International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I10P212

Volume 71 Issue 10, 133-145, October 2023

Original Article

A Hybrid Spatial-Temporal Approach to Pollution Forecasting with Dynamic Updates and Time Series Analysis

Snehlata Beriwal[1], A. John[2], Kavita[3], Avneesh Kumar[4]

[1,2,3,4]*School of Computer Science and Engineering, Galgotias University, Greater Noida, India.*

[1]*Corresponding Author : goelsneh@yahoo.com*

Received: 24 June 2023 Revised: 30 August 2023 Accepted: 15 September 2023 Published: 03 October 2023

Abstract *- The rapidly deteriorating air quality across the globe has increasingly become a challenge with far-reaching consequences. Hence, accurate air quality prediction, monitoring, and forecasting have become an intrinsic part of managing our living environment. Such advanced predictions and timely interventions thereof can aid in minimizing any untoward threats to our health and quality of life. The primary aim of this research is to enable effective time and location-based predictions and forecasting of air quality and pollution levels. To that end, a hybrid approach based on indexing and time series techniques has been proposed in this study. This hybrid approach is based on the D-Tree-based indexing method, SARIMA, Bidirectional LSTM, and the Pearson correlation. The D-Tree-based indexing method is used to manage current and previous data. The SARIMA is used to predict and forecast the future status of pollution particles based on current data as well as seasonal trends. The Bidirectional LSTM is utilized for Time Series Forecasting using current and past data managed by the D-tree indexing method. The Pearson correlation is used for measuring and managing the mean of two predicted outputs from inputs received concurrently from live environments. During implementation, live pollution data was concurrently collected from the different location-centric pollution sensing devices and updated using the indexing method. This current data was then appended to the previous year's data to improve accuracy further. Thus, using both past and live data, forecasts were made for the next 6, 12 hours, and 1, 2, and 3 days, respectively. Prediction accuracy was evaluated using various metrics such as accuracy, Air Quality Index (I), Mean Square Root (MSR), Mean Absolute Error (MAE), and correlation coefficient (R). The predicted results were found to produce higher accuracy (97.6%) across different time lags compared to other predominant forecasting methods. This approach, therefore, has been found to concurrently update the status of pollutant particles in dynamic environments effectively and consistently.*

Keywords *- Spatial and temporal data, Hybrid model, Pollution forecasting, SARIMA, LSTM, D-Tree.*

1. Introduction

Data predictions that consider both location and time to predict future states of events have evolved into a sophisticated area of research that allows for empowered decision-making that incorporates both time and geo-specific references. Such a spatial-temporal approach to future data prediction allows for effective determining and correlating the nearest similarities. The best of the contemporary spatiotemporal applications have found extensive implementation in areas like traffic prediction, advertising, and pollution forecasting, among others. Of these, pollution forecasting as an essential research area focused on spatiotemporal prediction is especially relevant since the Air Quality Index (AQI) is inherently dynamic in nature and undergoes substantial variations based on the geographical location due to multiple contributing factors such as population density, growth rate of vehicular usage, industrial presence, density of green cover and ever-changing weather data. Typically, the level of air pollution is gauged by the presence of a combination of particles and gases that are released into the atmosphere. More often than not, atmospheric air pollution is a direct result of human activities, though it can also stem from natural events such as forest fires and volcanic eruptions. Some man-made sources contributing to atmospheric pollution include industrial processes, the burning of fuels, transportation, and cooking [1]. Both natural and man-made processes contribute to creating CO2, NO2, SO2, CO, and sulphate that subsequently get trapped in the atmosphere. Although researchers have introduced multiple techniques for forecasting pollution levels and AQI, these prediction techniques often fail to consider time-based values. In this work, a comprehensive approach has been adopted that considers a host of parameters that are location-related (static),

temporal or time-related (dynamic), and dynamic criteria for forecasting the result. The surroundings that we live in are often steeped in particles such as SO2, NO2, CO2, CO, NO, PM10, and PM2.5. These pollutants and their concentrations recurrently change due to varying conditions such as the presence of industrial activities, vehicular movement, weather, population density, etc.

To tackle this, researchers have proposed various techniques for predicting AQI ranges. The prediction can be typically classified as basic and advanced prediction models. The basic techniques include the Lagrangian [3], Gaussian [2], Eulerian [3], Box [4], and Dense gas techniques [5]. The advanced techniques include Meteorological Forecasting, Chemical Transport Dispersion, Atmospheric chemical transport, Online and offline air quality, Computer programs for dispersion, Three-Dimensional, Data assimilation, and Hybrid models [6]. Of all these models, the intelligence models, along with the hybrid models, have been observed to have improved accuracy since they can be used for dynamic forecasting at any given point in time. In this proposed technique, a novel hybrid model for predicting and forecasting pollution particles using the D-Tree-based indexing method, SARIMA, Bidirectional LSTM, and Pearson correlation has been proposed. The following points can summarize the novelty of the proposed hybrid method:

- In this work, both live and historical data have been used to predict and forecast the future status of the pollution particles PM2.5, PM10, NO2, NH3, and CO, with different time lags such as 6 hours, 12 hours, 1 day, 2 days and 3 days.
- The proposed method has been dynamically updated using the indexing method and has been found to produce better forecasting accuracy in terms of time intervals.
- Due to ongoing data updates and the time series forecasting of the data, the prediction and forecasting results are seen to undergo continuous change.

The subsequent parts of this paper have been organized as follows. Section 2 discusses the current and related body of work with respect to various pollution forecasting methods, various works undertaken, and components considered for predicting pollution levels. Section 3 delves into the working methods of forecasting pollution particles. Section 4 provides details of the implementation and the performance analysis of the proposed work compared to other existing works and is followed by the conclusion.

2. Related work

The body of existing work related to this study can be divided into two distinct categories, viz., the various techniques proposed for predicting and forecasting pollution levels and the pollution collection systems proposed for live environments. Both have been discussed further in this section.

2.1. Techniques for Prediction and Forecasting

The authors of [1] have presented an overview of air pollution forecasting and its various techniques. In this article, the authors delved into various forecasting techniques and the issues identified with respect to these existing works. The authors of [6] propounded a spatial-temporal approach for forecasting pollution levels using a hybrid method in which the authors have used inverse distance and an ordinary kriging method to make multi-site air pollution predictions. The RMSE and MAE metrics were used to evaluate the performance. On the other hand, the authors of [7] proposed a hybrid model for spatial-temporal forecasting of PM 2.5 using graph-based CNN and LSTM. This method utilized the recall, the false alarm rate, and the correlation coefficient (R2) for performance evaluations. This work produced better results than the previous methods, such as MLR, FNN, and LSTM. The researcher of [8] similarly presented a model for a spatial-temporal based air quality prediction system using LSTM and a multi-index supervised learning algorithm. In this work, a one-year dataset was used during the implementation, and based on the implementation results, PM2.5, CO, NO2, O3, and SO3 were predicted. R2, MAE, and RMSE were used and compared to the SVM and ARIMA models for performance evaluation purposes. The authors of [9] alternately proposed a long-term pollution forecasting model using deep learning and statistical methods. In this work, the author collected data from Kolkata, India, and the auto AR, the SARIMA, and the Holt-winter deep learning method were used for long-term predictions. Using this method, seasonal forecasting was predicted, and RMSE and MAE were used for the evaluation.

The reference of [10] alternately proposed the LSTM neural to forecasting PM2.5 based on spatial-temporal data. In this work, the features were extracted using the LSTM model, and records from 35 air monitoring stations were used for the implementation. This proposed work was compared with LR, RF, SVM, and ARMA models. The [11] hybrid model with the help of data decomposition and different machine learning models. In this model, wavelet decomposition and low-frequency approximate sequences were proposed using LISTM and ARIMA for the sequence of predictions. In this model, RMSE, MAE, and R2 metrics were used for air quality prediction. The [12] proposed a hybrid approach to address different air pollutants such as O3, CO, SO2, and NO2. The GT-LSTM model was used for training and monitoring data for the next 24 hours. The experiment was conducted using data from Jan 2016 to 31 December 2019. The RMSE, MAE, R2, and NRMSE were the set of parameters used to evaluate the proposed model. It was observed that this model produced better accuracy and stability compared to various pollution predictions.

