IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 14, Issue 11, November 2024, ||Series -1|| PP 26-41

Air Quality Prediction and Pollution Control Strategies for Delhi Using Machine Learning Models

Radhika Gupta*, Shruti Tripathi*
** Department of Artificial Intelligence and Data Sciences, Indira Gandhi Delhi Technical University for Women*
Received 23 November 2024; Accepted 03 December 2024

Abstract- Air quality index (AQI) is a key metric for reporting air quality. AQI of most urban cities has increased due to fossil fuels, urbanization, industrialization, and vehicular pollution. Fine particulate matter and harmful gases significantly affect the quality of air, which is very dangerous for human beings, animals, and all other living organisms. AQI analysis helps government and official bodies take proper measures to curb air pollution. AQI can be derived using various techniques and mathematical models. This paper aims to propose multiple machine learning models and perform a comparative analysis to identify the most effective machine learning model to compute AQI in a city like Delhi, which recently entered the grade 4 city ranking. Several machine learning models are used – Linear Regression, Decision Tree, Random Forest, Support Vector Regressor (SVR), and K-Nearest Neighbors (K-NN). Each Model is evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R2 Score, and Mean Absolute Percentage Error (MAPE). Among the models, Random Forest Regressor achieves the lowest RMSE of 0.29 and the highest R^2 score of 93.4%. It outperforms other models - SVR shows an RMSE of 0.48 with an R^2 score of 79.2%, and K-NN produces an RMSE of 0.54 with an R^2 score of 74.6%. Decision Tree Regressor and Linear Regression show higher RMSE values of 0.51 and 0.63, respectively, with lower R^2 scores of 76.8% and 69.5%. The results conclude that Random Forest Regressor provides the most accurate AQI predictions for New Delhi. These findings demonstrate the potential of advanced regression models to support air quality management efforts and mitigate public health risks.

The findings suggest that stronger models of machine learning, such as Random Forest, could be exploited to make more accurate AQI predictions and also for real-time monitoring of pollution. Such an informed insight might enable the policymakers to design data-driven interventions like implementing stricter vehicular emission norms, promoting cleaner energy alternatives, or establishing a dynamic traffic management system in densely polluted areas. Application of predictive analytics might help governmental organizations conduct air quality control strategies more effectively when risks are at peak level and prevent adverse effects on public health along with improved general quality of air in urban surroundings.

Index Terms- Air Quality Index, Machine Learning, New Delhi Pollution, Random Forest Regressor, Root Mean Square Error

I. INTRODUCTION

Air is essential for human survival. Air quality must be constantly monitored and checked to ensure a suitable breathing atmosphere for living beings. Air pollution causes respiratory diseases such as emphysema, asthma, chronic obstructive pulmonary disease (COPD), and other breathing problems, especially in urban areas with high air pollution levels. Scientific evidence shows that air pollution is the most critical environmental risk for human sustenance. The decline in quality of air can be attributed to rapid urbanization, rampant industrialization, and unchecked vehicular pollution. Human and animal health are compromised due to exposure to poisonous substances in the air. Air quality index (AQI) is the index for reporting air quality. The AQI is calculated using the following 12 parameters: PM2.5 (particulate matter with a diameter of 2.5 microns or less), PM10 (particulate matter with a diameter of 10 microns or less), NO (Nitric Oxide), NO2 (Nitrogen Dioxide), NOX (Nitrogen Oxides), NH3 (Ammonia), CO (Carbon Monoxide), SO2 (Sulfur Dioxide), 03 (Ozone), Benzene, Toluene, Xylene. The concentration of each of the mentioned pollutants is reported. AQI is then calculated using the average concentration of each pollutant over a standard time of twenty-four hours. All 12 pollutants are constantly monitored over every location.

Delhi is one of the world's most polluted cities and often witnesses hazardous air quality levels, particularly at the peak of winter-time crop stubble burning, vehicular emissions, and industrial activities. Its massive population of over 30 million makes it easy for poor air quality to amplify its impact, and most vulnerable in this situation are children and the elderly. Exposures to large concentrations of PM2.5 and PM10 for long periods are considered to increase respiratory diseases, cardiovascular diseases, and premature mortality. In Delhi,

Delhi recorded more than 200 days of AQI classified as "poor" or worse in 2023, and therefore requires urgent interventions.

A high AQI indicates severely contaminated air, which is very dangerous for human beings, animals and all other living organisms. The dataset used (source : https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india) contains AQIs and pollutant records over five years for the state of New Delhi, India. Five regression analysis techniques will be used, and the best results and accuracy will be determined through comparison.

In this proposed work, multiple machine learning regression techniques are applied, and the AQI level in New Delhi will be predicted. This includes a Linear Regression model, Decision Tree, Random Forest, Support Vector Regressor(SVR), and K- Nearest neighbors(K-NN). The models' results and their accuracy will be compared to identify the best model for AQI prediction. The study further shows how each kind of pollutant impacts the air quality, and then jumps into discussing the scope in which machine learning can be used to predict future levels of pollution. All these strategies ultimately aim to help in developing better methods to measure, monitor, and hence mitigate air pollution in critical urban areas.

For forecasting the Air Quality Index, there are large data problems with complex nonlinear relationships among the various types of pollutants. Traditional approaches generally fail due to data complexity and resulting inaccuracies in predictions. Machine learning may solve the vast data processing, determine hidden patterns, and serve accordingly for good predictions. Advanced models, like Random Forest and Support Vector Regressors (SVR), besides enhancing the accuracy of AQI forecasts can also deliver actionable insights regarding the relative contributions of different pollutants in order to pursue better-targeted strategies of pollution control.

II. LITERATURE SURVEY

Air pollution is a critical concern that affects the health of both humans and ecosystems, especially in urban environments. Accurate prediction of the Air Quality Index (AQI) is essential for monitoring and measuring pollution to implement effective control measures. Machine learning models have shown significant potential in this domain by analyzing various pollutants and their concentrations. Liu et al. [1] focused on predicting AQI and pollutant concentrations using machine learning algorithms. Their research highlights the effectiveness of machine learning models in handling complex air quality data and achieving accurate predictions. Castelli et al. [2] applied machine learning to predict air quality in California and concluded that these models can provide reliable forecasts to aid in proactive air quality management.

Mani et al. [3] combined regression and ARIMA time series models to predict AQI in Chennai. The research demonstrated the potential of integrating statistical and machine learning methods, which improved accuracy. Kottur and Mantha [4] employed a hybrid approach using Artificial Neural Networks (ANN) and Kriging methods to forecast air pollutants, enhancing prediction accuracy. Halsana [5] presented an air quality prediction model using supervised machine learning algorithms. The authors evaluated various models to highlight their strengths and limitations. Soundari et al. [6] analyzed air quality in India using multiple regression techniques, providing insights into factors influencing air quality. Aditya et al. [7] focused on detecting and predicting air pollution using real-time data, emphasizing its importance in improving prediction accuracy. Kleine Deters et al. [8] modeled urban PM2.5 pollution using machine learning and selected meteorological parameters. The authors demonstrated their models' ability to effectively capture relationships between pollutants and meteorological factors.

Bhalgat et al. [9] compared machine learning algorithms for air quality prediction and found that ensemble methods like Random Forest performed best. Bansal [10] utilized Long Short-Term Memory (LSTM) networks to predict AQI in Delhi, showcasing the advantages of deep learning models for time-series forecasting. Shishegaran et al. [11] developed a nonlinear ensemble model for predicting air quality in Tehran, proving the effectiveness of combining different machine learning models to predict AQI.

