Sustainable Cities and Society 67 (2021) 102720

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

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

R. Janarthanan [a], P. Partheeban [b,*], K. Somasundaram [a], P. Navin Elamparithi [c]

[a] Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India
[b] Department of Civil Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India
[c] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirapalli, India

ARTICLE INFO

Keywords:
Air quality index
Deep learning
LSTM model
National Air Monitoring Program

ABSTRACT

In India, the Central and State Pollution Control Boards have commissioned the National Air Monitoring Program (NAMP) which covers 240 cities with 342 monitoring stations. Air Quality Index (AQI) has been categorized into different groups. To predict the AQI in Chennai city, the Dataset was collected, then preprocessed to replace missing values and remove redundant data. The mean, mean square error and standard deviation are extracted using the Grey Level Co-occurrence Matrix (GLCM). The combination of Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) based deep learning model is used to classify the AQI values. The proposed deep learning model gives an accurate and specific value for AQI on the city's specified location compared to the existing techniques. The prediction accuracy is improved in the proposed deep learning method, which will caution the public to reduce to an acceptable level. The deep learning mechanism predicts the AQI values accurately and helps to plan the metropolitan city for sustainable development. The expected AQI value can control the pollution level by incorporating road traffic signal coordination, encouraging the people to use public transportation, and planting more trees on some locations.

1. Introduction

In the 1950s, scientists began analyzing emissions in California, which was sponsored by the University of California. Albert Bush was the leading research scientist in Los Angeles who understood particles and their photochemical process. In 1966, a fog database was also built by the busy Initiative to collect air samples from cities in Australia, Africa, Asia, New Zealand, and Europe. Since then, work into air quality has culminated with the efforts of many scientists and scholars. Many of these scientists are transforming the world of science and the society. Each community in the world is slowly being integrated with data and connectivity technologies and the Internet of Things (IoT) systems to promote economic growth. The development of IoT can also be attributed to holding communities and infrastructure advancing as well. Air pollution regulations are strictly enforced in every nation to save people and the environment from poisonous gases. Metropolitan air quality is a significant cause of death as they are the reason for several respiratory diseases. According to the World Health Organisation, 70 % of the Metropolitan Areas of Emerging Countries' emissions are due to old cars, poorly maintained vehicles, poorly maintained roads, and low-quality fuel. With the rapid development and expansion of metropolitan cities' population, various environment-related problems have come to the forefront. Air pollution has adverse effects on people's health. Air quality monitoring systems are considerable effort towards reducing the impact of air pollution on people. Air Quality Index is a descriptive system and a tool used to communicate risks. It informs the public on the air quality level in their surroundings and, therefore the related potential health hazards, particularly for vulnerable groups like children, older people, and people with cardiovascular or respiratory diseases. The AQI is used to decide outdoor activities in general, particularly by individuals (Grace & Manju, 2019). Schools and sporting organizations can use these projects to check the latest AQI figures to decide whether sporting events can be held outdoors.

AQI usually (Grace & Manju, 2019) changes the weighted significance to a particular range or set of numbers for various pollution-related parameters (e.g., $PM_{2.5}$, SO_2, NO_2, CO_2, Ozone, visibility). This approach is used by many countries to express air quality and to make decisions. The Air Quality Index converts advanced air

* Corresponding author.
E-mail addresses: janarthananr@citchennai.net (R. Janarthanan), dean.pd@citchennai.net (P. Partheeban), soms72@yahoo.com (K. Somasundaram), parithi1999@gmail.com (P. Navin Elamparithi).

https://doi.org/10.1016/j.scs.2021.102720
Received 18 September 2020; Received in revised form 10 December 2020; Accepted 14 January 2021
Available online 20 January 2021

pollutant quality data into a number, word and value. AQI is measured by numerical values from 0 to 500. When the values are 0, it is the normal adequate air quality and 500 (higher AQI pollution) are the most hazardous air quality. AQI is classified into six groups centred on the environmental abundance and health impacts of air pollution (identified as safety interruptions). These six grades are the results: Good (safe), Moderate, Sensitive, Unhealthy, Very Unhealthy, Dangerous. Economy, food development, transport and its use are the human society's practices that influence the climate. The air quality is affected by these activities, either directly or indirectly.

The management of air quality comprises of all activities aimed at environmental air quality management. Control of air quality aims to ensure that the environment is clean enough to ensure the public health and environmental safety. This cycle removes the health impacts of air quality mentioned in the driving force, pressure, state-exposure, effect action chain from the roles of economies and society in general. The chain stresses that the community can take action on every link in the chain to minimize harmful health effects (NandigalaVenkatAnurag & Sharanya, 2021). Air quality monitors have a role to play in supplying knowledge on environmental quality concentrations. These are then used to evaluate population exposure and quality adverse health effects. When ecological threats are too high, the steps to reduce pollution and protect the atmosphere are appropriate. The significance of an accurate air quality measurement, the value of the AQI is used. This method renders the details of multiple contaminants as a single number or amount. AQI has six categories: fine, satisfactory, mildly infected, bad, very poor and serious - measures of Reform AQI. Hence to improve the air quality some measures are represented as follows,

- A clean house may be a healthier house, because good indoor hygiene can greatly cut down on dust and animal dander
- Keep the greenery outdoors.
- Change your filters.
- Invest in an air purifier.
- Let the fresh air in.
- Disclaimer:

2. Related works

(Al-Janabi, Mohammad, & Al-Sultan, 2020a) Suggested intelligent predictor for air pollutant concentrations over the next two days based on the Recurrent-Neural Network (RNN) profound learning techniques. The process was calculated using the PSO-algorithm. Smart Air Quality Prediction Model (SAQPM) is the latest predictive computing indicator focused on unattended data, i.e. Long-Short Term Memory (LSTM) and optimisation (i.e. PSO). The key goal is to estimate six rates of six forms of pollutants in air quality.

(Cabaneros, Calautit, & Hughes, 2019) Interactions are proposed between concentrations of air quality and other explanatory contaminants. Finally, powerful tools can develop and implement air quality control, which is less complicated. Predicting the ambient air quality is done by using Artificial Neural Networks (ANNs). The respiratory illnesses and early mortality of elderly members of the population have been caused by inadequate air quality in metropolitan areas. The factors involved, including the scale, quality and computational costs of the parameters concerned depend on large air quality databases.

(Kala, Joshi, Agrawal, Yadav, & Joshi, 2020) There has been a significant increase in population growth, urbanisation, and air pollutants (APs) in the environment, posing serious threats to health. The AP data offers details on environmental management's quality of air and health protection threats in the region. The whole world has a serious energy shortage and the air quality challenge is on the increase. Renewable energy production is thus the greatest priority above all other preventive capital.

(Yadav & Nath, 2020) The nature of the air reported has important environmental and human health impacts. Therefore, details on ambient emissions must be estimated to include previous safety reporting knowledge on their environmental concentration. Most urban areas were mainly affected rapidly by air quality and the energy shortage problem.

(Maleki et al., 2019) The potential for the hourly estimation of air contaminant concentration parameters and two air quality index, Air Quality Index (AQI) and Air Quality Health Index (AQHI) were tested by the artificial neural network (ANN) algorithm. Air quality often has negative health, socio-economic, agricultural and political consequences. Meteorology and pollutant sources are two fundamental factors affecting the air over quality. They may be used to predict spatiotemporal pollutant profiles and air quality index in computational approaches.

