OPEN ACCESS

Volume: 4

Issue: 4

Month: October

Year: 2025

ISSN: 2583-7117

Published: 15.10.2025

Citation:

Aesha Bhardwaj, Dr. Sanjay Silakar, Dr. Rajeev Pandey, Dr. Jashwant Samar "Accurate PM2.5 Prediction Using Machine Learning Ensembles for Sustainable Smart Cities" International Journal of Innovations in Science Engineering and Management, vol. 4, no. 4, 2025, pp. 10–17.

DOI:

10.69968/ijisem.2025v4i410-17

Accurate PM2.5 Prediction Using Machine Learning Ensembles for Sustainable Smart Cities

Aesha Bhardwaj[1], Dr. Sanjay Silakar[2], Dr. Rajeev Pandey[2], Dr. Jashwant Samar[3]

[1]Research Scholar, Uit Rgpv Bhopal, M.P.
[2]Professor, Uit Rgpv Bhopal, M.P.
[3]Assistant Professor, Uit Rgpv Bhopal, M.P.

Abstract

Particularly in India, where cities like Delhi suffer from serious air quality problems, air pollution is a major obstacle to urban sustainability and requires precise prediction models to help smart city projects. Using the Central Pollution Control Board's Air Quality Data in India (2015–2020) dataset, this study, Air Quality Prediction for Sustainable Smart Cities using Machine Learning, creates a reliable framework for predicting PM2.5 concentrations across 26 Indian cities. The dataset was optimised for modelling by careful preparation, which included one-hot encoding of city variables, IQR-based outlier treatment, SimpleImputer for missing values, and exploratory data analysis to find trends. Support Vector Regressor, Gradient Boosting Regressor, Random Forest Regressor, and Extra Trees Regressor were the four machine learning models that were trained. A hybrid ensemble that combined Random Forest and Extra Trees through a voting mechanism performed better (R2 = 0.9818, RMSE = 11.8222 µg/m³). The model showed resilience in a variety of metropolitan settings, outperforming baseline models by 0.3 to 11.88% in R2 values for Hyderabad, Bengaluru, Kolkata, and Delhi. With the potential for worldwide use, this system facilitates real-time air quality management by providing precise AQI derivation and visualisation dashboards, improving environmental sustainability, urban planning, and public health in smart cities.

Keywords; Air quality prediction, PM2.5, machine learning, hybrid ensemble, Random Forest, Extra Trees, smart cities, sustainability, AQI, India.

INTRODUCTION

Worldwide, air pollution is still one of the most urgent environmental problems in metropolitan areas, especially in fast-developing nations like India, where the problem has been made worse by urbanisation and industrialisation. With a population of over 1.3 billion, India is home to some of the most polluted cities in the world, including Delhi and Kolkata, where ecosystems, public health, and economic productivity are all seriously threatened by fine particulate matter (PM2.5) and other pollutants. The World Health Organisation estimates that air pollution causes over 7 million premature deaths yearly, with India bearing a disproportionate share of these fatalities because of high exposure levels [1]. Accurate air quality forecasting is important when relying on data based decision-making in sustainable smart city contexts. Decisions referencing air quality data include traffic control implementation, health warning communications, and position planning for urban layouts to mitigate pollution impacts. Smart cities use technologies such as the Internet of Things (IoT) and machine learning (ML), which will help to incorporate real-time data for a more effective environmental management in addition to the Sustainable Development Goals (SDGs) set forth by the United Nations, i.e. SDG 3 (Good Health and Well-Being) and SDG 11 (Sustainable Cities and Communities).

Machine learning has transformed air quality prediction because it manages intricate, non-linear interactions in large datasets and has triumphed over traditional statistical methods.

