Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, No. 5, October, 2025

LRX: A Hybrid-based Real-Time Air Quality Index Prediction and Visualization Model

J. Jayapradha [1, 2], Su-Cheng Haw [2*], Naveen Palanichamy [2], V. Arunesh [1], Surajith Pranav [1], T. Senthil Kumar [1]

[1] *Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu 603203, India.*

[2] *Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.*

Abstract

Accurately predicting the air quality index significantly reduces health risks and supports urban environmental planning. This paper presents LRX, a hybrid predictive model, for Air Quality Index (AQI) prediction. The model employs Long short-term memory to capture temporal dependencies, Random Forest to fine-tune the features, and Extreme Gradient Boosting to enhance the final predictions. The objective of the study is to build a model that can accurately predict air quality index numbers in real time for many cities in India. The proposed model LRX design influences the depth of each algorithm to enhance accuracy and generalization. The experimental results show the model's ability to predict the AQI forecast of various cities in India with a root mean square error of 0.014 and R^2 of 0.948, performing better compared to the models individually. To enhance this, a Stream lit-based user interface has been developed to enable real-time AQI predictions and visualization. The interface incorporates tabs for interactive inputs, model selection, graphical representation of predicted trends, ensuring accessibility and usability, and enhancing the practical applicability of the proposed model. This easy-to-navigate tool not only makes the prediction process more accessible but also helps bridge the gap between complex model results and practical environmental decision-making, enhancing the overall impact of the research. This research contributes to air quality prediction by presenting a robust modelling approach that can be applied in the real world.

Keywords:
Air Quality Index (AQI);
Long Short-Term Memory;
Random Forest;
Extreme Gradient Boosting;
Stream Lit;
Decision-Making.

Article History:

Received:	26	February	2025
Revised:	04	August	2025
Accepted:	11	August	2025
Published:	01	October	2025

1- Introduction

Air pollution is one of the most significant global problems, having disastrous effects on both human health and general environmental balance. Air pollution poses a significant threat and environmental challenge to public health, the atmosphere, and ecology. With the increasing sizes of cities and more industrial production coming into existence, there is an increase in pollutants, which leads to poor-quality air. Pollutants and fine particulate matter (PM) that contribute to air pollution include nitrogen dioxide (NO_2), carbon monoxide (CO), carbon dioxide (CO_2), ozone (O_3), and sulphur dioxide (SO_2) [1, 2]. Air quality prediction involves various factors, intricately connected to atmospheric conditions and exhibiting time dependencies. The Air Quality Index (AQI) is a vital tool since it gives a clearly defined way of measuring how bad the air is daily. Precise and appropriate estimates of AQI can enable experts, scholars, and parties to take preemptive measures to alleviate the hazards associated with deprived air quality.

The AQI could alert people to decide their daily activities, such as exercise, and help policymakers formulate laws and policies aimed at protecting the public. But achieving high accuracy in predicting AQI is not easy because air

* **CONTACT**: sucheng@mmu.edu.my

DOI: http://dx.doi.org/10.28991/ESJ-2025-09-05-010

© 2025 by the authors. Licensee ESJ, Italy. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC-BY) license (https://creativecommons.org/licenses/by/4.0/).

pollution is a highly variable and multi-factorial phenomenon. Generally, pollution depends on the flow of vehicular emissions, industrial releases, weather, and other chemical processes in the atmosphere. Moreover, there are very sharp changes that can be caused by unusual weather conditions (inflection points) and significant variations in air pollution because air itself changes over location and time [3, 4]. All these work in synergy, limiting the predictability of impacts on air quality, something that traditional models cannot capture adequately. Conventional air quality examination depends on installed ground-based instrument systems that prepare precise but repeatedly faltered and spatially constrained data. With collective suburbanization and developed movement, there is an expanding pressure for real-time, analytical, and spatially widespread air quality data. This has led to a rising importance of smart paradigms that merge ecological data cascades, machine learning systems, and sophisticated visualization methods.

The studies show that poor air quality can hamper these health problems and lead to increased hospitalization and healthcare attendance. Maximizing the understanding of AQI is vital to facilitate the appropriate formulation of public health policies and interventions. Subramanian et al. [5] proposed an Auto Regressive Integrated Moving Average model for statistical data prediction, which was incorporated with LSTM using encoder-decoder architecture, Quantum Particle Swarm Optimization, and XGBoost. The model showed great expertise by maintaining low error rates and high determination coefficients, maintaining high predictability. The authors implemented a neuro-fuzzy (neural network + fuzzy logic) logic to represent AQI data, training both deep neural networks (DNN) and the Markov model separately. The input dataset was fed to DNN and then to Markov, resulting in better accuracy than individual models. Sarkar et al. [6] proved that the traditional methods of autoregressive integrated moving average (ARIMA) and support vector regression were not able to identify the data series from the air-polluted data. As such, two models, namely EMD-SVR-Hybrid and EMD-IMFs-Hybrid, have been proposed to estimate the AQI.

Figure 1 reveals "Hazardous" AQI levels across northern regions; the central and western regions, including parts of Maharashtra and Uttar Pradesh, display "Unhealthy" levels. In contrast, southern India and northeastern areas report relatively cleaner air with "Good" AQI. The map highlights India's seasonal trend, where post-monsoon air quality dips drastically, particularly in northern areas, as pollutants get trapped near the surface. To address and overcome the challenges of enhancing the accuracy of the AQI predictions, this paper introduces a hybrid model called the LRX (which is derived from the combination of LSTM + RFR + XGBoost) model, which incorporates a combination of three machine learning methods. The outcome of this study can enhance decision-making in public health and urban planning to improve air quality standards and increase people's health in communities.

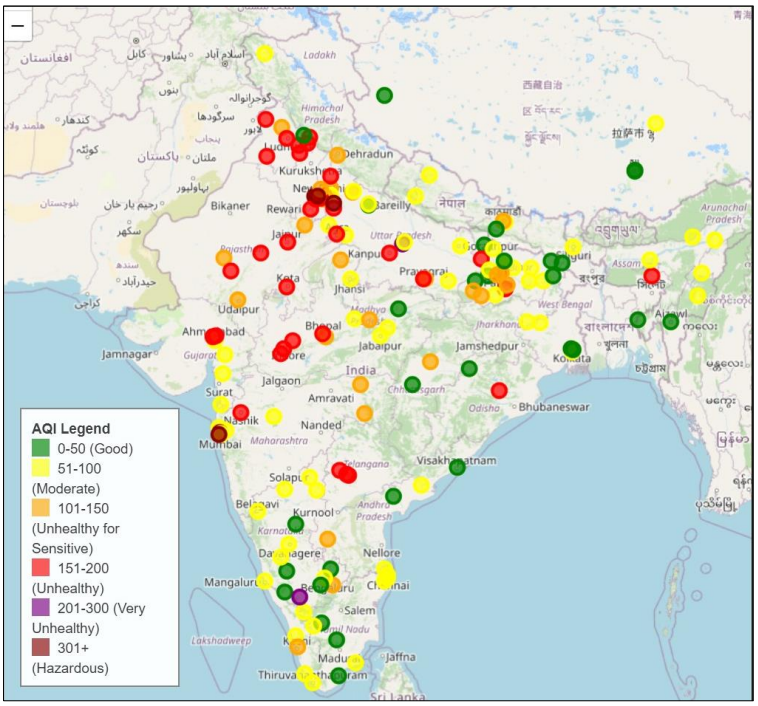


Figure 1. AQI map of India

2- Literature Survey

The AQI, as mentioned, is crucial from the public health perspective and relates to respiratory and cardiovascular ailments. The presence of high AQI values, particularly particulate matter (PM2.5 & PM10), increases the rates of asthma, bronchitis, and other respiratory diseases. Sarkar et al. [6] have been conducted on the discussion below. The research work proposed combining Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for

AQI prediction. The results showed that the proposed hybrid LSTM-GRU model outperforms other standalone machine learning models, achieving an MAE of 36.11 and an R^2 value of 0.84. The Random Forest Prediction algorithm [7] was employed in the study to predict carbon monoxide (CO) and nitrogen dioxide (NO_2) concentrations in the Amman-Zarqa region of Jordan. The key influencing factors, such as relative humidity, wind direction, and land surface temperature, are included with MAE values ranging from 0.11 to 0.18 for CO and 3.78 to 7.30 for NO_2 models. A comparative analysis of deep learning algorithms [8] evaluated LSTM and achieved varying Root Mean Square (RME) for daily and hourly data, highlighting its potential, especially for hourly predictions. The performance of the Random Forest model was compared [9] with Support Vector Regression (SVR) and Artificial Neural Networks (ANN) for AQI prediction, and it proved that Random Forest showed high accuracy.

The Grey Level Co-occurrence Matrix (GLCM) technique [10] was utilized for feature extraction in AQI prediction, optimized by the Moth Flame Optimization Algorithm (MFOA) and combined with SVR and LSTM for better predictive results. The LSTM model predicting AQI [11] used nine parameters in Shanghai to show high precision but faced overfitting due to sample imbalance, indicating that dataset choice impacts model performance. A comparison of SVR and Random Forest Regression (RFR) [12] revealed that RFR performed better, while SVR exhibited increasing time complexity with larger datasets, becoming impractical for extensive samples. The paper proposed a CNN-ILSTM model [13] for AQI prediction, outperforming traditional regression models. However, it struggled with extreme value predictions, highlighting a limitation in its applicability. A combination of grey wolf optimization and decision tree algorithms resulting [14] in high prediction accuracy (86%-98%) for AQI, emphasizing the importance of optimal feature selection, is proposed. Duan et al. [15] explored the ARIMA-DBO-CNN-LSTM model, showing potential in AQI prediction but facing issues with partial fitting, leading to re-averaging challenges. AQI prediction models using SVR, RFR, and Catboost Regression were tested in Indian cities [16], revealing a 6%-24% accuracy increase when applying the Synthetic Minority Oversampling Technique (SMOTE). The work highlighted conventional AQI prediction methods [17], recent advancements in deep learning applications, and future directions for the field's development and conducted [18] a general comparison of Random Forest, Linear Regression, and Naive Bayes, indicating that Random Forest achieved the highest accuracy, followed by Linear Regression. The potential of remotely sensed data for AQI prediction was explored [19], with single and multiple linear regression models demonstrating high predictability, which resulted in the multiple linear regression performing best. The work analyzed forecasting in Visakhapatnam using Random Forest and Catboost [20] and achieved high prediction accuracy, with minimal correlation differences in their performance.

