Hybrid RF-LSTM: An Advance Machine Learning Technique for Prediction of Air Pollution in Indian Cities

Dr. Pooja Singh
Assistant Professor, Department of Computer Science
Govt. College for Women, Faridabad, India - 121002
poojasinghpooja77@gmail.com
and
Dr. Sapna Nagpal
Professor, Department of Computer Science
Smt. Sushma Swaraj Govt. College for Girls, Ballabgarh, India – 121004
sapnanagpalg@gmail.com

Abstract

India is grappling with a serious air pollution crisis that threatens the health and well-being of its large population. Primary sources of air pollution include industrial emissions, vehicular exhaust, agricultural residue burning, and dust. Fine particulate matter (PM2.5), in particular, contributes to millions of deaths and substantial economic losses. Accurate prediction of the Air Quality Index (AQI) is vital for tackling air pollution, as it enables early warnings, supports data-driven decision-making and facilitates the formulation of effective control measures. To improve AQI forecasting, this paper presents a novel hybrid approach combining Random Forest (RF) and Long Short-Term Memory (LSTM) techniques. The hybrid RF-LSTM model utilizes the RF algorithm to extract relevant features from the AQI dataset, which are then input into the LSTM network to capture temporal dependencies and long-term patterns in the sequential data. The AQI data, with extracted features for five Indian cities, Delhi, Lucknow, Chandigarh, Jaipur and Gurugram, is input into the LSTM model, and the prediction results of the hybrid RF-LSTM model are then analysed and presented. The results indicate that the hybrid RF-LSTM approach outperforms both individual RF and LSTM models in predicting various AQI constituents, thereby offering improved support for addressing air pollution.

Keywords: AQI, AQI Prediction, Random Forest, LSTM, Machine Learning.

1. Introduction

India has implemented a variety of measures to monitor and control air pollution and AQI. These efforts are carried out at national, state, and city levels, involving government bodies, regulations, technology and public awareness. India has a National Air Quality Monitoring Programme (NAMP) run by the Central Pollution Control Board (CPCB), which monitors air quality in over 800 stations across 344 cities and measures pollutants like PM2.5, PM10, SO_2, NO_2, CO, O_3, NH_3, and Pb. There is SAFAR (System of Air Quality and Weather Forecasting and Research) developed by IITM (Indian Institute of Tropical Meteorology) Pune and IMD (India Meteorological Department), which provides real-time AQI values and 3-day forecasts for major cities like Delhi, Mumbai, Ahmedabad, and Pune. India uses a colour-coded AQI scale (0–500) for AQI standardization with categories from "Good" to "Severe" to communicate risk to the public. There are various policies and regulations to monitor and control the AQI levels in metropolitan cities such as National Clean Air Programme (NCAP) which was launched in 2019 with a goal to reduce PM2.5 and PM10 by 20–30% in 131 non-attainment cities by 2026; Graded Response Action Plan (GRAP) which was notified in Delhi NCR in January 2017 and includes a step-by-step action plan based on AQI levels containing ban on diesel generators, restrictions on construction, vehicle bans (odd-even, etc.), closure of schools during severe pollution, Bharat Stage Emission Norms (BS-VI) etc.

India is facing major concerns for PM2.5 and PM 10 levels due to stubble burning, vehicle emissions, industries, construction, open waste burning and poor urban planning which are causing various issues such as respiratory/heart disease, smog, seasonal pollution spikes in North India, toxic gases and particles, climate impact, reduced life span etc. Air pollution can lead to many health problems, such as respiratory problems, premature death, and hospitalization for heart and lung diseases [1]. Therefore, air pollution levels must be accurately predicted to work with governments and educate the public about the dangers of pollution [2]. Dealing with the severe AQI problem in India requires a combination of strong policy enforcement, adoption of clean technology, public awareness, and behaviour change. To ensure the success of these measures, precise AQI forecasting is necessary [8,9,10,12]. This paper proposes an advanced machine learning approach to predict AQI in five major cities in India: Delhi, Chandigarh, Jaipur, Lucknow and Gurugram. The proposed approach is based on the Random Forest technique and LSTM to make predictions from AQI data. The dataset was accessed from the kaggle.com server, which has collected data from government agencies, environmental monitoring stations, and satellite data from the years 2020 to 2023, having measurements of different air pollutants for various Indian cities. The dataset was refined through pre-processing to remove incomplete entries and ensure compatibility with machine learning algorithm. LSTM, along with the random forest technique, is applied to predict AQI for five India cities. The proposed hybrid RF-LSTM model works in two stages: the first stage applies the RF model to analyse the raw AQI data in combination with key pollutants such as CO (Carbon Monoxide), PM2.5 (Fine Particulate Matter), NO_2 (Nitrogen Dioxide), and SO_2 (Sulphur Dioxide) and to find the relation of these pollutants in escalating AQI and the second stage feeds the output from the RF model into the LSTM model which discovers temporal dependencies and long-term patterns in the sequential data. The results of the hybrid RF-LSTM model are presented and analysed in the subsequent sections.

2. Review of Literature

Drewil et.al. 2022 [2] have proposed a model based on a genetic algorithm and LSTM deep learning algorithm to predict the pollution level for the next day using four types of pollutants PM10, PM2.5, CO and NO_2. The paper has used an ideal representation of LSTM to find hyperparameters and to handle noisy data. The results of air pollution prediction are also based on meta-heuristic principles to provide more accuracy.
E. Sharma et.al. 2020 [3] have built a deep learning hybrid CLSTM model where convolutional neural network (CNN) is amalgamated with the LSTM network to forecast hourly Total Suspended Particulate (TSP). The hybrid CLSTM model is comprehensively benchmarked and is seen to outperform an ensemble of five machine learning models. The efficacy of the CLSTM model is elucidated in the model testing phase at study sites in Queensland, Australia. Using performance metrics, visual analysis of TSP simulations relative to observations, and detailed error analysis. This study ascertains the CLSTM model's practical utility for air pollutant forecasting systems in health risk mitigation.
S. Bouktif et. al. 2020 [4] have proposed an approach that provides valuable benefits by improving forecasting accuracy through the use of LSTM models optimized with metaheuristic techniques like GA and PSO. However, the method also presents some challenges, such as the reliance on zero-padding to manage variable-length sequences, which can introduce noise and affect performance. Additionally, the metaheuristic calibration process increases computational effort and may require significant processing time.
Li et al. (2017) [5] have investigated the impact of heating emissions on the Air Quality Index (AQI) using Principal Component Regression (PCR) analysis. The study analyzed large-scale environmental datasets to quantify the contribution of heating-related pollutants to overall air quality deterioration, particularly during colder seasons. Their results demonstrated that heating emissions significantly affect AQI levels and identified key pollutants driving this impact. The study provided valuable insights for policymakers to develop targeted emission control strategies to improve urban air quality, especially in heating-intensive regions.
Le et al. (2020) [6] proposed a spatiotemporal deep learning framework for citywide air pollution interpolation and prediction. The model effectively captures both spatial correlations across multiple monitoring stations and temporal dependencies in air quality data. By integrating deep learning techniques, the approach enables accurate estimation of pollutant concentrations at unsampled locations and reliable short-term forecasting. Their results demonstrate improved prediction accuracy compared to traditional methods, offering valuable support for urban air quality management and real-time pollution monitoring.
Zhan et al. (2020) [7] developed a deep learning-based model to predict air quality in major cities across China. The study focused on using historical air quality data to train predictive models capable of forecasting future AQI levels. Their approach demonstrated that deep learning techniques can effectively capture the complex, nonlinear patterns in air pollution data, leading to improved forecasting accuracy compared to traditional statistical methods. This work supports the application of deep learning for large-scale, city-level air quality management.