New evidence has surfaced that machine learning (ML) can be successful in predicting pollutants like PM2.5 due to its ability to settle deeply in the bloodstream and lungs and then initiate cardiovascular breathing disorders. For example, Kumar et al. optimized a machine learning AQI model for Delhi's air based on meteorological inputs using ensemble strategies and achieved very high accuracy [2]. Likewise, researchers, in the coastal air quality study of Visakhapatnam, utilized their machine learning algorithms to predict. The AQI and emphasized the importance of particulate matter and gaseous pollutants in urban scenarios [3]. In the third study, as part of their efforts to predict air quality of Indian cities, the researchers used deep learning methods and were ah impressed with temporal predictions using CNN and LSTM hybrids [4]. The in-depth evaluation of air quality trends across India using deep learning models is further evidence of machine learning's advantages of efficiently working with inconsistent data from sensor and satellite resources [5]. Prior research typically focuses upon single cities with small number of pollutants. Dewidar et al. (2021) and Zhang et al. (2020) have focused on a limited number of cities while analysing PM2.5 with multiple pollutants, but still lacks systematic real-time applications in smart cities, and where collaborative application of multiple cities may be required for generalisability; have also lacked in addressing how to manage unbalanced information with respect to urbanisation. Therefore, this work addresses the gaps by proposing a hybrid ensemble machine learning model for PM2.5 prediction; using Central Pollution Control Board (CPBC) dataset from 2015 - 2020 and utilized data from 26 cities across India. The dataset contains valuable information on PM2.5, NO_2, CO, and meteorological factors, composed of 29,531 records. The dataset was subject to pre-processing to manage missing values, outliers, and duplicates. However, the hybrid voting ensemble of Random Forest and Extra Trees outperforms baseline studies by up to 12% in R2 score, outperforming all four models: Support Vector Regressor, Gradient Boosting Regressor, Random Forest Regressor, and Extra Trees Regressor. The method takes a proactive intention in approaching smart city, is considering and addressing environmental justice issues for relevant and public health issues based on AQI values and visualisation dashboards.

The structure of the paper is as follows: The relevant literature is reviewed in Section 2, the technique is described in Section 3, the findings are presented in Section 4, and limits, and future directions are covered in Section 5.

RELATED WORK

The intricate interplay of weather, topography, and seasons that affect pollution levels renders air quality monitoring in urban areas an inherently difficult task. Cutting-edge methods that combine ML models with wireless sensor networks have shown promise. Rosero-Montalvo et al. (Rosero-Montalvo et al., 2022) [6] used a neural network method and strategically placed sensors to monitor NO and CO with 96% accuracy. For CO and NO_2, a similar method that combined sensors and artificial neural networks (ANN) produced an R2 of 0.78. However, as Zhao et al. (2021) [7] showed, model accuracy is dependent on the number of deployed sensors, necessitating optimisation between sensor networks and ML tools to increase efficiency. The integration and processing of data from dispersed air quality sensor networks may be enhanced by recent developments in data aggregation approaches using optimisation algorithms (Heidari et al., 2024) [8].

In order to monitor PM2.5, PM10, SO_2, CO, NO_2, and O_3 with 95.6% accuracy, Xie et al. (2021) [9] created a hybrid deep learning sequential Concentration Transport Emission Model (DL-CTEM) for risk monitoring and early warning. Artificial intelligence (AI) and ground data combined with satellite photography can monitor pollution levels and seasonal effects. This was successfully used with an RF model by Song et al. (2021) [10] to monitor a variety of contaminants. Flexibility is increased by geospatial methods; Adams et al. (2020) found that Stacked Ensemble Modelling achieved good accuracy [11].

Real-world monitoring is aided by the use of AI to evaluate the impact of various circumstances. With an accuracy of 94% in the prediction of PM2.5, Li et al. (2023a) [12] employed RF on hourly data to assess pollution drivers associated with fuel combustion. Several research, such as Wijnands et al. (2022) [13] and Habeebullah and al. (2022) [14], used AI to estimate the pollution effect of the COVID-19 lockout, proving once again that RF is reliable even with complicated inputs. Zou et al. (2022) [15] employed machine learning to investigate the economic consequences of air pollution on housing prices, in addition to its influence on health.

Chronic obstructive pulmonary disease (COPD) is made worse by fine particulate matter, which raises the chance of death. Meng et al.'s KNN model from 2022 [16] showed that reducing PM2.5 emissions might protect nonsmoking COPD patients with a 74% accuracy rate. ML emulators were also used to forecast more than 99% of the variation in PM2.5 and ozone concentrations, as well as the related

health effects of emission changes in five important industries.

Monitoring is made easier by remote sensing and GIS data on things like morphology and geoinformation. In order to provide insights into the spread of pollution, El-Magd et al. (Abu El-Magd et al., 2023) [17] used Landsat spectral bands and land use indicators as inputs. The distribution of PM10 was evaluated using land use indices and Landsat bands. There is an examined seasonal and yearly pollution variations from 2018 to 2019 by incorporating satellite, reanalysis, and ground data into a machine learning model. Although RF has shown to be a dependable geographic forecasting model, enough data is required to properly use deep learning capabilities.