The [13] present air quality forecasting using a hybrid deep learning approach. Using this approach, the authors proposed forecasting of PM2.5 with the help of spatial and temporal correlation features. In this approach, 1D-CNNS and Bi-LSTM

models were used for forecasting. The proposed work experiment was conducted using a Beijing dataset and evaluated with the help of RMSE and MAE. The authors of [14] proposed a domain-specific deep learning model for forecasting air pollution levels. In this model, air pollution over the long term was forecasted in China and the United Kingdom. This model proposed three novelties, such as a strong statistical relationship between PM2.5 and PM10, usage of historical features for temporal correlation, and combined historical certainty features.

The proposed work consisted of two types of predictions: one-time predictions and recursive predictions. The [16] forecasting over a period of three days using a neural network. The prediction and forecasting were performed using 3 to 15 training datasets from the past. The predicted output parameters, such as SO2, PM10, and CO, were presented on the website for real reference. The [17] forecasting weather and pollution levels simultaneously in Macedonia. In this work, date-wise and time-wise, PM10 was forecasted at different intervals. The [18] outdoor prediction and monitoring for healthy living and breathing. This work configured a PWP system using MQ07-CO, SDS021, NO2-B43F, and O3. With the help of these units, various pollutant levels were collected for a period of 90 days. The experiment was conducted using various machine learning models, and AQI levels were monitored.

The [19] forecasting model for preventive measures using ANN, ARIMA, TBATS, and FTS of Malaysia for the year 2017. In this model, fuzzy time series was the best forecasting model compared to other methods. The RMSE and MAPE were the two parameters used to evaluate the forecasting. The [20] PSO-SVM hybrid model for forecasting short-term pollution concentrations. Using this work, various factors that influenced pollution factors were considered for forecasting, using data from Beijing. The predicted forecasting variables were computed with regression analysis. The [21] combined system for forecasting using fuzzy theory and aggregation weight. This method used the Cuckoo search algorithm to find the optimal weight for aggregation. In this work, data pre-processing was performed with the help of a complimentary ensemble empirical mode, and individual forecasting was performed for BPNN and ELM. After that, combined forecasting was undertaken and evaluated using various metrics.

The authors of [22] undertook the forecasting of time series data in the Caribbean cities. The authors estimated missing values, selected the best values, and forecasted data for 24 hours. The proposed work was performed using SARIMA and evaluated using RMSE. The authors of [23] similarly proposed time series air pollution forecasting using the classification and regression tree (CART) model. Using this model, time series data was predicted a day before in the cities of Bulgaria. The CART model was built using the ARIMA model. The CART model was well fitted with 90% accuracy daily. Interestingly, the authors of [24] presented a wireless sensor network for monitoring air pollution. In this work, gas sensors were used to collect the data, and pollution was monitored and forecasted based on the collected data.

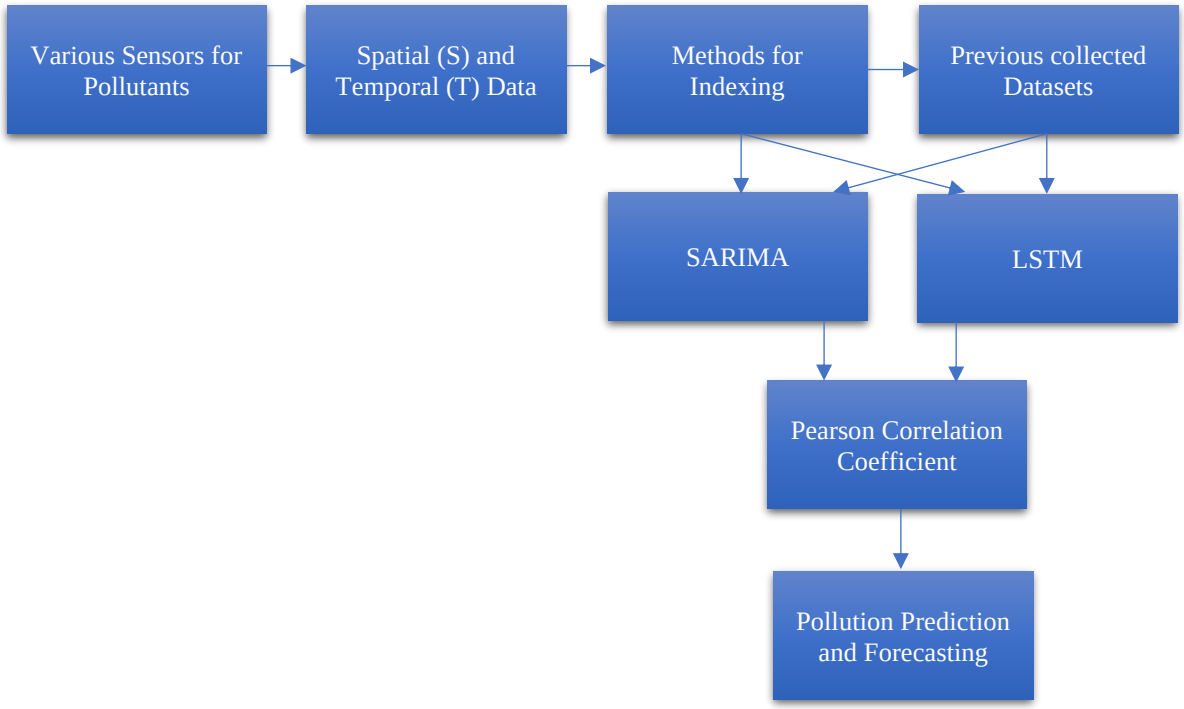

Fig. 1 Structure of the proposed work

Table 1. Dataset locations and interval Information

S. No	Locations	Duration	Collected Particles	Time Interval
1	Jawaharlal Nehru Stadium, Delhi.	1 January 2019 to June 2020	NO2, CO, NH3, Ozone, PM2.5 and PM10.	30 Minutes
2	Alipur, Delhi.	1 January 2019 to June 2020	NO2, CO, NH3, Ozone, PM2.5 and PM10.	30 Minutes
3	Dodhi Road, Delhi	1 January 2019 to June 2020	NO2, CO, NH3, Ozone, PM2.5 and PM10.	30 Minutes
4	Noida	1 January 2019 to June 2020	NO2, CO, NH3, Ozone, PM2.5 and PM10.	30 Minutes

Table 2. Features used for pollution forecasting

S. No	Features	Type
1	Longitude, Latitude, Altitude	Spatial
2	Time Interval, Temperature, Wind Direction, Wind Speed and Rainfall	Temporal

Table 3. Air quality ranges

AQI Category	Good /Low	Satisfactory	Moderate	Poor	Very poor/High	Severe/ Hazardous
Ranges	0-50	51-100	101-200	201-300	301-400	401-500

3. Materials and Methods

3.1. Problem Definition

The collected sequence input from different locations L1 = x {x1, x2, · · ·, xT} and past data L1 = P {x1, x2, · · ·, xT} where xt L1(X, P), T denotes the time interval, x denotes the live data collected from various sources from locations L1, L2..., Ln. The set of features is denoted by x1, x2…xn with respect to time t. For example, L1=x{x1} and L1=P{x1} set prediction parameters mapped with each of the features, such as wind speed (ws), rainfall rate (rf), etc. The targeted mapped features or AQI indexing rates Y= {y1, y2, · · ·, yT−1}, are mapped using the inputs and outputs using various techniques.

3.2. Materials and Features

Data was collected from the metropolitan Pollution Control Board of India from different locations in Delhi. For implementation, data was collected from four different locations such as {L1, L2, ..., L4}. Detailed information about the dataset and the corresponding attributes has been presented in Table 1. The different features and parameters used for the prediction and forecasting have been presented in Table 2.

The prediction features are classified into temporal and spatial. Longitude and latitude information are considered as the spatial features, and temperature, timing, direction of air movements and rainfall are considered temporal features. The spatial and temporal features are considered the timing and location-based vehicle movements and other data, such as rainfall in particular locations. In our proposed method, vehicle and tree densities are considered for pollution levels. The various considered features for forecasting pollution levels have been presented in Table 3.

3.3. Structure of the Hybrid Model

The proposed method consists of a physical layer, a communication layer, and a data analytics and intelligence layer. This structure has been illustrated in Figure 1. The data collection and the spatial-temporal data management are performed using the physical layer. The data changes, and updates are performed using the communication layer. Processing and decision-making are performed using the intelligence layer. The model utilizes SARIMA, LSTM, and the Pearson correlation coefficient for data processing.