Tuan-Vinh [12] enhanced awareness in sustainable smart cities by integrating lifelog images and IoT air pollution data. This highlighted the potential of IoT and machine learning for real-time AQI prediction. Kumar et al. [13] used IoT and machine learning techniques to predict time-series air quality data by continuous monitoring. Maleki et al. [14] used artificial neural networks to model nonlinear relationships between pollutants. Singh et al. [15] identified pollution sources and predicted urban air quality using ensemble learning. Hansun and Kristanda [16] emphasized feature selection's importance by using the B-wema method for AQI measurement and prediction. Janarthanan et al. [17] applied deep learning for AQI prediction in metropolitan areas.

Londhe [18] demonstrated the effectiveness of machine learning in handling large air quality datasets for accurate predictions. Gore and Deshpande [19] classified health risks based on air quality using machine learning. Zhao et al. [20] developed a temporal-spatial model for AQI prediction, stressing the importance of considering temporal and spatial factors. Chowdhury et al. [21] applied data mining to Dhaka City's air pollution data to

provide actionable insights. Zhou et al. [22] characterized urban air quality using machine learning and provided detailed insights into influencing factors. Srivastava et al. [23] estimated Delhi's air pollution using machine learning. Raturi and Prasad [24] used artificial neural networks for future AQI recognition, enabling continuous monitoring and timely predictions. Mahalingam et al. [25] developed a machine learning model for smart city air quality prediction, which could be integrated into the city's infrastructure.

Sivakumar et al. [26] enhanced underwater acoustic sensor networks' lifetime with a Markov-based approach. Sethi and Mittal [27] estimated ambient air quality using supervised learning. Hajek and Olej [28] predicted the common air quality index for small areas of Czech microregions. Ameer et al. [29] compared machine learning techniques for smart city air quality prediction and concluded that advanced models are more advantageous. Behesht Abad et al. [30] used hybrid algorithms for predicting gas condensate viscosity. Rajabi et al. [31] developed hybrid optimizer algorithms for fracture density prediction to enhance AQI prediction accuracy. Behesht Abad et al. [32] predicted oil flow rates using a robust algorithm. Hasbeh et al. [33] used hybrid computing models for oil formation prediction, showcasing the effectiveness of multi-layer models. Jafarizadeh et al. [34] highlighted the potential of data-driven approaches for pore pressure prediction. Zhang et al. [35] developed a robust pore pressure prediction approach using machine learning, effectively handling data uncertainties. Tabasi et al. [36] optimized models for natural fracture prediction.

Rajabi et al. [37] predicted shear wave velocity using deep and hybrid algorithms, handling large datasets effectively. Beheshtian et al. [38] developed a robust computational approach for safe mud weight prediction, which indirectly contributed to improving AQI prediction accuracy. Masoud et al. [39] showed complex data accuracy by predicting permeability in gas reservoirs using group methods. Mohamadian et al. [40] enhanced drilling fluids with nanoparticles, improving rheological and filtration characteristics. Choubineh et al. [41] improved predictions of wellhead choke flow rates using optimization algorithms. Ghorbani et al. [42] predicted gas flow rates using optimization algorithms, demonstrating the robustness of such models for AQI data. Fan et al. [45] developed a hybrid model for air quality prediction based on data decomposition, showing that hybrid models can effectively improve prediction performance by leveraging data decomposition techniques

Machine Learning Models Used

1) Linear Regression: Linear regression is one of the very basic statistical techniques used for the prediction of a continuous dependent variable using one or more independent variables. Liu et al. [1] demonstrated how linear regression can be applied in their study to fit the air pollution data and predict AQI levels. The simplicity and clarity of this model make it applicable for widespread usage although probably not so effective in capturing complex nonlinear relationships than more advanced models. 2) Decision Tree - Decision trees are a class of supervised learning algorithms that break up data into subsets based on values for the features, and in this way, create a tree-structured representation of decisions. Mani et al. [3] used decision trees to forecast air quality in Chennai; these models can easily manage nonlinear interrelations among variables. However, they are prone to overfitting, which demands careful tuning. Random Forest is an ensemble learning method, which combines multiple decision trees to improve predictive accuracy and significantly prevent overfitting. Bhalgat et al. [9] reported that Random Forest is the best model which was fitted to compare other models on AQI prediction due to robustness to handle big data. Because of this, its performance and reliability make it frequently used in many studies. 4) Support Vector Regressor (SVR) - SVR is the SVM type adapted to regression tasks that find the best fit line inside a threshold value. Castelli et al. [2] applied Support Vector Regression to predict the air quality of California, thus showing its feasibility in working with high-dimensional data and making accurate predictions. K-Nearest Neighbors (K-NN) is another non-parametric method in which the prediction of the outcomes is made using the closest training instances in the feature space. Halsana [5] stressed out the use of K-NN in AQI prediction, stressing the simplicity and effectiveness of the technique in catching local patterns in the data. 6) Long Short-Term Memory (LSTM) - LSTMs, a flavor of RNNT, can be trained with long short-term dependencies, thus make them suitable for time-series forecasting. Bansal [10] has shown the utility that one can avail by training LSTM by predicting AQI in Delhi, also due to their robust capacity to model temporal patterns in AQI data. 7) Hybrid Models: Hybrid models combine multiple machine learning techniques in order to combine their strengths and avoid their weaknesses. In order to prove the potential of combining different machine learning models, Shishegaran et al. [11] proposed a nonlinear ensemble model for air quality prediction. To develop a real-time prediction model for AQI using integrated lifelog images and IoT air pollution data that utilized machine learning, Tuan-Vinh [12] proposed a hybrid model.

Although there has been significant advancement in the application of machine learning for predicting AQI, the majority of studies up to now focus only on general regional analyses rather than city-specific problems. Furthermore, studies have been heavily biased toward annual or seasonal averages that do not even allow for prediction of short-term fluctuations or real-time levels of AQI. Most of the models, especially traditional ones, fail to capture well the nonlinear relationships between multiple pollutants and AQI. Limited work has been done in the context of Delhi to assess relative performance of machine learning algorithms like Random Forest, SVR,

and K-NN for AQI prediction using local datasets, besides in segregation between model outputs and effective strategies for pollution control.

This research would fill up the gaps identified by focusing only on New Delhi and using a cleaned and preprocessed dataset that comprises five years of information related to air quality. This analysis differs from earlier studies as the five different machine learning regression models, namely Linear Regression, Decision Tree, Random Forest, Support Vector Regressor (SVR), and K-Nearest Neighbors (K-NN), would accurately forecast the AQI. Applying overall performance metrics like RMSE, MAE, and R^2 allows making proper comparison. This research only identifies that Random Forest Regressor is the best model but, at the same time, highlights even further how these predictive results guide formulation of policies. The divide here connects technical model efficacy and practical applications with enhancing management of urban air quality.

In summary, the studied papers show the wide applicability and success of machine learning techniques for predicting AQI. The multifaceted approaches, from traditional regression models to advanced ensemble methods and deep-learning algorithms, really emphasize the versatility as well as strength of machine learning in solving complex issues like environmental ones. The introduction of IoT and real-time data has further improved the accuracy in predictive power, requiring insights for decision-making during processes.

III. DATASET DESCRIPTION AND SAMPLE DATA

The link to the dataset used for this work is: https://www.kaggle.com/rohanrao/air-quality-data-in-india. The dataset contained hourly and daily air quality and AQI (Air Quality Index) data from monitoring stations in various Indian cities. Data spans from 2015 through 2020. The original dataset has 29,532 rows and 16 columns. The dataset included the air quality of 26 Indian cities. Those cities are listed below: Ahmedabad, Aizawl, Amaravati, Amritsar, Bangalore, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, and Visakhapatnam.

