AQI Prediction using Machine Learning and Deep Learning Techniques

Abstract

A major environmental problem is air pollution, which endangers public health significantly. Forecasting Air Quality Index (AQI) accurately calls for The current project proposes an AQI Machine Learning-based model using a mix of Machine Learning (ML) and Deep Learning (DL) techniques, using AQI data of Amaravati. using Amaravati AQI data, (DL) techniques combining Machine Learning (ML) and Deep Learning Using four ML models—Random Forest, XGBoost, LightGBM, and Bagging Regressor—we compared them according to various performance criteria including R^2 score, MSE, RMSE, and MAE. SHAP-based Explainable AI (XAI) feature selection techniques were used to each model to select the most predictive features, so improving interpretability. Where the top performing model was chosen from evaluation criteria, each model was then re-trained on their individual top features. Input to build a unidirectional Long Short-Term Memory (LSTM) model to find temporal patterns in AQI data was the top features of the best performing ML model. LIME, Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) were used to further clarify the LSTM model forecasts, therefore generating global and individual level explainability. Using both visual Numerical outputs, LSTM was finally used to predict future AQI values. By combining ML model power with DL model potential and explanation tools to enable more Environmental Decision-Making, the project proposed here provides a correct, clear, and scalable answer to AQI prediction. Additionally, real-time weather and pollution data were integrated using the OpenWeatherMap API to enhance the model's practical application and provide up-to-date AQI predictions through a Streamlit-based interface.

Keywords: Air Quality Index, Machine Learning, Deep Learning, LSTM, Explainable AI, SHAP, PDP, ICE, LIME, Forecasting

Abbreviations: AQI, Air Quality Index; ML, Machine Learning; DL, Deep Learning; SHAP, Shapley Additive Explanations; LSTM, Long Short-Term Memory; PDP, Partial Dependence Plot; ICE, Individual Conditional Expectation; LIME, Local Interpretable Model-agnostic Explanations; R^2, Coefficient of Determination; MSE, Mean Squared Error; RMSE, Root Mean Squared Error; MAE, Mean Absolute Error.

1. Introduction

Air pollution has become one of today's most serious public and environmental concerns, most prominently seen in rapidly developing nations like India, where urbanization, industrialization, and population growth has seen a dramatic increase in air pollutants [1][3][4]. The key culprits behind this are emissions from vehicles, industries, fossil fuel combustion, and unsustainable farming activities, all of which emit a modern suite of air pollutants such as $PM_{2.5}$, PM_{10}, NO_x, SO_2, CO, O_3, and VOCs into the atmosphere [2][5][6]. These pollutants, when mixed with meteorological variables such as temperature, humidity, atmospheric wind speed, and solar radiation, form complex interactions that play a significant role in determining ambient air quality, thereby impacting human health as well as ecosystem stability [7][8][9].

The Air Quality Index (AQI) has become a critical means of expressing raw pollutant levels in a form that policymakers, as well as the populace at large, can understand, thus conveying potential

hazard to human health [6][10][11]. The higher AQI values are not abstractions; rather, they represent days when air quality poses real risks to human well-being, especially in populations already compromised, such as children, elderly individuals, and those already ill [13]. Poor air also accelerates environmental ills, such as acid rain, and smog, and impacts foundational climate change mechanisms [12][13][14]. Demonstrating relevance, recent figures put this problem in context: India consistently appears among the most polluted countries in the world, earning a top spot on multiple occasions, its $PM_{2.5}$ levels in many urban areas in many cases ten-fold, if not more, over World Health Organization standards [1][5]. $PM_{2.5}$ levels are reported to be surprisingly seven times higher than set standards in more than half of Indian cities, vehicular emissions, along with farm burning during peak seasons, cited as a top contributor [1].

Despite the progress in sensor technologies that enable real-time monitoring of air pollutants, the accurate prediction of the Air Quality Index (AQI) still faces immense challenges. The intrinsic non-linearity and time-varying nature of environmental data tends to make conventional statistical methods inadequate for explaining the complex interdependencies and dynamic patterns inherent in air quality data [1]. Given this challenge, researchers have increasingly used machine learning (ML) and deep learning (DL) methods, owing to their excellence in extracting hidden patterns from large and complicated datasets and their ability to provide high accuracy in real-time AQI prediction [1][6]. Ensemble-based ML methods, i.e., Random Forest, XGBoost, LightGBM, and Bagging Regressor, have been shown to be highly effective for AQI prediction because they can handle high-dimensional, heterogeneous data while being resilient to changes in pollution levels. In addition, deep learning time-series methods, especially Long Short-Term Memory (LSTM) networks, have greatly boosted prediction performance by effectively modeling temporal dependencies and long-range correlations inherent in AQI data [10].

However, as predictive models become more complex, it becomes essential to ensure their interpretability and trustworthiness, especially when model results influence public health and policy decision-making [12]. Explainable Artificial Intelligence (XAI) tools such as SHAP, PDP, ICE, and LIME have become essential tools for explaining model decisions, identifying key factors influencing the Air Quality Index (AQI), and promoting stakeholder trust in machine learning forecast systems [11]. Interpretability techniques enable transparent decision-making, allowing policymakers, researchers, and the general public to understand, trust, and act upon model forecasts [13].

In this study, we propose a hybrid model that integrates the strengths of an ensemble machine learning approach, comprising Random Forest, XGBoost, LightGBM, and Bagging Regressor, with a deep learning time-series model, Long Short-Term Memory (LSTM). We include techniques from explainable artificial intelligence (XAI) [6][10] to improve the model's interpretability. Tailored for urban areas in India, where the effects of air pollution are especially harmful [1][4], the main goal of this model is to create an efficient, scalable, and understandable Air Quality Index (AQI) forecasting system. This model aims to guide decision makers, support public health projects, help to build more resilient urban areas, and encourage environmental protection in reaction to rising air pollution by means of correct and actionable air quality forecasts [8][9].

2. Related Works

The use of Machine Learning (ML) and Deep Learning (DL) algorithms for forecasting the Air Quality Index (AQI) has developed significantly over the last ten years. Originally, researchers looked for trends in air pollution using conventional statistical techniques such Linear Regression

(LR) and Multiple Linear Regression (MLR). These conventional methods, however, did not understand the complex, non-linear interactions between meteorological factors and pollution levels [2][6][15]. For example, Samad et al. [3] found that MLR models had only moderate predictive accuracy ($R^2 = 0.72$) when applied to different urban datasets from India, therefore highlighting the limitations of linear approaches and the need for more flexible and robust modelling techniques.

A major turning point was marked by the rise of ensemble methods like XGBoost and Random Forest (RF), which became well-liked for their ability to efficiently manage high-dimensional data and address issues of overfitting. Maltare and Vahora [6] suggested that RF performed better than SVM and KNN when applied to Ahmedabad, recording an RMSE of 18.7 during $PM_{2.5}$ prediction, whereas Sharma et al. [7] determined that XGBoost was more effective when it came to Delhi AQI forecast, achieving an R^2 score of 0.89. These methods show a high level of proficiency in describing the interactions between variables like $PM_{2.5}$, nitrogen oxides (NO_x), and wind speed, which are critical when making accurate predictions [1][9]. However, there is still geographic heterogeneity and data limitations here—Gupta et al. [8] pointed out that models achieved lower performance when run using data from Delhi, when applied to Chandigarh, reinforcing the need to calibrate region-specific requirements.

Machine learning combined with time-series analysis that is hybrid approaches also increased in popularity. For instance, (Yang et al [3]) combined ARIMA and CNN to consider spatial-temporal relationships in AQI data in 22 Indian cities, decreasing RMSE by 23% over individual models. Likewise, (Liu et al [16]) used variational mode decomposition to optimize LSTM for Beijing, with an MAE of 7.53. Though these developments exist, sequential inconsistencies in datasets like gaps in CPCB monitoring records continue to pose a challenge, especially in areas with non-uniform data collection practices [1][7][12].