R. Janarthanan et.al. [2021] [8] have proposed the combination of Support Vector Regression (SVR) and LSTM based deep learning model to classify the AQI values in Chennai City. They have used the Grey Level Co-occurrence Matrix (GLCM) to extract the mean, mean square error and standard deviation from the dataset. This deep learning mechanism predicts the AQI values more accurately than previous techniques and helps to plan the metropolitan city for sustainable development.

Fei Xiao et.al. [2020] [9] has proposed a weighted long short-term memory neural network extended model (WLSTME), which addresses the issue of how to consider the effect of the density of sites and wind conditions on the spatiotemporal correlation of air pollution concentration. The proposed model considers the spatiotemporal correlation between the central site and surrounding sites. The model starts with choosing surrounding sites as the neighbour sites to the central site, and their distance, as well as their air pollution concentration and wind condition, were input to multilayer perception (MLP) to generate weighted historical PM2.5 time series data. After that, historical PM2.5 concentration of the central site and weighted PM2.5 series data of neighbour sites were input into a LSTM to address spatiotemporal dependency simultaneously and extract spatiotemporal features. Finally, another MLP was utilized to integrate spatiotemporal features extracted above with the meteorological data of the central site to generate the forecast's future PM2.5 concentration of the central site.

Mao et al. (2020) [10] proposed a deep learning-based framework for air quality prediction with a focus on optimizing model structure and improving forecasting performance. The study systematically evaluated different deep learning configurations and input features to identify the most effective modelling strategies. Their results confirmed that model optimization significantly enhances predictive accuracy and reliability in air quality forecasting. This work provides valuable insights into the practical deployment and fine-tuning of deep learning models for sustainable urban air management. Sun and Huang (2020) [11] introduced a hybrid air pollutant concentration prediction model that integrates secondary decomposition and sequence reconstruction techniques. By decomposing complex air quality time series and reconstructing them into more regular sequences, the model effectively enhances prediction accuracy. Their approach successfully captures both short-term fluctuations and long-term trends in pollutant concentrations, demonstrating improved forecasting performance over traditional single-stage prediction methods. This study highlights the benefits of combining advanced pre-processing with predictive modelling for air quality analysis. AlJanabi and Alkaim (2019) [12] proposed a novel collaborative approach called DRFLLS (Dynamic Rough Fuzzy Logic and Least Squares) for estimating missing values in datasets. The method integrates rough set theory, fuzzy logic, and least squares estimation to handle uncertainty and improve the accuracy of missing data imputation. Through extensive experiments, the DRFLLS model demonstrated superior performance compared to existing imputation techniques, offering a reliable solution for pre-processing incomplete datasets in various application domains.
Maleki et.al. [2019] [13] have presented an artificial neural network (ANN) algorithm to predict hourly criteria air pollutant concentrations and two air quality indices, AQI and air quality health index (AQHI), for Ahvaz, Iran, over the period 2009 and 2010. Ahvaz is known to be one of the most polluted cities in the world, mainly owing to dust storms. The applied algorithm involved nine factors in the input stage (five meteorological parameters, pollutant concentrations 3 and 6 h in advance, time, and date), 30 neurons in the hidden phase, and finally one output in the last level. This study proves that ANN has applicability to cities such as Ahvaz to forecast air quality with the purpose of preventing health effects. Alimissis A et.al. 2018 [14] have used interpolation methods for point estimations in the field of air pollution modelling to estimate the pollutant concentrations in metropolitan Athens in Greece. The main objective of the study is to evaluate two interpolation methodologies, Artificial Neural Networks and Multiple Linear Regression, using data from a real urban air quality monitoring network. The results for five regulated air pollutants (Nitrogen dioxide, Nitrogen monoxide, Ozone, Carbon monoxide and Sulphur dioxide) are compared through the use of a set of correlation and difference statistical measures and residuals distribution.

Xian Liu et.al. (2022) [15] have used ML methodology to process satellite data to obtain ground-level concentrations of atmospheric pollutants, pollution source apportionment, and spatial distribution modelling of water pollutants. In the past decade, the unprecedented accumulation of data, the development of high-performance computing power, and the rise of diverse machine learning (ML) methods provide new opportunities for environmental pollution research. The paper has discussed the challenges to use ML methods to revolutionize environmental science and its problem-solving scenarios. Kelly, Reff, and Gantt (2018) [16] developed a predictive method to estimate PM2.5 concentrations resulting from compliance with National Ambient Air Quality Standards (NAAQS). Their approach integrates regulatory modelling, emissions inventories, and air quality simulations to assess the potential impacts of policy-driven emission reductions on future PM2.5 levels. The study provides a valuable framework for evaluating how regulatory compliance can influence air quality outcomes, offering policymakers a practical tool for air quality management and planning. Anurag, Yagnavalkburra, and Selvaraj (2019) [17] proposed an Air Quality Index (AQI) prediction model using a feature-based weighted XGBoost approach that incorporates meteorological parameters. By applying feature selection and assigning optimized weights to key influencing factors, the model enhances prediction accuracy and computational efficiency. The study demonstrated that integrating meteorological data significantly improves the AQI forecasting performance, highlighting the effectiveness of ensemble learning methods like XGBoost in air quality prediction tasks.
Balaraman et al. (2021) [18] applied the LSTM model to predict PM2.5 concentrations in urban areas. The study leveraged the LSTM's capability to capture temporal dependencies in air quality time series data, leading to accurate short-term forecasting of particulate matter levels. The model demonstrated strong predictive performance and outperformed traditional statistical methods, emphasizing the potential of deep learning, particularly LSTM networks, in enhancing urban air pollution monitoring and early warning systems. Sharma et al. (2020) [19] developed a deep learning-based air quality forecasting model that combines Convolutional Neural Networks (CNN) and LSTM networks to predict suspended particulate matter (PM concentrations). The hybrid CNN-LSTM model effectively captures both spatial features and temporal dependencies in air quality data,

resulting in improved forecasting accuracy. The study demonstrated the superiority of this deep learning approach over traditional models, offering a robust solution for real-time air quality prediction and management.

Ayturan, Ayturan, and Altun (2018) [20] presented a comprehensive review of air pollution modelling techniques using deep learning methods. The study examined various deep learning architectures, including ANNs, CNNs, and RNNs, highlighting their growing application in air quality prediction. The authors emphasized that deep learning models offer significant advantages over traditional statistical and machine learning approaches, particularly in handling complex, nonlinear, and large-scale environmental datasets. The review concluded that deep learning holds strong potential for advancing air pollution forecasting and improving environmental decision-making.