3.3.1. Physical Layer

The physical layer collects all the required basic information, such as the ozone, nitrogen dioxide, carbon monoxide levels, and all other information, using the various sensors listed in Table 1. All information related to the spatial and temporal data is collected using these sensors.

3.3.2. Communication Layer

The communication layer transfers the data from the physical devices to the storage locations. In this work, the MQTT protocol has been used for communication and data transfer. This collected information can be stored in a cloud environment or any temporary storage location. Using this temporary storage location, the required data is then processed. Thus, the location information or spatial data and temporal or time-based data are collected using the physical and communication layers, and past data is stored in indexing.

3.3.3. Indexing

The data is arranged using an indexing method in the cloud environment or the temporary storage location for data

processing. In this case, the D-Tree-based indexing method has been used for temporary storage or data updates. Therefore, the indexing structure is based on the D-Tree-based storage. A multi-structure index method is used, which utilizes the D-Tree (D-Compose Tree), TB*-Tree (Trajectory Bundle Tree), NT-Tree (Network-limited TPR-Tree), and the hash table. The D-Tree is used to store the station information and current predicted data. The geographical information in the spatial data represents road information, spatial location, location time, etc. Such information is directly noted and managed by D-Tree. The multi-structure TB*-Tree is used to manage the dynamic pollution particles. The NT-Tree is used to find the pollution particles' present and future status from the pollution collecting devices. Finally, the hash table is used to update the data continuously. Since the hash tables are interconnected to the bottom of the node, it is updated quite easily.

3.3.4. Data Analytics and Intelligence Layer

The intelligence and the data analysis layer are used for processing, analysis, and decision-making. The intelligence layer processes data using the location information and the corresponding coordinates collected from dynamic environments. The timestamps and other Table 2 parameters are also dynamically inserted for the data processing. SARIMA, LSTM, and the Pearson correlation coefficient are used based on the location and timing data for the actual data processing and decision-making. The AQI data is also inserted for benchmarking and is presented in Table 3. The processed data is compared to the AQI data, and based on that, decision-making is performed.

3.3.5. SARIMA

The Seasonal autoregressive integrated moving average (SARIMA) method is used to trend the data based on seasonal data derived from frequent season-al effects and the time series data. The mathematical function of SARIMA has been represented in Equation 1.

$$Y = Y(p,d,q)(P,D,Q)m \quad (1)$$

In Equation 1, Y denotes output, p denotes the trend auto aggregate, d - trend difference, q - trend moving average, P, D, and Q denote the seasonable elements of auto aggregate, difference, and moving average, respectively, and m denotes the seasonable timestamp. Based on these parameters, the current and updated data from the past are trended based on the seasonal predictions in different timestamps.

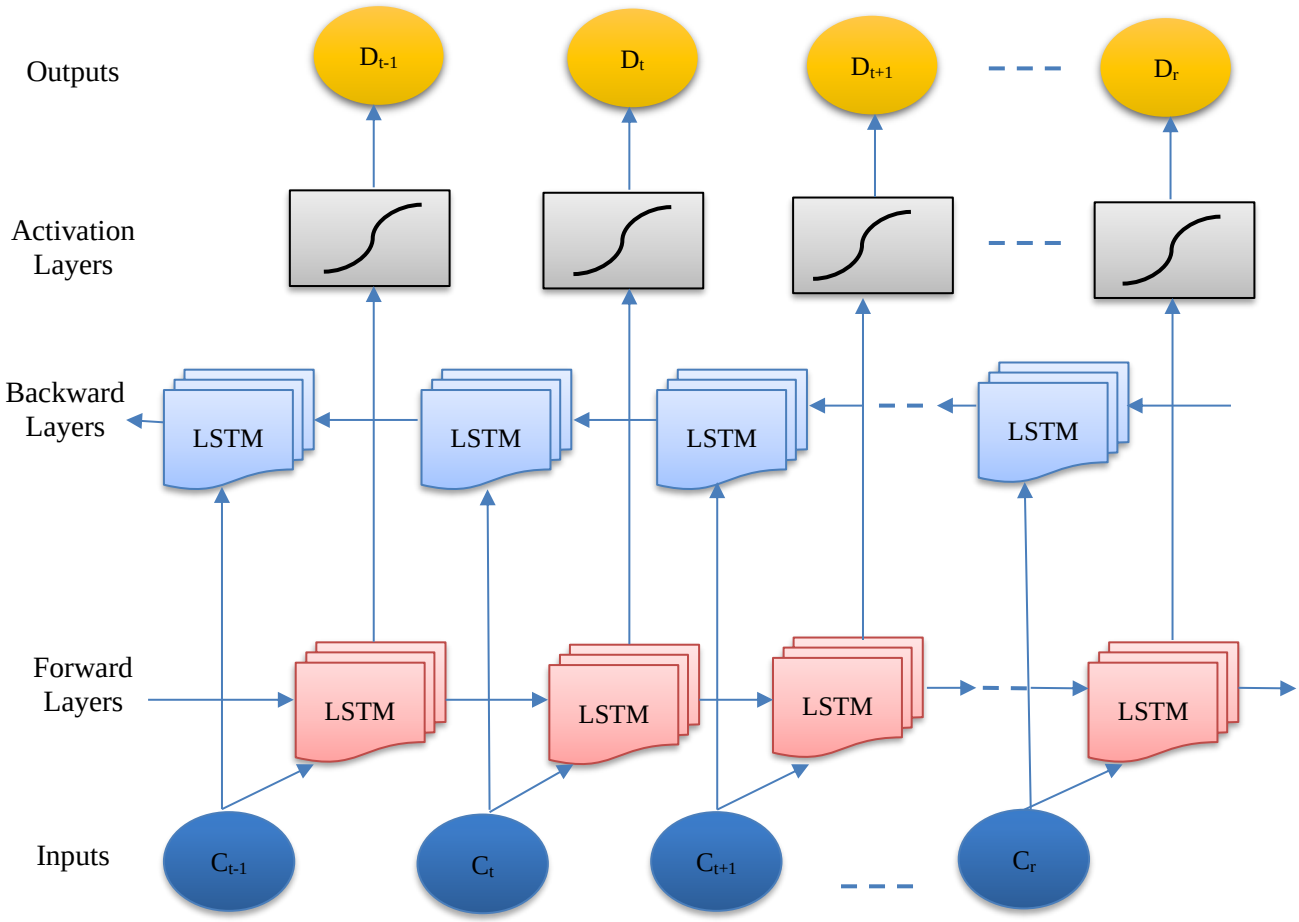

Fig. 2 Structure of a Bi-Directional LSTM

Bi-directional LSTM (Bi-LSTM): The LSTM is classified into different variants based on the situation, such as LSTM classic, LSTM peephole, Bi-directional LSTM, Multiplicative LSTM, and attention-based LSTM [25, 26]. Bidirectional LSTM is used to make sequences of information transferred in both directions, from past to future and future to past, since the input flow of a bi-LSTM is in both directions, such as forward and backboard. Using the forward direction helps manage the past predicted data and the current predicted data. In the structure of the proposed work, the current data is managed by using an indexing method, as is the past data, which is also provided as input. Thus, the past and current predicted data help easily manage future data. The structure of a bi-directional LSTM has been illustrated in Figure 2. The equations (2- 6) represent input representation, activation function, forward and backward layer representation, u the updating layer, and output representations.

The prediction features are classified into temporal and spatial. Longitude and latitude information are considered as the spatial features and temperature, timing, direction of air movements and rainfall are considered temporal features. The spatial and temporal features are considered the timing and location-based vehicle movements and other data, such as rainfall in particular locations. In our proposed method, vehicle and tree densities are considered for pollution levels. The various considered features for forecasting pollution levels have been presented in Table 3.

The input is represented using different variables, features, and time constraints. The input of the proposed work has been defined in the problem model and is also represented in Equations 2.

$$\{C_1, C_2, \ldots, C_T, D_1, D_2 \ldots D_T\} \tag{2}$$

Where C and D denote non-linear variables and denotes time. The activation function of the current state Bi-directional LSTM has been shown in Equation 3.

$$g_t = tan\, g\,(xg_{t-1} + xc_t) \tag{3}$$

Where denotes current state, represents the weight, denotes previous state while denotes various inputs with respect to time. The forward and backward activation functions have been shown in Equation 4.

$$D^T = h\big(x_d\big[b^{\rightarrow(t)}, b^{\leftarrow(t)}\big] + a_d\big) \tag{4}$$

The update, forget, and output gates of the bi-directional LSTM are represented by Equations (5-7).

$$E_v = \delta + x_v[b^{t-1}, c^t] + a_v \tag{5}$$

$$E_e = \delta + x_e[b^{t-1}, c^t] + a_e \tag{6}$$

$$E_o = \delta + x_o[b^{t-1}, c^t] + a_o \tag{7}$$

Hence, the differing status of bi-directional LSTM current inputs is updated using the function, and past data is updated using the function, whereas the output data is managed using the function. The current and past data have been represented in the Equations 3 and 4.

