EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology 2024, 25(1), 341–359
https://doi.org/10.12912/27197050/175720
ISSN 2719-7050, License CC-BY 4.0

Received: 2023.11.04
Accepted: 2023.11.24
Published: 2023.12.14

Optimal Prediction of Air Quality Index in Metropolitan Cities Using Fuzzy Time Series with Deep Learning Approach

Asha Unnikrishnan[1*], Subramanian Rajeswari[1]

[1] Department of Computer Science, Sree Saraswathi Thyagaraja College, Bharathiar University, Pollachi, Tamil Nadu, India

[*] Corresponding author's e-mail: asha.nkk@gmail.com

ABSTRACT

Predicting the air quality index (AQI) with high accuracy is just as crucial as predicting the weather. The research selected a few potential meteorological parameters and historical data after taking into account a variety of complex factors to accurately anticipate AQI. The dataset was gathered, pre-processed to substitute missing values (MV) and eliminate redundant information, and before being applied to predict the AQI. The data was collected from 2019 to 2022 to analyse the AQI founded on time series forecasting (TSF). Many AQI parameters, including accumulated precipitation, the daily normal temperature, and prevailing winds, are lacking in this study. To preserve the characteristics of the time series, kNN classification was implemented to fill in the MV and integrate Principal Component Analysis (PCA) to decrease the noise of data to recover the accuracy of AQI prediction. However, the majority of research is limited due to a lack of panel data, which means that characteristics such as seasonal behaviour cannot be taken into account. Consequently, the research introduced a TSF based on seasonal autoregressive integrated moving average (SARIMA) and stochastic fuzzy time series (SFTS). The stacked dilated convolution technique (SDCT) which effectively extracts the time autocorrelation, while the time attention module focuses on the time intervals that were significantly linked with each instant. To control the strongly connected features in the data set, the Spearman rank correlation coefficient (SRCC) was utilised. The selected features included SO_2, CO and O_3, NO_2, PM_{10} and $PM_{2.5}$, temperature, pressure, humidity, wind speed and weather, as well as rainfall. Additionally, to estimate the AQI and SO_2, PM_{10}, $PM_{2.5}$, NO_2, CO, and O_3concentration from 2019 to 2022, the data of climatological elements after PCA and historical AQI were input into the multiple linear regression (MLR) techniques with a temporal convolution network (TCN) built deep learning model (DLM). The proposed DLM springs a correct and detailed assessment for AQI prediction. The experimental results confirm that the expected background yields a stable forecasting result, that the pollutant concentration of the surrounding areas affects the AQI of a place, and that the planned model outperforms existing state-of-the-art models in terms of prediction of consequences. Consequently, utilising this presented innovative approach integrates fuzzy time series with deep learning, addressing missing values and noise reduction, incorporating seasonal behaviour, utilising the SRCC for feature control, employing a comprehensive set of meteorological parameters, and presenting a hybrid model that outperforms existing models. These aspects collectively contribute to the advancement of air quality prediction methodologies, particularly in metropolitan cities. However, this hybrid approach leverages the strengths of both traditional statistical methods and deep learning techniques, resulting in a robust and accurate assessment for AQI prediction as well as providing more stable and accurate forecasting results.

Keywords: air quality index, deep learning, northern and southern cities, stochastic fuzzy time series, SARIMA, temporal convolution network.

INTRODUCTION

The demand for financial expansion has resulted in severe pollution of air all across the world. Pollutants (harmful compounds) are presented into the air as gases and particle debris. Particulate matter, carbon dioxide, sulphur dioxide, and ozone are a few of the harmful air pollutants. The addition of these air contaminants degrades the air quality and renders it dangerous for human life [Abirami and Chitra, 2021]. Numerous cardiovascular and respiratory conditions, including ischemic heart disease, lung cancer, stroke, chronic obstructive pulmonary disorders (COPD), bronchial asthma, etc., are brought on by prolonged exposure to poor air quality. Over 90% of the global inhabitants reside in places with poor air quality, and the World Health Organization (WHO) estimates that air pollution causes about 4.2 million deaths annually [5]. However, action should be taken to remove hazardous pollutants when air pollution concentrations reach dangerous levels [Liu et al., 2021]. The small amount of pollutants in the air only has a minor impact on the human form and the surrounding environment. As a consequence, it is crucial to continuously assess and project a nation's air quality. Each country has its own rules and standards for measuring air quality, along with slightly varied terminology. In Malaysia, the air pollution index (API) is applied to assess the current state of air superiority consequently appropriate steps can be made to reduce the properties of air contamination on people and the atmosphere [Moscoso-López et al. 2020].

The concerns about the detrimental impact of air pollution have been expressed by regular people, researchers, policymakers, and operators of smart cities. Effective air quality prediction models (PM) are valuable for prompt air pollution prevention and control. However, the spatiotemporal delivery properties of air quality have not been properly taken into account. Using prior observations, the neural network-based bidirectional long short-term memory (BILSTM) network and Convolutional Neural Network (CNN) are exploited to spatially assess and forecast the pollutant level in the research region in advance [Samal et al. 2021, Zhang et al. 2021]. Custom DLM reproductions were constructed on spatiotemporal matrix to build a multi-time, multi-site predicting model of Beijing's air quality. The next-hour forecasting results were ranked as the best model for multiple-hour forecasting for overall forecasting [Yan et al. 2021]. Due to the relevance of the field, machine learning techniques (MLTs) have started to be actively researched in this area, and several studies as well as observations have been made. By combining all of the available data, one can identify trends and learn about new developments in the field, which will help plan and lead future studies [Iskandaryan et al. 2020].

However, there is still a possibility for improvement in the current methods for predicting long-term air quality. As an effect, a spatial-temporal DLM founded on a recurrent neural network (RNN-LSTM) and a bidirectional gated recurrent unit integrated with an l (BiAGRU) [Al-Janabi et al. 2020] were developed for more precise predicting of air quality. A spatial-temporal matrix obtained from previous air quality capacities and weather monitoring data can be functional as model input [Zhang et al. 2020]. The objective of the investigation was to estimate AQI in a location with varying concentrations of key air pollutants like RSPM, SO_2, NO_2, and SPM. The study aimed to create a daily AQI estimating model that can be utilised for regional and local air eminence organizations [Lin et al. 2020]. The predicting model depends on the symmetry idea of fuzzy data translation from a sharp single point to an ambiguous number. FTS is operated to estimate air pollution; nevertheless, it has a drawback for arbitrarily assigns vast majority of intervals [Alyousifi et al. 2020]. The rebuilt series is forecast using three hybrid models in the module for dynamic integration forecasting: an optimised extreme learning machine, an FTS model, and an ARIMA model. Time-varying parameters are applied to dynamically combine the forecasting results [Li et al. 2020]. The time series is divided into multiple levels using a decomposition approach, and it is reconstructed beforehand to better extract the sequence material and eliminate the random noise. To strengthen the fuzzy neural network, particle swarm optimization, genetic, and steepest descent back propagation methods are exploited [Tomar et al. 2020]. The classification of the AQI standards usages a DLM constructed on long LSTM and support vector regression (SVR). Associated with the current methods, the suggested DLM provides a correct and precise result for AQI on the designated city site. The suggested DLM has enhanced forecast accuracy, which will warn the public to lower to a satisfactory level. The DLT effectively forecasts the AQI ethics and aids in planning the urban growth of metropolitan area for sustainability [Alyousifi et al. 2020].

Monitoring of air pollution is continually expanding, with an emphasis on its effects on the well-being of people. As the two main contaminants, nitrogen dioxide (NO_2) and sulphur dioxide (SO_2), numerous models have been built to forecast the possible harm they may cause. Nevertheless, it is difficult to make precise predictions.

Heydari et al. [2022] describe the growth of a new hybrid intelligence model based on MVO and LSTM to predict and analyse the air contamination derived from joint cycle power plants. The collected findings demonstrate that, when diverse network input factors are taken into account, the advised model outpaces the previous combined benchmark model for forecasting in standings of accuracy. For the challenge of hour-ahead air worth forecasting, Zeng et al. [2022] familiarized a hybrid deep learning model (HDLM) that incorporates the advantages of the stationary wavelet transform (SWT) and the nested LSTM. This model aims to improve the forecast quality. Using an enhanced SWT algorithm, the suggested method divides the unique PM2.5 data into some additional stationary sub-signals with various resolves. To obtain predicting results for various sub-signals, a framework utilizing multiple NLSTM RNNs is built.

Two DLMs, APTR and DeepPM, were built and trained in [Wang et al. 2022] utilizing $PM_{2.5}$ and O_3 monitoring data besides WRF-Chem numerical predictions in the southeastern Beijing, Tianjin, and Hebei area to increase the estimating accuracy of numerical air excellence models. While the deep PM model makes better overall at medium-term – and optimising short forecasts, the APTR model excels at proximity forecasting optimization. DLTs using RNNs and STM models were utilized by Kristiani et al. [2022]. There were two different kinds of time-series data exploited: meteorological elements including temperature, humidity, wind speed, and direction, in addition to components of air pollution like SO_2, O_3, and CO_2 from 2017 to 2019. Within an eight-hour period, air pollution was predicted using a trained model. To increase the exactness of the shaped model, the MVs were filled in using the KNN approach. In the experiments, the model's presentation was assessed by expanding the normal complete error % value.

A two-stage method for forecasting when PM_{10} and O_3 concentrations will surpass the regulatory limitations set by the EU [Krylova and Okhrin, 2022]. The ensemble tree-based Stochastic Gradient Boosting Model demonstrated the best performance which is particularly practical because of its computational effectiveness and resilience to overfitting. Its aptitude to present outcomes was significantly improved than that of comparable research. It is essential to first comprehend air pollution, identify its patterns and correlations, as well as make predictions about it to lessen its negative impacts. To overcome this spatiotemporal problem, air contamination prediction requires highly complicated predictive algorithms. The Graph Convolutional Network (GCN) and CNN-LSTM are two cutting-edge DLMs utilised to understand the patterns of particulate matter 2.5 ($PM_{2.5}$) over temporal and spatial associations [Tiwari et al. 2021].

