

Original Article

Environmental Pollution Management through Air Quality Prediction- A Deep Learning Based Approach

Vetri Selvi R¹, Sathish Babu R²,

^{1,2}Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India.

¹Corresponding Author : vetriselvischolar@gmail.com

Received: 10 February 2026

Revised: 25 July 2026

Accepted: 07 August 2026

Published: 30 September 2026

Abstract - 'Good health and well-being' is the third Sustainable Development Goal formulated by the United Nations. Air pollution is one of the major issues that affects the good health of people across the globe. The quality of air we as 'humans' breathe is vital for leading a quality and healthy life. This research attempts to demonstrate the application of deep learning models in the prediction of the air quality index of select Indian cities. This research makes use of the techniques Random Forest, XGBoost, and LightGBM. It has also developed an advanced Deep Learning model for the prediction of the Air Quality Index. The developed models have been tested against a dataset on the air quality of multiple Indian cities. The performance evaluation revealed that the advanced deep learning model with additional layers has rendered better performance than the traditional model Random Forest, XGBoost and LightGBM. The research also gives insights into how deep learning models could be implemented on a real-time basis across all cities in India in order to prevent environmental pollution and ensure sustainable living of human beings on earth.

Keywords - Air Quality Index, Deep learning, Environmental Pollution, Indian Cities, SDG.

1. Introduction

Pollution has been studied and analyzed by different researchers in the past. It has been identified as a deadly-hazardous factor that affects the planet and humanity. Authors [1] have studied water pollution; The authors [2] studied radioactive pollution, and the authors [3] focused on noise pollution. Similarly, authors [4] studied in general how air pollution affects people and their lifespan more than other pollution factors globally. Authors [5] studied about environment pollution, causes, types, impact and health risks; the study insisted that air is the most dangerous pollutant factor. This is proved by authors [6, 7], who studied Indian air pollution specifically. Air pollution in India has drastically increased in the years 2017-2025 [8]. It directly affects the human population, causing approximately 7 million deaths in a year through severe lung illness and heart disease [9]. Globally, India is ranked 5th among the 138 countries, as a heavily polluted country [10]. It was ranked third in the world in terms of air pollution [11].

Prediction of air quality index using advanced technology like machine learning can help the nation in combatting the issue of air pollution and provide a healthy life to its citizens. This research has attempted to demonstrate the usage of machine learning techniques for the prediction of the Air Quality Index (AQI). Six machine learning models have been

developed and tested against the pollution data of major Indian cities, namely, Delhi, Bengaluru, Mumbai, Kolkata, Jaipur, Patna, Hyderabad, Lucknow, Gurugram, Chennai and Ahmedabad in this research. The models developed in this research are Random Forest (RF), k-NN (K Nearest Neighbour), eXtreme Gradient Boost (XGB), LightGBM (LGBM), a basic machine learning model and an advanced deep learning model. A comparison of the performances of the models has also been made and presented in this research. This research presents a novel approach that creates a complete machine learning pipeline for AQI prediction by developing six different models using a large dataset from major Indian cities.

2. Related Literature

2.1. Random Forest

Various models that could forecast AQI levels and categorize them into various impurity bands were tested, and their performance was evaluated. It was found that the decision tree and random forest models produced high accuracy in prediction, with 99.98 percent and 99.97 percent, respectively. On the other hand, logistic regression had low accuracy with 72.78 percent. It was clear from the experimental result that classifier models like random forest and decision tree performed better for predicting AQI [12].



The authors [13] developed a hybrid model, LSTM-Support Vector Regressor (LSTM-SVR), which showed effective performance. The ensemble model and the random forest both improved the accuracy rate in prediction with specific reference to heavily polluted areas. It was clear from the findings that ensemble and hybrid models are best when compared with individual models for prediction.

The authors [14] employed a hybrid framework that integrated the Autoregressive Integrated Moving Average (ARIMA) and Random Forest Regression (RFR) to enhance accuracy in forecasting AQI and maintain transparency. Non-linear relationships are captured by RFR among pollutants, whereas ARIMA is adopted to capture temporal patterns. The hybrid model used in this research is RFR for prediction and ARIMA for correcting residuals. The hybrid model achieved a lower MSE and a higher R² score than baseline models.

The authors [15] studied a prediction model for air quality using machine learning techniques. It was observed that the RFR algorithm performed better than other algorithms like linear regression, decision tree regression, and support vector regression. RFR had 0.99985 accuracy with a minimal MSE value = 0.00013 and MAE = 0.00373.

The authors [16] have stated that air quality acts as a main factor that influences well-being and public health, which shape the strategies of urban planning, environmental policies and practices to manage health. This research deployed two methodologies for prediction, namely LSTM and random forest regression. It was found that random forest regression performed well in AQI prediction based on a daily and hourly basis. Random Forest models showed minimum RMSE values and showed better position with real-time values for AQI, especially at the peak points.

2.2. Extreme Gradient Boost (XGBoost)

The authors [17] evaluated the performance of SVR and XGBoost using Exponentially Weighted Moving Average (EWMA) and individual charts. It was revealed that autocorrelation was perfectly done by the SVR model. At the same time, a combination of the EWMA chart and XGBoost outperformed well in monitoring air quality.

On the other hand, the authors [18] identified the factors influencing the AQI with specific reference to Changchun City with the help of the Pearson-MI method. Models adopted in this research entail the XGBoost model, Random Forest model, SVR model and LSTM model. This research developed a combined model with the help of the inverse variance method, which had the best accuracy rate in prediction when compared with the support vector regressor, LSTM, XGBoost and random forest method. The modified combined model minimizes the MAE by 0.6 percent and RMSE by 2 percent.

The authors [19] studied different algorithms used for predicting CO, PM₁₀, and PM_{2.5} that entail support vector machine, extreme gradient boosting, random forest, Artificial Neural Network, and multiple linear regression to determine the best machine learning algorithms for time scale and respective pollutants. Outcomes indicate that support vector machine, XGBoost, random forest, artificial neural network and multiple linear regression give the best performance in evaluating the performance metrics (R², RMSE, MAE) for 24hrs prediction.

2.3. Light Gradient Boosting Machine

The authors [20] introduced a new prediction methodology for air quality using machine learning models. The performance evaluation of different models in machine learning indicated that the LightGBM (Light Gradient Boosting Machine) classifier has the best accuracy rate with 97.5 percent. The LSTM model was considered superior and indicated goodness-of-fit (91.37 percent) for predicting AQI and O₃ prediction (90.46 percent). It was confirmed that the efficacy and consistency of deployed machine learning techniques predict air quality with high accuracy.

The authors [21] introduced an enhanced LASSA (Sparrow Search Algorithm), which optimized LightGBM for predicting the PM_{2.5} concentration. LightGBM is seen as the best approach in machine learning models and entails hyperparameters that could be mixed optimally for achieving the needed outcomes. From the findings of the research, it was observed that the proposed method performs better than others. MAPE error and RMSE indices were minimized from 3 percent to 16 percent.

The authors [22] adopted SSA for decomposing the AQI data into various arrangements like noise, oscillation component and trend. Bidirectional LSTM (long short-term memory) was added for predicting the decomposed AQI value, and LightGBM were adopted for integrating the estimated outcomes. From the outcomes of the research, the impact of combined methods (LightGBM, Bidirectional LSTM and SSA) had a good prediction rate.

The authors [23] developed a prediction system for air quality based on a framework using a machine learning model called LightGBM. The LightGBM classifier is trained in this model to predict weather and air quality. LightGBM is identified as the best framework for predicting air quality compared to previous models, with a high accuracy rate.

The authors [24] discussed gated recurrent units, LSTM, LightGBM, CatBoost, and XGBoost for predicting PM₁₀ levels. It was found that the LightGBM algorithm performed well in predicting PM₁₀. The performance metrics studied in this research are Nash-Sutcliffe efficiency = 0.09, MAE = 0.002, MSE = 0.02, R² = 0.99, mean absolute relative error = 1.0, RMSE = 0.1225.

Table 1. Air quality prediction models reviewed

Author and year	Method/Technique	Outcome
Nehete and Patil [12]	Random Forest	Classifier models like random forest and decision tree performed better for predicting AQI.
Zhang et al [13]	Ensemble model, random forest and LSTM-SVR	Ensemble and hybrid models are best when compared with an individual model for prediction.
Yenkikar et al [14]	A hybrid framework that integrated the Autoregressive Integrated Moving Average (ARIMA) and random forest regression	The hybrid model used in this research is RFR for prediction and ARIMA for correcting residuals.
Halsana [15]	RFR algorithm, linear regression, decision tree regression, and support vector regression.	The RFR algorithm performed better than other algorithms like linear regression, decision tree regression, and support vector regression.
Sadriddin et al [16]	Random Forest	Random Forest models showed minimum RMSE values and showed better position with real-time values for AQI, especially at the time of peak points.
Alfasanah et al [17]	Performance of SVR and XGBoost	The combination of the EWMA chart and XGBoost performed well in monitoring air quality.
Chen et al [18]	XGBoost model, Random Forest model, SVR model and LSTM model	The modified combined model minimizes the MAE by 0.6 percent and the RMSE by 2 percent
Lei et al [19]	Support vector machine, XGBoost, random forest, artificial neural network and multiple linear regression	Outcomes indicate that support vector machine, XGBoost, random forest, artificial neural network and multiple linear regression give the best performance in evaluating the performance metrics (R^2 , RMSE, MAE) for 24hrs prediction.
Liu et al [20]	LightGBM classifier	The LightGBM classifier has the best accuracy rate with 97.5 percent. The LSTM model was considered superior and indicated goodness-of-fit (91.37 percent) for predicting AQI and O_3 prediction (90.46 percent).
Liu et al [21]	LASSA optimized LightGBM	LightGBM is seen as the best approach in machine learning models and entails hyperparameters that could be mixed optimally for achieving the needed outcomes.
Zhang et al [22]	Bidirectional LSTM, Light GBM	The impact of combined methods (LightGBM, Bidirectional LSTM and SSA) had a good prediction rate.
Pillai et al [23]	LightGBM classifier	LightGBM is identified as the best framework for predicting air quality over previous models, with a high accuracy rate.
Mampitiya et al [24]	Gated recurrent unit, LSTM, LightGBM, CatBoost, XGBoost	The LightGBM algorithm performed well to predict PM_{10} .
Faldo et al [25]	APD-BayNet and TabNet	APD-BayNet, along with TabNet as a hybrid model, achieved the best performance with accuracy, F1-score, recall and precision of 98.17 percent, 95.79 percent, 95.28 percent and 96.64 percent, respectively.