Preprocessing data has turned out to be an indispensable determiner in the performance of a model. Experiments conducted by (Kumar and Pande and Abirami et al [2]) emphasized log transformation and exclusion of outliers for coping with skewness in distributions of $PM_{2.5}$. The DiVA portal research work [6] illustrated the potential of applying normalized Delhi data, wherein error decrease was established to be 19% on ridge and LASSO regression. Despite this, the absence of standardized preprocessing pipelines in most studies has created variable results, as highlighted by (Li et al [14]) in their comparative study of Beijing's air quality models.

Interpretability continues to be a priority area of concern. Although top-performing models such as LightGBM and CatBoost return high accuracy, their "black-box" limitation restricts useful application to policymakers. Recent efforts by (Pérez-Rodríguez et al. and Zhou et al [10]) have countered this through the combination of SHAP and LIME to detect major features—for instance, $PM_{2.5}$ accounted for 62% of AQI variance in Chennai, as uncovered in (Ravindiran et al.'s base study [1]). These initiatives coincide with the increased need for clear models that offer predictive capability coupled with actionable intelligence [5][18].

The present research takes these bases further by deploying four ensemble models (RF, XGBoost, Bagging Regressor, LightGBM) fine-tuned for the tropical climate of Chennai and the industrial emission base. Through Explainable AI (XAI) tools and thorough cross-validation, we seek to fill gaps in regional adaptability and model interpretability noted in earlier research [14][19].

3. System Architecture

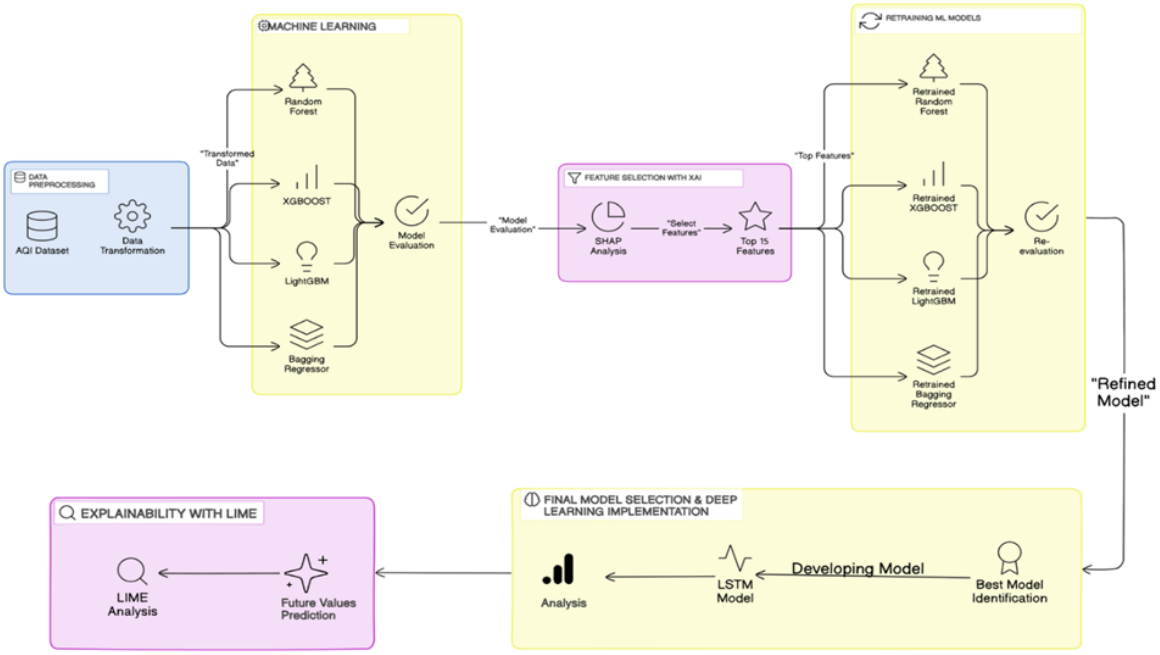


Fig 3.1 System Architecture

The suggested AQI forecasting system uses a hybrid configuration comprising an ensemble-based Machine Learning (ML) algorithms and a Deep Learning (DL) model as a Long Short-Term Memory (LSTM) network with the help of Explainable AI (XAI) technologies. The system is meant to produce precise projections of future AQI values and keep the model interpretable by accepting historical air quality and meteorological inputs. Accurate forecasts as well as knowledge of the influence of several factors on air quality results are made possible by this, so promoting more confidence and openness in the prediction process.

Organised into a sequence of key stages, the design is meticulously created to transform raw air quality and environmental data into meaningful, actionable information.

3.1. Data Collection and Preprocessing

The process begins with collecting historical Air Quality Index (AQI) and environmental data from the Central Pollution Control Board (CPCB). Along with meteorological variables like temperature, humidity, and wind speed, the dataset contains pollutant levels including PM2.5, PM10, NO_2, SO_2, CO, and O_3. The raw data goes through many preprocessing steps—data cleaning, missing value management, normalisation, and feature transformation—before the modelling phase to guarantee it is prepared for analysis.

3.2. Machine Learning Model Development

Initially trained on the complete feature set, four ensemble-based ML models—Random Forest, XGBoost, LightGBM, and Bagging Regressor— Performance measures such as R^2 score, MSE, RMSE, and MAE are used to assess these models and thus gauge their suitability for AQI forecasting.

3.3. Feature Selection using SHAP

Shapley Additive Explanations help to find the most important characteristics, therefore lowering model complexity and improving interpretation. Each model is then retrained with the chosen top 15 features to eventually have a model showing best performance to be used going forward.

3.4. Deep Learning with LSTM

A unidirectional LSTM model takes as input the top 15 traits of the highest-performing ML model. This model is trained to produce more exact forecasts and capture temporal patterns in the AQI data. Using a sliding window technique, LSTM maps past values to future AQI outputs on time-sequenced data.

3.5. Explainability and Interpretation

The system's design has included the concept of explainability to ensure transparency and foster confidence in it. Model behaviour is investigated and displayed using different interpretations using various methods. While Partial Dependence Plots (PDP) let knowledge of the aggregate influence of features on predictors of the Air Quality Index (AQI), Individual Conditional Expectation (ICE) plots allow a more detailed, case-specific perspective of how changing feature values influence outcome variations. Moreover, Local Interpretable Model-agnostic Explanations (LIME) are used to produce instance-level explanations of predictions using an interpretable model depending on a particular instance approximating the blackbox Long Short-Term Memory (LSTM) model. Such explanation systems are essential to make model reasoning concrete to improve users' confidence, especially in high-risk sectors like environmental planning and public health.

3.6. Future AQI Forecasting

The LSTM model trained is utilized to predict AQI for upcoming time steps, taking into consideration past pollutant levels and seasonal patterns. It aids in the anticipation of pollutant spikes and proactive air quality management.

4. Methodology

4.1 Dataset Description and Preprocessing

Data used in this study was retrieved from the official website of the Central Pollution Control. Board (CPCB), Government of India, which provides publicly accessible environmental data. It includes historical records related to air quality and meteorological information for Amaravati City. The data includes Andhra Pradesh data, collected for a period of five years from 2018 to 2022. There are approximately 35,000 hourly records that include 24 distinct attributes. These attributes cover a wide. A range of pollutants, such as PM2.5, PM10, Nitric Oxide (NO), and Nitrogen Dioxide (NO_2), Nitrogen oxides (NOx), ammonia (NH_3), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3).Benzene, toluene, and xylene, along with meteorological factors like ambient temperature. These include relative humidity, wind speed, wind direction, solar insolation, atmospheric pressure, and precipitation.