The health concerns brought on by air pollution will be reduced with accurate AQI prediction. Finding the best machine learning method for forecasting an accurate AQI in Colombo built on a specific $PM_{2.5}$ concentration was the main objective [Zhang et al. 2021]. The prediction simulations were qualified and evaluated using MLTs such as K-Nearest Neighbours (K-NN), MLR, SVM, and Random Forest (RF). After comparing the models, RF was shown to be the most accurate PM, with a superior accuracy of over 85%. Researching these modern risks is more successful with machine learning-based prediction technologies than with traditional approaches. Resampling is applied to the discussion on the subject of data imbalance, and five machine-learning models are exploited to project air quality [Seng et al. 2021]. Although there is still an opportunity to enhance estimating accuracy, it is also necessary to discuss the problem of lacking air pollution data for some target locations. Hence, a novel loaded ResNet-LSTM model was presented by [Cheng et al. 2022] to increase the precision of $PM_{2.5}$ concentration level forecasts.

Data mining (DM) constitutes a stated removal of hidden information about predicting from large files. From the large database, the necessary information is fetched using the DM process. Here, Neelaveni and Rajeswari [2016] discussed the DM procedures applied in the agriculture field. Accordingly, the prediction of air superiority using DLM is thoroughly investigated using a vast amount of environmental data. A thorough PM with multi-output and multi-index supervised learning (MMSL) was suggested and established on LSTM [Janarthanan et al. 2021]. The routine of the current model is often superior to other baseline models, according to experimental findings. Variational mode decomposition (VMD) and bidirectional LSTM are combined to generate the HDLM VMD-BiLSTM [Fernando et al. 2022], which is used to forecast the variations in $PM_{2.5}$ in Chinese cities. The innovative $PM_{2.5}$ complicated data in time series is broken down by VMD into various sub-signal components centred on the frequency area. Then, using BiLSTM to predict each sub-signal constituent, the prediction accuracy was greatly increased. The conclusions indicate that using the suggested forecasting framework greatly improves the prediction outcomes. A versatile and effective deep learning-driven model to predict ambient pollution attention was proposed by [Dairi et al. 2021]. The study first addresses the Variational AutoEncoder (VAE) and the attention mechanism (AM) to build a forecasting modelling approach based on the unique integrated multiple directed attention deep learning architecture (IMDA). Additionally, the predicted result of the submitted model outperforms that of GRU and LSTM using the AM [Kumar et al. 2022]. There are significant negative political, sociological, and economic implications of poor air value in addition to harmful effects on the condition of people and vegetation. To offer relevant and valuable solutions, attain acceptable air quality, and develop preventative strategies, it is crucial to place more effort into precise prediction of ambient air pollution.

Research problem

The rise in environmental issues has been attributed to the globalised development that is occurs in both industrialised and developing nations. The most common kind of pollution in the world is air pollution, which is mostly brought on by both natural phenomena like volcano eruptions and man-made ones like open burning and industrial processes. Air pollution affects people's health, the economy, and the advancement of civilization, in addition to upsetting the natural equilibrium. The AQI, as a single major indicator of air quality, is comprised of 6 major pollution components: reparable particulate matter, fine particulate matter, ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO). The research on AQI and forecasting methods allows for enhanced air contamination prevention and management. Data-driven TSF methods, including DLM and big data analysis, provide crucial forecasting results for several manufacturing areas. Traditional approaches to the AQI estimating difficulty comprise regression analysis, RF, and support vector machine (SVM). DLM approaches are utilised for the AQI projecting issue to manage environmental indices more precisely and produce a next-generation smart environment. The formation and testing of environmental pollution models, particularly those about atmospheric pollution have remained the subject of a wealth of research. However, the current and contemporary models do not examine the performances of pure modelling for forecasts and instead concentrate on determining the causes and their temporal correlations. Therefore, the study took into account three years' worth of data that show the hourly concentrations of SO_2, CO, O_3, NO_2, PM_{10}, and $PM_{2.5}$ in the air. Exposure to high concentrations of these pollutants has been related to an amount of respiratory, circulatory, and even neurological illnesses.

METHODOLOGY

With the advancement of science and technology, industry, transportation, and other sectors are utilised to release a significant amount of pollutants into the atmosphere, leading to air pollution. Human health will suffer greatly when air pollution reaches a critical level. High-accuracy estimation of the AQI is crucial to weather forecasting. People could plan their vacations and daily activities using the incredibly accurate prediction results, to better protect their health. The study selected several probable meteorological parameters and some historical data after taking into account several complex factors to accurately predict AQI. Figure 1 describes the block diagram of the proposed work.

This research presented the methods of predicting AQI using MLTs and forecasting the air pollution levels to take precautionary measures to minimise air pollution. Initially, the data were collected from the north and southern cities of India, this collected data were preprocessed using KNN classification and PCA methods. It replaces the MVs and reduces the noise of data. Accordingly, the TSFM was presented as built on SARIMA and SFTS. Furthermore, the SDCT efficiently extracts the time autocorrelation and the SRCC measures the grade of correlation. Subsequently, a multi-step prediction model (MPSM) was presented and constructed on DLM. It presents the combination of MLR techniques with a TCN-centered DLM for predicting AQI and $PM_{2.5}$, NO_2, SO_2, PM_{10}, O_3, and CO concentration. A novel time series forecasting model (TSFM) was introduced, constructed on a combination of SARIMA and SFTS. This fusion of traditional and fuzzy time series approaches provides a more nuanced and

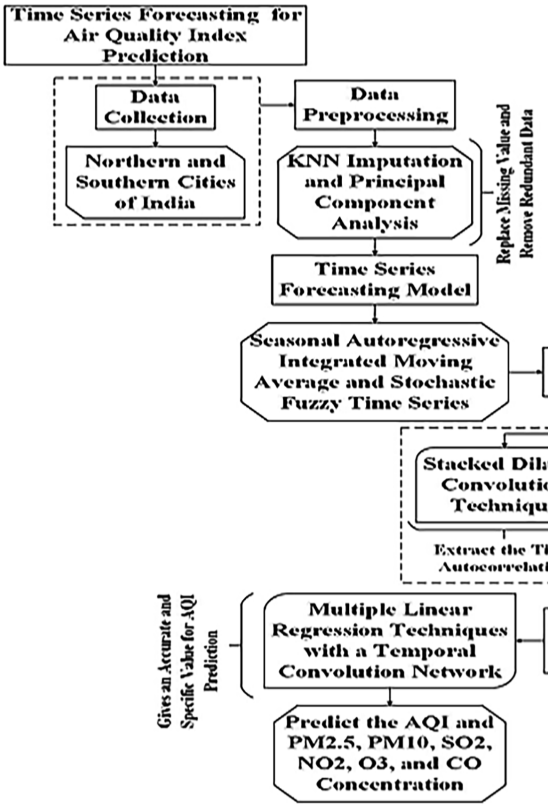

Time Series Forecasting for Air Quality Index Prediction
Data Collection
Northern and Southern Cities of India
Data Preprocessing
KNN Imputation and Principal Component Analysis
Replace Missing Value and Remove Redundant Data
Time Series Forecasting Model
Seasonal Autoregressive Integrated Moving Average and Stochastic Fuzzy Time Series
Extraction Model
Stacked Dilated Convolution Technique
Spearman Rank Correlation Coefficient
Extract the Time Autocorrelation
Measure the Degree of Correlation
Multi-Step Prediction Model
Multiple Linear Regression Techniques with a Temporal Convolution Network
Gives an Accurate and Specific Value for AQI Prediction
Predict the AQI and PM2.5, PM10, SO2, NO2, O3, and CO Concentration

Figure 1. Block diagram of the proposed work

accurate representation of spatio-temporal patterns in the AQI data. However, the proposed spatio-temporal feature extraction method enables the identification of interval autocorrelations and the assessment of weakly connected elements in the dataset, enhancing the model's capability to capture complex relationships. The ability of the proposed model to accurately predict AQI values is linked to potential interventions, such as coordinating traffic signals and promoting public transport, to reduce the pollution levels in the studied cities. Consequently, this work involves innovative hybrid DL-based technique, geographical diversity in data collection, consideration of comprehensive meteorological and historical factors, a unique TSFM, spatiotemporal feature extraction using advanced techniques, thorough performance evaluation metrics, and the model's practical implications for pollution reduction strategies in diverse Indian cities.

Data collection

The relationship between air quality and human health is tracked using the API, which is an example of crucial airworthiness measures. Government organisations, decision-makers, and citizens can all benefit from the API information in developing preventative steps to lessen the impact of episodes of air pollution. The dataset was formed and pre-processed to eliminate duplicate data and interchange MVs to conjecture the AQI. The data was collected from 2019 to 2022 to analyze AQI founded on TSF. Several AQI tenets, including total precipitation, the average daily temperature, and wind information, are lacking in this study.

Data sources

A diversity of air poisons, including particulate matter, sulphur dioxide, nitrogen dioxide, and ozone, were measured by sensors positioned across the city. In the meantime, a vast amount of data was collected from many sources. This investigation sought to validate the monthly API patterns in major cities of northern and southern India https://app.cpcbccr.com/AQI_India/. Other data applied for the study were composed of different regions tabulated in Table 1.

