

Research

A deep learning approach for prediction of air quality index in smart city

Adel Binbusayyis¹ · Muhammad Attique Khan² · Mohamed Mustaq Ahmed A³ · W. R. Sam Emmanuel⁴

Received: 23 December 2023 / Accepted: 2 May 2024

Published online: 14 May 2024

© The Author(s) 2024 [OPEN](#)

Abstract

Industrial developments and consumption of massive amount of fossil fuels, vehicle pollution, and other calamities upsurges the AQI (Air Quality Index) of major cities in a drastic manner. Owing to these factors, it is important to take proactive measures for reducing the air pollution in order to avoid life-threatening consequence. Therefore, prediction of air quality is significant for improving the health of living beings as highly polluted regions have a higher concentration of pollutants mixed in the air, affecting the respiratory system and reducing the lifetime. To control pollution, AQI is used as a measure for estimating the pollutant content in the air. Even though many existing techniques have predicted AQI, enhancement is required in prediction algorithms with minimized loss. To address the challenges in traditional algorithms, the proposed smart cities-based AQI prediction intends to utilize the proposed regression algorithm in the dataset, namely Air-Quality-Data, which collected harmful pollutants on an hourly and daily basis from multiple cities in India between 2015 to 2020. To achieve prediction efficiency with reduced loss, pre-processing of input data is being performed using Deep GAN (Generative Adversarial Network). It performs the imputation of data in place of missing values to improve accurate prediction. Additionally, feature scaling normalizes independent real-data features to a fixed scale. With the processed data, regression is done through modified Stacked Attention GRU with KL divergence, which predicts Ernakulam, Chennai and Ahmedabad cities with higher, medium, and low levels of AQI in India. The performance of the proposed regression algorithm is measured using metrics such as MAE (Mean Absolute Error), MSE (Mean Square Error), R2 (Coefficient of determination), MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Square Error) and better MAE, MSE, R2, MAPE and RMSE obtained by the model is 0.1013, 0.0134, 0.9479, 0.1152 and 0.1156. Internal assessment and comparative analysis performed with existing regression algorithms exhibit lower loss values obtained from the present research, which determines the efficacy of the proposed model.

Keywords Air quality index · Pollutants · Regression · Deep generative adversarial network · Modified stacked attention GRU · Chennai · Ernakulam · Ahmedabad

✉ Adel Binbusayyis, a.binbusayyis@psau.edu.sa; Muhammad Attique Khan, attique@ciitwah.edu.pk; Mohamed Mustaq Ahmed A, amohamedmustaqahmed@thenewcollege.edu.in; W. R. Sam Emmanuel, sam_emmanuel@nmcc.ac.in | ¹Department of Software Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, 11942 Al-Kharj, Saudi Arabia. ²Department of Computer Science, University of Engineering & Technology, Taxila, Pakistan. ³Department of Information Technology, The New College, University of Madras, Chennai 600 014, India. ⁴Department of Computer Science, Nesamony Memorial Christian College, Manonmaniam Sundaranar University, Tirunelveli 627 012, Tamil Nadu, India.



1 Introduction

The atmosphere is being polluted by air, water, and land for various reasons [1]. Particularly, polluted air affects human health to a greater extent [2]. Inhaling poor-quality of air damages both respiratory and cardiovascular organs [3–5]. Along with that, the pollutants also affect plants and other living beings (Kumar). Such air pollution is mainly caused by burning fossil fuels, which release an excess of sulphur dioxides and nitrogen oxides that gets mixed up with air [6]. In addition, the emission of smoke from vehicles and industries also becomes a source of air pollution. To protect the environment from danger, AQI (Air Quality Index) is being calculated to measure the emission level of gases which determines the highly polluted areas [7]. Based on the information collected from the index, appropriate actions can be taken to protect the health of humans.

Also, most countries utilize AQI to provide policies for better air quality [8, 9]. Identifying particular hotspots experiencing greater levels of air pollution is being controlled by targeted actions and considerably improves the future well-being of the environment [10]. Due to the serious impacts of air pollution, it is found to be very sensitive in predicting air quality. For this purpose, many AI (Artificial Intelligence) based algorithms have been used in prediction. Recently, with the progress of computing algorithm technology, deep learning models have begun to be used to analyze and forecast nonlinear relationships between data variables. In a similar way, deep learning models have significantly improved data analysis performance because they provide highly reliable results [11]. Correspondingly, RFR (Random Forest Regression) and SVR (Support Vector Regression) have been utilized in [12] for building the regression structure in the prediction of AQI. The concentration of nitrogen oxides in an Italian city from a public dataset has been analyzed. Coefficient correlation, RMSE, along with R^2 calculation, has measured the performance of both models and exhibited that both methods have achieved better prediction efficiency. Similarly, the suggested research [13] has utilized an optimal and hybrid method that has fused the benefits of SD (Secondary Decomposition) with an AI optimization algorithm. AQI data collected from china between 2016 and 2018 has been considered to verify the prediction model. Wavelet decomposition has been generated with higher frequency, lower frequency, and variational mode, along with sample entropy for smoothening the decomposition. Followed by that, LSTM (Long Short Term Memory) based NN (Neural Network) has been adopted for the prediction of AQI [14]. In addition, parameters of least-square SVM (Support Vector Machine) have been optimized by the BAT technique by considering the factors of air pollutants. Outcomes have revealed that it has captured characteristics of AQI data efficiently.

Recognizing the influence of pollution in air through the AQI algorithm has been performed through the considered model [15] depending on certain meteorological attributes. The collection of such parameters has been done through the inverse-distance-weighted statistical method. Analysis outcome has been achieved through measurements obtained from three different stations in Kuala Lumpur during the time interval between June and August 2018. In addition, OLS (Ordinary Least Squares) process in GIS (Geographical Information System) based on statistical analysis in recommended research [16] has been adopted for three months' time period, and the predicted value of AQI has been compared with the observed AQI. Outcomes revealed that the considered analysis [17, 18] had provided efficient monitoring and prediction of AQI with the expected accuracy level.

Moreover, the evaluation of the prediction of pollutant concentration on an hourly basis has been performed with the ANN (Artificial Neural Network) in Iran for the complete year from 2009 to 2010. The suggested algorithm [19] neurons in the hidden layer and a single output layer at the final stage. Six pollutants, such as O_3 , PM_{10} , SO_2 , CO, NO_2 , and $PM_{2.5}$, have been considered in [20, 21] AQI measurements from R^2 are significant in all three areas. RMSE value obtained through comparison has achieved better prediction on air quality which has assisted in making effective decisions. The earlier prediction of air quality aids in controlling pollution considerably [22, 23]. Equivalently, recommended research [24] has established multiple site quality prediction structures using the DL model based on a clustering strategy. The performance of the prediction of BPNN for four hours has been ranked in ascending order.

Additionally, CNN (Convolutional Neural Network) integrated with LSTM has been performed for multiple-hour analysis [25]. Performance outcome has indicated that CNN with LSTM has gained a better level of prediction than with CNN and BPNN. The emission of air pollutants has been monitored through the heuristic optimization method in [9], which has conducted pre-processing techniques in detecting outliers and has excavated significant characteristics in optimizing extreme learning parameters. Experimental analysis has emphasized the significance of outlier detection and provided a feasible solution for predicting AQI measures [26, 27]. It has also been investigated that the prediction methods have faced difficulty with time consumption and complexity. To tackle such issues, the adoption of SVM in the recommended study [28] has utilized three significant parameters such as regularization, penalty factor,

and kernel function, in which the attributes of kernel function have been improved to enhance the performance of the algorithm. Outcomes from the considered model have been analyzed with error metrics. Efficiency has been addressed through data screening techniques in the outlier-detection process.

Although existing studies have used numerous algorithms and methods in the prediction analysis of air quality in smart cities, accurate prediction of AQI with minimum error rate in the prediction process using regression algorithm still requires improvement. Moreover, the regression analysis on quality prediction alone has not assisted in minimizing the pollution rate, the missing and unstable conditions in input data also require an efficient data-pre-processing technique to prevent backdrops in AQI assessment. Further, AQI assist in predicting the health risks associated with air pollution as AQI plays a critical role as tool for assessing and monitoring the quality of air for protecting public health and also understand the root cause of pollution. Therefore, precise model is needed for effective AQI, hence accurate measurement of AQI and suitable processing of missing values together will reduce the backdrops considerably. To this purpose, the DL based regression method in AQI prediction with effective pre-processing of input data is proposed in the present study for controlling air pollution in the future. By adopting the effective integration of DL algorithms, the proposed research addresses the challenges of existing studies in attaining effective predictions on AQI measures.

1.1 Gaps identified

The evaluation of the existing works has used numerous intelligent models for prediction with AI (Artificial Intelligence) oriented algorithms for effective AQI prediction, however, there are significant gaps that has been identified such as imprecise AQI model, inability of the models to work with massive data, monitoring AQI effectively. Thus, effective models need to be used for overcoming these limitations. Hence, the objective of the present research is projected as follows,

- To pre-process the data using proposed Deep GAN (General Adversarial Network) by imputing the missing inputs in the dataset by effectively monitoring the quality of air.
- To perform regression with the modified Stacked Attention GRU (Gated Recurrent Unit) with modified KL (Kullback–Leibler) divergence to predict AQI (Air Quality Index) in three different cities like Ahmedabad, Chennai and Ernakulam with minimized regression loss.
- To evaluate the present system with performance metrics such as MSE, RMSE, R^2 , MAE, and MAPE internally and performs a comparative assessment with existing algorithms for validating the efficacy of the proposed model.

1.2 Paper organization

The research paper is organized as follows. Section 2 shows the review made on several existing studies related to pollution estimation of AQI in smart cities. Section 3 sketches the proposed methodology with the regression algorithm used. Section 4 discusses the system's performance through internal and comparative analysis with other techniques. Finally, the conclusion part is provided in Sect. 5.

2 Literature review

The following section discusses various techniques and methodologies used to identify AQI prediction.

