Received: 19 July 2023 | Revised: 22 September 2023 | Accepted: 3 November 2023

DOI: 10.4218/etrij.2023-0283

ORIGINAL ARTICLE

ETRI Journal WILEY

Air quality index prediction using seasonal autoregressive integrated moving average transductive long short-term memory

Subramanian Deepan | **Murugan Saravanan**

Department of Networking and Communications, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Correspondence
Murugan Saravanan, Department of Networking and Communications, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.
Email: saravanm7@srmist.edu.in

Funding information
This research received no external funding.

Abstract

We obtain the air quality index (AQI) for a descriptive system aimed to communicate pollution risks to the population. The AQI is calculated based on major air pollutants including O_3, CO, SO_2, NO, NO_2, benzene, and particulate matter PM2.5 that should be continuously balanced in clean air. Air pollution is a major limitation for urbanization and population growth in developing countries. Hence, automated AQI prediction by a deep learning method applied to time series may be advantageous. We use a seasonal autoregressive integrated moving average (SARIMA) model for predicting values reflecting past trends considered as seasonal patterns. In addition, a transductive long short-term memory (TLSTM) model learns dependencies through recurring memory blocks, thus learning long-term dependencies for AQI prediction. Further, the TLSTM increases the accuracy close to test points, which constitute a validation group. AQI prediction results confirm that the proposed SARIMA–TLSTM model achieves a higher accuracy (93%) than an existing convolutional neural network (87.98%), least absolute shrinkage and selection operator model (78%), and generative adversarial network (89.4%).

KEYWORDS
air pollutant, air quality index, seasonal autoregressive integrated moving average, time-series data, transductive long short-term memory

1 | INTRODUCTION

According to the World Health Organization, the guidelines for enhancing air quality are not being followed in 98% of urban areas in developing regions. Air quality can be reduced by various factors, such as the increasing use of motorized vehicles and fossil fuels, which may deplete the environment under various conditions [1]. Thus, severe consequences for human health may occur depending on the exposure time to high levels of air pollution [2]. The air quality index (AQI) is an important parameter because it allows to intuitively communicate air quality to the population. The AQI transforms the concentrations of various pollutants in the air with complex variations into a comprehensive value [3]. Various methods have been used to predict air quality across various cities of India, but they have limited functionality because they are costly and labor intensive [4].

Pollution levels are drastically increasing, especially in metropolitan areas, considerably affecting the environment. Thus, to establish a comfortable and healthy environment, pollution levels must be minimized [5]. Various factors influence air pollution [6]. Pollutants usually come from distinct domains such as transportation services, daily traffic, thermal power plants, garbage, home appliances, and hospitals [7]. High levels of air pollution harm animals, humans, and plants. For example, they generate new cases of respiratory diseases and affect the quality of crops and their overall production [8]. Therefore, under varying air quality, the pollution levels should be accurately monitored [9].

Time-series prediction has proven to be difficult when classical regression is performed depending on its complexity [10]. Deep learning models, including long short-term memory (LSTM), are often used to predict time series after learning features from representative data sequences to perform a task. Statistical approaches, such as autoregressive integrated moving average (ARIMA), seasonal ARIMA (SARIMA), and autoregressive models, are intended to detect periodicities [11–13].

It is necessary to apply forecasting and time-series approaches to air pollutant information collected by organizations such as the Central Pollution Control Board (CPCB) in India [14–17]. In this study, we analyzed the air quality in Indian cities, including Aizawl, Ahmedabad, Amritsar, Amaravati, Bengaluru, Brajrajnagar, Bhopal, Chandigarh, Coimbatore, Chennai, Delhi, Ernakulam, Patna, Shillong, Gurugram, Guwahati, Jaipur, Hyderabad, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Talcher, Thiruvananthapuram, and Visakhapatnam, for evaluating and predicting pollutant levels.

The contributions of this study are summarized as follows:

- SARIMA is used to predict values from historical data while accounting for seasonal changes.
- A transductive LSTM (TLSTM) learns dependencies using repeated memory segments, allowing to learn even long-term dependencies. TLSTM cells are employed for data-sequence learning and prediction with variable lengths.
- The integrated SARIMA–TLSTM model predicts the amount of air pollutants considering seasonality, possibly providing valuable information to the population.

The remainder of this paper is organized as follows. Section 2 describes existing methods for AQI prediction. Section 3 details the steps and methods devised in this study. Section 4 presents evaluation and comparison results of the proposed method. Finally, we draw conclusions and outline directions of future work in Section 5.

2 | RELATED WORK

Various prediction methods for AQI and harmful gases have been developed. In this section, we discuss their advantages and drawbacks.

Chauhan and others [18] developed a convolutional neural network (CNN) to contribute to building a sustainable urban environment and improve air quality forecasting. The model comprised preprocessing and data analysis as well as testing the model accuracy. However, CNN construction and large datasets were required to forecast patterns for knowledge discovery. Sethi and Mittal [19] developed a regression model based on the adaptive least absolute selection and shrinkage operator (LASSO) to improve the AQI prediction efficiency. However, the developed model was limited because the AQI was altered by additional temporal features, which required the hour of the day as another variable, and additional predictors must be added to the temporal factors for further analysis.

To improve air quality, Kothandaraman and others [20] introduced a method for intelligent air quality forecasting and pollution prediction using machine learning. However, alternative datasets and improved computational techniques must be used for prediction. Saha and others [21] developed an improved temporal prediction method by labeling time series based on matrix profiles and motifs. The model was simple and produced predictions that were comparable to those of sophisticated LSTM models, while time-series labeling reduced complexity.

Abirami and Chitra [22] developed an autoencoder-based generative adversarial network (GAN) to perform regional spatiotemporal forecasting of particulate matter. However, to evaluate prediction, the uncertainty of deep learning must be unveiled. Janarthanan and others [23] developed a deep learning method to forecast the AQI levels in cities including Chennai. Gray-level co-occurrence matrix features were combined with a support vector regression (SVR)–LSTM model to predict the AQI value. On the other hand, the AQI predictions did not identify pollutant-sensitive areas.

Cong and others [24] developed an LSTM model for feature awareness to predict concentrations of pollutant gases, including the common pollutants. The concentration data were collected in the field, and the environmental parameters were related to the prediction of multiple pollutant gas concentrations. However, the learning rate was high and few iterations for training were considered, leading to a continuous oscillation between normal and optimal convergence. By employing artificial intelligence, Gokul and others [25] conducted a spatiotemporal air quality study and predicted PM2.5 concentrations in Hyderabad, India. Various machine learning models, including multilinear regression, decision tree (DT), *k*-nearest neighbors, random forest, and extreme gradient boosting, were used to predict PM2.5 concentrations. This approach failed to identify any grouping of characteristics with clustered values, tested the set, and selected the closest value in a real-time setting.