From the studies and the literature, it is observed that several authors have used RF models, SVR, LightGBM and other neural networking-based models for AQI predictions. However, the application of different varieties of models, from basic classification models to advanced deep learning models, in the prediction of air quality with respect to a huge dataset has not been carried out yet, especially with respect to Indian cities.

2.4. Research Gap

The studies by authors Nehete and Patil [12], Halsana [15], and Sadriddin et al [16] focused on Random Forest

models in predicting air quality. Meanwhile, authors Zhang et al [13] and Yenikar et al [14] focused on hybrid and ensemble models for predicting the AQI. The authors Alfasanah et al [17], Chen et al [18] and Lei et al [19] used XGBoost, support vector machines and other boosting models.

Other studies, namely Liu et al [20], Liu et al [21], Pillai et al [23] and Mampitiya et al [24] studied AQI using the LightGBM models. Authors Faldo et al [25] used TabNet and BayNet combined models. This research is a novel approach that creates a complete machine learning pipeline for AQI

prediction by developing six different models, in addition to training them and testing them against large datasets of pollution levels of major Indian cities.

2.5. Novelty of the Proposed Approach

This research is novel because it has implemented a complete machine learning pipeline for air quality prediction, especially with respect to multiple Indian cities. This research is novel because it has trained and tested deep learning models for the prediction of AQI across Multiple Indian cities. The research conducted by [26] focused on four years of data obtained to predict the air quality of Gurugram city. The research conducted by [27] took into consideration fourteen locations across New Delhi, India. Likewise, the research conducted by [28] exclusively focused on the Indian city Vishakhapatnam.

The research by [30] focused on Anukul city of Odisha. The data used for the research was the data obtained from 17 cities in Odisha from 44 monitoring locations collected in the year 2024.

The current research, however, has trained and tested six models for the prediction of Air Quality Index by applying the dataset, which has air quality data at daily and hourly basis of numerous stations across several cities such as Delhi, Bengaluru, Mumbai, Kolkata, Jaipur, Patna, Hyderabad, Lucknow, Gurugram, Chennai and Ahmedabad in India. Therefore, this research is novel and unique.

The dataset collected from several Indian cities has been subject to training and testing using simple models like Random Forest, XGBoost, LightGBM and KNN. After that, machine learning models and advance deep learning models have been developed and tested. Further, a comparison of the performances of all the models has also been done using standard evaluation metrics.

In addition to that, there is novelty in the data preprocessing technique. The traditional models use k-nn approach, which fills missing values using nearest neighbours. The proposed approach uses Iterative Imputer, which is advanced and fills or predicts missing values by learning the relationship between the variables in the dataset.

The novelty lies in the progressive compression architecture: baseline vs improvised model. This deeper funnel-like design captures complex pollutant interactions and distils them into a compact representation, and offers better results when compared to traditional approaches.

This structure, theoretically, improves performance by capturing more complex non-linear pollutants interactions in the wider layer and enhancing generalization as compared to the shallow networks that underfit data, thereby rendering improved performance.

3. Materials and Methods

In this section, the proposed networking architecture, the algorithms adopted, the datasets used, data analysis and evaluation metrics used for measuring the model's performance have been explained in detail.

3.1. Research Framework

The following Figure 1 presents the entire framework of the research. As the figure illustrates, data on the pollutants that create air pollution in major cities is taken as input. The data is then subject to pre-processing. The pre-processed data is then fed into six models. The six models predict the Air Quality Index of the different Indian cities. Finally, the performances of the models are evaluated using Standard evaluation metrics.

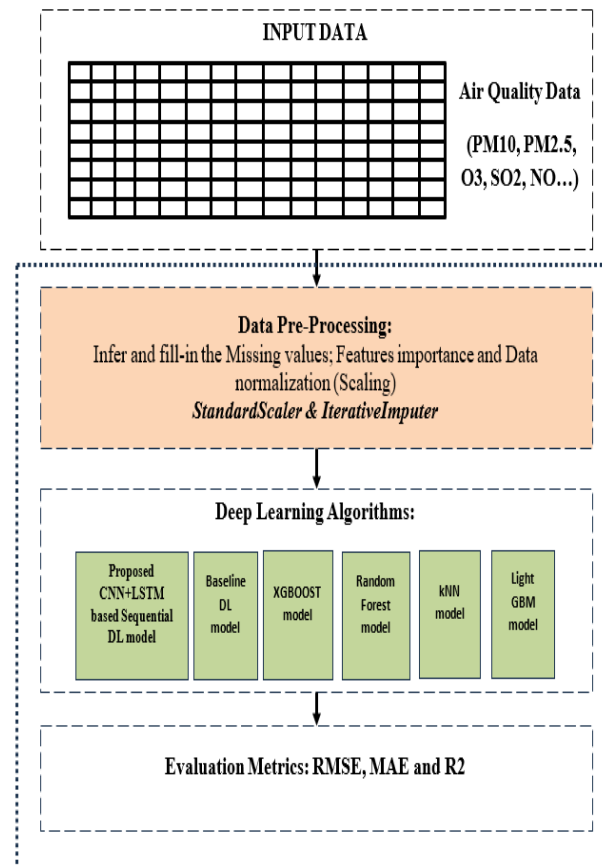


Fig. 1 Framework of the research

3.2. Architecture and Design

The proposed architecture consists of sequential layers with multiple processing trees (decision-making trees) that use the feature extraction approach. The novelty in this model is that the model uses iterative imputation for data preprocessing and identifies those missing values during the process of AQI prediction. Figure 2 presents the architecture of the machine learning model developed.

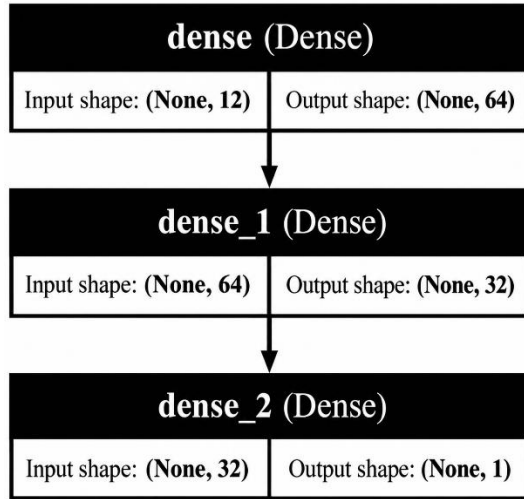


Fig. 2 Basic machine learning model architecture

The model is further improved by applying more layers for better performance. Figure 3 illustrates the architecture of the deep learning model.

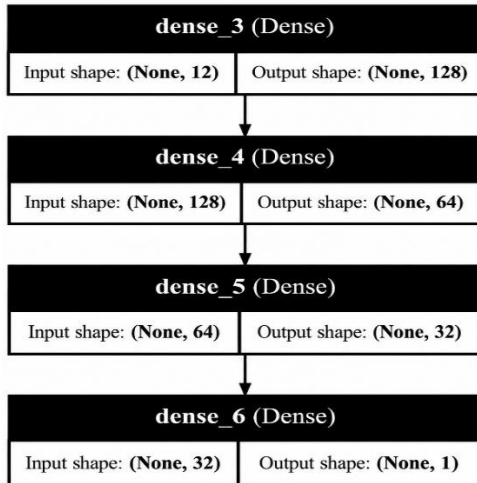


Fig. 3 Improved DL based dense network model architecture

As illustrated in Figure 2, the baseline DL architecture has three dense layers (the base model's blocks), whereas the improvised DL model has an additional block of a dense-dropout layer. The layer includes two batch normalization layers and two dropout layers. From the baseline model, which is developed by Google Cloud, the improved DL architecture is designed and developed to increase the baseline model's performance.

It can be identified from Figure 3 that the improvised 1Dimensional (1D) CNN + Bidirectional Sequential Bi- Long-Short Term Memory (LSTM) model with iterative imputation obtains the data in text form and passes it to the visual layers (CNN). Here, the feature extraction is carried out with a one-

dimensional vector that flattens the processed data. For pre-processing, the Iterative Imputer with Standard Scaler function is adopted here, which is a new approach for AQI prediction. Once the model extracts the features (pollutants), the flattened output from the visual layer is passed to the sequence learning layers, where the LSTM layers are present. Here, the data based on time-series are analyzed and forecasted to predict the AQI levels based on pre-defined AQI bucket values and classifications. Finally, the output layer labels the classified AQI values and classifies the predicted observations into respective classes. Figure 4 presents the detailed architecture of the deep learning model.

The improvised model includes a Convolutional Neural Networking (CNN) based architecture. For time-series data, sequential long-short-term memory learning is adopted. By combining these into a CNN+LSTM-Sequential model, the baseline DL model is executed. To improve performance and efficiency, hyper-parameter fine-tuning of reducing the epochs and early stopping is carried out in the improvised model. Similarly, an additional dense layer, as mentioned earlier, is included (dense-dropout layer). The CNN (visual features) includes convolution NN layers, max-pooling layers and batch normalization layers, where the pre-processed data is processed, and the features are extracted. Later, the processed data is passed through the sequence learning layers, where the temporal patterns of AQI (pollutants with AQI>150) are identified, and the final results are passed to the output layer.

The baseline model includes an input layer with 64 nodes (features) in size. Followed by are the hidden layers with 32 nodes (Dense + fully connected layer with ReLU function layer. Lastly, the third layer is a dense layer of size 1node. This is the output layer embedded with regression. The last and fourth layer is the output layer with 1node including the regression function. This progressive-compression architecture is the novelty here, where complex pollutant interactions are captured and distilled into compact representations. This improves the generalization that is lacking in the baseline model.