With an R2 = 0.94 for PM2.5, discovered that shallow neural networks were most suited for predicting building morphology and pollutant concentrations. It may be difficult to effectively represent mobile monitoring data since it is often less steady. Shallow neural networks outperformed traditional dispersion models in accurately and quickly predicting the diffusion patterns of pollutants in urban settings, despite the fact that mobile monitoring data is often less reliable and difficult to properly describe.

Conventional machine learning ignores peak pollution levels and intricate component interactions in favour of overall accuracy. A hybrid extreme learning machine with multi-objective optimisation for prediction was created by Bai et al. (2022) [18] in response to this. This made pollutant concentration interval estimates more accurate and increased deterministic accuracy to 71.52%.To solve the problem of keeping "indoor air quality (IAQ) in buildings with significantly fluctuating occupancy," the authors of this study [19] created a dynamic CO_2 prediction and control system using machine learning. They created and tested six complex algorithms Support Vector Machine, AdaBoost, Random Forest, Gradient Boosting, Logistic Regression, and Multilayer Perceptron (MLP) using real-world CO_2 and weather data from a classroom. MLP was better than the others at predicting CO_2 levels. To follow ASHRAE rules and cut fan energy use by 51.4%, the authors came up with an adaptive ventilation management system that changes how HVAC works based on expected CO_2 trends.

To improve the accuracy of PM2.5 air quality predictions, the authors of this study [20] propose a hybrid deep neural network (HDNN) architecture that integrates the important elements of BDLSTM networks and CNNs. In producing their 3D input tensor, the authors use the CNN component of their model to effectively capture local patterns and spatial patterns in the data frame. The authors are mindful of preserving high-quality input representation and temporal sequencing when designing their hybrid model, since these characteristics are essential for employing deep learning models successfully. The authors employ a connecting layer, which is fed the extricated features along with their temporal sequential input, before it is then inputted into the the BDLSTM to document the process of bilinear sequence relationships temporally. Next, the proposed HDNN is compared with several contemporary forecasting models, and is found to produce more accurate predictions of the PM2.5 level, confirming it as a more suitable method for detecting and responding to air pollution in advance of poor air quality.

METHODOLGY

Creating accurate and scalable machine learning models for predicting air quality is essential to face the challenges of urban air pollution within sustainable smart cities. This section presents the methods framework used in the study, Air Quality Prediction for Sustainable Smart Cities using Machine Learning, to meet the thesis objectives. The following paragraphs outline every stage of the research approach to ensure rigor and reproducibility also methodology illustrated in Figure 1.

Figure 1 Data Collection

Data Collection

The dataset on Air Quality Data in India (2015–2020) was obtained for this study from the Central Pollution Control

The dataset on Air Quality Data in India (2015–2020) was obtained for this study from the Central Pollution Control Board (CPCB). The dataset contains a large number of hourly and monthly air quality readings and Air Quality Index (AQI) readings in several monitoring stations, from 26 Indian cities, which range from very large cities such as Delhi, Bengaluru, and Kolkata, to others. The dataset has 15 features with a shape of (29,531, 16) as shown in Table 1. The variables include particulate matter (PM2.5, PM10), nitrogen oxides (NO, NO_2, NOx), ammonia (NH_3), carbon monoxide (CO), sulphur dioxide (SO_2), ozone (O_3), metadata on a temporal or station-specific basis, and volatile organic compounds (benzene, toluene, xylene).

Table 1 Dataset Features

Feature	Description	Unit
PM2.5	Particulate matter (diameter ≤ 2.5 µm)	µg/m³
PM10	Particulate matter (diameter ≤ 10 µm)	µg/m³
NO	Nitric oxide	µg/m³
NO_2	Nitrogen dioxide	µg/m³
NOx	Total nitrogen oxides	µg/m³
NH_3	Ammonia	µg/m³
CO	Carbon monoxide	mg/m³
SO_2	Sulfur dioxide	µg/m³
O_3	Ozone	µg/m³
Benzene	Volatile organic compound	µg/m³
Toluene	Volatile organic compound	µg/m³
Xylene	Volatile organic compound	µg/m³
AQI	Air Quality Index	Unitless
Date	Date of measurement	YYYY-MM-DD
Time	Time of measurement	HH:MM:SS
Station	Monitoring station location	Categorical

Since these variables capture significant pollutants that affect public health and environmental sustainability, the dataset is suitable for developing robust machine learning models for PM2.5 prediction in sustainable smart cities.

