OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Comparative Analysis On Different Modeling Techniques Reliability For Predicting PM 2.5 Values

Aadrit Nandy
aadritnandy10@gmail.com

ABSTRACT

Fine particulate matter ($PM_{2.5}$) is an air pollutant that affects public health and outcomes and is an environmental threat. According to research from the National Library of Medicine (NIH), "7.3 billion people are directly exposed to unsafe average annual $PM_{2.5}$ concentrations" (Rentschler and Leonova). The ability to forecast future pollutant levels such as $PM_{2.5}$ would improve public health by warning people to be wary and minimize exposure adequately. It is fundamentally impossible to have a guaranteed prediction, though models have the ability to be exceptionally close. Therefore, utilizing the model with the best performance at predicting $PM_{2.5}$ values by objective standards offers the best protection for respiratory systems. The dataset was built by collected information from an OpenAQ sensor in Mumbai, India. The dataset has 102,492 instances. The best model for predicting $PM_{2.5}$ was determined by creating a ranking system and point allocation method: Linear Regression, Decision Tree, Deep Neural Network, XGBoost, and Random Forest. The study determined that random forest was most suitable for $PM_{2.5}$ prediction in Mumbai.

INTRODUCTION

Fine particulate matter, $PM_{2.5}$, is defined as particles in air with diameters of 2.5 micrometers or smaller. It is one of the most harmful measures of pollution as it can penetrate the respiratory system and potentially enter the bloodstream, compromising health. *Our World In Data* mentions that "in total 8.8 million people die prematurely due to air pollution every year" (Roser). Human activity like rapid urbanization, increasing commute, and industrial activity have increased harmful particle emission and pollution levels. As these levels rise, $PM_{2.5}$ levels are an important field of study. Consequently, the ability to reliably forecast $PM_{2.5}$ concentration is an important research objective since it can help governments and individuals take caution against these through measures, preventing a global public health risk.

Although multiple machine learning models have been trained for predicting environmental features, determining the most ideal model for a specific metric like $PM_{2.5}$ concentration forecasts hasn't been focused on. Atmospheric pollution is a combination of features that is influenced and determined by complex interactions between conditions and pollutant concentrations. Therefore, the relationship is extremely non-linear and deeper models with the ability to recognize these intricate patterns may be needed. Commonly, deeper models that can track and make accurate predictions are computationally

expensive, and $PM_{2.5}$ forecasting needs immediate action and therefore a balance between accuracy and efficiency is ideal. Which machine learning model has the highest level of predictive accuracy and computational efficiency for forecasting $PM_{2.5}$ concentrations using a dataset from an OpenAQ sensor in Mumbai, India? To answer this question, a comparative analysis is done between commonly used models with different abilities. The models were an extreme gradient boost model, a
decision tree, a linear regression model, a deep neural network, and a random forest model. They were trained with the same data set and evaluated by objective standards.

Therefore, the need for machine learning models with high levels of accuracy in $PM_{2.5}$ prediction is necessary. OpenAQ is an NGO that provides free data regarding various AQI factors like emission rates and general environment conditions (OpenAQ, 2026) . These features include and can help predict $PM_{2.5}$ values. Different machine learning models have strengths in prediction and classification of various fields. As $PM_{2.5}$ prediction is such a contemporary issue, finding the best algorithm for predicting $PM_{2.5}$ is apposite. A comparative analysis of popular algorithms can indicate which algorithms are most apt for $PM_{2.5}$ prediction.

