BreatheSafe – Predictive Analysis of Air Pollution Levels

Anushka Jadhav[1], Indrajit Joshi[2], Sampada Gupta[3], Siddharth Joisar[4], and Shweta S. Ashtekar[5*]

[12345]Dept. of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, India

{[1]anu.jad.rt22@dypatil.edu, [2] ind.jos.rt22@dypatil.edu, [3]sam.gup.rt22@dypatil.edu, [4]sid.joi.rt22@dypatil.edu, [5*]shweta.ashtekar@dypatil.edu }

*Corresponding Author

Abstract: The Air Quality Index is a good indication tool for the monitoring of air quality in smart cities and the assessment of the cleanliness or pollution level of air. Predictions of AQI values can facilitate people and authorities in taking precautionary steps like avoiding exposure to the outdoors on days when pollution levels are high. This study will look into the analysis of data and machine learning to forecast AQI by using previous pollution data, weather patterns, and environmental factors. A number of models have been trained and compared, which include a neural network to capture intricate non-linear correlations, XGBoost for efficient gradient boosting, Random Forest Regressor for robustness, and Linear Regression as the baseline. Results prove the potential of advanced models in real-time environmental monitoring and public health awareness, where models from an ensemble and deep learning yield better accuracy and reliability in predicting trends in air quality. Future iterations may consider regional or seasonal differences in air quality.

Keywords: Pollution levels, Linear Regression, Random Forest Regressor, Neural Network, Forecast AQI.

1. Introduction

Air pollution is a major threat to our everyday health and the environment, which is why our research focuses on creating an **early warning system** to predict the Air Quality Index (API)—giving cities time to issue warnings or manage traffic, and allowing people to take simple steps like wearing a mask or staying indoors.

1.1. Data Source

Data used here was compiled from records provided by the Central Pollution Control Board (CPCB) on official portal of the Government of India.Time Range: Dataset

© The Author(s) 2026
P. Goyal et al. (eds.), *Proceedings of the Conference on Technologies for Future Cities (CTFC 2025)*, Atlantis Highlights in Sustainable Development 7,
https://doi.org/10.2991/978-94-6239-650-0_13

covers data from 2015 up to that of 2020. It provides all daily measurements of different pollutants, aggregated at the city level from monitoring stations. Key Pollutants include metrics used PM2.5,PM10,NO,NO2,CO,SO and O3 among others.

1.2. Data Processing:

Data Preconditioning: In this step data is cleaned (using mean/median or dropping them) or removing outliers (via IQR or Z-score), and standardizing or normalizing features. IQR and Z-score were used as those are robust methods that used for identifying data points that can deviate significantly from rest of distribution.

Principal Component Analysis: Principal Component Analysis(PCA) is used in some models to reduce dimensionality and improve efficiency. PCA is linear technique that can transform large set of correlated variables into that of smaller set of uncorrelated components. It is used as we deal with highly collinear pollutant groups.

Model Training: Multiple models are trained and compared, including: Linear Regression as baseline model,Random Forest Regressor for ensemble method for robustness XGBoost gradient boosting-based model ,Neural Network for deep learning model for complex patterns Hyperparameter tuning is performed. Performance of all implemented models here was evaluated using standard regression metrics like R^2 score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Neural Network achieved highest accuracy with RMSE of 24.9424 and Random Forest achieved R^2 of 0.9490, among all of tested models which shows most predictive capability compared to other traditional machine learning approaches. XGBoost (Extreme Gradient Boosting) was used over simpler algorithms because it is a advanced ensemble technique which is designed for high predictive power.It is highly effective for capturing complex non-linear relationships and robust for noisy data, often making it best choice before implementing of deep learning.

Frontend Interface: A Streamlit-based web application allows users to input pollutant values and receive predicted AQI in real time. It's user-friendly and accessible to non-technical users.