Farooq et al. [21] utilized quantum computers in conjunction with SVM for AQI prediction, achieving 97% accuracy with quantum SVM, indicating future potential for complex computations. Six machine learning models, including SVM, RF, and AdaBoost [22], were applied to predict AQI using data from Taiwan, revealing that stacking ensemble and AdaBoost outperformed traditional models. This paper proposed a comparative analysis of Seasonal Autoregressive Integrated Moving Average (SARIMA), SVM, and LSTM models [23] for AQI prediction using data collected from multiple air quality monitoring stations in Ahmedabad from January 2015 to January 2021. The authors also evaluated the models using performance metrics such as R2 Score, MSE, RMSE, and MAE. They showed that the SVM model with the RBF kernel outperforms other models in predicting AQI for Ahmedabad city. This model uses an LSTM layer to analyze Taiwan's AQI dataset and predict the concentrations of four key pollutants (PM2.5, PM10, O_3 , and NO_2) [24]. The study compared LSTM_ON to the existing XGB_NON model and found LSTM_ON to be more accurate. The research paper proposed a unique approach for predicting AQI using pollutant concentrations and meteorological data [25]. The researchers developed a three-module system that first predicts the concentration of eight key pollutants using ARIMA and ANN, then uses these values along with meteorological factors like wind speed and temperature to forecast the AQI. This system was tested against several variations of Support Vector Regression (SVR) models with different kernel functions and input parameters, resulting in improved accuracy of the predictions, with the linear kernel function performing best overall.

A hybrid framework combined several machine learning techniques [26], utilizing wavelet decomposition for time series data and employing a BiLSTM model optimized with Particle Swarm Optimization, improving prediction accuracy. The Cuckoo Search algorithm optimized the LSTM model for AQI prediction [27], demonstrating improved accuracy over SVR, BP neural networks, and standard LSTM models. Daily air quality data analysis from Henan Province evaluated ten regression models [28] and highlighted RF and Gradient Boosting as superior in predictive accuracy and generalization. Gupta et al. [29], in their study of air quality trends in India, emphasized seasonal changes and the impact of the COVID-19 lockdown on AQI levels, with PM2.5 and PM10 showing the most significant effects. A Bidirectional LSTM (BiLSTM) [30] with an attention mechanism was proposed to enhance feature extraction for AQI prediction, resulting in improved accuracy compared to standard LSTM models. Though various studies and experiments have been carried out, the accurate prediction of the AQI value remains challenging. As such, in this paper, we propose the LRX model to improve the accuracy of prediction.

3- Implementation of LRX (LSTM + RFR + XGB) Model

This paper proposes a hybrid machine learning model for AQI forecasting to improve the accuracy and reliability of predictions. The model begins with collecting and preprocessing historical AQI data and other relevant features. The stages involved in the LRX model are (i) cleaning the data, (ii) encoding categorical variables, (iii) scaling the features to ensure consistency, and (iv) readiness for model training.

3-1-Air Quality Dataset

The experiment is carried out in the city_day.Csv dataset. The dataset consists of 29,531 entries, each representing daily air quality data for various cities in India. The data consists of multiple pollutants and their concentrations, as well as calculated AQI values. The dataset is available at the link below: city_day.csv (*kaggle.com*). The specific features used in the dataset are PM2.5, PM10, NO, NO_2, NH_3, CO, SO_2, O_3, BENZENE and XYLENE, where AQI is the target. In the dataset no meteorological features such as temperature and humidity are included in the dataset, which restricts the proposed model range and generalizability. The sample raw dataset is shown in Table 1, while the different levels of AQI for the public are shown in Table 2.

Table 1. AQI value and conforming ambient concentrations for the identified pollutants

City	Date	PM 2.5	PM 10	NO	NO_2	NH_3	CO	SO_2	O_3	BENZ ENE	XYL ENE	AQI	AQI_ BUCKET
AMD	1/1/15	73.24	141.6	0.92	18.22	23.48	0.92	27.64	133.36	0	0	166.46	Poor
BLR	2/1/15	30.65	70.46	3.26	17.33	20.36	0.33	3.54	10.73	0.56	0	91	Moderate
MAA	3/1/15	173.5	48.55	16.3	15.39	4.59	1.17	9.2	11.35	0.17	0	333	Poor
DELHI	4/1/15	313.2	607.9	69.16	36.39	33.85	15.2	9.25	41.68	14.36	9.84	472	Severe
HYD	5/1/15	47.03	93	3.7	17.19	24.94	0.3	2.58	30.34	0.41	1.11	120	moderate

Table 2. AQI index value, rating and their health impacts

AQI	Rating	Health impact
0-50	GOOD	It's a great time for outdoor activities!
51-100	MODERATE	You can go outside, but sensitive groups should take precautions.
101-150	SATISFACTORY	Low immune people should limit prolonged outdoor exertion.
151-200	POOR	Everyone should limit prolonged outdoor exertion.
200-300	VERY POOR	Avoid outdoor activities.
More than 300	SEVERE	Stay indoors and avoid all outdoor activities, use medical standard mask

In Figure 2, the two images compare the summary statistics of the raw and pre-processed air quality data. Forward-filling gaps were used to fill in the missing values in the raw data. The overall structure of the dataset appears stable post-preprocessing, with no major changes in statistical properties such as means, standard deviations, and quartile ranges. This suggests that the pre-processing has not dramatically altered the underlying distribution of the data.

Summary statistics for Raw Data:

	PM2.5	PM10	NO	NO2	NOx
count	24933.000000	18391.000000	25949.000000	25946.000000	25346.000000
mean	67.450578	118.127103	17.574730	28.560659	32.309123
std	64.661449	90.605110	22.785846	24.474746	31.646011
min	0.040000	0.010000	0.020000	0.010000	0.000000
25%	28.820000	56.255000	5.630000	11.750000	12.820000
50%	48.570000	95.680000	9.890000	21.690000	23.520000
75%	80.590000	149.745000	19.950000	37.620000	40.127500
max	949.990000	1000.000000	390.680000	362.210000	467.630000

	NH3	CO	SO2	O3	Benzene
count	19203.000000	27472.000000	25677.000000	25509.000000	23908.000000
mean	23.483476	2.248598	14.531977	34.491430	3.280840
std	25.684275	6.962884	18.133775	21.694928	15.811136
min	0.010000	0.000000	0.010000	0.010000	0.000000
25%	8.580000	0.510000	5.670000	18.860000	0.120000
50%	15.850000	0.890000	9.160000	30.840000	1.070000
75%	30.020000	1.450000	15.220000	45.570000	3.080000
max	352.890000	175.810000	193.860000	257.730000	455.030000

	Toluene	Xylene	AQI
count	21490.000000	11422.000000	24850.000000
mean	8.700972	3.070128	166.463581
std	19.969164	6.323247	140.696585
min	0.000000	0.000000	13.000000
25%	0.600000	0.140000	81.000000
50%	2.970000	0.980000	118.000000
75%	9.150000	3.350000	208.000000
max	454.850000	170.370000	2049.000000

Summary statistics for Pre-Processed Data:

	PM2.5	PM10	NO	NO2	NOx
count	29531.000000	29531.000000	29531.00000	29531.000000	29531.000000
mean	67.450578	118.127103	17.57473	28.560659	32.309123
std	59.414476	71.500953	21.35922	22.941051	29.317936
min	0.040000	0.010000	0.02000	0.010000	0.000000
25%	32.150000	79.315000	6.21000	12.980000	14.670000
50%	58.030000	118.127103	11.53000	25.240000	27.550000
75%	72.450000	118.127103	17.57473	34.665000	36.015000
max	949.990000	1000.000000	390.68000	362.210000	467.630000

	NH3	CO	SO2	O3	Benzene
count	29531.000000	29531.000000	29531.000000	29531.000000	29531.000000
mean	23.483476	2.248598	14.531977	34.491430	3.280840
std	20.711370	6.715753	16.909088	20.163443	14.226364
min	0.010000	0.000000	0.010000	0.010000	0.000000
25%	12.040000	0.540000	6.090000	20.740000	0.240000
50%	23.483476	0.950000	10.480000	34.491430	1.840000
75%	23.483476	1.710000	14.531977	42.730000	3.280840
max	352.890000	175.810000	193.860000	257.730000	455.030000

	Toluene	Xylene	AQI
count	29531.000000	29531.000000	29531.000000
mean	8.700972	3.070128	166.463581
std	17.034769	3.932426	129.064348
min	0.000000	0.000000	13.000000
25%	1.280000	2.000000	88.000000
50%	6.930000	3.070128	138.000000
75%	8.700972	3.070128	179.000000
max	454.850000	170.370000	2049.000000

Figure 2. Summary statistics of raw and processed data

3-2-Pre-Processing

Label encoding converts non-numeric variables, like city names, into numerical values, as shown in Equations 1 and 2. The gaps present in the data are pre-processed through the forward-fill ($ffill()$) technique to ensure the dataset has no gaps present, as shown in Equations 3 and 4.

$$lab_{encod} = LabEncod() \quad (1)$$

$$data_{cityencod} = lab_{encod}.fit_{transform}(data_{city}) \quad (2)$$

Let the dataset, d = {x_1, x_2,…,x_n} where some of the values x_i may be missing (NaN).

$$X_i = \begin{cases} x_{i-1}, & X_i = NaN \\ x_i, & X_i \neq Nan \end{cases} \quad (3)$$

$$inp_{data} = inp_{data}.ffill() \quad (4)$$

The MinMaxScaler technique is used to normalize the data, rescaling the feature values between 0 and 1 for the scaling process. This technique enhances LRX performance by ensuring all features are on a similar scale, especially when working with multiple machine learning models. By minimizing the scale difference, as shown in Equations 5 to 7 these models can better learn the underlying patterns.