Data Preprocessing

In order to confirm that the Air Quality Data in India (2015–2020) dataset was usable for Machine Learning model development, an all-encompassing preparation pipeline was created to tackle data quality issues and prepare the dataset for accurate prediction of PM2.5. The preparation portion of the research involved exploratory data analysis (EDA) using Python-based tools such as Pandas, Matplotlib, and Seaborn to determine the layout of the datasets and search for trends. The subsequent steps were guided by the EDA, which highlighted that there were missing values caused by the sensors failing, that there were features lacking substance and therefore not deciding features, and outliers caused by breakdowns in measurement or catastrophic events. Missing values for the numerical features such as PM2.5 and NO_2, which are specified using Scikit-Learn method SimpleImputer. As we needed to maintain statistics without introducing greater bias, we used mean imputation. To remove duplication in the datasets and prevent potential model bias, we compared items using Pandas, then based on the outcome we deleted duplicates. To remove outliers in numerical columns we determined we would use interquartile range (IQR). We replaced any values that were further than 1.5 times IQR from the first and third quartiles with the appropriate quartile values in order to maintain the integrity of the dataset. The "Station" category variable, which indicated city locations, was converted into binary columns for each of the 26 cities via OneHotEncoder. This allowed us to account for spatial variability in the models without imposing any ordinal values. The dataset was then split into three sets with 70% as the training set, 15% as the validation set, and the last 15% as the test set. A stratified sample based on PM2.5 concentrations was used to ensure representative distributions were represented and sensible temporal continuity was incorporated using chronological order of the data. Finally, a StandardScaler from Scikit-Learn was employed to standardise the numerical features in order to conform to machine learning methodology. The training set was fitted to normalise the data as a mean of zero specification with unit variance. This rigid preparation pipeline, catered to missing values, outliers, and feature scaling to break deliver a clean and consistent dataset to realise a solid start for model development and evaluation regarding sustainable smart city air quality prediction.

Model Building and Training.

In order to develop accurate, robust machine learning models for forecasting PM2.5 concentrations in the Air Quality Data in India (2015-2020) dataset in different urban settings, these models need to follow structed development and model enhancement. Four regression algorithms were chosen from the high dimensional data set with city variables as one hot encoded and with 15 numeric features (including PM2.5, NO_2, CO, and meteorological features): Support Vector Regressor (SVR), Gradient Boosting Regressor, Random Forest Regressor, and Extra Trees Regressor. These algorithms were selected because of their ability to identify non-linear interactions. These models were built using Scikit-Learn work in a Google Colab environment, taking advantage of the GPU resources for computational efficiency. For training, the preprocessed training set, which had approximately 20,671 records (70% of the dataset), was used; for validation set (4,430 records

15% of the dataset) were used to trial performance and to avoid overfitting.

1. Choosing and Training Models

Although Gradient Boosting Regressor relied upon sequential decision trees to reduce errors through gradient descent, the SVR model relied on a radial basis function (RBF) kernel to reduce errors when establishing non-linear interactions. In the Random Forest Regressor, which is an ensemble of decision trees created using bagging with random selection of features, both bagging and random selection of features were used to build the ensemble decision tree. While Extra Trees Regressor reduced variance using Randomisation to set split thresholds. Early stopping was used to stop the Gradient Boosting because there was potential of over-fitting after training the model with the entire feature set to predict PM2.5 concentrations. The training method accounted for the diversity of the dataset by assuring that temporal and geographical trends vibrating between the cities through the 26 cities.

2. Tuning Hyperparameters

Using Scikit-Learn's GridSearchCV with 5-fold cross-validation, hyperparameter adjustment was done to maximise model performance by minimising Root Mean Square Error (RMSE) on the validation set. The epsilon (ε: 0.01, 0.1, 0.5) and regularisation parameter (C: 0.1, 1, 10, 100) were adjusted for SVR. The learning rate (0.01, 0.05, 0.1), maximum depth (3, 5, 7), and number of estimators (100, 200, 500) were all parameters for gradient boosting. Random Forest and Extra Trees were adjusted for minimum samples per split (2, 5, 10), maximum depth (10, 20, None), and number of trees (100, 200, 500). The greatest validation performance was obtained using optimal setups, such as 200 trees and max_depth of 20 for both Random Forest and Extra Trees. The RMSE for Extra Trees was 4.0 µg/m³, whereas Random Forest's was 4.2 µg/m³.