(Kelly et al., 2019) The National Air Quality Standards (NAAQS) have been identified as useful in regulatory assessment exposure characteristics. Air quality is impossible to prepare specifically for NAAQS for certain situations because of the computational costs involved with Photochemical Grid Models (PGM) widely used for air quality prediction. Emission is reduced only while air quality modelling models for both decreased and enhanced pollution encourages wider use. In combination pollution reductions case applications, which showed health risk assessment of air quality through the typical prediction of toxicity demonstrated for quality system prediction neglect nonlinear interactions in chemistry (Maciąg, Kasabov, Kryszkiewicz, & Bembenik, 2019). It suggested inadequate air pollution would kill several thousand people. Air quality studies and their effect on people's wellbeing show a strong association between high accumulation and decreased mortality. To forecast high-efficiency air quality data, one might suggest utilising ensemble models that considered a much better output prediction/classification relative to current prediction models.

(Wen et al., 2019) The forecast of mass concentrations has a crucial role to play in taking decisions on atmospheric resources. This results in adverse health effects, such as excessive cardiovascular and respiratory disease morbidity and mortality. For this cause, it is crucial to avoid air pollution in advance by improving air quality protection and ensuring effective environmental monitoring. This is very critical for people's everyday safety and the government's decision-making on air quality regulation.

(Ghasemi & Amanollahi, 2019) It is reported that development and production of seeds is decreased by air pollutant under the O_3 influence. Several reports also identified serious questions regarding trans-boundary air quality in multiple areas of the world. The results of the air quality surveillance model for air quality prediction are still not fully convincing. Most recently, the emphasis was on evaluating the outcomes of different air quality prediction models (Honarvar & Sami, 2019). It is very important to measure air quality and use the results of air quality to predict and detect the relations between different urban problems. Air quality has influenced urban sustainability. The data on air quality in real-time, including Ozone's, particulates, carbon monoxide, sulphur dioxide and nitrous oxide, are vital for regulating air pollution and shielding citizens from harm.

(Li, Dong, Zhu, Li, & Yang, 2019) The sensitivity study has demonstrated that the air pollutant concentration is commendably configured. The dilemma of various international air quality standards avoiding the breaking-point effects of AQI can guide human health against air quality exposure.

(Liu et al., 2019) People have stated that their air quality is severely poor. Nonetheless, there are no successful methods to tackle heavy smog completely despite the present state of technologies. Therefore, it is possible to forecast how much PM2.5 is absorbed and to inform citizens about taking steps in time to stop damaging residents by excessive smog. Air quality data is sequential, so models which are good for sequence processing data like a recurrent neural network (RNN). A dynamic prediction model was developed to forecast the PM2.5 in urban areas of cold regions (Xiao et al., 2020). This research systematically examines the vertical distribution of particulate matter in intensely cold urban

settlements. It offers a theoretical basis and technical support for further improvement of the air quality of urban settlements.

(Lei, Monjardino, Mendes, Gonçalves, & Ferreira, 2019) This proposed cumulative access to air pollution and NO2 particulate matter. Natural environmental quality also has morbidity consequences. Multiple neurological abnormalities have been linked to air pollution treatment. Nevertheless, even low rates of indoor contaminants, such as carbon monoxide (CO), are related to neurological symptoms. Exposure to air Quality with construction properties such as geometry and design, permeability and ventilation components affect the penetration indoors of external quality and removal of internal quality.

A review conducted on smart building and its performance to review the features and function of smart buildings (Dakheel, Del Pero, Aste, & Leonforte, 2020). They have also reviewed the method of achieving smart building goals. Further, the identified nine groups of performance indicators required improvement. They concluded that there is a gap to improve the performance of smart buildings.

(Xu, Shan, Li, & Zhang, 2020) Including air polluting criteria and comprehensive indexes, the popular air quality evaluation indices stated. Air pollutants including PM2.5, PM10, SO_2, NO_2, CO and O_3, are described in micrograms per cubic meter or part by million, estimated in new standards of air quality and higher in number, higher risks to health and prevention. A study conducted for Dublin city to model the different traffic management strategies on air pollution and public health. They have identified the possible areas affected in Dublin city due to road traffic (Tang, McNabola, & Misstear, 2020). Further, they established the relation between health and traffic pollution.

Air pollution in major cities in India is an important ecological issue. The problem of air pollution has significant environmental and human health effects (Zhang, Wargocki, Lian, & Thyregod, 2017). The main concern is atmospheric pollution caused by industrialised behaviour, traffic and road obstruction. Specific combinations of contaminants and their mechanisms or physical encounters with certain environmental materials, earth's atmosphere and the spatial properties are the main explanation for the extent of air emissions. The three principal pollutants causing environmental air quality damage include PM10, SO_2 and NO_2 (Agarwal & Melkania, 2018). The neural network method was used to identify the extremely complicated model and to overcome the question in the context of loud datasets. ANN's models have been successful in combining them with conventional deterministic modelling approaches. This model and the estimation of normal specific concentrations of PM10, NO_2 and SO_2 through the use of meteorological variables (input) was used for backpropagation algorithms (Asghari & Nematzadeh, 2016). (Gia et al., 2019) A network that consists of a sensor node, Edge gateway, LoRa repeaters, Fog gateway and Cloud servers and terminal apps for end-users are introduced and implemented. At the edge layer, a CNN-based image compression method is used to send information on hundreds or thousands of sensor nodes within the gateway range in a single message. They use sophisticated compression strategies in a recent IoT storage scheme to minimise data size to 67 % by less than 5% with a decompression flaw.

In another study conducted using a Lag-FLSTM (Lag layer-LSTM-Fully Connected network) model based on Bayesian Optimization (BO) for multivariant air quality prediction (Ma et al., 2020). They have arrived at the results of 23.86 % lower RMSE than other methods. Their method can automatically optimise the different metrological features and other pollutants affect the prediction of PM2.5. A research was carried out on deep learning based IoT for secure smart City infrastructure (Singh, Jeong, & Park, 2020). Al-Janabi and Alkaim (2020) here the author attempts to discover the potential to construct a new method to approximate missing data settings called Random Forests and Local Minimum Squares (DRFLLS). Seven categories of similarity measures were described by designing random forest algorithms. These types are personal coefficient of similarity, basic resemblance and fluid similarity (M1, M2, M3, M4 and M5). They are enough to predict the optimum number of missed values neighbourhoods in this application. The missing values have been determined by local minimum squares (LLS). In different ways, the precision of imputation can be measured: Pearson correlation (PC) and NRMSE. Then the higher PC value and lowest NRMSE value are correlated with the optimum number of neighbourhoods. In this research, they evaluated the performance of infrastructure for a smart city. Further, they compared quantitative analysis and security and privacy analysis with different measures. Finally, it is concluded that the evaluation implementation improved the performance of infrastructure (Al-Janabi, Mohammad, & Al-Sultan, 2020b). The goal is to establish an insightful prediction of air pollutant levels over the next two days, using a recurrent neural network (RNN) to deep learning techniques. A particle swarm optimization (PSO) algorithm is used to decide the better structure for its operation. Smart air quality prediction model (SAQPM) is the latest smart estimation forecast focused on unattended, e.g. long-term storage (LSTM) and optimization (i.e., PSO). (Al-Janabi & Mahdi, 2019) Presents each of the general features, grammar, benefits and drawbacks. The research relies, most notably, on the parameters used to construct a prediction model for each of them. Furthermore, it is another job to identify the strategies by their primary and secondary criteria. In order to identify the best sharing of these parameters between the strategies, the presence and absence of parameters is also compared. Al_Janabi, Yaqoob, and Mohammad (2019) here the main purpose of this proposal is design predictor to accurately forecast air quality indices (AQIs) of the future 48 h. Accurate predictions of AQIs can bring enormous value to governments, enterprises, and the general public -and help them make informed decisions. Here, there are several other existing methodologies on indicating the air quality but none of them shows accurate prediction of air quality. Here the air quality prediction can be affected by the meteorological conditions, atmospheric diffusion, and geographical features. Hence there is a need of an effective system to overcome all the existing issues on predicting the quality of the air.