2. Literature Survey

To forecast pollutants like PM2.5 and PM10, S. B. Sonu and A. Suyampulingam presented a linear regression model based on Python. Their method is appropriate for educational and policy applications because it emphasizes interpretability and simplicity.[1]

S. Sunori et al. compared Adaptive Neuro-Fuzzy Inference System (ANFIS) with Linear Regression to predict Air Quality Index (AQI). The study discovered that while regression is straightforward, ANFIS increases accuracy by modeling non-linear interactions.[2]

B. D. Parameshachari et al. investigated some number of machine learning techniques, such as Random Forest and Linear Regression, for AQI forecasting. This study showed how ensemble learning techniques can raise prediction accuracy and dependability.[3]

R. Renugadevi et al. employed Random Forest algorithm to predict the AQI and the pollutant concentrations. This model offered scalability and also accurate environmental monitoring and it worked well with big and also complicated datasets.[4]

To predict the AQI levels in Jakarta, R. Muljana et al. A Random Forest Classifier was used. In line with the study, localized air quality was supported by the model analysis by effectively identifying patterns of pollution in real-world, noisy data.[5]

S. Al-Eidi et al. estimated the air quality in smart cities contrasted regression methods like Linear, Ridge and Support Vector Regression (SVR). Their findings provided valuable insights into the efficacy, efficiency, and flexibility of the strategy for monitoring the air in cities.[6]

C. Li et al. employed environmental data from the real world to examine different AQI machine learning models and forecasting the level of pollutants. The results revealed that forecasting is enhanced by data-driven techniques accuracy and support environmental planning.[7]

To find trends over time in data on air quality, Long Short-Term Memory was used by Dhilsath Fathima M et al. Recurrent neural networks and (LSTM) models (RNN). Regarding anticipated accuracy, their in-depth. The learning approach outperformed the traditional machine learning techniques.[8]

Graph neural networks were used by Ditsuhi Iskandaryan et al. GNNs for modeling temporal and spatial dependencies in Madrid's air quality data. The GNN-based model outperformed conventional techniques by utilizing spatial correlations between observation points.[9]

For AQI forecasting, Cheng Zhao et al. used feedforward and recurrent neural networks. Their study showed how neural models can better represent non-linear relationships, leading to more precise forecasts of air quality.[10]

3. Proposed Methodology

3.1. Steps of Methodology

This section will elaborate various steps involved in the proposed methodology as shown in Figure 1. The steps in the suggested methodology, as depicted in Figure 1, will be explained in detail in this section.