Surono and others [26] used *k*-means clustering and principal component analysis for dimensionality reduction to construct an optimized RNN for AQI prediction. As there was no established method for choosing the parameter, experimentation on the data and model was required. The parameters of earlier studies could be used for guidance, but an experiment was required to determine suitable settings. Similarly, Ni and others [27] developed a deep belief network based on a CNN to predict toxic gas dispersion. For various evaluation indices, the CNN achieved the highest performance among various models, but it required a larger inference time.

A summary of existing work is provided in Table 1. LSTM has been widely used for prediction. Nevertheless, several factors affect pollution when attempting to predict air quality. We analyze the proposed SARIMA–TLSTM model in various regions to evaluate the AQI. TLSTM is widely used to process data sequences and generate predictions. It is a type of RNN that excels in resolving various problems. Compared with previous models, it selects new information substantially faster and enhances inference. Furthermore, it completes complex and long-term tasks that cannot be performed by existing learning algorithms.

2.1 | Research objective

Air pollution has increased significantly in recent years. Each year, hundreds of fatalities are related to air pollution, which threatens human health and the environment. In addition to contributing to global warming and greenhouse effects, pollution worsens respiratory conditions such as asthma and lung cancer. The AQI quantifies the degree of air pollution, and deep learning can contribute to its prediction. We aimed to create and train a model using deep learning and determine which model is more accurate in predicting the AQI. Experiments were conducted to determine the most reliable deep learning model for AQI prediction.

3 | PROPOSED AQI PREDICTION METHOD

We consider major air pollutants, such as O_3, CO, SO_2, NO, NO_2, benzene, and PM2.5 to compute the AQI. From time-series data, the automated evaluation of air pollutants using deep learning can reveal a sequence that captures datapoints in constant time intervals. The SARIMA model is used to predict future values from historical values considering seasonal trends. In addition, a TLSTM model reflects long-term dependencies and learns dependencies when memory segments are repeated. Figure 1 shows a diagram of the proposed method.

3.1 | Data collection

The data are publicly available on the CPCB website at https://cpcb.nic.in/. To predict the AQI, data were gathered from various parts of Chennai, Manali, Alandur, and Velachery. The measured pollutants in the dataset were PM2.5, PM10, NO, NO_2, NOx, NH_3,

TABLE 1 Summary of related studies.

Study	Method	Drawbacks
[18]	CNN	Fixed; an additional model is required to handle large datasets
[19]	LASSO	More predictors are required to account for temporal factors
[20]	Linear regression, random forest, *k*-nearest neighbors, reinforcement learning, extreme gradient boosting, and Adab models	Needs additional computational techniques for testing
[21]	Time-series labeling	Time complexity issues
[22]	GAN	Prediction uncertainty
[23]	SVR–LSTM	AQI predictions do not identify pollutant-sensitive areas
[24]	LSTM	Large learning rate leads to oscillation in training
[25]	LSTM	Cannot identify clusters of important characteristics in real time
[26]	Optimized RNN	Not accurate for similar datasets during comparison
[27]	Deep belief network	Long computation time

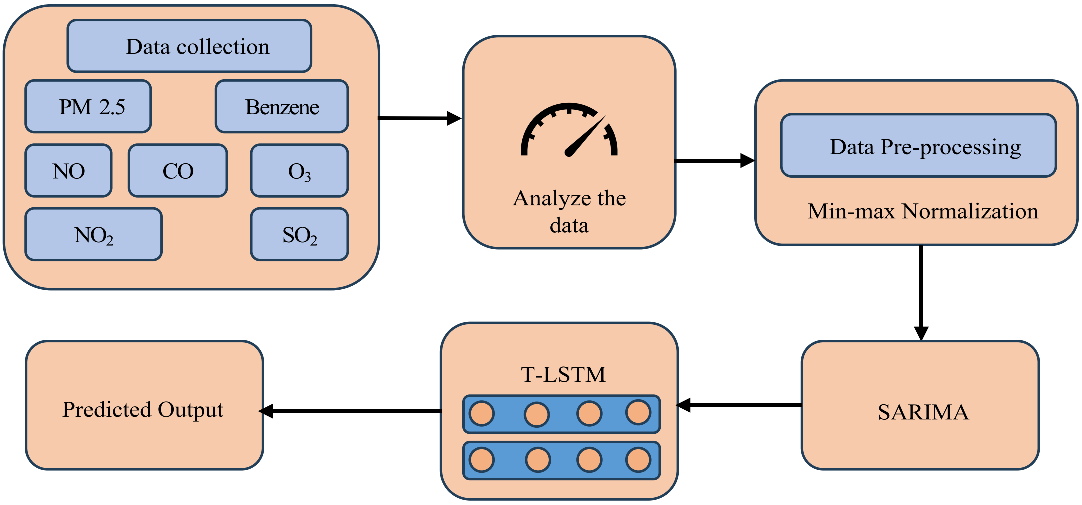


FIGURE 1 Block diagram of proposed hybrid feature selection method for AQI prediction.

CO, and SO_2. The data were measured every 15 min from May 1, 2019, to April 30, 2020. By excluding missing values, 22 827 data rows remained in the dataset. PM2.5, NO_2, SO_2, CO, and ozone were meteorological features gathered from three stations at intervals of 15 min in Chennai. The data were collected from May 1, 2019, to April 30, 2020, given the data availability at all the stations, obtaining 35 039 datapoints with 490 546 data rows.

3.2 | Classification using SARIMA–TLSTM model

The SARIMA model analyzes data to identify patterns by modeling relations between past and present values in a time series. To capture patterns and seasonality of data, SARIMA combines autoregression, moving average, and differencing models. Then, TLSTM predicts the AQI by considering the sources of pollution, weather, and temporal information. The SARIMA–TLSTM model analyze impact patterns from several event categories and accept inputs of various lengths. Hence, the SARIMA–TLSTM model is suitable for AQI prediction.

3.2.1 | ARIMA model

The ARIMA model is used for prediction on univariate time-series data [28] using moving average and an autoregressive model. However, the ARIMA model does not support seasonal data, that is, the time at which the time series starts a repeating cycle. On the other hand, the SARIMA model [29] extends ARIMA also for univariate data.

3.2.2 | SARIMA model

The SARIMA model captures seasonal effects and adjusts the predictions accordingly. The SARIMA parameters are required to provide two types of orders, similar to ARIMA with explanatory variable, represented as (p, d, q). The SARIMA model can be expressed as

$$\Phi_p(L)\overline{\phi}_P\left(L^S\right)\Delta^d\Delta_s^D y_t = A(t) + \theta_q(L)\overline{\theta}_Q\left(L^S\right)\epsilon_t, \qquad (1)$$

where $\Phi_p(L)$ is a non-seasonal autoregressive lag polynomial, $\overline{\phi}_P\left(L^S\right)$ is a lag polynomial, which is a seasonal autoregressive parameter, $\Delta^d\Delta_s^D y_t$ is a time-series expression that is differenced with d times seasonally up to D times, $A(t)$ is a trend polynomial included as the intercept, $\theta_q(L)$ is a non-seasonal parameter, b, which is evaluated based on the moving average, and $\overline{\theta}_Q\left(L^S\right)$ is a seasonal moving average lag polynomial.