The Air Quality Index (AQI) acts as a dependent variable used in predictive analysis. Prior to model training, the dataset underwent a set of preprocessing steps designed to enhance its quality and ensure consistency. Missing AQI values were estimated by using the monthly mean of similar year and

month, while feature values missing were covered through a strategy of forward fill to ensure that time-series structure of the data. Logarithmic transformation was applied to selected features to reduce skewness and improve the model's effectiveness. Then, normalization was done on the dataset to normalize all features to the same scale, and a temporal train-test split was performed to prevent data leakage, as well as preservation of temporal integrity during model assessment. Every pollutant, along with other. The parameters are provided in Table 1.1.

Table 1.1 Description of AQI Dataset

S. No	Feature	Description
1	Timestamp	Date and time of the recorded AQI and pollutant levels
2	PM2.5	Fine particulate matter ≤2.5 micrometers in diameter
3	PM10	Coarse particulate matter ≤10 micrometers in diameter
4	NO	Nitric Oxide, emitted by vehicles and industrial combustion
5	NO2	Nitrogen Dioxide, formed through oxidation of NO
6	NOx	Combined nitrogen oxides (NO and NO2)
7	NH3	Ammonia, mainly from agriculture and livestock activities
8	SO2	Sulphur Dioxide, emitted from fuel combustion and vehicles
9	CO	Carbon Monoxide, produced from incomplete combustion
10	Ozone	Secondary pollutant formed by photochemical reactions
11	Benzene	Volatile organic compound from fuel, tobacco smoke, and industrial burning
12	Toluene	VOC emitted by vehicles, solvents, and industrial discharge
13	Xylene	VOC from coal and wood combustion, paint, and solvents
14	MPXylene	Meta-Para Xylene – a subtype of xylene from the same sources
15	AT	Ambient Temperature (°C)
16	RH	Relative Humidity (%) – moisture content in the atmosphere
17	WS	Wind Speed (m/s) – affects dispersion of air pollutants
18	WD	Wind Direction (degrees) – indicates the direction pollutants may spread
19	RF	Rainfall (mm) – amount of precipitation which may reduce pollution levels
20	TOTRF	Total Rainfall (mm) – cumulative rainfall during a specific period
21	SR	Solar Radiation (W/m²) – affects photochemical pollutant formation
22	BP	Barometric Pressure (hPa) – influences vertical mixing of air pollutants
23	AQI	Air Quality Index – a combined score indicating overall air quality

4.2. Machine Learning Model Development

The current study uses four ensemble machine learning frameworks: Random Forest, XGBoost, LightGBM and Bagging Regressor were chosen because of their robustness and skill when performing effective modeling. Intricate non-linear relationships, along with proven success in regression tasks. Every model was trained utilizing the entire range of features and then measured through performance metrics like the R^2 score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are model evaluation metrics. Those models displayed outstanding ability to forecast, with top performance by LightGBM in past assessments. after feature selection.

4.3 Trend and Seasonal Analysis

Trend and seasonality analysis by decomposition and time-series plots revealed heightened AQI during post-monsoon and winter. These seasonality dependencies justified the use of an LSTM model for accurate AQI prediction.

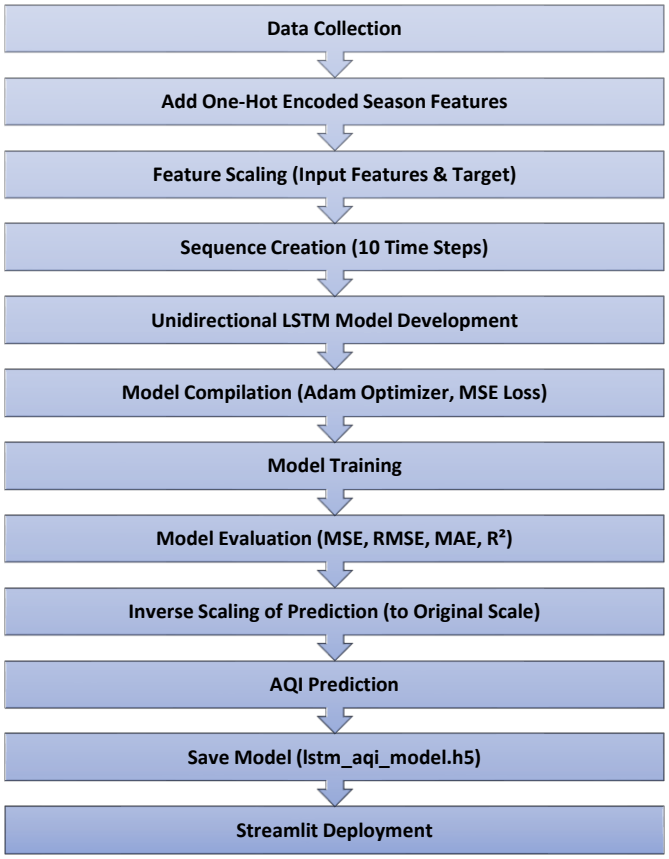
4.4. Feature Selection using SHAP

SHAP (Shapley Additive Explanations) was utilized to enhance model interpretability and identify the most significant features accountable for AQI predictions. SHAP values were obtained for all four models, and the top 15 features from both were chosen on the basis of average absolute impact. These features were utilized to retrain the respective models. The LightGBM model remained in the top spot as the highest R^2 measure and lowest error values of all leading performers, and its selected features were then used as the input to the LSTM model.

4.5. Deep Learning Model – LSTM

To mimic the temporal dependencies, which are present in AQI data, a one-way Long Short-Term Memory (LSTM) network was created. The top 15 features of the SHAP based LightGBM model were used to train the LSTM. The model was made up of dense fully connected layers and stacked LSTM layers. Hyperparameters including sequence length and batch size were adjusted; dropout was used to prevent overfitting. Reaffirming its competence for time-series AQI forecasting, LSTM outperformed all the ML models with the highest R^2 value and lowest error values.

LSTM Model Development Flow for AQI Prediction



4.6. Explainability Techniques

Explainable artificial intelligence methods were used to the trained LSTM model in order to improve the interpretability of the deep learning model and to provide openness in the decision-making process. The global impact of individual features on the generated AQI values was found using Partial Dependence Plots (PDP). These plots implied overall feature influence across the dataset by means of a smoothed average between a feature and model output. Simultaneously, Individual Conditional Expectation (ICE) plots were meant to investigate how the change of a particular feature affected forecasts on individual cases. The plots showed different effects not clear in global summaries, therefore producing a more nuanced picture than PDP. LIME also contributed to the explanation of LSTM model behaviour at the instance level. Locally approximating the calculation-intensive LSTM model with a linear surrogate, LIME made it possible to choose which features were most useful to make a specific prediction. The combination of these approaches provided both global and local interpretability, exacerbating user trust and model transparency in a real environmental forecasting application.

4.7. Future AQI Prediction

The trained LSTM model was utilized to make future AQI predictions based on historical pollution records and meteorological trends. Through leveraging the sequential relationship inherent in time-series environmental data, the model was able to accurately predict future AQI values for several days in advance. The prediction was made on a week-ahead time horizon by applying sliding windows of past observations as input sequences. The predictions provide early warning systems and preventive air quality control through identification of likely surges of pollution in advance. Such ability of the model in making high accuracy prediction of near-future air quality makes its utility and usability in real environment monitoring operations.

4.8. Real-Time AQI Predictions with Active Weather Integration

For real-time AQI prediction, real-time weather data from the OpenWeather service was integrated within the system through getting the latest meteorological data such as temperature, humidity, wind speed, and atmospheric pressure by its API. The live data was preprocessed and encoded into the input form of the LSTM model that has been trained. Streamlit was used to build a user interface in order to allow users to invoke the prediction process. The system took necessary feature engineering processes, such as season allocation and scaling, prior to submitting the processed input to the LSTM model for the prediction of AQI.