Materials

This section outlines the suggested process for generating TSFMs for the following concentrations: AQI, $PM_{2.5}$, PM_{10}, O_3, NO_2, CO, and

Table 1. Data sources

Regions	Link
Kochi	https://www.aqi.in/dashboard/india/kerala/kochi
Trivandrum	https://www.iqair.com/in-en/india/kerala/thiruvananthapuram
Kozhikode	https://aqicn.org/city/india/kozhikode/palayam
Thrissur	https://air-quality.com/place/india//053c2703?lang=en&standard=aqi_us
Kanyakumari	https://www.wunderground.com/health/in/kanyakumari/8.07999992,77.54000092
New Delhi	https://www.oneindia.com/air-quality-index/delhi
Kanpur	https://airpollutionapi.com/aqi/india/uttar-pradesh/iit-kanpur-airstrip
Hyderabad	https://aqicn.org/city/hyderabad
Mangalore	https://air.plumelabs.com/air-quality-in-Mangalore-5iLy

SO_2. It also compares the models and determines which model is optimal when taking 24-step-ahead predictions into account. Comparing various DLMs with more traditional machine learning techniques and with each other is the primary departure point of the suggested methodology.

Datasets and study area

Cities in northern and southern India, including Delhi, Kanpur, Hyderabad, and Mangalore, make up the research region. The experimental datasets were made available for retrieval by the Pollution Control Department of the Ministry of Natural Resources and Environment (Noise and Air Superiority Management Division, 2020). Air pollution (nitrogen dioxide (NO_2), floating debris with a size less than in aerodynamic diameter (PM_{10}), carbon monoxide (CO), suspended particles of a size less than in aerodynamic diameter ($PM_{2.5}$), ozone (O_3), and sulphur dioxide (SO_2) data for the period 2019 to 2022 were collected at an hourly sampling rate from four air class checking places are employed in this analysis.

Meteorological data

The monthly averaged surface meteorological variables included relative humidity (RH), rainfall, station level pressure (SLP), minimum temperature (t_{min}), mean temperature (t_{mean}), daily temperature range (difference among daily minimum and maximum temperature, t_{range}), vapour pressure (vp), weather, cloud cover (cc), radiation (rd), sunshine hours (ssh), wind speed (wsp), daily maximum temperature (t_{max}), rainfall (rf), visibility (v), and wind direction index (wdi). The information from 2019 to 2022 in Delhi, Kanpur, Hyderabad, and Mangalore has been acquired from the Indian Meteorological Department (IMD).

Air quality index

In many countries, AQI is generally employed and it is a universal format. The AQI level of India is based on the level of six pollutants like CO, PM_{10}, $PM_{2.5}$, O_3, SO_2, and NO_2, and throughout these cities, it is measured at the measuring stations. AQI is used for the effective assessment of air quality. The data of various pollutants are transformed into a single quantity or values through this process. Moderately polluted, good and satisfactory, very poor, poor, and severe, these six different categories, of AQI are categorized. For tracking these pollutants (PM_{10}, $PM_{2.5}$, O_3, SO_2, CO, and NO_2), the AQ sub-index was developed. The AQ sub-index calculation is constructed on the ambient concentration of air contaminants.

Data preprocessing

Outlier identification and missing data imputation (MDI) were carried out during the pre-processing stage. The information submitted for processing is not an average or mean, but rather the actual hourly value of each type of air pollution is determined by measurements taken at nearby stations. To preserve the characteristics of the time series, implement kNN classification to fill in the MV and participate in PCA to reduce the noise of data to progress the accuracy of AQI prediction. Real data processing applications have successfully utilised the KNN and PCA techniques for their simplicity, ease of understanding, and reasonably high accuracy.

KNN estimator for missing data imputation

The NN method is one of the hot deck strategies intended to make up for nonresponse (MVs) in the sample data. It needs to be among the main options for MDI when there is minimal or no prior knowledge regarding the dissemination of the data. It was extended to the kNN method for the NN method often leads to over-fitting. With the mainstream midst, its *k* nearest neighbours, a categorical MV is attributed with the kNN method, and for a numerical MV, the middling value (mean) of the *k* nearest neighbours is observed as the prediction, termed as the majority/mean rule. The formal definition of it is as follows. The kNN classification is defined as follows, given (X, Y, 0) and the set of its kNN $D_k = \{(X_j, Y_j, 1) \mid j = 1, 2, \dots, k\}$.

$$Y = \{\arg max_v \left\{ \sum_{(X_j, Y_j, 1) \in D_k} 1\left(Y_j = v\right) \right\}, \quad (1)$$

$$if\ Y\ is\ categorical\ \frac{1}{k} \sum_{j=1}^{k} \quad Yj,\ if\ Y\ is\ numerical$$

where: $1(Y_j = v)$ is an indicator function that returns the value 1 if its disagreement is true and 0 otherwise, and is a value in the domain of the target feature Y.

kNN classification is a model-free technique as a result. The kNN approach faced two difficult problems, despite the simplicity of the kNN classification (majority/mean rule): (1) figuring out the ideal value of k beforehand, and (2) choosing the k nearest neighbours. Since there is no prior knowledge regarding the ideal k for a given application, the optimal worth of k was determined via experimental testing.

From the differences between instances, the resemblance between an example and its NNs was determined, and should certainly be maximal (or, the variance among them is minimal) for selecting kNNs. The Euclidean distance may be employed to calculate the distance measure. Choose its nearest neighbours from the training subset, after calculating the detachment from to all the working out instances as demonstrated in equation (2).

$$A_x = \{v_k\}_{k=1}^{K} \quad (2)$$

The K nearest neighbours of x, arranged in ascending order of distance, are represented by the set in equation (2). Therefore, v_1 was the NN of x. By comparing the distance with the inputs that are present in the imputed incomplete feature, the K nearest neighbours were selected. By an approximation from the th feature standards of , the unidentified value was imputed after its kNNwas selected. The assessment of its K nearest neighbours aimed to acquire the imputed value $\widetilde{x_j}$, if the th feature was a numeric variable. Established on its distance to x, one significant change was to weigh the influence of each comment, that is more weight is provided to the nearest neighbours.

$$\widetilde{x_j} = \frac{1}{KW} \sum_{k=1}^{K} \quad w_k v_{kj} \quad (3)$$

The main drawback of this approach is that the algorithm practices the full dataset when the KNN classification determines the most similar samples. For large databases, this restriction can be very serious. With high-dimensional data, the main disadvantage of KNN classification is that it can suffer greatly from the lack of significant differences between the nearest and farthest neighbours. The most significant selected features were only utilised in this study instead of utilising all the characteristics. The feature assortment system is commonly employed in machine learning for many applications. Using a few appropriate features instead of a huge quantity of redundant and irrelevant data to train the model can first increase the correctness of the conventional approach. Addressing the overfitting issue is the second and most important reason, since it increases the likelihood of overfitting as the amount of immaterial features increases. With this model, the most significant features are exploited to switch the missing data. This achieves the goal of mitigating the disadvantages of the KNN classification technique. According to the magnitude ranging from 0 to 1, this method predicts missing attribute values.

Principal component analysis

The correlation matrix U' of U is obtained by eigenvalue $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_i\}$ and the influence rate of the eigenvalue, $c = \{c_1, c_2, \dots, c_i\}$. The PCA is carried out on the sample data set = $\{U_1, U_2, \dots, U_i\}$. The components with an 85% cumulative contribution rate are extracted as the major component and the appropriate eigenvector $d = \{d_1, d_2, \dots, d_k\}$ is generated based on the eigenvalue λ and the component α eigenvalue. Following the acquisition of PCA from the sample data set U and $D^T = \{d_1, d_2, \dots, d_k\}$, equation (4) displays the dimensionality reduction data u of the sample data set U.

$$u = [u_1, u_2, \dots, u_k] = D^T \quad (4)$$

The computational complexity of the classic is effectively reduced by using this PCA method to remove overlapping feature information and the presentation of the classical is improved. On the other hand, it can improve the AQI accuracy due to its denoising characteristics.

Time series forecasting model

The lack of panel data limits the majority of studies, making them unable to take into consideration the elements, including seasonal behaviour. Consequently, the research introduces a TSFM constructed on SARIMA and SFTS. Fuzzy theory has been functional in particular by time series models to address the forecasting issues including financial forecasting and air excellence forecasting, temperature forecasting, and rainfall forecasting. These methods may effectively expand the accuracy of forecasting. SARIMA and Fuzzy Time Series models often possess complementary strengths. SARIMA is effective in capturing linear patterns and seasonality, while fuzzy time series models can handle non-linearity and uncertainty. By combining these approaches, the research leverages the strengths of each model to create a more robust forecasting framework. The utilisation of both SARIMA and Stochastic Fuzzy Time Series suggests an intention to enhance the predictive capability of the model. By tapping into the strengths of diverse methodologies, the research aimed to create a more accurate and reliable model for predicting AQI, considering both deterministic and uncertain components.

SARIMA model

SARIMA is particularly effective in capturing seasonal variations and trends in time series data. Its inclusion in the methodology allows the research to address and model the influence of seasonal patterns on the air quality index, providing a more accurate representation of the data. In the multiplicative model, the SARIMA is applied to model a time series that includes both seasonal and non-seasonal variables. The standard notation for the SARIMA is:

$$ARIMA\ (p, d, q) \times (P, D, Q)_s \tag{5}$$

where: s is the period of recurring seasonal pattern, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA imperative, p and is the non-seasonal AR order, d is the non-seasonal differencing, and q is the non-seasonal MA direction.

For the time series X_t (t = 1, 2, …, T), the SARIMA is provided by equation (6).

$$\varphi_p(B)\Phi_P(B^s)\ \Delta^d\Delta_s^D X_t = \theta_q(B)\Theta_Q(B^s)a_t, \tag{6}$$

$$t = 1, 2, \dots, T$$

where:

- non-seasonal AR component:

$$\varphi_p\ (B) = 1 - \varphi_1\ B - \dots - \varphi_p\ B^P \tag{7}$$

- seasonal AR component:

$$\Phi_p\ (B^S) = 1 - \Phi_1\ B^S - \dots - \Phi_p\ B^{PS} \tag{8}$$

- non-seasonal MA component:

$$\theta\ (B) = 1 + \theta_1\ B + \dots + \theta_q\ B^q \tag{9}$$

- seasonal MA component:

$$\theta\ (B^S) = 1 + \theta_1\ B^S + \dots + \theta_Q\ B^{QS} \tag{10}$$

$$\Delta^d\ \Delta_s^{\ D}\ X_t = (1–B)(1–B^s)\ X_t \tag{11}$$

where: B is the backshift worker, $\varphi_p < 1$, $\Phi_p < 1$, $\theta_q < 1$, $\Theta_Q < 1$ and a_t follows white noise $(0, \sigma^2)$.