3. Proposed work

This section is the deliberation of the detailed implementation of the proposed mechanism. Initially, the Dataset collected has been preprocessed to replace the missing values and to remove redundant data. Then the feature extraction is carried out with the use of GLCM technique to extract the features like mean, standard deviation and so on. The extracted features are then optimised with the help of Modified Fruit fly optimisation (MFOA) scheme to get an optimised selection of extracted features. Then the classification mechanism is carried with the use of deep learning mechanism SVR with LSTM model, which in turn predicts the AQI of the preferred location.

The Air Quality Index might be a describing structure and an indispensable mechanism for risk communication. It reports the common public regarding the ambient air quality level. As a result, the promising health risk might inflict, particularly on susceptible groups like children, the people, and older with living diseases like respiratory and cardiovascular.

AQI is regarded as 'One Number- One Colour-One Description' for an ordinary man for judging the air quality in their respective district. Different health impressions from exposure to the outdoor air pollutants were the intricate function of pollutant concentrations and compositions. The major outdoor air pollutants in the cities are Sulphur dioxide (SO_2), Particulate matter (PM) or Particle, Ozone (O_3), Carbon Monoxide (CO), volatile organic compounds (VOCs), pesticides and metals, Nitrogen Oxide (NO_2). The index formulation was an idea in Swachh Bharat Mission (Cleanness Mission). The air quality measurement depends on eight pollutants, which are

- Particulate matter (size less than 2.5 μm) or ($PM_{2.5}$),
- Ozone (O_3),
- Sulphur Dioxide (SO_2),
- Particulate matter (size less than 10 μm) or (PM_{10}),

Table 1
AQI value and conforming ambient concentrations for the identified eight pollutants.

AQI Category, Pollutants and Health Breakpoints								
AQI Category (Range)↓	→ Categories for the various readings of the pollutant based on the health breakpoints or health impacts							
	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5 –1.0
Moderately polluted(101–200)	101–250	61–90	81–180	101–168	2.1- 10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748*	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

The AQI Index values and their accompanying health impacts are as given in Table 2.

Table 2
AQI index value and their associated health impacts.

AQI	Associated Health Impacts
Good (0–50)	Minimal Impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people.
Moderately polluted (101–200)	May cause breathing discomfort to people with lung disease such as asthma, and discomfort to people with heart disease, children and older adults.
Poor (201–300)	May cause breathing discomfort to people on prolonged exposure, and discomfort to people with heart disease
Very Poor (301–400)	May cause respiratory illness to the people on prolonged exposure. The effect may be more pronounced in people with lung and heart diseases.
Severe (401–500)	May cause respiratory impact even on healthy people, and serious health impacts on people with lung/heart disease. The health impacts may be experienced even during light physical activity.

- Nitrogen Dioxide (NO_2),
- Carbon Monoxide (CO),
- Lead (Pb), and
- Ammonia (NH_3)

3.1. Pre-processing

Initially, the data are retrieved from the Dataset and are pre-processed to remove unwanted data. Thus, a filtering technique is applied to pre-process the input data, which eliminates irrelevant data. For the pre-processing stage, the normalisation approach is used, which is much more effective in the removal and replacement of unwanted and missing data. The main advantage of these approaches is based on the hypothesis that collecting predictions from classifiers could provide better class noise detection and is suitable for input data. Thus, the raw data is pre-processed to get noiseless data for further processing which suppress valuable information every so often or leads to the information loss (Table 1).

3.2. Feature extraction using GLCM

Feature extraction approach is computed in which the features are extracted from the data. In the analysis of statistical texture, texture features are evaluated from the numerical distribution of the observed intensity combinations in a particular place on comparing each one in the data. As per the number of intensity pixels (points) in all grouping, data are categorised as first-order, second order and higher-order statistics. The method Gray Level Co-occurrence Matrix (GLCM) is a technique of extracting a second-order statistical texture feature. This has been employed in a wide range of applications. Higher and Third-order textures are regarded as the relationships between three or more pixels. These are possible hypothetically but not implemented commonly because of interpretation difficulty and calculation time. Hence, the feature extraction based on GLCM is an effective approach for extracting features.

The features can be extracted by using the GLCM. The Grey level Co-occurrence Matrix (GLCM) approach is a way to remove statistical texture characteristics of second order. The technique has been used in many applications. In contrast, third-and higher-order textures find the interaction of three or more pixels. The GLCM is an arithmetic purpose, which normally can effectively remove the artefacts. The picture quality can also be clearly distinguished. The picture can be isolated for the process of study. Use the GLCM with this to remove the feature's functionality. In an exact differential area, GLCM can specify the frequency of the pixels. The single-pixel is to be questioned here and another pixel is to be known as the Ø route l and the adjacent value detachment of m. Usually, m obtains a single value, and Ø can benefit directionally. Then the obtained directional value can remove the attributes of the images used for the segmentation process. The GLCM process may be set as follows:

$$\mathrm{R}(\mathrm{m},\mathrm{n}) = \mathrm{G}(\mathrm{m},\mathrm{n},\mathrm{o},\ \emptyset) \Big/ \sum_{m=1}^{H}\sum_{n=1}^{H} G(\mathrm{m},\mathrm{n},\mathrm{o},\ \emptyset) \tag{1}$$

Where G is the frequency vector, m, n, o is the frequency of the particular component will generally having the pixel values of l and mR represents the features of an image, (m,n) is the component of the m and l, $\emptyset$ represents the normalised constant.

By implementing the GLCM, the different attributes can be obtained.

Entropy:

This includes general information about the objects used to compact the images.

$$\text{Entropy } = -\sum_{m=1}^{H-1}\sum_{n=1}^{H-1} R(\mathrm{n},\ \mathrm{o}) * \log(\mathrm{R}(\mathrm{n},\ \mathrm{o})) \tag{2}$$

Where R(n, o) is the frequency of the features R, H represents the fixed constant.