Pearson correlation: The Pearson correlation is also known as a bi-variant correlation and is used to find the relationship between two linear data sets. The result of the variance is always between -1 and 1. In this work, the Pearson correlation has been used to find the correlation between SARIMA and B-LSTM. Equations 8 to 9 have been used to find the relationship between the two from the outputs. The pair of random variables are defined as (A, B) in Equation (8).

$$\rho_{A,B} = \frac{Cov(A,B)}{\delta_A \delta_B} \tag{8}$$

The Pearson correlation's expression and mean of variance have been represented by Equation (9).

$$Cov(A,B) = F[(A - \mu A)(B - \mu B)] \tag{9}$$

In Equation 10, μA denotes mean of A , and μBdenotes mean of B. The function for the correlation coefficient of expression and mean has been represented using Equation (10).

$$\rho_{C,D} = \frac{F[(A-\mu B)(B-\mu B)]}{\delta_C \delta_D} \tag{10}$$

Thus, with the help of different non-linear inputs, two linear outputs are received, and any correlation between them is predicted using the Pearson correlation.

3.4. Working of the Proposed Hybrid Model

In this work, two methods, SARIMA and Bi-LSTM, have been used to process the inputs. Various input features are considered using these two models, as mentioned in Table 3. These inputs are processed and collected from the various physical devices and locations, as mentioned in Tables 1 and 2. This data is collected in 30-minute time intervals and is stored and arranged using the D-Tree-based indexing method. This index method is used to manage both past and present data at the time. The arranged data is then stored in a temporary location. From this temporary location, the data is processed using SARIMA and Bi-LSTM. Using these methods, non-linear data is processed, and linear data is produced in the specified time interval. The received discrete linear data is then processed using the Pearson correlation. Algorithm 1 illustrates the overall processing of the proposed work. The pollutant particles are forecasted using 30-minute time intervals. Using forecasting (F) of F SARIMA and F Bi-LSTM, the new correlation (F Pearson) is predicted at different time intervals. In this work, the pollution particles are updated at 30-minute intervals and predicted using the models.

Algorithm 1: Hybrid Model

Input: Current and Past pollution particles, different features, and time interval
Models: SARIMA, Bi-LSTM, and Pearson correlation
Output: Hybrid forecasted values

1. Begin
2. Interval ← T
3. While (Interval ≤ 30 min) do
4. Indexing ← Temp (Present and Past Data)
5. F SARIMA ← SARIMA (Temp (Present and Past Data))
6. F Bi-LSTM ← LSTM (Temp (Present and Past Data))
7. F Pearson ← { F_{SARIMA}, $F_{Bi\text{-}LSTM}$ }
8. F interval ← Max Time Interval
9. Prediction ← { Interval of Prediction}
10. hr (30min) ← 30min +1
11. While (hr≤ 12) do
12. $Hybrid^{h}_{rt}$ ← Updation (30 min Interval)
13. End

4. Implementations

This section presents the implementation details of the proposed hybrid model using different intervals. The Python programming language is used for prediction and forecasting. The data collection devices and the study area or locations of the proposed work have been mentioned in Table 1 and Table 2.

In this work, two types of data have been used: live data and previously collected datasets mentioned in Table 2. The different dynamic features have been mentioned in Table 3. The data received in 30-minute time intervals, and the data arranged using indexing methods have been described in Tables 2 and 3.

The indexing method arranges the past and present data in temporary memories. From the temporary locations, the received data and previous past data at the specified time intervals are taken for processing. Apart from these inputs, several dynamic features are also considered for the processing, as presented in Table 2. The forecasted results are then mapped with the In-dia Air Quality indexing (AQI) data, and based on that, the comparisons and decision-making are performed. The AQI data for Delhi has been shown in Table 3, and with the help of this data, future air pollutant data is forecasted.

4.1. Metrics for Forecasting

In this work, the following metrics have been used to evaluate the prediction and forecasting: Air Quality Index (I), mean square root (MSR), mean absolute error (MAE), correlation coefficient (R), normalizing average (NA), and Index agreements (IA) and have been represented in the Equations 11-16.

$$\mathrm{I} = \frac{I_{High} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low} \quad (11)$$

$$MSE = 1/n + \sum_{i=1}^{N}(F_i - A_i)^2 \quad (12)$$

$$MAE = 1/n + \sum_{i=1}^{N}|F_i - A_i \quad (13)$$

$$\mathrm{R} = \frac{(F_i - F)(A_i - A)}{\sigma_A C \sigma_F} \quad (14)$$

$$NA = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (15)$$

$$IA = 1 - \sum_{t=1}^{n}(F_i - A_i)^2 \,/\, \sum_{t=1}^{n}(|F_i - A| + |A_i - A|)^2 \quad (16)$$

Forecasting Model for 6 hours to 3 days: The collected dataset samples are modeled, and the samples have been summarized in the annexure. In particular, from location 1 (Jawaharlal Nehru Stadium, Delhi), old data (8256 data entries), as well as the collected live data (1440 data entries), have been summarized in the annexure. Based on these two data sets, data is forecasted every 6 hours, 8 hours, 12 hours, 1 day, and 2 days. The forecasted data has been summarized in Tables 3 and 4 with the help of different supporting polluting particles. The sample of the Jawaharlal Nehru Stadium lo-cation and the supporting factors used for forecasting data for 1 year have been shown in Figures 6 to 11.

As illustrated by the supporting figures of 6 to 11, different sets of data forecasting were undertaken. This work predicted data at specific intervals starting from 6 hours to 3 days, as summarized in Table 5. In this table, four different predictions have been summarized, with the predicted values changing based on the parameters mentioned in Table 2 and Figures 6 to 11. For example, in these four locations, the prediction range of the morning pollution was moderate because the number of moving objects and human migration in these areas normally decreased during morning hours compared to other times.

Table 4. Prediction and forecasting of AQI indexing ranges

Locations	CO	SO2	NO2	O3	PM10	PM2.5
Jawaharlal Nehru Stadium, Delhi.	10	5	7	6	73	137
Alipur, Delhi.	7	3	13	2	78	151
Dodhi Road, Delhi	0	5	30	0	134	159
Noida	24	7	0	12	230	157