To fit the time series after transforming the series into stationary and eliminating the seasonal component for the series, the model SARIMA was utilised. A logarithm is generally taken to the series to obtain the stationary series and for the constant variance, differencing is among the methods which is frequently applied. Using first difference (d = 1), the AQI time series was transformed into stationary sequences and partial autocorrelation and autocorrelation functions identify the orders of seasonal and non-seasonal autoregressive (AR) as well as MA models. Further, various forecast error methods checked the prediction accuracy of the reproductions.

Forecast accuracy of the models

Various forecast error measures have been exploited to check the forecast accuracy in this article. The most frequently applied two categories of estimate error procedures are given as follows:

- Measures based on scale-dependent errors,
- Measures based on percentage errors.

The calculation errors' size is equal to that of the data, and as the accuracy measures are based only on the in the scale-dependent errors technique, they cannot be utilised to compare time series with different units of measurement. The most widely exploited metrics focused on scale-dependent errors are MAE, root mean square error (RMSE), and mean absolute scaled error (MASE), which are explained as follows:

- Mean absolute error:

$$MAE = \frac{\sum_{t=1}^{n}\ e_t}{n} = \frac{\sum_{t=1}^{n}\ (|X_t - \hat{X}_t|)}{n} \tag{12}$$

- Root mean squared error:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\ e_t^2}{n}} = \sqrt{\frac{\sum_{t=1}^{n}\ (X_t - \hat{X}_t)^2}{n}} \tag{13}$$

- Mean absolute scaled error:

$$MASE = \frac{\sum_{t=1}^{n}\ \frac{|X_t - \hat{X}_t|}{Q}}{n} \tag{14}$$

where: Q is a stable measure of the gauge of the series of time X_t and for spring time series, the value of Q is:

$$Q = \frac{\sum_{t=m+1}^{n} \frac{|x_t-\hat{x}_t|}{Q}}{n-m} \quad (15)$$

For comparison of the forecast presentation of different data series in the measures based on the percentage errors, procedure measures are unit free and frequently exploited. Mean absolute percentage error (MAPE) is the most commonly utilised measure based on percentage errors.

$$MAPE = \frac{\sum_{t=1}^{n} |p_t|}{n} = \frac{\sum_{t=1}^{n} \frac{(|x_t-\hat{x}_t|)}{x_t}}{n} \times 100 \quad (16)$$

where: X_t the actual time series value, n is the sample size, and $\hat{x}_t$ is the prediction for the t^{th} month.

The main drawback of the measures for any period of interest t is that they have dangerous standards if any X_t is close to zero. This is based on the premise that percentage errors are infinite and undefined for $X_t = 0$. The inapplicability of percentage non-error cases when the units of measurement lack meaningful zeros is another drawback. For instance, when assessing the accuracy of a temperature prediction, a percentage error does not distinguish when the temperature has arbitrary zero values in either the Celsius or Fahrenheit scales.

The precision of the time series models applied to the projection of the air superiority data is therefore checked in this article using the measures based on scale-dependent errors (i.e., MAE, RMSE, and MASE) to control whether the AQI has meaningful units as zero (AQI are recorded as zero or close to zero in some of the months) in a few of the months in the data series.

Stochastic fuzzy time series forecasting model

In the field of air pollution cases, the SFTS has been extensively functional. In this method, the seven-step procedure of this method is presented. The stochastic fuzzy time series forecasting model indicates recognition of the uncertainty inherent in environmental and meteorological data. Fuzzy time series models, especially when stochastic elements are incorporated, are useful in handling uncertainty and variability, contributing to a more robust prediction framework. The development of fuzzy TSFM can be undertaken as follows using these seven steps:

- Define the dissertation V universe so that it can contain time series data that covers the entire production history.
- Divide the studied space into seven comparable length intervals.
- Define the membership purposes in the corresponding divisions to describe the fuzzy set.
- Finding correlations between historical data values and fuzzy sets is a step in the fuzzification process. The greatest degree of membership for each historical value has been fuzzed. Fuzzification is carried out in preparation for the subsequent computation.
- Fuzzy associations are discovered using the fuzzified historical data. Suppose $F(t–1)$ as A_i and A_{ij} fuzzy logical relationship (FLR) can be articulated as $A_i \rightarrow A_j$, where A_i and A_j are called the left-hand side and right-hand side of the FLR correspondingly.
- Develop the FLRs into the centred on the comparable fuzzily-defined integer to the left of FLRs. Construct a distinct logical relationship group if the transition takes place to another fuzzy set that is similar.
- The predicted value at t, $F(t)$, is demonstrated by three heuristic rules. Assuming the fuzzified API of $F(t–1)$ is A_j, the predicted results of $F(t)$ is resolute consuming the subsequent instructions:

Rule 1: In the event where $A_i \rightarrow A_k$ is a one-to-one connection group and A_k's highest degree of belongingness occurs at interval u_k, before the projected outcome of $F(t)$ equaling the midpoint of u_k.

Rule 2: When A_j is the empty set, that is, $A_j \rightarrow \phi$, and v_j is the interval where A_j has the highest degree of fit, then the expected outcomes are equal to v_j's midpoint.

Rule 3: If there exists a one-to-many relationship group of A_j, say $A_j \rightarrow A_1, A_2, \ldots, A_n$ and the maximum degree of fit occurs at a set $v_1, v_2, \ldots, v_n$, at that period the output predicted is completed as the normal of the midpoints $m_1, m_2, \ldots, m_n$ of $v_1, v_2, \ldots, v_n$.

However, the combination of SARIMA and stochastic fuzzy time series represents methodological innovation in the context of air quality prediction. This approach goes beyond traditional modelling techniques, showcasing a commitment to advancing the field and improving the accuracy of predictions in complex environmental scenarios.

Feature extraction model

SDCT effectively extracts the time autocorrelation whereas the time attention module focused on the periods that were highly correlated with each moment. The features in the data set that were weakly linked were removed using the SRCC method, which measures the point of association between the historical sequences of neighbouring stations and the historical arrangement of the board station. The selected parameters included SO_2, CO and O_3, NO_2, PM_{10} and $PM_{2.5}$, temperature, humidity, pressure, wind speed and weather, and rainfall.

Stacked dilated convolution technique

To improve the greatest descending window for the SDC process, this paper employed the time AM by the random searching method. To extract the time-dependent features of the two-dimensional eigenmatrix, the SDC model and the aggregated spatial eigenmatrix as the dilated convolution's input were used.

The eigenvalues were first determined by nonlinearly calculating the activation function, after which the shared convolution kernel (CK) was utilised for operation. The four input layer data modules were combined into a single group for the first hidden layer (HL) using dilated convolution. The dilation rate was included, allowing the CK to bypass the data modules with distinct dilation rates at the processing stage. This is the dilated convolution that differed from regular convolution. The input layer and the first HL had a first dilation rate of 1, while the second HL and the input layer had a second dilation rate of 2. The second HL and the output layer had a third dilation rate of 4, which meant that CK was sampled every four intervals. The unique structure of dilated convolution made it possible to generate a greater Receptive Field (RF) without significantly increasing network depth. This approach improved the comprehensiveness of feature extraction while also decreasing information loss and space loss. Here is the formulation for dilated convolution;

$$(F_d^* X)\,(x_t) = \sum_k^K \quad f_k x_{t-\,(K-k)d} \tag{17}$$

The typical sequence of input is represented by x_t in equation (17) and the filter is denoted by f_k. In addition, d denotes the dilation rate, k the size of the convolution kernel, and K the numeral of nodes engaged in the dilated convolution process of local RF. The SDC RF had a size of $(K–1)d + 1$. The model could achieve a broader RF by raising the CK or dilation rate. The SDC dilation rate was set to $d = 2^i$, and the network layer depth was set to 4, 4, after the initial values of the convolution kernel, where i represents the network layer depth, were established**.**

Spearman rank correlation coefficient

To assess the degree of association between the historical sequences of nearby stations and the target station (the AQI sequence that was going to be predicted), the SRCC was utilised, and the correlation coefficient (CC) was between 0 and 1. The more strongly the two stations' association is the closer the CC was to 1. The calculation method is as follows:

$$\rho(Y, Y_k) = 1 - \frac{6\sum_{i=1}^{N} \left(Y^i - Y_k^i\right)}{N(N^2-1)} \tag{18}$$

The Spearman CC of the two sequences is represented by $\rho(Y,\ Y_k)$ in equation (18), where, Y represents the historical IAQI data sequence of the target station organized in numerical downward order, Y_k signifies the arrangement of IAQI historical data of a surrounding station fixed in numerical descendant order, and N represents the integer of samples in the sequence. For example, ρ_list in equation (19) is the CC calculated between the target station and all stations.

$$\rho\_list = \ [\rho(Y^*, Y_1), \rho(Y^*, Y_2), \ldots., \rho(Y^*, Y_S)] \tag{19}$$

Compare the CC with the threshold ρ_{th}, and finally obtained the set M of stations with a CC greater than ρ_{th} as follows:

$$X = \ \{X_i | \rho(Y^*, Y_1) > \rho_{th}, i \in \ 1, \ldots, M\}x \tag{20}$$

The feature matrix of the *i*-the station with a strong spatial correlation with the target station is represented by $X_i \in R^{T \times L}$ in formula (20). The three-dimensional feature matrix with a strong spatial correlation with the target station is represented by $\in R^{M \times T \times L}$, where $M \leq S$, M denotes the number of stations with a strong spatial correlation with the target station, T is the time step of historical data, and L is the feature dimension.