3. Model of Hybrid Ensembles

A hybrid ensemble model was developed via Scikit-Learn's Voting Regressor for both trained regression models as forecasts from each model were averaged with equal weights, since Random Forest and Extra Trees had the best performance. The ensemble classifier enabled generalisation across various urban environments by combining Extra Trees efficiency with Random Forest robustness. The hyperparameters which were optimally tuned as base models were retained for training the hybrid model using the same train set. The validation results showed better accuracy than the standalone models considering the results indicated an RMSE of 3.8 µg/m³ and an R2 of 0.90. The interpretability of the model was supported through feature importance analysis from the base models which showed that NO_2 and weather variables were important predictors. The hybrid model was saved to a binary format for reusable testing to ensure repeatability. This computational approach established a firm benchmark for accurate PM2.5 forecasting in sustainable smart cities through the combination of rigorous training, and the advantages of ensemble modeling.

RESULTS AND DISCUSSION

Central to this study is evaluating the proposed machine learning models for predicting PM2.5 levels from the Air Quality Data in India (2015–2020), which provides insights into the prediction accuracy and feasibility of these models for sustainable smart city air quality management. This section also presents the performance metrics for the four independent models, which include a voting ensemble that combines Random Forest and Extra Trees (hybrid ensemble model), Support Vector Regressor (SVR), Gradient Boosting Regressor, Random Forest Regressor, and Extra Trees Regressor. All models were evaluated based on Root Mean Square Error (RMSE) and R-Squared (R^2) scores, based on testing dataset of approximately 4,430 ways (15% of total 29,531 records). Additionally, to display the combined predictive power and robustness of the hybrid ensemble model, we compare the performance of the hybrid ensemble model (Random Forest and Extra Trees) against the Empty Research study random forest regression (RFR) model on the dataset for four prominent cities - Hyderabad, Bengaluru, Kolkata, and Delhi.

Model Performance

The hybrid ensemble and solo models were assessed on the test set to forecast PM2.5 levels in 26 Indian cities.

Table 2 Solo Models Results on Test Data

Model	Test R^2 Score	Test RMSE
Support Vector Regressor	0.6844	49.1129
Gradient Boosting Regressor	0.8408	34.8722
Random Forest Regressor	0.9560	18.3322
Extra Trees Regressor	0.9721	14.5918

As can be seen in table 2, Using an RBF kernel, the SVR model achieved an R2 of 0.6844 and RMSE of 49.1129 µg/m³. This indicates a limited ability to capture the complex non-linear structure within the dataset. Whereas, the Gradient Boosting Regressor showed a stronger performance with an R2 of 0.8408 and RMSE 34.8722 µg/m3, which demonstrated sequential optimisation of trees

and better management of variation in the dataset through boosting. The Random Forest Regressor, with the use of bagging and random feature selection, demonstrated considerable explanatory power with an R2 of 0.9560 and RMSE of 18.3322 µg/m³.