Let x be the feature value, x_{min} be the minimum value of the feature, and x_{max} be the maximum value of the feature.

$$X_{scaled} = \frac{(X - X_{min})}{X_{max} - X_{min}} \quad (5)$$

$$scaler_{feature} = MinMax_{scaler}() \quad (6)$$

$$Xscaled = scaler_{feature}.fit_{transform}(x) \quad (7)$$

The advantage of using this approach is that it eliminates noisy data from the raw data while maintaining data consistency, ensuring that LRX receives pre-processed and normalized data input, which benefits its prediction accuracy.

Table 3 shows the dataset after being pre-processed to get its features extracted and fed into the LRX model.

Table 3. Dataset after preprocessing

City	Date	PM2.5	PM10	NO	NO_2	NH_3	CO	SO_2	O_3	BENZ ENE	XYL ENE	AQI	CITY_ ENCODE
AMD	1/1/15	73.24	141.6	0.92	18.22	23.48	0.92	27.64	133.36	0	0	166.46	0
BLR	2/1/15	30.65	70.46	3.26	17.33	20.36	0.33	3.54	10.73	0.56	0	91	1
MAA	3/1/15	173.5	48.55	16.3	15.39	4.59	1.17	9.2	11.35	0.17	0	333	2
DELHI	4/1/15	313.2	607.9	69.16	36.39	33.85	15.2	9.25	41.68	14.36	9.84	472	3
HYD	5/1/15	47.03	93	3.7	17.19	24.94	0.3	2.58	30.34	0.41	1.11	120	4

3-3- Model Execution

The execution of LRX occurs in three stages (LSTM→RFR→XGB). Each stage has a specific role in improving the overall accuracy and predictability.

3-3-1- LSTM

LSTM is the first model in the LRX pipeline. As a recurrent neural network (RNN), LSTM excels at capturing long-term dependencies in the data. LSTM is configured with 100 hidden units (hid_{units}), 4 classes (Num_{class}), 1 input (tr_{size}), and trained for 10 epochs (ep_{size}) with a batch size (bh_{size}) of 32. The LSTM captures temporal patterns by analyzing historical AQI data and predicting future values based on past trends.

3-3-2- Random Forest

After LSTM generates the initial predictions, the output is fed into RFR. In general, RFR is an ensemble learning method that uses multiple decision trees to improve prediction accuracy. Each tree is trained on a different subset of data, and their results are averaged to increase generalizability and prevent overfitting. The model uses 100 trees to refine the LSTM predictions. The RFR ensures that tabular data is processed and enhances the interpretability of the results by handling non-linear relations between features.

3-3-3- XGB

The final step involves training XGB on the output of RFR. XGB is a gradient boosting algorithm known for its performance and speed. It optimizes the predictions by applying boosting techniques to reduce errors iteratively. XGB is configured with a learning rate 0.1 and a maximum depth of 6 for each tree. XGB refines the predictions further, accounting for any patterns overlooked by LSTM and RFR.

3-4- Creation of the Streamlit user Interface:

A web-based interface was developed using Streamlit to enhance accessibility and interaction with the hybrid LRX model for AQI prediction. This interface allows users to input relevant data, execute the model, and view the prediction

results clearly and intuitively, as shown in Figure 3. The design focused on providing both functionality and customization, ensuring the application is user-friendly for individuals with varying levels of technical expertise.

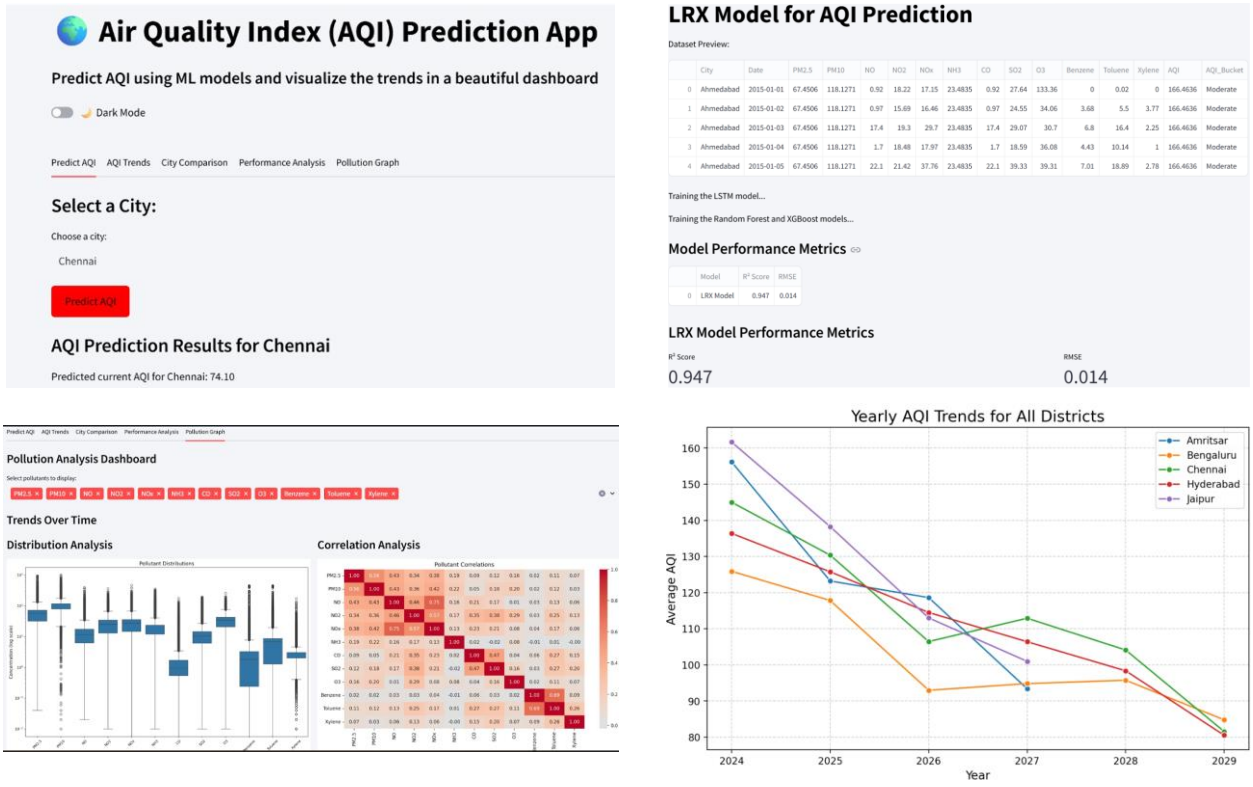


Figure 3. Streamlit user interface tabs for AQI prediction

- **Page Configuration**

The UI was designed to enhance user experience along with tabs to enhance navigation through the many input and output parts, improving the interface's usability.

- **Dark and Light Theming**

The interface allows users to toggle between dark and light modes, catering to user preferences. This toggle was implemented to improve user comfort. At the same time, the dark mode is used to save battery and is easy on the eyes, and while used in the dark, the light mode provides a bright contrast, which makes it easy to see the screen under bright environments or while using outside.

- **User Interactivity and Input Handling**

The interface allows users to choose the data (city) for AQI prediction, which the LRX model processes. The results are presented in the interface, enabling immediate feedback. The interface also ensures that users can modify inputs and view updated predictions without needing to refresh or restart the application.

- **Display of AQI Predictions and Visualizations**

The results of the AQI predictions are presented in graphical formats, providing an understanding of the prediction trends over time or across various locations. Additionally, performance metrics, such as R-squared and RMSE values, offer insights into the model's accuracy and reliability.

The flowchart in Figure 4 provides a clear and concise visual representation of the application's structure and features, making it easy to understand how the system operates. It helps outline the functionalities such as loading the data, building the model, predicting AQI, comparing cities, and analyzing trends, giving the users a roadmap of how to interact with the interface. This helps navigate and ensure users know where to go for specific tasks such as predicting AQI or comparing cities. The architecture diagram shown in Figure 5 provides a high-level overview of an AQI prediction system, illustrating the end-to-end workflow. It shows how data is collected from a database, pre-processed, and then input into LRX model that collectively uses models like LSTM, Random Forest, and XGBoost for training and prediction. The prediction module processes incoming data to forecast the city's AQI, which is then deployed via Streamlit for user interaction through a web interface. Ensuring a seamless integration from data input to real-time AQI prediction and user access. Algorithm LRX mode explains the functionality of the model.