Multi-step prediction model

The focus on developing a multi-step prediction model demonstrates a forward-thinking approach to forecasting. This is particularly relevant in air quality prediction, where anticipating future values is crucial for effective planning and regulatory decision-making. The MPSM built on the DLT is fed the data of atmospheric factors after PCA and historical AQI. Accordingly, the research employed the combination of MLR techniques with a TCN-centered DLM to foresee AQI as well as the $PM_{2.5}$, O_3, SO_2, CO, PM_{10}, and NO_2 concentrations from 2019 to 2022. The planned DLM stretches a specific value and is accurate for AQI prediction. The combination of DLM, TCN, and MLR suggests a thoughtful balance between accuracy and interpretability. While deep learning contributes to capturing intricate patterns, MLR provides a transparent framework for understanding the linear relationships between predictor variables and the predicted outcome. However, the use of a deep learning model (DLM) suggests a commitment to leveraging the capabilities of neural networks to capture intricate patterns within the data. The combination of TCN and MLR adds layers of thoroughness, ensuring that the model is equipped to handle different aspects of data complexity.

Multiple linear regression

MLR is a statistical method that allows the expectation of variability among a dependent variable and self-governing variables. Environmental data, especially in the context of air quality, can exhibit intricate patterns influenced by various factors. The chosen methodologies demonstrate the adaptability of the study to handle such complexity, combining deep learning with linear regression for interpretability. For investigating the statistical relationship among a reliant mutable and several independent variables, this method is widely applied in atmospheric modelling by fitting a linear equation to actual data and providing the contribution percentage of each parameter to atmospheric pollution. To measure the goodness of fit for developed MLR, the performance indicators were utilised. These indicators are the coefficient determination (R^2), an adjusted measure of determination (R^2_{adj}), and RMSE.

Temporal convolutional network autoencoder

For sequential modelling tasks, TCN was established on RNN and 1D CNN structure. TCN achieves better sequential modelling errands than the LSTM model, as demonstrated in recent studies. For better sequential modelling tasks, TCN employs causal convolution, which makes an expenditure of both historical and current data in each layer. To grow the RF with less computational cost, dilated convolution is presented in TCN. The widened convolution filter of the TCN model is practical by combining input standards with a specified step over a region greater than its size. The RF is guaranteed to cover previous information input using the dilation factor, which rises with network depth.

For an input sequence x of an element y, the convolution operation is defined as follows:

$$F(y) = \sum_{i=0}^{j-1} \quad f(i)\, x_{y-i} \qquad (21)$$

Both the previous input and convolution results are combined by TCN applying residual connection, for deep networks, it is very beneficial. The mapping of residuals from sequential input x to the transformation $f(x')$ is carried out via the residual block. This function has the following definition:

$$O = \varphi\ (f(x') + x) \qquad (22)$$

where: φ represents the nonlinear activation function.

To condense the several types of information in the sequential data set into a fixed-length vector, this research study modified the TCN model to function as both an encoder and a decoder, improving the stability of the tensor flow. Meteorological factors and geographical characteristics are crucial for AQI, $PM_{2.5}$, O_3, SO_2, NO_2, PM_{10}, and CO prediction. The spatial information and the available meteorological data are not measured in the traditional statistical, empirical PMs. To improve the accuracy of prediction as well as predict AQI, and O_3, SO_2, PM_{10}, NO_2, $PM_{2.5}$, and CO concentration values, this research proposed a model that makes use of geographical and meteorological data.

From the corrupted version, the information was reconstructed using the auto-encoder a neural network. Feature learning $h = f_\lambda(x)$ is defined by the autoencoder network, where, f_λ is the encoder function, the original information is reconstructed by the decoder $\hat{x} = g_\lambda(h)$.

The TCN encoder determines the local association of affecting meteorological and geographical variables by encoding consecutive inputs. Before modelling both global and local relationships, a decoder excerpts the encoded vectors for AQI, $PM_{2.5}$, O_3, SO_2, NO_2, PM_{10}, and CO forecasting. Hence, sequential modelling of AQI, $PM_{2.5}$, PM_{10}, CO, NO_2, O_3, and data from time series data on SO_2 concentration was performed more thoroughly than with a single neural network model. Accordingly, using his method, the AQI is accurately predicted. The integration of deep learning and traditional methods implies the potential for improved multi-step prediction accuracy. Deep learning models are known for their ability to learn complex representations, and the addition of TCN and MLR components enhances the model's capacity to capture both short-term and long-term dependencies. This multifaceted strategy enhances the model's ability to capture the temporal and interconnected nature of environmental dynamics, contributing to more accurate and reliable predictions.

RESULTS AND DISCUSSION

The presentation analysis of the suggested strategy using the applied dataset is represented in this section. The predictions are produced by utilising the MLR techniques with a TCN model for both the testing and training data to assess the routine of the model. In this, the data was gathered from the monitoring stations in Kanpur, Chennai, Hyderabad and Mangalore. The stations are situated at Jawaharlal Nehru Stadium, IITK, Kanpur, Kadari and Sanathnagar. The following research variables were obtained from these locations: wind speed (WS), atmospheric pressure, $PM_{2.5}$ levels, relative humidity (RH), SO_2, CO, O_3, NO_2, PM_{10}, temperature, humidity, weather, and rainfall. The results and conclusions were thoroughly examined and given in the following order: model assessment, temporal stability of the models, spatial stability of the models, and summarisation.

The first dataset employed is the Jawaharlal Nehru Stadium, Chennai, API data that was gathered over three years, 2019–2022. Figure 2 shows the plots for the testing and training data. A few steps are needed for training and prediction calculation, and the recommended classifier maintains training procedures. The first phase is entering the data for the second step, and the construction of a classifier system with a membership utility is the input stage. The third phase involves using a chosen learning algorithm to exercise presumptive input data. Iterative computations for test data and train scores were then performed following learning.

The goodness-of-fit between the MLR-TCN-AQI prediction for the AQI is shown in Figure 3. The estimated worth and the actual value were typically relatively fitted. However, there were occasionally certain variations in the predicted values for the abrupt deviations in AQI at specific times (1800 and 2200 in Figure 3). As can be seen in Figure 3, where the precise values of the API are extremely similar to the anticipated values, it can be realised that the suggested estimating model built on the challenging dataset of the API has very well modelled the air pollution. Overall, the model's prediction presentation

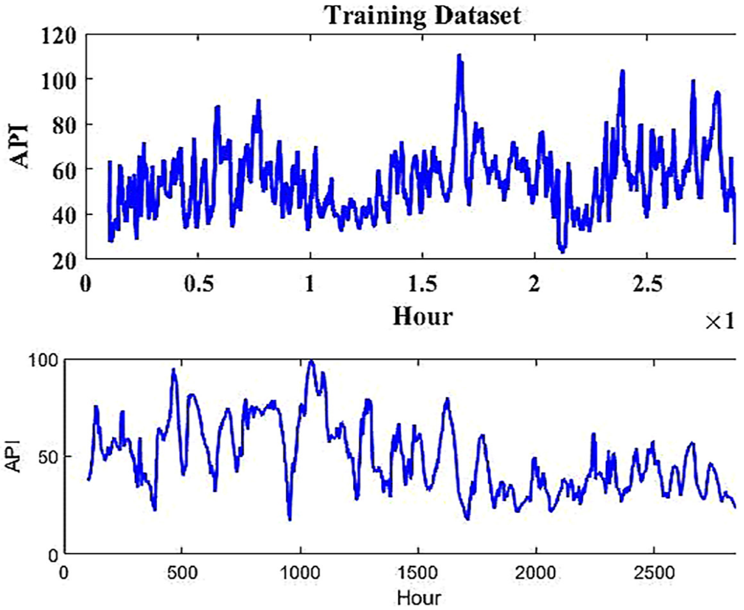


Figure 2. Time sequences plots the API standards of training and testing and datasets

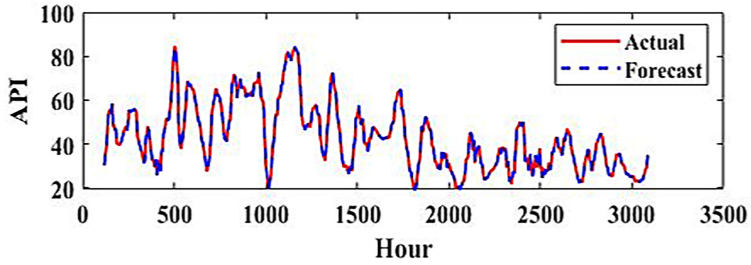
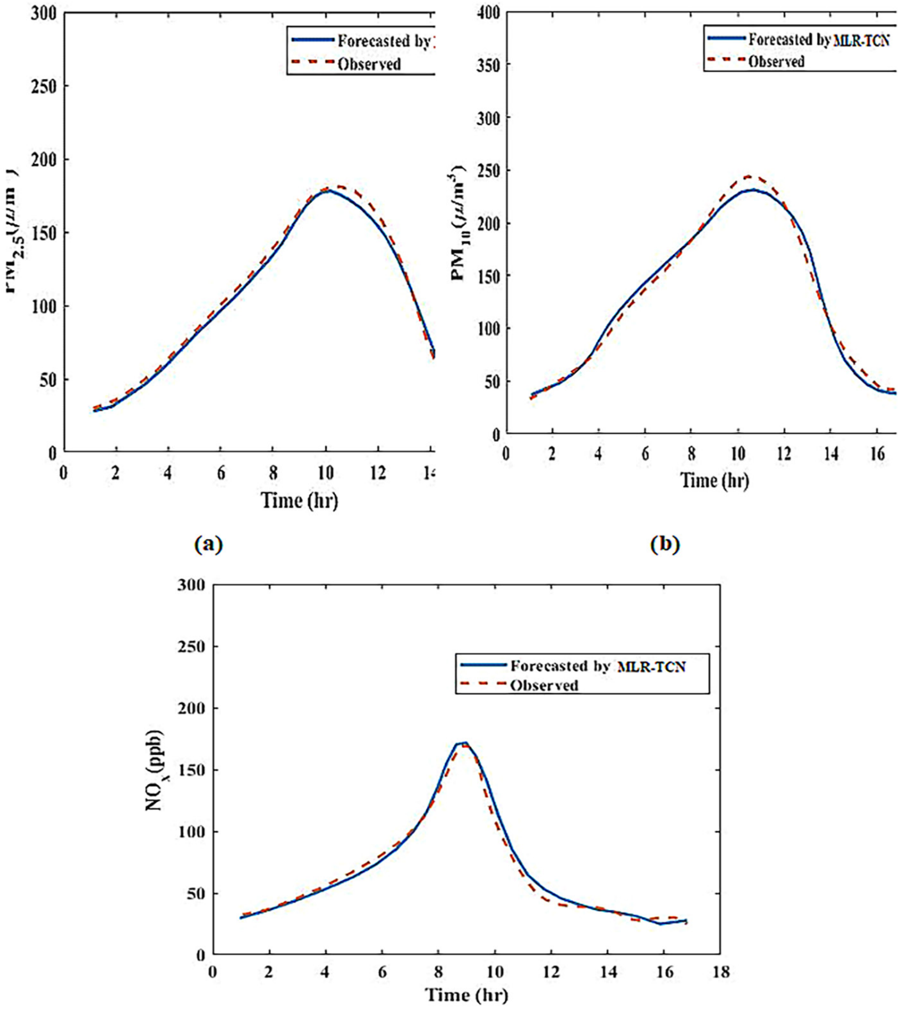