Sachdeva et al. (2023) [21] proposed an AQI prediction model that integrates meteorological parameters and pollutant concentration data to enhance forecasting accuracy. The study focused on developing a data-driven approach that combines real-time atmospheric conditions with pollutant levels to predict the AQI more reliably. Their findings demonstrated that incorporating meteorological factors significantly improves the model's precision, supporting more effective air quality monitoring and early warning systems in urban environments. El-Latef, Zaki, and Issa (2018) [22] developed a Traffic Air Quality Health Index (TAQHI) to assess the health impacts of traffic-related air pollution on a selected street in Alexandria. The study measured concentrations of key traffic pollutants and combined them into a health-focused index that reflects the potential risk to public health. Their analysis emphasized the significant influence of vehicular emissions on local air quality and highlighted the need for targeted traffic management and pollution control strategies in urban areas to mitigate health risks.

3. Proposed Model

The proposed model is based on the short-term accuracy of Random Forest and the LSTM sequence modelling. The RF and LSTM algorithms are combined in this hybrid model in a two-stage pattern. The model used air quality and meteorological data from Kaggle.com. Feature engineering and normalization were performed to enhance model learning, followed by splitting the dataset into training, validation, and testing subsets. The model implements Random Forest regression to capture nonlinear relationships and assess feature importance in AQI prediction using ensemble-based decision trees. A Long Short-Term Memory (LSTM) network is developed to model temporal dependencies and sequential patterns in AQI time-series data, enabling effective short-term and long-term forecasting. The model performance is evaluated using statistical metrics such as RMSE and MAE. A comparative analysis of Random Forest and LSTM models is conducted to assess prediction accuracy, robustness, and suitability for AQI forecasting. The results help identify the most effective model for real-time air quality prediction and early warning systems. This section explains the detailed operation of the proposed hybrid RF-LSTM model.

3.1 Pre-process Dataset

Data pre-processing is a crucial step in the machine learning pipeline. It works on raw data to make it clean, consistent, and ready for model training. The dataset used was in '. csv' format and has details about the air pollution levels in various Indian cities, having parameters: AQI, PM2.5, PM10, SO_2, NO_2 etc., regarding the air pollution level at various times of day. Statistical methods were used to handle missing values, outliers, noisy observations, duplicate values, and invalid formats to improve data quality and reliability. All pre-processing steps were applied consistently across the dataset, ensuring the robustness and reproducibility of the proposed prediction model. The pre-processed dataset was divided into training, validation, and testing sets for model development and evaluation using a time-aware strategy.

3.2 First Stage of Hybrid RF-LSTM Model: Random Forest for Feature Extraction

The Random Forest is applied to analyse the feature importance and extract the relevant features from the dataset. Feature extraction is used to prepare data for either classification or regression. It is the process of transforming raw data into a set of measurable, informative features that can be used to train a model. In traditional ML, this is often done manually, but in deep learning, it's mostly automated and learned from the data. The feature extraction is used for dimensionality reduction and understanding which input influences prediction. The Random forest technique has been used to extract most relevant features to make predictions about the target variable. The RF process is applied not to extract some new features but to rank the existing factors to find out the most driving one that can predict the target output. RF is an ensemble learning method that uses multiple decision trees to make predictions. Random Forest calculates feature importance based on how much a feature contributes to reducing impurity (like Gini Impurity or Entropy) across all trees [23]. Gini Impurity is used to measure how "pure" a node is in a decision tree. It helps to split data by quantifying the impurity of a node's class distribution. A low gini impurity indicates the node to be homogeneous whereas a high gini impurity specifies heterogeneous node i.e. the classes are more evenly distributed. The following equation (1) gives the formula to calculate the gini impurity.

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

Where pi is the proportion of class i instances in the node.

The Gini Impurity for a node is calculated as (2):

$$\text{Feature Importance (f)} = \sum_{t \in trees} \sum_{n} \in nodes\ (t,f) \frac{Nn}{N} \Delta i(n) \tag{2}$$

Where:

- t: a tree in the forest
- n: a node where feature f is used for splitting
- Nn: number of samples at node n
- N: total number of samples
- Δi(n): decrease in impurity at node n

Decrease in Impurity Formula (3) is given below:

$$\Delta i(n) = i(n) - pL\ .\ i(nL) - pR\ .\ i(nR) \tag{3}$$

Where:

- i(n): Gini impurity at the parent node
- i(nL),i(nR): Gini impurity of the left and right child nodes
- $pL = \frac{nL}{Nn}$, $pR = \frac{nR}{Nn}$: Proportions of samples that go left or right

This process measures how much cleaner the data got when split on that feature. If a feature is used at a node that greatly reduces impurity, it's considered important. If it's used in many such nodes across the forest, its total importance increases. Features used in root nodes (with many samples) usually carry more weight. Random Forest aggregates all these decreases (across all trees and all nodes where the feature is used) and normalizes the values to give a final importance score between 0 and 1 for each feature.

In the proposed hybrid RF-LSTM model, the first stage involves the Random Forest model analysing the raw AQI data in combination with key pollutants such as CO, PM2.5, NO_2 and SO_2.

The RF model, using AQI and CO features, captures short-term AQI fluctuations in response to CO spikes. This type of RF model can be applied in traffic management systems. Using AQI in combination with CO as input features for a Random Forest model is particularly useful in urban traffic zones, where vehicle emissions are a major source of pollution.

The Random Forest, when combining AQI with PM2.5 levels offers the most accurate and reliable results. PM2.5, or fine particulate matter, is one of the primary contributors to poor air quality, especially in areas affected by heavy pollution from construction activities, dust storms, or wildfire smoke. In such environments, the relationship between PM2.5 and AQI is both strong and direct, allowing the Random Forest model to effectively learn and map this correlation. Among all single-feature combinations tested, the (AQI , PM2.5) pairing consistently delivers the highest prediction accuracy, making it an ideal choice for modelling air quality in regions where particulate matter is a dominant pollutant.

Combining AQI with NO_2 as input features in a Random Forest model is particularly effective in high-traffic areas where combustion-based emissions are prevalent. NO_2, a by-product of vehicle exhaust and industrial activity, has a well-established impact on air quality, contributing to both short-term pollution spikes and longer-term atmospheric degradation. The Random Forest model is adept at capturing these complex patterns, making it well-suited to model the nuanced influence of NO_2 on AQI. This pairing offers a good balance, effectively capturing both immediate fluctuations and broader trends in air pollution, resulting in reliable AQI predictions in urban and industrial environments.

Using AQI in combination with SO_2 as input features in a Random Forest model is most relevant in industrial regions and areas surrounding coal-fired power plants, where SO_2 emissions are more prominent. The model can detect changes in AQI driven by elevated SO_2 levels, especially during industrial activity peaks. However, the correlation between SO_2 and AQI tends to be weaker compared to other pollutants like PM2.5 or NO_2, which can limit the model's predictive accuracy. Additionally, in cleaner or residential zones where SO_2 levels are generally low or stable, this feature provides minimal contribution to AQI forecasting, making its utility highly location-dependent.