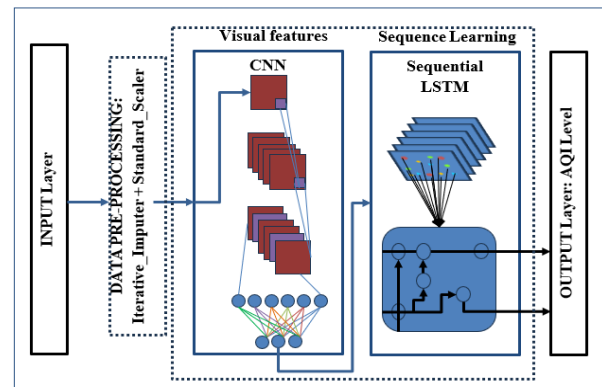


Fig. 4 Architecture diagram of CNN + Sequential LSTM model

The performance of the AQI prediction models developed is then measured by adopting the R^2 and Mean-Squared-Error (MSE) evaluation techniques, while the classifier algorithm-based models (LightGBM, XGBoost, and Random Forest) are

measured by adopting the metric evaluation techniques like R^2 , Mean-Absolute-Error (MAE) and MSE. The flow of the entire research is illustrated in Figure 5 as follows.

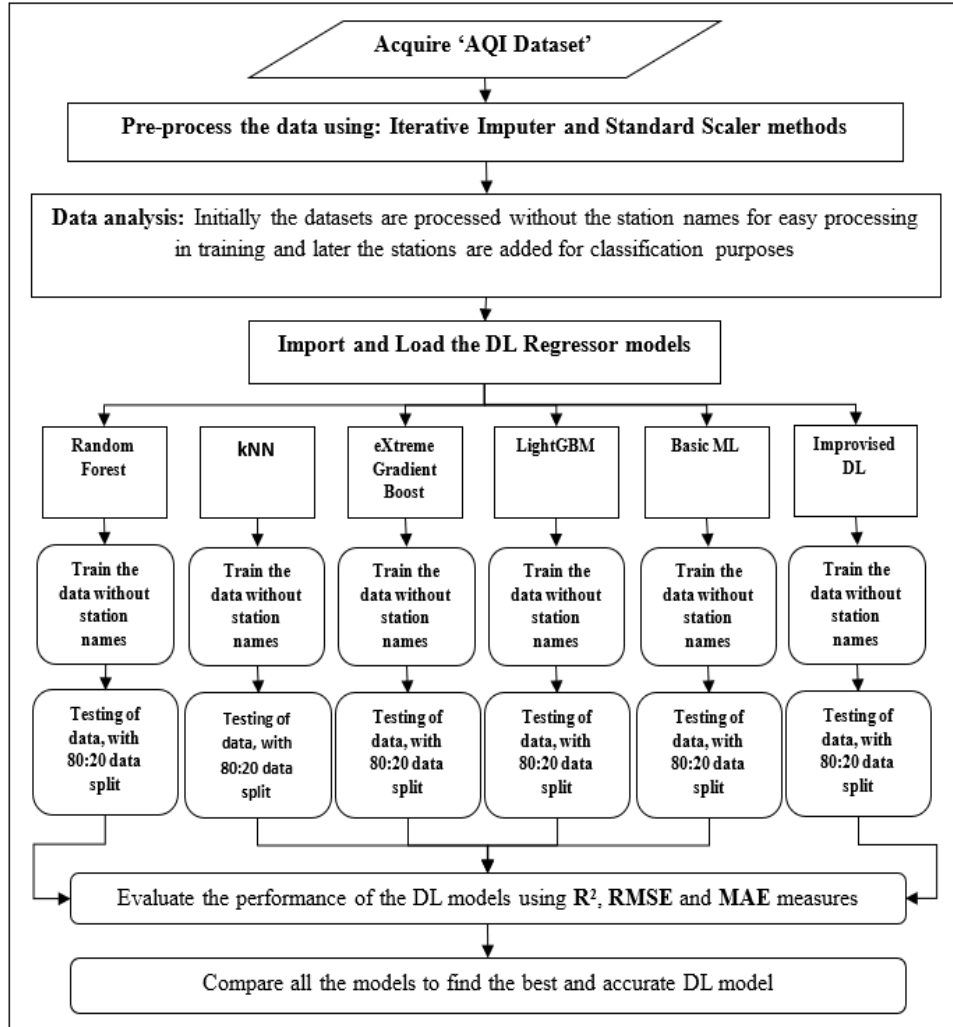


Fig. 5 Proposed research flow

3.3. Dataset

In this section, the data adopted from the open source database Kaggle, the parameters used and the pre-processing of the datasets are explained as follows.

3.3.1. Data

In prediction models, different data types are used, such as texts, images, and numeric value. In this study, the data used are the numeric values.

The numeric values denote air quality indices, which indicate the hazard level; where the higher indices (301+) show that the air is heavily polluted and consists of various health risks, while the lesser indices show the air quality is safer to breathe (0-50).

3.3.2. Parameters

The data includes 1,08,034 samples in the repository, which are used to the full extent and missing values are filled in by using the Iterative-Imputer function from sklearn with random state set as '0'. The data is accumulated from the time period of 2015-2020. The data frequency is acquired hourly in Indian cities across different stations.

The parameters used here are obtained from the Kaggle open-source database [30]. The data based on the stations (station ID) and date are pre-processed and used for the prediction model. These two (station ID from the City and Date-time from the Date) parameters are later used as base parameters for city-wise performance analysis. The 14 parameters of the dataset have been presented in Table 2.

Table 2. Parameter and data types used

S. No	Parameter	Data Type	Description
1	PM2.5	Integer64	Fine particles: Primary pollutants that cause sever lungs and heart disease.
2	PM10	Integer64	Inhalable particles: The second primary pollutant that causes respiratory irritation, mainly in high-dust areas.
3	NO	Integer64	Nitrogen-Oxide (NO), Nitrogen-Oxides (NOx) and Nitrogen-Dioxide (NO2): Highly sensitive gas that aggravates lung and heart disease.
4	NO2	Integer64	This contributes to smog.
5	Nox	Integer64	This aggravates lung disease.
6	NH3	Integer64	Ammonia: Colorless and pungent gas that aggravates asthma.
7	CO	Integer64	Carbon monoxide: Drastically reduces oxygen supply.
8	SO2	Integer64	Sulfur-Dioxide: Fossil-fuel based combustions related and causes respiratory system irritations.
9	O3	Integer64	Ground-Level-Ozone: Product of VOCs + NOx that results in asthma.
10	BENZENE	Integer64	C ₆ H ₆ : The most dangerous carcinogen that may result in leukemia and potentially death.
11	TOLUENE	Integer64	C ₇ H ₈ : Toxic solvent that attacks the nervous system, resulting in potential kidney failure/damage.
12	XYLENE	Integer64	C ₈ H ₁₀ : It causes chronic neurological impact and also irritations to the skin, respiratory system, and eyes.
13	AQI	Integer64	Air Quality Index
14	AQI_BUCKET	Float64	Target Output: Classification of AQI

The parameters chosen here are the key pollutants, namely particulate pollutants, namely PM: 10 and 2.5; gas pollutants as NO, SO, CO and O; and Volatile-Organic-Compounds (VOCs)/ hazardous air pollutants, namely benzene, xylene and toluene.

These are the most sensitive group-based pollutants chosen concerning human health, where severe lung disease and heart disease are caused, which leads to death [31]. Though other pollutants like Suspended Particulate Matter (SPM) or Respirable SPM (RSPM) are considered in different studies, this is not applied here due to data scarcity in the focused cities.

3.3.3. Data Pre-Processing

Using the “null value analysis” in Python, the parameters are pre-processed, in which the missing data or null (0) data are defined. In this research, the date, time (hours), and stations (ID) are pre-processed since there were null values (undefined data/ 0) found in the database. By using the Iterative Imputer (Multivariate Imputation by Chained Equations: MICE) method, the study handled the missing values, i.e. filling in data for undefined or null value data.

3.3.4. Data Handling

The novelty here is the KNNImputer function in sklearn. This is a more advanced function than the k-folds and cross-validation techniques in AQI predictions. By utilizing the KNNImputer along with the StandardScaler method, the model pipeline is designed. The IterativeImputer fills in the

missing values and null values (NaN). The neighbour value for KNNImputer in the pipeline is set to 3, where the algorithm fills the missing values based on the three closest data points for imputing the missing value.

This is more effective and significant than k-fold, which assumes the nearest/closest value available. This function uses the built-in ‘Euclidean distance’ calculation for missing values in Python.

3.3.5. Data Normalization

By using the Iterative Imputer (Multivariate Imputation by Chained Equations: MICE) method, the study handled the missing values, i.e. filling in data for undefined or null value data. This technique enhances a model’s performance.

3.3.6. Feature Engineering

For feature scaling, Standardization from sklearn (StandardScaler) is adopted. This technique overcomes outlier issues. Thus, pre-processing is carried out with the imputation’s random state set to ‘0’. Non-numeric data (day and station ID) are dropped for model prediction and applied later for city-wise station predictions.

The station names are added using the station ID and city datasets. Similarly, the date-time is converted into a processable format using the “Date format” function. The rows with missing values are dropped (refer to Figure 6) and processed in the PM2.5 distribution column:

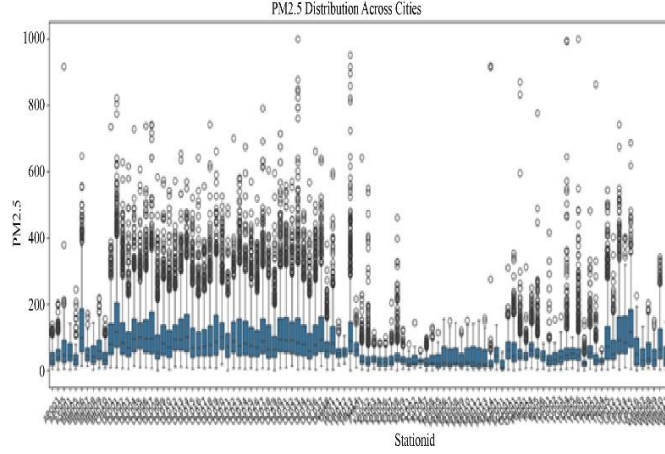


Fig. 6 PM2.5 distribution city-wise

3.4. Algorithms Applied

3.4.1. Random Forest

It is the ensemble method that builds and integrates various decision trees and enhances their predictions to improve consistency and accuracy. It adopts bootstrap aggregating, where some decision trees are given individual training on a random samples dataset. The overall result is identified by regression and classification in predicting independent trees.

Working Principle of Random Forest

Bootstrap sampling: Subsets are constructed randomly by selecting data points and also include duplicates.

Construct tree: A decision tree is constructed on each data subset, adopting a randomly chosen set of attributes at each tree at every partition to encourage diverse model paths.