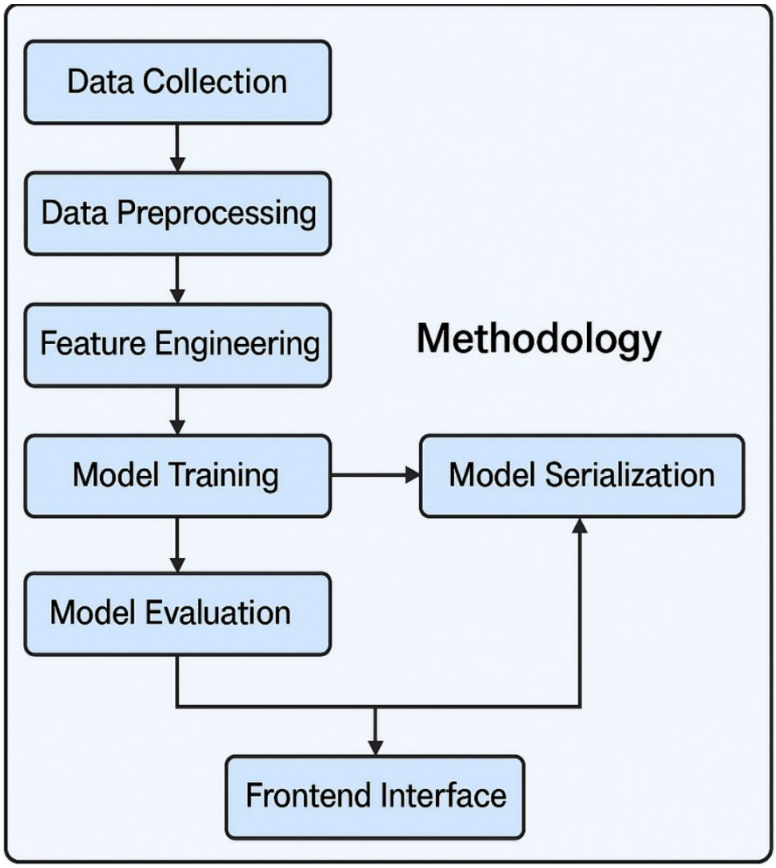

Fig. 1: Methodology

User Interface: On the homepage, users are prompted to input current pollutant concentration values in an intuitive form, such as PM2.5 PM10 NO SO CO O.
Data Entry: Users can input current concentrations of contaminants recorded in their area. There is a clear label in each input field to prevent confusion.
Forecasting AQI: After filling the fields, the user clicks "Predict" or "Submit" button. The trained model will process data after app sends these inputs to backend.

Display Results: Air Quality Index (AQI) results shows up on screen in seconds. With colour coded visuals and warnings to indicate severity, with help of AQI score, app will provide a category name like "Good," "Moderate," or "Unhealthy," .
Examining Outcomes: To help users understand AQI number, app may include a scale or legend. Health advice such as "Reduce outdoor activity" for high AQI levels may also be included.
Repeat or Modify the Inputs: As pollution levels change throughout day, users can see different scenarios or get updated predictions by adjusting input parameters.
Reactivity and Accessibility: Viewing of air quality on forecasts on PCs, tablets, and also smartphones is simple by using the design of website.

3.2. Features

Data used here was compiled from records provided by the Central Pollution Control Board (CPCB) on official portal of the Government of India, consisting of 52,874 records. Dataset used in project is "Air Quality Data in India (2015 - 2020)" used from Kaggle, which covers 20+ cities of India. Time Range: Dataset covers data from 2015 up to that of 2020. Data Type: It provides all daily measurements of different pollutants, aggregated at the city level from monitoring stations. Key Pollutants: Features include metrics used PM2.5,PM10,NO,NO2,CO,SO and O3 among others. Target Variable: Primary target for prediction is Air Quality Index (AQI) value and all of corresponding bucket (e.g., Good, Moderate).

Backend logic and Streamlit based interface are features of the implementation that make it simple for non-tech users to engage with system. The interface includes dynamic model switching, color coded AQI output, dropdown menus for category levels like Moderate and Satisfactory. There's interactive sliders for entering pollutant concentrations, and visual aids like scatter plots, pie charts, and bar charts to show data patterns and model performance. Regression diagnostics that compare actual and expected values in real time. Users, environmental scientists, and policymakers, can respond to changes in air quality with help of these features.

4. Findings and Discussions

4.1. Results

Because of complete modularity, this research project is seems ideal for environmental studies, policymaker tools, and campaigns for public awareness . It provides clear visualizations, real-time prediction, and support for multiple model types. The project shows how to successfully integrate data science, machine learning, and frontend development to produce a solution that is both technically stunning and useful in real-world scenarios.

The Air Quality Index (AQI) Prediction offers approach to air quality forecasting by combining deep learning and also conventional machine learning techniques. An interactive and user friendly web interface that has strong feature engineering, data pre-

processing, and model training procedures was made by using Streamlit. Technical precision and usefulness are also prioritized in this work.