Angular moment:

It can be computed by summarising the gained values using the GLCM to determine the high or low homogeneity of the image. If the precision is low, the angular moment will be increased. The images will usually be measured for uniformity.

$$\text{Angular moment} = \sum_{m=1}^{H-1}\sum_{n=1}^{H-1} R(\mathrm{n},\mathrm{o})\hat{2} \tag{3}$$

Contrast:

The amount of intensity of the images is calculated. In general, the difference between the areas is assessed

$$\text{Contrast} = \sum_{n=0}^{H-1} n\hat{2}\{\sum_{n=1}^{H-1} R(n,o)\} \tag{4}$$

Inverse difference moment:

It is also used to measure homogeneity in general.

$$\mathbf{IDM} = \sum_{m=1}^{H-1}\sum_{n=1}^{H-1}\frac{1}{1} + (n-o)\hat{2}R(n,0) \tag{5}$$

Energy:

It could be used to evaluate whether the return is feasible with as many of the square components.

$$\mathbf{E} = \sum_{0=1}^{H-1}\sum_{n=1}^{H-1}R(n,o)\hat{3} \tag{6}$$

Variance:

The deviation of the grey level values from the mean can be calculated directly

$$\mathbf{VAR} = \sum_{m=1}^{H-1}\sum_{n=1}^{H-1}(R(n,o))\hat{2} - \emptyset\hat{2} \tag{7}$$

Sum average:

The frequency connections between pixels may normally be determined

$$\sum_{n=0}^{2H-1} nR_{x+y(l)} \tag{8}$$

After extracting the features, the features can be visualised. The key element number is less than or equivalent to the initial parameters. Correlation says that information is redundant, and data can be compressed if this redundancy is reduced.

3.3. Classification using deep learning mechanism SVR with LSTM model

The major downside with the Recurrence Neural Networks (RNN) is their incapability to preserve memory. On behalf of extensive sequences, they might have a hard time bringing information from the previous phase to deal with the later ones. Therefore, the fading gradient difficulty is faced by the recurrent neural networks. In this learning phase, weights of the network are restructured concerning the partial derivative of error function which concerns the current weight. At some iteration, this updating of weight at which the gradient might be small extremely is prohibited from doing so. This, in turn, forces the network to stop learning more. Therefore, the recurrent neural network, in turn, loses its memory.

This drawback of a recurrent neural network is overcome utilizing LSTM. The models of LSTM are very powerful, mostly for the retention of long short-term memory. It is an effectual learning algorithm which can seem at the data series past and forecast what the future elements of series will be in an accurate manner. SVR is developed using basic SVM. The basic idea of SVM is to map the training data through a function from the input space into a higher dimensional feature space, and then build a separate hyperplane with a maximum margin in the feature space. The regression problem idea is to determine a function which can precisely approximate future values. SVM and SVR were widely used in the prediction models.

LSTM has the memory blocks which relate to each other over layers. The block contains gates that choose the block state. The gates are accountable for forgetting or remembering the information at training. This will be accomplished with the use of a sigmoid function. This function value squishes among 0 and 1. Once the data is multiplied by 0, it is forgotten and while multiplied by 1, it is remembered.

The LSTM algorithm used in our proposed system is described as follows:

Training the LSTM algorithm

Step 1: Preprocessing of particulate matter and meteorological data

- Inspect, visualise and clean the Dataset
- Normalise the Dataset and set the look back

LSTM training

Step 2: Construct LSTM network with one input, four hidden layers and an output layer with single value prediction

- Apply the sigmoid function for LSTM layer
- Train the network with epochs = 64 and batch size = 32

Step 3: Obtain prediction results for test dataset using the trained model

LSTM Network for Regression

The network consists of three layers having the visible layer with one input. The block hidden is pretended of four LSTM units and output layer creates a single value prediction. The information from the Dataset is afterwards fit into the model. From this, the performance of the test and train datasets could be estimated. Henceforth, the model is employed to formulate predictions on the train and test datasets together. From this, the model's visual skill is identified.

Algorithm 1 (SVR with LSTM model)

Input: Processed data S_{data}
Output: Classified data C_{data}
Initialize the multi-Network layers
Initialize train features T_{fea}
input size i_{size}***=1***
No of hidden units h_{units} ***=100***
No of classes N_{class}***=4***
maxEpochs ϵ_{size}***=100***
minibatch size bat_{size}***=27***
Initialize label I_{label}
Train label =80%
Tet label =20%
initialize the layers I_{layers}
initialize the options $I_{options}$
Label=unique(label)
For ii=1:lengh(Lab)
Class=find(label== Lab (ii))
label I_{label}***=categorical(***I_{label}***)***
net=trainNetwork(T_{fea}***,*** I_{label}***,*** $I_{options}$***)***
Traincut=length(class)-traincut
Traindata=[traindata; trainfeatures;class(1: Traincut)end-5:end]

Predict label=classify(net,traindata, bat_{size}***)***
End
End
For ii=1:size(traindata,1)
Traindata=[traindata; trainfeatures;class(1: Traincut)end-5:end]
End
For ii=1:size(trainfeatures,1)
Traindata=[trainfeatures;
trainfeatures;class(1: Traincut)end-5:end]
End

4. Performance analysis

This section is the representation of performance analysis of the proposed scheme with the Dataset used.

4.1. Dataset description

In this, the data was gathered from the 3 Central Pollution Control Board (CPCB) monitoring stations in Chennai city. The stations are situated at Manali, Velachery and Alandur. The investigative variables collected from these places are relative humidity (RH), PM2.5 values, atmospheric pressure (BP), wind speed (WS) and wind degree (WD). The collected data are presented in 15 min intervals intended for the time of 00:00, 01 May 2019 to 23:59, 30 April 2020 and all station yielded a dataset containing 35039 data rows which total a 105117 data row. The

	PM2.5 (Std60)	NO2 (Std80)	SO2 (Std80)	CO (std4)	Ozone (Std180)	AQI
0	32.20	13.81	4.13	0.99	38.44	24.439444
1	30.38	14.42	2.56	1.14	56.70	26.371667
2	25.83	14.16	1.47	1.02	57.37	23.991944
3	23.02	13.83	0.37	1.10	55.94	22.938889
4	28.47	17.14	0.05	0.66	54.39	23.130833