Grouping: Prediction is done based on the most commonly predicted trees in classification; on the other hand, regression is estimated by means of the results obtained from individual trees.

Algorithm

The random forest algorithm used in this research is adopted from [32]. The pseudocode for the random forest model is presented below:

Algorithm 1: Random Forest

```

For d=1 to N, do
    Bd = Bootstrap(B)
    Id = Randomtree(Bd, √k)
    Forest = U Id
    For each Id ∈ Forest, do
        For each yj ∈ B, do
            NPd(yj) = NodesOnPath(yj, Id)
            For each i ∈ NPd(yj), do
                node_id = GetNodeID(i)

```

feature_id = Hash(node_id, HF)

$y'_{j, \text{feature_id}} = y'_{j, \text{feature_id} + 1}$

Return B' = U y'_j

Mathematical Representation

Let $R_1, R_2, R_3, \dots, R_n$ denotes the individual trees, and you be the model input data

$$\hat{G} = \text{mode}(R_1(g), R_2(g), R_3(g), \dots, R_n(x)) \quad (1)$$

Regression is done using the final prediction.

$$\hat{G} = 1/n \sum_{i=1}^{no} R_i(g) \quad (2)$$

Where mode depicts grouping and $\sum$ average prediction

3.4.2. Extreme Boost Gradient (XGBoost)

XGBoost sequentially includes decision trees; in each step, it corrects the errors made in the previous one to improve the prediction of the model. Every tree is given training to rectify the mistakes made by the previous tree, and it is known as boosting. Parallel processing is inbuilt which allows effective training on huge volumes of data. It supports tailoring and permits users to tune parameters of the model to get the exact performance on the issue.

Working Principle of Extreme Boost Gradient

Initialize with base learner: training starts with the first decision tree and is developed using training data. The base model in regression tasks estimates the average of target variables.

Estimating the error: First tree is given training, then the errors between actual and predicted values are estimated.

Next tree training: The next decision tree is given training to predict the previous tree error. It rectifies the errors made by the initial tree.

Reiterate the process: It moves on with each tree attempting to rectify the previous tree's errors till it meets the criteria.

Integrate predictions: The final step is the cumulative outcome of predictions produced by each tree.

Algorithm

The pseudocode for the XGB-DL model is given below.

Algorithm 2: XGB algorithm

Input: C, current mode's instance set,

Input: fd, feature dimension

Gain=0

$XG \leftarrow \sum_{j \in C} h_j$, $K \leftarrow \sum_{j \in C} g_j$

For i=1 to n, do

$XG_L \leftarrow 0$, $K_L \leftarrow 0$

For l in sorted (C, by y_{li}), do

$XG_L \leftarrow XG_L + h_l$, $K_L \leftarrow K_L + g_l$

$XG_R \leftarrow XG - XG_L$, $K_R \leftarrow K - K_L$

$Score \leftarrow \max \left(\frac{XG_L^2}{K_L + \lambda} + \frac{XG_R^2}{K_R + \lambda} + \frac{XG^2}{K + \lambda} \right)$

End

End

Output: Split data with max_scores

Mathematical representation

This method operates iteratively, initialized with a starting prediction commonly zero and trees are added sequentially to reduce the residual errors.

The mathematical representation of XGBoost is

$$\hat{y}_i = \sum_{s=1}^S A_s(j_i) \quad (3)$$

Where

$\hat{y}_i$ represent the model's final estimate for i^{th} point

s denotes the number of base learners

$A_s(j_i)$ denotes the j^{th} tree prediction for i^{th} point

3.4.3. LightGBM

It is a gradient boosting-based model and training is given to decision trees, provide best performance across unstructured and structured data types. It includes rapid speed for training, better accuracy and minimal usage of memory.

Working Principle of LightGBM

Construct initial tree: It initiates the process by training an initial decision tree to map the target variable using input features.

Reduce the errors: It continues by incorporating new trees; each tree is built to minimize the errors of the preceding tree.

Algorithm

The LightGBM algorithm used in this research is given below.

Algorithm 3: LightGBM DL algorithm

Input: P: training data, I: iterations

Input: l: Sampling_Ratio, (larger-gradient data)

Input: s: Sampling_Ratio, (smaller-gradient data)

Input: loss: loss_function, W: Weaker_Learner

Models $\leftarrow \{\}$, fact $\leftarrow \frac{1-l}{s}$

topM $\leftarrow (l \times (\text{len}(P)))$, ranM $\leftarrow (s \times (\text{len}(P)))$

For k=1, to n, do

preds $\leftarrow$ models.predict(P)

gb $\leftarrow$ loss(P, preds), x $\leftarrow \{1, 1, \dots\}$

Sorted $\leftarrow$ Get_Sorted_Indices(abs(gb))

TopSet $\leftarrow$ sorted[1:topM]

RandSet $\leftarrow$ RandomPick(sorted[topM:len(P)],

randM)

UsedSet $\leftarrow$ RandSet + topSet

x[randSet] X = Fact $\triangleright$ (Weight fact for the smaller-gradient data: "s" is assigned)

newModel $\leftarrow$ W(P[usedSet], gb[usedSet],

x[usedSet])

Models.append(newModel)

Mathematical Representation

The LightGBM model is represented as G (y) for given input y are denoted as the cumulative sum of predictions from each tree, $d_i(y)$ and w is individually adjusted by the learning rate.

$$G(y) = \sum_{i=1}^{Tr} w \cdot d_i(y) \quad (4)$$

3.5. Pseudocode of the Proposed Model

The pseudocode of the proposed approach is presented as follows.

Pseudocode (Complete AQI prediction pipeline)

Input: Dataset 'd' with features a and b (AQI)

1. Step 1: Initialize the Data processing,

Data pre-processing

Remove the columns with non-numeric data (here string): 'StationId and AQI_Bucket';

Data handling

Missing values are handled using Iterative Imputations, where for each missing feature of a_i , using the features from a_j , the missing value is filled;

Feature Scaling

Here, for feature processing, the 'standardize feature' method is adopted, where $a_scaled = (a - \text{mean}) / \text{std}$ is applied.

2. Step 2: Later allocate the Train-Test split: Split data into training and testing sets as "(a_train, b_train) and (a_test, b_test)" respectively;

3. Step 3: Next, Train the six models,
 - Baseline model development to predict the mean of b_{train} ;
 - Random forest ensemble model built with decision trees with number-of-jobs (n_{jobs}) as 1;
 - XGBoost model built and trained with sequential boosting (with weak learners), where the objective is set to “reg:squarederror”;
 - LightGBM: model developed is trained with leaf-wise gradient boosting with the number of leaves set to 31;
 - kNN: model is developed with neighbour set to 5,
 - Improvised DL: hidden layered feed-forward neural networking model optimized with Adam optimization; fine-tuning hyperparameters [epochs, lr] and using early stopping.
4. Step 4: Now, the prediction for each model is set to ‘b_pred’ on ‘a_test’ to obtain results;
5. Step 5: Next, the evaluation of the models is computed using the metrics “Mean_Squared_Error/ MSE and R^2 fit score”;
6. Step 6: Lastly, Models are compared for the best scores based on their performances [lowest MSE and Highest R^2];

Output: The best AQI prediction model with evaluation metrics is obtained.

Thus, the six models are trained, and the outputs are obtained. The best model is achieved with good-fit scores.

3.6. System Specification

The coefficient of determination is commonly referred to as R^2 for the selected model, a machine learning model on training data entries, to how closely predictions match with the actual values in the training set. It is recommended that the R^2 value is greater values (toward 1) implies a more accurate fit, denoting that model explains stronger explanatory power to which input features over the chosen variable.

The main Python libraries used for research are sklearn, Pandas, tensorflow, Numpy, and Seaborn. The popular Panda tools included for analyzing data are Jupyter Notebook and Anaconda.

The effectiveness of high-powered libraries and Python is found to be fit for carrying out experiments for analyzing the data, from covering traditional statistical approaches to advanced machine learning models and processing the big data.

3.7. Evaluation Metrics

R^2 , RMSE, and MSE are considered as performance evaluation metrics for this research. Scikit-learn in Python library provide built-in functions to compute the metrics.

3.7.1 R^2 Score

The R^2 value represented the variance ratio in AQI. The optimal score for R^2 value is 1.0, and in some cases, negative values also happen due to poor performance of the model. It shows how values are predicted accurately with actual values. The equation is presented as follows:

$$R^2 = 1 - \frac{\text{Residual-variance}}{\text{Total-variance}} \Rightarrow 1 - \frac{\sum_{i=1}^m (h_i - \bar{h})^2}{\sum_{i=1}^m (h_i - \bar{h})^2} \quad (5)$$

Here, h_i denotes actual data, $\hat{h}$ denotes ‘predicted data of i ’, and ‘mean value of i ’ is represented as $\bar{h}$.

3.7.2 Rooted-Mean Square Error (RMSE)

RMSE value denotes the standard deviation of the residuals. It calculates the gap between observed and predicted values. The square root of MSE is referred to as rooted mean square error:

$$MSE = \frac{1}{ME} \times \sum_{y=1}^{ME} (h_i - \hat{h}_i)^2 \quad (6)$$

$$RMSE = \sqrt{MSE} \quad (7)$$

3.7.3 Mean Absolute Error (MAE)

This value is the estimated difference between actual and predicted values that shows how closely the result of the model is aligned with observed data, referring to the same condition.

$$MAE = \frac{1}{MA} \times \sum_{y=1}^{MA} |y_i - \hat{h}_i| \quad (8)$$

3.8. Deployment and Impact

The DL model CNN + sequential LSTM built along with the other five models can be practically deployed via cloud platforms, smart-city infrastructure, IoT systems and mobile applications. The model developed can be deployed in real-time using sensors, IoT data acquisition, and pre-processing of data (data collection). Once the primary phase is completed, the model-built CNN + Sequential LSTM is loaded, and the AQI predictions are made in the second phase (prediction). The third phase is the warning phase, where the model is built and sends an alert to the users via mobile applications or through dashboards, which is later used as environmental monitoring datasets.

The scalability in AQI research refers to the ‘AQI prediction ability’, where real-time predictions and monitoring, user requests for data, geographic coverage of climatic regions, and data volume via sensors. In the current research, the model developed increases the real-time prediction and monitoring, which impacts the environmental monitoring of changes in climatic regions. The edge computing (for example: Raspberry Pi and Jetson Nano) sends

the data to the nearest data source rather than sending it to the centralized clouds. This reduces latency, improves privacy, lowers bandwidth, faster AQI predictions and offline predictability. Here, the model uses NVIDIA Nano with IoT controllers.