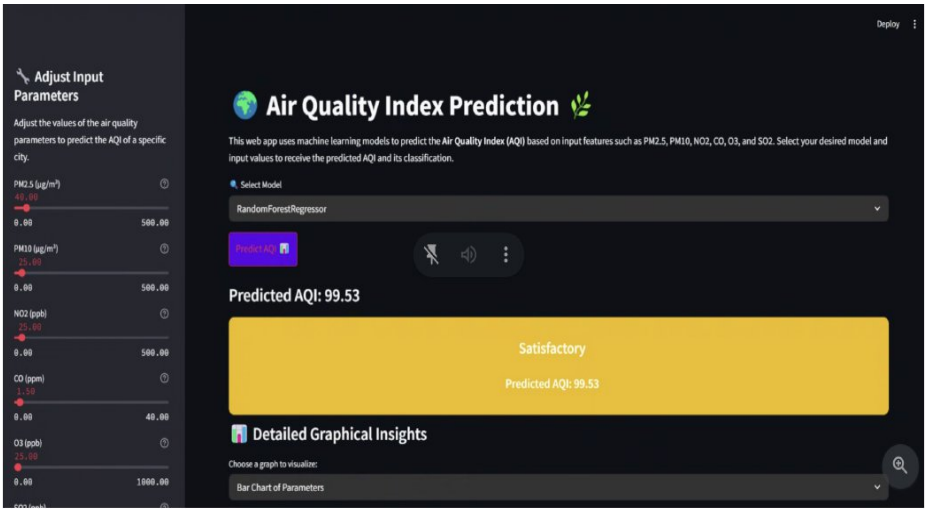

Fig. 2: Basic Prediction Interface

To obtain AQI predictions and also health based classifications like "Satisfactory" or "Moderate" immediately, you can select from a variety of pre trained models (like Linear Regression, Random Forest, XGBoost, and Neural Networks) and input various pollutant concentrations (like PM2.5, PM10, NO2, CO, O3, and SO2).

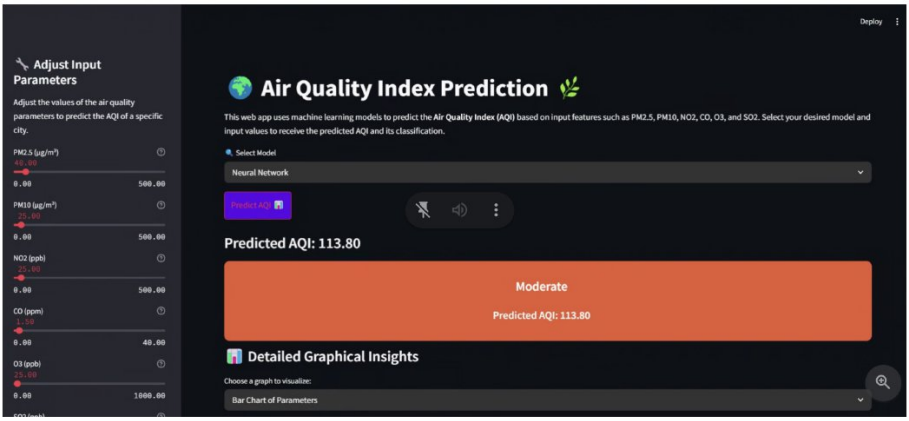

Fig. 3: Switching Models

This application also provides complete graphical insights using regression analysis, line plots, pie charts, and bar charts to help users better understand significance of each.

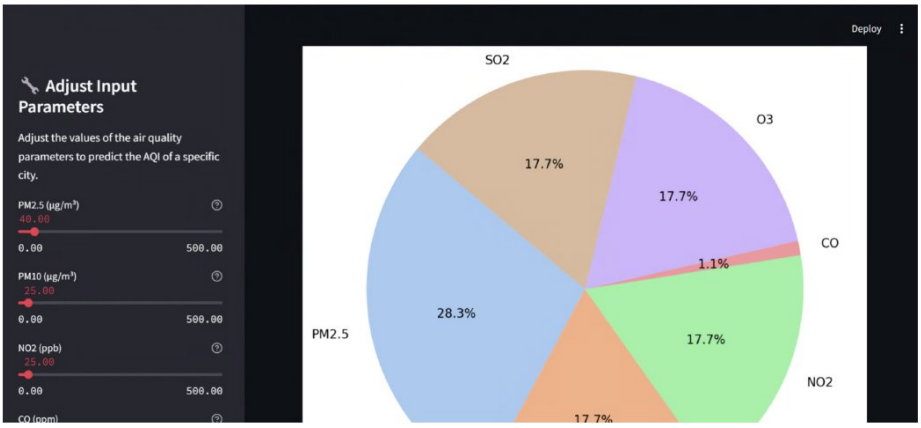

Fig. 4: Pie Chart of Pollutant Contribution

Users can view the relative contributions of each pollutant to the total input data in Figure 4. PM2.5 is the biggest contributor (28.3), followed by SO, NO, and O, each of which makes up about 17.7 percent. This function aids users in comprehending the main elements influencing air quality.