3.3 | TLSTM model

TLSTM model is used to train a deep neural network to classify a data sequence [30]. The LSTM output is denoted as h_t, and h_{t-1} is the previous cell output. In LSTM, b_c is a bias, $\tilde{c}_t$ represents the memory cell, and W_c is a weight matrix. These elements are related by

$$\tilde{c}_t = \tanh(W_c.[h_{t-1}, x_t] + b_c). \qquad (2)$$

The input gate is denoted as i_t, and the current input manages a state value with the help of i_t. In addition, b_i is a bias, W_i is a weight matrix, and σ denotes the sigmoid function. The input gate is expressed as.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i). \tag{3}$$

The forgetting the gate is denoted as f_t and assesses the memory based on the predictions. This gate is updated as follows:

$$f_t = \sigma\left(W_i.[h_{t-1}, x_t] + b_f\right). \tag{4}$$

The memory cell is denoted as c_t and its previous state is denoted as c_{t-1}. The cell is expressed as

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t, \tag{5}$$

where * denotes the dot product. o_t denotes the output gate, which is computed by the memory cell state as Equation (6). Output h_t of the LSTM model is given by Equation (7):

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0). \tag{6}$$

$$h_t = o_t * \tanh(c_t). \tag{7}$$

Hidden unit function $\overrightarrow{h}$ of a hidden forward layer at each timestep t is calculated based on current input x_t and previous hidden state h_{t-1}. Hidden unit function $\overleftarrow{h}$ of a hidden backward layer is calculated using x_t and future hidden state $\overleftarrow{h}_{t+1}$. Forward and backward expressions are created through $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$, which are concatenated into a vector [31].

3.3.1 | TLSTM model for time-series prediction

TLSTM allows process specific data sequences and make predictions, which has been used in various applications. Figure 2 shows the structure of the TLSTM model.

TLSTM learns faster and provides better inference than similar models. In addition, it solves complicated tasks that others RNNs cannot handle [32]. Hence, all linear models can rely on the test point [33]. Let $Z^{(\eta)}$ be a hidden state. The state-space model of the TLSTM model is given by Equation (8):

$$\begin{cases} C_{t,\eta} = f\left(C_{t-1,\eta}, h_{t-1,\eta}, x_t; \omega_{\text{lstm},\eta}, b_{\text{lstm},\eta}\right) \\ h_{t,\eta} = g\left(h_{t-1,\eta}, \ c_{t-1,\eta}, x_t; \omega_{\text{lstm},\eta}, b_{\text{lstm},\eta}\right) \end{cases}. \tag{8}$$

The linear models rely on the newly provided datapoint, $Z^{(\eta)}$, where η identifies the datapoint. Prediction is expressed as follows (9) by using a dense layer:

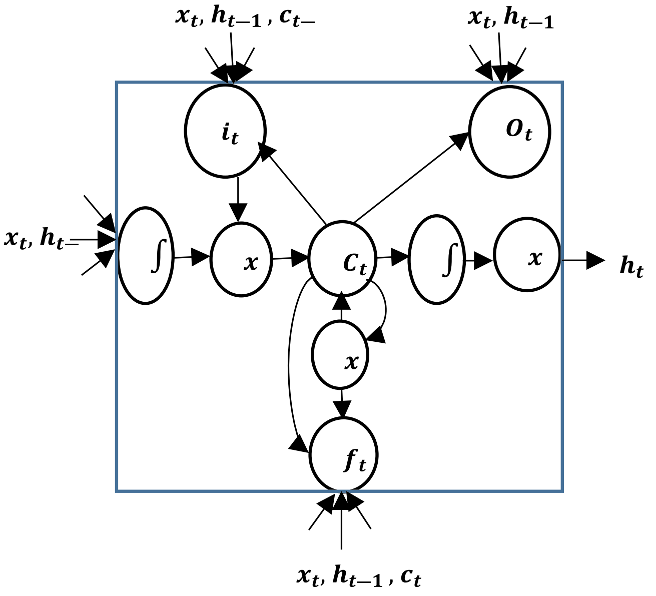


FIGURE 2 Structure of TLSTM cell.

$$\widehat{y}_\eta^{(t)} = \omega_{\text{dense},\eta}^T h_{t+T-1,\eta} + b_{\text{dense},\eta,} \ t = 1, ..., N, \tag{9}$$

where $\omega_{\text{dense},\eta}^T \in \mathbb{R}^{n\times 1}$ and $b_{\text{dense},\eta,} \in \mathbb{R}$ represent weights and bias, respectively.

To determine the needs on a fresh hidden point, consider $s_{t,\eta}$ as the resemblance among T, $Z^{(t)}$, $Z^{(\eta)}$, ω_η, and b_η, denoting every constraint in $(\omega_{\text{lstm},\eta,} \omega_{\text{dense},\eta})$ and $(b_{\text{lstm},\eta,} b_{\text{dense},\eta})$. The objective function is expressed as

$$\left(\widehat{\omega}_{\text{lstm},\eta}, \widehat{\omega}_{\text{dense},\eta}, \widehat{b}_{\text{lstm},\eta}, \widehat{b}_{\text{dense},\eta}\right) = \left(\widehat{\omega}_\eta, \ \widehat{b}_{\eta,}\right) = \arg\min_{\omega_\eta, b_{\eta,}} \text{J}_\eta$$
$$J_\eta = \frac{1}{N}\sum_{t=1}^{N} S_{t,\eta}\left(\widehat{y}_\eta^{(t)} - y^{(t)}\right)^2 + \Upsilon_\eta \omega_\eta^T \omega_\eta \tag{10}$$

where $S_{t,\eta} \in \mathbb{R}^+$.

Υ_η in Equation (10) is a tuning parameter, and the number of neurons in the LSTM gates is determined as a tuning parameter for the transductive method. The TLSTM parameters depend on $Z^{(\eta)}$. Therefore, each unseen subsample is rehabilitated. Hence, constraints $\widehat{\omega}_{\eta,}$ and $\widehat{b}_{\eta,}$ are diverse across test points.

On the other hand, when the model can be retrained, it is suitable for time-series prediction. The TLSTM model is also appropriate when the relations between factors in various areas of the input space differ.

For $Z^{(\eta)}$, the hidden state update is described as

$$h_{t',\eta} = \text{g}\left(h_{t'-1,\eta}, \ c_{t'-1,\eta,Z_{t'}^{(\eta)}}; \ \widehat{\omega}_{\text{lstm},\eta}, \widehat{b}_{\text{lstm},\eta,}\right), \tag{11}$$

where $t' = \eta$, ..., $(\eta + T - 1)$. The prediction is given by

$$\widehat{y}_{\eta}^{(\eta)} = \widehat{y}_{\text{dense},\eta}^{T} h_{\eta+T-1,\eta} + \widehat{b}_{\text{dense},\eta} 0. \tag{12}$$

$Z^{(\eta)}$ is the input. Still, the hidden state update and prediction in TLSTM is based on new point $Z^{(\eta)}$, where the model constraints are optimized between the training points and $Z^{(\eta)}$ for points η related to $h_{t',\eta}$ and $\widehat{y}_{\eta}^{(\eta)}$ in Equations (11) and (12). A flowchart of the proposed method is shown in Figure 3.