In conclusion, among all the pollutants analysed, PM2.5 emerges as the most reliable and influential predictor of AQI when using Random Forest, due to its strong and direct correlation with air quality fluctuations. While Random Forest models are effective in capturing non-linear relationships between pollutants and AQI, integrating them into a hybrid modelling approach—such as combining RF with LSTM—can significantly enhance predictive accuracy. This hybrid method leverages the strength of Random Forest in feature extraction and the capability of LSTM to learn temporal patterns, offering a more robust forecasting framework. Additionally, it is important to consider city-specific modelling, as the relevance and impact of each pollutant can vary widely based on local geography, climate, industrial activity, and traffic patterns. Tailoring models to individual cities ensure more accurate and context-aware AQI predictions.

3.3 Second Stage of Hybrid RF-LSTM Model: LSTM for Capturing Time Series Dependencies

The extracted features from the Random Forest algorithm are fed in the LSTM model to learn temporal dependencies. LSTM networks are a type of Recurrent Neural Network (RNN) designed to handle long-range dependencies in sequence data. Unlike standard feedforward neural networks, LSTMs have feedback connections, allowing them to exploit temporal dependencies across sequences of data. To find temporal dependencies using features with the LSTM algorithm, the LSTM model learns patterns in sequential data such as time series by processing inputs in order and using internal memory (cell state) to capture information across time steps. It has memory cells and gates to control the flow of information. Fig. 1 shows the model for the working of the LSTM algorithm.

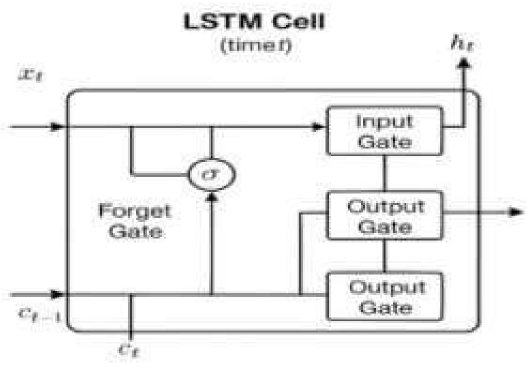


Figure 1: Working of LSTM Algorithm

Inputs:

- x_t: current input
- c_{t-1}: previous cell state

Gates:

- **Forget Gate** σ: decides what to forget
- **Input Gate** i_t: decides what new information to store
- **Output Gate** o_t: decides what to output

Outputs:

- c_t: new cell state
- h_t: new hidden state

The LSTM algorithm works like a tiny decision maker. Here is the list of steps which are followed in LSTM for making prediction using features in AQI database -

1. Look at current input and past memory and decide which old information to erase.
2. Decides which new parts of the input to store.
3. Mixes the remembered past and the new important information and updates its internal memory (called the "cell state").
4. Decides what part of memory to share with the next step or final answer.
5. Repeat the steps using the new input and updated memory.
6. After all the time steps, use the last result (or the full sequence of outputs) to make a prediction or classification.

LSTM can take multiple features per time step and learn how they interact over time. Using an LSTM model to make predictions on AQI data is a powerful approach because LSTM is well-suited for time series data and can capture temporal patterns, trends, and dependencies in pollution levels over time. AQI data is time-dependent, having past values influence future values. LSTM is good at remembering patterns like daily or seasonal pollution cycles, levels of PM2.5, PM10 or SO2 etc. The next section shows the result of the LSTM model for the feature extracted AQI data of five Indian cities and reports the results of the RF-LSTM hybrid model on AQI data.

4. Result Analysis and Discussions

In this section, the performance of the proposed two-stage hybrid RF–LSTM model used for AQI prediction is evaluated. The analysis of the experimental data used city-wise datasets for the years 2020–2023. The predictive performance of the hybrid RF–LSTM framework was compared with that of the standalone Random Forest (RF), and LSTM. The assessment emphasizes the model's adequacy in describing the nonlinear relationships between pollutants and the temporal dependencies of AQI changes.

The hybrid RF–LSTM model prediction outperformed the individual models in forecasting AQI levels. The first stage of the Random Forest model was able to appropriately capture such complex nonlinear patterns of AQI dependence on important pollutants such as CO, PM2.5, NO_2, and SO_2. RF is robust to missing values, reduces overfitting using an ensemble learning method, and provides a mechanism for characterizing feature importance, making it appropriate for pre-processing schemes and preliminary prediction tasks. Then, the RF output is passed to a sequence model based on LSTM that learns long-term dependencies in the data. This two-step architecture allows the system to utilize the advantages of both models: RF principal nonlinear feature interactions and noise reduction, and LSTM, which represents time trends and dynamic variations.

The dataset used in this study was obtained from Kaggle and consists of comprehensive Air Quality Index (AQI) records collected over a four-year period from 2020 to 2023. To evaluate the robustness and generalization capability of the model, data from five distinct Indian cities—Lucknow, Chandigarh, Delhi, Gurugram, and Jaipur—were selected. These cities exhibit diverse pollution characteristics influenced by population density, industrial activities, vehicular emissions, and seasonal meteorological variations.

To ensure uniformity and improve model convergence during training, the AQI values were normalized and scaled within the range of 0–5. Feature scaling reduces the numerical imbalance among variables, preventing any single parameter from dominating the learning process and thereby stabilizing the overall model performance. The primary input features considered in this study included major air pollutants, such as carbon monoxide (CO), particulate matter (PM2.5), nitrogen dioxide (NO_2), and sulphur dioxide (SO_2), as these pollutants significantly influence AQI levels. The dataset was organized in a time-series format to facilitate sequential modelling and enable effective learning of temporal patterns in air quality variations.

The implementation follows a cascaded architecture beginning with Random Forest regression. In this stage, nonlinear regression was performed to generate preliminary AQI predictions using the four selected pollutant features. These variables were chosen based on their strong correlation with the AQI (Pearson correlation coefficient > 0.7) and their significance within India's National Air Quality Monitoring Programme (NAMP). Collectively, these pollutants represent major emission sources, including vehicular traffic (CO and NO_2), industrial and fuel combustion (SO_2), and airborne particulate matter (PM2.5).

For each city, the dataset was partitioned into an 80:20 train-test split, while maintaining temporal continuity. The most recent 20% of the data were reserved for out-of-sample testing to simulate real-time forecasting conditions. The Random Forest regressor was trained exclusively on the 80% training subset, with hyperparameters optimized through a grid search. The tuning range included 200–500 trees, a maximum tree depth between 10 and 20, and a minimum of 2–10 samples per leaf node. Optimization was performed using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics under 5-fold cross-validation.

After training, the RF model generated denoised preliminary predictions for the held-out test set and produced feature importance rankings, with PM2.5 typically contributing the highest share (approximately 40–50%). These refined predictions were then structured sequentially and passed to the LSTM stage for further temporal modeling and prediction refinement.