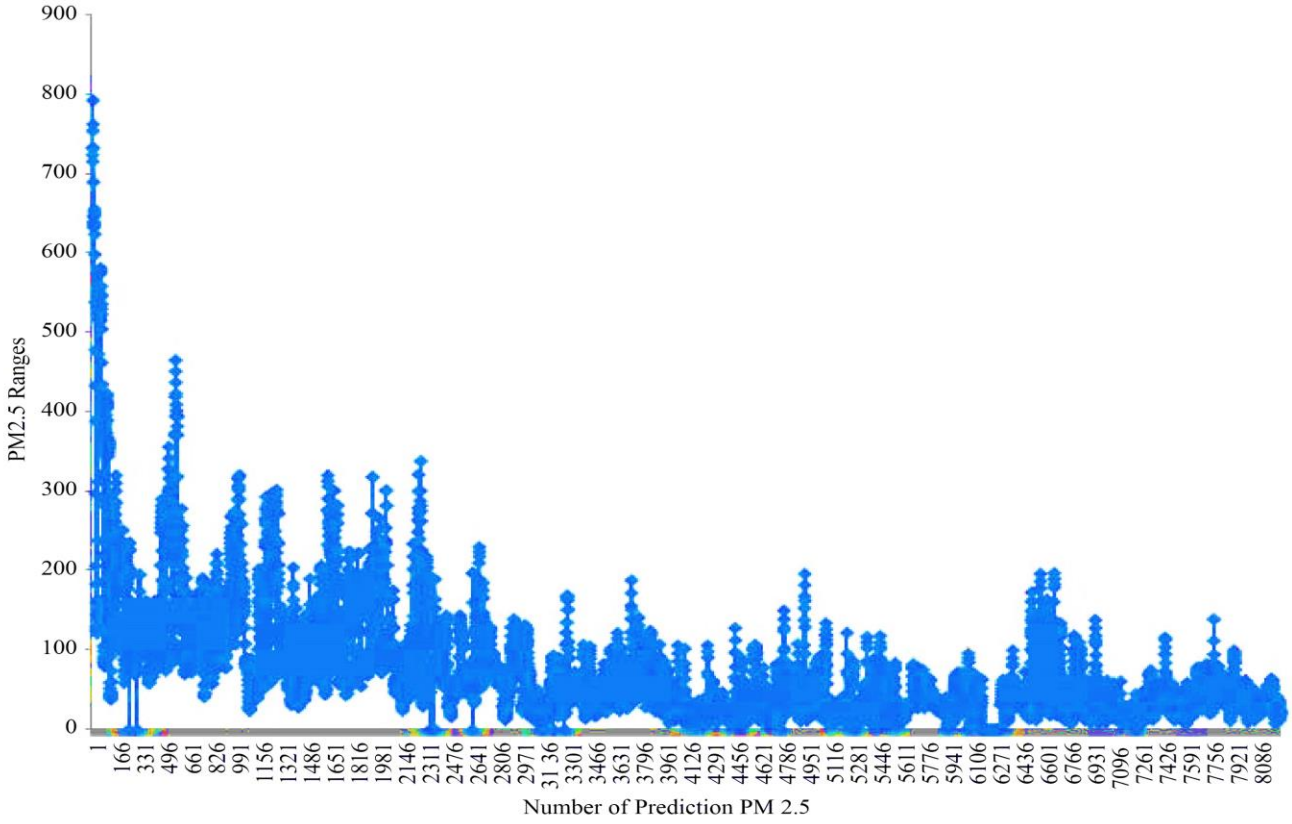

Fig. 3 Forecasting of PM 2.5 in jawaharlal nehru stadium

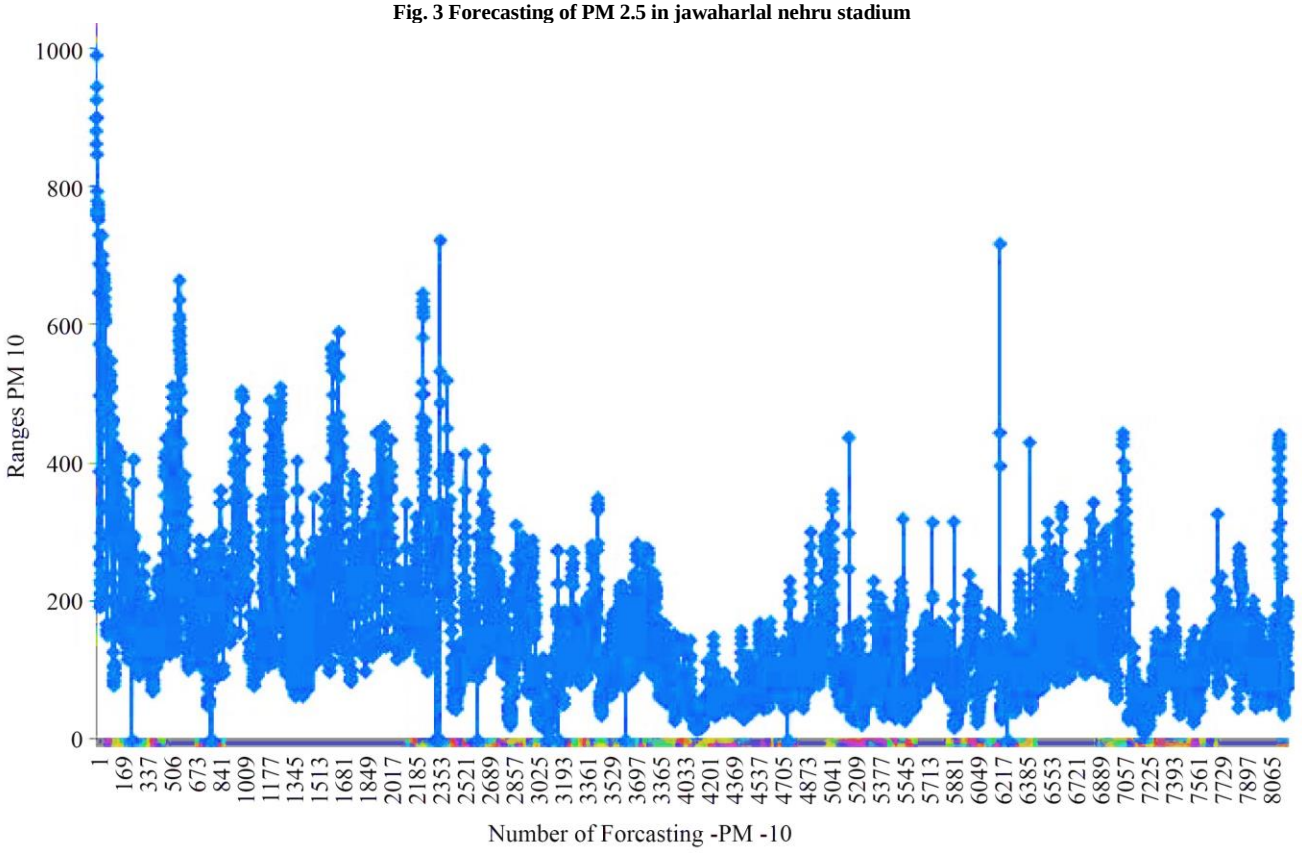

Fig. 4 Forecasting of PM 10 in jawaharlal nehru stadium

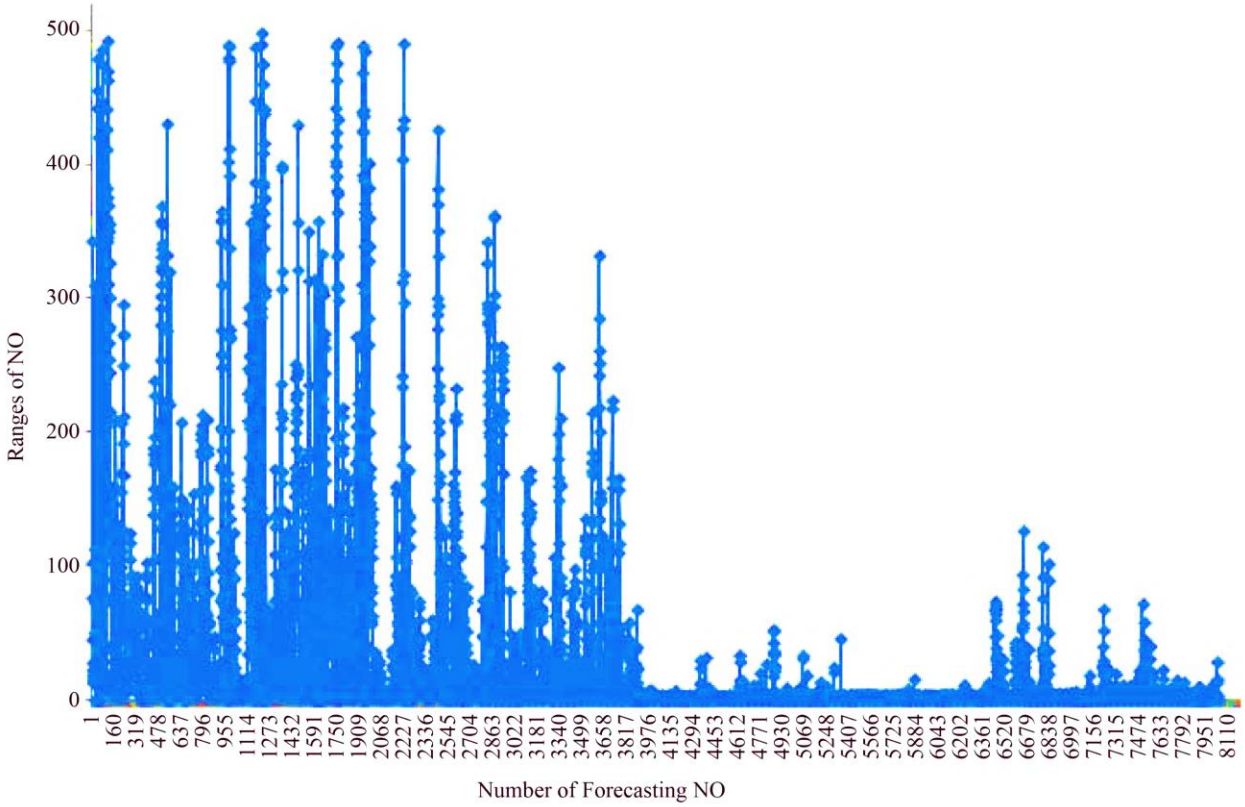

Fig. 5 Forecasting of NO in jawaharlal nehru stadium

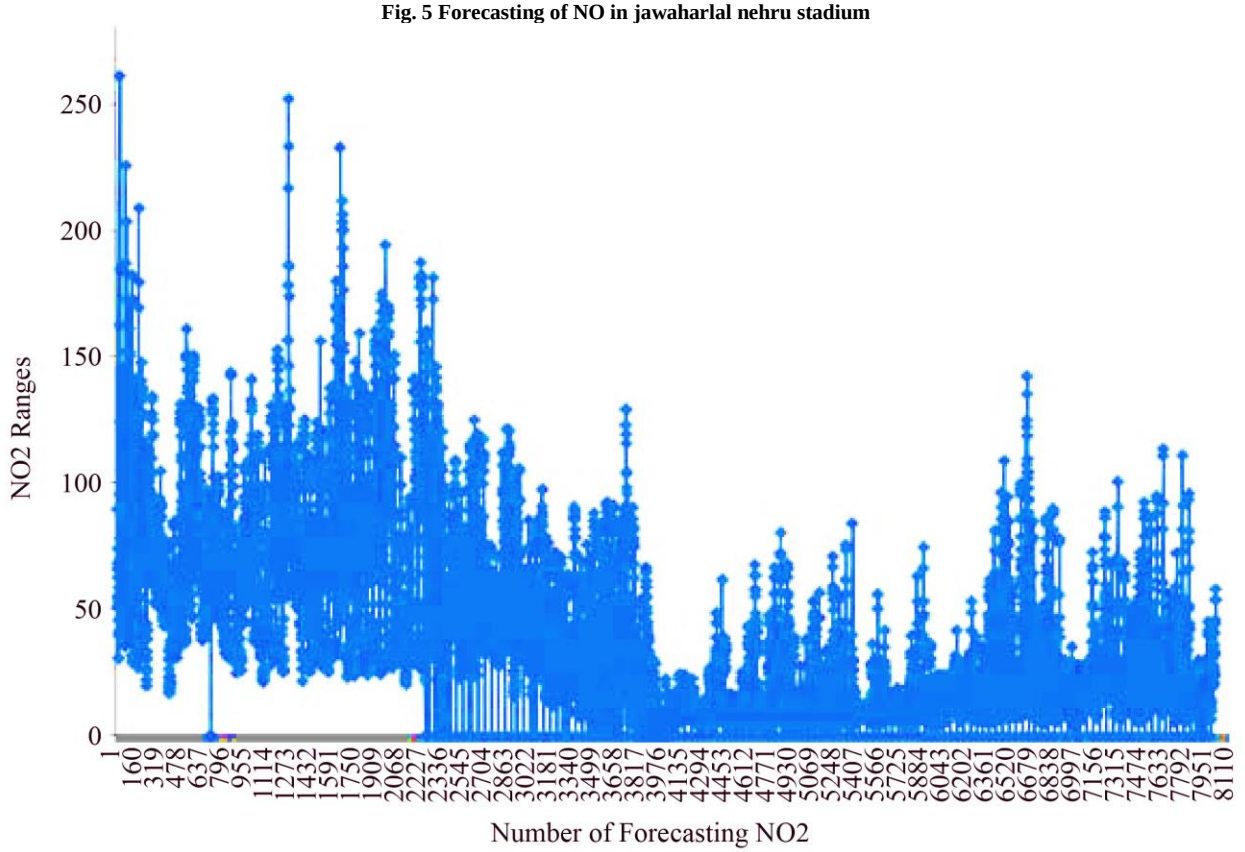

Fig. 6 Forecasting of NO2 in jawaharlal nehru stadium

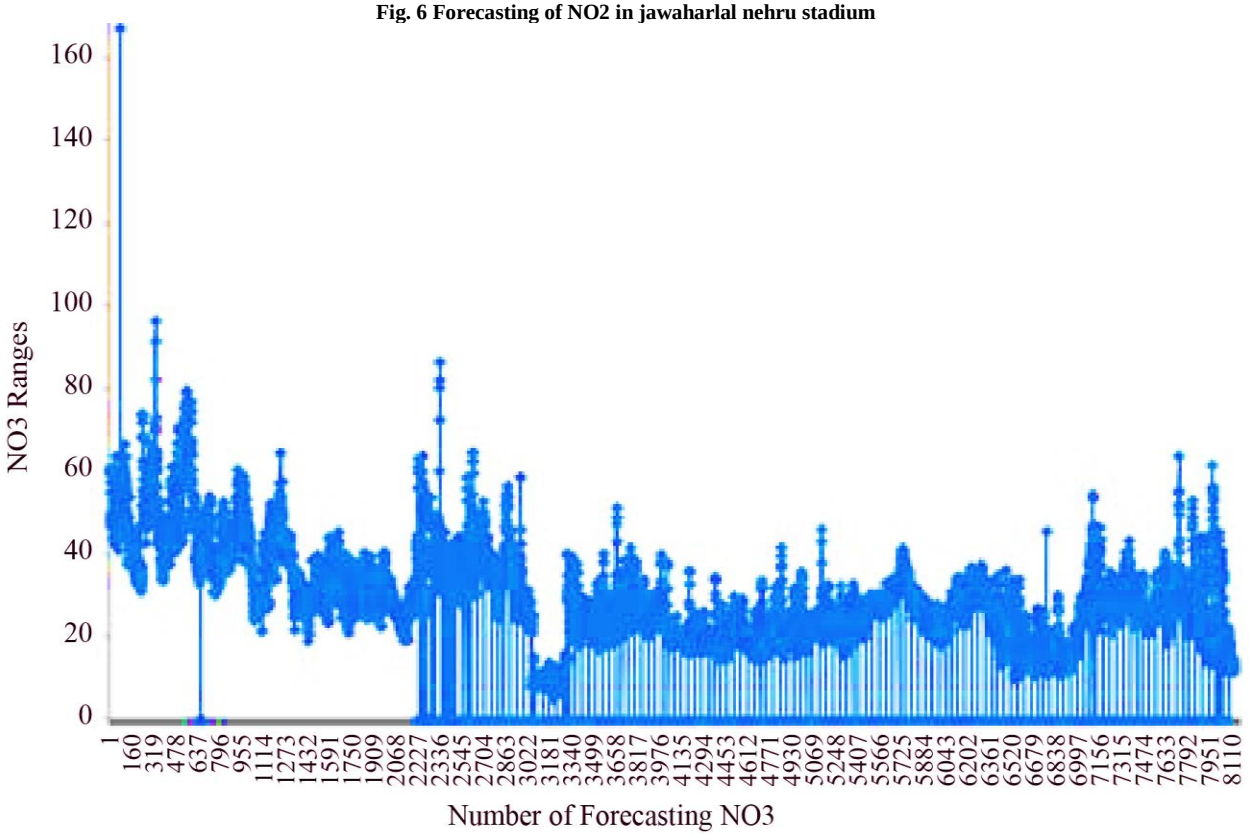

Fig. 7 Forecasting of NO3 in jawaharlal nehru stadium

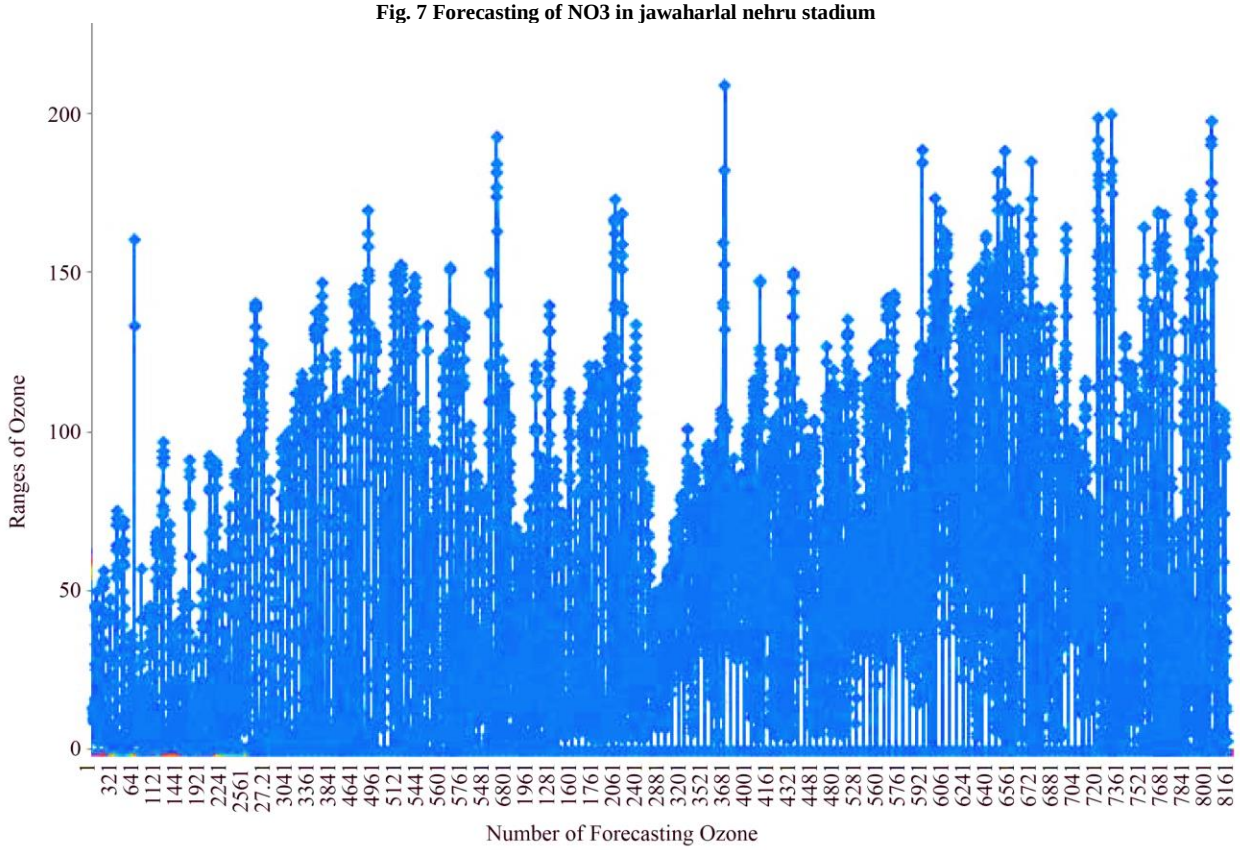

Fig. 8 Forecasting of ozone in jawaharlal nehru stadium

Table 5. The Average forecasted ranges for Various particles from 6 Hours to 3 Days

Prediction Locations / forecasting ranges	Current Prediction	Forecasting	Forecasting	Forecasting	Forecasting
	6:00 AM	At 12 PM	1 day	2 days	3 days
Location 1: Jawaharlal Nehru Stadium, Delhi.	120	204	180	183	196
Location 2: Alipur, Delhi.	118	181	170	176	160
Location 3: Dodhi Road, Delhi	115	158	143	160	157
Location 4: Noida	106	208	198	186	198

However, the same day at 12 PM, the forecasting data automatically increased due to different temporal properties such as vehicle movement, people migration, industry running time, etc. These reasons led to an automatic increase in the AQI indexing rates. The Prediction and Forecasting of AQI indexing ranges are shown in Table 4. In particular, in the Jawaharlal Nehru Stadium and Noida, the status of the AQI index was found to be very poor, since at 12 PM, both vehicle movement and people movement in both these locations were very high. Similarly, In the Noida location, the automatic increase of the AQI index at 12 PM was also due to the high number of industries.