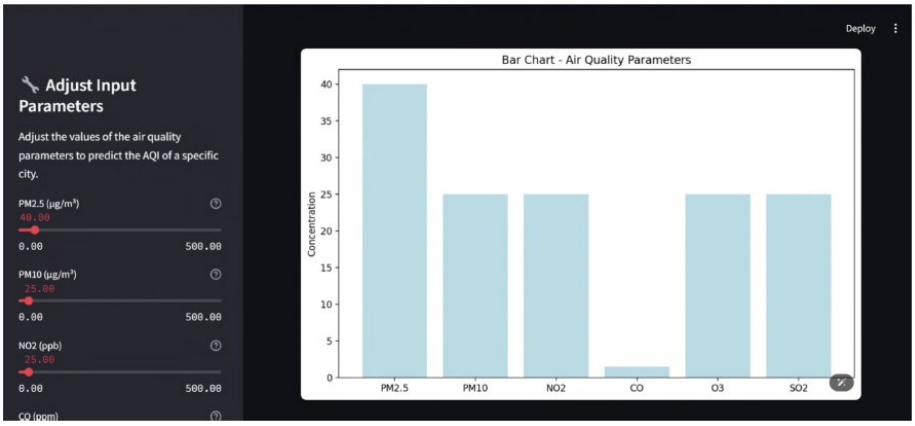

Fig. 5: Bar Chart Visualization

Figure 5 displays a bar chart of the air quality parameters that graphically depicts the concentration levels of each pollutant. By making it easier to compare pollutants side by side, it increases the accuracy of input values before forecasting.

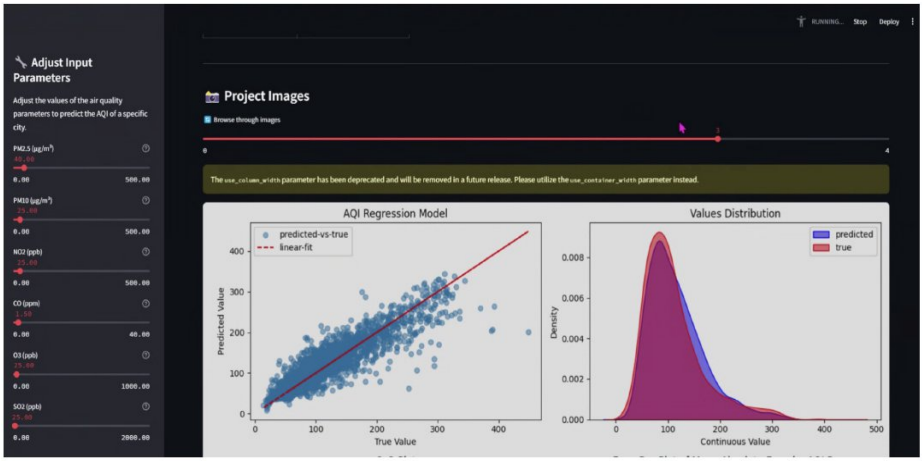

Fig. 6 : Regression Fit and Distribution

The interface shown in Figure 6 includes two plots: a distribution plot that illustrates how predicted and actual values match, and a scatter plot that contrasts predicted and actual AQI values.