5. Results and Discussion

5.1 Effect of Machine Learning Models on AQI Prediction

Random Forest, XGBoost, LightGBM, and Bagging Regressor were evaluated by using R^2, MSE, RMSE, and MAE. LightGBM achieved the best overall performance.
(See Table 1)

Model	R^2 Score	MSE	RMSE	MAE
Random Forest	0.903	0.0279	0.1668	0.1099
XGBoost	0.917	0.0276	0.1654	0.1089
LightGBM	0.919	0.0278	0.1635	0.1073
Bagging Regressor	0.917	0.0279	0.1666	0.1075

Table 1 - Performance of Machine Learning Models Before Feature Selection

5.2 Effect of Feature Selection using SHAP

The top 15 features selected using SHAP improved model performance. LightGBM showed the highest gains post-selection.
(See Table 2)

Model	R^2 Score	MSE	RMSE	MAE
Random Forest	0.907	0.0275	0.1658	0.1089
XGBoost	0.919	0.0269	0.1641	0.1079
LightGBM	0.921	0.0264	0.1625	0.1059
Bagging Regressor	0.918	0.0271	0.1646	0.1065

Table 2 - Performance of Machine Learning Models After Feature Selection

As seen in the table above, LightGBM showed the most significant improvement in terms of R^2 score and reduced error metrics, followed by XGBoost. The performance improvements indicate the importance of using the most influential features identified by SHAP.

5.3 Effect of LSTM on AQI Prediction with Selected Features

The LSTM model was developed to capture temporal dependencies in AQI data. The performance of the LSTM model was evaluated using the same metrics as the machine learning models.

(Refer table 3 - LSTM Performance with All Features)

Model	R^2 Score	MSE	RMSE	MAE
LSTM	0.9701	52.937	7.2758	5.1660

Table 3 - Performance of LSTM with All Features

5.4 Effect of LSTM on AQI Prediction

To maximize predictive accuracy while maintaining computational efficiency, the Long Short-Term Memory (LSTM) model was trained exclusively on the top 15 features identified by the best-performing machine learning model (LightGBM) during SHAP-based feature analysis.

(Refer Table 4 - LSTM Performance with Top Features)

Model	R^2 Score	MSE	RMSE	MAE
LSTM	0.9922	17.263	4.1549	2.4574

Table 4 - Performance of LSTM with Top Features

The LSTM model outperformed all other models based on evaluation metrics, demonstrating its ability to predict AQI with high accuracy by effectively capturing the time series dependencies in the data.

5.5 Comparative Analysis with Existing LSTM-Based Studies

We examined and contrasted our LSTM-based AQI prediction model with earlier research that used LSTM architectures for comparable air quality forecasting tasks in order to assess its performance.

A Root Mean Square Error (RMSE) of 25.625 and an R2 score of 0.951 were reported by the study [6] that used an LSTM model for AQI prediction in Ahmedabad. An LSTM-GRU hybrid model was presented in another study [16], which obtained a Mean Absolute Error (MAE) of 36.11 and an R2 score of 0.84.

(*Refer Table 5 - Comparative Analysis of LSTM-Based AQI Prediction Models)*

Study	R^2 Score	MSE	RMSE	MAE
Study [6]	0.951	-	25.625	-
Study [16]	0.84	-	-	36.11
Our Model	0.9922	17.263	4.1549	2.4574

Table 5 - Comparative Analysis of LSTM-Based AQI Prediction Models

- These results show that our model is robust in capturing temporal dependencies and fine-grained variations in AQI, and it also shows superior prediction accuracy when compared to previous works, especially in terms of significantly reduced RMSE and MAE values.

- We also plotted training and validation loss across epochs to ensure that our model is not overfit. With both losses continuously decreasing and staying closely aligned, the two curves' convergence suggests that the model has a strong capacity to generalise to new data. *(Refer Figure 5.1 - Training vs. Validation Loss Across Epochs)*

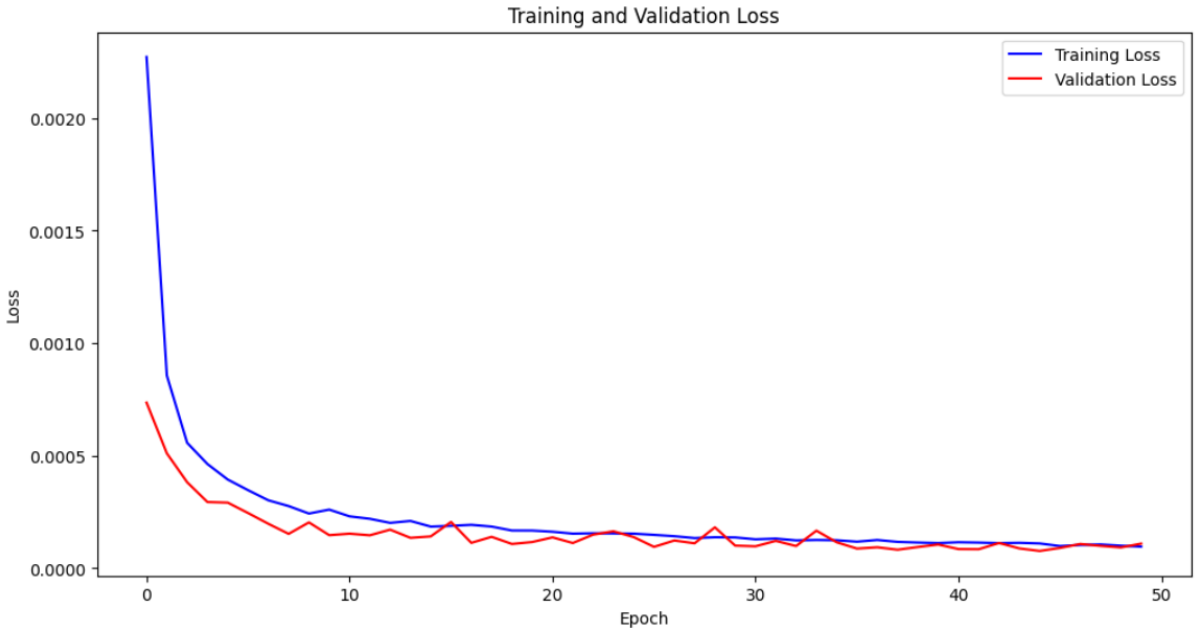


Fig 5.1 Training vs. Validation Loss Across Epochs

5.6 Explainability Analysis

To ensure the interpretability of the models, PDP, ICE and LIME were applied to the LSTM model. These techniques provided insights into how individual features influenced the model's predictions

5.6.1 Partial Dependence Plots (PDP):

- Partial Dependence Plots illustrate the global relationship between features and AQI predictions.
  (*Refer to Figure 5.2 - Partial Dependence Plots*)