Additionally, compared to the other locations, Noida and Jawaharlal Nehru Stadium reported higher pollution ranges due to the absence of adequate green cover in these areas, leading to a rapid spread of pollution particles to surrounding areas. Similarly, the prediction ranges of various pollution particles such as CO, SO2, NO2, and O3 particles have been summarized in Table 5.

4.2. Prediction Performance Analysis using MAE and RMSE

The performance of the proposed work was compared using MAE and RMSE. Using these two metrics, the performance of the proposed work was evaluated. The proposed work was compared to previously performed standard works such as GBT, LSTM, and ALSTM [27]. The comparison of the proposed work using MAE and RMSE has been presented in Figures 9 and 10. Both MAE and RMSE average prediction error values were observed to have decreased in the proposed work.

The proposed work was directly compared with ALSTM. The predicted error rates of MAE of the proposed hybrid method were 6 hours (0.8), 12 hours (0.4), 24 hours (0.2), and 48 hours (0.3) respectively. Similarly, the error rates of RMSE of the proposed method were 6 hours (1.2), 12 hours (1), 24 hours (1), and 48 hours (1.8), respectively. A comparison of the MAE and RMSE error rates revealed that the MAE produced lesser prediction errors. Comparison of the proposed method with existing methods such as GBT, LSTM, and ALSTM also revealed that it produced lesser errors in terms of hours and days.