Many existing prediction methods have lacked efficiency because of locally optimal solutions trapped with feature selection along with over-fitting complications. The suggested model [29] tackled overcoming the optimal local problem with the BSMO (Balanced Spider Monkey Optimization) in selecting the best features. Air quality data has been acquired from the CPCB (Central Pollution Control Board) from four different cities: Chennai, Bangalore, Cochin, and Hyderabad. The Min–Max normalization method has been performed to fill up missing values in the dataset. CNN model has been applied in the deeper representation of the input. BSMO methodology has selected features based on balancing attributes and passed the features to BI-LSTM (Bidirectional LSTM), which has predicted air quality with index measure. Suggested hybrid BSMO with BI-LSTM has attained effective performance in predicting AQI. Similarly, AQI for Delhi has predicted using LSTM-GRU model [30]. Implementation of the model has resulted in considerable outcome for AQI prediction, in which the R Square value obtained by the model was 0.84. Likewise, AQI for Bangladesh [31] has been notified using LSTM-GRU model. The outcome of the model suggested that, hybrid model has delivered better outcome than standalone models.

To protect society with a pollution-free environment, the considered model [32] has used an attention-based AQI predictor, following sequence-to-sequence modeling. Moreover, it has exploited historical quality information in predicting future indexes. To overcome the minimum training speed [33], the encoder in RNN has been replaced with a fully connected encoder to enhance the training. Position-level embedding has been used to predict the relationship between the sequences. Recurrent levels of prediction have caused an increase in error, and it has been resolved through the n-step process of RNN. Experimental outcomes with the air quality predictor and n-step recurrent process have attained reduced training time in the prediction of the index and decreased rate of error. Prediction of AQI acts as the alarm when the concentration level of gases has exceeded a certain limit and contributes to taking effective measures to the appropriate departments. Traditional methods have faced difficulty utilizing temporal features from correlations acquired from various monitoring regions and have attained poor predictions above 24 h. Such issues have been handled through the recommended research [34] using the DL model in the prediction of quality levels in the air all through the day. TS-LSTME (Temporal Sliding Long Short Term Memory Extended) model. It has integrated optimized time lag for the realization of prediction with the multilayer involved with the hourly-based concentration of temporal, $PM_{2.5}$, and meteorological information. Outcomes exhibited a better level of stability and obtained a higher correlation coefficient. The considered model has predicted the quality index with an O_3 level of concentration.

Effective decisions on protecting the atmosphere have been performed through significant index measurements of $PM_{2.5}$ (Particulate Matter), which acts as the indicator of controlling the frequency of air pollution. Semisupervised learning model has been designed in [35], which has included EMD (Empirical Mode Decomposition) with BI-LSTM networks. It has taken the level of $PM_{2.5}$ as the input data, which has been considered the signal through the adoption of EMD as a feature-learning technique that has decomposed data and has extracted both amplitude and frequency features. Certain sudden variations have been recorded and have improved short-term predictions. Bi-LSTM has been applied during the supervised learning phase. Likewise, study [36] has compared DL algorithm like LSTM and ML algorithms like ARIMA, DT, KNN, Dummy regressor, Adaptive Boosting, gradient and extreme gradient boosting, Huber regressor. Analytical outcome has depicted that, DL based LSTM has delivered better outcome than ML models for forecasting the quality of air. Similarly, GA-KELM (Genetic Algorithm Based Improved Extreme Learning Machine) [37] has used in the paper for predicting the quality of the air, along with other models like SVM and DBN-BP. However, GA-KELM model has delivered rapid outcome and precise outcome than existing model.

The rapid development of technologies in smart cities leads to different kinds of pollution in the environment. More specifically, air pollution has increased due to transportation and the excretion of smoke in manufacturing industries [38]. It is highly recommended by the government to measure AQI in polluted areas, and the suggested model [39] has provided an ML algorithm with intelligent based data –analysis in the prediction of air quality. It has utilized the integration of the DT (Decision Tree) algorithm with the NN for an accurate level of prediction. Similarly, integrating LSTM with DAE (Deep Auto Encoder) is considered research [40] have predicted PM concentration from 2015 to 2018 in South Korea, analyzing the quality of air hourly. On average, the concentration level of PM has been predicted from continuous 10 days with the optimal learning rate with 100 epochs for 0.01 for the size of batch in LSTM as 32. The best performance of DAE has been achieved with a batch size of 64. Experimental outcomes with the considered methodology have exhibited better performance in prediction.

Air pollution has been one of the issues in recent decades, thus, forecasting the quality of air play a vital role in controlling the air pollution. Therefore, SVR (Support Vector Regression) and RFR (Random Forest Regression) has used for building the regression model for predicting the AQI. Despite the performance of the model, SVR model is not suitable for processing huge number of samples [12]. The development of IoT (Internet of Things) in the prediction of quality measures in the air has assisted in improving the quality of life of humans. In the same way, the recommended research [41] has utilized sensors for the monitoring process and predicts the emission gas level of concentration, which affects the purity of air. The initial process has been handled with sensors, values stored in a cloud-based server have been measured using LSTM, and obtained results have been compared with SVR. It has measured the AQI value based on the concentration of harmful substances present in the air. With the assistance of DL algorithms, the concentration of pollutants has been predicted for two years. Intelligent prediction models in [42] have used RNN (Recurrent NN) with PSO (Particle Swarm Optimization). The Computation of the model has been performed through the LSTM structure. It has considered six different kinds of pollutants mixed in the air for the prediction of AQI. It has collected data from multiple stations and handled missing values in pre-processing using normalization with the MinMax-scalar approach and evaluated the AQI for every hour in 30 days.

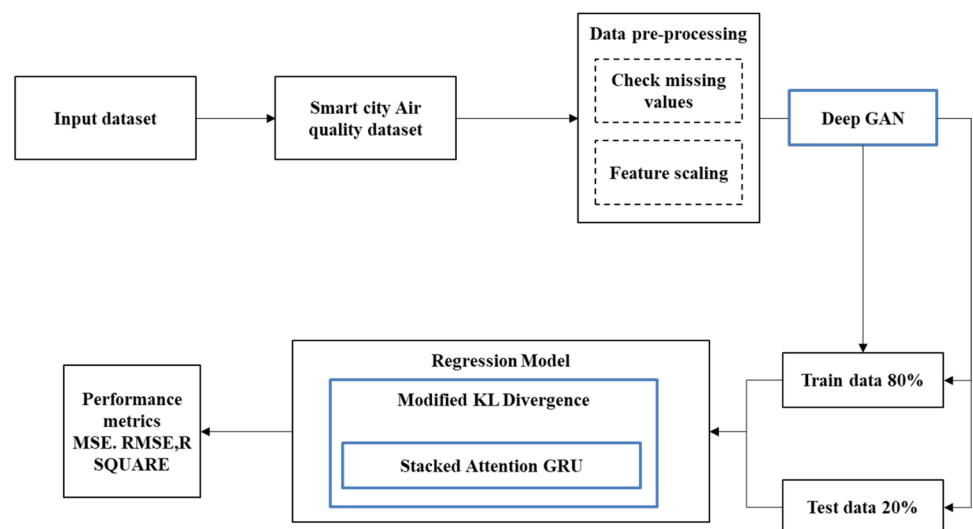
3 Proposed methodology

The proposed regression model employs the DL algorithm for analyzing the concentration level of pollutants mixed in the air to predict the pure quality of air through AQI measures. Continuous monitoring of air quality in different cities in India assists the government in taking appropriate quality measures to control pollution. Existing literature on quality prediction has suffered with maximum loss, and multiple missing values in the dataset have created biased outcomes and unstable prediction results. Moreover, it has undertaken limited areas for analysis and lacked generalized prediction. Hence, the present study is motivated by all such factors and intends to enhance the efficiency of prediction through an effective regression technique with the enhancements performed at the pre-processing process using Deep GAN and regression process accomplished using Modified Stacked Attention GRU with Modified KL Divergence. The architecture diagram of the proposed model is depicted in Fig. 1.

As illustrated in Fig. 1, the proposed working mechanism is divided into two stages: pre-processing and regression. The input dataset considered for computation is Air-quality-data in India dataset. The pollutant data are taken from different cities in India concerning the information about the concentration level of various pollutants mixed in the air. With that inputs, the possibilities of missing values are greater in range. To make an effective outcome on prediction, missing values are handled through Deep GAN, and feature scaling is processed to minimize the complexity of the model. Independent features are standardized to a fixed range in the dataset during the scaling process. Higher variances in the magnitude of values are handled with Deep GAN. With the assistance of generators and discriminators, it generates data samples from the statistical distribution of information. It outputs samples in the dataset which are closely associated with the real distribution of values in the dataset. If missing values are untreated, it creates complications in real-time datasets and produces biased outcomes. After that, the pre-processed data are splitted up into 80% training and 20% testing of the model, which then passes on to the regression phase. Finally, effective prediction is being performed with the modified Stacked Attention GRU with KL divergence for efficient prediction of AQI. It is mainly used for mapping selected pollutant attributes with the collected pollution data due to the self-feedback connections and improved training methods. To manage longer sequences of input data, KL divergence is incorporated, which minimizes the parameters and enhances the model's efficiency. Since Stacked Attention GRU operates with less memory consumption and a faster training method, the proposed system can perform effective prediction, and the complexity of the model is reduced considerably through the proposed regression algorithm. The illustrative diagram of the proposed working structure is illustrated in Fig. 2.