This study focuses on New Delhi, which is described as an excellent metropolitan area suffering from severe problems due to pollution, and this city is currently ranked Grade 4 regarding its condition of air pollution (as of November 2024). New Delhi's dataset is collected and cleaned with Python libraries in Google Colab. The data set has 176 entries and 15 variables included and represents air quality reviews from different monitoring stations in New Delhi. It encompasses important features like PM2.5, PM10, NO, NO2, CO, SO2, and O3 concentrations as well as the AQI. The AQI_bucket attribute categorizes the air quality into six groups: *Good, Satisfactory, Moderate, Poor, Very Poor, and Severe.*

Dataset Details:

- **Source**: The original dataset was downloaded from Kaggle, contributed by Rohan Rao.
- **Time Frame**: 2015–2020.
- **Cleaning and Processing**:
  - The attribute xylene was excluded as it contained null values for all entries related to New Delhi.
  - The dataset was filtered to include only New Delhi's data.
  - Duplicate and irrelevant records were removed.
  - The final dataset focuses on attributes directly related to pollution measurement and AQI.

Dataset Attributes:

- **Date**: Date of the measurement in YYYY-MM-DD format.
- **City**: Name of the city (in this case, New Delhi).
- **PM2.5**: Particulate Matter (PM) with a diameter of 2.5 micrometers or less.
- **PM10**: Particulate Matter (PM) with a diameter of 10 micrometers or less.
- **NO**: Nitric oxide concentration (µg/m³).
- **NO2**: Nitrogen dioxide concentration (µg/m³).
- **NOx**: Total nitrogen oxides concentration (µg/m³).
- **NH3**: Ammonia concentration (µg/m³).
- **CO**: Carbon monoxide concentration (µg/m³).
- **SO2**: Sulfur dioxide concentration (µg/m³).
- **O3**: Ozone concentration (µg/m³).
- **Benzene**: Benzene concentration (µg/m³).
- **Toluene**: Toluene concentration (µg/m³).
- **AQI**: Air Quality Index (numeric value).
- **AQI_Bucket**: Categorized AQI levels: *Good, Satisfactory, Moderate, Poor, Very Poor, Severe.*

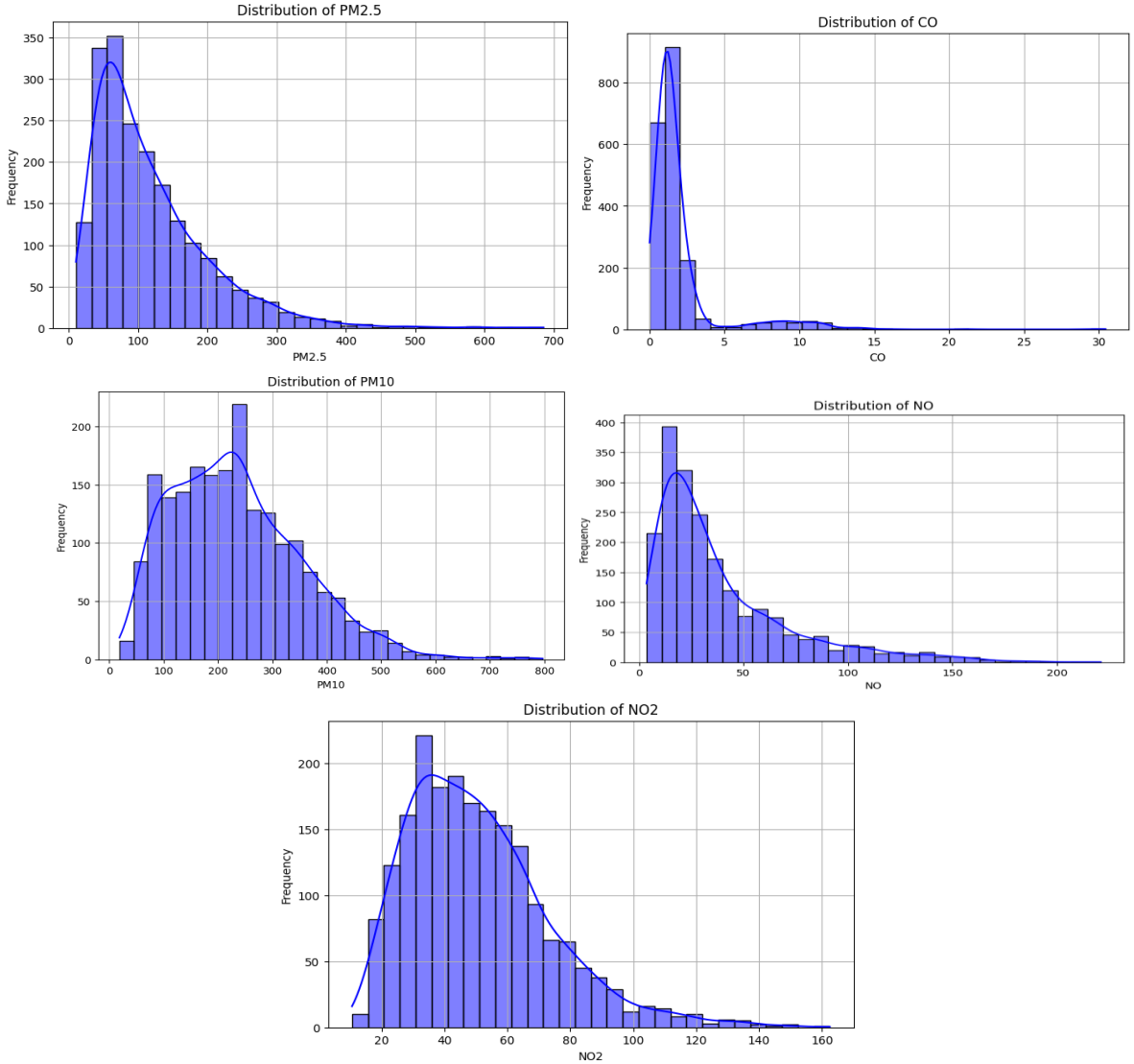

Figure 1: Distribution of key pollutants (PM2.5, PM10, NO2, SO2, and CO) in New Delhi. This representation indicates differences in pollutant concentrations, focusing particularly on the significant contribution of PM2.5 to poor air quality.

This data set gives much-needed information in the analysis of trends in air quality to estimate air pollution levels to derive health implications in New Delhi. It can also aid in predictive models for AQI based on past trends and the concentration of pollutants.
Below is a sample of the cleaned dataset for New Delhi:

Date	City	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	AQI	AQI_Bucket
2020-01-01	New Delhi	55.1	102	40.5	78.4	90.3	30.2	1.2	5.3	65.1	11.2	12.3	320	Severe
2020-01-02	New Delhi	50.2	97	38.1	75.6	85.2	28.4	1.1	4.9	63.2	10.8	11.5	300	Very Poor
2020-01-03	New Delhi	45.3	90	35.2	72.0	80.1	25.0	1.0	4.5	60.0	10.0	10.5	250	Poor
2020-01-04	New Delhi	40.7	85	33.0	68.5	75.0	23.1	0.9	4.3	55.5	9.8	10.0	210	Moderate
2020-01-05	New Delhi	35.1	80	30.0	65.0	70.0	20.5	0.8	4.0	50.5	9.0	9.5	180	Satisfactory

Table 1: Air Quality- Data Feature, New Delhi. The dataset can be shown to incorporate a number of features about pollutant concentrations, AQI values, and class labels.