4 | RESULTS AND DISCUSSION

The proposed method was implemented using the Python application programming interface interlinked by a local server running Windows 10 Pro. The server was equipped with 16 GB NVIDIA and Geo-force graphics processors and an Intel i9 processor operating at 2.5 GHz. To predict the AQI, the levels of noise were also estimated. The results were obtained using a dataset of air quality in neighboring Indian cities.

4.1 | Experimental data

The dataset was collected from three stations located across Chennai (Tamil Nadu, India) around **13.067439°N in** latitude **and 80.237617°E in longitude**. Weather forecasting factors such as SO_2, PM2.5, CO, NO_2, and ozone were measured every few minutes. The data were collected from May 1, 2019, to April 30, 2020, at each station, obtaining 490 546 rows and columns of information in the datasets over the 1-year study period.

4.2 | Performance measures

We considered various measures to evaluate the performance of the proposed method.

Accuracy reflects the prediction effectiveness of a machine learning model by

$$\text{Accuracy}\,(\%) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100, \tag{13}$$

where TP, TN, FP, and FN represent the rates of true positives, true negatives, false positives, and false negatives, respectively.

Precision is the ratio of predicted positive observations to the total number of predicted positive observations:

$$\text{Precision}\,(\%) = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100. \tag{14}$$

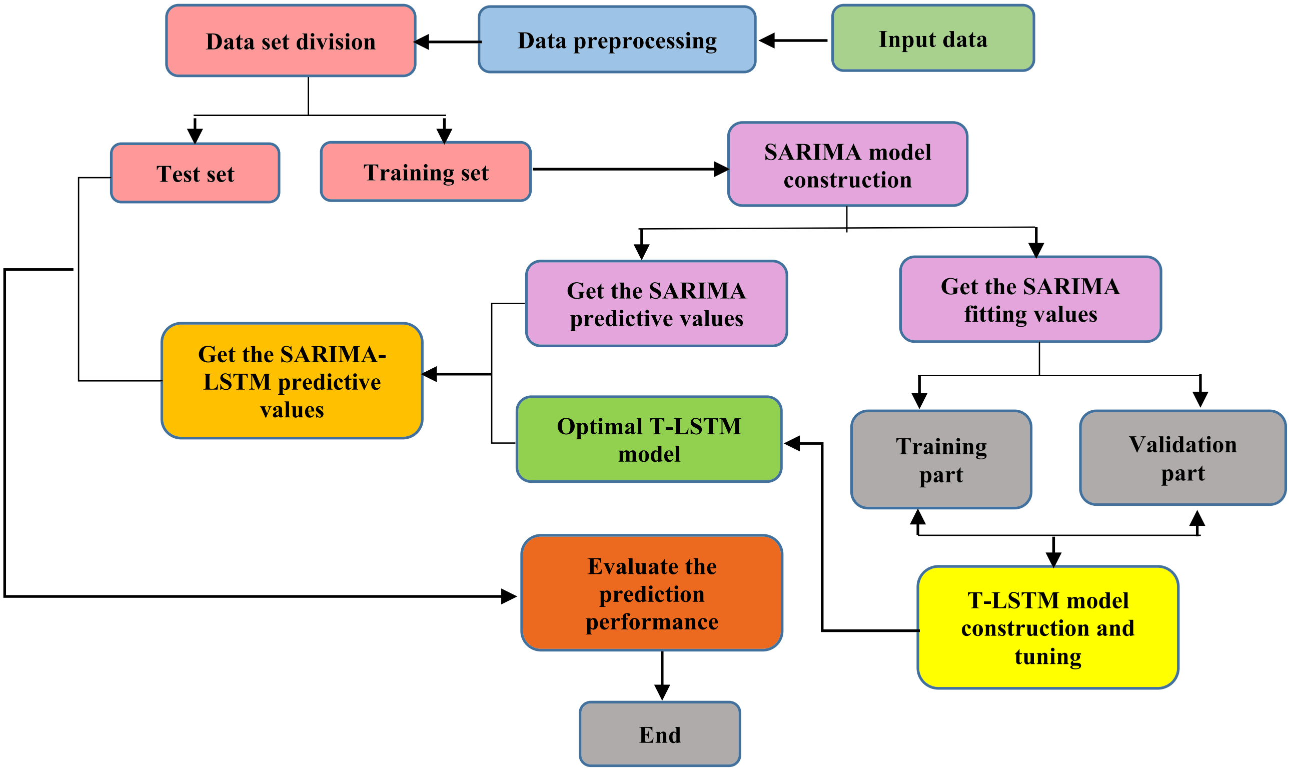

Data set division
Data preprocessing
Input data
Test set
Training set
SARIMA model construction
Get the SARIMA predictive values
Get the SARIMA fitting values
Get the SARIMA-LSTM predictive values
Optimal T-LSTM model
Training part
Validation part
Evaluate the prediction performance
T-LSTM model construction and tuning
End

FIGURE 3 Flowchart of proposed method for predicting AQI.

Recall is the sum of true positives TP that are present across the classes divided by the true positives and false negatives for all the classes:

$$\text{Recall}\,(\%) = \frac{\text{TN}}{\text{TP}+\text{FN}} \times 100. \tag{15}$$

The F1-score combines the recall and precision as follows:

$$F1-score(\%) = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100. \tag{16}$$

The root-mean-square error (RMSE) is calculated by taking the square root of the result and averaging the squared discrepancies between the actual and predicted values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\left(\sum\nolimits_{j=1}^{n}\left(y_j - y_j'\right)^2\right)}. \tag{17}$$

The coefficient of determination, R^2, provides the square of the prediction errors divided by the total sum of squares, replacing predictions with the mean value as follows:

$$R^2 = 1 - \frac{\sum_i \left(y_j - y_j'\right)^2}{\sum_i \left(y_j - \overline{y_j}\right)^2}. \tag{18}$$

The MAE is given by the average of the absolute discrepancies between the actual and predicted observations in a test sample:

$$\text{MAE} = \frac{1}{n}\sum\nolimits_{j=1}^{n}\left(y_j - y_j'\right), \tag{19}$$

where n is the number of observations, y_j is the actual value, $\overline{y_j}$ is the mean value, and y_j' is the predicted value. To estimate the PM2.5 concentration and predict the air quality, we use

$$\text{AQR} = \frac{\text{AQR}_{\text{high}} + \text{AQR}_{\text{low}}}{\text{PC}_{\text{high}} - \text{PC}_{\text{low}}}\left(\text{PC} - \text{PC}_{\text{low}}\right) + \text{AQR}_{\text{low}}, \tag{20}$$

where AQR is the air quality range, PC is the pollutant concentration, PC_{low} is the concentration break point below PC, PC_{high} is the concentration break point above PC, AQR_{low} is the AQR break point corresponding to PC_{low}, and AQR_{high} is the AQR break point corresponding to PC_{high}.

