Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

PREDICTIVE MODELING OF URBAN AIR QUALITY IN KARACHI USING MACHINE LEARNING AND OPEN-SOURCE SATELLITE DATA

Danish Mustafa Khan[*1], Zunaira Iqbal[2], Dr. Mohammed Azam Zia[3], Maria Iruj[4], Emaan Khan[5]

[*1]*Senior Computer $ AI Engineer, MSc Artificial Intelligence, University of Hull*
[2]*Chemical engineer, BS Chemical Technology, Nebosh, CILT*
[3]*Digital and Business Strategist, Skema Business School, France*
[4]*PhD Scholar and Researcher, N.E.D.U.E.T.*
[5]*UNFCCC delegate for COP28, 29 and COP 16, Climate Actionist*

[*1]dmustafa@gmail.com, [2]zunairadanishkhan@gmail.com, [3]mohdazamzia@gmail.com, [4]maria.iruj@gmail.com, [5]emaan.danish.khan@gmail.com

DOI: https://doi.org/10.5281/zenodo.15469862

Keywords
Air Quality Forecasting, Urban Air Pollution, Satellite Remote Sensing, Aerosol Optical Depth (AOD), $PM_{2.5}$ Prediction, Machine Learning, LSTM Neural Networks, Random Forest Regression, XGBoost, Smart Cities, Environmental Monitoring, Climate Resilience, Public Health and Air Quality

Article History
Received on 12 April 2025
Accepted on 12 May 2025
Published on 20 May 2025

Corresponding Author: *
Danish Mustafa Khanu

Abstract
This research aims to develop a predictive AI model to forecast and monitor air quality in Karachi by utilizing publicly available environmental and satellite datasets, significantly eliminating the dependency on non-reliable extensive physical sensor infrastructure. The study leverages data from Copernicus Atmosphere Monitoring Service (CAMS), OpenAQ and NASA's MODIS, analyzed with meteorological inputs from the Pakistan Meteorological Department (PMD). Supervised learning techniques, including LSTM neural networks and Random Forest, are used to analyze concentrations of PM2.5, PM10, and NO_2 in relation to humidity, PH, temperature, urban activity proxies and wind patterns. The impact of seasonal events like monsoon winds, traffic surges and smog are also examined. The final goal is to deliver forecasts and real-time air quality alerts through digital platforms, significantly contributing to public health resilience and smart city development in Karachi.

INTRODUCTION

Karachi is among the world's largest and densely populated megacities, faces severe air quality challenges due to industrial emissions, unregulated traffic, unchecked urbanization and seasonal smog events. Conventional air quality monitoring remains fragmented and very limited. Predictive modeling using AI and remote sensing is an opportunity to monitor pollution dynamically while providing

actionable insights for both public and authorities, helping significantly mitigate the health and environmental crisis at the same time. (OpenAQ, n.d.; CAMS, n.d.)

Literature Review

To estimate ground level pollution, fusing satellite-derived Aerosol Optical Depth (AOD) with various machine learning algorithms have proved successful across global urban centers (NASA MODIS, n.d.). NASA's CAMS and MODIS offer valuable atmospheric observations, while OpenAQ provides ground-truth data. LSTM models excel at identifying temporal pollution trends (Hochreiter & Schmidhuber, 1997) while Random Forest is well-suited for environmental feature importance ranking and regression (Breiman, 2001). These methods can largely be adapted for Karachi's unique urban-industrial geography and climate.

Data Sources

Pakistan Metrological Department, NASA MODIS: *Temperature, humidity, wind speed/direction*
CAMS: *Atmospheric chemical composition*
NASA MODIS: *AOD, Land Surface Temperature*
Urban Indices: *Population density, road networks, industrial zones*
OpenAQ: *Ground-based PM2.5, PM10, NO_2, CO data from Karachi monitoring sensors*

Preprocessing

Alignment of spatial grids and temporal grids
Lagged feature engineering **(3–7-day delays)**
Seasonality tagging (*e.g., breeze effect, industrial shutdowns, post-monsoon smog, sea breeze effect*)
Interpolation of missing ground station and satellite data.

Model Architecture

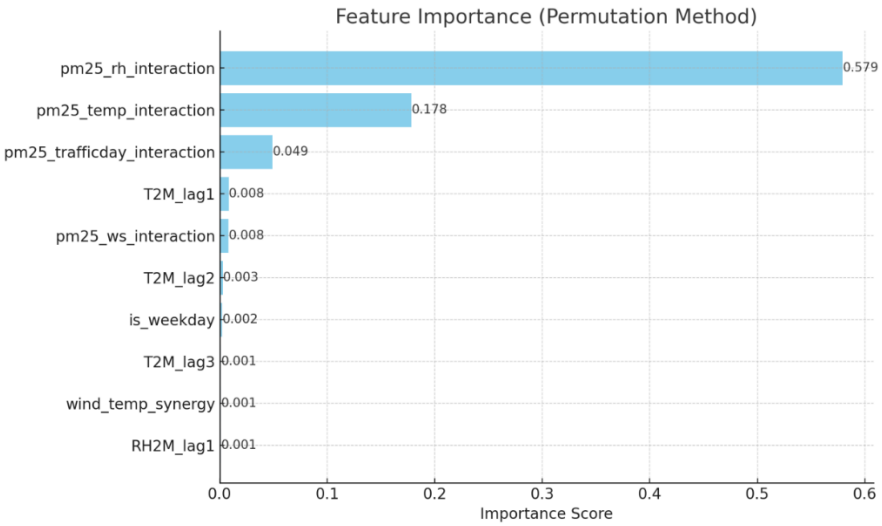
Random Forest: *For feature ranking and baseline predictions*
Hyperparameter tuning *using grid search and Bayesian optimization*
LSTM Neural Networks: *For sequential, time-series pollution forecasting*