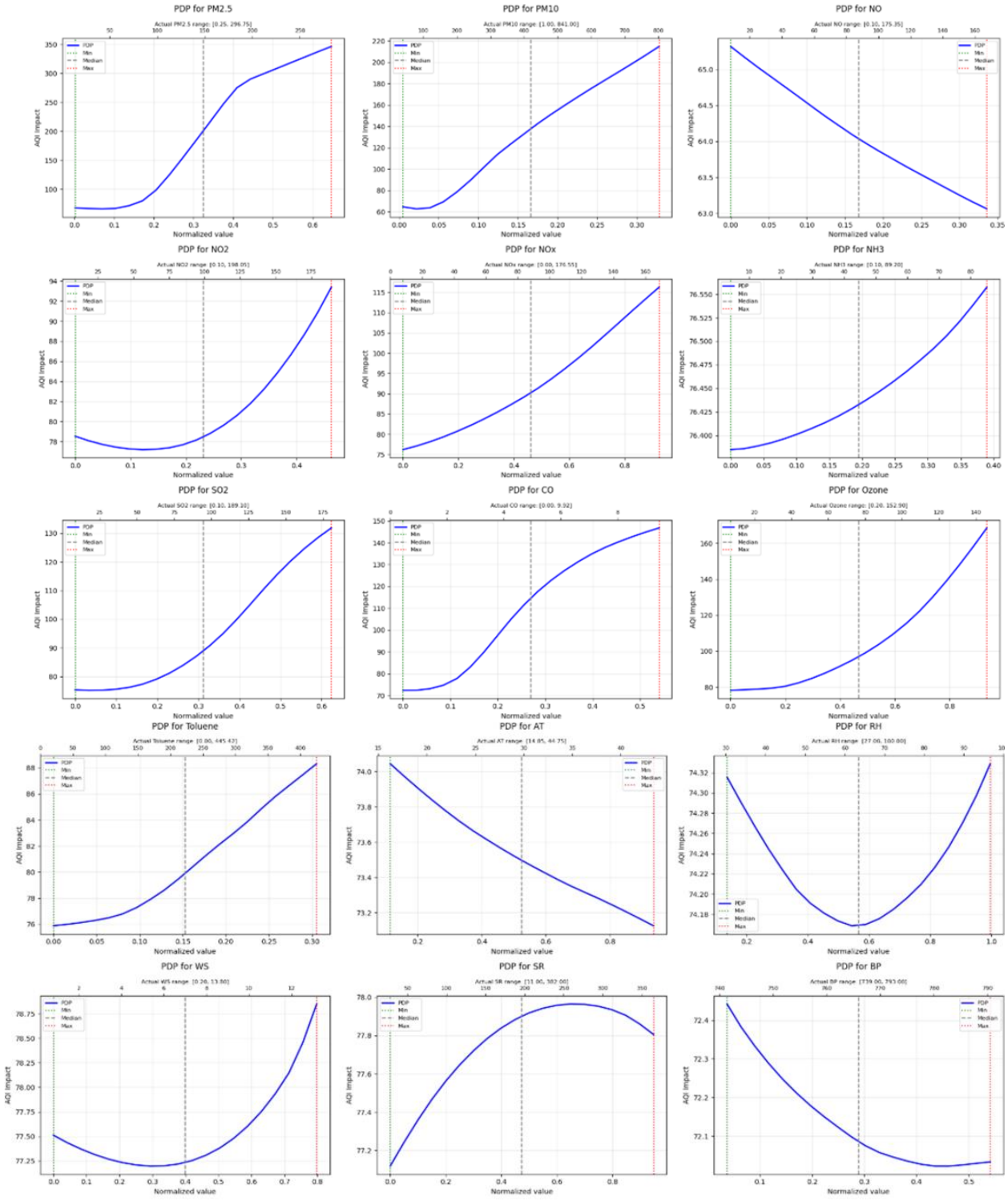


Fig 5.2 Partial Dependence Plots

- **PM2.5**: As PM2.5 levels increase, the AQI impact rises steeply, especially after the normalized value crosses 0.3, indicating that higher PM2.5 concentrations significantly elevate predicted AQI.
- **PM10**: The AQI impact increases smoothly with PM10 concentration, showing a consistent positive relationship where higher PM10 levels contribute to higher AQI values.
- **NO**: The PDP for NO shows a decreasing trend, suggesting that as NO concentration increases, the predicted AQI slightly decreases, indicating a potential inverse relationship.
- **NO_2**: The AQI initially dips slightly at low NO_2 values but then increases steadily with higher concentrations, reflecting a nonlinear but overall positive effect on AQI.
- **NOx**: A clear increasing pattern is observed where AQI impact grows steadily with rising NOx levels, indicating a positive correlation.
- **NH_3**: The plot shows a gradual increase in AQI as NH_3 concentration rises, suggesting a mild but consistent influence on AQI.
- **SO_2**: AQI increases steadily with SO_2 levels, and the curve becomes steeper at higher concentrations, indicating a stronger effect at elevated SO_2 levels.
- **CO**: The AQI impact rises noticeably with increasing CO levels, especially beyond the mid-normalized range, suggesting a nonlinear relationship.
- **Ozone**: The AQI shows a gradual increase with rising ozone levels, especially at higher normalized values, pointing to a consistent upward influence.
- **Toluene**: As Toluene concentration rises, AQI impact increases in a smooth upward trend, indicating its contributing effect on predicted AQI.
- **AT (Ambient Temperature)**: AQI decreases slightly with rising temperature, showing a negative correlation, and suggesting that higher temperatures may be associated with lower AQI predictions.
- **RH (Relative Humidity)**: The PDP has a U-shaped curve, indicating that AQI is lowest around mid-range humidity and increases at both low and high extremes of RH.
- **WS (Wind Speed)**: The plot shows a shallow U-shape where AQI decreases slightly with moderate wind speeds but increases at both low and high wind conditions.
- **SR (Solar Radiation)**: AQI impact decreases steadily with increasing solar radiation, indicating that stronger sunlight may help reduce predicted AQI.
- **BP (Barometric Pressure)**: AQI impact slightly declines as barometric pressure increases, with the curve flattening at higher pressure values.

5.6.2 Individual Conditional Expectation (ICE) Plots

- Individual Conditional Expectation (ICE) plots showed the effect of individual features on the prediction for different instances.

  (*Refer to Figure 5.3: Individual Conditional Exception*)