Fig. 1. Dataset used for study – First Five rows of the Dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21707 entries, 0 to 21706
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   PM2.5 (Std60)   21707 non-null  float64
 1   NO2 (Std80)     21707 non-null  float64
 2   SO2 (Std80)     21707 non-null  float64
 3   CO (std4)       21707 non-null  float64
 4   Ozone (Std180)  21707 non-null  float64
 5   AQI             21707 non-null  float64
dtypes: float64(6)
memory usage: 1017.6 KB
```

Fig. 2. Details of the Dataset used in the study.

PM2.5 missing values were 78.28 % approximately. The data was then processed for removing any rows that had empty columns and the information was then limited to rows that had levels of PM 2.5 below 250 μg/m^3. This data left reduced to 22827 data rows as certain elements were not there in all the other rows.

The meteorological features like $PM_{2.5}$, NO_2, SO_2, CO and Ozone were gathered for every 15 min from 3 stations all through Chennai. The collected data for one year from a period 01 May 2019, to 30 April 2020, along with each station that contributes 35039 information rows and therefore 490546 rows totally in the Dataset. The details of the Dataset which are used for analysis is shown in Figs. 1 and 2. The implemented Data sets can hold information to be used by a program running on the system. Data sets are also used to store information needed by applications or the operating system itself, such as source programs, macro libraries, or system variables or parameters.

4.2. GLCM based feature extraction

The dataline time component contained in our dataset was used to obtain new features, valuable to help tease out series seasonality information. Considering all the below features can be considered as a input features for the purpose of the classification

4.3. Performance analysis of the proposed SVR and LSTM scheme

The statistical summary of the Dataset like the count of the mean of the values, standard deviations of the observations, nun null observations, minimum value, maximum value, and percentiles (upper 75 %, median 50 % and lower 25 %) are as revealed in Table 3.

5. Result analysis

The predictions are generated using the LSTM model for both the training and testing data to assess the performance of the model. In Fig. 4, the original Dataset is shown in red colour for the $PM_{2.5}$ versus AQI index value, the prediction of the AQI value is shown in green colour for the $PM_{2.5}$. It is seen that the model performs as a good job in fitting both training and test data.

The seasonal factors dependent upon the time was depicted in Fig. 4 (a). The continuous monitoring of the seasonal factors was compulsory for effective prediction of the quality of the air.

The RMSE & R^2 values are obtained and described in the following Table 4. Training algorithms are maintained in the suggested classifier and a few steps are required for training and prediction computation. The first step is data entry of inputs stage, and the second step is creating classifier system with membership function. Training of given input data with selection of learning algorithm is the third step, and after learning, performing iterative calculations for test data and train score.

The LSTM model for Regression is shown in Fig. 3. The result predicts that the model has an average error of 7.804 μg/m^3 for the training set and 10.995 μg/m^3 for the test set. The value of R^2 obtained with training and testing set is 0.632 and 0.570, respectively, which indicates that the model is the best fit for PM2.5 prediction.

The performance analysis of the $PM_{2.5}$ for observed data is shown in red and predicted AQI value for $PM_{2.5}$ is shown in green and plotted in Fig. 4(b). The LSTM method predicts the AQI value for $PM_{2.5}$ by applying the proper training of the model and predicts the correct

Table 4
RMSE & R^2.

category	Train Score	Test Score
RMSE	7.804	10.995
R^2	0.632	0.570

Table 3
Statistical summary of the exploratory variables in the Dataset.

	PM2.5 (Std60)	NO2 (Std80)	SO2 (Std80)	CO (std4)	Ozone(Std180)	AQI
count	21707.000000	21707.000000	21707.000000	21707.000000	21707.000000	21707.000000
mean	30.321609	12.088319	6.528234	0.703744	37.195119	22.412852
std	38.133978	9.771051	5.782149	0.403747	36.083979	14.346564
min	0.010000	0.010000	0.010000	0.000000	0.010000	2.871111
25 %	12.710000	5.670000	3.300000	0.450000	11.050000	15.11153
50 %	22.700000	10.310000	5.950000	0.660000	26.320000	20.075278
75 %	36.580000	16.170000	8.370000	0.860000	48.960000	26.445000
max	999.990000	312.930000	179.350000	10.000000	199.550000	359.836667

Features	Pollutants and particulates				
	CO	NO_2	SO_2	Ozone	$PM_{2.5}$
CO	*				
NO_2		*			
SO_2			*		
Ozone				*	
$PM_{2.5}$					*
Wind speed	*	*	*	*	*
Relative humidity	*	*	*	*	*
Outdoor temperature	*	*	*	*	*
CO roll mean	*				
CO lag features	*	*	*	*	*
SO_2_roll_mean			*		
SO_2 lag features	*	*	*	*	*