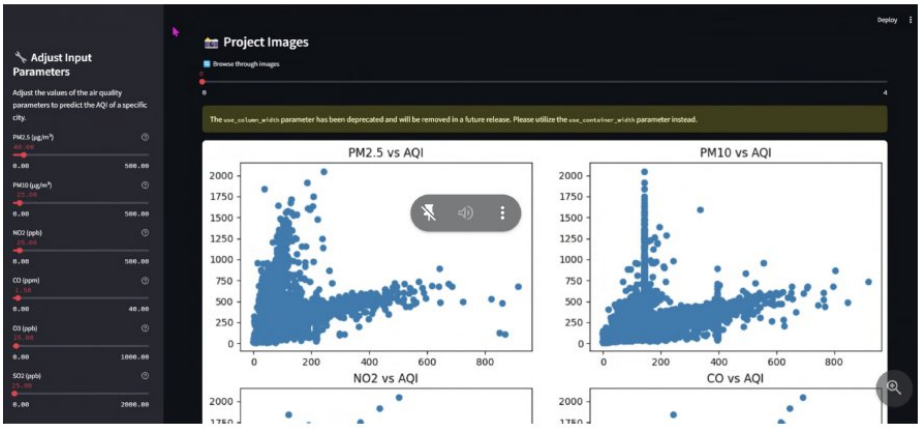

Fig. 7 : Feature vs. AQI Relationships

Relationships such as PM2.5 vs. AQI and PM10 vs. AQI are shown in multiple scatter plots in Figure 7.

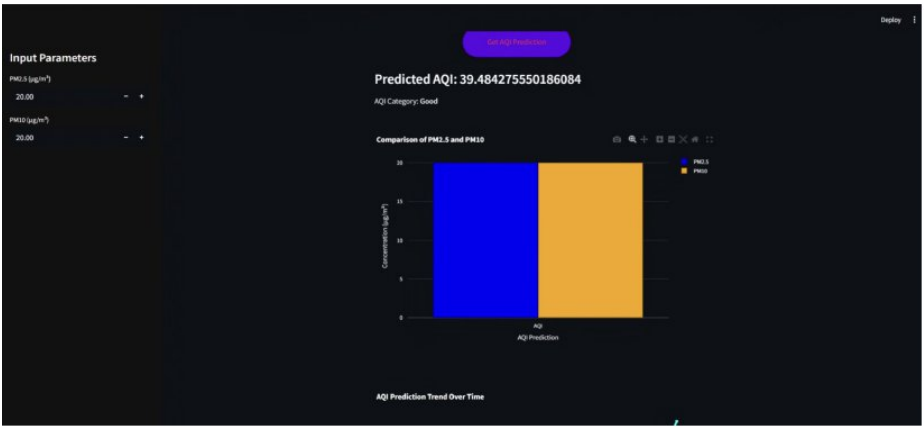

Fig. 8: Model Comparison Chart

As indicated in Figure 8, the performance of several models (Linear Regression, Random Forest, XGBoost, and Neural Network) is compared in a table or bar graph. This graphic helps choose the top-performing model by showing how each model performed according to metrics like RMSE and R2 score.

This research introduces novel framework for AQI prediction by combining advanced data cleaning, statistical outlier detection, and dimensionality reduction prior to model training. Unlike several studies that already exist that depend solely on either linear regression or a single machine learning model, this work evaluates multiple models, including Random Forest, XGBoost, Support Vector Regression, and a Neural Network–based deep learning model. Novelty further lies in the systematic benchmarking of these models using consistent evaluation metrics to determine most optimal architecture. Additionally, inclusion of real-time prediction interface demonstrates practical applicability of system beyond theoretical analysis, addressing real-world deployment feasibility for Indian environmental monitoring needs.

4.2. Model Limitations

Temporal Constraints: Higher complexity models like XGBoost and Neural Networks can have possibility to overfitting despite of cross-validation and of early stopping that can limit performance on new data.

Feature Deficiency: Reliance on only six input features may sometime result in discarded temporal information of which can limit model's ability to capture important seasonal or long term trends.

Data Drift/Non-Stationarity: Non stationary nature of AQI time series data can mean relationships between pollutants and of AQI can change over time as due to policy or climate shifts. A static model can experience data drift leading to degraded prediction accuracy over long time.

Data Quality Dependence Imputation Bias: Accuracy of model is constrained by residual bias introduced by during statistical imputation of missing values.

Input Reliability: Accuracy is strictly dependent on the quality and completeness of the sensor input data, as processing artifacts or remaining noise will always limit predictive ceiling.