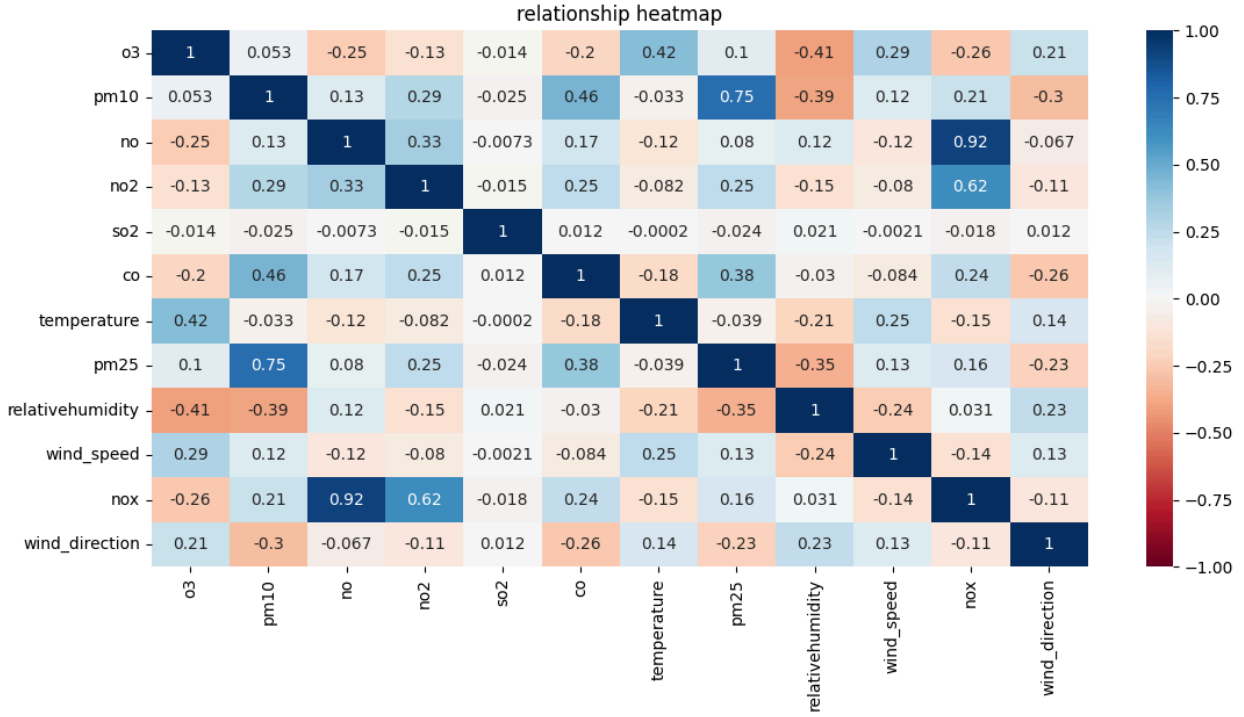

Figure 1: A heatmap of the OpenAQ feature's relationships

"An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution" by Di and team focuses on ensemble models and doesn't compare against baseline standard regression models. The research is conducted across the United States, compared to this paper's focus area of Mumbai, India. In Mumbai, air pollution is a more relevant issue, though combining

this paper's findings with Di and his team brings the findings to a global level (Di et al.). Di and team focus on building a model for prediction while this research compares existing ones. "Development of a daily PM_{10} and $PM_{2.5}$ prediction system using a deep long short-term memory neural network model" focuses on comparing neural networks with different settings abilities' to predict $PM_{2.5}$ and PM_{10}. $PM_{2.5}$ is more invasive than PM_{10} due to it being easier for smaller particles to invade the body, hence why this research doesn't focus on its prediction. Despite Kim and team's research being conducted in South Korea which typically has less arbitrary environment conditions, some of their accuracy standards like R were quite low (Kim et al.). Such low reports are expected in this research as Mumbai has unpredictable $PM_{2.5}$ spikes frequently.

Unlike previous studies which predict different values and a combined output, like AQI predictors, this research hones down to a single metric that is pivotal in determining public health risk and pollution levels. The analysis considers computational efficiency and predictive quality. The findings provide guidance for selecting models and emphasize the unpredictability of real-world settings by comparing benchmarks with and without outliers. This research can help organizations and individuals forecast $PM_{2.5}$ and lower public health risk.

METHODS

Setup and data sources:
The data is a table of environmental sensor reports organized as a time series. There are 102,492 instances in the dataset. OpenAQ limits the time span of data downloadable directly from their web application (OpenAQ, 2026) . Still free, they provide an API key plus instructions for accessing data from a larger frame like one year. The API was used to access data from a 1 year time span in Mumbai. Subsequently, the data was loaded onto a Google sheet for inspection. There were rows missing due to past issues of emission and environment sensors having temporary collection breakdowns. The missing data was estimated by averaging the previous and following records (interpolation). The features were, O_3, PM_{10}, NO, NO_2, SO_2, CO, temperature, $PM_{2.5}$, relative humidity, wind speed, NOX, and wind direction. From the original data set, there were 12 missing values, from faulty sensors or gaps in recording. For $PM_{2.5}$ concentrations, which almost conform to a uniform curve through a given day's progression, interpolation does a good job estimating these null values.