From Fig. 2, the overall process is very clearly represented in which the pollutant sources from different cities in India are being collected and stored in a dataset, namely Air Quality data in India. AQI values are generally categorized as good, moderate, or severe base on the concentration level of pollutants. The dataset holds the data of AQI on an hourly and daily range of various stations over multiple cities in India. The concentration of harmful gases and the AQI index are stored in the dataset. Input data is being processed by the proposed model. Missing values are imputed using GAN, and after filling up input values, the process of training and testing is performed. Followed by that, the regression model plays

Fig. 1 Overall architectural diagram of the proposed method



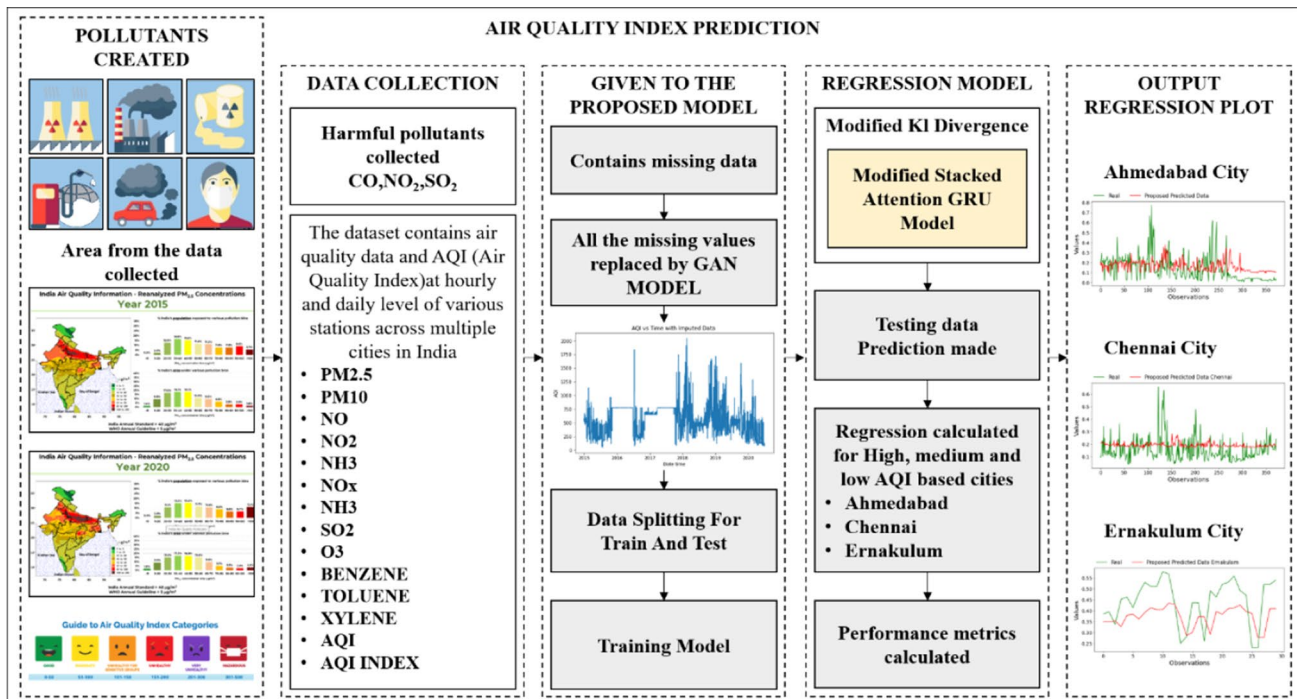


Fig. 2 Illustrative diagram of the proposed model

its role in predicting AQI using the modified stacked attention GRU with KL divergence. From the regression calculation of the higher, medium, and low levels of AQI measured in different cities with the proposed algorithm, it is capable of predicting the three cities such as Ahmedabad with high AQI, Chennai with medium level, and Ernakulum with lower AQI which is being evaluated through error metrics such as MAPE, MSE, R^2 , RMSE, and MAE. It can provide a lower error in the difference between predicted and actual values.

3.1 Dataset description

The dataset preferred for the proposed algorithm analysis is the Air Quality Data in India which possess the data of AQI calculated for hourly and daily level in multiple cities and various stations in India. It measures data between 2015 and 2020. Values of AQI occurred in data at a Particular city containing two different stations are being compared based on hour, city, day, and station level for confirming that the calculations are valid. Calculation of AQI utilizes seven different measures such as NH_3 , SO_2 , CO, $\text{PM}_{2.5}$, O_3 , NO_2 , and PM_{10} . Average AQI values over the previous 24 h are used with the criteria of holding 16 different values. Specifically for O_3 and CO, the maximum concentration level in the last eight hours is being considered. Every measure is converted into a sub-category index depending on the pre-defined clusters. At times, measures many measures are not available due to the absence of measuring or might be due to a lack of specified data points. The final level of AQI amounts to the maximum level of the index with the criteria of presenting at least the concentration of three factors from the seven attributes. The AQI level categorizes the measure as good for (0–50), which shows minimal impact, satisfactory for (51–100), which depicts a minor level of breathing discomfort in old people, moderate (101–200), which affects lungs, children, and leads to heart diseases, poor (201–300) exhibits high breathing difficulties, very poor (301–400) leads to respiratory illness on prolonged exposure and severe (>400) range indicates that it affects the respiratory system entirely even for healthy and young people. The dataset link is <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>.

3.2 Pre-processing with deep GAN imputation

Data preprocessing is performed by checking missing values and feature scaling. Pollution data are stored as discrete values in a dataset in which the occurrence of missing values is higher. It is being handled through an effective missing data imputation strategy. Scaling of features minimizes the complexity. The proposed GAN model in data pre-processing

imputes the missing information directly by considering spatial and temporal relationships with the remaining-observed input values already collected. Table 1 shows the Attribute, missing values of the dataset.

Table 1 depicts the attributes, missing values and the total % values present in the dataset. As the missing values of the dataset is collectively in large number, it is important to replace the missing values in order to improve the performance of the model. GAN is a kind of unsupervised generative structure for summarizing data distribution. It has better generalization capability with limited parameters for processing. The two sections in GAN are the generator and discriminator, in which the generator is trained to generate real data illustrations.

In contrast, the discriminator is trained to differentiate the generated examples from real values. Following that, the generator segment takes the original distribution with a random mask and noise matrix to indicate the real and fake values and then performs the imputation process. The next section of the discriminator obtains the generated information, which determines the components imputed. The generator tries its best to impute the lost information in such a way as to confuse the discriminator.

Similarly, the discriminator puts maximum effort into identifying the imputed information from real values.

Finally, the imputed values are successfully generated when the discriminator is inefficient in finding the imputed data from real data. Considering 15 random variables in the dataset represented as DAT with attributes P_1 to P_{15} as $DAT = (P_1, P_2, \dots, P_{15})$ which is being arranged in the form of $DAT = (P_1, P_2, \dots, P_{15})^K$ where K is referred to as a total number of sampling points and if suppose $M_{vec} = (M_{vec1}, M_{vec2}, \dots, M_{vec15})$ is 15 random attributes with a masking vector representing the missing values of 0 or 1. The incomplete dataset is being defined as $\tilde{D} = (\tilde{P}_1, \tilde{P}_2, \dots, \tilde{P}_n)^K$ in which $\tilde{P} = (\tilde{P}_1, \tilde{P}_2, \dots, \tilde{P}_{15})$ is considered as the random variable masked by the M_{vec} and denoted as

$$\tilde{P}_i = \begin{cases} P_i, & \text{if } M_{veci} = 1 \\ nan, & \text{if } M_{veci} = 0 \end{cases} \quad (1)$$

where *nan* denotes the missing data. The generator takes samples of \tilde{P}_i , N , and M_{vec} represented as $N = (N_1, N_2, \dots, N_{15})$ in which randomized noise is being included, and defining of complete and imputed data are given as

$$\bar{P} = G(\tilde{P}, M_{vec}, (1 - M_{vec}) \otimes N) \quad (2)$$

$$\hat{P} = M_{vec} \otimes \bar{P} + (1 - M_{vec}) \otimes \tilde{P} \quad (3)$$

\bar{P} is the imputed value vector, and G denotes computation operation in a generative network. Finally retrieved complete value is referred to as \hat{P} . Particularly, the data observed directly are outputted as given without changes,

Table 1 Attributes, missing values

Attributes	Missing values	% of total values
Xylene	16989	60.3
PM10	11005	39.1
NH3	10121	35.9
Toluene	7715	27.4
Benzene	5386	19.1
AQI	4575	16.2
AQI_Bucket	4575	16.2
PM2.5	4552	16.2
NOx	4142	14.7
O3	3888	13.8
SO2	3798	13.5
NO2	3465	12.3
NO	3459	12.3
CO	2020	7.2

and missing values are processed with the generator and then outputted. The operator \otimes determines the element level multiplication. The architecture of GAN is self-possessed with two fully connected layers. The activation functions for both hidden and output layers are Relu (Rectified Linear Unit) and the sigmoid functions, respectively. The operations of both sigmoid and Relu are mathematically computed as follows:

$$Relu : f(x) = \max(0, x) = \begin{cases} p_i, & \text{if } p_i \geq 0 \\ 0, & \text{if } p_i < 0 \end{cases} \tag{4}$$

$$Sigmoid : f(x) = \frac{1}{1 + \exp^{-x}} \tag{5}$$

To ensure the imputation quality, the hint method is used for delivering partial information on loss values to the discriminator, which takes both imputed data and hint-matrix for predicting the mask-matrix. The features of real data are captured using the hint mechanism. After obtaining information from the mask-matrix, it provides a hint as 0 as missing data and 1 as observed data. The discriminator tries to output the mask matrix to distinguish the missing value 0 from the real value 1. Satisfactory outcomes are achieved even with more missing values in the dataset. In the imputation process, the discriminator maximizes the level of accuracy in output, whereas the generator minimizes the accuracy of discriminator prediction. The adversarial process of training ensures the quality of imputation. Hence, binary cross-entropy obtains the distance between the actual and predicted possibilities representing 0 or 1. With the help of the hint matrix, the estimated value of the mask matrix is outputted from the discriminator, and the objective function is defined as

$$\min_{Gen} \max_{Dis} \mathbb{E}_{\hat{p}, M_{vec}, H} [M_{vec}^T \log \widehat{M}_{vec} + (1 + M_{vec})^T \log (1 - \log \widehat{M}_{vec})] \tag{6}$$

where \log denotes the element – wise logarithm function, and \widehat{M}_{vec} is considered as the projected mask matrix, which is represented as $\widehat{M}_{vec} = D(\hat{P}, H)$. In addition to that, binary cross – entropy based loss – function is defined as a negative expectation of a log of corrected – predicted possibilities, which is in the form of

$$\mathcal{L}(x, y) = \sum_{i=1}^n [x_i \log(y_i) + (1 - x_i) \log(1 - y_i)] \tag{7}$$

Followed by that, Eq. (6) is being simplified and rewritten as

$$\min_{Gen} \max_{Dis} \mathbb{E}[\mathcal{L}(M_{vec}, \widehat{M}_{vec})] \tag{8}$$

Through the efficiency of GAN, the incomplete data in the dataset are handled through imputation, and with the pre-processed data, regression is processed. Table 2 shows before and after pre-processing.