Figure 3. AQI actual and predicted results

in the stationary and nonstationary periods of the AQI was steady and accurate, indicating that the model has a well-predicted AQI.

Figure 4 illustrates that both models were tested by determining the goodness-of-fit between observations and forecasts at horizon t⊆4, an air pollution event with peak PM_{10}, $PM_{2.5}$, and NOx concentrations higher than 150 mg/m^3, 200 mg/m^3, and 150 ppb, respectively, at a traffic station. According to the consequences, the SM-LSTM model was unable to accurately forecast due to clear time lag phenomena and wider gaps between observations and forecasts, whereas the projected model of MLR-TCNA was able to predict the air value at horizon tξ4. The created model seemed to be able to make much more accurate and dependable regional multistep-ahead air superiority prediction, in addition to following the trails of air-worth events and significantly reducing time-lag effects.



(a)

(b)

(c)

Figure 4. $PM_{2.5}$, PM_{10} and NO_x forecasted results

Regression analysis

The regression models' presentation was assessed using RMSE, *r*, and R^2. The subsequent criteria could be employed as a guide to get a good regression model: low RMSE, high *r*, and high R^2. The standard nonconformity of the sample between the expected and observed values is represented by the RMSE. In basic statistics, the correlation and dependence type are quantified by the CC (*r*) value, which suggests the statistical relationships between two or more variables. According to how much the independent variable(s) can predict the difference in reliance on mutable, the percentage of determination (R^2) increases.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \quad (\widehat{y_i} - y_1)^2} \tag{23}$$

$$r = \frac{\sum_{i=1}^{n} \quad (y_i - \underline{y})(\widehat{y_i} - \underline{\hat{y}})}{\sqrt{\sum_{i=1}^{n} \quad (\widehat{y_i} - y_1)^2 \sum_{i=1}^{n} \quad (\widehat{y_i} - \underline{\hat{y}})^2}} \tag{24}$$

$$R^2 = 1 - \frac{\sum_i^n \quad (y_i - \widehat{y_i})^2}{\sum_i^n \quad (y_i - \underline{y})^2} \tag{25}$$

The prediction impact of the TCNN-AQI model predictive modelling is shown in Figure 5, demonstrating that the spatial-temporal dependencies among stations should not be disregarded in the study. Furthermore, the model exhibited a significant expectation error and deviation in fitting degree when TCNN-AQI was utilised to expect the AQI-PM2.5 of some interior positions. The station's more complicated environmental conditions, which are prone to several external disturbances like traffic conditions and manufacturing operations, could be one explanation. Future AQI PMs should consider these elements as they may have contributed to certain variations in the model's predictions.

The time chain plots of the AQI of the four stations are shown in Figures 6a to 6d. The time arrangement plot showed that data is non-stationary in all cases. The stationary data series for the seasonal pattern was obtained by taking D = 1 with S = 12, and the non-seasonal pattern was formed by taking D = 1. This predicts the seasonal pattern of four stations (Jawaharlal Nehru stadium, IIT Kanpur, Kadari and Sanathnagar) AQI with month-year 2019–2020.

The model was employed to estimate the AQI for the years 2019–2022, after the identification of the best-fit SARIMA. The actual and predicted values for the year 2019–2020 for the stations (Jawaharlal Nehru stadium, IIT Kanpur, Kadari and Sanathnagar) are publicized in Figure 7 (a-d). Error i.e. the alteration among most of the stretch, there is a difference between the actual and predicted values of –3 to +3, with larger errors noted in August 2019 (Error = -6) and August 2020 (Error = 8).

The seasonal variables that were time-dependent are presented in Figure 8 (a-d). Here, statistical analysis of the sample reveals that the dataset exhibits monthly seasonality. To effectively extrapolate the perfection of the air, seasonal parameters had to be continuously monitored. This seasonality idea is utilised in this case for the suggested seasonality-based imputation method. Two different MV types' single-hour MV and

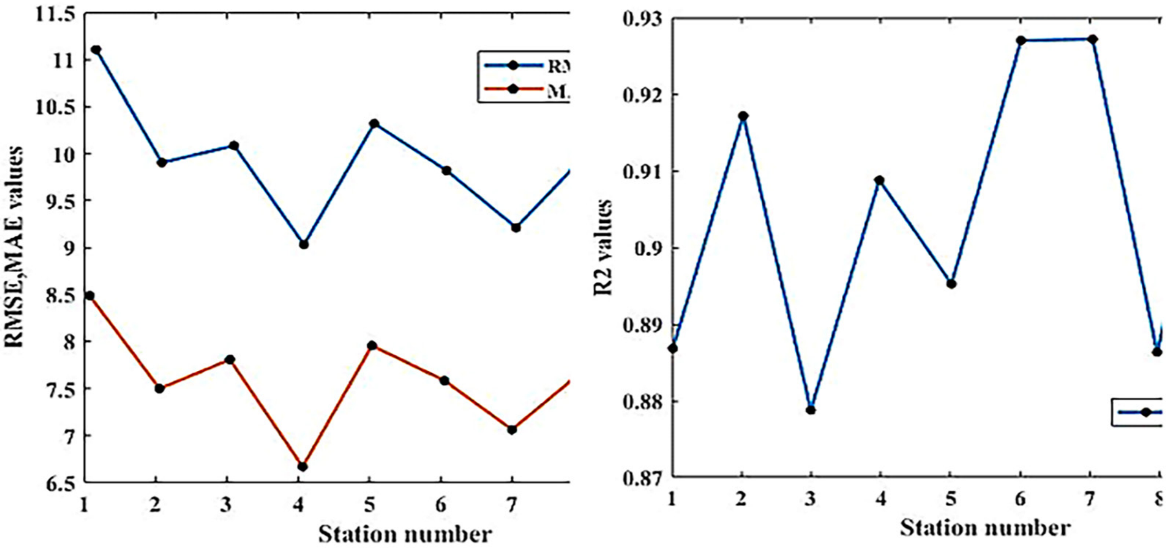
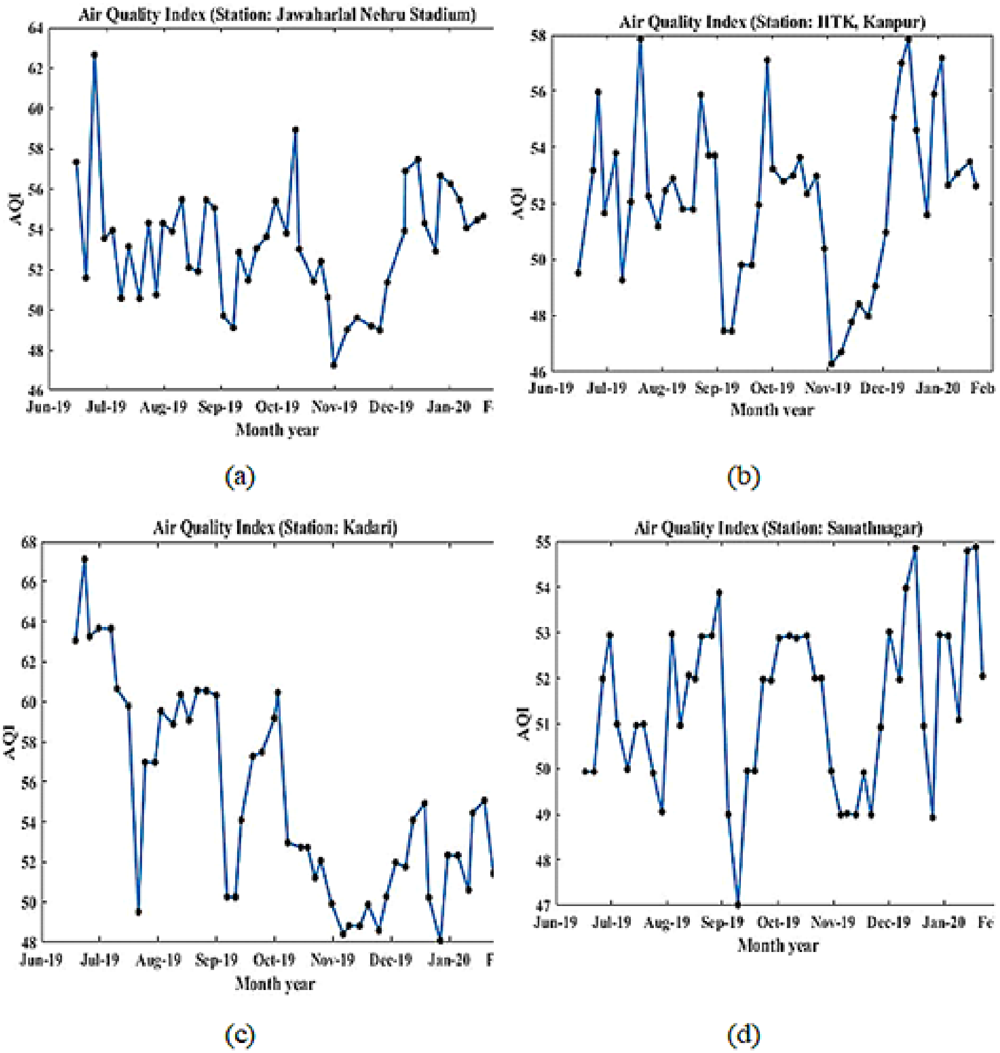