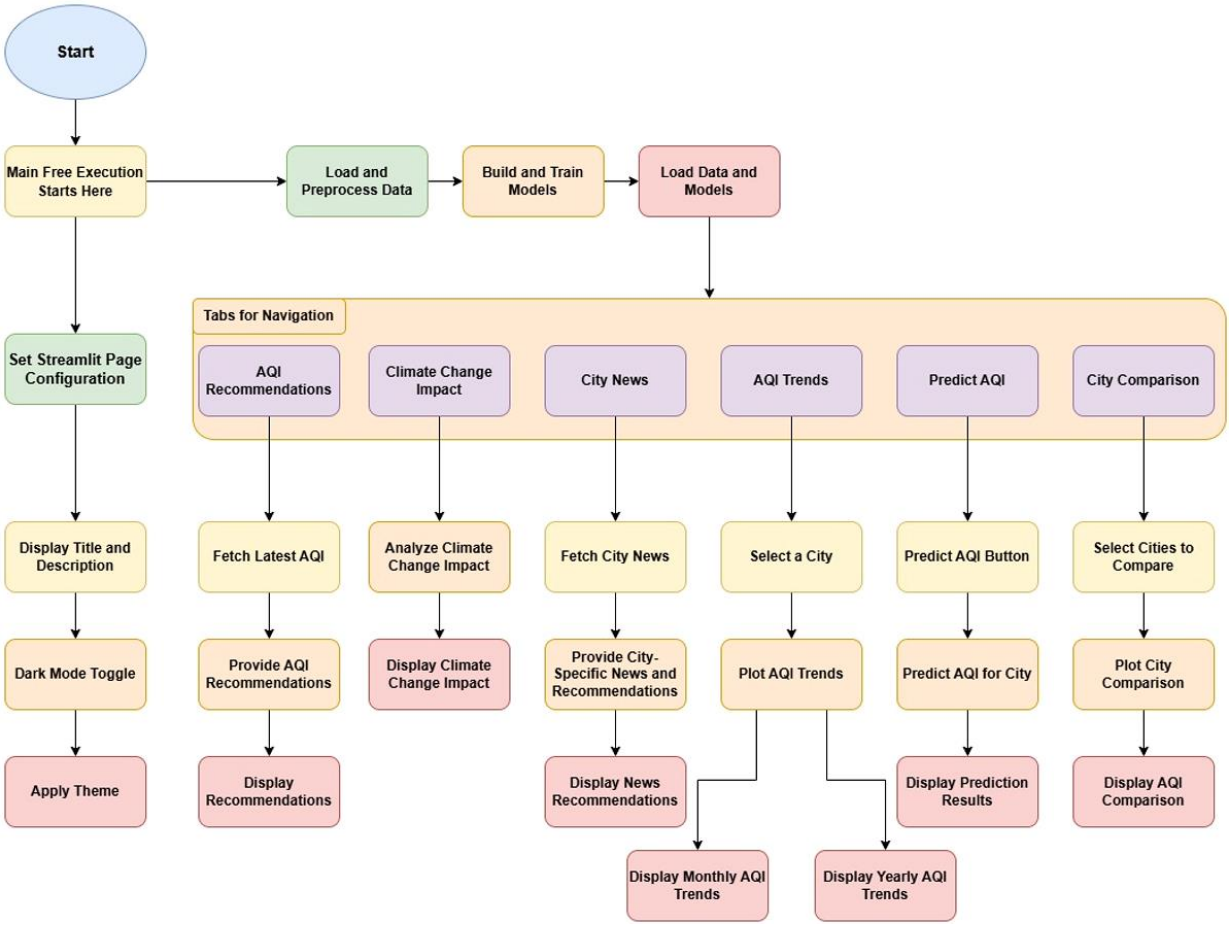


Figure 4. General flow of working project

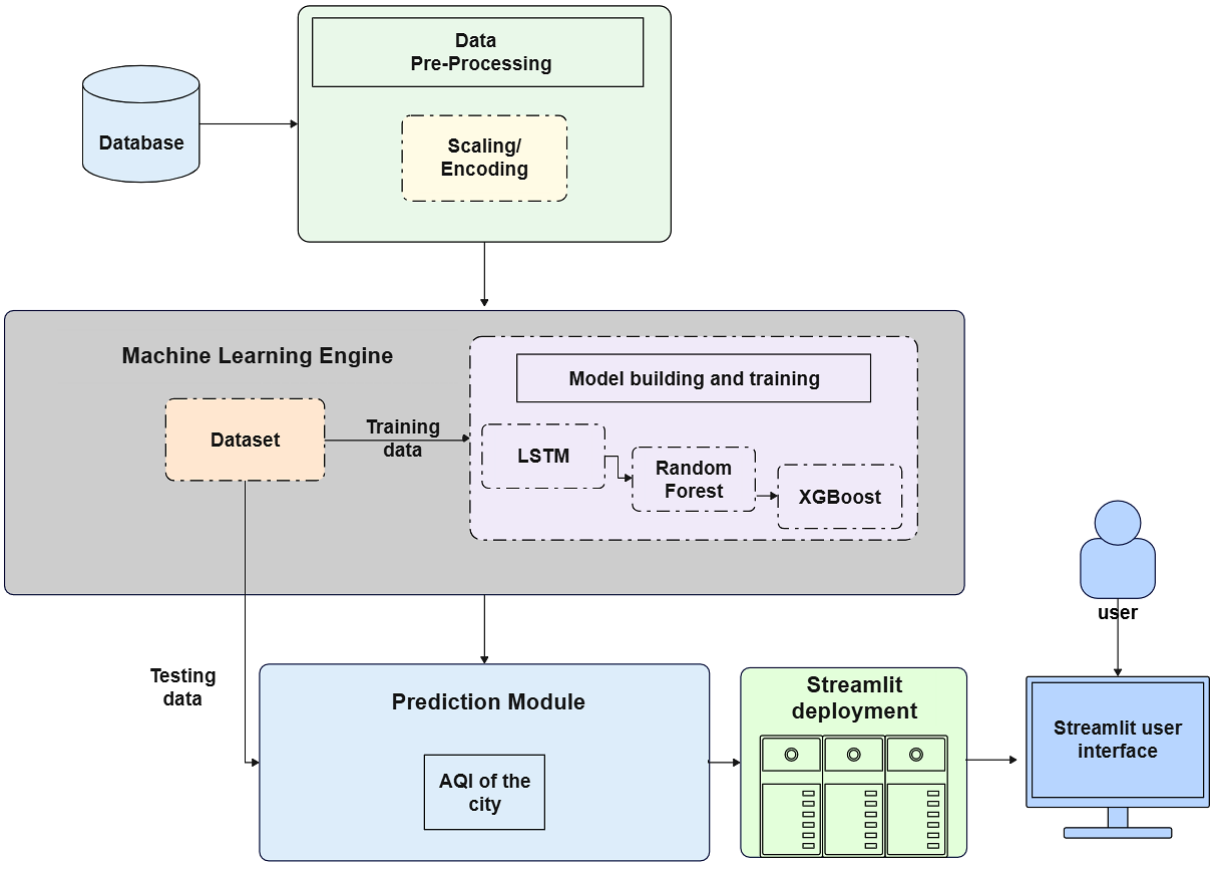


Figure 5. Architecture diagram of hybrid AQI prediction model

3-5-Algorithm 1: LRX Model (Combining LSTM + Random Forest + XGBoost Appraoches)

1. Input: processed data Si_{data}
2. Output: final predictions Pi_{data}
3. Initialise LSTM
4. Initialise Tr_{fea}, tr_{size} = 1, hid_{units} = 100, Num_{class} = 4, ep_{size} = 10, bh_{size} = 32, $lstm_{label}$
5. Train label = 80%, Test label = 20%
6. Initialize the LSTM layers $lstm_{layers}$
7. Initialize LSTM training options $lstm_{options}$
8. Train LSTM
9. Label = unique(label)
10. For xx= 1 : leng(Label):
    Class = find(label == Label(xx))
    Convert $lstm_{label}$ = cat($lstm_{label}$)
    Net = $train_{Net}$(Tr_{fea},$lstm_{label}$,$lstm_{options}$)
    $Train_{cut}$ = length(class) – $train_{cut}$
    $Train_{data}$ = [$train_{data}$; $train_{fea}$; class(1:$Train_{cut}$)]

#Predict LSTM

11. $Predict_{LSTM}$ = classify(Net, $Train_{data}$, bh_{size})

#Initialize RFR with trees

12. N_{trees} = 100

#Train RFR on LSTM output

13. T_{forest} = train(RF,$Predict_{LSTM}$)

#Predict RFR

14. $Predict_{RF}$ = RF.predict(T_{forest})

#Initialize XGB with learning rate lr = 0.1, max depth d = 6

#Train XGB on RFR

15. T_{XGB} = train(XGB,$Predict_{RF}$)

#Predict XGB

16. $Predict_{XGB}$ = XGB.predict(T_{XGB})

#Final prediction

17. P_{data} = $Predict_{XGB}$

The processed data are fed as input in line 1 and the expected final output of the algorithm is mentioned in line 2. The LSTM is initialized in line 3. To make predictions, the LRX algorithm combines the strengths of LSTM, RFR, and XGB and the first, parameters of the LSTM are initialized in line 4, i.e., number of training features (Trfea), number of hidden units (hidunits), number of classes (Numclass), batch size (bhsize), and number of epochs (epsize). The data is then divided into a training set and a testing set with 80% and 20% samples respectively in line 5 to ensure that there is a sufficient amount of data for training as well as validating the model. The two LSTM layers are initialized in lines 6 & 7 and three training options are configured in them.

The rigorous training of the LSTM model is done on the training data and special class labels (Label) from lines 8 to 10, ensure that it is able to learn the temporal dependencies in the data. A LSTM network is created by the trainNet function and a specific number of features (Trfea) are combined with their corresponding class labels (lstmlabel). The output of the LSTM network is used by the classify function in line 11 to assign labels to the test data (Traindata). The RFR model is initialized on line 12 with a hundred decision trees. Line 13 trains the RFR model on the LSTM predictions. This model is an ensemble-based model that uses at least two non-linear relationships in the data to improve its stability and the highly accurate predict function of the Random Forest model generates predictions from the RFR in line 14.

The output of the RFR model is passed to the XGBoost (XGB) model, which is initialized in line 15 with hyperparameters such as a learning rate (lr = 0.1) and a maximum depth (d = 6). The XGB model is trained on the RFR predictions using gradient boosting, which fine-tunes the predictions and reduces errors. Final predictions are obtained in line 16, where the predict function of XGBoost generates the results. In line 17, these final predictions (Pdata) represent the ensemble output of the hybrid LRX model. This step-by-step approach enables the LRX model to harness the temporal analysis capability of LSTM, the non-linear modeling strength of RFR, and the fine-tuning power of XGBoost, providing an effective and scalable solution for predictive tasks.

4- Results and Experiments

4-1-Analysis of Dataset

In Figure 6, the heatmap reveals the correlation between various air pollutants. To examine the interrelation among various air contaminants, a correlation heatmap was generated by Pearson correlation co-efficient. Figure 6 depicted the intensity and paths of linear correlation amongst the eleven contaminants: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, BENZENE and XYLENE. Pollutants such as NO & NO2 with r = 0.46 and NO & NOx with r=0.75 has a strong positive correlation due to the strong substance response of all three component NO, NO2 and NOx. Also, the reaction between NO2 and NOx with r=0.57 indicate a good correlation. When few components such as PM 2.5 & PM10 with r=0.56 and NH3&NO2 with r= 0.35 are correlated, it results in a moderate correlation. Two pollutants SO2 and Benzene with r=0.02 results in low correlation as per the heatmap generated. High positive correlations (closer to 1, in dark red) indicate pollutants that increase or decrease together and strong negative correlations (closer to -1, in dark

blue) indicate pollutants that move in opposite directions. This suggests some pollutants behave similarly under specific environmental conditions, which might hint at their common sources or atmospheric interactions. The pollutants showing high correlations can be further investigated to determine if they arise from similar sources or environmental processes. This helps in designing targeted interventions to reduce specific types of pollution.