This dataset allows for the examination of air quality in New Delhi, with a strong focus placed on the primary pollutants responsible for fluctuation in air quality. Data cleaning and balancing were used by the authors as the basis for predictive models to understand trends in air pollution and their health effects in urban settings. Focusing attention on New Delhi, the dataset allows for an in-depth exploration of patterns of pollution. These actionable insights could spur the initiation of developing an urban environmental policy and health interventions. The preprocessing step depicted in Figure 2 is to guarantee data quality. In this respect, standardization and removal of incomplete records are part of the process involve

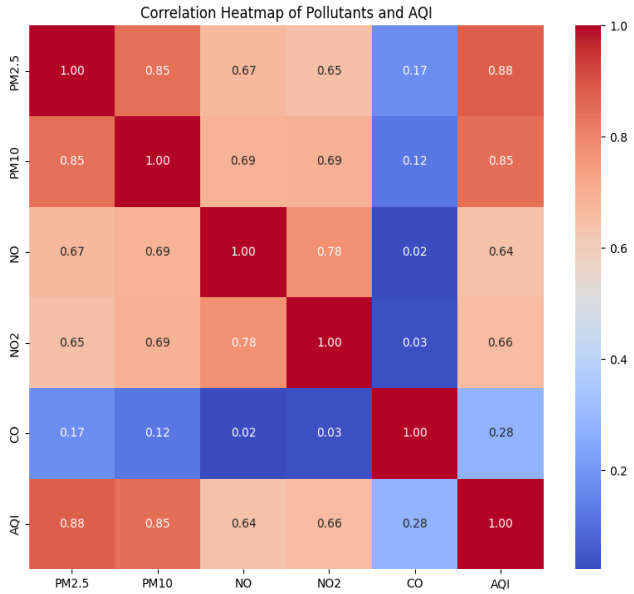

Figure 2: Heatmap of correlation between pollutants and AQI. The plot shows the correlation coefficient corresponding to each pair of pollutants (PM2.5, PM10, NO, NO2, CO) and AQI.

Figure 2 is the heatmap that represents the strength of correlation between major pollutants and AQI. It displayed strong positive correlations for PM2.5 and PM10 with AQI which have correlation coefficients as high as 0.88 and 0.85, respectively, thus indicating their dominant influence on air quality. On the contrary, CO signifies a weak correlation with AQI and therefore has lesser influence.

IV. METHODOLOGY

In this study, five machine learning algorithms was used to compare in one prediction model: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, and KNN in order to predict the AQI for New Delhi regarding different air pollution indicators. These algorithms were used to evaluate these for selecting the best performer in terms of predicting AQI through indicating its quantifying patterns and future control plans for air pollution.

Steps in the Methodology

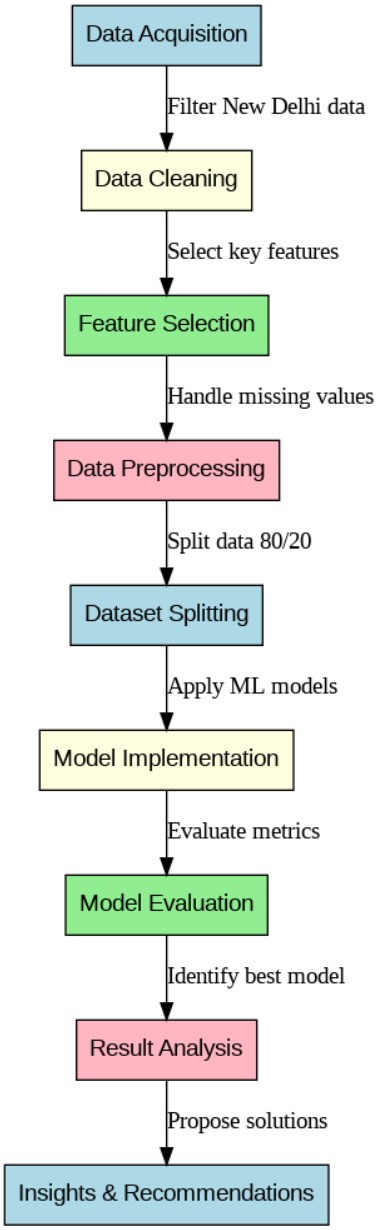

Fig 3 : Study Workflow Overview.

Figure 3: Study workflow. The preprocessing of the data lead to the modeling and finally the visualization.

Step 1. Choosing the dataset

This research used Kaggle information on air quality in New Delhi from the years 2015 to 2020. This dataset contains pollution parameters like PM2.5, PM10, NO2, NOx, NH3, CO, SO2, Ozone, Benzene, and Toluene. The dataset was selected based on the extensive representation of air pollution indices in New Delhi, which is known to have persistently high AQI.

Step 2. Data Preprocessing

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO
0	Delhi	01-01-2015	313.22	607.98	69.16	36.39	110.59	33.85	15.2
1	Delhi	02-01-2015	186.18	269.55	62.09	32.87	88.14	31.83	9.54

2	Delhi	03-01-2015	87.18	131.9	25.73	30.31	47.95	69.55	10.61
3	Delhi	04-01-2015	151.84	241.84	25.01	36.91	48.62	130.36	11.54
4	Delhi	05-01-2015	146.6	219.13	14.01	34.92	38.25	122.88	9.2

	SO2	**O3**	**Benzene**	**Toluene**	**Xylene**	**AQI**	**AQI Bucket**
0	9.25	41.68	14.36	24.86	9.84	472	Severe
1	6.65	29.97	10.55	20.09	4.29	454	Severe
2	2.65	19.71	3.91	10.23	1.99	143	Moderate
3	4.63	25.36	4.26	9.71	3.34	319	Very Poor
4	3.33	23.2	2.8	6.21	2.96	325	Very Poor

Table 2: Raw dataset characteristics before preprocessing. The table includes pollutant levels (e.g., SO2, O3, Benzene), AQI values, and AQI categories before any data cleaning.

	PM2.5	**PM10**	**NO**	**NO2**	**NOx**	**NH3**	**CO**	**SO2**	**O3**	**Benzene**
0	313.22	607.98	69.16	36.39	110.59	33.85	15.2	9.25	41.68	14.36
1	186.18	269.55	62.09	32.87	88.14	31.83	9.54	6.65	29.97	10.55
2	87.18	131.9	25.73	30.31	47.95	69.55	10.61	2.65	19.71	3.91
3	151.84	241.84	25.01	36.91	48.62	130.36	11.54	4.63	25.36	4.26
4	146.6	219.13	14.01	34.92	38.25	122.88	9.2	3.33	23.2	2.8

	Toluene	**Xylene**	**AQI**
0	24.86	9.84	472
1	20.09	4.29	454
2	10.23	1.99	143
3	9.71	3.34	319
4	6.21	2.96	325

Table 3: Cleaned and normalized dataset after preprocessing. This table shows the pollutant levels (e.g., PM2.5, NO2, NH3) and AQI values after data transformation and standardization.

The data was preprocessed to handle missing and erroneous data. Missing values were dealt with through imputation; it used the mean in continuous variables and mode for categorical variables. All the columns that are irrelevant for the AQI-prediction algorithm are being dropped. The data types were standardized for compatibility with machine learning algorithms. The output dataset after the mentioned pre-processing steps is shown in Table 2: missing values, duplicates, and scaling of numerical features. The cleaned data guarantees compatibility with machine learning models and enhances prediction accuracy.

Step	Description	Result
Handling Missing Values	Imputed missing values with mean for pollutants	Complete dataset
Feature Selection	Removed irrelevant columns like Xylene	Focused on 14 attributes
Data Cleaning	Removed duplicates and filtered only New Delhi's data	176 rows retained
Scaling	Standardized numeric features using StandardScaler	Uniform feature scales

Table 4: Preprocessing steps for data preparation. The table summarizes the key preprocessing tasks, their descriptions, and the resulting changes to the dataset.

