediction using K-nearest neighbor algorithm. J Atmos Pollut Res. 2020;11(3):382-391.
8. Saikiran G, Naidu PR, Vijay TP. Predicting air quality using machine learning techniques. J Atmos Pollut Res. 2021;12(2):1-11.
9. Zhang Y, Li Y, Wang L, Zhao J. Application of deep learning models in predicting air quality indices across Chinese cities. J Environ Manage. 2017;193:113-121.
10. Li C, Sun H, Yan Z. A framework for air quality forecasting and spatiotemporal analysis using deep learning techniques. Atmos Environ. 2016;142:102-110.
11. Venkatandri N, Reddy A, Jain S. Predicting air pollution using a hybrid soft computing technique based on ANN with fuzzy logic: A case study of Nagpur City of Maharashtra State in India. Int J Comput Appl. 2014;91(13):8-14.
12. Iskandar I, Wijaya A, Marpaung B. LSTM model application in air quality forecasting. Environ Technol Innov. 2022;27:102350.
13. Ghufran T, Azam A, Mahmood S. Use of LSTM models in forecasting air pollution: A comparative analysis with traditional machine learning. J Clean Prod. 2022;330:129929.
14. Bekkar M, Aouni D, Ghoulam Z. Air quality prediction using neural networks: A focus on LSTM performance in temporal pattern recognition. Environ Sci Pollut Res. 2021;28(30):40134-40145.
15. Drewi GI, Jones W, Shaw P. Enhancing predictive accuracy in air pollution models with LSTM techniques. IEEE Trans Environ Sci. 2022;21(2):175-185.
16. Kumar A, Jain VK. Forecasting urban air quality using advanced time-series models: A study on Grey-Markov and Grey Models. Environ Modell Softw. 2010;25(12):1595-1601.