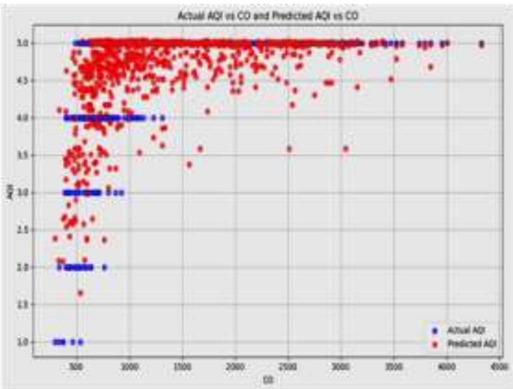


Figure 2 (a): Performance Analysis: AQI vs. CO at first stage

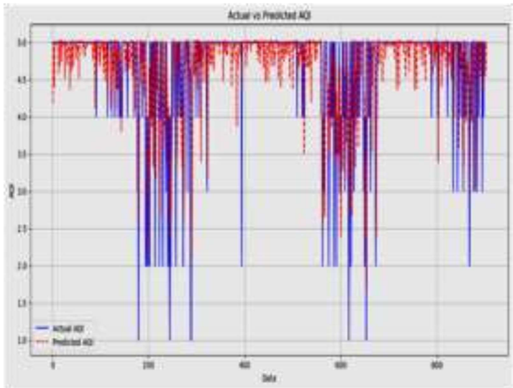


Figure 2 (b): Performance Analysis at second stage

Figure 2(a) illustrates the relationship between the AQI and CO values provided as inputs to the Random Forest model. As shown in Figure 2(a), AQI and CO exhibited a strong positive correlation, indicating that variations in the CO concentration significantly influenced the AQI. The plotted results demonstrate that the Random Forest model effectively captures this nonlinear association between AQI and CO, validating the feature selection for the first stage of the proposed framework.

Similarly, Figure 2(b) presents the comparison between actual AQI values and the predicted AQI values generated in the second stage of the hybrid model. The graph shows that the predicted values closely follow the actual AQI trend over time, indicating that the LSTM model successfully captures temporal dependencies and sequence-level variations. The close alignment between actual and predicted curves demonstrates the effectiveness of the RF–LSTM hybrid architecture in improving forecasting accuracy.

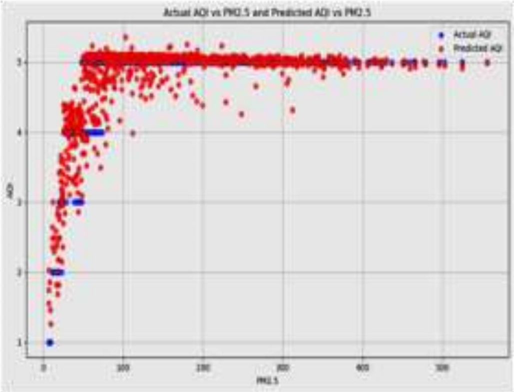


Figure 3 (a): Performance Analysis: AQI vs. PM2.5 at first stage

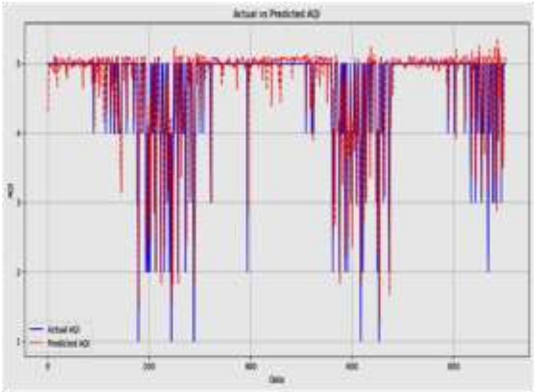


Figure 3 (b): Performance Analysis at second stage

Similarly, the hybrid framework was evaluated using AQI and PM2.5 as input features. In this case, the first stage employs the Random Forest (RF) model to capture the nonlinear relationship between AQI and PM2.5 and to generate a denoised sequence of preliminary AQI predictions. These refined predictions were subsequently passed to the second-stage LSTM model, which further processed the sequential data to produce more accurate and temporally consistent AQI forecasts. The corresponding results are shown in Figures 3(a) and 3(b).

In Figures 4(a) and 4(b), the analysis of AQI in relation to SO_2 is presented, following the same two-stage hybrid approach. Similarly, Figures 5(a) and 5(b) depict the results for AQI and NO_2, demonstrating the effectiveness of the cascaded RF–LSTM architecture across different pollutant inputs. Across all cases, the results consistently indicate that the proposed hybrid model outperforms the standalone approaches, providing improved prediction accuracy and better alignment between the actual and predicted AQI values.

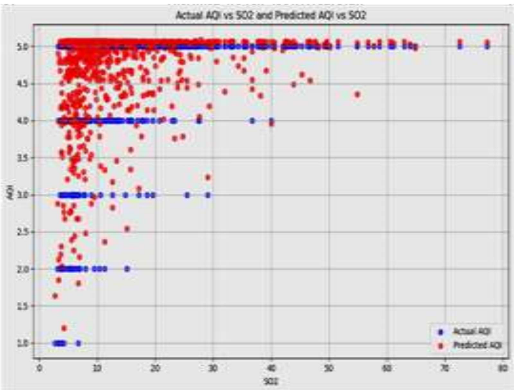


Figure 4 (a): Performance Analysis: AQI vs. SO_2 at first stage

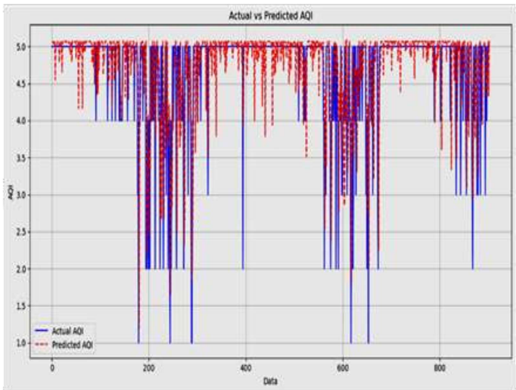
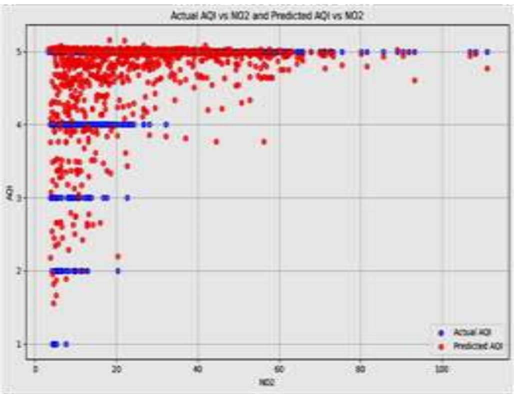