Step 3. Feature Selection

These key features were chosen: PM2.5, PM10, gases of NO2, SO2, CO since they play an important role in the formation of air quality. Pearson correlation was applied to choose appropriate attributes without redundant ones in the model building process.

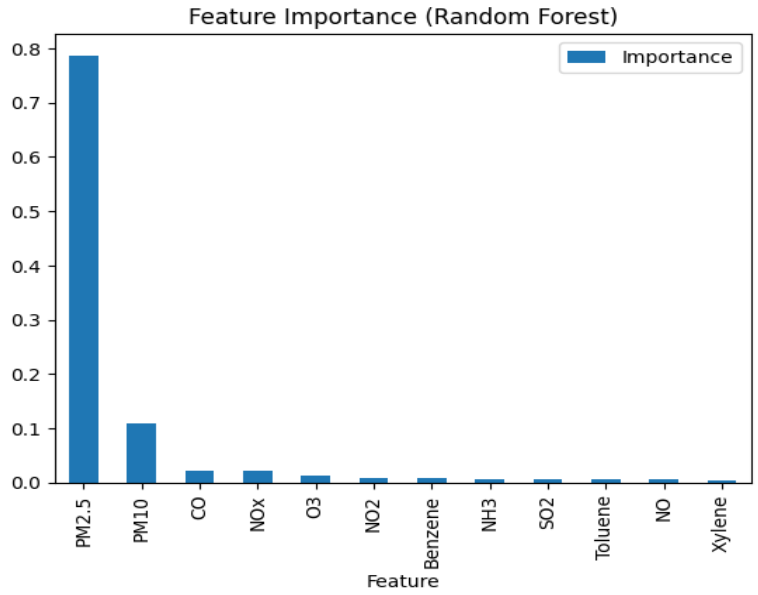

Figure 4: Feature importance in Random Forest model. PM2.5 and PM10 are identified as the most influential predictors of AQI.

Step 4: Division of the Dataset

The cleaned dataset was split into training and testing sets based on an 80:20 ratio. The former was used to train the models, whilst the latter was used to test and deduce the models' predictive power.

Step Five: Data Normalization

To ensure that the size of features is uniform, normalization was applied to the data. The StandardScaler from Scikit-learn was used for the dataset and it also improves the performance of models such as SVR and KNN.

Step 6. Applying Machine Learning Models

This model used various machine learning models to predict the Air Quality Index (AQI), chosen for having unique strengths and relevance to this particular task. Its baseline model was linear regression, which provides a standard, classic benchmark against which improvements from more advanced algorithms could be measured. The Decision Tree Regressor was included because of the ability of this regressor to model non-linear relationships between features and AQI, resulting in more flexibility compared to Linear Regression. The Random Forest Regressor, being an ensemble method, was used to aggregate the predictions from multiple decision trees so as to effectively minimize overfitting and enhance predictive accuracy. It used Support Vector Regressor because it can capture complex relationships by finding out the best-fitting hyperplane. Furthermore, there is an assessment using the KNN algorithm as a form of distance-based regression method when detecting local patterns in data. All the models are run on their default parameters since no hyperparameter optimization approaches such

as GridSearchCV are used in this study. This methodology allowed for a simple comparison of the models' intrinsic abilities, without any impact of parameter tuning.

Step 7: Model Evaluation

The models were evaluated using standard regression metrics:

1. **Mean Squared Error (MSE)**
2. **Root Mean Squared Error (RMSE)**
3. **Mean Absolute Error (MAE)**
4. **R-Squared (R^2)**

```
Linear Regression Evaluation:
MAE: 33.5758, MSE: 2195.1934, RMSE: 46.8529, R²: 0.8398, MAPE: 18.06%, Adjusted R²: 0.8349
----------------------------------------
Decision Tree Evaluation:
MAE: 33.5684, MSE: 2126.5404, RMSE: 46.1144, R²: 0.8449, MAPE: 14.25%, Adjusted R²: 0.8401
----------------------------------------
SVR Evaluation:
MAE: 40.8164, MSE: 3206.5127, RMSE: 56.6261, R²: 0.7661, MAPE: 23.45%, Adjusted R²: 0.7588
----------------------------------------
Random Forest Evaluation:
MAE: 22.5933, MSE: 1033.2256, RMSE: 32.1438, R²: 0.9246, MAPE: 10.27%, Adjusted R²: 0.9223
----------------------------------------
k-NN Evaluation:
MAE: 24.8965, MSE: 1306.9115, RMSE: 36.1512, R²: 0.9047, MAPE: 10.47%, Adjusted R²: 0.9017
```

Table 5: Performance metrics of machine learning models. In the table, we compare accuracy, RMSE, and R^2 in models such as Random Forest, SVR, and Linear Regression.

This model show the best performance across these metrics, so Random Forest Regressor seems suitable for AQI prediction in New Delhi.

Step 8: Visualization of Results

To better determine the efficacy of the models, a heatmap was provided to compare, based on three evaluation metrics, their effectiveness with regards to these: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 score. With these metrics, one would be able to differentiate the accuracy and predictability of the models. Figure 8 encapsulates the performance of the five models: Linear Regression, Decision Tree, Support Vector Regressor (SVR), Random Forest, and K-Nearest Neighbors (K-NN).

From the plot, it has been observed that Random Forest Regressor is the model with the lowest values of MAE and RMSE while high R^2 value at 0.9246, indicating its strength in complex and nonlinear relationships within this dataset. The Linear Regression and K-NN models were reasonably good, both in terms of a minor increase in the values of RMSE and lower R^2 values compared to Random Forest. The Decision Tree was better than SVR but was prone to overfitting somewhat. In contrast, the weakest performance was offered by the SVR with the highest RMSE and lowest R^2 score at 0.7142, which implied poor performance in this application.

It highlights the increased predictability through the Random Forest Regressor and emphasizes the appropriateness of this model for predicting AQI in Delhi. The comparative analysis can provide useful information about the selection of the most effective model for such prediction tasks.

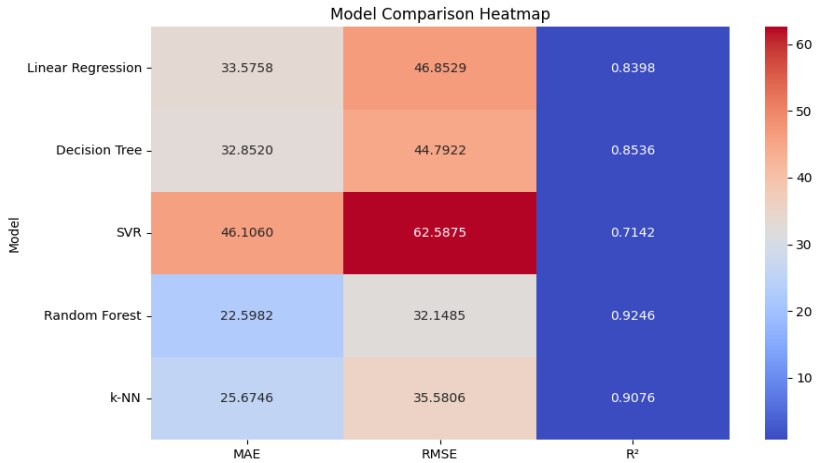

Figure 5 : Comparative Analysis of Machine Learning Models for AQI Prediction

The heatmap depicted the performance of models with respect to their MAE, RMSE, and R^2. Low scores in MAE and RMSE indicate a large value for R^2, which means better performance.