Before and after pre-processing has been depicted in Table 2 where the undefined numbers are replaced by using GAN model, which helps in precise prediction of AQI.

Table 2 Before and after pre-processing

Before pre-processing	After pre-processing
2015-02-24 6.05	2015-02-24 6.05
2015-02-25 0.81	2015-02-25 0.81
2015-02-26 NaN	2015-02-26 0.82
2015-02-27 NaN	2015-02-27 0.8354
2015-02-28 NaN	2015-02-28 0.8145
2015-03-01 1.32	2015-03-01 1.32
2015-03-02 0.22	2015-03-02 0.22
2015-03-03 2.25	2015-03-03 2.25
2015-03-04 1.55	2015-03-04 1.55

3.3 Modified regression using stacked attention GRU with KL divergence

The regression process integrates KL (Kullback–Leibler) with the Stacked Attention-based GRU. KL divergence reduces the loss of information while approximating the distribution. Combining the divergence with NN assists in learning the complex approximate distribution of data. It measures the difference between the probability distribution over a similar variable. Specific functions for assessment of similarity between the two different distributions are referred to as relative entropy and called KL divergence, in which two specific probability-mass functions are specified as

$$MKL(p(x) \parallel q(x)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (9)$$

The above Eq. (9) determines 0 if both points of distribution, such as $p(x)$ and $q(x)$, have the same quantity of data. The function $MKL(p(x) \parallel q(x))$ provides whether the data is lost and $q(x)$ approximates the value of $p(x)$. When both distribution value outputs a positive value, the KL-divergence value is always a positive number. When the value of $p(x)$ yields 0 value, then the equation attains an infinity value. Similarly, when $q(x)$ is zero, it provides an undefined value. Minimization of KL divergence is equivalent to a reduction of negative-log likelihood.

$$MKL = KL(p(x) \parallel q(x)) = \mathbb{E}_p \left(\log \frac{p(x)}{q(x)} \right) \quad (10)$$

The above Eq. (10) determines the relationship ratio with likelihood, which is projected as the expected date of log-likelihood, which assist in estimating the data distribution. GRU is a kind of RNN (Recurrent Neural Network) which solves the complication of vanishing gradient. Both the encoder and decoder are designed with GRU. The encoder section combines the bidirectional stack of GRU, and the decoder section combines unidirectional GRU. Moreover, the attention method is involved with the recurrent-based sequence generator. The attention process learns the alignment within the input and output sequence. In the case of input mathematical denotation consisting of a sequence of points with variable N represented as

$$\{ [p_1, q_1, r_1], [p_2, q_2, r_2], \dots, [p_L, q_L, r_L] \} \quad (11)$$

In which p_i and q_i are p-q coordinates, and s_i determines the stroke at which the i-th value belongs and addresses the problem of non-uniform based sampling by variations in size and speed of coordinates in which the feature-vectors for every point is represented as

$$[p_k, q_k, \Delta p_k, \Delta q_k, \Delta^2 p_k, \Delta^2 q_k, \delta(s_i = s_i + 1), \delta(s_i \neq s_i + 1)] \quad (12)$$

where $\Delta p_k = p_{k+1} - p_k$, and $\Delta q_k = q_{k+1} - q_k$. The attributes of the equation is represented as $\Delta^2 p_k = p_{k+2} - p_k$ and $\Delta^2 q_k = q_{k+2} - q_k$ with $\delta(\cdot) = 1$ when condition is 0 or 1. The final terms represent flags denoting the status. The mathematical representation is denoted by $P = (P_1, P_2, \dots, P_L)$ representing the input-sequence of the encoder as $P_i \in \mathbb{R}^d$ ($d = 8$). The sequences are represented as (P_1, P_2, \dots, P_L) in which RNN is applied as the encoder for computing the hidden sequences such as (h_1, h_2, \dots, h_L) such that.

$$h_t = \tanh(W_{ph}P_t + W_{hh}h_{t-1}) \quad (13)$$

In which W_{ph} is represented as the connection weight of the network matrix within the hidden and input layers, and W_{hh} is denoted as the weight matrix of recurrent level connections within hidden layers. It performs RNN mapping from the previous input data to every output. The encoder segment in the hidden state is computed using the following equation as

$$z_t = \sigma(W_{pz}P_t + W_{hz}h_{t-1}) \quad (14)$$

$$m_t = \sigma(W_{pm}P_t + W_{hm}h_{t-1}) \quad (15)$$

$$\tilde{h}_t = \tanh(W_{ph}P_t + U_{mh}(m_t \otimes h_{t-1})) \quad (16)$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tilde{h}_t * MKL \tag{17}$$

In which σ is represented as the sigmoid function and \otimes is represented as an element-level operator for multiplication. $z_t, r_t \rightarrow m_t$, and \tilde{h}_t are the update gate, reset gate, and candidate activation. The decoder is equipped with an attention process which is generating a sequence with formulations, and accordingly, the output is encoded as vectors which are signified as

$$Y = \{q_1, \dots, q_n\}, q_i \in \mathbb{R}^k \tag{18}$$

where the parameter k is represented as the total number of attributes in AQI and L is considered as the length with D as the dimensional vector and A as the matrix-vector in which the bi-directional encoder in GRU creates annotation presented as

$$A = \{a_1, \dots, a_{len}\}, a_i \in \mathbb{R}^D \tag{19}$$

To estimate the learning of variable length and associate the length with the computation of intermediate fixed size vectors, the probability of every prediction is denoted as the following equation.

$$p(q_t|P, q_{-(t-1)}) = s(W_o(Eq_{(t-1)} + W_s c_t + W_c sta_t)) \tag{20}$$

The variable s represents the softmax-activation function, c_t explores the present status of hidden states, and sta_t denotes the stacked vector. The decoder also considers the previous stage target as input and adopts the single directional GRU for calculating the hidden state with the representation as follows

$$z'_t = \sigma(W_{qm}Eq_{t-1} + U_{sm}c_{t-1} + C_{cm}sta_t) \tag{21}$$

$$m'_t = \sigma(W_{qm}Eq_{t-1} + U_{sm}c_{t-1} + C_{cm}sta_t) \tag{22}$$

$$\tilde{c}_t = \tanh(W_{qz}Eq_{t-1} + U_{mz}(m'_t \otimes c_{t-1}) + C_{cz}sta_t) \tag{23}$$

$$c_t = (1 - m'_t) \otimes c_{t-1} + m'_t \otimes \tilde{c}_t \tag{24}$$

where m'_t, a'_t , and \tilde{c}_t are considered as he updates, reset along with candidate-activation parameters, respectively. $W_{qm}, W_{qz}, U_{sm}, U_{zm}, C_{cm}$, and C_{cz} are weighted matrices. For every prediction from the decoder section, a subset of adjacent points is contributed to the computation of vectors at every time. Hence, the decoder adopts the attention model in linking the related sequence and assigns a greater weight to the appropriate annotation-based vector. The parameterization of attention is being trained with an encoder and decoder as

$$W_{ti} = V_{att}^T \tanh(W_{att}c_{t-1} + U_{att}a_i) \tag{25}$$

$$\alpha_{ti} = \frac{\exp(W_{ti})}{\sum_{k=1}^{len} \exp(W_{tk})} \tag{26}$$

where α_{ti} represents the attention probability with n' as the attention dimension in which $V_{att} \in \mathbb{R}^{n' \times n}, U_{att} \in \mathbb{R}^{n' \times D}$. With weights as w_{ti} and stack vector sta_t is being calculated as

$$sta_t = \sum_i^{len} \alpha_{ti} a_i \tag{27}$$

The attention model is being used for providing specific concentrations to input pollutants to determine AQI accurately. Computation on coverage-related attention method with the sum of previous attention-probabilities is defined as

$$\rho_t = \sum_l^{t-1} \alpha_l \tag{28}$$

$$cov = Q * \rho_t \quad (29)$$

$$W_{ti} = V_{att}^T \tanh(W_{att} S_{t-1} + U_{att} a_i + U_f f_i) \quad (30)$$

In which α_i is considered the attention probability at the initial step, and the variable ρ_t is also initialized as a zero vector. Such vectors are produced with a convolutional layer with relative attention-based probabilities. The decoding process is accomplished throughout the entire input by assigning a higher level of attention. The proposed stacked attention-based GRU integrating the KL divergence method effectively predicts the AQI measure in a dataset. It determines the high, medium, and low-level polluted cities in India.

4 Results and discussion

The regression performance attained from the computation of the proposed algorithm is evaluated with the rate of prediction and loss obtained. The experimental outcomes are described in the succeeding section and perform a comparison of values internally and externally with several other algorithms to identify the effectiveness of the enhanced algorithm used in present research concerning evaluation metrics like RMSE, MAPE, MSE, MAE, and R^2 .

4.1 Experimental results

The competence estimation of the proposed system is being evaluated with various performance metrics, and this section demonstrates the exploration made with the experimentation of the model along with analyzing comparative assessments are being performed. From regression outcomes, it is found that high, medium and low level polluted cities in India are found. Simulation outcomes achieved through the Ahmedabad, Chennai and Ernakulam are projected in subsequent section.