Models:
Linear regression was the first model tested. It assumed a linear relationship between the features and predictors for $PM_{2.5}$. It is extremely fast in terms of training time and prediction time. Due to the simplicity, it serves as a good baseline for comparison between other models as it becomes more evident if higher complexity leads to better results.

A decision tree was tested afterwards. It is flowchart-esque and splits the data recursively. It is easy for visualization and overall simplicity, because it is possible to trace the model's decision path that led to the

final prediction. It typically performs well on training data yet it is highly susceptible to overfitting, leading to weaker results on new data.

Deep neural networks contain multiple hidden layers of connected artificial neurons. Each neuron receives weighted inputs and applies non-linear activation functions. Subsequently, it passes the output as input to the following layer. The multifaceted approach due to the presence of multiple hidden layers allows for complex-non linear patterns and relationships to be captured. Air quality and $PM_{2.5}$ levels are affected by intricate interactions with the weather and the natural environment as well as human activity. The complicated factors suggest non-linear relationships and might require multiple layers to capture. Therefore, a deep neural network might be ideal for environmental predictions. However, neural networks and especially DNNs require longer training times and prediction times.

Random forest models break the data set into multiple individual trees. It averages their outputs leading to more accurately weighted decisions. Environmental data sets commonly contain unpredictable and complex non linear relationships. By aggregating multiple sub trees, random forest models are less prone to random noise and overfitting, which is important due to erratic real-life conditions. However, the synthesis of data for better results is time-consuming. This results in higher training times and prediction times.

Extreme Gradient Boost (XGBoost) is a machine learning algorithm that constructs trees sequentially, each time adapting to the residuals from the previous build. By correcting prediction errors, XGBoost models respond well to datasets containing outliers, preventing overfitting. This leads to higher accuracy making it ideal for tabular data. Previous research has demonstrated XGBoost having strong performance in forecasting applications. *Optimal starting point for time series forecasting* depicted how XGBoost and LightGBM models were efficient in forecasting intervals (Zhong et al. 12). Its ability to capture non-linear interactions between variables makes it suitable for $PM_{2.5}$ forecasting. It requires longer training time, but its performance makes up for computational costs.

Model	Training Split % / Testing Split %	Validation	Hyper Parameters
Linear Regression	80/20	N/A	Default settings
Decision Tree	80/20	N/A	Criterion = “squared-error”, max depth = None, min_samples_split = 2, min_samples_leaf = 1
XGBoost	80/20	N/A	objectives= ‘reg:squarederror’,

Model	Training Split % / Testing Split %	Validation	Hyper Parameters
			n_estimators = 100
Deep Neural Network	80/20	20%	Arc: 128-64-32-1, epochs=50, b_size = 32, Loss = MSE, Metric = MAE, Dropout=0.2
Random Forest	80/20	N/A	Default settings

Metrics:
No single metric can determine which model has the best performance for $PM_{2.5}$ prediction. However, some metrics are more relevant than others. For example, while lower training times are preferred, accuracy is way more significant since lowering error as best as possible is optimal for $PM_{2.5}$ prediction.

Mean Absolute Error is a measure of the average magnitude of prediction errors. Every residual contributes equally to the final reported MAE. MAE was chosen as an evaluation metric because it allows a straightforward comparison of accuracy.

RMSE measures accuracy by taking the square root of the average prediction error squared. Squaring the residuals punishes outliers by causing larger errors to contribute to the score disproportionately. Therefore, RMSE is more sensitive to significant prediction errors. This is valuable in the context of $PM_{2.5}$ prediction since underestimations or overestimations may lead to inaccurate reports of public health safety or risk. As a result, RMSE compliments MAE by emphasizing not just the models' average errors, but their ability to avoid large prediction errors.

R^2, the coefficient of determination, is a measure of the proportion of variability for $PM_{2.5}$ predictions that is explained by the predictive model. Unlike MAE and RMSE which quantify residuals in predictions, R^2 captures how well the variation in $PM_{2.5}$ concentration is captured by the given algorithm. r^2 provides insights on the explanatory power of models.