Figure 5. RMSE, MAE and R^2 analysis graph



(a)

(b)

(c)

(d)

Figure 6. Time sequences plot of different stations

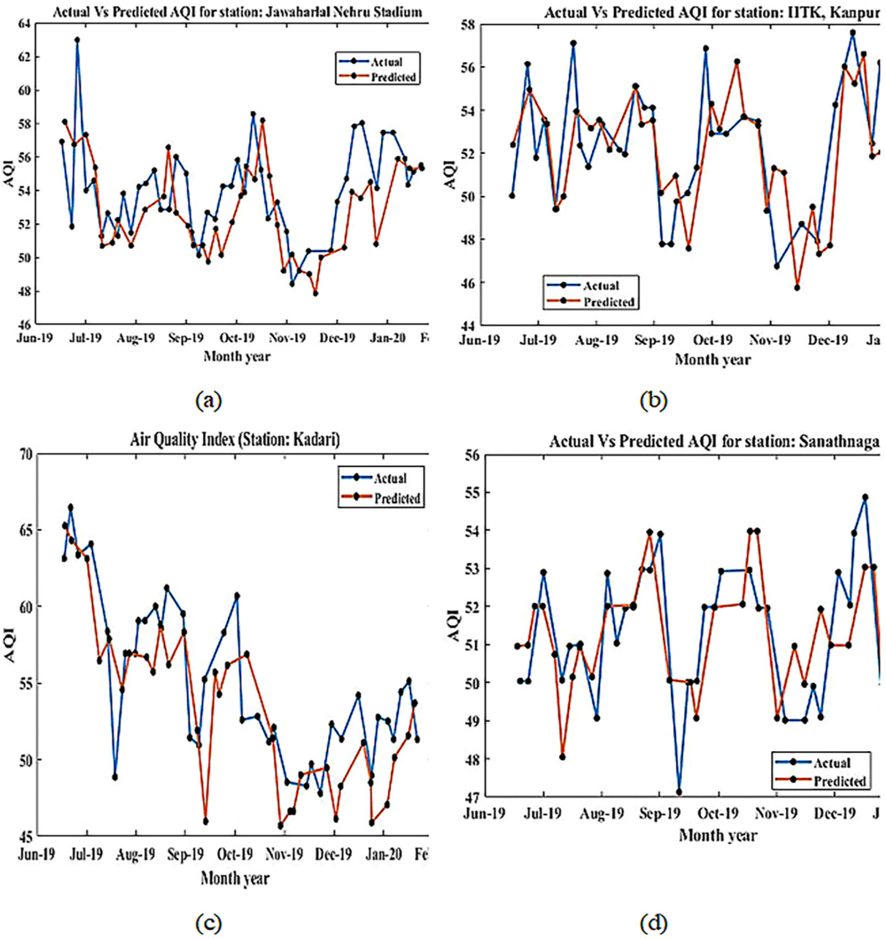
consecutive-hour MV are handled by this suggested algorithm.

Figure 9 shows the ACF plots for the training dataset to demonstrate the autocorrelation of the data. It can be perceived that the SARIMA models utilize the difficult dataset and training dataset, respectively. Figure 5 demonstrates the autocorrelation grid of the AQI series over time for the interval period of l (l = 0, 1, 2... 45) hours. As it can be observed from this graph, ACF continuously declines up to a lag of 17 hours before increasing up to a lag of 25 hours. It once more starts to decrease up to a lag of 34 hours starting at the 25th hour. This leads to the supposition that the AQI time series exhibits a 24-hour seasonality pattern.

Figure 10 shows the estimation of the AQI value for the SO_2 and the initial dataset for the SO_2 versus AQI index value in green and red, respectively. It was found that the model obtained a excellent fitting of both preparation and test data. For SO_2, the MSE and R^2 values reached 1.005 and -0.005, respectively. Figure 10 shows the AQI principles for the actual data and anticipated data for the SO_2.

Using the training and testing data, the suggested LSTM model will evaluate the NO_2 AQI value. Figure 11 shows the NO_2 versus AQI index value with the unique dataset highlighted in red, and NO_2 with the predicted AQI value highlighted in green. It was noted that the model achieved the fit that was excellent in both test and training data. For NO_2, the MSE value and R_2 criteria were determined to be 0.956 and 0.098, respectively.

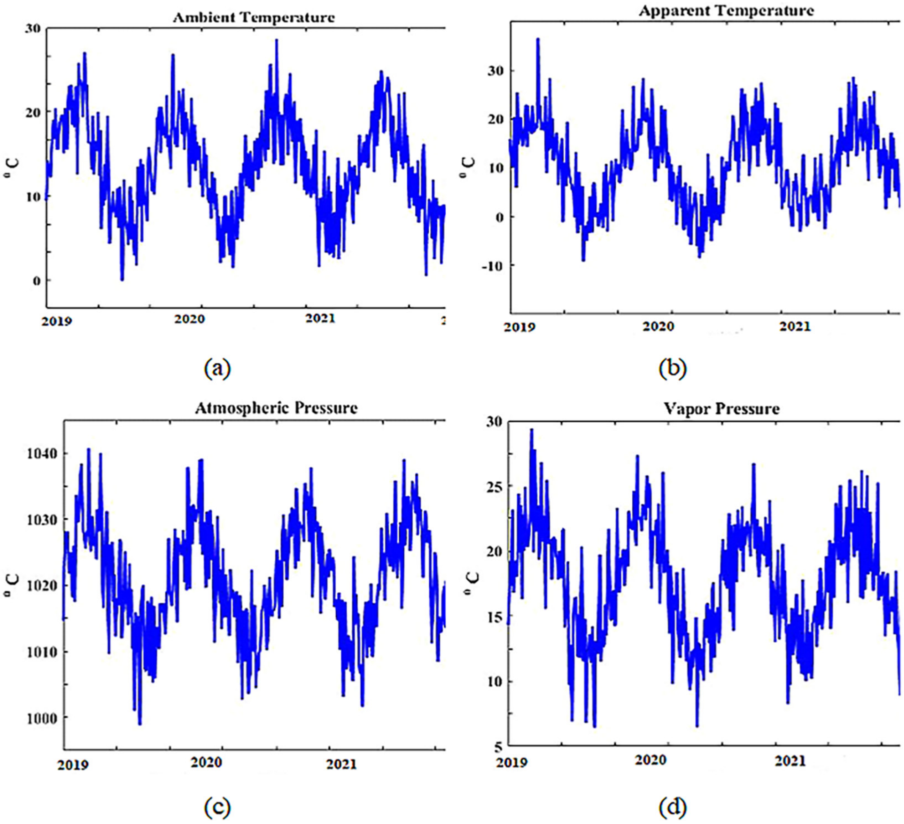
The planned model was utilised for testing and training data to estimate the success of the AQI for CO. Figure 12 shows the initial Dataset for the CO against the AQI index value in red and the predicted AQI value for the CO in green. It was observed that the model achieved an excellent fitment of both test and training data. The



(a) (b)

(c) (d)

Figure 7. Actual and predicted results of AQI with different stations



(a) (b)

(c) (d)