4.1.1 Ahmedabad

Figure 3a shows the output obtained 'without imputation' and Fig. 3b AQI with imputed data for Ahmedabad.

Figure 3a indicates AQI prediction with the proposed mechanism yielding efficient outcomes. Figure 3b depicts that the proposed imputation by Deep GAN for filling missing values exhibits continuous data distribution and leads to attaining greater prediction efficiency. The highest AQI measure in the Ahmedabad city is given in Fig. 4.

From Fig. 4a, b, the AQI predicted for Ahmedabad city is being validated. It found that the predicted values using the proposed regression algorithm are much similar to the observed values. The green line shows higher volatility and momentous fluctuations compared to the red line. Notable underestimation occurs when the green line peaks while the red line remains lower at the same observation point, indicating that the red line underestimates the values compared to the green line's data. The red line exhibits a smoother progression and gradual decrease in values over time, consistently lagging behind the green line. Notable overestimation instances, where the red line surpasses the green line, are minimal or non-existent based on the provided image.

The reasons behind the prominent underestimation observed in the Fig. 4b may stem from differences in data sources, measurement techniques, or inherent variability in the datasets represented by the green and red lines. Further analysis and comparison of the data sources and methodologies used to generate the green and red lines could provide insights into the reasons behind the observed underestimation. Figure 5 shows the estimation of pollutant contents in the air.

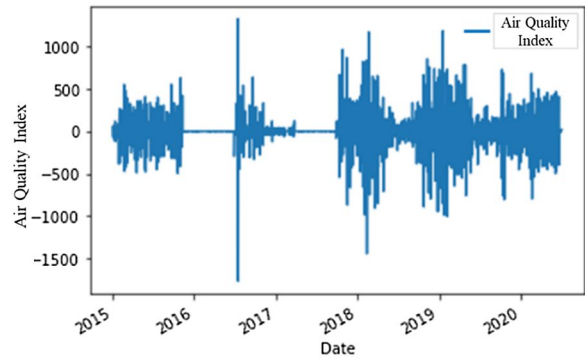
From Fig. 5, it is obvious that the prediction of AQI is based on the number of pollutants present in the air, which data collected from Ahmadabad assists in predicting AQI. Table 3 depicts the 'after imputation' using 0, mean and GAN model.

MSE, RMSE, MAE, MAPE and R Square obtained by using after imputation technique using 0, mean and GAN is depicted in Table 3, where the error rate using proposed technique is less than the other 'after imputation' approaches.

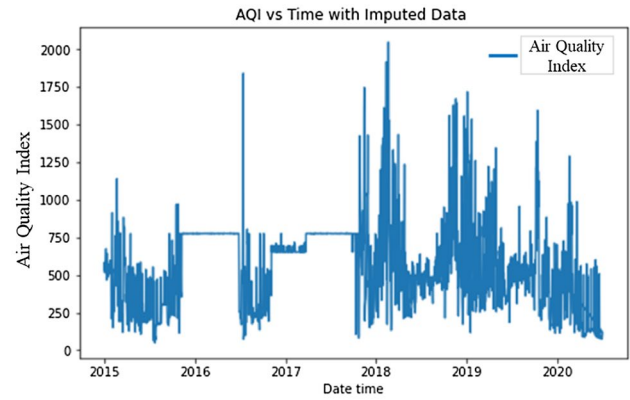
4.1.2 Chennai

The concentration level of harmful gases obtained in Chennai is projected in Fig. 6.

Fig. 3 **a** Without Imputation.
b AQI with Imputed Data

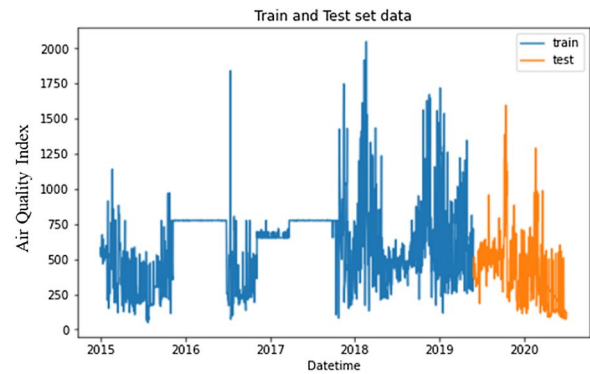


(a) Without Imputation

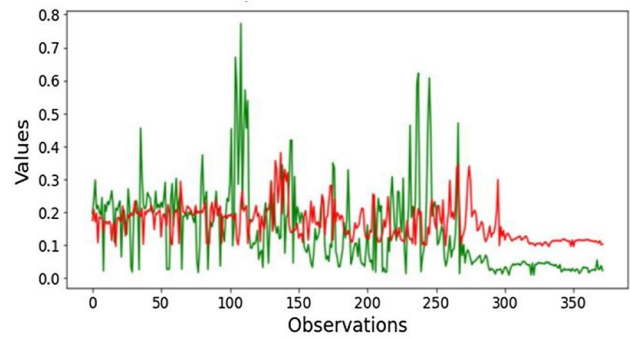


(b) AQI with Imputed Data

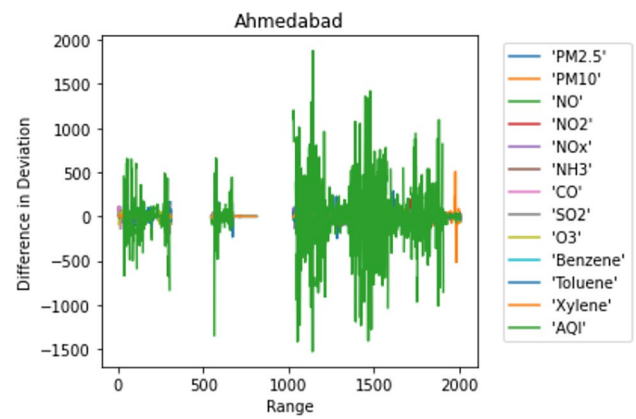
Fig. 4 **a** Prediction Validation.
b Proposed Predicted data in Ahmedabad



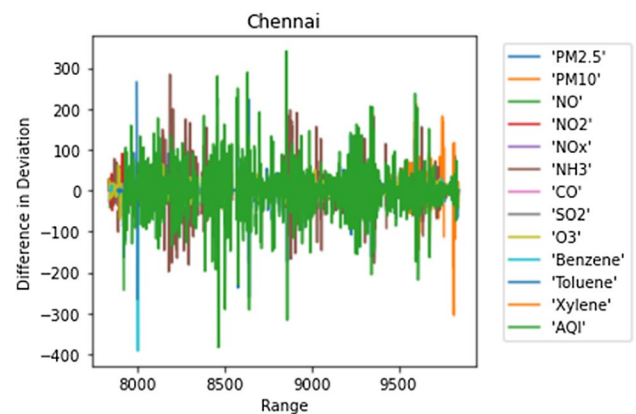
(a) Prediction Validation



(b) Proposed Predicted data in Ahmedabad

Fig. 5 Estimation of pollutant contents in the air**Table 3** Imputation for Ahmedabad

Ahmedabad city	After imputation (0)	After imputation (mean)	After imputation (GAN)
MSE	0.2574	0.3547	0.0161
RMSE	0.5073	0.5956	0.1271
MAE	0.2489	0.3481	0.0979
R-square	0.3567	0.4248	0.8942
MAPE	0.2148	0.3417	0.0587

Fig. 6 Prediction of Pollutant Contents in Chennai City

From Fig. 6, it is evident that the amount of pollutant present in the air is being predicted by estimating the AQI level, which indicates the efficiency of the regression model in determining the degree of polluted air present in Chennai. The experimental results in Chennai city are illustrated in Fig. 7.

From Fig. 7a, b, the prediction of AQI measure in Chennai city is being calculated at different dates and times with imputed data to provide effective prediction outcomes. The observation from the prediction is presented in Fig. 8.

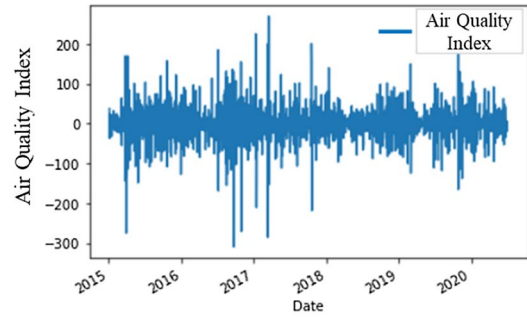
The proposed validation on Chennai city observed from Fig. 8a, b indicates that the observed outcomes are similar to the predicted values, which shows better prediction performance experienced in Chennai city. Similarly, Table 4 illustrates the 'After imputation technique' for Chennai city.

RMSE, MSE, MAE, MAPE, R Square values obtained using GAN model are 0.0977, 0.0095, 0.0795, 0.1219 and 0.9064, which is better than the 'After imputation technique using 0 and mean'.

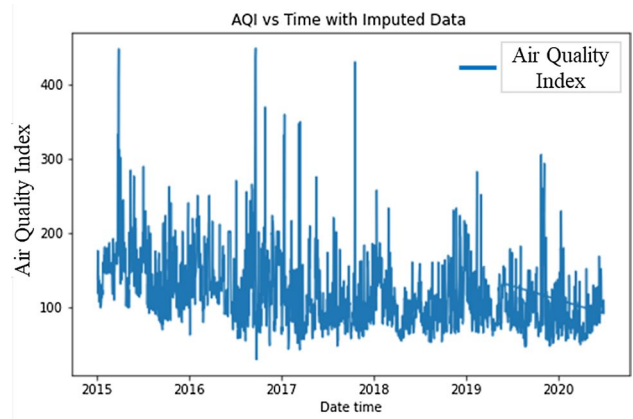
4.1.3 Ernakulam

The concentration level of pollutants in Ernakulam is depicted in Fig. 9.