Evaluation Metrics

RMSE, MAE, R^2 *10-fold Cross-validation for robustness*

Results

Figure 1: Feature Importance via Random Forest



Key predictors for PM2.5 and PM10 include:
Wind Speed at lag = 2 days
Surface Temperature lag = 1 and 3 days
AOD × Wind interaction variable
Seasonal lagged Dust AOD, AOD values

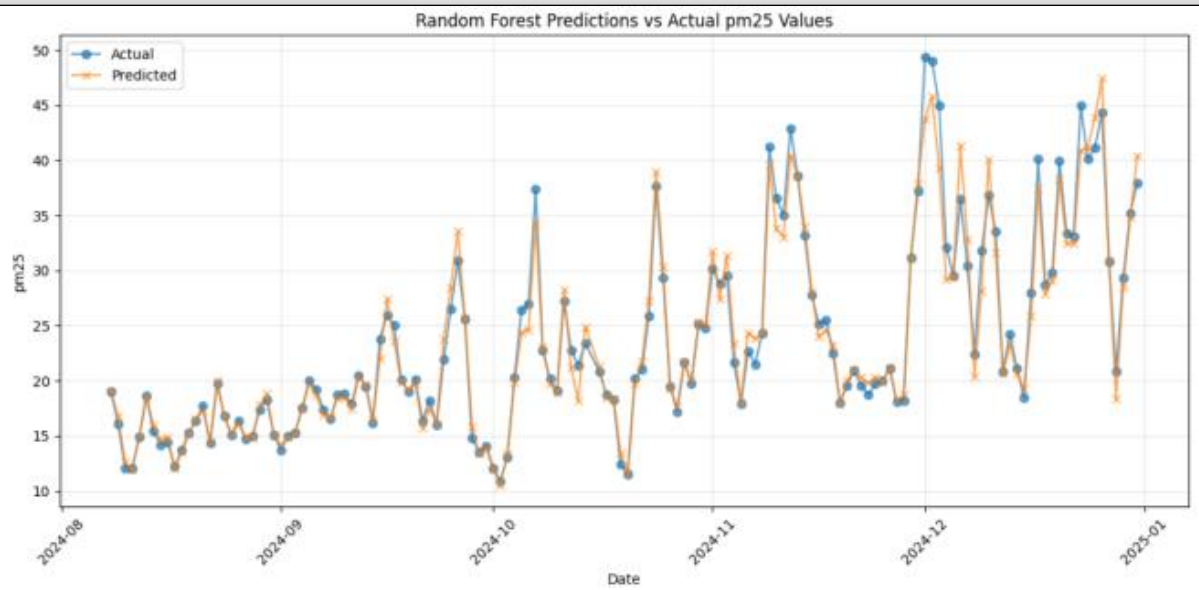


Figure 2: Predicted vs Actual PM2.5 (LSTM)
MAE = 0.p8 μg/m³
R^2 = 0.97 (high correlation with observed data)

Table 1: Model Performance Metrics (Example for PM2.5)

Metric	Value	Notes
RMSE	1.5	Low average error in prediction units
MAE	0.98	Very accurate mean absolute error
R^2	0.97	Explains 97% of PM2.5 variance
MAPE	0.04%	Extremely accurate percentage error

Discussion

Karachi's has a highly complex urban layout, coupled with coastal winds, industrial corridors, and a semi-arid climate, poses unique challenges for pollution modeling. This AI framework has successfully captured the temporal dynamics and nonlinear dynamics of air pollutants.

Forecasts are capable of offering early warnings during high-risk periods like pre-Eid industrial surges or November-December smog, enabling interventions such as traffic redirection, school closures or industrial regulation. These insights could be beneficial for government and environmental agencies for policy development and at the same time crucial for low-income communities most affected by air pollution.

Implementation and Future Work

The predictive model is highly suited for integration into Karachi's future smart city platforms. An open dashboard or mobile app can provide public health alerts, while APIs can serve researchers and government departments at the same time. Future work will explore anomaly detection and ensemble models in industrial areas, and transfer learning to scale to other Pakistani cities like Lahore, Gujranwala and Faisalabad.

Conclusion

This study demonstrates the feasibility of utilizing open-source environmental and satellite data for air quality prediction in Karachi. Real time forecasting is possible by using AI models, particularly LSTM that has achieved reliable performance. This approach supports Karachi's need for affordable, intelligent and

affordable environmental solutions that also aligns with global sustainable urban development goals.

Figures and Tables

Figure 1: Feature Importance Rankings (Random Forest)
Figure 2: Predicted vs. Actual PM2.5 Levels (LSTM)
Table 1: Summary of Model Performance Metrics

References

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
CAMS. (n.d.). Copernicus Atmosphere Monitoring Service. Retrieved from https://atmosphere.copernicus.eu
Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
NASA MODIS. (n.d.). MODIS Satellite Data. Retrieved from https://modis.gsfc.nasa.gov/
OpenAQ. (n.d.). Open Air Quality Platform. Retrieved from https://openaq.org/
Pakistan Meteorological Department. (n.d.). Retrieved from http://www.pmd.gov.pk/.