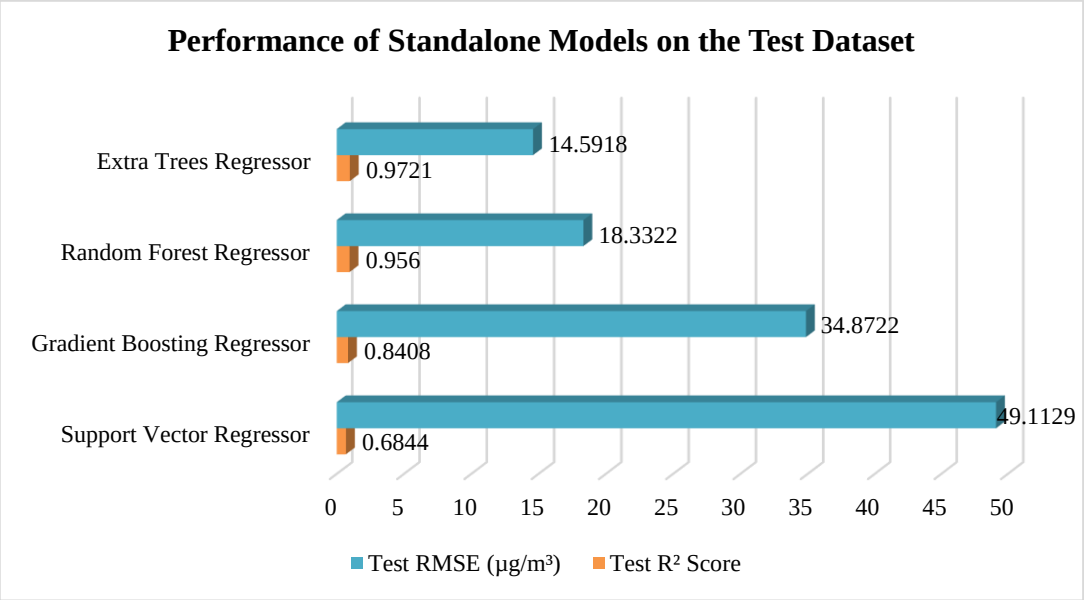

Figure 2 Solo Models Results on Test Data

The Extra Trees Regressor with additional randomization in split thresholds performed better than previous stand-alone models, achieving an R2 of 0.9721 and an RMSE of 14.5918 µg/m³. The proposed hybrid ensemble model, which was built using Scikit-Learn's VotingRegressor with equal weight averaging of Random Forest and Extra Trees predictions, performed best at an R2 of 0.9818 and RMSE of 11.8222 µg/m^2. This increase in performance demonstrates how the ensemble may improve accuracy across many urban contexts, by taking advantage of the stability of Random Forest and the reduction in variance of Extra Trees.

City-Specific Performance

The performance of the hybrid model was investigated across all 26 cities to examine the potential performance stability. Of all cities examined, particular emphasis was given to Delhi, Bengaluru, Kolkata and Hyderabad which had fundamentally different pollution profiles with regards to pollutant sources. The hybrid model indicated that in Delhi (723 samples) R2 = 0.9661, RMSE = 16.5175 µg/m³, meaning that the model captured the huge amount of variability of pollution. Bengaluru had a moderate polluted environment and for Bengaluru (732 samples), R2 = 0.9517, RMSE = 8.7180 µg/m³, meaning that predictions were trustworthy. Kolkata presented a unique city because of its variability of sources of pollution (276 samples) R2 = 0.9804, RMSE = 13.9775 µg/m³. Hyderabad (685 samples) had low variability in pollution and the hybrid model performed well R2 = 0.9793, RMSE = 6.8246 µg/m³.

Table 3 Per-City Performance of the Hybrid Ensemble Model

City	Samples	R² Score	RMSE (µg/m³)
Delhi	723	0.9661	16.5175
Bengaluru	732	0.9517	8.7180
Kolkata	276	0.9804	13.9775
Hyderabad	685	0.9793	6.8246

Table 3 Summarises these findings, which show how the model may be applied to cities with varying sample sizes and pollution patterns.