Fig. 7 **a** AQI Prediction with Dat. **b** AQI versus Time with Imputed Data

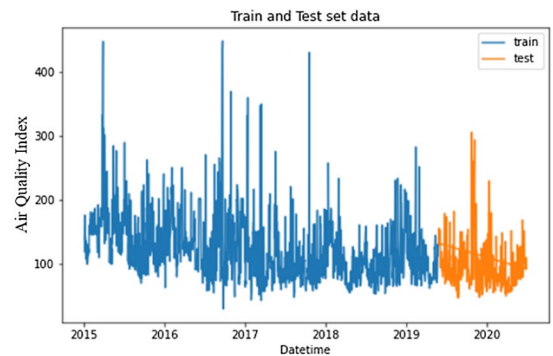


(a) AQI Prediction with Date

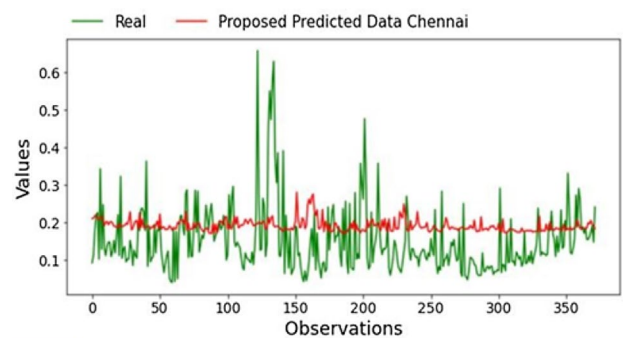


(b) AQI versus Time with Imputed Data

Fig. 8 **a** Validation with Chennai data. **b** Proposed Evaluation with Chennai



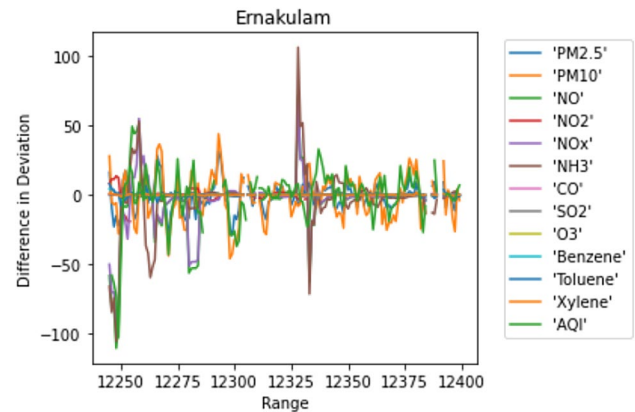
(a) Validation with Chennai data



(b) Proposed Evaluation with Chennai

Table 4 Imputation for Chennai city

Chennai city	After imputation (0)	After imputation (Mean)	After imputation (GAN)
MSE	0.3482	0.4986	0.0095
RMSE	0.5901	0.7059	0.0977
MAE	0.2148	0.3597	0.0795
R-square	0.3975	0.5749	0.9064
MAPE	0.2745	0.3169	0.1219

Fig. 9 Estimation of pollutant contents in the air

From Fig. 9, it is obvious that the prediction of AQI is based on the number of pollutants present in the air, which data collected from Ernakulam assists in predicting AQI. The simulation results of the proposed model are illustrated in Fig. 10.

From Fig. 10a, b, prediction with imputed data determines that Ernakulam shows a low level of pollutant content with the measure of AQI. Imputation with Deep GAN in the proposed work achieves greater efficiency.

The prediction competence is projected in Fig. 11.

From Fig. 11, it is evident that the predicted values with the utilization of Stacked Attention GRU with KL divergence are very similar to observed values, which exhibits the efficiency of the proposed regression algorithm. 'After imputation' outcome for Ernakulam city is depicted in Table 5.

Table 5 shows that MSE, MAPE, R square, RMSE and MAE values obtained using GAN model is better than 'After imputation' technique using 0 and Mean. This is because, when a missing value is simply replaced with 0 and mean values, significant information and variability in the data maybe lost, which leads to biased and inaccurate predictions, however, imputation using GAN model is more effective as it can learn the underlying structure of the data and generate realistic values in order to fill the missing data.

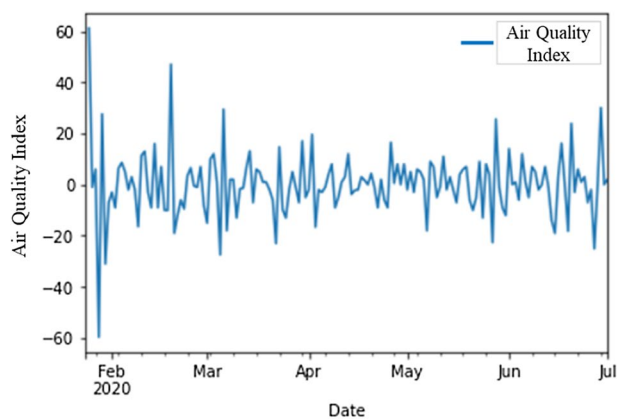
4.2 Performance analysis

The proposed model is evaluated with performance metrics such as RMSE, MAPE, MSE, R-Square, and MAE for estimating the effectiveness of regression. The performance of the proposed approach in terms of loss values for AQI predicted in the city of Ahmedabad is tabulated in Table 6.

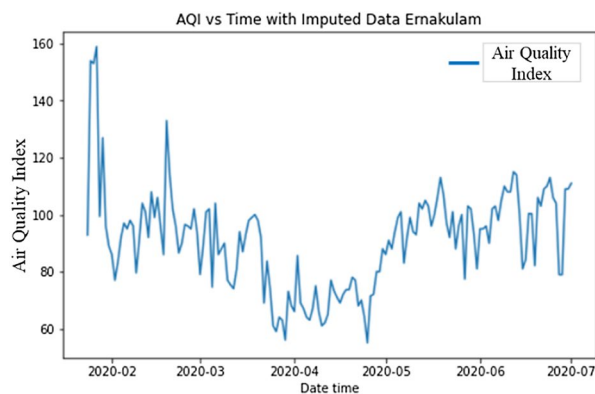
From Table 6, it is clear that the proposed algorithm is being evaluated with regression error and estimated concerning MAE of 0.0979, MSE with 0.0161, RMSE of 0.1271, MAPE with 0.0587 and R-Square with 0.8942, which exhibits an efficiency of the regression method through minimum loss acquired from the analysis. The medium level of AQI in Chennai city is being predicted and tabulated in Table 7.

From Table 7, it is clear that the prediction made on AQI in Chennai city acquires minimum loss values calculated with MSE as 0.0095, MAE with 0.0795, MAPE with 0.1219, R-Square with 0.9064 and RMSE with 0.0977. The minimum values of loss calculated from Chennai prediction data indicated the efficiency of the proposed regression method. Along with that, the regression analysis of the city of Ernakulam is tabulated in Table 8.

Fig. 10 a AQI Prediction. **b** AQI vs time in Ernakulam with Imputed data



(a) AQI Prediction



(b) AQI vs time in Ernakulam with Imputed data

Fig. 11 Proposed Prediction Efficiency

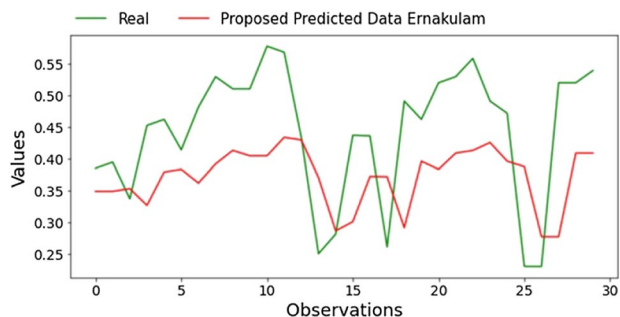


Table 5 Imputation for Ernakulam City

Ernakulam city	After imputation (0)	After imputation (mean)	After imputation (GAN)
MSE	0.42158	0.59412	0.0134
RMSE	0.6493	0.7712	0.1156
MAE	0.3658	0.4561	0.1013
R-square	0.4198	0.5186	0.9479
MAPE	0.3254	0.6497	0.1152

From Table 8, inference from loss values predicted in terms of MSE with 0.0134, MAE with 0.1013, MAPE with 0.1152, R-Square with 0.9479, and RMSE with 0.1156. Figure 12 depicts the graphical representation of the performance metrics for different cities.

Table 6 Performance evaluation for Ahmedabad city

Metrics	Ahmedabad
MAE	0.0979
MSE	0.0161
RMSE	0.1271
R-Square	0.8942
MAPE	0.0587

Table 7 Performance Efficiency of AQI Predicted in Chennai

Metrics	Chennai
MAE	0.0795
MSE	0.0095
RMSE	0.0977
R-Square	0.9064
MAPE	0.1219

Table 8 Performance evaluation of AQI Measure in Ernakulam

Metrics	Ernakulam
MAE	0.1013
MSE	0.0134
RMSE	0.1156
R-Square	0.9479
MAPE	0.1152

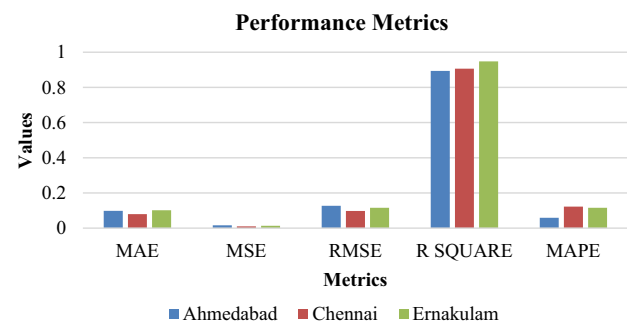
Fig. 12 Comparative Plot

Figure 12 showcases the performance metrics of obtained for different cities. Different metrics are used for assessing the efficacy of the model for precise and effective AQI.

4.3 Comparative analysis

The efficiency of the proposed methodology is exhibited by comparing the calculated loss values with the existing techniques. A comparison of different existing algorithms with the proposed algorithm is illustrated in the upcoming section.