4.3 | Quantitative analysis

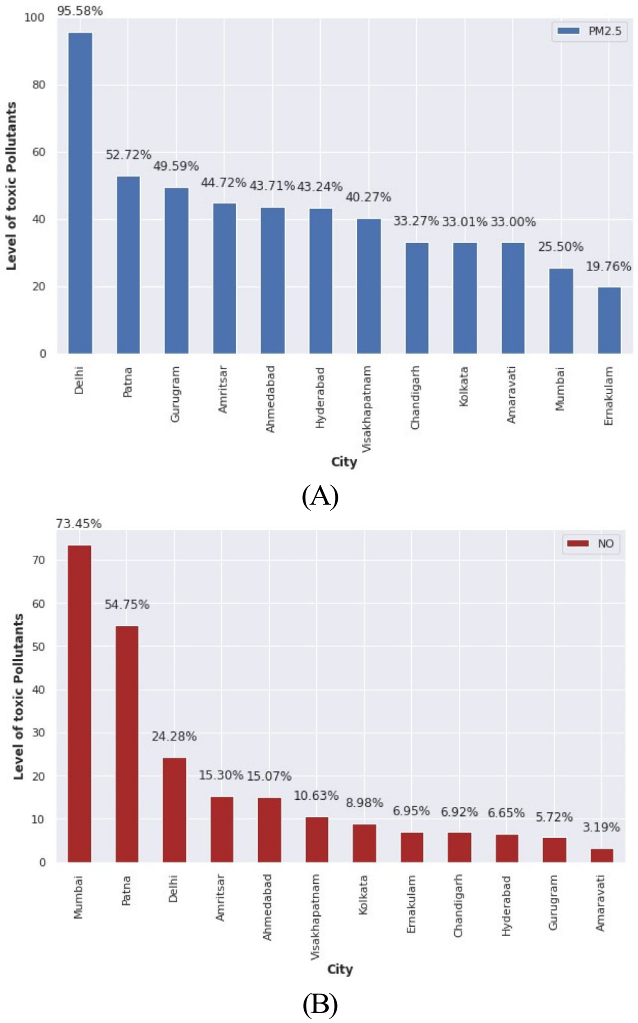
The evaluation measures for various classifiers, including naïve Bayes, LASSO–naïve Bayes, CbAL–naïve Bayes, DT, LASSO–DT, CbAL–DT, ANN, LASSO–ANN, CbAL–ANN, SARIMA, LSTM, SARIMA–LSTM, and SARIMA–TLSTM are listed in Tables 2 and 3. To locate and predict the AQI, different combinations of classifiers and deep learning models were tested, but the models performed below expectations. Tables 2 and 3 show that proposed

TABLE 2 Recall and F1-score of evaluated prediction models.

Parent classifier	Classifier	Recall (%)	F1-score (%)
Naïve Bayes	Naïve Bayes	54	54
	LASSO–naïve Bayes	57	56
	CbAL–naïve Bayes	63	61
DT	DT	57	56
	LASSO–DT	67	67
	CbAL–DT	70	70
ANN	ANN	49	25
	LASSO–ANN	50	29
	CbAL–ANN	63	33
Deep learning models	SARIMA	71	70
	LSTM	87	87
	SARIMA–LSTM	92	92
	SARIMA–TLSTM	94	96

TABLE 3 Accuracy and precision of evaluated prediction models.

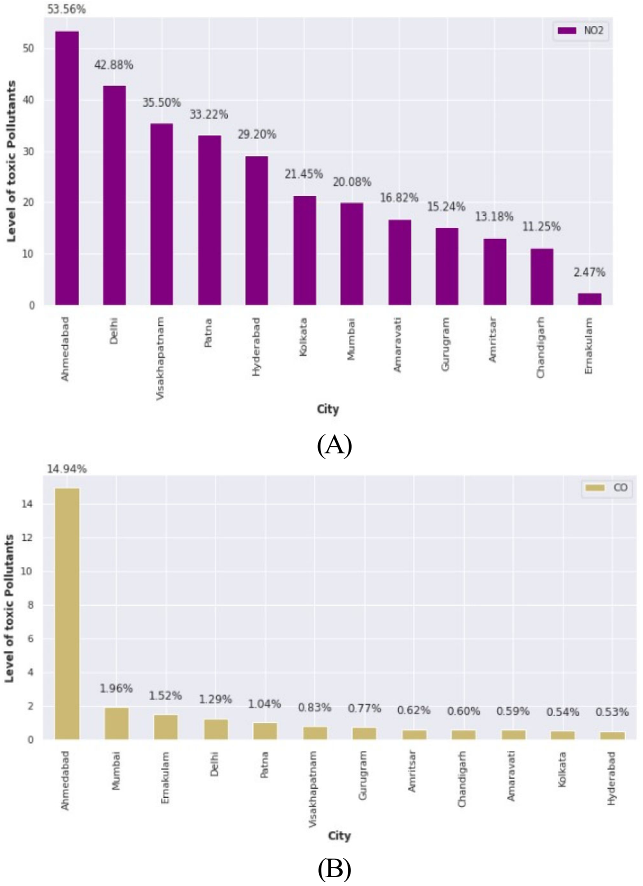
Parent classifier	Classifier	Accuracy (%)	Precision (%)
Naïve Bayes	Naïve Bayes	69	55
	LASSO–naïve Bayes	70	57
	CbAL–naïve Bayes	74	60
DT	DT	70	57
	LASSO–DT	80	67
	CbAL–DT	83	70
ANN	ANN	65	31
	LASSO–ANN	69	38
	CbAL–ANN	77	45
Deep learning models	SARIMA	85	70
	LSTM	88	87
	SARIMA–LSTM	91	93
	SARIMA–TLSTM	93	95



(A)

(B)

FIGURE 4 Pollutants levels: (A) PM2.5 and (B) NO.