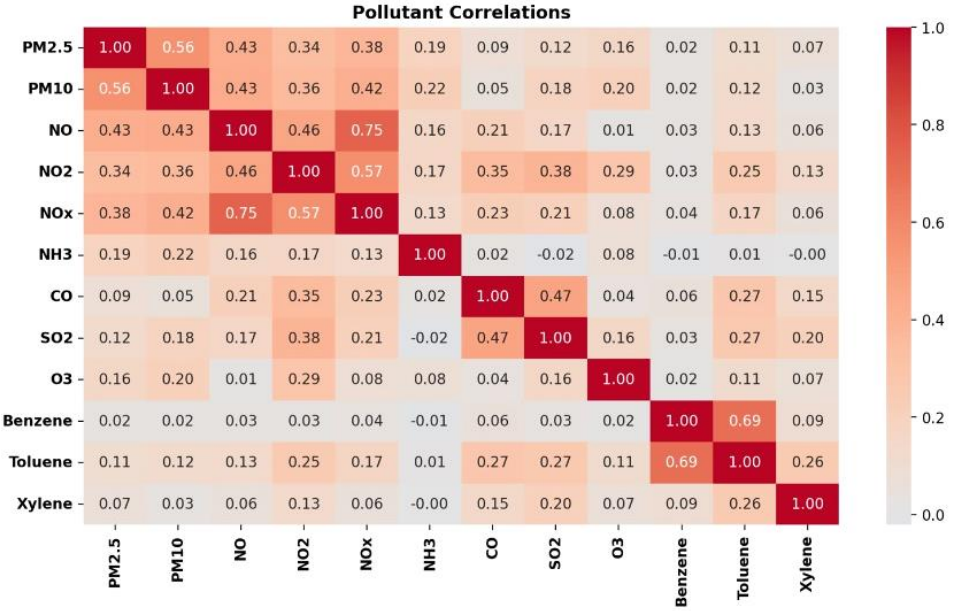


Figure 6. Heatmap of pollutants correlations

The time series plot in Figure 7 shows the fluctuations of different pollutants from January to April 2020. It depicts the time-series image of the intensity readings of various 12 atmosphere contaminants specifically for PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, TOLUENE, BENZENE and XYLENE. The graph is plotted with days' timelines in the x-axis and range of intensity values varying from 0 to 100. Each contaminant is embodied by a highlighted line, lead to substantially dense and intersecting form because of the occurrence and inconsistency of the data. The pattern formed shows that contaminants were captured on a regular basis either daily or weekly, indicating variations in concentrations over time because of various factors such as meteorological conditions, transportation or industrial happenings. Though the graph portrays the dynamics of the contaminants in a certain period no clear cyclical or seasonal pattern exists, but all pollutants exhibit much variability across the time frame. This could suggest varying environmental or anthropogenic influences during this period, leading to irregular peaks and troughs in pollutant concentrations. No distinct seasonal trends are immediately visible, indicating the need for more granular analysis (e.g., looking at specific meteorological data) to identify the drivers behind these fluctuations and reduce them to improve AQI.

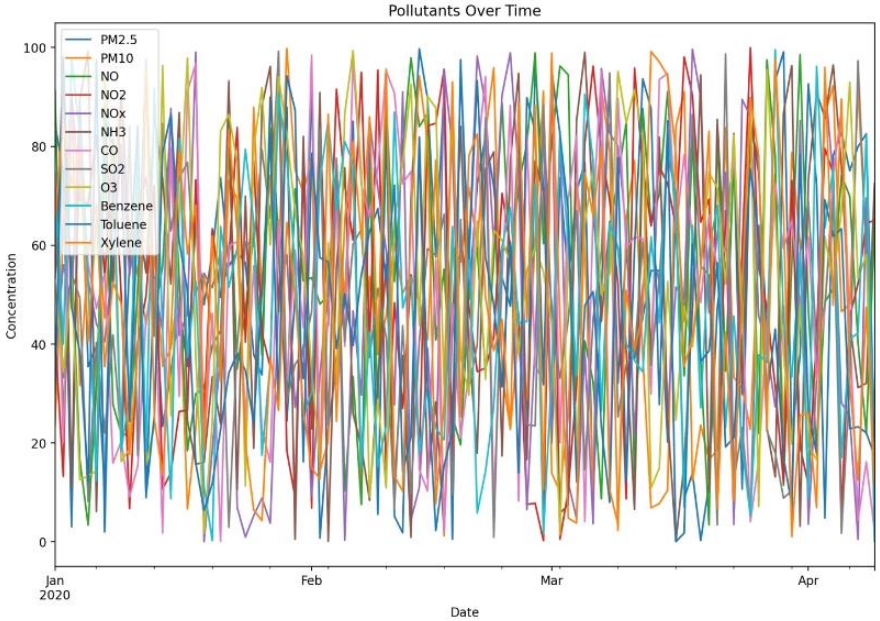


Figure 7. Pollutants graph over time

The plot shown in Figure 8 compares the pollutant distributions before and after pre-processing using a log scale for concentration. Using logarithmic range on the x-axis and density on the y-axis, the graph has compared original dataset and the pre-processed dataset. The solid lines correspond to the original dataset, while the dashed lines represent the pre-processed dataset. For most pollutants (e.g., PM2.5, PM10, NOx, Benzene, CO, SO2, O3), the solid and dashed lines are quite close, indicating that pre-processing has preserved the core distributions of the data. There is a slight difference at lower concentrations in pollutants like PM2.5, NOx, and Xylene. The distribution of the original dataset regularly appears crooked or uneven spreading with hard points or extended lines representing the incidence of outliers. Pre-processed dataset seems to have smoothed or broadened some peaks compared to the original dataset. For NH3 and CO, the original dataset shows sharp spikes that are less pronounced in pre-processed dataset. Pre-processing may have reduced some of these spikes, possibly by handling outliers or filling missing data, which makes the distribution of pre-processed dataset more uniform.

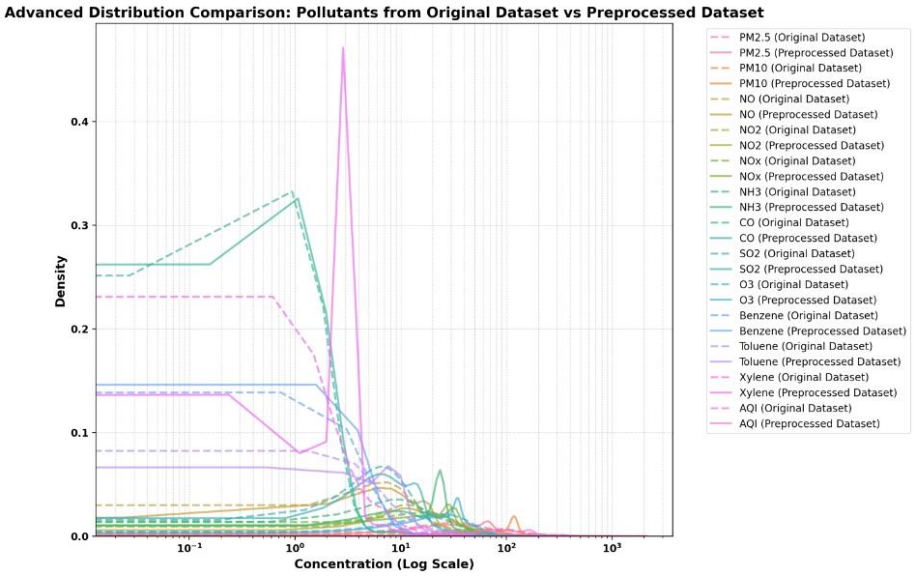


Figure 8. Pollution concentration distribution variability

The density plot in Figure 9 compares the Air Quality Index (AQI) distribution for raw and pre-processed data. The x-axis embodies the AQI values and the y-axis indicates the pollution density distribution variability, which replicates how regularly values hit in the data. Both raw data and pre-processed data are skewed on the right, yet the pre-processed data is less spread out and appears smoother, so it has fewer noisy values. The pre-processed data shows a smoother distribution with fewer outliers, suggesting that the pre-processing removed noise or extreme values. Both distributions indicate a high concentration of AQI values, implying that most of the recorded air quality is within the "good" or "moderate" range. There are some instances of poor air quality (higher AQI), but they appear less frequent after pre-processing. The pre-processed AQI data reduces noise and gives a clearer picture of air quality trends, which is important for model prediction.

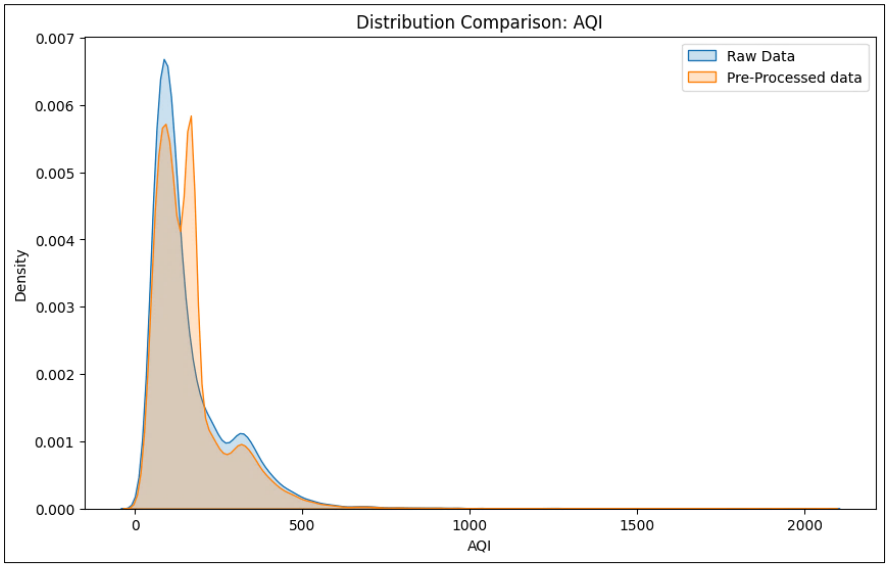


Figure 9. Distribution of AQI

In summary, pre-processing has preserved the fundamental characteristics of the dataset while making slight adjustments that improve data consistency and reduce the impact of outliers or missing data, making the dataset more suitable for analysis and model training.