4.3.1 Ernakulam

Different existing work has been compared with the proposed model for observing the loss rate in terms of AQI. Therefore, Table 8 depicts the MSE, RMSE and MAE value obtained by the existing SVR, ESVR, CNN and ECNN model along with proposed model.

From Table 9, comparing the loss obtained with regression calculation on lower AQI in Ernakulam by the proposed technique is found to be lower than considered SVR, CNN, ESVR, and ECNN. It demonstrates that the proposed regression algorithm is efficient in AQI prediction. The graphical representation of the comparison of algorithms is given in Fig. 13.

From comparing the loss acquired from the proposed with existing methods in Fig. 13, the proposed method is identified to be efficient in regression and yields effective prediction output.

4.3.2 Chennai

The existing research obtained AQI prediction for only Chennai city, which is being compared with the prediction values obtained in Chennai by the present research and tabulated in Table 10.

From Table 10, the inferred loss values on the prediction made from Chennai city with both proposed and existing regression methods exhibit that the proposed Stacked Attention-based GRU model with KL divergence has acquired minimum loss measured in terms of R², MAE, and RMSE. The graphical analysis is depicted in Fig. 14.

From Fig. 14, comparing loss calculated from existing and proposed regression algorithms in predicting AQI measures signifies that the proposed regression methodology is efficient in prediction with minimum loss values. Likewise, Table 11 shows the AQI value obtained using MAE, MSE, RMSE and R Square for Chennai.

From Table 11 it can be observed that, MAE, MSE, R square and RMSE obtained using the proposed model is better than the existing model for AQI prediction and the graphical representation is shown the Fig. 15.

4.3.3 Ahmedabad

Similarly, RMSE and R Square value for Ahmedabad is depicted in Table 12 for existing works like SARIMA, SVM-Linear, SVM-RBF, SVM-polynomial and LSTM.

From Table 12, it can be identified that, RMSE, R-Square value attained for Ahmedabad city is better than the existing model as the RMSE value obtained by the proposed model is 0.1271 and R square value obtained is 0.8942. This primarily

Table 9 Comparative analysis of Loss values of Stacked Attention GRU with KL Divergence [43]

Model	Technique	MSE	RMSE	MAE
Existing model	SVR	0.787	0.28	0.284
	ESVR	0.672	0.259	0.211
	CNN	0.723	0.268	0.312
	ECNN	0.401	0.2	0.243
Proposed model	Modified Stacked Attention GRU with modified KL divergence	0.0134	0.1156	0.1013

Fig. 13 Comparison of Loss Efficiency from Proposed with Existing Regression Algorithms [43]

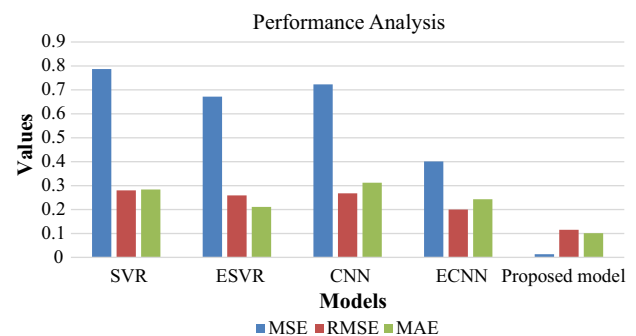


Table 10 Comparing the Performance of AQI predicted in Chennai [44]

AQI DATASET	MAE	RMSE	R SQUARE
Existing model for Chennai	0.8054	0.91538	0.9
Proposed model Chennai	0.0795	0.0977	0.9064

Fig. 14 Comparative analysis of Regression Loss Obtained in Chennai by Proposed and Existing models [44]

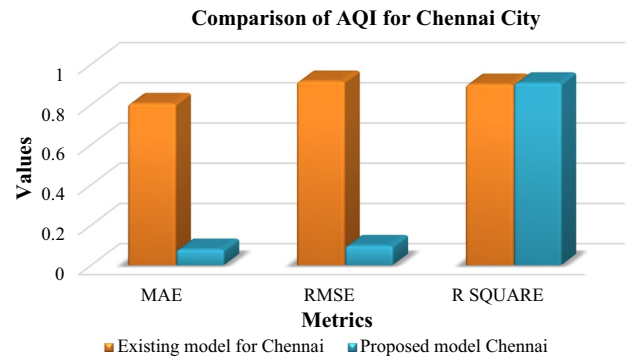


Table 11 Comparative analysis for Chennai [45]

Air Quality Index Dataset	MAE	MSE	RMSE	R SQUARE
Existing method	0.0478	0.1224	0.1365	0.886
Proposed method	0.0795	0.0095	0.0977	0.9064

Fig. 15 Comparative analysis

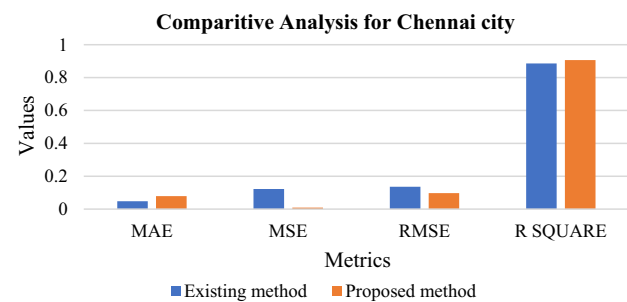


Table 12 Comparative analysis for Ahmedabad [46]

Method	Techniques	RMSE	R-square
Existing	SARIMA	29.75	0.824
	SVM—Linear	26.76	0.968
	SVM—RBF	4.94	0.998
	SVM—Polynomial	22.09	0.978
	LSTM	25.62	0.951
Proposed	Proposed model	0.1271	0.8942

due to the incorporation of Deep GAN imputation and regression using Stacked Attention GRU with KL divergence in the proposed mechanism. Figure 16 shows the graphical depiction of table.

With the assistance of proposed Deep GAN imputation and regression using Stacked Attention GRU with KL divergence, efficient prediction on AQI with minimum loss is obtained. The better level of prediction efficiency with lower loss is the advantage of the proposed system. From AQI measured using the proposed model, both government and private institutions can focus on initiating activities to improve air quality.

5 Discussion

Analytical outcome of the proposed mechanism shows that proposed work has delivered better performance than the existing works. Table 13 shows the overall values obtained by the proposed model using different loss rate metrics for different cities.

Fig. 16 Comparative analysis

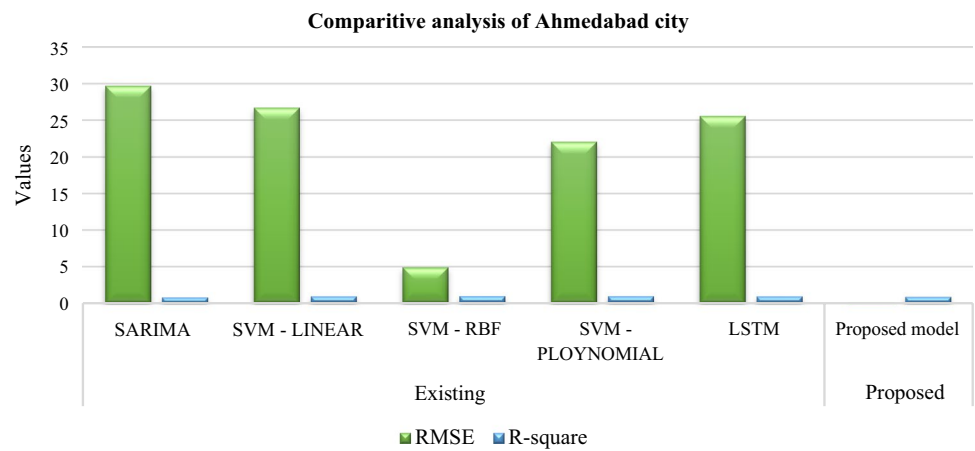


Table 13 Values of proposed model

AQI DATASET	MAE	MSE	RMSE	R SQUARE	MAPE
Ahmedabad	0.0979	0.0161	0.1271	0.8942	0.0587
Chennai	0.0795	0.0095	0.0977	0.9064	0.1219
Ernakulam	0.1013	0.0134	0.1156	0.9479	0.1152

RMSE value obtained by the proposed model is 83.1143% difference lower than SVR, 76.5617% than ESVR, 79.4578 than CNN, 53.4854% difference than ECNN. Likewise, MSE values of the proposed model is lower than existing models, thereby implying the efficacy of the proposed mechanism. Similarly, MAE value of the proposed model is, RMSE and R square obtained by the proposed model is 164.064% lower than existing model, likewise, RMSE value of the proposed model obtained is 161.425% lower than the prevailing model, eventually, R Square value of the proposed model is 0.7085%. Moreover, R square obtained by LSTM-GRU model [30] for AQI prediction is 0.87. However, R square attained by the model for AQI for Chennai city is 0.9064, Ernakulam 0.9479 and Ahmedabad is 0.8942, this shows that, R square of the proposed work is better than the existing LSTM-GRU for AQI. Further, SVR employed [12] in the study possess the limitation of not working effectively for large data, however, this limitation is primarily overcome by proposed work as it employs Deep GAN model and modified stacked attention GRU with KL divergence. Though different cities are covered in the proposed work, Chennai, Ahmedabad, Ernakulam have not covered in other studies, thereby making the paper distinct and diverse. Further, in order to depict the overall work of the proposed mechanism, a SWOT analysis has been provided in Table 14.