Training time represents the amount of time needed for each algorithm to learn from the training data set and build the prediction model. However, predictive accuracy is the main focus. That being said, training time is still considered as it gives us insight on the computational costs associated. As long as training time length isn't at the point where constructing new models for $PM_{2.5}$ prediction takes time remarkably longer than any other model for lower performance, training time isn't too much of a factor for comparison.

Prediction time measures the amount of time required for a trained model to generate predictions on never seen data. For operational $PM_{2.5}$ forecasting systems, forecasts need to be rapidly produced in order to alert public health updates timely. Consequently, a model with an extremely high degree of precision yet long prediction times can be rendered useless in urgent settings, due to a lack of practicality. Measuring prediction time helps compare efficiency between models.

Data processing:
For an objective comparison, all models were trained with the same dataset and evaluation metrics. Predictive accuracy is the most important, with RMSE being more important than MAE due to RMSE's ability to reflect error rate. R^2 follows since it demonstrates the effectiveness of features. Computational efficiency is a secondary criteria, and prediction time is more important than training time. The comparison system that emphasizes accuracy over computational efficiency follows:

Metric	RMSE	MAE	R^2	Prediction time	Training time
Maximum Score	25	20	15	10	5

The model ranking the highest in each category gets the maximum score, and the model in the next position gets a fraction of the maximum score ⅘, ⅗, ⅖, ⅕. The total points for each metric are added and the best model for this case is determined. It follows this pattern and descending order based on importance in order to more heavily punish accuracy errors.

Results & Discussion

Model	MAE	RMSE	R^2	Training Time (seconds)	Prediction Time (seconds)
Linear Regression	10.0768	28.2366	0.2806	0.0185	0.0126
Decision Tree	10.3111	28.1712	0.2839	0.0994	0.0023
XGBoost	8.4348	25.3106	0.4220	0.2995	0.0052
Deep Neural	8.9599	26.2462	0.3784	19.3528	0.1905

Network					
Random Forest	7.8415	24.9598	0.4379	7.6482	0.0969

Figure 2: Results by performance standards for each model in $PM_{2.5}$ prediction

Model	MAE	RMSE	r^2	Training time	Prediction time	Total
Linear regression	8	5	3	5	6	27
Decision Tree	4	10	6	4	10	34
XGBoost	16	20	12	3	8	59
Deep neural network	12	15	9	1	2	39
Random Forest	20	25	15	2	4	66

Figure 3: Scoring each model based on previous standards

According to figure 2, Random forest demonstrated the overall best performance by achieving the lowest MAE, RMSE and highest R^2. Although computation efficiency was below average, shown by the higher training times and prediction times, it was superior in prediction accuracy. XGBoost followed closely in each accuracy standard and ranked second between all the models. However, its computational costs were a little bit above average, yet computational efficiency is second to accuracy for an environmental prediction model like $PM_{2.5}$ prediction.

The DNN achieved an intermediate $PM_{2.5}$ prediction performance, ranking third at 39 overall points. It outperformed both linear regression (27) and the decision tree (34)in accuracy metrics and overall score. However, due to its multiple hidden layers its computational costs were high. Despite relatively strong predictive capability, its efficiency wouldn't be very practical in real life. Linear regression and the decision tree model had similar accuracy standards, but the decision tree was computationally cheaper, with an efficiency score of 14, vs linear regression's 11.

Overall, random forest performed the best, but XGBoost was still very reliable. Both models were a good balance of efficiency and accuracy. Therefore, Random Forest is the best algorithm for $PM_{2.5}$ prediction.