NO_2_roll_mean		*			
NO_2 lag features	*	*	*	*	*
Ozone lag features	*	*	*	*	*
Ozone roll mean				*	
$PM_{2.5}$ lag features	*	*	*	*	*
$PM_{2.5}$ roll mean					*

Fig. 3. Output of the feature extraction.

values. The Mean squared Error and R^2 value are measured for $PM_{2.5}$ are 0.179 and 0.821, respectively.

The proposed LSTM model, the training data and testing data to assess the performance of AQI for NO_2. In Fig. 5, the original Dataset is shown in red for the NO_2 versus AQI index value, the prediction of the AQI value is shown in green for the NO_2. It is seen that the model performs a good job in fitting both training and test data (Fig. 6). The MSE value and R^2 values are obtained for NO_2 as 0.908 and 0.092, respectively. The AQI values for the observed data and predicted data for the NO_2 is depicted in Fig. 7.

LSTM model is used for the training data and testing data to assess the performance of AQI for SO_2. In Fig. 8, the original Dataset is shown in red for the SO_2 versus AQI index value, the prediction of the AQI value is shown in green for the SO_2. It is seen that the model performs a good job in fitting both training and test data. The MSE value and R^2 values are obtained for SO_2 as 1.005 and -0.005, respectively. The AQI values for the observed data and predicted data for the SO_2 is depicted in Fig. 9.

The proposed model is used for the training data and testing data to assess the performance of AQI for CO. In Fig. 10, the original Dataset is shown in red for the CO versus AQI index value, the prediction of the AQI value is shown in green for the CO. It is seen that the model performs a good job in fitting both training and test data. The MSE value is 0.920 and R^2 value is 0.080. The AQI values for the observed data and predicted data for the CO is depicted in Fig. 11.

The model is used for the training data and testing data to assess the performance of AQI for Ozone. In Fig. 12, the original Dataset is shown in red for the Ozone versus AQI index value, the prediction of the AQI value is shown in green for the Ozone. It is seen that the model performs a good job in fitting both training and test data. The MSE value is 0.971 and R^2 value is 0.029. The AQI values for the observed data and predicted data for the Ozone is depicted in Fig. 13. Fig. 13 indicates that the the actual value is the value that is obtained by observation or by measuring the available data. It is also called the observed value. The predicted value is the value of the variable predicted based on the analysis. If the difference is zero, then that data points lie on the regression line. Here, there is a high range of difference.

The summary of Mean Squared Error and R^2 value for the AQI category of $PM_{2.5}$, NO_2, SO_2, CO and Ozone is depicted in Table 5.

The LSTM model is based mainly on the use of dependence between consecutive events on a relevant time. The LSTM neural network model is a part of deep recurrent neural network (RNN). In the RNN mechanism, the hidden units share information based on a time index. The sharing process helps in building memory blocks of long time series which help the model to recognize and predict the sequences. The feedback loop provides the units with memory, in which the previous states of the current neuron can be used as input parameters when updating memory. LSTM model contains three main units which are defined as input, output and forget gates. The units construct the memory blocks to provide an ability to update and filter information fow in different blocks. The proposed LSTM for AQI is depicted in Fig. 14. The graph is plotted against the Dataset versus AQI. This graph provided detailed information about the training plot, testing plot and predicted data plot is drawn in Fig. 14.

From the comparison it will reveal that the proposed method outperforms well when compared to other existing methodologies (Table 6).

6. Conclusions

Predicting the air quality is a complex task due to the dynamic nature, volatility, and high variability in space and time of pollutants and particulates. At the same time, being able to model, predict, and monitor air quality it is becoming more and more important, especially in urban areas, due to the observed critical impacts of air pollution for populations and the environment. In this approach, AQI has been estimated and predicted to control air pollution in the metropolitan city. At first, the collected Dataset has been pre-processed to replace the missing values and to remove redundant data. Then the feature extraction is carried out with the use of GLCM technique to extract the features like mean, standard deviation and so on. The extracted features are then optimised to get an optimised selection of extracted features. In this work various climatic conditions were considered and AQI is predicted. Then the classification mechanism is carried with the use of deep learning