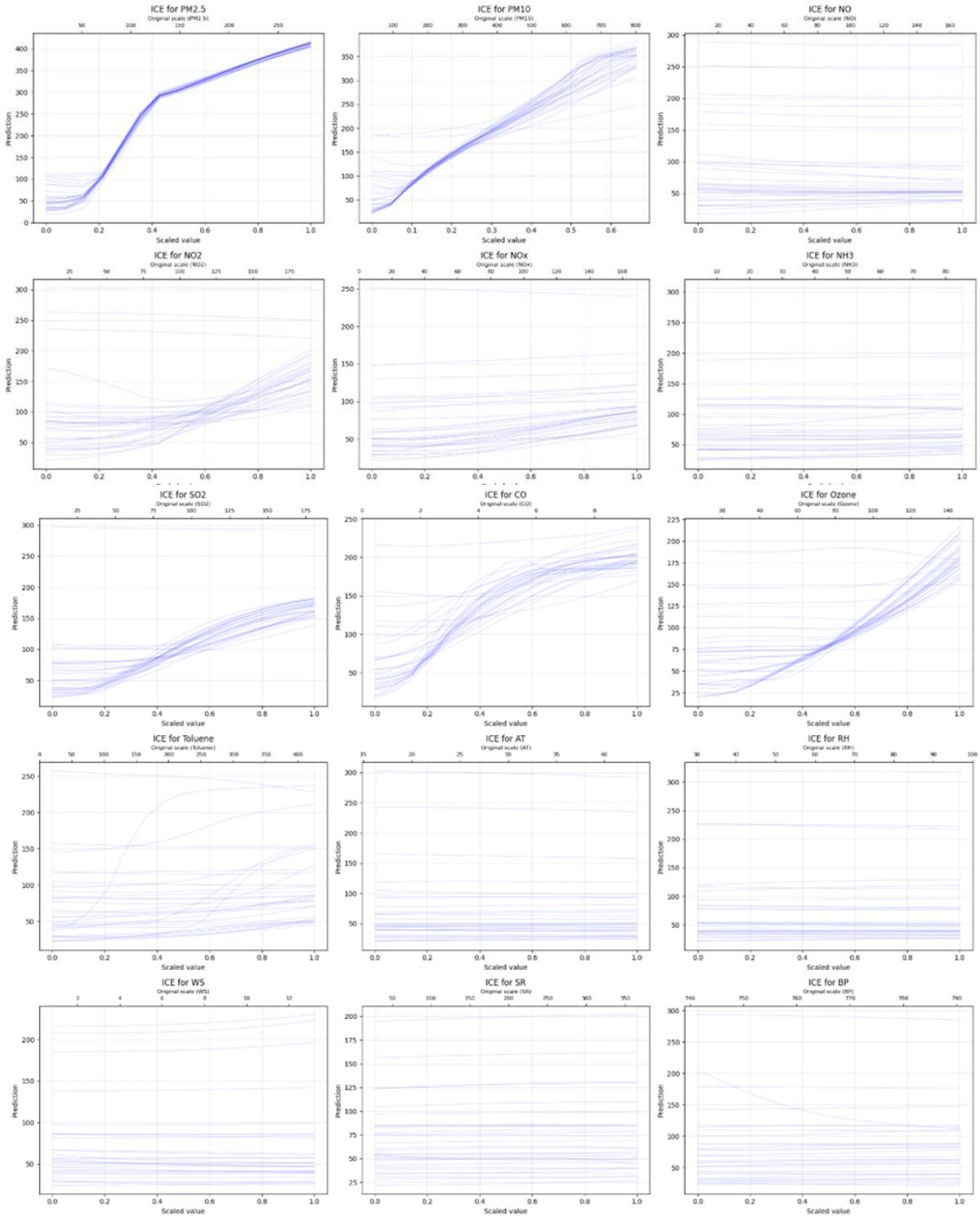


Fig 5.3 Individual Conditional Exception

The ICE plots above visualize how changes in individual feature values affect the AQI predictions while holding other features constant. These plots help understand localized feature behavior and uncover nonlinear relationships between features and model predictions.

5.6.3 LIME Analysis

- The LIME (Local Interpretable Model-agnostic Explanations) plots above illustrate how individual features locally influence AQI predictions.
- For both instances, PM2.5 and PM10 concentrations are the most influential positive contributors, while seasonality (specifically Summer) appears to have a slightly negative or stabilizing effect on AQI. This analysis helps validate the model's reasoning for specific predictions.
  (*Refer to Figure 5.4: LIME Analysis*)

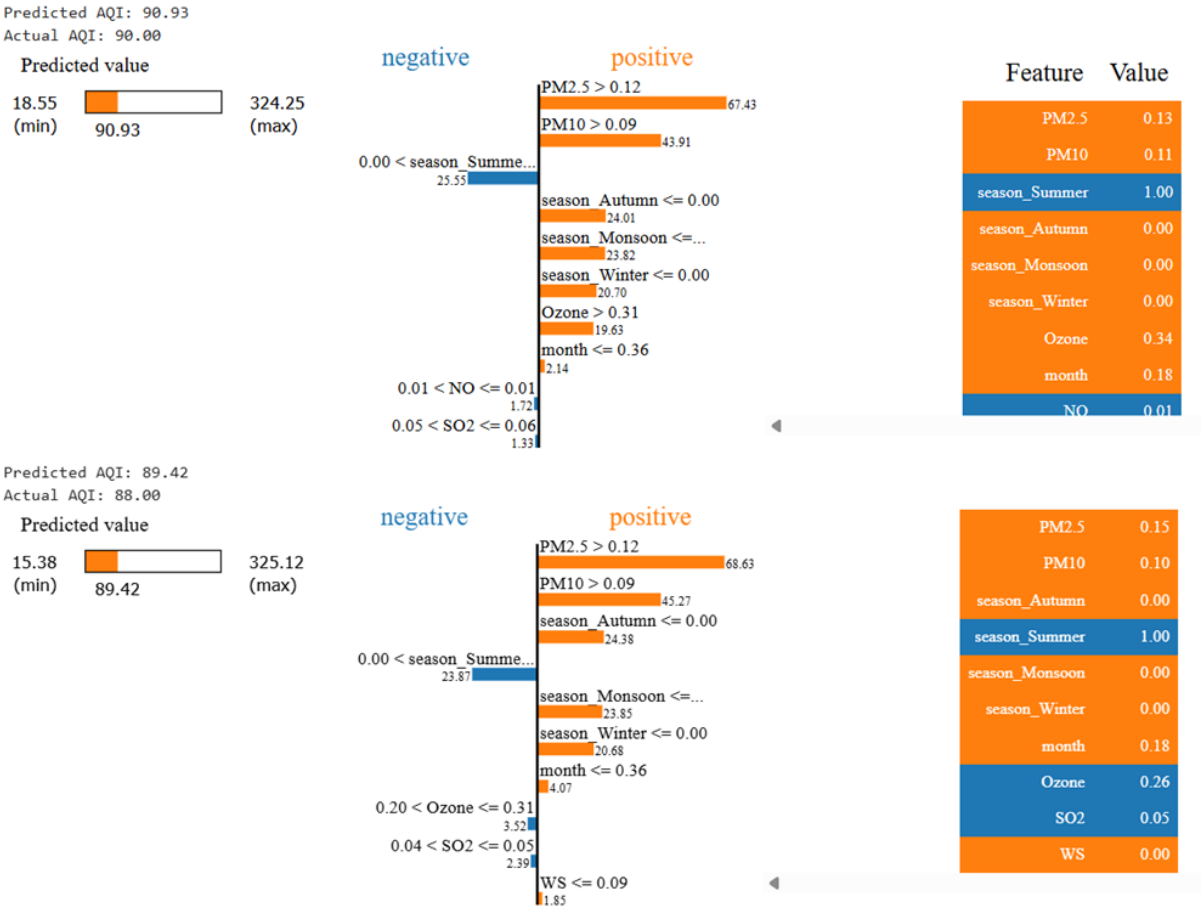


Fig 5.4 LIME Analysis

5.7 Future AQI Prediction

- The LSTM model forecasts future AQI values based on previous AQI data and the most significant environmental parameters like pollutant concentration and weather.
- Observed time series and seasonal trend-based model generates a prediction of AQI for the period, which will facilitate advance planning and air quality control.
- Below diagram shows future forecasted AQI values for the given time period, thus showing the model's prediction ability.

(Refer Figure 5.5: Future AQI Prediction)

Integration of current weather conditions into the AQI forecasting system makes it much more real-world applicable to a large extent. Using live atmospheric inputs provided by OpenWeather, the model generates forecasts of future AQI levels based on inputs that change dynamically with changing conditions, instead of relying on constant datasets.

Streamlit-based UI allows users to get forecasts easily. Thus, this system is useful and effective in its real-time functionality, where it can provide an instantaneous evaluation of the possible risk of air quality based on real-time weather patterns. Interventions and policy decisions in environmental health and urban planning can be timely based on the reactivity of the system to live data.