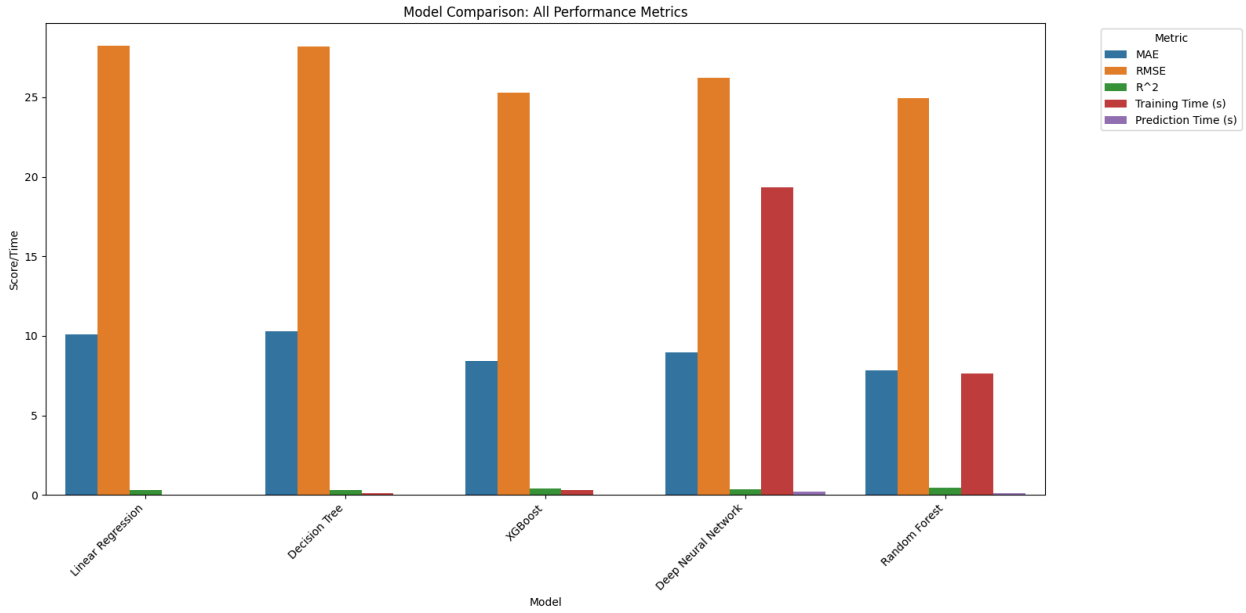

Figure 4: Comparison of all models by metric in a bar graph

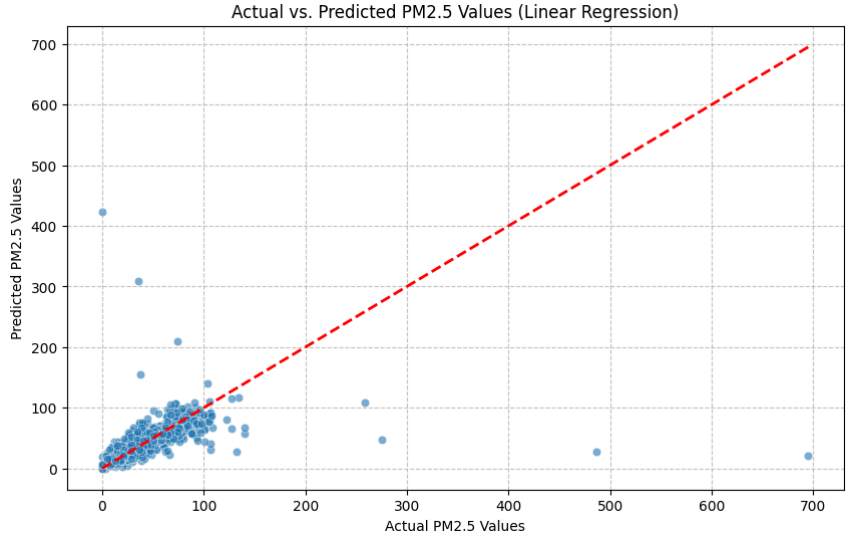

Figure 5: Graph of linear regression predicted $PM_{2.5}$ values vs actual values

In the linear regression graph, it becomes clear that the model was relatively accurate in predicting $PM_{2.5}$ values for smaller values, within the expected range. However, larger values and outliers compromised the accuracy.