6 Conclusion

The proposed regression methodology utilized Deep GAN for imputing missing values in the dataset during preprocessing. With the processed input data, regression was performed using modified Stacked Attention GRU with KL divergence for predicting AQI in multiple cities of India. The regression process performed with the combined approach was evaluated to estimate the model's efficiency. From the experimental outcome, it was found that Ahmedabad was possessed with higher AQI in terms of MSE, RMSE, R Square, MAE, and MAPE with values 0.0161, 0.1271, 0.0152, 0.0979, and 1.3862, respectively and Ernakulam city was predicted with lowest AQI in terms of 0.0134, 0.1156, 0.2946, 0.1013 and 0.4839 respectively. The regression loss obtained was also compared with other traditional algorithms. As a result, the proposed AQI prediction was considered efficient since it achieved reduced error values than other existing regression algorithms. Despite the advantages of the proposed model, there are few limitations of the model which includes complexity of the model, as the model complexity can pose the challenges in terms of computational resources, besides, proposed model only focuses on predicting AQI in specific cities in India and may not be applicable to other regions with different environmental factors, this generalizing the findings top broader geographic area requires careful consideration of local factors and validation against relevant datasets. As future work, the present research can focus on predicting AQI after

Table 14 SWOT analysis

Strength	Weakness	Opportunities	Threats
<ul style="list-style-type: none"> Implementation of Deep GAN for missing data imputation improves data quality and ensures more robust predictions Incorporation of modified Stacked Attention GRU with KL divergence enhances the capability of the model to capture temporal dependencies and spatial relationships in the dataset Complete evaluation using numerous performance metrics offers a whole assessment of model effectiveness 	<ul style="list-style-type: none"> Dependency on the obtainability and quality of input data, which may vary across different monitoring stations and cities Complexity of the proposed model architecture may require substantial computational resources and expertise for implementation 	<ul style="list-style-type: none"> Potential for further refinement and optimization of model parameters to improve predictive accuracy Expansion of the study to include additional variables or features could enhance the model's predictive capabilities Application of the proposed methodology to other geographic regions or environmental datasets may yield valuable insights into air quality prediction 	<ul style="list-style-type: none"> Potential challenges related to data privacy and accessibility may limit the availability of comprehensive datasets for model training and validation External factors such as changes in environmental policies or regulations could impact the relevance and applicability of the proposed methodology over time

COVID 19 in India with the involvement of additional factors after air pollution, further associating diseases through air pollution will be focused on future.

Acknowledgements This project is sponsored by Prince Sattam Bin Abdulaziz University (PSAU) as part of funding for its SDG Roadmap Research Funding Programme project number PSAU-2023- SDG-86).

Author contributions The authors confirm contribution to the paper as follows: study conception and design: Binbusayyis and Khan; data collection: Ahmed and Emmanuel; analysis and interpretation of results: Ahmed and Emmanuel; draft manuscript preparation: Binbusayyis, Khan, Ahmed and Emmanuel. All authors reviewed the results and approved the final version of the manuscript.

Data availability The dataset used in our study can be obtained from <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>. The data generated during and/or analyzed during the current study are available from the corresponding author upon request.

Declarations

Competing interests The authors declare that there is no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Aram S, et al. Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. *Int J Environ Sci Technol.* 2024;21(2):1345–60.
2. Coccia M. The effects of atmospheric stability with low wind speed and of environment pollution on the accelerated transmission dynamics of COVID-19. *Int J Environ Stud.* 2020; 1802937.
3. Coccia MJSF. New directions of technologies pointing the way to a sustainable global society. *Sustain Futures.* 2023;5: 100114.
4. Coccia MJAPR. How do low wind speeds and high levels of air pollution support the spread of COVID-19? *Atmos Pollut Res.* 2021;12(1):437–45.
5. Mihăiță AS, Dupont L, Chery O, Camargo M, Cai CJJOCP. Evaluating air quality by combining stationary, smart mobile pollution monitoring and data-driven modelling. *J Cleaner Prod.* 2019;221:398–418.
6. Coccia MJSOTTE. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. *Sci Total Environ.* 2020;729:138474.
7. Lee M, et al. Forecasting air quality in Taiwan by using machine learning. *Sci Rep.* 2020;10(1):4153.
8. Jiang X, Wei P, Luo Y, Li YJA. Air pollutant concentration prediction based on a CEEMDAN-FE-BiLSTM model. *Atmosphere.* 2021;12(11):1452.
9. Wang J, Du P, Hao Y, Ma X, Niu T, Yang WJJOEM. An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. *J Environ Manag.* 2020;255:109855.
10. Li H, Wang J, Li R, Lu HJJOCP. Novel analysis–forecast system based on multi-objective optimization for air quality index. *J Cleaner Prod.* 2019;208:1365–83.
11. Kim D, Han H, Wang W, Kang Y, Lee H, Kim HSJAS. Application of deep learning models and network method for comprehensive air-quality index prediction. *Appl Sci.* 2022;12(13):6699.
12. Liu H, Li Q, Yu D, Gu YJAS. Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Appl Sci.* 2019;9(19):4069.
13. Wu Q, Lin HJSOTTE. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Sci Total Environ.* 2019;683:808–21.
14. Phruksahiran NJUC. Improvement of air quality index prediction using geographically weighted predictor methodology. *Urban Climate.* 2021;38: 100890.
15. Jumaah HJ, Ameen MH, Kalantar B, Rizzei HM, Jumaah SJJG. Air quality index prediction using IDW geostatistical technique and OLS-based GIS technique in Kuala Lumpur, Malaysia. *Geomat Nat Haz Risk.* 2019;10(1):2185–99.
16. Maleki H, et al. Air pollution prediction by using an artificial neural network model. *Clean Technol Environ Policy.* 2019;21:1341–52.
17. Janarthanan R, Partheeban P, Somasundaram K, Elamparithi PNJSC. A deep learning approach for prediction of air quality index in a metropolitan city. *Sustain Cities Soc.* 2021;67:102720.
18. Ma J, Li Z, Cheng JC, Ding Y, Lin C, Xu ZJSOTTE. Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. *Sci Total Environ.* 2020;705:135771.
19. Insani F, Sari APJJOAI. Optimzation of interval fuzzy time series with particle swarm optimization for prediction air quality on Pekanbaru. *Ind J Artif Intell Data Mining.* 2020;3(1):36–41.
20. Kumar K, Pande BJJOES. Air pollution prediction with machine learning: a case study of Indian cities. *Int J Environ Sci Technol.* 2023;20(5):5333–48.
21. Shang K, et al. Haze prediction model using deep recurrent neural network. *Atmosphere.* 2021;12(12):1625.

22. Alkabbani H, Ramadan A, Zhu Q, Elkamel AJA. An improved air quality index machine learning-based forecasting with multivariate data imputation approach. *Atmosphere*. 2022;13(7):1144.
23. Ding W, Zhu YJA. Prediction of PM_{2.5} concentration in Ningxia Hui autonomous region based on PCA-attention-LSTM. *Atmosphere*. 2022;13(9):1444.
24. Yan R, Liao J, Yang J, Sun W, Nong M, Li FJESWA. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Syst Appl*. 2021;169:114513.
25. Ho JY, et al. Towards a time and cost effective approach to water quality index class prediction. *J Hydrol*. 2019;575:148–65.
26. Fan J, Wu L, Ma X, Zhou H, Zhang FJRE. Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. *Renew Energy*. 2020;145:2034–45.
27. Sun Y, Liu JJS. AQI prediction based on CEEMDAN-ARMA-LSTM. *Sustainability*. 2022;14(19):12182.
28. Leong W, Kelani R, Ahmad ZJJOECE. Prediction of air pollution index (API) using support vector machine (SVM). *J Environ Chem Eng*. 2020;8(3):103208.
29. Aarathi C, Ramya VJ, Falkowski-Gilski P, Divakarachari PBJs. Balanced spider monkey optimization with Bi-LSTM for sustainable air quality prediction. *Sustainability*. 2023;15(2):1637.
30. Sarkar N, Gupta R, Keserwani PK, Govil MCJEP. Air quality index prediction using an effective hybrid deep learning model. *Environ Pollut*. 2022;315: 120404.
31. Hossain E, Shariff MA, Hossain MS, Andersson K. A novel deep learning approach to predict air quality index. *Proc Int Conf Trends Comput Cogn Eng*. 2020. https://doi.org/10.1007/978-981-33-4673-4_29.
32. Liu B, et al. A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. *IEEE Access*. 2019;7:43331–45.
33. Wu Q, Lin H. Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. *Sustain*. 2019;50:101657.
34. Mao W, Wang W, Jiao L, Zhao S, Liu A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustain Cities Soc*. 2021;65:102567.
35. Zhang L, Liu P, Zhao L, Wang G, Zhang W, Liu JJAPR. Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmos Pollut Res*. 2021;12(1):328–39.
36. Mishra A, Gupta YJSIR. Comparative analysis of air quality index prediction using deep learning algorithms. *Spat Inf Res*. 2024;32(1):63–72.
37. Praveen KJTJJOA and EM analysis. Air quality index forecasting via genetic algorithm-based improved extreme learning machine. *IEEE access*. (2023) vol. 26, no. 1, pp. 17–27, 2024.
38. Lu H, Ma XJC. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*. 2020;249: 126169.
39. Gunasekar S, Kumar GJ, Vijayakumar K, Agbulu GP. Development of hybrid artificial neural network and decision tree algorithm for air quality prediction. *J Innov Technol Exploring Eng*. 2021.
40. Xayasouk T, Lee H, Lee GJS. Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability*. 2020;12(6):2570.
41. Parkavi P and Rathi S. LSTM Based Air Quality Prediction Model In Smart Cities. *IEEE international conference on big data (big data) 2021*.
42. Al-Janabi S, Mohammad M, Al-Sultan AJSC. A new method for prediction of air pollution based on intelligent computation. *Soft Comput*. 2020;24(1):661–80.
43. Zahid M, et al. Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids. *Electronics*. 2019;8(2):122.
44. Mani G, Viswanadhapalli JKJJOER. Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models. *J Eng Res*. 2022;10(2A):179–94.
45. Natarajan SK, Shanmurthy P, Arockiam D, Balusamy B, Selvarajan SJSR. Optimized machine learning model for air quality index prediction in major cities in India. *Sci Rep*. 2024;14(1):6795.
46. Maltare NN, Vahora SJDCE. Air quality Index prediction using machine learning for Ahmedabad city. *Digit Chem Eng*. 2023;7: 100093.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.