Figure 4 (b): Performance Analysis at second stage



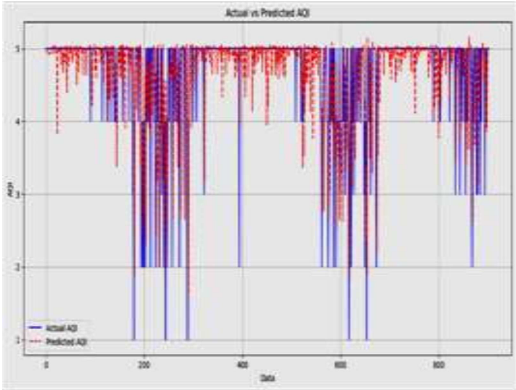


Figure 5 (a): Performance Analysis: AQI vs. NO_2 at first stage

Figure 5 (b): Performance Analysis at second stage

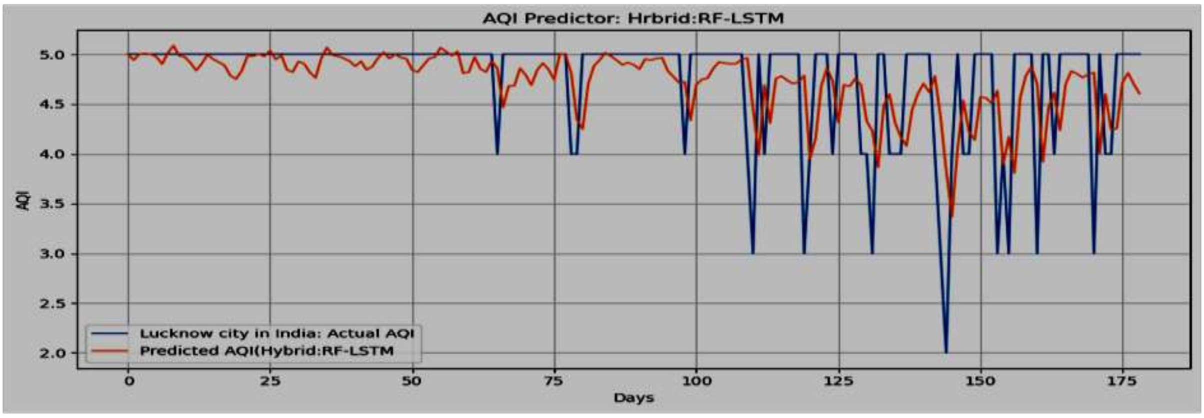


Figure 6: AQI prediction for Lucknow city

Figure 6 shows the predicted AQI values for Lucknow obtained using the proposed hybrid RF–LSTM model. The graphical comparison between the actual and predicted AQI values demonstrated a strong agreement, with the predicted curve closely following the observed AQI trend over time. The model captures both short-term fluctuations and broader temporal variations, indicating effective learning of pollutant interactions and sequential dependencies. Overall, the predicted values were approximately 85% close to the actual AQI measurements for Lucknow, reflecting the robustness of the forecasting framework.

Quantitatively, the performance metrics further validated the accuracy of the model. The Root Mean Square Error (RMSE) was 0.154756, indicating a low deviation between the predicted and actual AQI values, while the Mean Absolute Error (MAE) was 0.083212, demonstrating a minimal average prediction error. These relatively small error values confirm that the proposed hybrid model achieves high precision and reliable predictive performance for AQI forecasting in Lucknow, India.

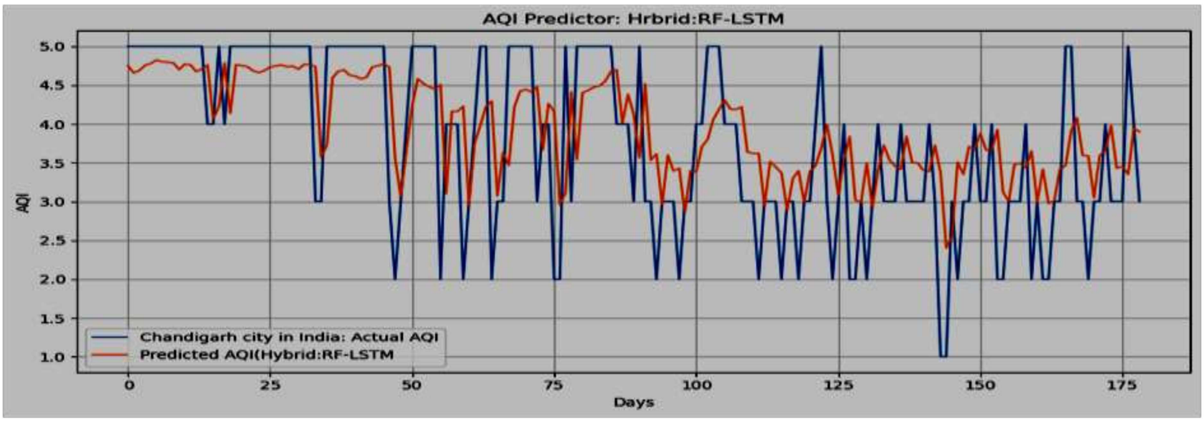


Figure 7: AQI prediction for Chandigarh city

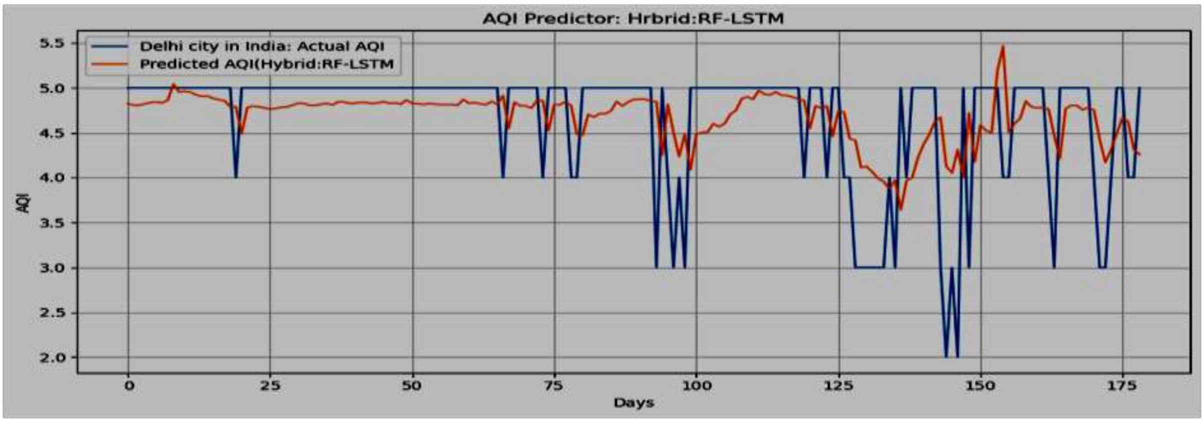


Figure 8: AQI prediction for Delhi city

Figure 7 illustrates the predictive performance of the proposed hybrid RF–LSTM model using the Chandigarh dataset. Similar to the results observed for Lucknow, the predicted AQI values closely followed the actual AQI trend, demonstrating the model's ability to generalize across different urban environments. For Chandigarh, the predicted values were approximately 75% close to the observed AQI levels. The performance metrics further support this observation, with a Root Mean Square Error (RMSE) of 0.241346 and a Mean Absolute Error (MAE) of 0.179433. Although the error values were slightly higher than those for Lucknow, they remained within an acceptable range, indicating stable and reliable model performance.