mechanism of SVR with LSTM model, which in turn predicts the AQI of the preferred location of the metropolitan city. The AQI gives the report to the common public regarding the ambient air quality level of the metropolitan city. The estimation of AQI is done in the metropolitan city and accurate value was obtained by using the deep learning approach. The predicted values are helpful to the city planning committee to suggest in installing road traffic signal coordination and promote to use public transport for their commutation. Also, electrical vehicle technology/non-motorised vehicles may be implemented 100 % where the pollution level is higher than ambient air quality. These predictions will be useful in developing a sustainable community in urban areas, particularly for developing countries. In future, the deep learning algorithm can be used to suggest incorporating air pollution controlling mechanism to improve the city's AQI values. Based on the predicted AQI values, it is possible to identify the vulnerable areas of pollutants.

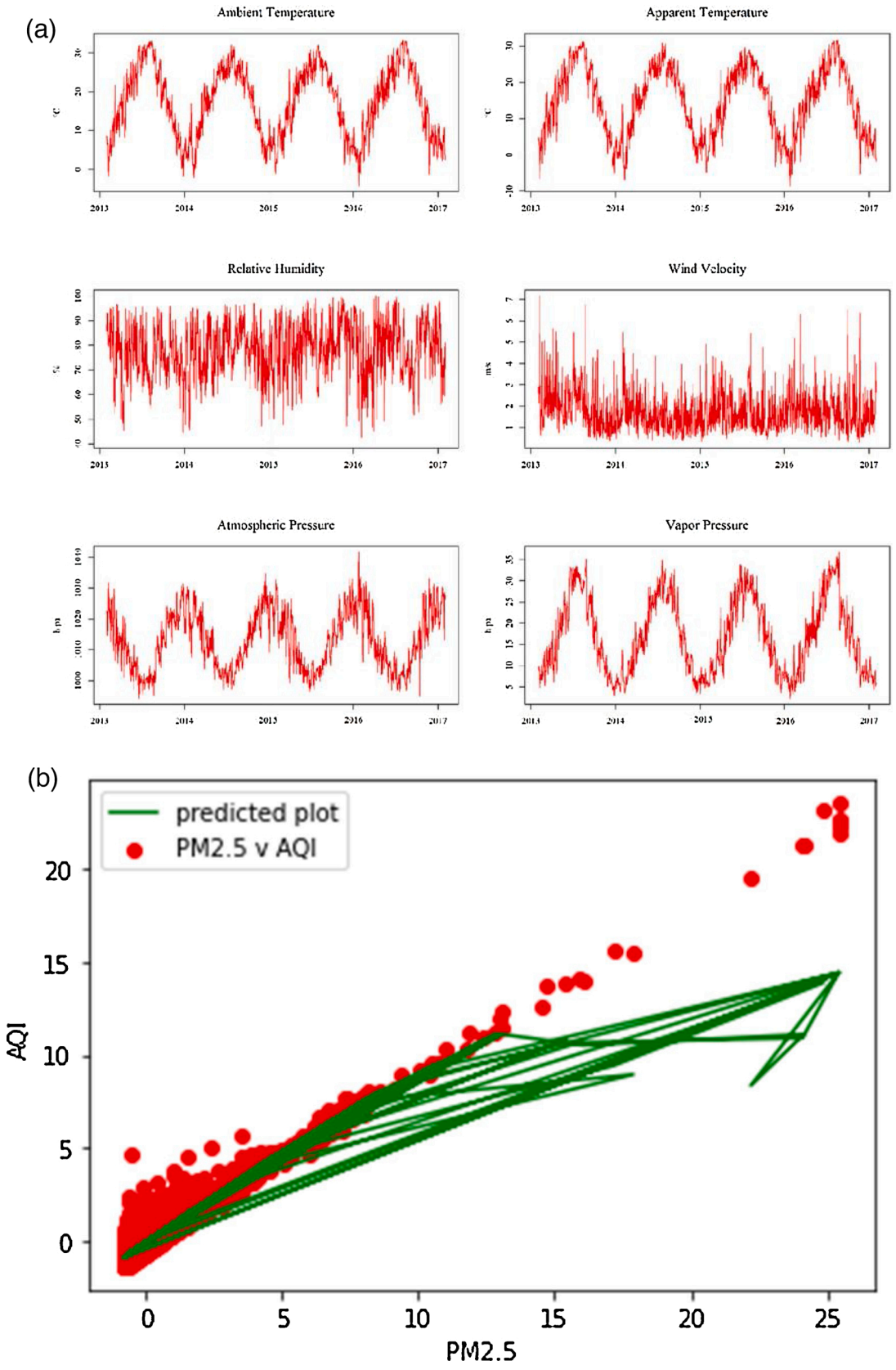

Fig. 4. a) Seasonal factors. b): Performance analysis of $PM_{2.5}$ Vs AQI.

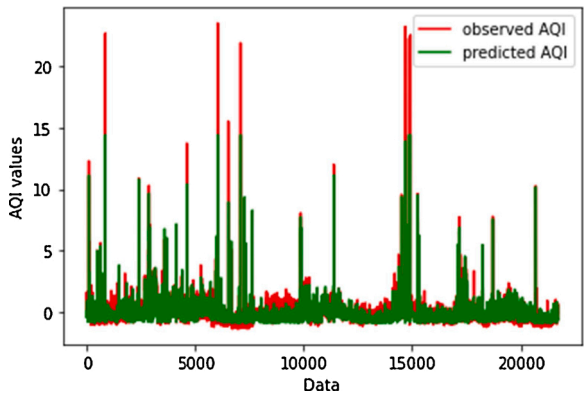

Fig. 5. Performance analyses of Observed data vs Predicted data for $PM_{2.5}$.

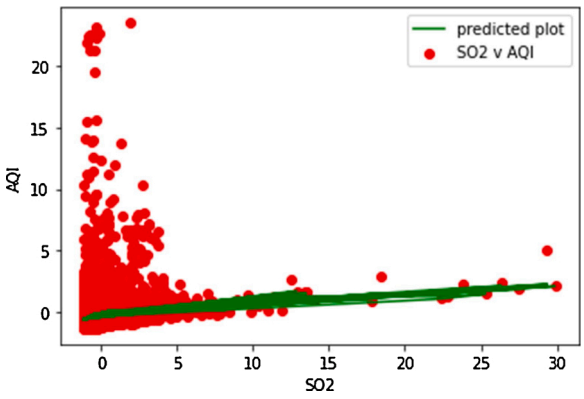

Fig. 8. Performance analysis of SO_2 Vs AQI.

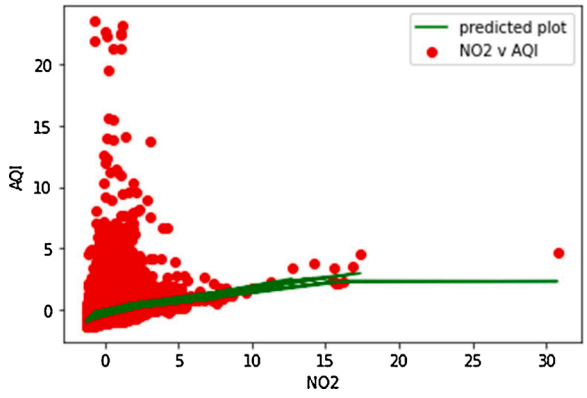

Fig. 6. Performance analysis of NO_2 vs AQI.

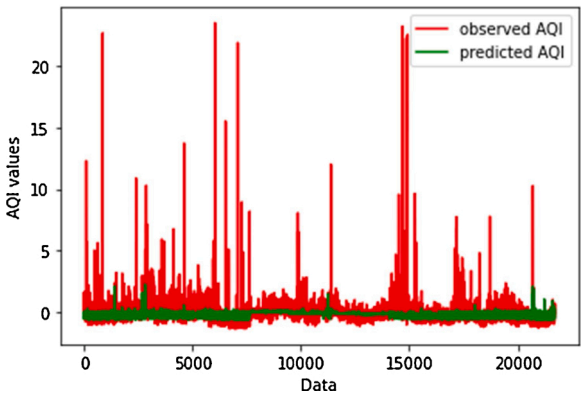

Fig. 9. Performance analyses of Observed data vs Predicted data for SO_2.

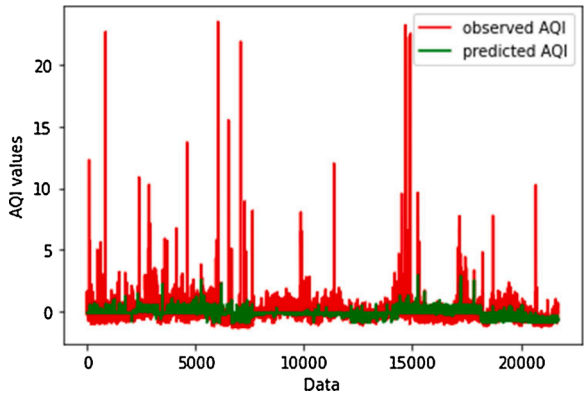

Fig. 7. Performance analyses of Observed data vs Predicted data for NO_2.

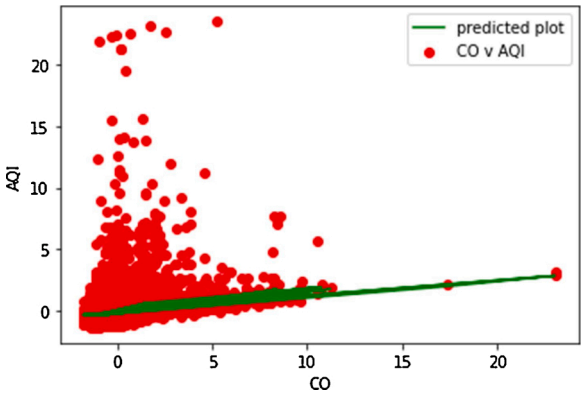

Fig. 10. Performance Analysis of CO vs AQI.

Declaration of Competing Interest

The authors report no declarations of interest.

References

Agarwal, P., & Melkania, U. (2018). Assessment of the ambient air quality at the industrial area using the air quality index method (AQI). *International Journal of Agriculture Environment and Biotechnology, 11*, 227–233. https://doi.org/10.1002/apj.450

Al_Janabi, S., Yaqoob, A., & Mohammad, M. (2019). Pragmatic method based on intelligent big data analytics to prediction air pollution. In *International conference on big data and networks technologies* (pp. 84–109). https://doi.org/10.1007/978-3-030-23672-4_8

Al-Janabi, S., & Alkaim, A. F. (2020). A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. *Soft Computing, 24*(1), 555–569. https://doi.org/10.1007/s00500-019-03972-x

Al-Janabi, S., & Mahdi, M. A. (2019). Evaluation prediction techniques to achievement an optimal biomedical analysis. *International Journal of Grid and Utility Computing, 10*(5), 512–527. https://doi.org/10.1504/IJGUC.2019.102021

Al-Janabi, S., Mohammad, M., & Al-Sultan, A. (2020a). A new method for prediction of air pollution based on intelligent computation. *Soft Computing, 24*, 661–680. https://doi.org/10.1007/s00500-019-04495-1

Al-Janabi, S., Mohammad, M., & Al-Sultan, A. (2020b). A new method for prediction of air pollution based on intelligent computation. *Soft Computing, 24*(1), 661–680. https://doi.org/10.1007/s00500-019-04495-1

Asghari, M., & Nematzadeh, H. (2016). Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network. *Journal of AI and Data Mining, 4*, 49–54. https://doi.org/10.5829/idosi.JAIDM.2016.04.01.06

Cabaneros, S. M. S., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software, 119*, 285–304. https://doi.org/10.1016/j.envsoft.2019.06.014

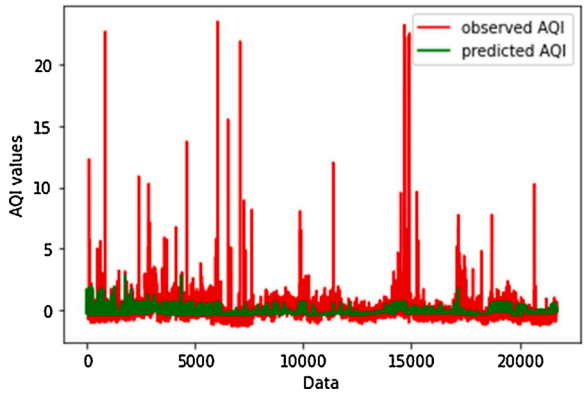

Fig. 11. Performance analyses of Observed data vs Predicted data for CO.

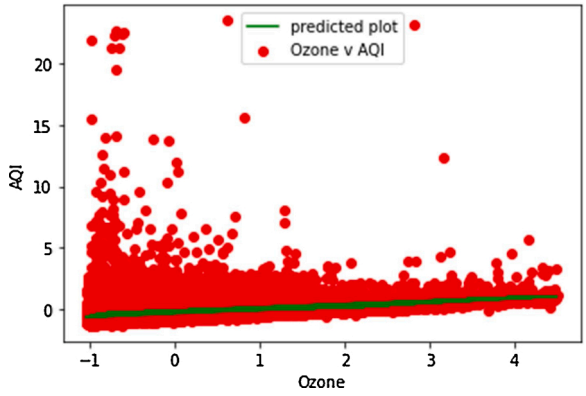

Fig. 12. Performance Analysis of Ozone vs AQI.

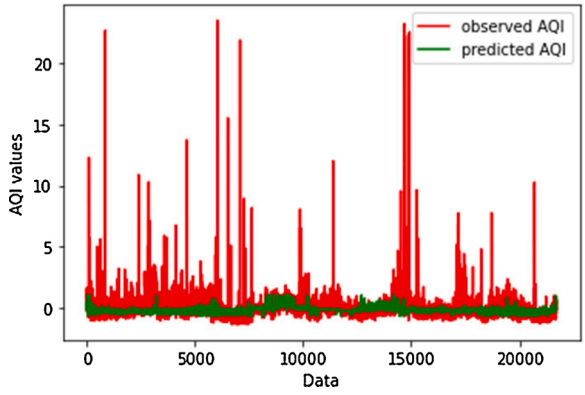

Fig. 13. Performance analyses of Observed data vs Predicted data for Ozone (O_3).

Table 5
Summary of Mean Squared Error and R^2.

AQI category	Mean Squared Error	R^2
$PM_{2.5}$	0.179	0.821
NO_2	0.908	0.092
SO_2	1.005	−0.005
CO	0.920	0.080
Ozone (O_3)	0.971	0.029

Dakheel, J. A., Del Pero, C., Aste, N.ò, & Leonforte, F. (2020). Smart buildings features and key performance indicators: A review. *Sustainable Cities and Society, 61*, 102328. https://doi.org/10.1016/j.scs.2020.102328

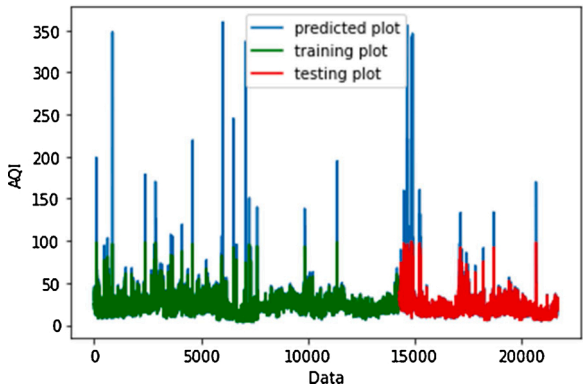

Fig. 14. Analysis of LSTM with Training data, Testing data and Predicted data.

Table 6
Comparative predictive analysis.

Model	RMSE	R
RNN (Jin et al., 2020)	64.05	0.66
LSTM (Jin et al., 2020)	65.4	0.64
GRU (Jin et al., 2020)	63.1	0.65
EMDCNN-RNN (Jin et al., 2020)	54.55	0.74
EMDCNN-LSTM (Jin et al., 2020)	51.17	0.77
EMD-CNN (Jin et al., 2020)	46.26	0.81
Proposed (LSTM)	10.9	0.97

Ghasemi, A., & Amanollahi, J. (2019). Integration of ANFIS model and forward selection method for air quality forecasting. *Air Quality, Atmosphere, & Health, 12*, 59–72. https://doi.org/10.1007/s11869-018-0630-0