In a real-world case scenario or setting, all the above-mentioned components and systems are obtained along with data and assembled to predict the AQI. This provides public health benefits, government impact (policies), environmental sustainability, societal impact, research impact and industrial applications (green manufacturing and safety).

4. Findings and Results

This research adopts a heat-map, which visually represents the attribute wherein the attribute is mapped with rows and columns. Each cell denotes the value that describes the interaction between respective attributes. Each cell in the grid is represented as a color map. The outcomes and comparison of machine learning models such as extreme gradient boosting, random forest, LightGBM, baseline deep learning model and improved deep learning model are discussed in this section. To predict air quality from all major cities, i.e., Delhi, Bengaluru, Mumbai, Kolkata, Jaipur, Patna, Hyderabad, Lucknow, Gurugram, Chennai, and Ahmedabad.

4.1. Training

The data split ratio is set to 80:20 (training: testing=10 & validation=10). To optimize the improvised model, ‘Adam Optimizer’ is adopted. The hyperparameters adopted for each model are given as follows:

RF regressor model: The estimator is set to 200; random state is set to 42, and n_jobs is set to 1 (for accelerating the training).

kNN regressor model: The neighbor is set to 5 to eradicate the overfitting and underfitting issues, with no distance weighing value here.

XGB regressor model: The estimators are set to 100 with the learning rate as 0.1. The random state is set to 42 with a max-depth of 4. The objective is set to “reg:squarederror”

LightGBM regressor model: The objective is set to “regression”. Estimators, lr and random state are the same as XGB (100, 0.1 and 42), and the num_leaves is set to 31, respectively. Here, to remove the overhead issue, “force_col_wise” is set to ‘True’ with total bins as 3060 and number of data points in training as 87507 with features 12.

Baseline_Deep Learning model: Epochs = 75; batch size of 32, and the verbose is set to “True”. The layers are of size 64→32→1.

Improved model: The improvised model includes a Convolutional Neural Networking (CNN) based architecture. For time-series data, sequential long-short-term memory learning is adopted. By combining these into a CNN+LSTM-Sequential model, improved results are obtained.

4.2. Testing

The testing of the different models is carried out with the remaining 20% data-split, from the training. The station names using the “city” column are executed where city_valid is replaced with actual cities. Similarly, StationId is used to map the cities from station IDs. Once the replacement of city names and station IDs are done, the results are predicted, and the performances of each model are evaluated using MAE, RMSE and R². Figures of the air quality indices obtained through the data analysis in machine learning are as follows:

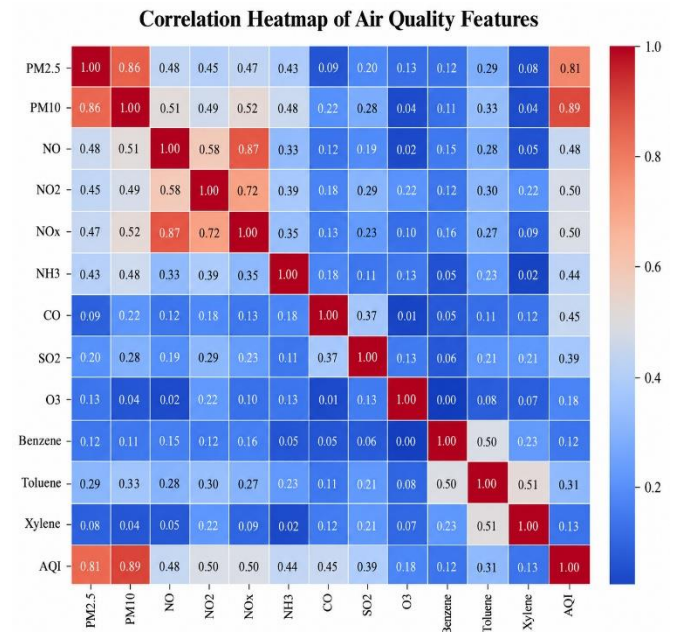


Fig. 7 Correlation heat-map of air quality features

In Figure 7, the heat map shows the darkest/brightest with the hottest and the lightest/dull colour as the coldest associations between the variables analysed here. The hottest association (≥ 0.80) exists between: PM 2.5 & PM 10; PM 2.5 & AQI; PM 10 & AQI, and NO & NOx, respectively. Similarly, the hot association (≥ 0.70) exists only between the variables NO2 and NOx.

Overall, these interpretation shows that the pollutants PM 2.5, PM 10, NO, NOx and AQI are significantly associated. Pollutants PM 2.5 and PM 10 are positively associated with AQI and are identified as the primary contributors to deteriorated air quality in India. Contrarily, the coldest (dark blue: $0.00 \geq 0.05$) shows that there exists no significance association between the variables (negative relationship). The values obtained are:

O₃ & Benzene with 0.00;

CO & O₃ with 0.01;

NO & O₃ and NH₃ & Xylene with 0.02;

PM 10 & O₃ and PM 10 & Xylene with 0.04; and

NO & Xylene, Benzene & NH₃ and Benzene & CO with 0.05.

Overall, these values show that there exists a negative association (no significant relationship) among the variables PM 10, O₃, CO, NO, NH₃, Xylene and Benzene, respectively. Next, to evaluate and plot graphs for the AQI predictions using the samples for the list of cities that achieved AQI_Bucket value >150 as the threshold value. In the following Figures (8-18), the blue color represents the 'Actual AQI', orange represents the 'Baseline DL', green represents the 'Improved DL', Red represents the 'RF', purple represents the 'XGB', and lastly brown represents the 'LGBM' models, respectively.

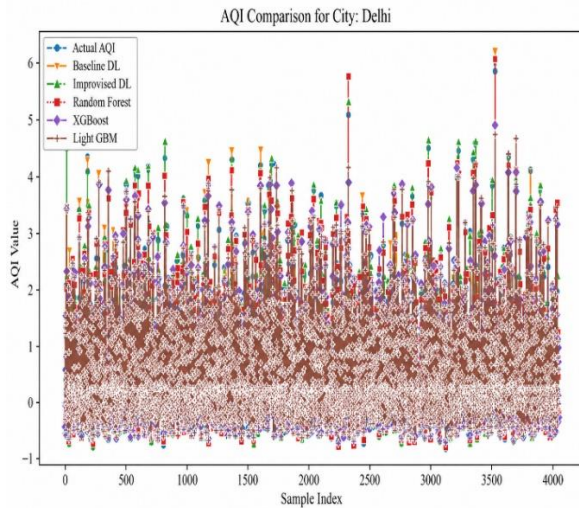


Fig. 8 AQI comparison for city: Delhi

Figure 8 illustrates the AQI comparison for Delhi. It is inferred from the graph that AQI fluctuates significantly, indicating a very high peak in pollution with 4-6 AQI ranges. The best prediction model is the Improved DL model, followed by XGB. The RF model performed moderately, while the LightGBM and Baseline models performed poorly with scattered predictions for Delhi.

Figure 9 illustrates the AQI comparison for Bengaluru. It is inferred from the graph that AQI overall is predicted reasonably well by all six models. The Improved model predicts accurately, while the decision tree models (RF, XGB and LGBM) are less responsive and less accurate in predicting pollution dynamically for Bengaluru.

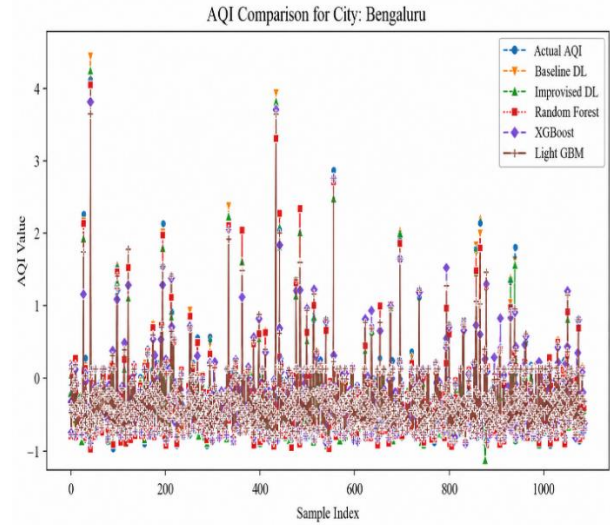


Fig. 9 AQI comparison for city: Bengaluru

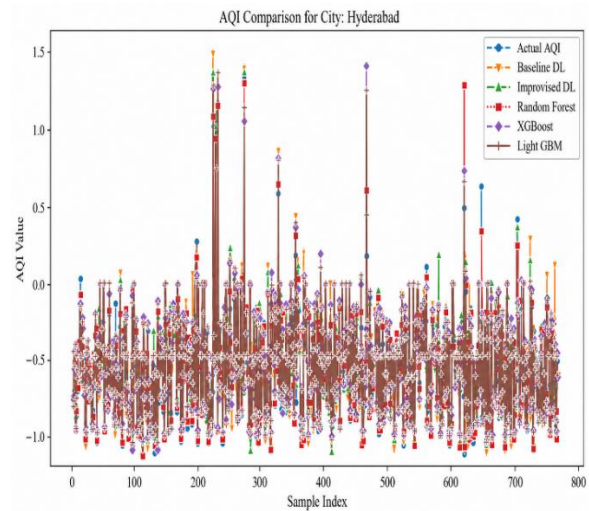


Fig. 10 AQI comparison for city: Hyderabad

Figure 10 depicts the AQI comparison for Hyderabad. It is inferred from the graph that AQI overall is predicted to be fairly or moderately polluted by all six models. The pollution level (AQI) is generally low (0 to -1.2), which indicates air quality is stable and occasionally spikes in Hyderabad.

Figure 11 depicts the AQI comparison for Lucknow. It is inferred from the above graph that AQI fluctuates significantly, indicating a moderately high peak in pollution with 3-4 AQI ranges. The best prediction model is the Improved DL, followed by XGB, LightGBM and RF in Lucknow AQI predictions.

Figure 12 illustrates the AQI comparison for Chennai. It is inferred from Figure 12 that AQI is moderately high in Chennai, with peaks varying from 2-3.5 AQI ranges. The best prediction model is the Improved DL, followed by XGB and RF in Chennai AQI predictions.