Figures 8, 9, and 10 show the corresponding results for Delhi, Gurugram, and Jaipur, respectively. In each case, the predicted AQI curve aligned closely with the actual AQI values, capturing both peak pollution episodes and low-pollution intervals. Across all city datasets, the proposed hybrid model consistently achieved an overall prediction accuracy of approximately 85 %, while maintaining low RMSE and MAE values. These results confirm the robustness and generalization capability of the model across diverse pollution profiles.

The accuracy and consistency of the predicted AQI values demonstrate that the proposed framework is suitable for real-world air quality forecasting applications. Reliable AQI predictions can support environmental monitoring agencies and policymakers in implementing timely preventive measures, thereby contributing to more effective pollution control strategies in major urban centres.

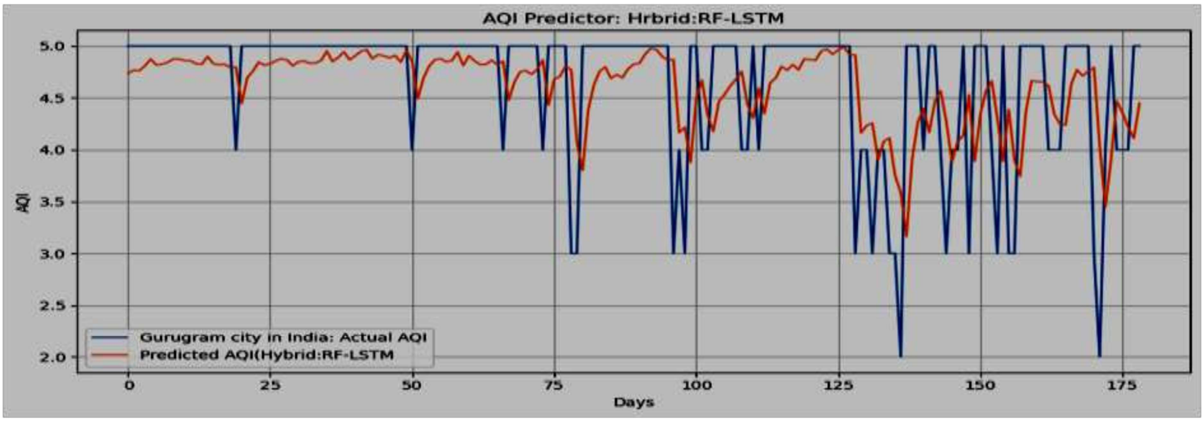


Figure 9: AQI prediction for Gurugram city

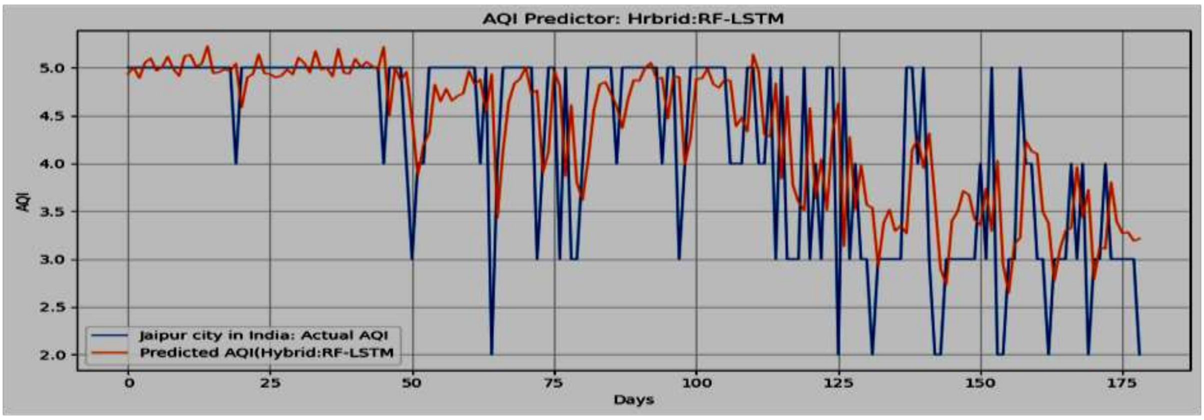


Figure 10: AQI prediction for Jaipur city

City	RF_RMSE	LSTM- RMSE	Hybrid- RMSE
Lucknow	0.456655	0.541426	0.154756
Chandigarh	1.122889	0.975295	0.241346
Delhi	0.529038	0.6139	0.153836
Gurugram	0.526762	0.634765	0.147025
Jaipur	0.825043	0.79024	0.210758

Table 1: RMSE values of standalone RF, standalone LSTM and the proposed hybrid model

City	RF_MAE	LSTM- MAE	Hybrid- MAE
Lucknow	0.188726	0.383749	0.083212
Chandigarh	0.84879	0.757345	0.179433
Delhi	0.232561	0.417332	0.091815
Gurugram	0.254816	0.434653	0.100536
Jaipur	0.530768	0.599814	0.146776

Table 2: MAE values of standalone RF, standalone LSTM and the proposed hybrid model

The performance of the proposed hybrid RF–LSTM model was further evaluated through a comparative analysis with standalone Random Forest (RF) and LSTM models. Tables 1 and 2 summarize the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values obtained for all three models across the five selected city datasets. This comparative assessment enables a systematic evaluation of each model's predictive capability and helps identify the most suitable approach for AQI forecasting in the future.

The results clearly indicate that the proposed hybrid model consistently outperforms both the standalone RF and LSTM models across all cities. In each case, the hybrid framework achieved lower RMSE and MAE values, demonstrating reduced prediction error and improved alignment between actual and predicted AQI values. Although the standalone RF model effectively captures the nonlinear relationships among pollutants, it lacks temporal modelling capability. Conversely, the standalone LSTM model captures time-dependent patterns but may be sensitive to the noisy input features. By integrating the strengths of both approaches, the hybrid RF–LSTM model successfully combined nonlinear feature interaction with temporal sequence learning, resulting in enhanced overall prediction accuracy.

The superior performance of the hybrid model confirms its robustness, stability, and generalization capability across diverse urban pollution profiles. The improved accuracy contributes to higher reliability in AQI forecasting, making the model suitable for practical environmental monitoring. Considering the increasing governmental emphasis on environmental governance and the persistent deterioration of air quality in major cities, the proposed RF–LSTM framework can be used as an effective decision-support tool. Accurate and timely AQI predictions can assist policymakers and regulatory agencies in implementing proactive measures to mitigate air pollution and reduce its adverse health and environmental impacts.

5. Conclusions

In the current environmental conditions, the early prediction of air quality plays a crucial role in air monitoring systems, weather forecasting, and the protection of public health. Accurate AQI forecasting allows authorities to issue timely warnings, reduce exposure to harmful pollutants, and implement preventive measures to manage environmental risks. However, predicting air quality is a complex and challenging task because of the dynamic and nonlinear nature of pollutant concentrations, which vary significantly across time and geographic locations.