4-2-Performance Metrics

The metrics used in our research work are R^2 and RMSE.

R^2 measures how well the regression model aligns with the observed data. A higher R^2 value indicates a strong model fit, as shown by Equation 8.

$$R^2 = \frac{S_{reg}}{S_{tl}} \tag{8}$$

The sum of squares due to regression, represented as S_{reg} (explained sum of squares), indicates how well the regression model fits the data. The total sum of squares (S_{tl}) reflects the overall variation in the observed data used in the regression model. S_{reg} measures the model's explanatory power, while S_{tl} captures the total variability in the data.

RMSE reflects how closely the data clusters around the line of best fit shown by Equation 9.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Y_i - Y_i')^2}{n}} \tag{9}$$

where Y_i is the observed value, Y_i' is the corresponding predicted value, and n is the number of observations used.

4-2- Models Performance

The hybrid LRX model implemented to predict AQI values has achieved an impressive performance, with an R^2 of 0.948 and a Root Mean Squared Error (RMSE) of 0.014 as shown in Figure 10. The high coefficient of determination R^2 of 0.948 indicates that the hybrid model explains 94.8% of the variance in the AQI data, which suggests that the model has successfully captured the underlying relationships in the data. The low RMSE of 0.014 shows that the error in predicting AQI values is minimal, demonstrating the LRX's precision in AQI forecasting, bolstering the viability of hybrid approaches in addressing multi-faceted environmental problems, like AQI forecasting.

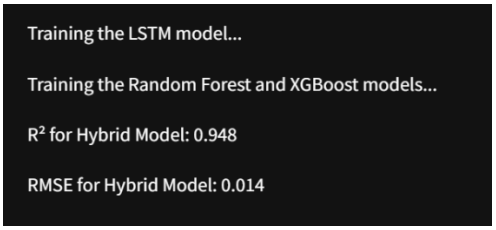


Figure 10. Performance metrics

This figure also highlights the efficiency of the training process for the hybrid model, as it successfully integrates the temporal dependencies captured by LSTM with the robustness of Random Forest and the fine-tuning capabilities of XGBoost. By combining these complementary strengths, the hybrid LRX model effectively overcomes the limitations of standalone approaches. Figure 10 is a testament to the practicality of leveraging hybrid methodologies in addressing multi-dimensional environmental challenges, such as air quality management and prediction. The results validate the LRX model as a promising tool for real-world applications where accurate AQI forecasting is critical for policymaking and public health.

The hybrid approach takes advantage of the strengths of all three models (LSTM, RFR, and XGBoost). The proposed LRX model is compared with three individual models: LSTM, RF, and XGBoost. When the data is fed into all four models, the results achieved are shown in Table 4. As per the results achieved, the LSTM scored the least R2. The three remaining models, RF, XGBoost, and LRX, achieved almost similar R^2 values; however, the proposed LRX model is likely to score the highest R^2 of 0.948. Regarding the RMSE value, it should be lesser for the model to perform better. As per Table 4, LSTM has the highest RMSE value, and the proposed model LRX has the least RMSE value compared to the other models. By comparing the performances of R2 and RMSE, the proposed model LRX is efficient compared to the other individual models. The Streamlit interface is structured to facilitate a seamless experience when comparing model performances. Dedicated tabs highlight detailed metrics for each model, including a side-by-side visualization of R^2 and RMSE scores in interactive charts. A custom feature enables users to upload their datasets and visualize predicted AQI trends across selected time intervals, emphasizing the strengths of the LRX model. Unlike traditional static result presentations, this interactive approach empowers users to dynamically validate the model's efficiency. Beyond its impressive R^2 of 0.948 and low RMSE of 0.014, the hybrid LRX model exhibits several distinct advantages that enhance its performance and applicability in AQI forecasting.

Table 4. Model performances

Model	R^2	RMSE
LSTM	0.873	0.021
Random Forest	0.946	0.016
XGBoost	0.943	0.014
LRX (proposed)	0.948	0.014

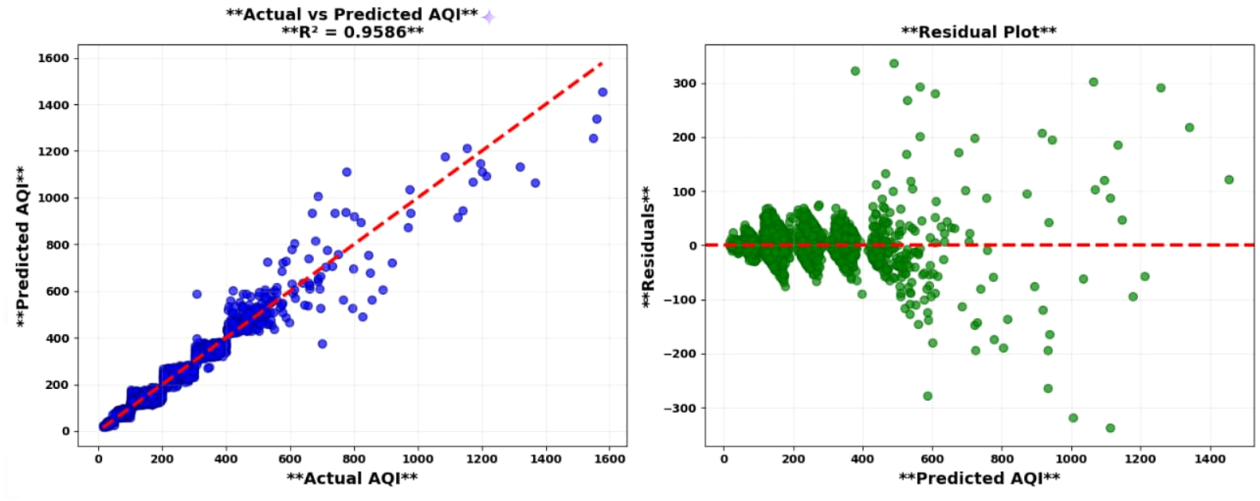
The LRX model's accuracy and real-time interactivity enable timely and data-driven decision-making. Environmental managers or policymakers can rely on these predictions to take proactive measures to improve air quality, ensuring better public health outcomes by preventing harmful AQI levels before they occur. Unlike standalone models that might overfit or underperform when exposed to new data, the hybrid approach improves generalization. The LRX model better adapts to different environmental conditions and datasets by combining multiple learning strategies, making it more reliable across diverse scenarios. The significant reduction in RMSE indicates that the LRX model's predictions are far more precise than individual models. This ensures that forecasted AQI values closely align with real-world data, offering high accuracy and reducing the margin of error in important environmental predictions, such as air quality management.

Integrating multiple models within the hybrid structure makes the LRX system highly scalable. As AQI datasets grow or new data sources are incorporated (e.g., additional environmental factors), the model can be retrained or updated without losing performance. Its flexibility in handling varied data types (e.g., time-series, categorical) ensures applicability in a wide range of forecasting tasks. The user-friendly Streamlit interface enhances accessibility by allowing users to upload their datasets, visualize predictions, and explore performance metrics interactively. The real-time visualization allows users to test and validate the model's efficiency dynamically, providing an intuitive and informative user experience that traditional methods of static result presentation lack. The interactivity enables stakeholders to make better decisions based on the model's outputs, promoting trust in the model's predictive capabilities.

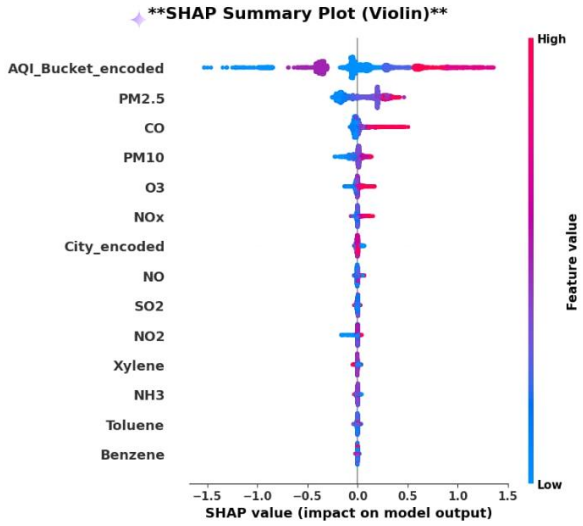
4-3-SHAP -Built Interpretability and the Model Valuation

Even though the suggested hybrid model is intricate, its results remain easy to interpret because of SHAP (Shapley additive explanations). To explain an individual prediction, SHAP created a waterfall plot, and to show how all features contribute, SHAP used beeswarm, summary, and feature importance plots. Figure 11 makes it very clear which characteristics influence the predicted AQI. The level of clear information from this data makes it easy for decision-makers to understand the importance of different factors impacting pollution in policymaking. Figure 11 comprises 4 graphs depicting (a) Actual vs. Predicted AQI, (b) SHAP summary plot, (c) Feature Importance, and (d) SHAP waterfall model. Figure 11-a shows that the predicted AQI values are very close to the real ones, reflected in $R^2 = 0.9588$, and shows the gaps between the values that are expected and the values that are measured. Also, the residues are almost near to zero, indicating that the proposed model is not inclined, and thus the proposed model can be applied to the new dataset also. Figure 11-b shows how each feature affects the way the model makes predictions, and the values in AQI_Bucket_encoded, PM2.5, and CO tend to have the most impact on what the prediction model says about AQI. Figure 11-c depicts the bar chart that emphasizes the significance of the feature centered on the SHAP importance scale. Figure 11-d depicts the SHAP waterfall plot that demonstrates what different variables played in the prediction that the AQI reached a value of 1454.97. The baseline (average log-AQI) is used to predict, and most of the increase is driven by Feature 13 and Feature 6, which impact it the most positively. Certain features contribute less, but nonetheless}, a small number have very weak negative impacts. It clearly shows what leads to high AQI, which helps make focused decisions on environmental protection activities. Collectively, Figure 11 indicates that the proposed model has both a high accuracy (R^2) and is understandable through SHAP values. The top three pollutants, PM2.5, CO, and PM10, combined with the location and the AQI categories, are shown to be the most significant factors in determining AQI—giving useful advice for making data-based AQI policies.