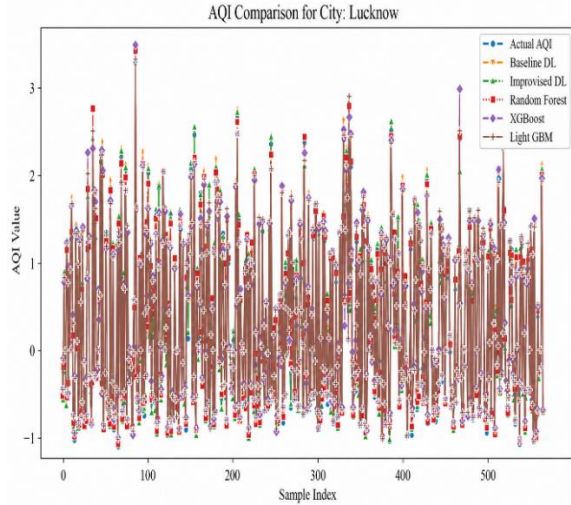


Fig. 11 AQI comparison for city: Lucknow

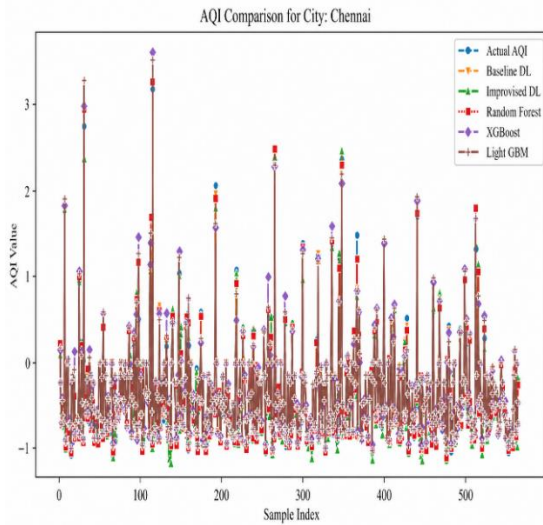


Fig. 12 AQI comparison for city: Chennai

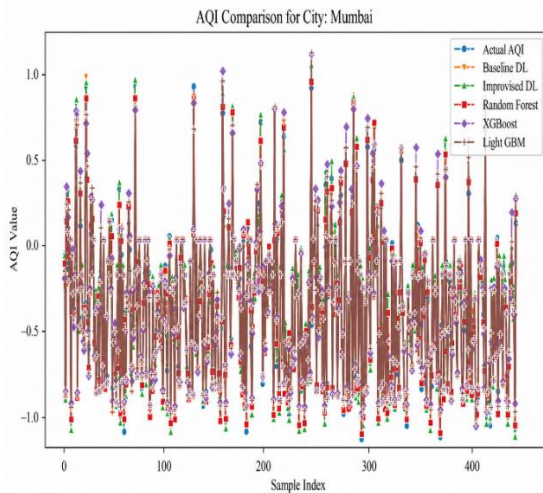


Fig. 13 AQI comparison for city: Mumbai

Figure 13 illustrates the AQI comparison for Mumbai. It is inferred from the above graph that AQI fluctuates significantly; however, it indicates that it is a moderately high pollution range, from 0.5 to 1.2 AQI ranges. It is inferred that the Mumbai AQI predictions made by the Improved model are the most accurate.

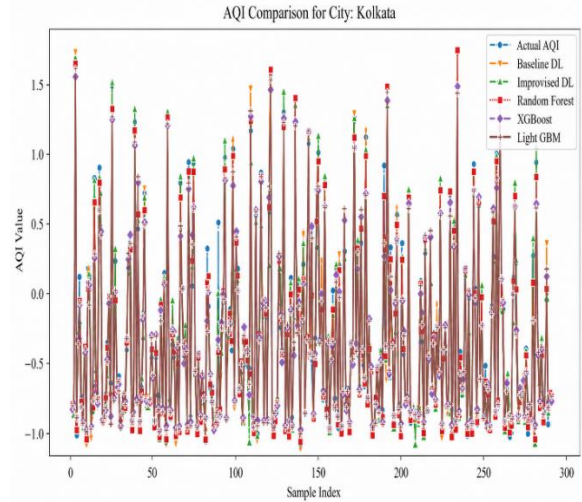


Fig. 14 AQI comparison for city: Kolkata

Figure 14 illustrates the AQI comparison for Kolkata. It is inferred from the graph 14 that AQI fluctuates significantly, indicating a moderately high peak in pollution with 1-2 AQI ranges. The best prediction model is the Improved DL, followed by Baseline and RF in Kolkata AQI predictions.

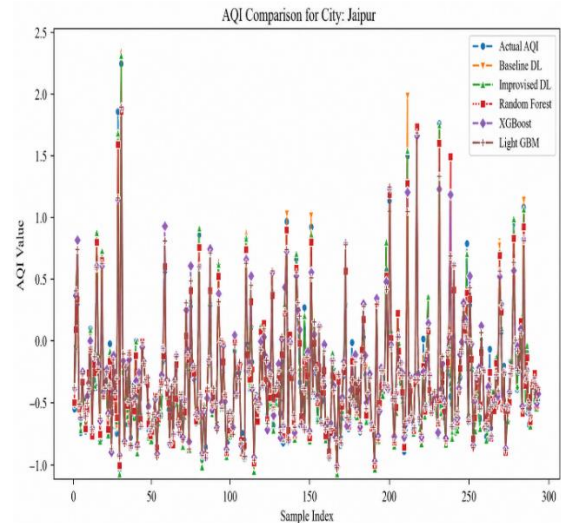


Fig. 15 AQI comparison for city: Jaipur

Figure 15 illustrates the AQI comparison for Jaipur. It is inferred from the graph 15 that AQI fluctuates significantly with a high peak in pollution, with AQI ranges of 1.5-2.5. The best prediction model is the Improved DL, followed by the Baseline and RF models in Jaipur-based AQI predictions.

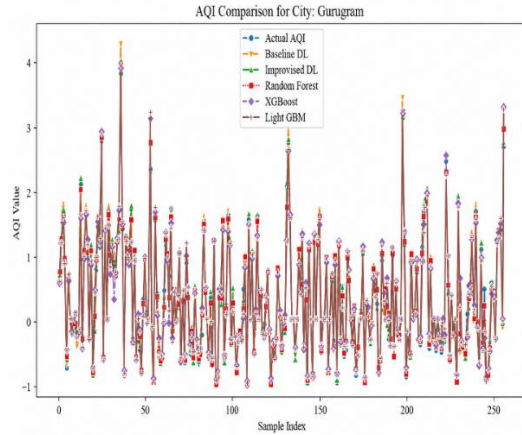


Fig. 16 AQI comparison for city: Gurugram

Figure 16 illustrates the AQI comparison for Gurugram. It is inferred from the above graph that AQI ranges moderately from the peak pollution range of 2-4.2. The best prediction model is the Improved DL, followed by the Baseline and RF models in Gurugram AQI predictions.

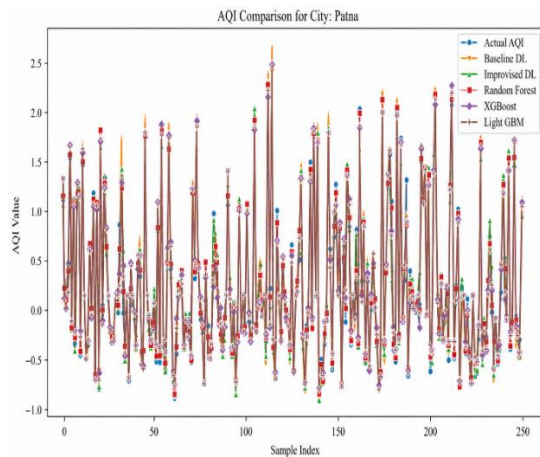


Fig. 17 AQI comparison for city: Patna

Figure 17 illustrates the AQI comparison for Patna. It is inferred from the above graph that AQI fluctuates significantly, indicating a moderately high peak in pollution with 1.5-2.7 AQI ranges. The best prediction model is the Improved DL, followed by RF and Baseline in Patna AQI predictions.

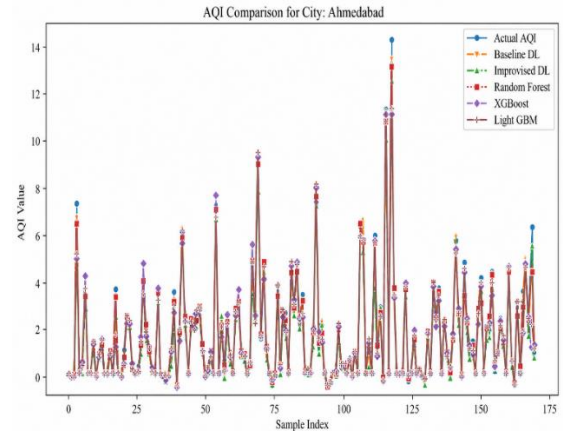


Fig. 18 AQI comparison for city: Ahmedabad

Figure 18 illustrates the AQI comparison for Ahmedabad. It is inferred from the above graph that AQI is significantly higher, indicating a very high peak in air pollution, with the highest AQI ranges of 8-14. The best prediction model is the Improved DL, followed by Baseline, RF and XGB, in Ahmedabad AQI predictions. Thus, overall, the Improved DL is the best predictive AQI model while Baseline DL, RF and XGB models are the next best (decision tree models) predictive models in terms of predicting Air Quality Index of the select Indian cities.

Figure 19 depicts the model performance comparison on the validation set. Performance is evaluated for different machine learning techniques adopted in this research, using R2, RMSE and MAE.