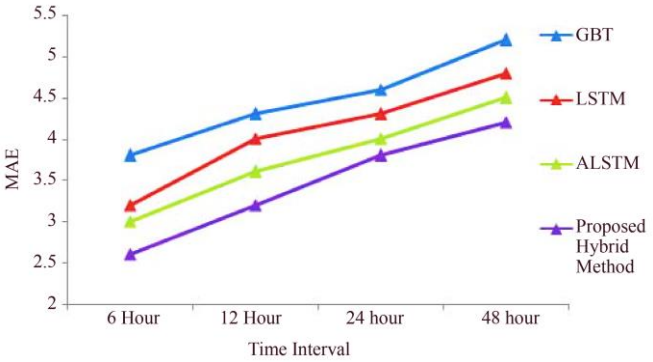

Fig. 9 Comparison of MAE using different methods

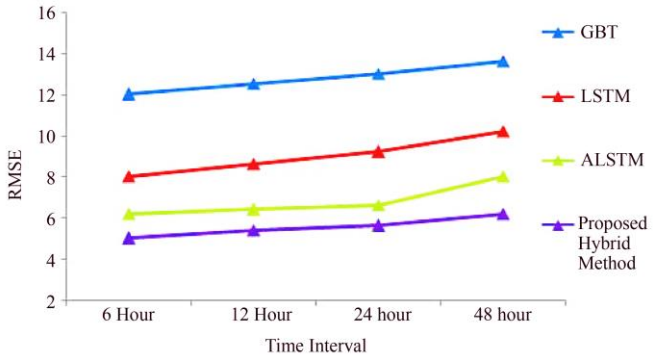

Fig. 10 Comparison of RMSE using different methods

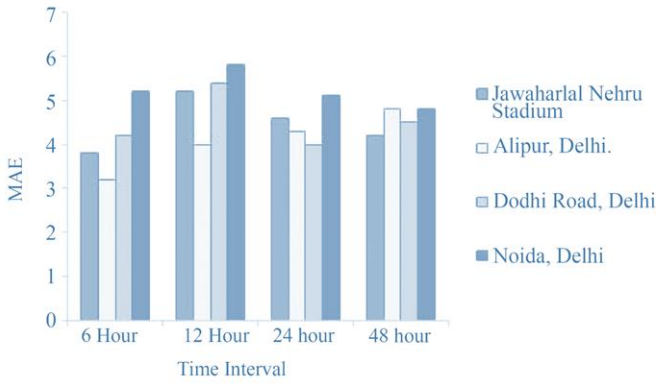

Fig. 11 Comparison of MAE using different locations

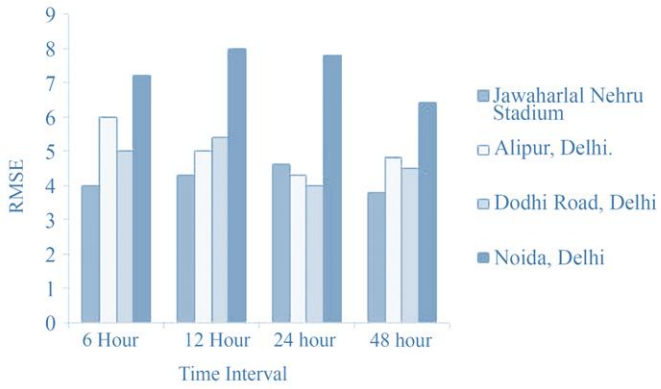

Fig. 12 Comparison of RMSE using different locations

Location-wise performance analysis using MAE and RMSE: Location-wise error was also predicted using the MAE and RMSE. The proposed work used four locations, as mentioned in Table 2, and their corresponding errors were also predicted using intervals of 6 hours, 12 hours, 1 day, and 2 days, respectively. The predicted error rates corresponding to different locations have been presented in Figures 11 and 12. The errors for each location were calculated separately, but the consolidated error rates of all four locations have been presented here. The consolidated error rates were found to be 6 hours (4.1), 12 hours (5.1), 24 hours (4.5), and 2 days (4.58) respectively. Similarly, the RMSE error rates were 6 hours (5.5), 12 hours (5.1), 24 hours (5.17), and 2 days (4.80) respectively. Thus, the proposed work reduced location-wise prediction error rates compared to other methods such as GBT, LSTM, and ALSTM. Hence, an increased prediction accuracy was noted in terms of different hours.