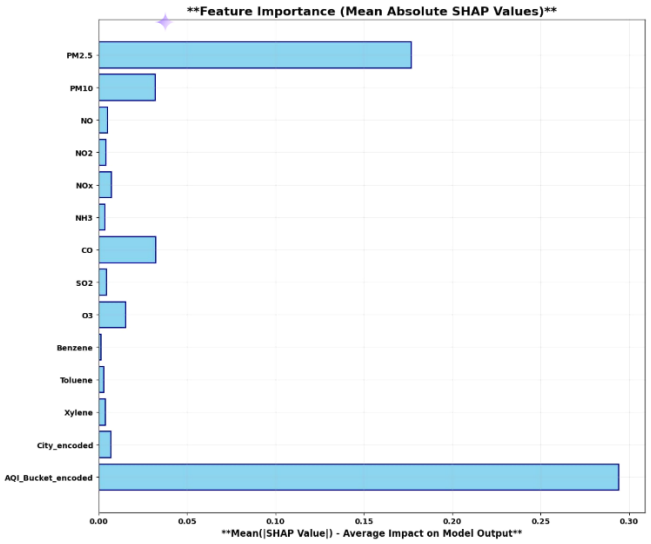
Figure 12 shows a comparison of the performance metrics of the individual models versus the LRX model. The graph clearly shows the LRX model achieving the highest R^2 and lowest RMSE, outperforming all standalone models tested. The orange line in Figure 12 tracks the coefficient of determination (R^2) for each model. It rises steadily from LSTM (0.873) to random forest (0.946), then remains constant for XGBoost (0.943) before reaching its peak with the hybrid LRX model (0.948). The upward trend demonstrates how the hybrid model improves upon the variance explanation capabilities of standalone models. Likewise, the blue line represents the RMSE for each model. The values start higher for LSTM (0.021) and drop progressively for Random Forest (0.016) and XGBoost (0.014), with the hybrid model maintaining the same minimal RMSE as XGBoost (0.014). The declining trend reflects the reduction in prediction errors as we move to more advanced or hybrid techniques. The lines illustrate the complementary strengths of the hybrid approach, thus demonstrating its robustness and efficiency.



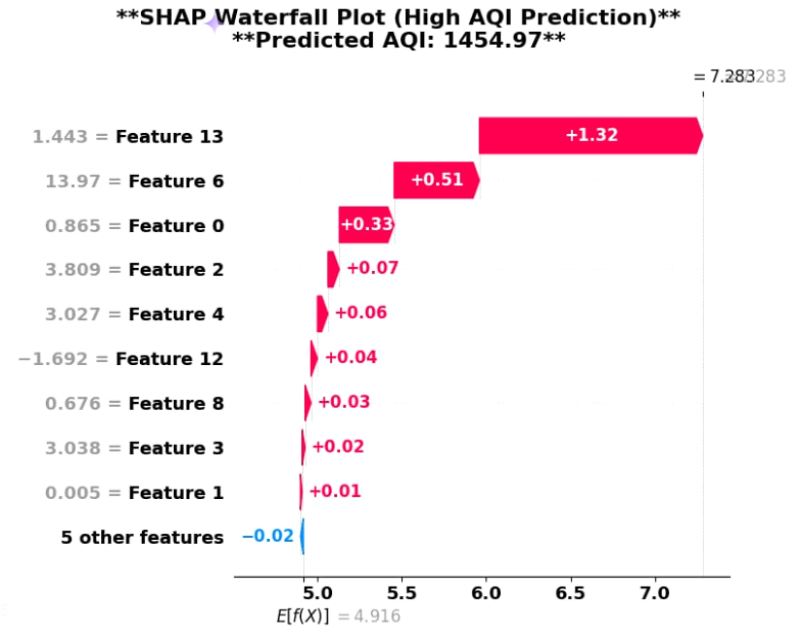
(a) Actual Vs Predicted



(b) SHAP summary plot



(c) Feature Importance



(d) SHAP Waterfall model

Figure 11. Model Evaluation using SHAP

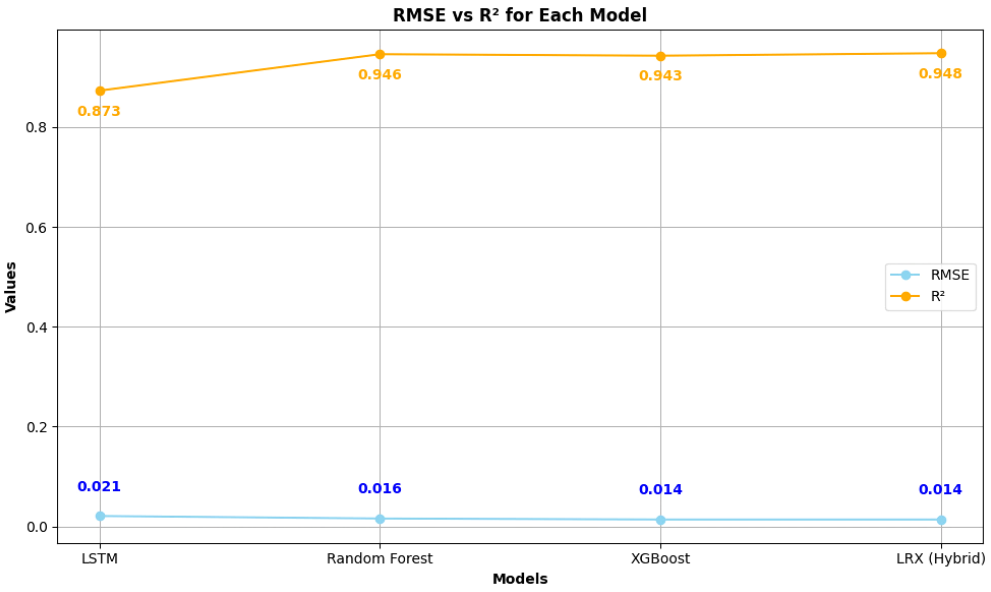


Figure 12. Performance comparison graph

The integration of LSTM's strength in temporal sequence learning, Random Forest's robustness in handling non-linear relationships, and XGBoost's ability to fine-tune and enhance prediction have enabled the LRX model to leverage the best of all three algorithms. This synergy highlights the importance of adopting hybrid approaches that work in sync to achieve Great results in predictive modelling

The analysis of future AQI trends, depicted in Figure 13, provides valuable insights into air quality patterns across six major cities, Bengaluru, Chennai, Chandigarh, Delhi, Lucknow, and Mumbai, over the forecasted years 2024 to 2029. The graph highlights variations in AQI trends among these cities, underscoring the importance of localized interventions in air quality management. Delhi and Lucknow emerge as significant concerns, with high AQI values throughout the observation period. Despite this, an encouraging downward trend by 2029 indicates potential improvements, likely due to the implementation of air quality control measures. Bengaluru and Chennai consistently exhibit better air quality, as shown by lower AQI values and a steady decline over the years. This trend reflects the effectiveness of existing measures and a potentially cleaner environment in these regions. The general declining trend of AQI in most cities toward 2029 may suggest implementing effective air quality control measures. However, the persistent challenges in cities like Delhi necessitate a more focused and aggressive approach, including stricter regulations, enhanced monitoring, and innovative solutions to address local and regional pollution sources. Such efforts are critical to achieving sustainable improvements in air quality for heavily polluted urban centres.

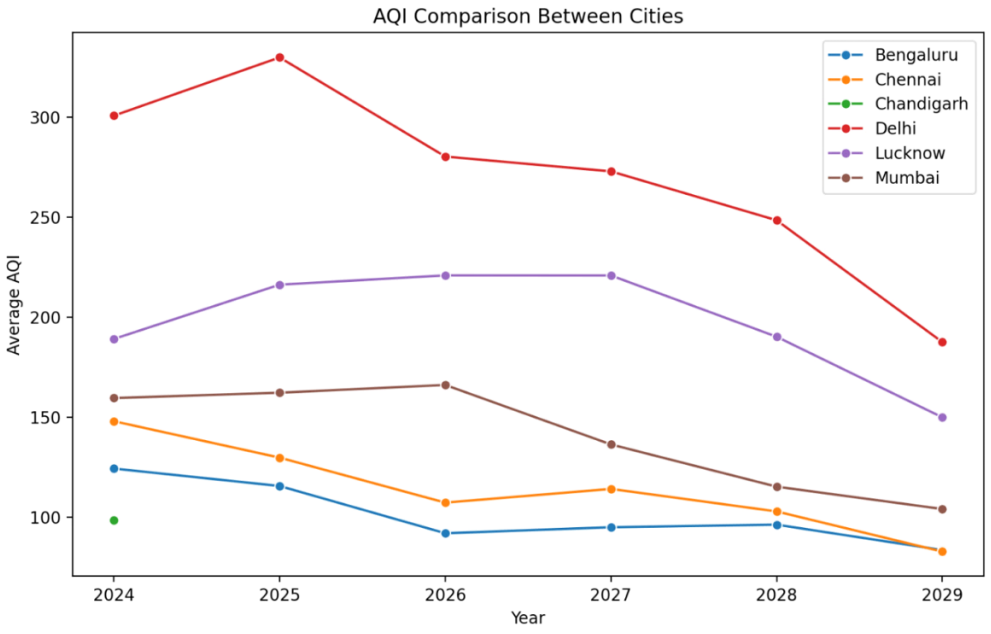


Figure 13. Future AQI predictions

The key findings from this study are as follows:

Hybrid Modelling: The LRX model proposed in this study can combine the strengths of LSTM, RFR, and XGBoost, allowing it to perform better in terms of prediction accuracy and prediction error than any of these models separately. This is good because it allows the LRX model to overcome limitations related to temporal constraints and feature selection, which enhances the accuracy of the AQI forecasting models.