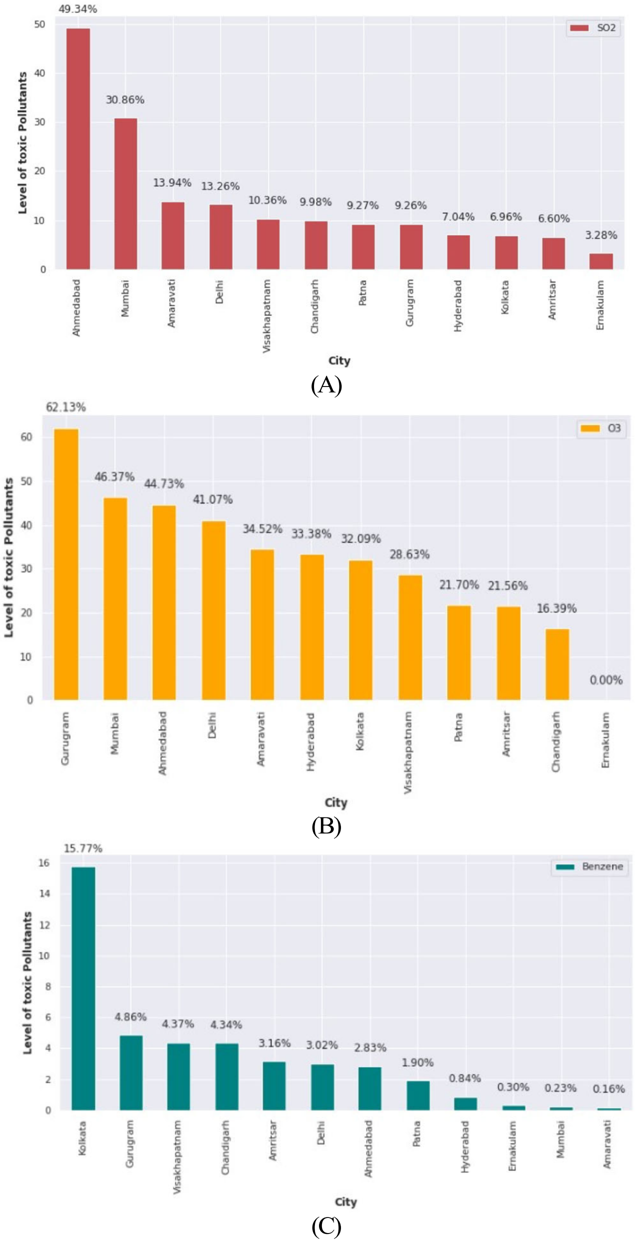
(A)

(B)

FIGURE 5 Pollutants levels: (A) NO_2 and (B) CO.

SARIMA–TLSTM model achieves better results in all the performance measures.

The pollutant level analyses for PM2.5 and NO are shown in Figure 4A,B, respectively. The TLSTM model predicts the AQI based on the moving average values obtained from SARIMA. The model was trained to predict the values correctly. For PM2.5, the cities of Emakullan and Delhi provided toxic pollutant levels of 20 and 95, respectively. Similarly, the TLSTM model provided the AQI for NO, as shown in Figure 4B, which shows the predicted pollutant levels. Amaravati provided five AQI levels with lower pollutant concentrations, and Mumbai obtained 75 levels with the highest pollutants levels. Hence, the proposed model performed well in fitting, training, and testing. The obtained values for R^2 were evaluated for NO_2. The observed and predicted data for NO_2 and CO are shown in Figure 5. The proposed model suitably fits the training and test data, as shown in Figure 5A,B for NO_2 and CO, respectively. The analysis of the levels of pollutants SO_2, O_3, and benzene are



(A)

(B)

(C)

FIGURE 6 Pollutant levels: (A) SO_2, (B) O_3, and (C) benzene.

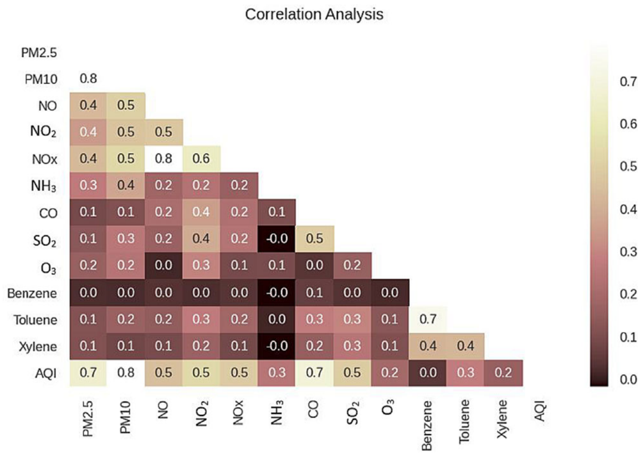


FIGURE 7 Correlation between pollutants.

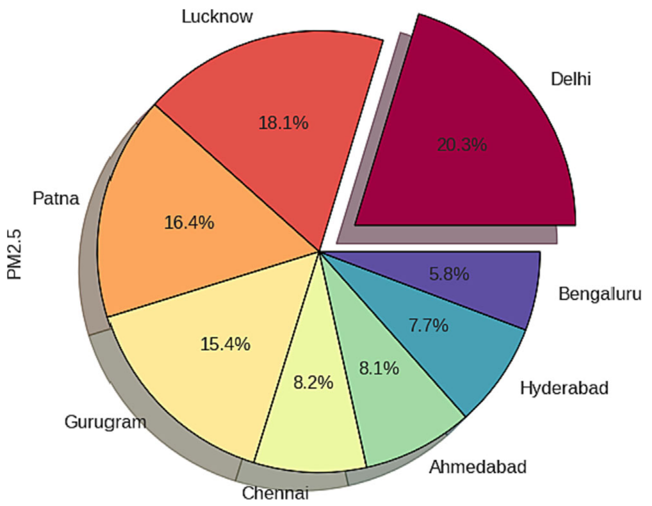


FIGURE 8 Pollutant levels in considered Indian cities.

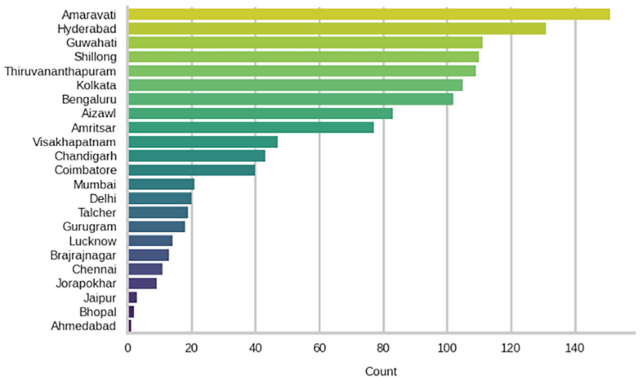


FIGURE 9 Presence of pollutants over days.

shown in Figure 6A–C, respectively. The AQI of SO_2 was assessed as part of the training and testing of the TLSTM model.

Figure 7 shows the correlation between pollutants considered in the AQI calculated by the proposed SARIMA–TLSTM model based on similarity values ranging from 0.0 to 0.7. A high correlation indicates high similarity. PM2.5 and PM10 have a high correlation of 0.8, whereas benzene has low correlations with PM2.5, PM10, NO, NOx, NH_3, CO, SO_2, and O_3. Figure 8 shows the pollutant levels in the cities, and Figure 9 shows the pollutant levels over days. Figure 10 shows the AQI predictions for 2019–2020, and Figure 11 shows those for 2019–2021.