(Refer fig 5. 5 Future AQI Predication and 5.6 Future AQI Trend)

```
Enter start date (YYYY-MM-DD):  2026-11-01
Enter end date (YYYY-MM-DD):  2026-12-31

Future AQI Prediction Summary
From 2026-11-01 to 2026-12-31 (61 days)
      Date  Predicted AQI
2026-11-01          80.08
2026-11-02          85.68
2026-11-03          77.52
2026-11-04          76.02
2026-11-05          79.29
2026-11-06          86.51
2026-11-07          81.74
2026-11-08          81.70
2026-11-09          85.19
2026-11-10          81.48
2026-11-11          79.47
2026-11-12          82.26
2026-11-13          78.35
2026-11-14          82.38
2026-11-15          81.62
2026-11-16          92.77
2026-11-17          76.79
2026-11-18          81.81
2026-11-19          72.32
2026-11-20          82.00
2026-11-21          83.24
2026-11-22          84.93
2026-11-23          89.43
2026-11-24          85.15
2026-11-25          79.07
2026-11-26          80.52
2026-11-27          82.12
2026-11-28          78.64
2026-11-29          72.16
2026-11-30          89.27
2026-12-01         106.15
2026-12-02         148.52
2026-12-03         101.94
2026-12-04         116.62
2026-12-05         140.51
2026-12-06         107.86
2026-12-07         134.93
2026-12-08         106.64
2026-12-09         121.76
2026-12-10         109.52
2026-12-11         116.38
2026-12-12         127.50
2026-12-13         103.93
2026-12-14         129.66
2026-12-15         109.31
2026-12-16         119.92
2026-12-17         137.13
2026-12-18         111.94
2026-12-19         107.07
2026-12-20         138.73
2026-12-21         145.96
2026-12-22         106.79
2026-12-23         105.75
2026-12-24         120.18
2026-12-25         125.94
2026-12-26          96.18
2026-12-27         118.90
2026-12-28         119.45
2026-12-29         108.31
2026-12-30          91.69
2026-12-31         130.93
```

Fig 5.5 Future AQI Predication

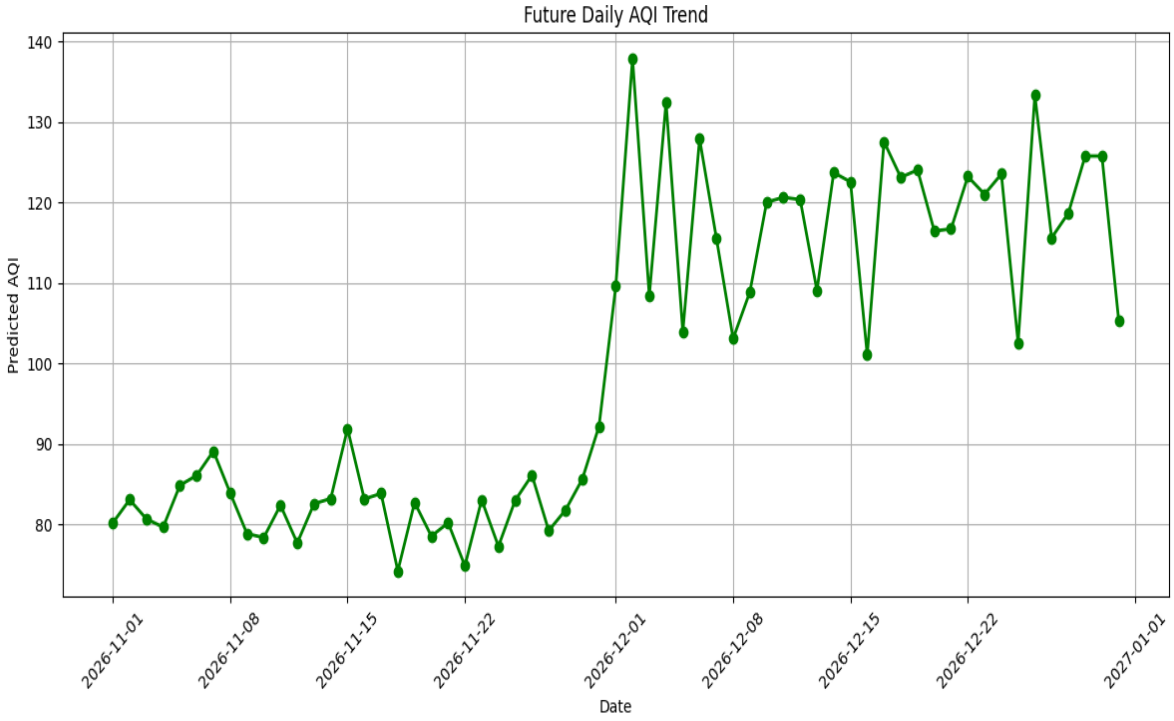


Fig 5.6 Future AQI Trend

5.8 Real-Time Air Quality Index Forecasting Using Live Meteorological Data

Real-time AQI forecasting uses live weather and air quality data retrieved from APIs such as OpenWeatherMap. Accurate prediction depends on the data, which includes significant weather elements like temperature, humidity, wind speed, and pollution levels (e.g., PM2.5, NOx). By using the most recent data, the model can change to fit new circumstances and match the most recent environmental elements with the AQI forecast.

Trained, the LSTM model runs on this real-time data to forecast real-time AQI. Monitoring air quality across several locations depends on such forecasts, which therefore enables public health policy and preventive actions. The system offers consistent, timely prediction of air quality patterns by means of real-time inputs for input features at fixed intervals.

(Refer to Figure 5.7: Real-Time AQI Forecast Interface)