Gia, T. N., Qingqing, L., Queralta, J. P., Zou, Z., Tenhunen, H., & Westerlund, T. (2019). *Edge AI in smart farming IoT: CNNs at the edge and fog computing with LoRa*. https://doi.org/10.1109/AFRICON46755.2019.9134049

Grace, R. K., & Manju, S. (2019). A comprehensive review of wireless sensor networks based air pollution monitoring systems. *Wireless Personal Communications, 108*, 2499–2515. https://doi.org/10.1007/s11277-019-06535-3

Honarvar, A. R., & Sami, A. (2019). Towards sustainable smart city by particulate matter prediction using urban big data, excluding expensive air pollution infrastructures. *Big Data Research, 17*, 56–65. https://doi.org/10.1016/j.bdr.2018.05.006

Jin, X.-B., Yang, N.-X., Wang, X.-Y., Bai, Y.-T., Su, T.-L., & Kong, J.-L. (2020). Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction. *Mathematics, 8*(2), 214. https://doi.org/2227-7390/8/2/214.

Kala, P., Joshi, P., Agrawal, S., Yadav, L. K., & Joshi, M. (2020). Introduction to condition monitoring of PV system. *Soft computing in condition monitoring and diagnostics of electrical and mechanical systems*, 169–187. https://doi.org/10.1007/978-981-15-1532-3_7

Kelly, J. T., Jang, C. J., Timin, B., Gantt, B., Reff, A., Zhu, Y., et al. (2019). A system for developing and projecting PM2.5 spatial fields to correspond to just meeting national ambient air quality standards. *Atmospheric Environment: X, 2*, 100019. https://doi.org/10.1016/j.aeaoa.2019.100019

Lei, M. T., Monjardino, J., Mendes, L., Gonçalves, D., & Ferreira, F. (2019). Macao air quality forecast using statistical methods. *Air Quality, Atmosphere, & Health, 12*, 1049–1057. https://doi.org/10.1007/s11869-019-00721-9

Li, R., Dong, Y., Zhu, Z., Li, C., & Yang, H. (2019). A dynamic evaluation framework for ambient air pollution monitoring. *Applied Mathematical Modelling, 65*, 52–71. https://doi.org/10.1016/j.apm.2018.07.052

Liu, B., Yan, S., Li, J., Qu, G., Li, Y., Lang, J., et al. (2019). A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. *IEEE Access, 7*, 43331–43345. https://doi.org/10.1109/ACCESS.2019.2908081

Ma, J., Ding, Y., Cheng, J. C. P., Jiang, F., Gan, V. J. L., & Zherui, X. (2020). A Lag-FLSTM deep learning network based on Bayesian optimisation for multi-sequential-variant PM2.5 prediction. *Sustainable Cities and Society, 60*, 102237. https://doi.org/10.1016/j.scs.2020.102237

Maciąg, P. S., Kasabov, N., Kryszkiewicz, M., & Bembenik, R. (2019). Air pollution prediction with clustering-based ensemble of evolving spiking neural networks and a case study for London area. *Environmental Modelling & Software, 118*, 262–280. https://doi.org/10.1016/j.envsoft.2019.04.012

Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y. T., & Rahmati, M. (2019). Air pollution prediction by using an artificial neural network model. *Clean Technologies and Environmental Policy, 21*, 1341–1352. https://doi.org/10.1007/s10098-019-01709-w

NandigalaVenkatAnurag, Y., & Sharanya, S. (2021). *Air quality index prediction with meteorological data using feature based weighted xgboost.* https://doi.org/A3492058119/19©BEIESP.

Singh, S. K., Jeong, Y.-S., & Park, J. H. (2020). A deep learning-based IoT-oriented infrastructure for secure smart City. *Sustainable Cities and Society, 60*, 102252. https://doi.org/10.1016/j.scs.2020.102252

Tang, I. J., McNabola, A., & Misstear, B. (2020). The potential impacts of different traffic management strategies on air pollution and public health for a more sustainable city: A modelling case study from Dublin. *Sustainable Cities and Society, 60*, 102229. https://doi.org/10.1016/j.scs.2020.102229

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., et al. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *The Science of the Total Environment, 654*, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086

Xiao, Y., Zhao, J., Liu, H., Wang, L., Yue, M., & Liua, J. (2020). Dynamic prediction of PM2.5 diffusion in urban residential areas in severely cold regions based on an improved urban canopy model. *Sustainable Cities and Society, 62*, 102352. https://doi.org/10.1016/j.scs.2020.102352

Xu, Z., Shan, J., Li, J., & Zhang, W. (2020). Extending the theory of planned behavior to predict public participation behavior in air pollution control: Beijing, China. *Journal of Environmental Planning and Management, 63*, 669–688. https://doi.org/10.1080/09640568.2019.1603821

Yadav, V., & Nath, S. (2020). Artificial neural networks based condition monitoring of air pollutants for Allahabad cities in India. *Soft computing in condition monitoring and diagnostics of electrical and mechanical systems*, 423–437. https://doi.org/10.1007/978-981-15-1532-3_19

Zhang, X., Wargocki, P., Lian, Z., & Thyregod, C. (2017). Effects of exposure to carbon dioxide and bioeffluents on perceived air quality, self-assessed acute health symptoms, and cognitive performance. *Indoor Air, 27*, 47–64. https://doi.org/10.1111/ina.12284