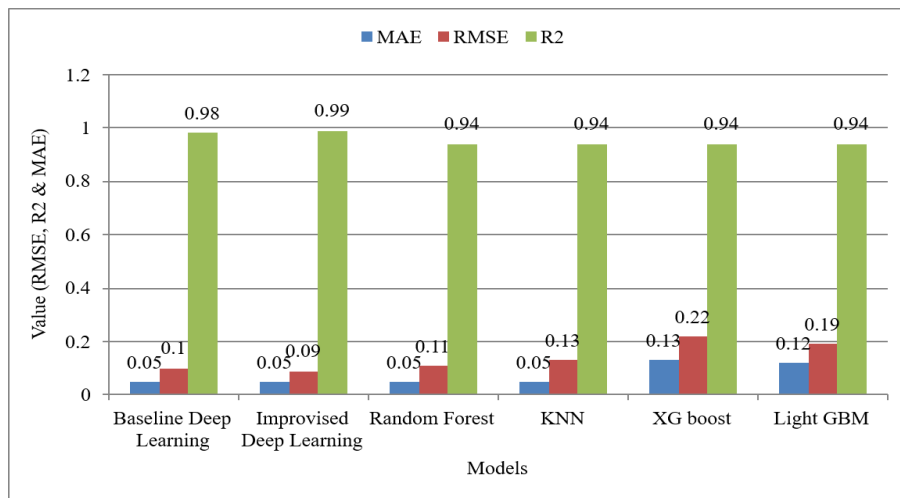


Fig. 19 Model performance comparison on the validation set

Inference: From Figure 19, it is evident that the improvised DL model gained the highest R2 fit score of 0.99, while the second-highest is the baseline DL model with 0.98 R² fit score, followed by RF, kNN, LightGBM and XGBoost models with 0.94 each, respectively. The MAE and RMSE are observed to be higher in the latter models (RF, XGB, kNN and LightGBM) than the baseline (MAE = 0.05 and RMSE = 0.10) and improved (MAE = 0.05 and RMSE = 0.09) DL models. Thus, it is proven significantly that the improvised DL model has outperformed the baseline and other regressor models adopted, significantly. Table 3 represents the comparison of machine learning techniques in predicting air quality:

Table 3. Comparison of machine learning techniques in predicting air quality

Machine learning techniques	MAE	RMSE	R2
Baseline deep learning	0.05	0.10	0.98
Improved model	0.05	0.09	0.99
Random Forest	0.05	0.11	0.94
kNN	0.05	0.13	0.94
XGBoost	0.13	0.22	0.94
LightGBM	0.12	0.19	0.94

4.3. Statistical Interpretation

The statistical interpretations and the significance of the six models have been tested and assumed using ANOVA with repeated runs (cross-validation folds), using Python. The methods used are paired t-tests, one-way ANOVA and the Friedman test to find the significance between models developed. The statistical interpretations in Table 4 show that the improvised model scored the highest variance, whereas the baseline model also had high predictive accuracy. However, the lowest is observed in XGB, which is weaker than

LightGBM. Similarly, RF and kNN are observed to have less error and better predictive accuracy than XGB.

Table 4. Significance analysis

Models	p-value	Statistical interpretation
Baseline deep learning	<0.05	High predictive accuracy and low error
Improved model	<0.01	Highest variance and best performing model
Random Forest	<0.05	Moderate error and good predictive accuracy
k-NN	<0.05	Higher RMSE but acceptable predictive accuracy
XGBoost	>0.05	A very high prediction error rate compared to other DLs
LightGBM	=0.05	Weaker than DL models but better than XGB

4.4. Comparative Analysis

Existing studies on AQI prediction mostly use CNN, LSTM and sometimes Transformer architectures. The DL models developed are unique because of the truncation and imputation methods used here. Similarly, the datasets used here are large are not been used in any other studies yet. Hence, comparing the models is not significant, but comparing the error and accuracy of the models gives a basic understanding of the improved DL prediction variance and accuracy. Table 5 shows some studies and their performance:

Table 5. Comparative analysis with benchmarking results

Author and Year	Model/ Architecture	Performance Metrics	Findings
Wang et al., [33]	CNN + Improved LSTM	R ² = 0.9601	AQI sequences based on temporal dependencies are captured in long-term basis
Guo et al., [34]	CNN + LSTM	RMSE = 5.236	Combines temporal and spatial learning in AQI forecasting
Sun et al., [35]	Transformer + BiLSTM	R ² = 0.9781	Self-attention that captures long-range dependencies
Donget al., [36]	EMD Transformer + BiLSTM	RMSE = 5.6853	Non-stationary and non-linear based AQI data with strong performance
Hu et al., [37]	CNN + LSTM + KAN Hybrid approach	R ² = 0.92 → 0.99	Established reliable accuracy across multiple climatic-regions
Proposed model	CNN + Sequential LSTM	R ² = 0.99, RMSE = 0.10	Large linear data with very high variance and accuracy in AQI prediction.

From Table 5, it is observed and apparent that the proposed CNN + Sequential LSTM model gained the highest accuracy with a fit measure of 0.99 (R²) and with minimal

error (RMSE = 0.10) compared to other DL models. The benchmarking results from Table 5 show that the improvised CNN + sequential LSTM models achieved the highest

variance and accuracy. When temporal and spatial patterns are identified and recognized in AQI predictions, the models' performance declines while CNN models struggle with long-term dependencies, yet improves features extraction via multiple sensor inputs. Thus, a hybrid approach is found significant than adopting traditional models. This proves that the developed CNN + Sequential LSTM model is more significant in AQI predictions than other DL models.

5. Discussion and Conclusion

5.1. Discussion

The authors [38] compared and predicted AQI using various regression techniques, namely decision tree regression, linear regression and random forest regression. Decision tree models are evaluated with these metrics R^2 , RMSE and MAE values, such as 93.21, 7.073 and 0.738, respectively.

Random Forest model is evaluated with these metrics R^2 , RMSE and MAE values, such as 93.08, 7.141, and 0.795, respectively. Linear regression models had low performance with a high value RMSE of 42.7 percent and MAE of 32.19 percent, respectively. Linear regression is not suitable for predicting air quality.

The authors [39] proposed a new Multimodal Imputer (MI) multi-modal autoencoder and also incorporated the XGB algorithm to predict AQI. From the findings of the research, it was noted that the proposed method achieved a remarkable RMSE value of 93.87 percent and R^2 score as 97.14 percent. Novel method performed well in regression and classification in prediction. The researchers [40] combined an LSTM + Light GBM model to predict AQI. Performance evaluation metrics studied in this research include R^2 and RMSE. Values of R^2 and RMSE are 0.969 and 5.09, respectively. Overall, the proposed method's improvised deep learning model with additional dense layers has performed well when compared with other models; acquiring the best R^2 fit of 0.99, MAE of 0.05 and RMSE score of 0.10.

Table 6. Findings of existing models' performances

Author and Year	Machine learning techniques	R^2 value	RMSE	MAE
Al-Eidi et al [38]	Random Forest	93.08	7.141	0.795
Rao et al [39]	XGB algorithm	97.14	93.87	-
Huang et al [40]	LSTM + LightGBM	96.9	5.09	-
Proposed research	Improvise model	99	0.10	0.05

Table 6 shows the derived score as the model's performance for prediction model evaluation.

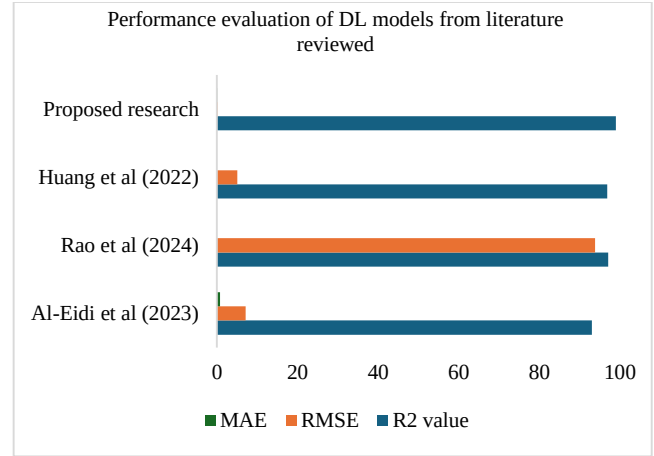


Fig. 20 Performance evaluation from literature

Figure 20 portrays the performance evaluation analysis of a few studies reviewed from existing literature, with a focus on air quality predictions using DL models.

5.2. Significance of DL Models

Traditional models like CNNs, Recurrent NN (RNNs), Deep Belief Networks (DBNs), Deep NNs (DNNs), and more lack spatial patterns and temporal dependencies in complex non-linear associations. These spatial and temporal correlations, hybridization, feature extraction and non-linear mapping make deep learning more significant and effective than other models, especially in AQI prediction for sensitive datasets. Deep learning models are well-suited for AQI forecasting, monitoring and help towards better environmental policy-based decision making.

5.3. Environmental Implications

Deep learning models using the CNN and LSTM as a hybrid approach outperform the traditional AQI prediction methods via complex non-linear association capturing. By utilizing the feature scaling and data normalization methods: iterative imputation with a standard scaler, this study provides a better predictive model in AQI forecasting, which is useful in early health warnings (environmental hazards), emission controls, and urban planning sustainability, especially in poor countries and developing countries.

5.4. Practical Applications

The developed model is effective and significant, as it allows the investigators to monitor and predict the high alerts instantaneously in cities and heavy-industrial regions, when the air quality is ≥ 150 . Similarly, in sensor optimization and data-driven decision support areas, the Air Pollution Monitoring (APM) systems are installed and tested for analyzing the AQI's impact on environmental policies, prior to implementing the real-time monitoring. The current DL model, which is improvised is significant and forecasts the AQI levels effectively.

5.6. Conclusion

Air pollution is one of the major health hazards faced by people in daily life. Thus, air quality is important to combat people's health. The intention of the study is to predict AQI using various machine learning techniques in selected cities. Delhi, Bengaluru, Mumbai, Kolkata, Jaipur, Patna, Hyderabad, Lucknow, Gurugram, Chennai, and Ahmedabad are included in the research for predicting AQI.

Machine learning-based regressor techniques used in this research are Random Forest, XGBoost, and LightGBM, along with the developed baseline deep learning model and the improvised deep learning models, respectively. AQI prediction is done using all selected machine learning models.

It is evident that the improvised model acquired very high R^2 scores of 0.99, which is revealed as the best AQI predictive model. Followed by is the baseline model with second highest AQI prediction, with R^2 fit measure of 0.98. Lastly, the RF and LightGBM showed a bit less predictive capability than other models, with R^2 fit measures of 0.94 each, and lastly, the XGBoost model was the least predictive among the other three models, with 0.93 R^2 fit score, respectively.

It is thus apparent from the analysis that the improvised deep learning model with additional dense layers has outperformed well (5% higher than RF and LightGBM models; and 6% higher than the XGBoost model) when compared with other regressor models, with the best R^2 fit

score of 0.99 and RMSE score of 0.05 with MAE as 0.10. In future, this work can be expanded by combining advanced machine learning techniques to predict AQI.