Model	MAE	RMSE	R^2	Training Time (seconds)	Prediction Time

					(seconds)
Linear Regression	9.005176	15.389379	0.631561	0.037332	0.001973
Decision Tree	9.736683	19.004586	0.438125	0.132679	0.002165
XGBoost	7.205762	11.693901	0.787263	0.491374	0.006898
Deep Neural Network	7.883252	12.865722	0.742492	29.199261	0.251411
Random Forest	6.863251	11.837405	0.782010	6.216028	0.057680

Figure 6: Filtered table after removing outliers

Model	MAE	RMSE	r^2	Training time	Prediction time	Total
Linear regression	8	10	6	5	10	39
Decision tree	4	5	3	4	8	24
XGBoost	16	25	15	3	6	65
Deep neural network	12	15	9	1	2	39
Random Forest	20	20	12	2	4	58

Figure 7: Scoring table updated after filtering

XGBoost performed better when removing the 10 outliers in both directions. However, outliers do occur in real life since the environment is unpredictable. Therefore, random forest is still slightly superior for $PM_{2.5}$ prediction, yet the updated results show the drastic effect of outliers on environmental prediction, especially if you look at the accuracy standards. The outliers were determined by using manual inspection to locate the 10 lowest $PM_{2.5}$ reports and the 10 highest $PM_{2.5}$ reports.

CONCLUSION

Air pollution remains one of the leading threats to public health, emphasizing the importance of accurate $PM_{2.5}$ predictions. The study compared 5 models' (Extreme Gradient Boost, Deep Neural Network, Random Forest, Decision Tree, Linear Regression) performances. The data was collected from a one year span from OpenAQ environment trackers in Mumbai, India. Each model was judged based on weighted standards and an objective ranking system. The system prioritized predictive accuracy over computational efficiency as reliability matters more for models that concern public health safety and risk.

The results indicate that random forest achieved the overall best performance by having the highest weighted score. XGBoost performed very strongly as well. Although the deep neural network had relatively strong results it was too computationally expensive. Simpler models that couldn't capture complex curves and patterns like decision trees and standard linear regression are not ideal. When removing outliers XGBoost had a higher weighted score and all accuracy metrics improved greatly. However, in the real environment due to unpredictable conditions outliers can't be discounted. Therefore, random forest is the strongest model for $PM_{2.5}$ prediction. The findings also suggest that ensemble-based machine learning models are more suitable due to their ability to capture environmental patterns, making them favorable compared to simpler approaches. These findings additionally demonstrate how random forest predicts more accurately on full, natural datasets, and filtered or limited datasets are best handled with XGBoost for prediction.

The study is limited in the location analyzed. The data was tracked from a single monitoring station, and this doesn't provide a comprehensive overview of global $PM_{2.5}$ concentrations. However, the location chosen, Mumbai, is very relevant since India notably has hazardous $PM_{2.5}$ levels and AQI. This research can help people subject to public health risk from harmful emissions in Mumbai due to traffic, smoking and general pollution. For future study, a more comprehensive study that aggregates multiple sensors' data from spots globally would provide more representative insights.

REFERENCES

Di, Qian, et al. "An Ensemble-Based Model of $PM_{2.5}$ Concentration across the Contiguous United States with High Spatiotemporal Resolution." *Environment International*, vol. 130, Sept. 2019, p. 104909, https://doi.org/10.1016/j.envint.2019.104909.

K S, Harishkumar, et al. "Forecasting Air Pollution Particulate Matter ($PM_{2.5}$) Using Machine Learning Regression Models." *Procedia Computer Science*, vol. 171, Elsevier, June 2020, pp. 2057–66, https://doi.org/10.1016/j.procs.2020.04.221. Accessed 8 Aug. 2026.

Kim, Hyun S., et al. "Development of a Daily PM $_{10}$ and PM $_{2.5}$ Prediction System Using a Deep Long Short-Term Memory Neural Network Model." *Atmospheric Chemistry and Physics*, vol. 19, no. 20, Copernicus GmbH, Oct. 2019, pp. 12935–51, https://doi.org/10.5194/acp-19-12935-2019. Accessed 7 Aug. 2026.

Rentschler, Jun, and Nadezda Leonova. "Global Air Pollution Exposure and Poverty." *Nature Communications*, vol. 14, no. 1, July 2023, p. 4432.

Roser, Max. "Data Review: How Many People Die from Air Pollution?" *Our World in Data*, 25 Nov. 2021, https://ourworldindata.org/data-review-air-pollution-deaths. Accessed 7 Aug. 2026.

Zhong, Y., Ren, Y., Cao, G., Li, F., & Qi, H. (2024). *Optimal starting point for time series forecasting*. arXiv. https://arxiv.org/abs/2409.16843

OpenAQ. (2026). *OpenAQ Air Quality Data Explorer*. Retrieved July 22, 2026, from OpenAQ Platform.