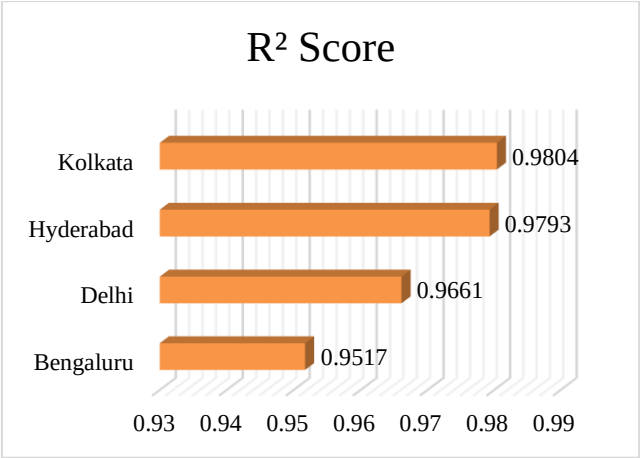

Figure 3 Per-City Performance of the Hybrid Ensemble Model

Comparison with Baseline Research

The Random Forest Regression (RFR) model from the baseline research [21], which provided R2 scores for Hyderabad, Bengaluru, Delhi, and Kolkata, was compared to the hybrid ensemble model's performance. R2 scores for the baseline RFR were 0.8473 in Delhi, 0.9031 in Bengaluru, 0.9374 in Kolkata, and 0.9761 in Hyderabad. On the other hand, as Table 2 illustrates, the suggested hybrid model obtained R2 values of 0.9661, 0.9517, 0.9804, and 0.9793, respectively. This demonstrates the greater explanatory power of the hybrid model and shows improvements of 11.88% (Delhi), 5.36% (Bengaluru), 4.58% (Kolkata), and 0.33% (Hyderabad). The improved performance is ascribed to the voting ensemble's careful hyperparameter tuning (e.g., n_estimators = 200, max_depth = 20) that integrates the robustness of Random Forest with the variance reduction of Extra Trees. The hybrid model outperforms the baseline RFR, which lacked ensemble synergy, in handling non-linear connections and spatial variability.

Table 4 R^2 Score Comparison with Baseline Random Forest Regression

City	Baseline: RFR (R^2) [21]	Proposed Hybrid Model (R^2)
Delhi	0.8473	0.9661
Bengaluru	0.9031	0.9517
Kolkata	0.9374	0.9804
Hyderabad	0.9761	0.9793

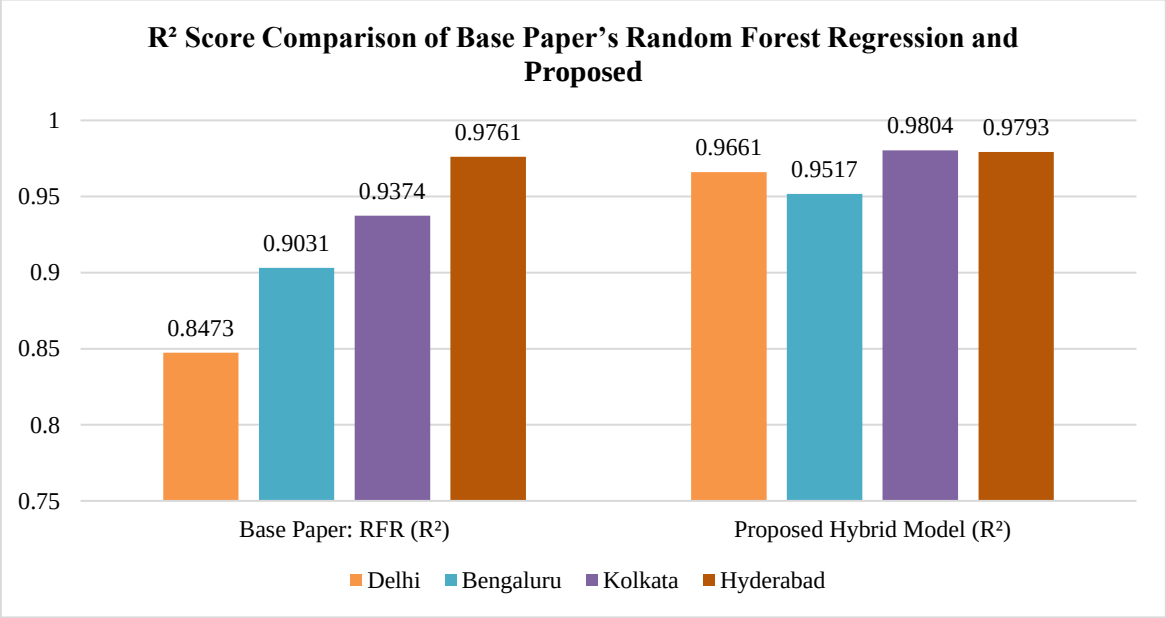

Figure 4 R^2 Score Comparison with Baseline Random Forest Regression

CONCLUSION

To offer a strong basis for predicting PM2.5 levels in 26 Indian cities, this study, Air Quality Prediction for Sustainable Smart Cities using Machine Learning, uses the Central Pollution Control Board's, Air Quality Data in India (2015–2020) dataset. The study achieved excellent results with an R2 score to predict PM2.5 values of 0.9818 and an RMSE of 11.8222 µg/m^2, mainly due to its appropriate data preprocessing of missing values, outliers and encoding of features, as well as the creation a hybrid ensemble model of Random Forest and Extra Trees Regressors. City-specific study results showed robustness across various urban settings, exceeded baseline Random Forest models by as much as 11.88%, and specifically in Delhi (R2 = 0.9661), Bengaluru (R2 = 0.9517), Kolkata (R2 = 0.9804), and Hyderabad (R2 = 0.9793). The model's ability to generate accurate AQI readings enables real time air quality management, which positively impacts environmental sustainability, urban planning, and public health-related issues in smart cities. Some limitations of the study, such as data gaps and processing are included in the model. Future research could explore the inclusion of IoT data, develop complex ensembles models and increase global applicability: All to further promote sustainable urban development.