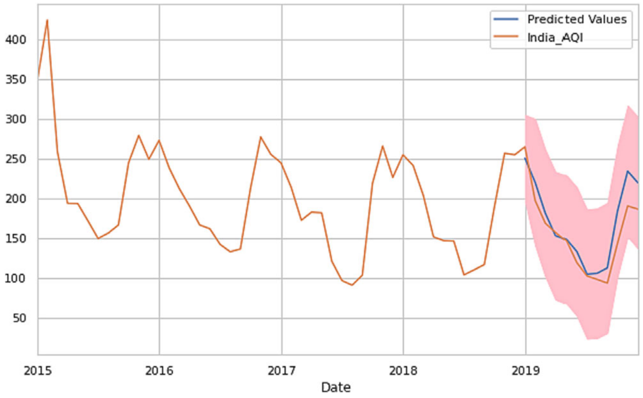


FIGURE 10 Predicted AQI values for 2019–2020.

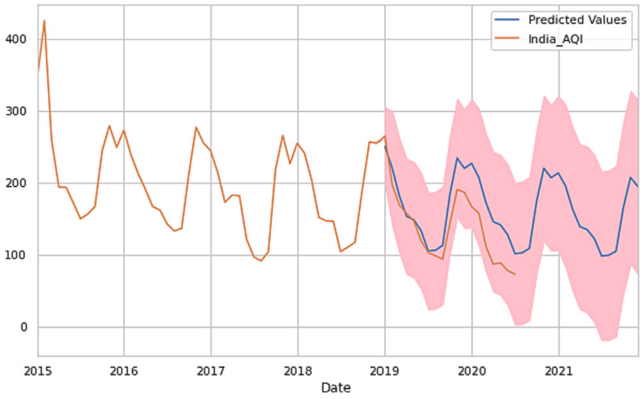


FIGURE 11 Predicted AQI values for 2019–2021.

4.4 | Statistical analysis of historical pollutant and meteorological datasets

This study included data from 35 air quality monitoring stations in Chennai from 2019 to 2021. The data included identifiers, timestamps, and concentrations of PM2.5, PM10, NO_2, CO, SO_2, and O_3 calculated at the stations. The minimum, mean, and maximum values are listed in Table 4 along with other statistical measures. PM2.5 had the highest concentrations from 2 to 1004 g/m^3, exceeding the upper limit of severe air pollution.

4.5 | Comparative analysis

Table 5 lists the evaluation results of the proposed and existing models. The CNN shows 87.98% accuracy, while LASSO regression fails to consider the temporal factors that influence AQI, yielding an accuracy of 78%. This model lacks predictors and requires additional temporal factors for further analysis. The autoencoder-based GAN performs regional spatiotemporal forecasting and exhibits an accuracy of 89.4%. The gray-level co-occurrence matrix was combined with SVR and LSTM. The obtained AQI values can be used to detect areas susceptible to pollutants. The proposed SARIMA–TLSTM model obtains an accuracy of 93%, outperforming the existing models. Table 5 also lists the RMSE and R^2 values for 2019 and 2020, respectively.

TABLE 4 Statistical characteristics of air quality factors.

Characteristic	PM2.5	PM10	NO_2	CO	O_3	SO_2
Count	290 621	227 747	292 359	268 197	290 589	292 462
Mean	56.13	87.04	44.72	0.97	54.72	9.01
Standard deviation	63.15	89.64	33.11	1.00	53.82	11.70
Minimum	2.0	5.0	1.0	0.1	1.0	1.0
25% percentile	15.0	38.0	19.0	0.5	2912.0	3.0
50% percentile	39.0	70.0	39.0	0.7	45.0	5.0
75% percentile	77.0	113.0	66.0	1.2	79.0	11.0
Maximum	1006	3000	300	14	499	312

TABLE 5 Comparative analysis results.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	RMSE	R^2
CNN [18]	87.98	—	—	—	—	—
Adaptive LASSO regression [19]	78	58.33	65.33	54.66	—	—
Extreme gradient boosting [20]	—	—	—	—	13.85	0.999
Air-GAN [22]	89.4	—	—	—	4.21	0.972
SVM–LSTM [23]	—	—	–	–	10.9	0.970
LSTM [25]	—	—	—	—	—	0.89
Optimized RNN [26]	—	—	—	—	0.83	—
SARIMA–LSTM	91	93	92	92	0.188	0.528
Proposed SARIMA–TLSTM	93	95	94	96	0.179	0.512

5 | CONCLUSION

We consider various measured climatic factors for AQI prediction. The predictions can support activities such as planning highway traffic signals and supporting the exchange of public transport. Moreover, electrical wheel machinery and nonmechanized vehicles may be deployed when pollution levels exceed the limits to ensure an adequate air quality. The proposed model outperforms other solutions and may be used in developing countries for sustainable urban areas. Deep learning algorithms have been combined to detect air pollution to control and improve AQI prediction in cities. Based on the predicted AQI values, susceptible areas can be identified. The proposed SARIMA–TLSTM model obtained an accuracy of 93%, outperforming the existing CNN (87.98% accuracy), LASSO regression (78% accuracy), and GAN (89.4% accuracy). The proposed method may be enhanced by considering time complexity parameters in future work.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during this study are available from the Central Pollution Control Board, Ministry of Environment, Forests, and Climate Change, India at https://app.cpcbccr.com/AQI_India.

ORCID

Murugan Saravanan https://orcid.org/0000-0003-4001-2818

REFERENCES

1. M. K. AlOmar, M. M. Hameed, and M. A. AlSaadi, *Multi hours ahead prediction of surface ozone gas concentration: robust artificial intelligence approach*, Atmos. Pollut. Res. **11** (2020), 1572–1587.
2. R. Sharma, R. Kumar, D. K. Sharma, L. H. Son, I. Priyadarshini, B. T. Pham, D. T. Bui, and S. Rai, *Inferring air pollution from air quality index by different geographical areas: case study in India*, Air Qual. Atmos. Health **12** (2019), 1347–1357.
3. H. Amini, N. T. T. Nhung, C. Schindler, M. Yunesian, V. Hosseini, M. Shamsipour, M. S. Hassanvand, Y. Mohammadi, F. Farzadfar, A. M. Vicedo-Cabrera, J. Schwartz, S. B. Henderson, and N. Künzli, *Short-term associations between daily mortality and ambient particulate matter, nitrogen dioxide, and the air quality index in a Middle Eastern megacity*, Environ. Pollut. **254** (2019), 113121.
4. S. Zhu, L. Yang, W. Wang, X. Liu, M. Lu, and X. Shen, *Optimal-combined model for air quality index forecasting: 5 cities in North China*, Environ. Pollut. **243** (2018), 842–850.
5. C. W. Chen and L. M. Chiu, *Ordinal time series forecasting of the air quality index*, Entropy **23** (2021), 1167.
6. Q. Wu and H. Lin, *A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors*, Sci. Total Environ. **683** (2019), 808–821.
7. R. Yan, J. Liao, J. Yang, W. Sun, M. Nong, and F. Li, *Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering*, Expert Syst. Appl. **169** (2021), 114513.
8. H. Li, J. Wang, R. Li, and H. Lu, *Novel analysis–forecast system based on multi-objective optimization for air quality index*, J. Clean. Prod. **208** (2019), 1365–1383.
9. A. Cascallar-Fuentes, J. Gallego-Fernández, A. Ramos-Soto, A. Saunders-Estévez, and A. Bugarín-Diz, *Automatic generation of textual descriptions in data-to-text systems using a fuzzy temporal ontology: application in air quality index data series*, Appl. Soft Comput. **119** (2022), 108612.
10. L. A. Gladson, K. R. Cromar, M. Ghazipura, K. E. Knowland, C. A. Keller, and B. Duncan, *Communicating respiratory health risk among children using a global air quality index*, Environ. Int. **159** (2022), 107023.
11. L. D. Perlmutt and K. R. Cromar, *Comparing associations of respiratory risk for the EPA air quality index and health-based air quality indices*, Atmos. Environ. **202** (2019), 1–7.