5. Conclusion

Accurate pollution forecasting and updates of the same at specific time intervals is an important area of research since it affects the population's health and quality of life in the immediate surroundings. Effective and accurate prediction and monitoring of the air quality of our surroundings can greatly benefit the living environment around us and help avoid unfortunate health issues and distress.

In this work, a hybrid model was proposed for forecasting the level of pollution particles at different time intervals. The main novelty of the proposed work is the automatic and frequent update of the pollution particle levels as well as of other dynamic parameters such as vehicle movement, weather status, wind speed, and rainfall rates, which were all considered during the forecasting. This proposed hybrid method used the D-based indexing method, SARIMA, Bi-LSTM, and the Pearson correlation. This D-Tree-based indexing method was used to manage past and current collected data. The SARIMA is used to predict and forecast the future status of the pollution particles by trending seasonal and current data. The Bidirectional LSTM was used to predict the time series forecasting based on the current and past data managed by the D-tree indexing method. The Pearson correlation was used to manage the mean values of the two predicted outputs, SARIMA and Bi-LSTM. The experiment was performed using four locations and their data for the past one year, as well as the live data received using different sensors. The collected data was managed using indexing methods and forecast-ed using SARIMA and Bi-LSTM. The errors in the forecasted results were predicted using MAE and RMSE. The hourly and day-wise error prediction rates for MAE were 6 hours (0.8), 12 hours (0.4), 24 hours (0.2) and 48 hours (0.3) respectively.

Similarly, the RMSE error prediction rates were 6 hours (5.5), 12 hours (5.1), 24 hours (5.17), and 2 days (4.80) respectively. Compared to the existing methods, such as GBT, LSTM, and ALSTM, the proposed hybrid method produced reduced error rates and better prediction accuracy. Due to the indexing and SARIMA methods, the seasonal data was updated continuously with respect to specified time intervals and threshold values. The future direction and scope of the research would include various other temporal parameters like vehicle traffic flow and direction, wind speed, etc., to strengthen further the current research based on location and time. Additional temporal data such as rainfall, growth, and development of new industries and their active hours can also be considered for forecasting the level of pollution particles.

References

[1] Lu Bai et al., "Air Pollution Forecasts: An Overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 780, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] C.H. Bosanquet, and J.L. Pearson, "The Spread of Smoke and Gases from Chimneys," *Transactions of the Faraday Society*, vol. 32, pp. 1249-1263, 1936. [CrossRef] [Google Scholar] [Publisher Link]

[3] European Union Joint Research Centre (JRC), "Features of Dispersion Models," 2004.

[4] Nolan Atkins, Air Pollution Dispersion: Ventilation Factor, NVU-Lyndon Atmospheric Sciences, 2008. [Online]. Available: https://apollo.nvu.vsc.edu/classes/met130/notes/chapter18/dispersion_intro.html.

[5] Donald Ermak, "*User's Manual for Slab: An Atmospheric Dispersion Model for Denser-than-Air Releases*," Technical Report, OSTI.GOV, UCRL-MA-105607, 1990. [Google Scholar] [Publisher Link]

[6] Pritthijit Nath et al., "Spatio-Temporal Pollution Forecasting using Hybrid Networks," *Research Square*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yanlin Qi et al., "A Hybrid Model for Spatiotemporal Forecasting of $PM_{2.5}$ based on Graph Convolutional Neural Network and Long Short-Term Memory," *Science of the Total Environment*, vol. 664, pp. 1-10, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[8] Dewen Seng et al., "Spatiotemporal Prediction of Air Quality Based on LSTM Neural Network," *Alexandria Engineering Journal*, vol. 60, no. 2, pp. 2021-2032, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[9] Pritthijit Nath et al., "Long-Term Time-Series Pollution Forecast using Statistical and Deep Learning Methods," *Neural Computing and Applications*, vol. 33, pp. 12551-12570, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[10] Fang Zhao et al., "Research on PM2.5 Spatiotemporal Forecasting Model Based on LSTM Neural Network," *Computational Intelligence and Neuroscience*, vol. 2021, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[11] Shurui Fan et al., "A Hybrid Model for Air Quality Prediction Based on Data Decomposition," *Information*, vol. 12, no. 5, p. 210, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[12] Wenjing Mao et al., "A Hybrid Integrated Deep Learning Model for Predicting Various Air Pollutants," *GIScience & Remote Sensing*, vol. 58, no. 8, pp. 1395-1412, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[13] Shengdong Du et al., "Deep Air Quality Forecasting using Hybrid Deep Learning Framework," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2412-2424, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[14] Yang Han et al., "A Domain-Specific Bayesian Deep-learning Approach for Air Pollution Forecast," *IEEE Transactions on Big Data*, vol. 8, no. 4, pp. 1034-1046, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[15] Van-Duc Le, "Spatiotemporal Graph Convolutional Recurrent Neural Network Model for Citywide Air Pollution Forecasting," *arXiv*, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[16] Atakan Kurt et al., "An Online Air Pollution Forecasting System using Neural Networks," *Environment International*, vol. 34, no. 5, pp. 592-598, 2008. [CrossRef] [Google Scholar] [Publisher Link]
[17] Vlado Spiridonov et al., "Development of Air Quality Forecasting System in Macedonia, based on WRF-Chem Model," *Air Quality, Atmosphere & Health*, vol. 12, pp. 825-836, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[18] Sharnil Pandya et al., "Pollution Weather Prediction System: Smart Outdoor Pollution Monitoring and Prediction for Healthy Breathing and Living," *Sensors*, vol. 20, no. 18, pp. 1-25, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[19] Jian Wei Koo et al., "Prediction of Air Pollution Index in Kuala Lumpur using Fuzzy Time Series and Statistical Models," *Air Quality, Atmosphere & Health*, vol. 13, pp. 77-88, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[20] Shuixia Chen, Jian-qiang Wang, and Hong-yu Zhang, "A Hybrid PSO-SVM Model Based on Clustering Algorithm for Short-Term Atmospheric Pollutant Concentration Forecasting," *Technological Forecasting and Social Change*, vol. 146, pp. 41-54, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[21] Hufang Yang et al., "A Novel Combined Forecasting System for Air Pollutants Concentration based on Fuzzy Theory and Optimization of Aggregation Weight," *Applied Soft Computing*, vol. 87, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[22] Angel Cujia et al., "Forecast of PM10 Time-Series Data: A Study Case in Caribbean Cities," *Atmospheric Pollution Research*, vol. 10, no. 6, pp. 2053-2062, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[23] Snezhana Georgieva Gocheva-Ilieva et al., "Regression Trees Modeling of Time Series for Air Pollution Analysis and Forecasting," *Neural Computing and Applications*, vol. 31, pp. 9023-9039, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[24] Zena A. Aziz Aziz, and Siddeeq Y. Ameen Ameen, "Air Pollution Monitoring using Wireless Sensor Networks," *Journal of Information Technology and Informatics*, vol. 1, no. 1, pp. 20-25, 2021. [Google Scholar] [Publisher Link]
[25] Yang Yurong et al., "A Study on Water Quality Prediction by a Hybrid CNN-LSTM Model with Attention Mechanism," *Environmental Science and Pollution Research*, vol. 28, pp. 55129-55139, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[26] Gaurav Anand, Sharda Kumari, and Ravi Pulle, "Fractional-Iterative BiLSTM Classifier: A Novel Approach to Predicting Student Attrition in Digital Academia," *SSRG International Journal of Computer Science and Engineering*, vol. 10, no. 5, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[27] Yue-Shan Chang et al., "An LSTM-based Aggregated Model for Air Pollution Forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451-1463, 2020. [CrossRef] [Google Scholar] [Publisher Link]