Moreover, the predicted AQI values can be effectively integrated into intelligent traffic control and urban planning systems. For instance, prior knowledge of high pollution levels in specific areas can help divert traffic from heavily polluted zones, thereby reducing emissions and improving public safety.

To address these challenges, the proposed hybrid RF-LSTM model combines the strengths of two powerful approaches: RF for short-term pattern-based prediction and LSTM networks for capturing temporal dependencies in time-series data. The model operates in a two-stage architecture, where RF is used in the first stage to extract features and reduce noise, and the output is passed to the LSTM in the second stage for sequential learning.

This hybrid approach leverages the robustness and interpretability of RF, along with the deep learning capabilities of LSTM. As a result, it delivered more accurate AQI predictions than either model alone. The experimental results across multiple Indian cities demonstrate that the proposed hybrid model consistently outperforms the standalone RF and LSTM models in terms of RMSE and MAE, making it a promising tool for real-time air quality forecasting and public health management.

References

1. I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Frontiers in Public Health,* vol. 8, 2020, doi: 10.3389/fpubh.2020.00014.
2. G. I. Drewil and R. J. Al-Bahadilli, "Air Pollution Prediction using LSTM Deep Learning and Metaheuristics Algorithms," *Measurement: Sensors*, 29 Oct. 2022, doi: 10.1016/j.measen.2022.100546.
3. E. Sharma, R. C. Deo, R. Prasad, A. V. Parisi, and N. Raj, "Deep Air Quality Forecasts: Suspended Particulate Matter Modeling with Convolutional Neural and Long Short-Term Memory Networks," *IEEE Access*, vol. 8, pp. 209503–209516, 2020, doi: 10.1109/ACCESS.2020.3039002.
4. S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting," *Energies*, vol. 13, no. 2, pp. 1–23, 2020, doi: 10.3390/en13020391.
5. H. Li et al., "Analyzing the Impact of Heating Emissions on Air Quality Index based on Principal Component Regression," *Journal of Cleaner Production*, vol. 171, pp. 1577–1592, 2017, doi: 10.1016/j.jclepro.2017.10.106.
6. V.-D. Le, T.-C. Bui, and S.-K. Cha, "Spatiotemporal Deep Learning Model for Citywide Air Pollution Interpolation and Prediction," *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Busan, Korea (South), 2020, pp. 19–26, doi: 10.1109/BigComp48618.2020.00-99.
7. C. Zhan, S. Li, J. Li, Y. Guo, Q. Wen, and W. S. Wen, "Prediction of Air Quality in Major Cities of China by Deep Learning," *Proceedings of the 16th International Conference on Computing Intelligent Security (CIS)*, 2020, pp. 68–72, doi: 10.1109/CIS52066.2020.00023.
8. R. Janarthanan, P. Partheeban, K. Somasundaram, and P. N. Elamparithi, "A Deep Learning Approach for Prediction of Air Quality Index in a Metropolitan City," Sustainable Cities and Society, 2021, doi: 10.1016/j.scs.2021.102720.
9. F. Xiao, M. Yang, H. Fan, G. Fan, and M. A. A. Al-qaness, "An Improved Deep Learning Model for Predicting Daily PM2.5 Concentration," Scientific Reports, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-77757-w.
10. W. Mao, W. Wang, L. Jiao, S. Zhao, and A. Liu, "Modeling Air Quality Prediction using A Deep Learning Approach: Method Optimization and Evaluation," Sustainable Cities and Society, 2020, Article 102567, doi: 10.1016/j.scs.2020.102567.
11. W. Sun and C. Huang, "A Hybrid Air Pollutant Concentration Prediction Model Combining Secondary Decomposition and Sequence Reconstruction," Environment Pollution, vol. 266, p. 115216, 2020, doi: 10.1016/j.envpol.2020.115216.
12. S. Al-Janabi and A. F. Alkaim, "A Nifty Collaborative Analysis to Predicting a Novel Tool (DRFLLS) for Missing Values Estimation," Soft Computing, vol. 24, no. 1, pp. 555–569, 2019, doi: 10.1007/s00500-019-03972-x.
13. H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. T. Birgani, and M. Rahmati, "Air Pollution Prediction by using an Artificial Neural Network Model," Clean Technologies and Environmental Policy, vol. 21, pp. 1341–1352, 2019, doi: 10.1007/s10098-019-01709-w.
14. A. Alimissis, K. Philippopoulos, C. G. Tzanis, and D. Deligiorgi, "Spatial estimation of urban air pollution with the use of artificial neural network models," Atmospheric Environment, vol. 191, pp. 205–213, 2018.
15. X. Liu, D. Lu, A. Zhang, Q. Liu, and G. Jiang, "Data-Driven Machine Learning in Environmental Pollution: Gains and Problems," Environmental Science & Technology, vol. 56, no. 4, 2022.
16. J. T. Kelly, A. Reff, and B. Gantt, "A Method to Predict PM2.5 Resulting from Compliance with National Ambient Air Quality Standards," Atmospheric Environment, vol. 162, pp. 169–187, 2018, doi: 10.1016/j.atmosenv.2017.05.009.

17. N. V. Anurag, S. Yagnavalkburra, and S. Selvaraj, “Air Quality Index Prediction with Meteorological Data Using Feature Based Weighted XGBoost,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 1, May 2019.
18. S. Balaraman, P. Partheeban, P. N. Elamparithi, and M. S. Manimozhi, “Application of LSTM Models in Predicting Particulate Matter (PM2.5) Levels for Urban Area,” Journal of Engineering Research, Nov. 2021, doi: 10.36909/jer.11781.
19. E. Sharma, R. C. Deo, R. Prasad, A. V. Parisi, and N. Raj, “Deep Air Quality Forecasts: Suspended Particulate Matter Modeling with Convolutional Neural and Long Short-Term Memory Networks,” IEEE Access, vol. 8, pp. 209503–209516, 2020, doi: 10.1109/ACCESS.2020.3039002.
20. Y. A. Ayturan, Z. C. Ayturan, and H. O. Altun, “Air Pollution Modelling with Deep Learning: A Review,” International Journal of Environment Pollution and Environment Modelling, vol. 1, no. 3, pp. 58–62, 2018.
21. S. Sachdeva, R. Kaur, Kimmi, and S. Kharb, “Meteorological AQI and Pollutants Concentration-Based AQI Predictor,” International Journal of Environmental Science and Technology, vol. 21, no. 11 (Special Issue), Nov. 2023.
22. E. M. A. El-Latef, G. R. Zaki, and A. I. Issa, “Traffic Air Quality Health Index in a Selected Street, Alexandria,” Journal of High Institute of Public Health, vol. 48, no. 2, pp. 67–76, 2018.
23. W. Zhu, J. Feng, and Y. Lin, “Using Gini-Index for Feature Selection in Text Categorization,” 3rd International Conference on Information, Business and Education Technology (ICIBET 2014), pp. 76–80, doi: 10.2991/icibet-14.2014.22.