12. K. R. Cromar and L. D. Perlmutt, *EPA Air Quality Index: Limitations as a risk communication tool*, (American Thoracic Society 2016 International Conference; A106 Epidemiology and risk factors of Asthma: From the Crib to Adulthood, San Francisco, CA, USA), 2016, art. no. A2774.
13. X. Zhang and Z. Gong, *Spatiotemporal characteristics of urban air quality in China and geographic detection of their determinants*, J. Geogr. Sci. **28** (2018), 563–578.
14. W.-Z. Huang, W.-Y. He, L. D. Knibbs, B. Jalaludin, Y.-M. Guo, L. Morawska, J. Heinrich, D.-H. Chen, Y.-J. Yu, X.-W. Zeng, H.-Y. Yu, B.-Y. Yang, L.-W. Hu, R.-Q. Liu, W.-R. Feng, and G.-H. Dong, *Improved morbidity-based air quality health index development using Bayesian multi-pollutant weighted model*, Environ. Res. **204** (2022), 112397.
15. L. Ye and X. Ou, *Spatial-temporal analysis of daily air quality index in the Yangtze River Delta region of China during 2014 and 2016*, Chin. Geogr. Sci. **29** (2019), 382–393.
16. C.-J. Liang, J.-J. Liang, C.-W. Jheng, and M.-C. Tsai, *A rolling forecast approach for next six-hour air quality index track*, Eco. Inform. **60** (2020), 101153.
17. Y. Alyousifi, M. Othman, R. Sokkalingam, I. Faye, and P. C. Silva, *Predicting daily air pollution index based on fuzzy time series Markov chain model*, Symmetry **12** (2020), 293.
18. R. Chauhan, H. Kaur, and B. Alankar, *Air quality forecast using convolutional neural network for sustainable development in urban environments*, Sustain. Cities Soc. **75** (2021), 103239.
19. J. K. Sethi and M. Mittal, *An efficient correlation based adaptive LASSO regression method for air quality index prediction*, Earth. Sci. Inform **14** (2021), 1777–1786.
20. D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. M. Rao, D. Dhabliya, S. Satla, and W. Abera, *Intelligent forecasting of air quality and pollution prediction using machine learning*, Adsorpt. Sci. Technol. **2022** (2022), 5086622.
21. P. Saha, P. Nath, A. I. Middya, and S. Roy, *Improving temporal predictions through time-series labeling using matrix profile and motifs*, Neural Comput. Applic. **34** (2022), 13169–13185.
22. S. Abirami and P. Chitra, *Regional spatio-temporal forecasting of particulate matter using autoencoder based generative adversarial network*, Stoch. Environ. Res. Risk Assess. **36** (2022), 1255–1276.
23. R. Janarthanan, P. Partheeban, K. Somasundaram, and P. N. Elamparithi, *A deep learning approach for prediction of air quality index in a metropolitan city*, Sustain. Cities Soc. **67** (2021), 102720.
24. Y. Cong, X. Zhao, K. Tang, G. Wang, Y. Hu, and Y. Jiao, *FA-LSTM: a novel toxic gas concentration prediction model in pollutant environment*, IEEE Access **10** (2022), 1591–1602.
25. P. R. Gokul, A. Mathew, A. Bhosale, and A. T. Nair, *Spatio-temporal air quality analysis and $PM_{2.5}$ predictions over Hyderabad City, India using artificial intelligence techniques*, Eco. Inform. **76** (2023), 102067.
26. S. Surono, K. W. Goh, C. W. Onn, and F. Marestiani, *Developing an optimized recurrent neural network model for air quality prediction using K-means clustering and PCA dimension reduction*, Int. J. Innov. Res. Sci. Stud. **6** (2023), 330–343.
27. J. Ni, H. Yang, J. Yao, Z. Li, and P. Qin, *Toxic gas dispersion prediction for point source emission using deep learning method*, Hum. Ecol. Risk Assess. Int. J. **26** (2020), 557–570.
28. A. G. Salman and B. Kanigoro, *Visibility forecasting using autoregressive integrated moving average (ARIMA) models*, Proc. Comput. Sci. **179** (2021), 252–259.
29. D. Doreswamy, K. S. H. Kumar, and I. Gad, *Time series analysis for prediction of $PM_{2.5}$ using seasonal autoregressive integrated moving average (SARIMA) model on Taiwan air quality monitoring network data*, J. Comput. Theor. Nanosci. **17** (2020), 3964–3969.
30. T. Xayasouk, H. Lee, and G. Lee, *Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models*, Sustainability **12** (2020), 2570.
31. J. Ma, Z. Li, J. C. Cheng, Y. Ding, C. Lin, and Z. Xu, *Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network*, Sci. Total Environ. **705** (2020), 135771.
32. M. Ebrahimi, J. F. Nunamaker Jr., and H. Chen, *Semi-supervised cyber threat identification in dark net markets: a transductive and deep learning approach*, J. Manag. Inf. Syst. **37** (2020), 694–722.
33. Z. Karevan and J. A. Suykens, *Transductive LSTM for time-series prediction: an application to weather forecasting*, Neural Netw. **125** (2020), 1–9.

AUTHOR BIOGRAPHIES

Subramanian Deepan completed his B.E. and M.E. studies and is pursuing his Ph.D. at the Department of Networking and Communications, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. He worked as a teaching associate in the Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. He received an M.E. degree in computer science and engineering at the SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India, and the B.Tech. degree from Valliammai Engineering College, Kattankulathur, Tamil Nadu, India. His research interests include machine learning, Internet of Things, and deep learning.

Murugan Saravanan is working as an associate professor in the Department of Networking and Communications, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. He received the

M.E. degree in computer science and engineering from the Government College of Technology, Coimbatore, Tamil Nadu, India, and the Ph.D. degree in computer science and engineering from Periyar Maniammai University, Thanjavur, Tamil Nadu, India. His research interests include cloud computing, fog computing, and cloud security.

How to cite this article: S. Deepan and M. Saravanan, *Air quality index prediction using seasonal autoregressive integrated moving average transductive long short-term memory*, ETRI Journal **46** (2024), 915–927, DOI 10.4218/etrij.2023-0283