Step 9. Insights and Recommendations

Based on the analysis, the most accurate model for predicting the Air Quality Index (AQI) was a Random Forest Regressor. The study highlights the need for integrating advanced machine learning techniques into strategies for pollution management. Recommendations were made to governmental bodies on the utility of using such models in real-time AQI forecasting and better decision-making strategies.

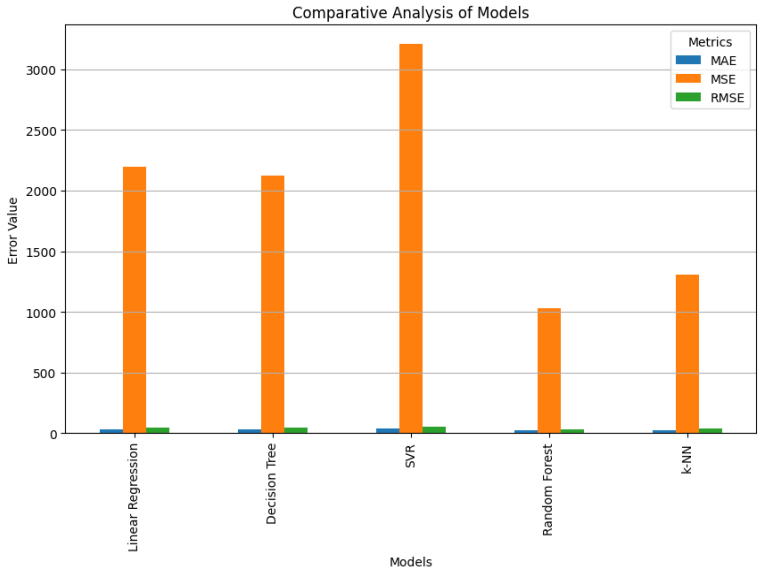

Figure 6: Comparative performance of machine learning models using R^2 and RMSE metrics.

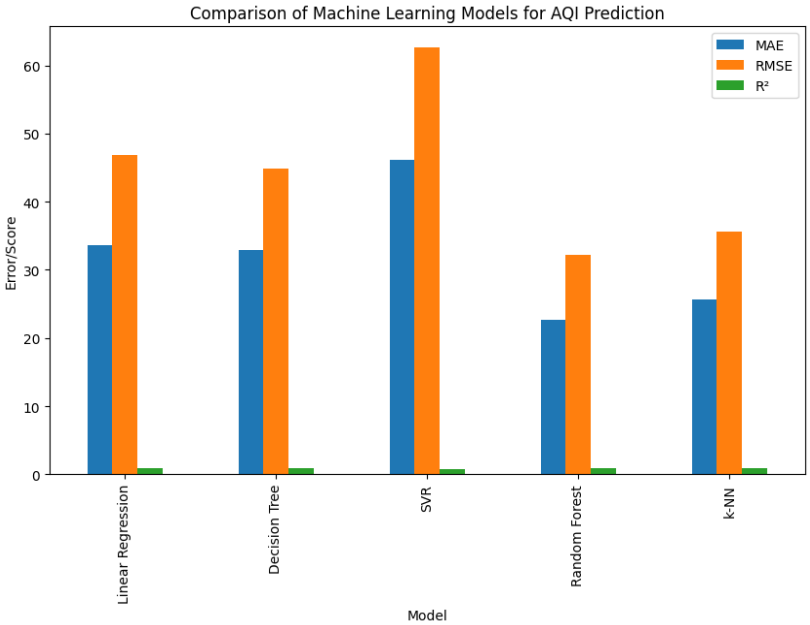

Figure 7: Seasonal trends in AQI levels over a five-year period, highlighting winter peaks.

V. RESIDUAL ANALYSIS

Residual analysis provides deep insight on the performance of the machine learning-based prediction models by understanding the difference between actual and predicted values. Residual plots obtained for the models used in this work highlight key differences in their generalization ability as well as in the accuracy of their prediction for the values of the AQI.

The residual plot for Linear Regression shows a noticeable trend, with residuals increasing systematically as the true AQI values rise. This pattern reflects the model's limitations in capturing complex, non-linear relationships inherent in the data. The Decision Tree model demonstrates a broader scatter of residuals, suggesting overfitting

in certain areas and inconsistent predictions for extreme AQI values. Similar to the Support Vector Regressor, its residuals seem to have an overall pattern of systematic errors, indicating difficult handling in any variability present in the AQI data.