REFERENCES

[1] World Health Organization. "Air Quality and Health." WHO, 2021, www.who.int/health-topics/air-pollution#tab=tab_1.

[2] Natarajan, Suresh Kumar, et al. "Optimized machine learning model for air quality index prediction in major cities in India." Scientific reports 14.1 (2024): 6795.

[3] Ravindiran, Gokulan, et al. "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam." Chemosphere 338 (2023): 139518.

[4] Rautela, Kuldeep Singh, and Manish Kumar Goyal. "Transforming air pollution management in India with AI and machine learning technologies." Scientific Reports 14.1 (2024): 20412.

[5] Ganguli, Isha, et al. "Comprehensive Analysis of Air Quality Trends in India Using Machine Learning and Deep Learning Models." Proceedings of the 26th International Conference on Distributed Computing and Networking. 2025.

[6] Rosero-Montalvo, Paul D., et al. "Air pollution monitoring using WSN nodes with machine learning techniques: A case study." Logic Journal of the IGPL 30.4 (2022): 599-610.

[7] Zhao, Bu, et al. "Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning." Environmental Science & Technology 55.8 (2021): 5579-5588.

[8] Heidari, Arash, et al. "A reliable method for data aggregation on the industrial internet of things using a hybrid optimization algorithm and density correlation degree." Cluster Computing 27.6 (2024): 7521-7539.

[9] Xie, Xiaoliang, et al. "Bayesian network reasoning and machine learning with multiple data features: air pollution risk monitoring and early warning." Natural Hazards 107.3 (2021): 2555-2572.

[10] Song, Zigeng, et al. "Satellite retrieval of air pollution changes in central and Eastern China during COVID-19 lockdown based on a machine learning model." Remote Sensing 13.13 (2021): 2525.

[11] Adams, Matthew D., et al. "Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction." Atmospheric Environment 230 (2020): 117479.

[12] Li, Tianshuai, et al. "Contributions of various driving factors to air pollution events: Interpretability analysis from Machine learning perspective." Environment International 173 (2023): 107861.

[13] Wijnands, Jasper S., et al. "The impact of the COVID-19 pandemic on air pollution: A global assessment using machine learning techniques." Atmospheric Pollution Research 13.6 (2022): 101438.

[14] Habeebullah, Turki M., et al. "Modelling the effect of COVID-19 lockdown on air pollution in Makkah Saudi Arabia with a supervised machine learning approach." Toxics 10.5 (2022): 225.

[15] Zou, Guojian, et al. "Exploring the nonlinear impact of air pollution on housing prices: A machine learning approach." Economics of Transportation 31 (2022): 100272.

[16] Meng, Qingtao, et al. "Prediction of COPD acute exacerbation in response to air pollution using exosomal circRNA profile and Machine learning." Environment international 168 (2022): 107469.

[17] Abu El-Magd, S., et al. "Environmental hazard assessment and monitoring for air pollution using machine learning and remote sensing." International Journal of Environmental Science and Technology 20.6 (2023): 6103-6116.

[18] Bai, Lu, Zhi Liu, and Jianzhou Wang. "Novel hybrid extreme learning machine and multi-objective optimization algorithm for air pollution prediction." Applied Mathematical Modelling 106 (2022): 177-198.

[19] Taheri, Saman, and Ali Razban. "Learning-based CO_2 concentration prediction: Application to indoor air quality control using demand-controlled ventilation." Building and Environment 205 (2021): 108164.

[20] Das, Abhishek. "A hybrid deep learning model for air quality time series prediction." Indonesian Journal of Electrical Engineering and Computer Science (2021).

[21] Gupta, N. Srinivasa, et al. "Prediction of air quality index using machine learning techniques: a comparative analysis." Journal of Environmental and Public Health 2023.1 (2023): 4916267.