5.7. Limitations and Future Work

Spatio-Temporal data have not been used in this research due to the huge size of the dataset. Similarly, advanced techniques (Explainable AI) can be applied in the future to explain the performance of the models proposed. In the future, the Spatio-Temporal data, along with Explainable AI, will be given preference in AQI prediction model development research works.

Similarly, different datasets would be tested using advanced models in the future. Multiple sources, like mobile sensors and satellites, will be taken into consideration, and their impact on air quality prediction will be tested in the future. Further, this research could be extended by developing advanced models like transformer models, and the same dataset could be tested against these models to evaluate their performance in terms of the prediction of the air quality index.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research received no funding from any external agency.

References

- [1] Raymond Shakespeare, and Sowmya Uttham, Environmental Pollution and its Types, African Journal of Water Conservation and Sustainability, vol. 9, no. 2, pp. 1-2, 2021. [Online]. Available: <https://www.internationalscholarsjournals.com/articles/environmental-pollution-and-its-types.pdf>
- [2] Shabiha Hossain, Amit Hasan Anik, and Mahbub Alam, "Radioactivity and its Possible Impact on Environment and Human Health: A Review," *International Research Journal of Environmental Sciences*, vol. 10, no. 4, pp. 24-34, 2021. [[Publisher Link](#)]
- [3] Walihabib Hemmat, Atiq Mohammad Hesam, and Hamza Atifnigar, "Exploring Noise Pollution, Causes, Effects, and Mitigation Strategies: A Review Paper," *European Journal of Theoretical and Applied Sciences*, vol. 1, no. 5, pp. 995-1005, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Franklin Akwasi Adjei, and Augustine Afriyie, "Air Pollution as a Global Health Risk: Emerging Evidence, Interventions and Policy Insights," *World Journal of Advanced Research and Reviews*, vol. 27, no. 1, pp. 1925-1940, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Satsita Khasanova, Elina Alieva, and Aishat Shemilkhanova, "Environmental Pollution: Types, Causes and Consequences," *BIO Web Conferences*, vol. 63, pp. 1-7, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Sumit Kumar Mavi, "Analyzing Air Pollution in India: Causes, Impacts, Government Policies and Challenges," *International Journal of Geography, Geology and Environment*, vol. 4, no. 2, pp. 254-259, 2022. [[Publisher Link](#)]
- [7] Pratik D. Sengani, and Gaurav Agarwal, Air Pollution its Source and Effects on the Environment: Present Scenario in India-A Review, *International Journal of Research and Analytical Reviews*, vol. 11, no. 1, pp. 550-562, 2024. [Online], Available: <https://ijrar.org/papers/IJRAR24A3260.pdf>
- [8] Kiera Jefferson, Air Pollution's History in India: A Troubled Past, 2025. [Online]. Available: <https://shunwaste.com/article/when-did-air-pollution-start-in-india>
- [9] Bridget M. Kuehn, "WHO: More than 7 Million Air Pollution Deaths Each Year," *JAMA*, vol. 311, no. 15, 2014. [[Google Scholar](#)] [[Publisher Link](#)]

- [10] Sarath K. Guttikunda, Rahul Goel, and Pallavi Pant, "Nature of Air Pollution, Emission Sources, and Management in the Indian Cities," *Atmospheric Environment*, vol. 95, pp. 501-510, 2014. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] WAQR, Worldwide Air Quality Report 2024, Global Air Quality Trend in 2024 by Highlighting the Continents, Countries and Cities in the World with the Highest AQI Levels, 2025. [Online]. Available: <https://www.aqi.in/assets/files/world-air-quality-report-2024-by-aqi.in.pdf>
- [12] Ruchita Nehete, and D.D. Patil, "Air Quality Prediction using Machine Learning Algorithms," *International Journal of Creative Research Thoughts*, vol. 9, no. 6, pp. 365-368, 2021. [[Publisher Link](#)]
- [13] Binzhe Zhang et al., "Air Quality Index Prediction in Six Major Chinese Urban Agglomerations: A Comparative Study of Single Machine Learning Model, Ensemble Model, and Hybrid Model," *Atmosphere*, vol. 14, no. 10, pp. 1-22, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Anuradha Yenikar et al., "Explainable Forecasting of Air Quality Index using a Hybrid Random Forest and ARIMA Model," *MethodsX*, vol. 15, pp. 1-18, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Suprateek Halsana, "Air Quality Prediction Model using Supervised Machine Learning Algorithms," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 6, no. 4, pp. 190-201, 2020. [[CrossRef](#)] [[Google Scholar](#)]
- [16] Zuhra Sadridin, Remudin Reshid Mekuria, and Mekia Shigute Gaso, "Machine Learning Models for Advanced Air Quality Prediction," *Proceedings of the International Conference on Computer Systems and Technologies*, Association for Computing Machinery, New York, NY, United States, pp. 51-56, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Zulfani Alfasanah et al., "Monitoring Air Quality Index with EWMA and Individual Charts using Xgboost and SVR Residuals," *MethodsX*, vol. 14, pp. 1-21, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Feng Chen, Lei Wang, and Hongyu Deng, "A Novel Combined Model for Air Quality Index Forecasting in Changchun," *Atmosphere*, vol. 14, no. 10, pp. 1-20, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Thomas M.T. Lei, Stanley C.W. Ng, and Shirley W.I. Siu, "Application of ANN, XGBoost, and other ML Methods to Forecast Air Quality in Macau," *Sustainability*, vol. 15, no. 6, pp. 1-17, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Qian Liu, Bingyan Cui, and Zhen Liu, "Air Quality Class Prediction using Machine Learning Methods based on Monitoring Data and Secondary Modeling," *Atmosphere*, vol. 15, no. 5, pp. 1-23, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Xuehu Liu et al., "PM_{2.5} Concentration Prediction based on LightGBM Optimized by Adaptive Multi-Strategy Enhanced Sparrow Search Algorithm," *Atmosphere*, vol. 14, no. 11, pp. 1-24, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Xiaowen Zhang, Xuchu Jiang, and Ying Li, "Prediction of Air Quality Index based on the SSA-Bilstm-Lightgbm Model," *Scientific Reports*, vol. 13, no. 1, pp. 1-14, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Narayani Sunil Pillai et al., "Air Quality Prediction System using LightGBM," *International Research Journal of Engineering and Technology*, vol. 7, no. 7, pp. 3861-3866, 2020. [[Publisher Link](#)]
- [24] Lakindu Mampitiya et al., "Machine Learning Techniques to Predict the Air Quality using Meteorological Data in Two Urban Areas in Sri Lanka," *Environments*, vol. 10, no. 8, pp. 1-18, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Raey Faldo et al., "APD-BayNet: Jakarta Air Quality Index Prediction using Bayesian Optimized Tabnet," *IEEE Access*, vol. 13, pp. 57734-57752, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Anil Utku, "Hybrid CNN-LSTM Model for Air Quality Prediction: A Case Study for Gurugram," *Journal of Soft Computing and Artificial Intelligence*, vol. 5, no. 1, pp. 33-40, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Anurag Barthwal, and Amit Kumar Goel, "Advancing Air Quality Prediction Models in Urban India: A Deep Learning Approach Integrating DCNN and LSTM Architectures for AQI Time-Series Classification," *Modeling Earth Systems and Environment*, vol. 10, no. 2, pp. 2935-2955, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] T. Sreenivasulu, and G. Mokesh Rayalu, "Accurate Hourly AQI Prediction using Temporal CNN-LSTM-MHA+GRU: A Case Study of Seasonal Variations and Pollution Extremes in Visakhapatnam, India," *Results in Engineering*, vol. 27, pp. 1-22, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Ekata Mohapatra, Mira Das, and Smita Rath, "Deep Learning-based AQI Forecasting: A CNN-LSTM Model with Visual Insights from SHAP-LIME and PDP," *Discover Applied Sciences*, vol. 7, no. 11, pp. 1-31, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Vopani, Air Quality Data in India (2015-2020): Air Quality Index (AQI) and Hourly Data Across Stations and Cities in India, 2020. [Online]. Available: <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india/data>
- [31] Seth A. Horn, and Purnendu K. Dasgupta, "The Air Quality Index (AQI) in Historical and Analytical Perspective a Tutorial Review," *Talanta*, vol. 267, pp. 1-15, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Nenny Anggraini et al., "A Comparative Analysis of Random Forest, XGboost, and LightGBM Algorithms for Emotion Classification in Reddit Comments," *Journal of Informatics Engineering*, vol. 17, no. 1, pp. 88-97, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] Jingyang Wang et al., "An Air Quality Index Prediction Model based on CNN-ILSTM," *Scientific Reports*, vol. 12, no. 1, pp. 1-16, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [34] Zicheng Guo et al., “Air Quality PM2.5 Index Prediction Model based on CNN-LSTM,” *arXiv preprint arXiv*, pp. 1-7, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Zhuoran Sun, Qing Zhang, and Guici Chen, “Air Quality Index Prediction based on Transformer Encoder-CNN-BiLSTM Model,” *Atmosphere*, vol. 17, no. 3, pp. 1-15, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] Jie Dong, Yaoli Zhang, and Jiang Hu, “Short-Term Air Quality Prediction based on EMD-Transformer-Bilstm,” *Scientific Reports*, vol. 14, no. 1, pp. 1-17, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [37] Yue Hu, Yitong Ding, and Wenjing Jiang, “Geographically Aware Air Quality Prediction through CNN-LSTM-KAN Hybrid Modeling with Climatic and Topographic Differentiation,” *Atmosphere*, vol. 16, no. 5, pp. 1-22, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [38] Shorouq Al-Eidi et al., “Comparative Analysis Study for Air Quality Prediction in Smart Cities using Regression Techniques,” *IEEE Access*, vol. 11, pp. 115140-115149, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [39] Routhu Srinivasa Rao et al., “Multimodal Imputation-based Multimodal Autoencoder Framework for AQI Classification and Prediction of Indian Cities,” *IEEE Access*, vol. 12, pp. 108350-108363, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [40] Wenchao Huang et al., “Research on Air Quality Prediction based on Improved Long Short-Term Memory Network Algorithm,” *PeerJ Computer Science*, vol. 8, pp. 1-16, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]