Figure 8. Seasonal factors estimation

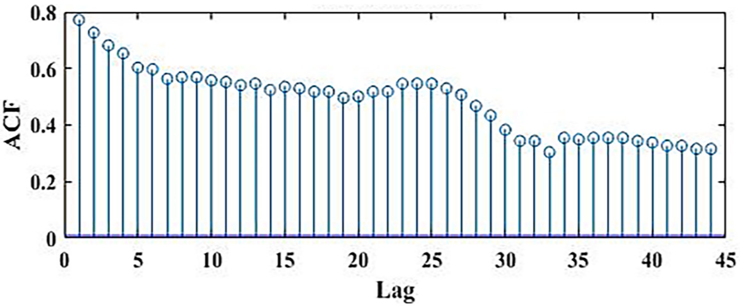


Figure 9. Autocorrelation for the training dataset

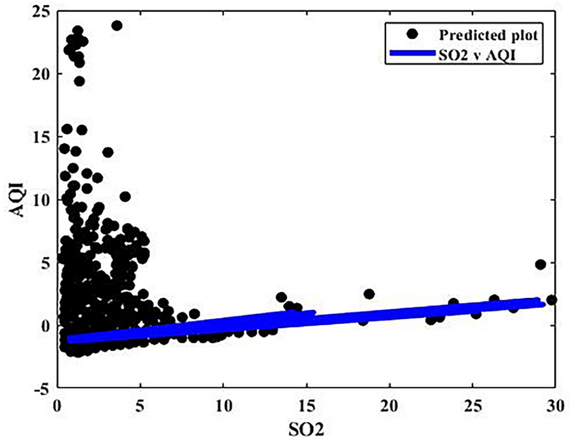


Figure 10. AQI SO_2 prediction

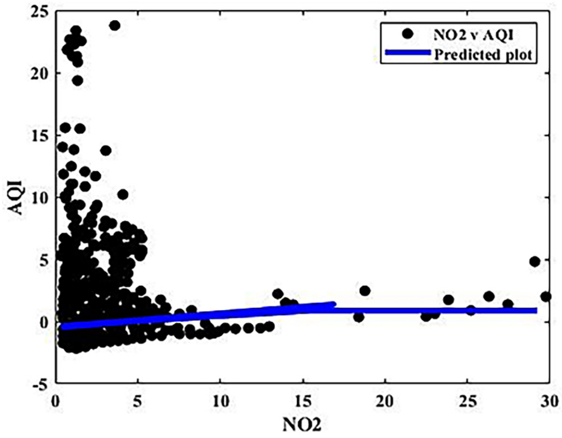


Figure 11. AQI NO_2 prediction

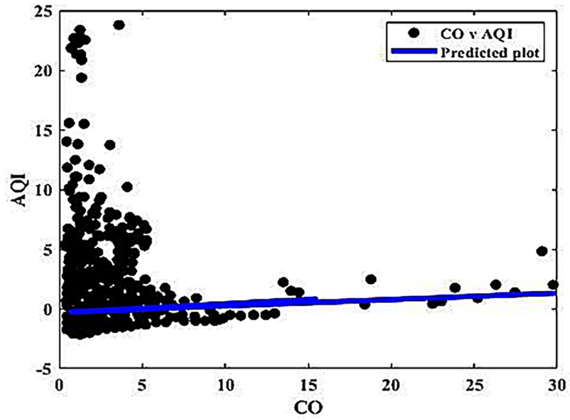


Figure 12. AQI CO prediction

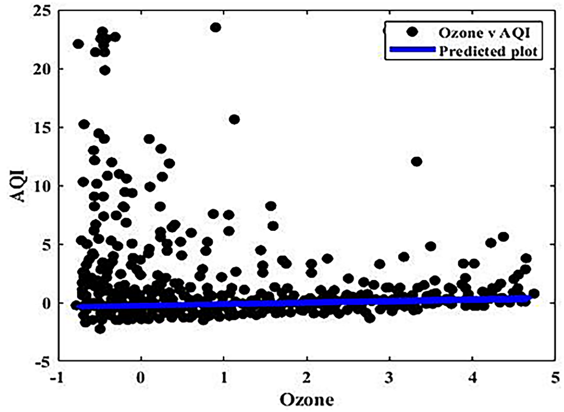


Figure 13. AQI O_3 prediction

MSE value was 0.920 and the R^2 value was 0.080. Figure 12 shows the CO AQI prices for the experimental data and predicted data.

To estimate the usefulness of AQI for ozone, the prototypical is employed using training data and testing data. In Figure 13, the actual dataset for the ozone versus AQI index charge is displayed in red, and the predicted AQI value for the ozone is displayed in green. It can be observed that the model does a fantastic job of fitting both the exercise and test data.

CONCLUSIONS

Accurate AQI prediction is the essence of the AQI, It has major implications for the regulation of air class and human fitness. Accordingly, the article planned a hybrid DL-based technique for AQI prediction. Data were collected from northern and southern cities of India including, Chennai, Kanpur, Hyderabad and Mangalore. From these cities, the data from four stations were collected and analysed using metrological as well as historical factors of SO_2, CO and O_3, NO_2, PM_{10} and $PM_{2.5}$, temperature, pressure, moisture, wind speed and weather, and rainfall. A TSFM constructed on SARIMA and SFTS was obtained. Spatio-temporal feature extraction was processed by utilising the SDCT and SRCC. These professionally extract the interval autocorrelation and identify the elements in the data set that were weakly connected by measuring the amount of relationship among the history sequences of nearby stations and the target station. The presentation evaluation metrics for the models are the MAPE, root RMSE, and computational period; these values were determined for every PM that was selected. The findings show that the recommended model outperforms the current traditional and sophisticated time series representations in terms of forecasting accuracy. Additionally, the proposed model demonstrates the capacity to resolve forecasting issues successfully to increase model accuracy, which is investigated in difference to the existing models to support its superiority. DLM accurately predicts the AQI values in northern and southern cities in India. By implementing coordinated traffic signals for the roads and promoting the custom of public transport, the predicted AQI value can reduce the degree of pollution.

REFERENCES

1. Abirami S. and Chitra P. 2021. Regional air quality forecasting using spatiotemporal deep learning. Journal of Cleaner Production. 283, 125341.
2. Al-Janabi S., Mohammad M. and Al-Sultan A. 2020. A new method for prediction of air pollution based on intelligent computation. Soft Computing. 24(1), 661-680.
3. Alyousifi Y., Othman M., Faye I., Sokkalingam R. and Silva P.C. 2020. Markov weighted fuzzy time-series model based on an optimum partition method for forecasting air pollution. International Journal of Fuzzy Systems. 22(5), 1468-1486.
4. Alyousifi Y., Othman M., Sokkalingam R., Faye I. and Silva P.C. 2020. Predicting daily air pollution index based on fuzzy time series markov chain model. Symmetry. 12(2), 293.
5. Iskandaryan D., Ramos F. and Trilles S. 2020. Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Applied Sciences. 10(7), 2401.
6. Cheng X., Zhang W., Wenzel A., and Chen J. 2022. Stacked ResNet-LSTM and CORAL model for multi-site air quality prediction. Neural Computing and Applications. 1-18.
7. Dairi A., Harrou F., Khadraoui S. and Sun Y. 2021. Integrated multiple directed attention-based deep learning for improved air pollution forecasting. IEEE Transactions on Instrumentation and Measurement. 70, 1-15.
8. Fernando R.M., Ilmini W.M.K.S. and Vidanagama D.U. 2022. Prediction of Air Quality Index in Colombo.
9. Heydari A., Majidi Nezhad M., Astiaso Garcia D., Keynia F. and De Santoli L. 2022. Air pollution forecasting application based on deep learning model and optimization algorithm. Clean Technologies and Environmental Policy. 24(2), 607-621.
10. Janarthanan R., Partheeban P., Somasundaram K., and Elamparithi P.N. 2021. A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society. 67, 102720.
11. Koo J.W., Wong S.W., Selvachandran G., Long H.V. and Son L.H. 2020. Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models. Air Quality, Atmosphere & Health. 13(1), 77-88.
12. Kristiani E., Lin H., Lin J.R., Chuang Y.H., Huang C.Y. and Yang C.T. 2022. Short-term prediction of PM2. 5 using LSTM deep learning methods. Sustainability. 14(4), 2068.
13. Krylova M., and Okhrin Y. 2022. Managing air quality: Predicting exceedances of legal limits for PM10 and O_3 concentration using machine learning methods. Environmetrics. 33(2), e2707.
14. Kumar K. and Pande B.P. 2022. Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology. 1-16.
15. Li H., Wang J. and Yang H. 2020. A novel dynamic ensemble air quality index forecasting system. Atmospheric Pollution Research. 11(8), 1258-1270.
16. Lin Y.C., Lee S.J., Ouyang C.S. and Wu C.H. 2020. Air quality prediction by neuro-fuzzy modeling approach. Applied soft computing. 86, 105898.
17. Liu B., Jin Y. and Li C. 2021. Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR–SVR–ARMA combined model. Scientific reports 11(1), 1-14.
18. Moscoso-López J.A., Urda D., González-Enrique J., Ruiz-Aguilar J.J., and Turias I.J. 2020, September. Hourly air quality index (AQI) forecasting using machine learning methods. In: International Workshop on Soft Computing Models in Industrial and Environmental Applications, Springer, Cham. 123-132.
19. Neelaveni N., and Rajeswari S. 2016. Data mining in agriculture-a survey. International Journal of Modern Computer Science—Revista da Faculdade de Serviço

Social da UERJ, Rio de Janeiro. 4(4), 104-107.

20. Samal K., Babu K.S. and Das, S.K. 2021. Spatio-temporal prediction of air quality using distance based interpolation and deep learning techniques. EAI Endorsed Transactions on Smart Cities. 5(14), e4.
21. Seng D., Zhang Q., Zhang X., Chen G. and Chen X. 2021. Spatiotemporal prediction of air quality based on LSTM neural network. Alexandria Engineering Journal. 60(2).
22. Tiwari A., Gupta R. and Chandra R. 2021. Delhi air quality prediction using LSTM deep learning models with a focus on COVID-19 lockdown. arXiv preprint arXiv:2102.10551.
23. Tomar N., Patel D. and Jain A. 2020, February. Air Quality Index Forecasting using Auto-regression Models. In: IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE. 1-5.
24. Wang W., An X., Li Q., Geng Y.A., Yu H. and Zhou X. 2022. Optimization research on air quality numerical model forecasting effects based on deep learning methods. Atmospheric Research. 271, 106082.
25. Yan R., Liao J., Yang J., Sun W., Nong M. and Li, F. 2021. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Systems with Applications 169, 114513.
26. Zeng Y., Chen J., Jin N., Jin X. and Du Y. 2022. Air quality forecasting with hybrid LSTM and extended wavelet transform. Building and Environment. 213, 108822.
27. Zhang K., Thé J., Xie G. and Yu H. 2020. Multi-step ahead forecasting of regional air quality using spatial-temporal deep neural networks: a case study of Huaihai Economic Zone. Journal of Cleaner Production. 277, 123231.
28. Zhang L., Liu P., Zhao L., Wang G., Zhang W. and Liu J. 2021. Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmospheric Pollution Research. 12(1), 328-339.
29. Zhang Z., Zeng Y. and Yan K. 2021. A hybrid deep learning technology for PM2. 5 air quality forecasting. Environmental Science and Pollution Research. 28(29), 39409-39422.