Performance Superiority: Performance metrics such as R^2 and RMSE showcase the superiority of the LRX model over each of the models on its own. In this case, even though LSTM achieves an R^2 of 0.873 and RMSE stands at 0.021, the hybrid model does much better on both statistics, significantly illustrating its usability on difficult tasks such as forecasting AQI values.

Scalability and Practical Application: The LRX model's low error rates and high accuracy suggest its practical applicability in real-world air quality monitoring systems. This research demonstrates the potential of hybrid machine learning approaches to provide actionable insights into urban air quality management. The results presented in this paper contribute to the growing field of air quality modelling by demonstrating the effectiveness of hybrid techniques.

5- Conclusion

This paper proposes a hybrid AQI prediction model based on the LRX approach, which combines LSTM, Random Forest Regressor (RFR), and XGBoost. Compared to standalone models, the LRX model has shown great promise in predicting AQI values. The hybrid approach exploits the strengths of each component; as a result, the LRX model achieves an R^2 of 0.948 and an RMSE of 0.014, exceeding its standalone counterparts. Furthermore, the comparative analysis of standalone models LSTM, RFR, and XGBoost reinforces the hybrid model's superiority, as evidenced by its ability to achieve higher accuracy and lower error rates. This highlights the LRX model's robustness in capturing complex temporal patterns and non-linear relationships, demonstrating its adaptability to real-world AQI prediction scenarios. The hybrid approach mitigates the limitations of individual techniques, maximizing predictive accuracy by combining the temporal pattern recognition of LSTM, the stability of RFR, and the fine-tuning capability of XGBoost. This suggests that the LRX model is effective in AQI forecasting and could be further enhanced by incorporating additional features such as meteorological data. Future research could focus on refining the model's scalability for real-time prediction, making it more applicable for continuous monitoring and actionable air quality management strategies. In addition to the predictive model, a Streamlit-based interface was implemented to provide a user-friendly platform for exploring AQI trends and making predictions. The interface enables real-time interaction, visualization of data trends, and accessibility for non-technical users, ensuring the model's applicability in practical urban planning and public health scenarios. This integration bridges the gap between advanced predictive techniques and real-world usability. Despite promising results, limitations must be acknowledged, and future research is proposed to enhance the LRX model further. A drawback of the LSTM model is that it relies on a large amount of historical data for training. The LRX model is computationally demanding since it integrates three different machine learning approaches. Further research should concentrate on enhancing LRX for operating under limited resources or finding lighter alternatives that can attain performance and computational efficiency close to LRX.

6- Declarations

6-1- Author Contributions

Conceptualization, J.J. and S.C.H.; methodology, A.V.; software, S.P.; validation, S.K.T., J.J., and A.V.; formal analysis, J.J.; investigation, S.C.H.; resources, A.V.; data curation, S.K.T.; writing—original draft preparation, A.V.; writing—review and editing, J.J.; visualization, S.P.; supervision, S.C.H.; project administration, N.P. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: [*https://www.kaggle.com/datasets/hirenvora/city-daycsv*].

6-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Nguyen, A. T., Pham, D. H., Oo, B. L., Ahn, Y., & Lim, B. T. H. (2024). Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. Journal of Big Data, 11(1), 71. doi:10.1186/s40537-024-00926-5.

[2] Palaniappan, S., Logeswaran, R., Khanam, S., & Zhang, Y. (2025). Machine Learning Model for Predicting Net Environmental Effects. Journal of Informatics and Web Engineering, 4(1), 243–253. doi:10.33093/jiwe.2025.4.1.18.

[3] Zayed, R., & Abbod, M. (2024). Air Quality Index Prediction Using DNN-Markov Modeling. Applied Artificial Intelligence, 38(1). doi:10.1080/08839514.2024.2371540.

[4] Palaniappan, S., Logeswaran, R., Velayutham, A., & Bui, N. D. (2025). Predicting Short-Range Weather in Tropical Regions Using Random Forest Classifier. Journal of Informatics and Web Engineering, 4(1), 18–28. doi:10.33093/jiwe.2025.4.1.2.

[5] Subramanian, A., Palanichamy, N., Ng, K.-W., & Aneja, S. (2025). Climate Change Analysis in Malaysia Using Machine Learning. Journal of Informatics and Web Engineering, 4(1), 307–319. doi:10.33093/jiwe.2025.4.1.22.

[6] Sarkar, N., Gupta, R., Keserwani, P. K., & Govil, M. C. (2022). Air Quality Index prediction using an effective hybrid deep learning model. Environmental Pollution, 315, 120404. doi:10.1016/j.envpol.2022.120404.

[7] Alzu'bi, F., Al-Rawabdeh, A., & Almagbile, A. (2024). Predicting air quality using random forest: A case study in Amman-Zarqa. Egyptian Journal of Remote Sensing and Space Science, 27(3), 604–613. doi:10.1016/j.ejrs.2024.07.004.

[8] Mishra, A., & Gupta, Y. (2024). Comparative analysis of Air Quality Index prediction using deep learning algorithms. Spatial Information Research, 32(1), 63–72. doi:10.1007/s41324-023-00541-1.

[9] Gaikar, D., Patel, U., Vispute, O., Singh, S., Sanghvi, T., & Professor, A. (2023). Prediction of Air Quality Index using Random Forest Algorithm. International Research Journal of Engineering and Technology, 10(4), 1248–1252.

[10] Janarthanan, R., Partheeban, P., Somasundaram, K., & Navin Elamparithi, P. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society, 67, 102720. doi:10.1016/j.scs.2021.102720.

[11] Jiao, Y., Wang, Z., & Zhang, Y. (2019). Prediction of air quality index based on LSTM. Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019, 17–20. doi:10.1109/ITAIC.2019.8785602.

[12] Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences (Switzerland), 9(19), 4069. doi:10.3390/app9194069.

[13] Wang, J., Li, X., Jin, L., Li, J., Sun, Q., & Wang, H. (2022). An air quality index prediction model based on CNN-ILSTM. Scientific Reports, 12(1), 8373. doi:10.1038/s41598-022-12355-6.

[14] Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. Scientific Reports, 14(1), 6795. doi:10.1038/s41598-024-54807-1.

[15] Duan, J., Gong, Y., Luo, J., & Zhao, Z. (2023). Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer. Scientific Reports, 13(1), 12127. doi:10.1038/s41598-023-36620-4.

[16] Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. Journal of Environmental and Public Health, 2023, 1–26. doi:10.1155/2023/4916267.

[17] Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning. Alexandria Engineering Journal, 93, 128–141. doi:10.1016/j.aej.2024.03.031.

[18] Kumar Singh, R., Raghav, S., Maini, T., Kumar Singh, M., & Arquam, Md. (2022). Air Quality Prediction using Machine Learning. SSRN Electronic Journal. doi:10.2139/ssrn.4157651.

[19] Anggraini, T. S., Irie, H., Sakti, A. D., & Wikantika, K. (2024). Machine learning-based global air quality index development using remote sensing and ground-based stations. Environmental Advances, 15, 100456. doi:10.1016/j.envadv.2023.100456.

[20] Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam. Chemosphere, 338, 139518. doi:10.1016/j.chemosphere.2023.139518.

[21] Farooq, O., Shahid, M., Arshad, S., Altaf, A., Iqbal, F., Vera, Y. A. M., Flores, M. A. L., & Ashraf, I. (2024). An enhanced approach for predicting air pollution using quantum support vector machine. Scientific Reports, 14(1), 19521. doi:10.1038/s41598-024-69663-2.

[22] Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. Applied Sciences (Switzerland), 10(24), 1–17. doi:10.3390/app10249151.

[23] Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering, 7, 100093. doi:10.1016/j.dche.2023.100093.

[24] Wang, S.-J., Huang, B.-J., & Hu, M.-H. (2023). A Deep Learning-based Air Quality Index Prediction Model Using LSTM and Reference Stations: A Real Application in Taiwan. 33rd International Telecommunication Networks and Applications Conference, 204–209. doi:10.1109/itnac59571.2023.10368496.

[25] Sachdeva, S., Kaur, R., Kimmi, Singh, H., Aggarwal, K., & Kharb, S. (2024). Meteorological AQI and pollutants concentration-based AQI predictor. International Journal of Environmental Science and Technology, 21(5), 4979–4996. doi:10.1007/s13762-023-05307-8.

[26] Chang, W., Chen, X., He, Z., & Zhou, S. (2023). A Prediction Hybrid Framework for Air Quality Integrated with W-BiLSTM(PSO)-GRU and XGBoost Methods. Sustainability (Switzerland), 15(22), 16064. doi:10.3390/su152216064.

[27] Zhongjie, Y., Shengwei, W., & Ze, W. (2022). Air quality prediction method based on the CS-LSTM. 5th International Conference on Data Science and Information Technology (DSIT), 1–5. doi:10.1109/dsit55514.2022.9943922.

[28] Li, C., Li, Y., & Bao, Y. (2021). Research on Air Quality Prediction Based on Machine Learning. 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), 77–81. doi:10.1109/ichci54629.2021.00022.

[29] Gupta, V., Kapadia, S., & Bhadane, C. (2023). Time Series Analysis and Forecasting of Air Quality in India. 5th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2023, 1–5. doi:10.1109/ICECCT56650.2023.10179673.

[30] Zhou, Z. (2023). Air Quality Prediction Based on Improved LSTM Model. 4th International Conference on Computer Engineering and Application (ICCEA), 392–395. doi:10.1109/iccea58433.2023.10135512.