5. Conclusion

This air quality index prediction utilizes real-time data. The driven system proves as an efficient and useful for Forecasting quality of air using both ML and deep learning models. Each detail is carefully integrated within an end-to-end pipeline that prioritizes accuracy and usability, from feature engineering and Data pre-treatment to model training, evaluation and deployment. The very interactive system Streamlit online interface, which places a major emphasis on public usefulness, enables the user to enter pollutant concentrations and obtain AQI estimates and associated health-based classifications quickly. Users can compare the results of Linear Regression, Random Forest, XGBoost, and Neural Networks thanks to the app supporting multiple model options. The platform also provides users with rich graphical insights that enable them to observe parameter contributions and model behavior.

Future Scope include:

Integrating External Weather Data: We can incorporate some real time meteorological features such as wind speed, humidity, and temperature as these are some variables that can strongly influence of pollutant dispersion and it can significantly improve predictive accuracy.
Develop Spatio - Temporal Models: Shift towards some advanced architectures like (e.g., Graph Neural Networks or Convolutional LSTMs) to model that of geographic influence of neighboring cities and sources which can also enhance prediction of localized pollution events.
Deploy as a Real-Time Forecasting Service: Transition current Streamlit app to a fully automated cloud-based service that provides AQI forecasts 24/7 that is complete with automated model retraining that can address data drift.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. S. B. Sonu and A. Suyampulingam, “Linear Regression Based Air Quality Data Analysis and Prediction using Python,” in Proc. 2021 IEEE Madras Section Conference(MASCON), Aug. 2021, doi: 10.1109/MASCON51689.2021.9563432.
2. S. Sunori, D. Verma, P. B. Negi, and P. K. Juneja, “Air Quality Index Prediction using Linear Regression and ANFIS,” in Proc. 2024 Int. Conf. Inventive Computation Technologies (ICICT), Apr. 2024, doi: 10.1109/ICICT60155.2024.10544842.
3. B. D. Parameshachari, G. M. Siddesh, V. Sridhar, M. Latha, K. N. A. Sattar, and G. Manjula, “Prediction and Analysis of Air Quality Index using Machine Learning Algorithms,” 29–30 Jul. 2022.
4. R. Renugadevi, T. Vyshnavi, T. P. Reddy, and P. S. Lahari, “Air Quality Prediction using Random Forest Algorithm,” in Proc. 2023 Int. Conf. Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), 01–02 Nov. 2023, doi: 10.1109/RMKMATE59243.2023.10369180.
5. R. Muljana, L. D. Ayuningtyas, R. P. Daksa, S. F. Djamhari, M. A. Fiezayyan, and N. T. M. Sagala, “Air Pollution Prediction using Random Forest Classifier: A Case Study of DKI Jakarta,” 16 Feb. 2023.
6. S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani, and N. Niveshitha, “Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques,” IEEE Access, vol. PP, no. 99, pp. 1–1, Jan. 2023, doi: 10.1109/ACCESS.2023.3323447.
7. C. Li, Y. Li, and Y. Bao, “Research on Air Quality Prediction Based on Machine Learning,” 17–19 Nov. 2021.
8. Dhilsath Fathima M, Sashank Donavalli, and Harshitha Kambham, “Air Quality Prediction using Deep Learning Models,” in Proc. 2024 Int. Conf. Advancements in Power, Communication and Intelligent Systems (APCI), Jun. 2024, doi: 10.1109/APCI61480.2024.10616969.
9. Ditsuhi Iskandaryan, Jose Francisco Ramos, and Sergi Trilles Oliver, “Graph Neural Network for Air Quality Prediction: A Case Study in Madrid,” IEEE Access, vol. PP, no. 99, pp. 1–1, Jan. 2023, doi: 10.1109/ACCESS.2023.3234214.
10. Cheng Zhao, Mark van Heeswijk, and Juha Karhunen, “Air Quality Forecasting Using Neural Networks,” in Proc. 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 06–09 Dec. 2016, published 09 Feb. 2017, doi: 10.1109/SSCI.2016.7850128.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.