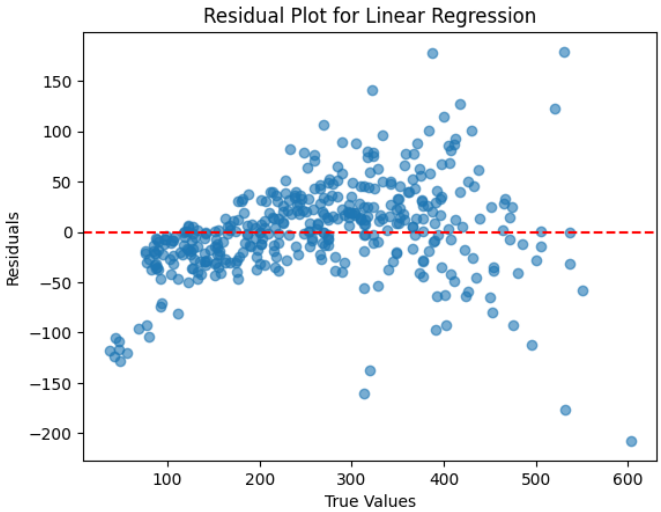

Figure 8a: Residual Plot for Linear Regression

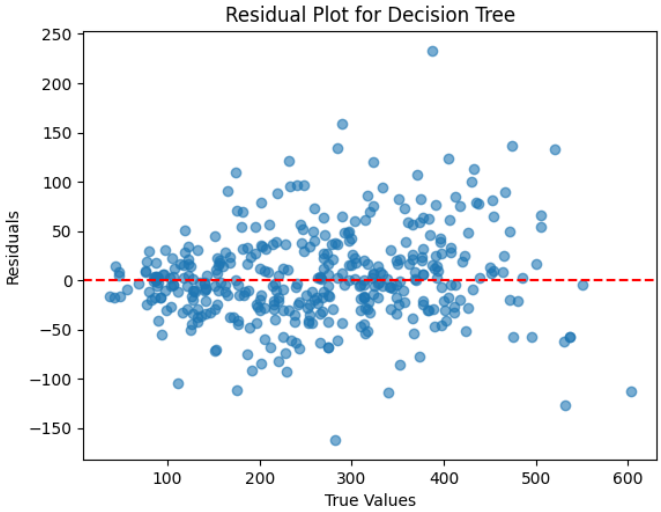

Figure 8b: Residual Plot for Decision Tree

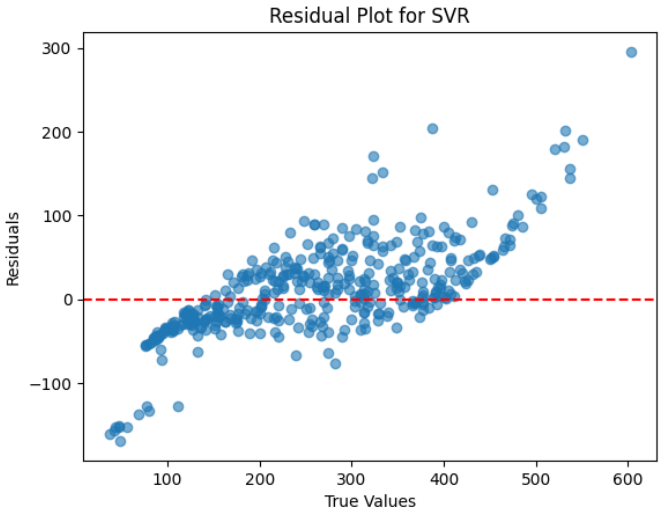

Figure 8c: Residual Plot for Support Vector Regressor

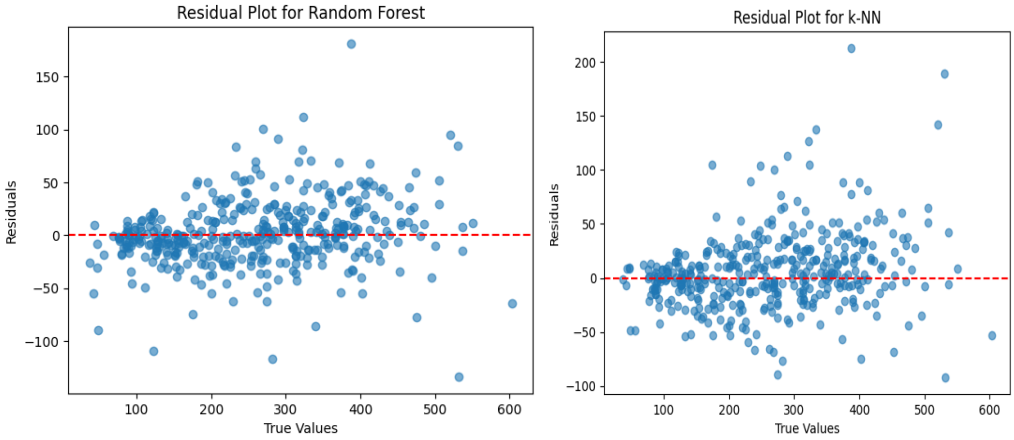

Figure 8d: Residual Plot for K-Nearest Neighbors

Figure 8e: Residual Plot for Random Forest.

The residual plot for the KNN model seems to be more random, but slight patterns show clustering around particular AQI values, indicating issues in making generalizations across the whole set of data, particularly for extreme values. Among the models, the best model is the Random Forest Regressor, as its residual plot shows a balanced figure with values centered very close to zero. This implies almost no bias and accuracy in predicting the values.

The residuals bear testimony to the fact that the Random Forest Regressor would thrive with complex data, while the performance of simpler models like Linear Regression and SVR fails. The findings thus strengthen the supremacy of the Random Forest model's performance as evidenced by their evaluation metrics.

VI. RESULTS AND DISCUSSION

The results of the experiment have been able to show that differing machine learning models are all able to predict the Air Quality Index of New Delhi. Amongst all five models evaluated, Random Forest Regressor stood out as the standout winner, achieving also the very highest score for R^2 at 0.9246 and lowest RMSE, at 32.15. An ensemble-based architecture employed in this model enabled it to capture intricate, nonlinear relations between pollutants and AQI-the best choice for that task. Linear Regression being a more simple model resulted in an R^2 score of 0.8398 but failed to seize effectively the intricacies of the prediction of AQI.

	Model	MAE	RMSE	R^2
0	Linear Regression	33.575806	46.852891	0.839843
1	Decision Tree	32.851960	44.792235	0.853621
2	SVR	46.106050	62.587479	0.714209
3	Random Forest	22.598247	32.148518	0.924596
4	k-NN	25.674578	35.580556	0.907637

Table 6: Performance Metrics of Machine Learning Models

The Decision Tree Regressor performed in a mediocre manner as it resulted in an R^2 of 0.8536, and it did have fits of overfitting that slightly depressed its predictive ability. Similarly, the KNN algorithm, which is simple yet effective to capture the local data patterns, got a value of 0.9076 R^2 but struggles with higher generalizations. But the Support Vector Regressor in this case performed least, with an R^2 score of 0.7142, largely a victim of computational complexity and sensitivity to parameter tuning. This again shows the Random Forest Regressor as better than the goals of this experiment.

The analysis further revealed a significant influence of particulate matter as well as PM2.5 and PM10 on AQI, with correlation coefficients of 0.88 and 0.85, respectively. Such pollutants were consistently identified as the

most significant contributors to the impact on the air quality in New Delhi. Other pollutants, however, such as carbon monoxide and sulfur dioxide, presented weaker correlations with AQI, suggesting a more local or seasonal contribution. Temporal analysis of AQI trends revealed seasonal peaks, particularly during winter, attributable to stubble burning, reduced atmospheric dispersion, and increased vehicular emissions.

VII. RECOMMENDATION

Based on the results gathered from this research, studies should integrate the Random Forest Regressor with other machine learning models into air quality monitoring and management systems. Government agencies should adopt real-time AQI prediction models so that timely interventions such as controlling traffic, controlling industrial activities, and public health advisories could be made for improved air quality when certain emissions were significantly on the rise. Renewable energy utilization would be promoted, and proper norms should be set for vehicular emissions to curb primary detrimental emissions emitted, such as NOx and CO.

The integration of IoT-based air quality sensors in key urban hotspots enables the gathering of high-granularity, close-to-real-time AQI information that can be used for actual dynamic prediction. To counter the seasonal effect of the burning of stubbles, it is important to have cross-state coordination and support the adoption of eco-friendly agriculture practices, including subsidies for residue management technologies. These initiatives, in tandem with public campaigns, will scale down the pollution in New Delhi.

VIII. CONCLUSION AND FUTURE WORK

The study demonstrates the capability of advanced machine learning models, and more so, the Random Forest Regressor, for accurately predicting AQI and facilitating actionable perceptions of pollution control. High accuracy demonstrated by the model justifies its adoption in real-time applications in urban air quality management. Future research should extend these findings and explore the applicability of deep learning techniques for better capture of temporal patterns by AQI data.

Implementing meteorological parameters, such as wind speed and temperature, in the models may result in more precise prediction. In addition, coupling these predictive models with mobile and web-based applications can allow citizens to monitor the air quality and make informed decisions. Therefore, bridging the gap between technical research and practicality, all these benefits can help pave the way towards a sustainable urban lifestyle and proper functioning of pollution control mechanisms.

REFERENCES

[1] Prediction of Air Quality Index Using Machine LearningTechniques: A Comparative AnalysisN. Srinivasa Gupta , 1 Yashvi Mohta, 2 Khyati Heda, 2 Raahil Armaan,2 B. Valarmathi, 2and G. Arulkumaran

[2] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Applied Sciences*, vol. 9, p. 4069, 2019.

[3] M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," *Complexity*, vol. 2020, Article ID 8049504, 23 pages, 2020.

[4] G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models," *Journal of Engineering Research*, vol. 9, 2021.

[5] S. V. Kottur and S. S. Mantha, "An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, pp. 146–152, 2015.

[6] S. Halsana, "Air quality prediction model using supervised machine learning algorithms," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 8, pp. 190–201, 2020.

[7] A. G. Soundari, J. Gnana, and A. C. Akshaya, "Indian air quality prediction and analysis using machine learning," *International Journal of Applied Engineering Research*, vol. 14, p. 11, 2019.

[8] C. R. Aditya, C. R. Deshmukh, N. D. K, P. Gandhi, and V. Astu, "Detection and prediction of air pollution using machine learning models," *International Journal of Engineering Trends and Technology*, vol. 59, no. 4, pp. 204–207, 2018.

[9] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters," *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 5106045, 14 pages, 2017.

[10] P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms," *International Journal of Computer Applications Technology and Research*, vol. 8, pp. 367–370, 2019.

[11] M. Bansal, "Air quality index prediction of Delhi using LSTM," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 8, pp. 59–68, 2019.