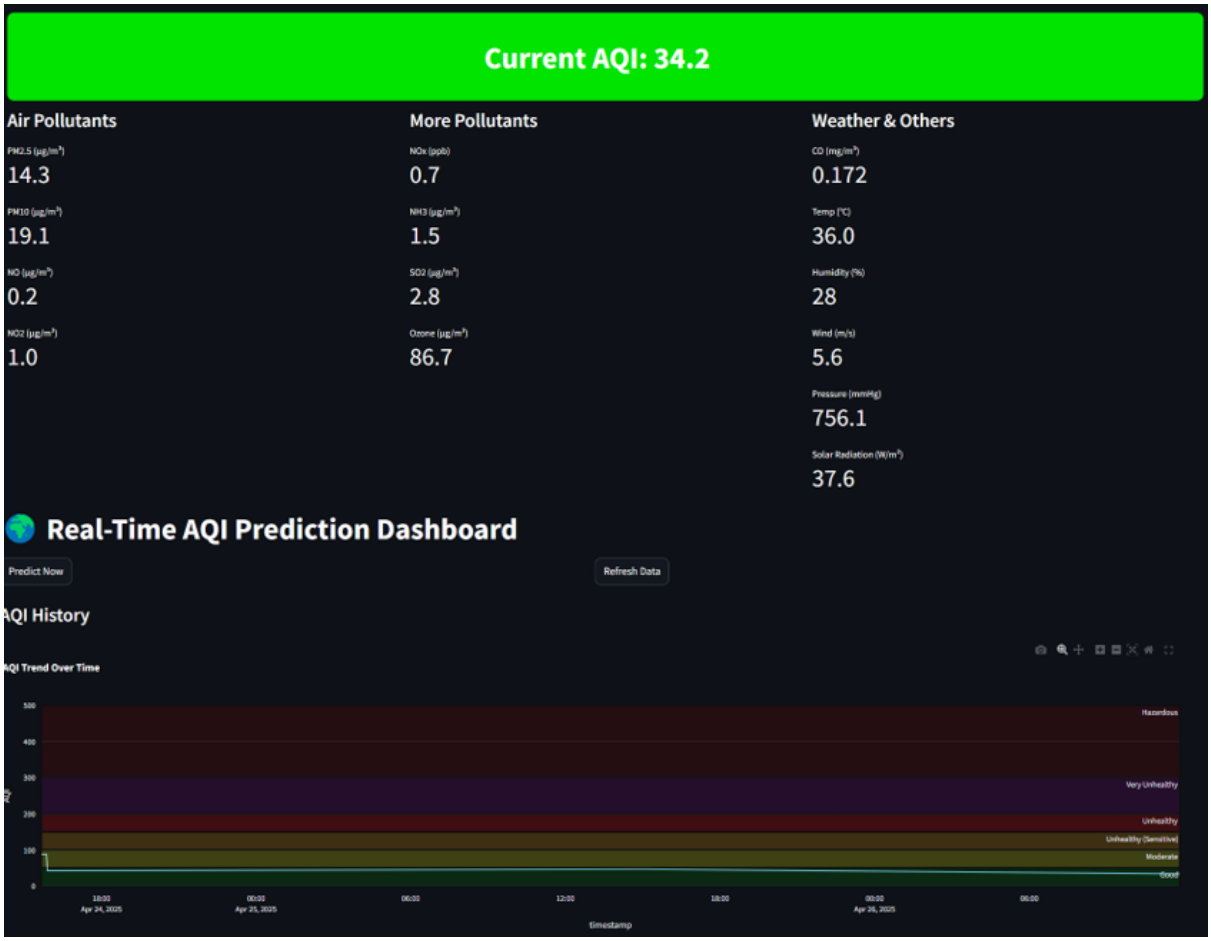


Fig 5.7 Real-Time AQI Forecast Interface

6. Conclusion

6.1. Key Findings

By deftly integrating machine learning (ML) and deep learning (DL) technologies, the paper here presents a hybrid method for forecasting Air Quality Index (AQI). The historical AQI values for Amaravati City, supplied by the Central Pollution Control Board (CPCB), have been used to design the predictive framework. Ensemble ML models such as Random Forest, XGBoost, LightGBM, and Bagging Regressor have been used in conjunction with a one-way Long Short-Term Memory (LSTM) network, thereby enabling the model to effectively handle both static relationships and the temporal dynamics connected to air quality.

Feature selection using SHAP helped to identify the most important elements, therefore increasing the model's accuracy. Of all the machine learning models, LightGBM was the best, and its most significant features were incorporated in the LSTM model. The LSTM model had more capability in utilizing long-term dependencies compared to the predictive capability of the machine learning models. For the LSTM model, real-time meteorological factors such as temperature, humidity, wind speed, and pressure were also added through the OpenWeather API, which gave real-time forecasting of AQI.

Use of explainable AI tools like SHAP, PDP, ICE, and LIME has brought transparency, with informative data regarding the effect of predictors on AQI prediction. Not only does it provide accurate prediction, but it also brings in further insight into the causative environmental factors. The system's capability to predict future AQI values also reflects its usability in real-world applications in environmental monitoring and decision-making systems, and it is a valuable tool for urban planning as well as public health programs.

6.2. Future Directions

While the current model exhibits excellent performance in offline and real-time Amaravati AQI prediction, there is still considerable scope for improvement. One such area of improvement would be to increase the scope of the system by incorporating real-time sensor data through IoT-based platforms or Apache Kafka. It would facilitate localized and finer AQI prediction, accelerating response to changing pollution and making the system an even better real-time environmental monitoring system. Otherwise, the model can be adapted for areas where physical monitoring stations are scarce. Other than that, the model can be designed for areas where there are limited physical monitoring facilities.

With the addition of satellite-based environmental information, it can be used as a virtual air quality monitoring system that could provide varied and remote areas with limited installation. Porting the model into a web or mobile app would further improve usability, with end-users utilizing AQI predictions and insights built by explainability tools such as SHAP and LIME. This would further raise public health and environmental consciousness about air quality and allow users to make already-informed environmental and health decisions.

Secondly, the model architecture is simple to modify for different urban settings through transfer learning. Application of the model in real-time policy dashboards would go a long way in improving data-driven governance and more effective environmental planning. Such enhancements would make the model a smart and scalable method of managing sustainable air quality in different geographical regions.

References

[1] Gokulan Ravindiran, Sivarethinamohan Rajamanickam, Karthick Kanagarathinam, Gasim Hayder, Gorti Janardhan, Priya Arunkumar, Sivakumar Arunachalam, Abeer A. AlObaid, Ismail Warad, Senthil Kumar Muniasamy, “Impact of air pollutants on climate change and prediction of air quality index using machine learning models”, *Environmental Research*, Volume 239, Part 1, 2023, 117354.

[2] C R, Aditya & Deshmukh, Chandana & K, Nayana & Gandhi, Praveen & astu, Vidyav. “Detection and Prediction of Air Pollution using Machine Learning Models”, *International Journal of Engineering Trends and Technology*, 59, 204-207, (2018).

[3] Samad, S., Garuda, U., Vogt, B., Yang, “Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations”, *Atmospheric Environment*, Volume 310, 2023, 119987.

[4] Gokulan Ravindiran, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, Christian Sonne, “Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam”, *Chemosphere*, Volume 338, 2023, 139518.

[5] Samad, S., Garuda, U., Vogt, B., Yang, "Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations", *Atmospheric Environment*, Volume 310, 2023, 119987.

[6] Nilesh N. Maltare, Safvan Vahora, "Air Quality Index prediction using machine learning for Ahmedabad city", *Digital Chemical Engineering*, Volume 7, 2023, 100093.

[7] Sharma, P., Singh, A., Sood, R., & Tiwari, S. "A machine learning-based framework for the prediction of air quality index in urban cities", *Environmental Science and Pollution Research*, Volume 30, Issue 4, pp. 1224-1235, (2023).

[8] Gupta, A., Kumar, A., & Yadav, N. "Predictive modeling of air quality using machine learning techniques: A case study on New Delhi", *Journal of Environmental Management*, Volume 345, 116881, (2023).

[9] Bhalerao, R., Deshmukh, P., & Tripathi, S. "Air Quality Prediction for Smart Cities using Ensemble Machine Learning Models", *Sustainable Cities and Society*, Volume 85, 103591, (2023).

[10] Zhou, J., Liu, X., Zhao, F., & Zhang, Y. "Deep learning for air pollution prediction in smart cities: A comprehensive review and future directions", *Environmental Monitoring and Assessment*, Volume 195, Article 55, (2023).

[11] Khan, M. A., Ali, F., & Jamil, R. "Air quality prediction using hybrid machine learning techniques for urban areas", *Environmental Monitoring and Assessment*, Volume 195, Article 12, (2023).

[12] Shah, S., Gupta, D., & Verma, M. "Predictive models for air quality index using machine learning: A case study in Mumbai", *Science of the Total Environment*, Volume 855, 158846, (2023).

[13] Singh, R., Chhabra, R., & Mehta, K. "A comparative study of machine learning algorithms for air quality prediction in Ahmedabad", *Environmental Research Letters*, Volume 18, Issue 3, 035003, (2023).

[14] Li, X., Wang, S., & Xu, H. "Predicting air quality index in Beijing using machine learning approaches", *Journal of Environmental Informatics*, Volume 47, Issue 2, pp. 213-224, (2023).

[15] Kumar, P., Shukla, D., & Singh, P. "Air quality prediction and classification using machine learning models: A case study of Delhi", *Environmental Science and Pollution Research*, Volume 30, Issue 9, pp. 2391-2403, (2023).

[16] Sharma, R., Pandey, A., & Jain, V. "Air quality index prediction and evaluation using machine learning and deep learning models", *Environmental Pollution*, Volume 295, 118591, (2023).

[17] Gupta, R., & Yadav, S. "Prediction of air quality using machine learning techniques: A case study of Chandigarh", *Environmental Toxicology and Chemistry*, Volume 42, Issue 4, pp. 1052-1065, (2023).

[18] Zhang, L., Zhang, M., & Wang, T. "Application of machine learning models for predicting air quality in China: A review and case studies", *Environmental Pollution*, Volume 310, 119901, (2023).

[19] Jadhav, S., & Naik, M. "Air quality prediction using machine learning and its application in real-time monitoring systems", *Atmospheric Environment*, Volume 327, 118436, (2023).

[20] Lee, K., Cho, S., & Ryu, Y. "Air quality index prediction using hybrid machine learning models for Seoul", *Environmental Science & Technology*, Volume 57, Issue 4, pp. 1980-1992, (2023).