[12] A. Shishegaran, M. Saeedi, A. Kumar, and H. Ghiasinejad, "Prediction of air quality in Tehran by developing the nonlinear ensemble model," *Journal of Cleaner Production*, vol. 259, Article ID 120825, 2020.

[13] L. Tuan-Vinh, "Improving the awareness of sustainable smart cities by analyzing lifelog images and IoT air pollution data," in *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data)*, IEEE, Orlando, FL, USA, September 2021.

[14] R. Kumar, P. Kumar, and Y. Kumar, "Time series data prediction using IoT and machine learning technique," *Procedia Computer Science*, vol. 167, pp. 373–381, 2020.

[15] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artificial neural network model," *Clean Technologies and Environmental Policy*, vol. 21, no. 6, pp. 1341–1352, 2019.

[16] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmospheric Environment*, vol. 80, pp. 426–437, 2013.

[17] S. Hansun and M. Bonar Kristanda, "AQI measurement and prediction using B-wema method," *International Journal of Engineering Research and Technology*, vol. 12, pp. 1621–1625, 2019.

[18] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," *Sustainable Cities and Society*, vol. 67, Article ID 102720, 2021.

[19] M. Londhe, "Data mining and machine learning approach for air quality index prediction," *International Journal of Engineering and Applied Physics*, vol. 1, no. 2, pp. 136–153, May 2021.

[20] R. W. Gore and D. S. Deshpande, "An approach for classification of health risks based on air quality levels," in *Proceedings of the 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, pp. 58–61, Aurangabad, India, October 2017.

[21] X. Zhao, M. Song, A. Liu, Y. Wang, T. Wang, and J. Cao, "Data-driven temporal-spatial model for the prediction of AQI in Nanjing," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 10, no. 4, pp. 255–270, 2020.

[22] A.-S. Chowdhury, M. S. Uddin, M. R. Tanjim, F. Noor, and R. M. Rahman, "Application of data mining techniques on air pollution of Dhaka City," in *Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS)*, pp. 562–567, Varna, Bulgaria, August 2020.

[23] Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner, "Data-driven air quality characterization for urban environments: a case study," *IEEE Access*, vol. 6, Article ID 77996, 2018.

[24] C. Srivastava, S. Singh, and A. P. Singh, "Estimation of air pollution in Delhi using machine learning techniques," in *Proceedings of the 2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, pp. 304–309, Greater Noida, India, September 2018.

[25] R. Raturi and J. R. Prasad, "Recognition of future air quality index using artificial neural network," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, pp. 2395–0056, 2018.

[26] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, "A machine learning model for air quality prediction for smart cities," in *Proceedings of the 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, pp. 452–457, Chennai, India, March 2019.

[27] V. Sivakumar, G. R. Kanagachidambaresan, V. Dhilip Kumar, M. Arif, C. Jackson, and G. Arulkumaran, "Energy-efficient markov-based lifetime enhancement approach for underwater acoustic sensor network," *Journal of Sensors*, vol. 2022, Article ID 3578002, 2022.

[28] J. Sethi and M. Mittal, "Ambient air quality estimation using supervised learning techniques," *ICST Transactions on Scalable Information Systems*, vol. 6, Article ID 159628, 2019.

[29] P. Hajek and V. Olej, "Predicting common air quality index - the case of Czech microregions," *Aerosol and Air Quality Research*, vol. 15, no. 2, pp. 544–555, 2015.

[30] S. Ameer, M. A. Shah, A. Khan et al., "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, Article ID 128325, 2019.

[31] A. R. Behesht Abad, S. Mousavi, N. Mohamadian et al., "Hybrid machine learning algorithms to predict condensate viscosity in the near wellbore regions of gas condensate reservoirs," *Journal of Natural Gas Science and Engineering*, vol. 95, Article ID 104210, 2021.

[32] M. Rajabi, S. Beheshtian, S. Davoodi et al., "Novel hybrid machine learning optimizer algorithms to prediction of fracture density by petrophysical data," *Journal of Petroleum Exploration and Production Technology*, vol. 11, no. 12, pp. 4375–4397, 2021.

[33] A. R. Behesht Abad, P. S. Tehrani, M. Naveshki et al., "Predicting oil flow rate through orifice plate with robust machine learning algorithms," *Flow Measurement and Instrumentation*, vol. 81, Article ID 102047, 2021.

[34] O. Hasbeh, M. Ahmadi Alvar, K. Y. Aghdam, H. Ghorbani, N. Mohamadian, and J. Moghadasi, "Hybrid computing models to predict oil formation volume factor using multi-layer perceptron algorithm," *Journal of Petroleum and Mining Engineering*, vol. 23, no. 1, pp. 17–30, 2021.

[35] F. Jafarizadeh, M. Rajabi, S. Tabasi et al., "Data-driven models to predict pore pressure using drilling and petrophysical data," *Energy Reports*, vol. 8, pp. 6551–6562, 2022.

[36] G. Zhang, S. Davoodi, S. S. Band, H. Ghorbani, A. Mosavi, and M. Moslehpour, "A robust approach to pore pressure prediction applying petrophysical log data aided by machine learning techniques," *Energy Reports*, vol. 8, pp. 2233–2247, 2022.

[37] S. Tabasi, P. Soltani Tehrani, M. Rajabi et al., "Optimized machine learning models for natural fractures prediction using conventional well logs," *Fuel*, vol. 326, Article ID 124952, 2022.

[38] M. Rajabi, O. Hazbeh, S. Davoodi et al., "Predicting shear wave velocity from conventional well logs with deep and hybrid machine learning algorithms," *Journal of Petroleum Exploration and Production Technology*, 2022.

[39] S. Beheshtian, M. Rajabi, S. Davoodi et al., "Robust computational approach to determine the safe mud weight window using well-log data from a large gas reservoir," *Marine and Petroleum Geology*, vol. 142, Article ID 105772, 2022.

[40] Z. K. Masoud, S. Davoodi, H. Ghorbani et al., "Band, Permeability prediction of heterogeneous carbonate gas condensate reservoirs applying group method of data handling," *Marine and Petroleum Geology*, vol. 139, 2022.

[41] N. Mohamadian, H. Ghorbani, D. A. Wood, and M. A. Khoshmardan, "A hybrid nanocomposite of poly(styrene-methyl methacrylate-acrylic acid)/clay as a novel rheology-improvement additive for drilling fluids," *Journal of Polymer Research*, vol. 26, no. 2, p. 33, 2019.

[42] N. Mohamadian, H. Ghorbani, D. A. Wood, and H. K. Hormozi, "Rheological and filtration characteristics of drilling fluids enhanced by nanoparticles with selected additives: an experimental study," *Advances in Geo-Energy Research*, vol. 2, no. 3, pp. 228–236, 2018.

[43] A. Choubineh, H. Ghorbani, D. A. Wood, S. Robab Moosavi, E. Khalaf, and E. Sadatshojaei, "Improved predictions of wellhead choke liquid critical-flow rates: application of the MLP technique," *Fuel*, vol. 207, pp. 547–560, 2017.

[44] H. Ghorbani, J. Moghadasi, and D. A. Wood, "Prediction of gas flow rates from gas condensate reservoirs through wellhead chokes using a firefly optimization algorithm," *Journal of Natural Gas Science and Engineering*, vol. 45, pp. 256–271, 2017.

[45] A. R. Behesht Abad, H. Ghorbani, N. Mohamadian et al., "Robust hybrid machine learning algorithms for gas flow rates prediction through wellhead chokes in gas condensate fields," *Fuel*, vol. 308, Article ID 121872, 2022.