Computer Systems Science & Engineering
DOI: 10.32604/csse.2022.023882
Article

Tech Science Press

Multi-Site Air Pollutant Prediction Using Long Short Term Memory

Chitra Paulpandi*, Murukesh Chinnasamy and Shanker Nagalingam Rajendiran

Anna University, Chennai, 600066, Tamil Nadu, India
*Corresponding Author: Chitra Paulpandi. Email: chitrapaulpandi09@gmail.com

Received: 25 September 2021; Accepted: 08 November 2021

Abstract: The current pandemic highlights the significance and impact of air pollution on individuals. When it comes to climate sustainability, air pollution is a major challenge. Because of the distinctive nature, unpredictability, and great changeability in the reality of toxins and particulates, detecting air quality is a puzzling task. Simultaneously, the ability to predict or classify and monitor air quality is becoming increasingly important, particularly in urban areas, due to the well documented negative impact of air pollution on resident's health and the environment. To better comprehend the current condition of air quality, this research proposes predicting air pollution levels from real-time data. This study proposes the use of deep learning techniques to forecast air pollution levels. Layers, activation functions, and a number of epochs were used to create the suggested Long Short-Term Memory (LSTM) network based neural layer design. The use of proposed Deep Learning as a structure for high-accuracy air quality prediction is investigated in this research and obtained better accuracy of nearly 82% compared to earlier records. Determining the Air Quality Index (AQI) and danger levels would assist the government in finding appropriate ways to authorize approaches to reduce pollutants and keep inhabitants informed about the findings.

Keywords: LSTM; epochs; deep learning; air quality index; particulates; neural networks

1 Introduction

It is due of air that we are living today. Every month, we breathe roughly 1 million times without realizing the consequences of the air pollution we inhale. Over 93 percent of the world's population is exposed to dangerous air pollution chemicals such as Nitrogen Oxides (NOx), Carbon Oxides (COx), Sulphur Oxides (SOx), Particulate Matter (PM), Ozone (O_3), and Ammonia (NH_3) on a daily basis. Indoor air pollution is also much worse than outdoor pollution. Everyday products contain toxic compounds.

Noise, land, water, and air pollution are all major pollutants that influence humans and other living things. Among the several types of pollution, air pollution is the most serious. Natural disasters, automobiles, industries, crop fires, dust storms, man-made smokes such as burning of wood, plastics, natural gas, and coal, deforestation, population, and other factors all contribute to air pollution in India and is typically lower in summer than in the winter. Air pollution increases the risk of a variety of health problems, including arrhythmia, ischemia, heart failure, and stroke and so understanding and monitoring air pollution is critical for our well-being. The government employs the Air Quality Index (AQI) concept to forecast air pollutant levels and inform citizens.

AQI is a tool that displays the current state of air quality in six categories based on ambient concentration levels of air contaminants. Good, satisfactory, moderate, poor, very poor, and severe are the six classifications. An increase in the AQI level implies that there is a chance of breathing polluted air, which can have serious health consequences. The AQI is calculated using eight primary pollutants: Particulate Matter less than 2.5 microns ($PM_{2.5}$), Particulate Matter less than 10 microns (PM_{10}), Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), Carbon Monoxide (CO), O_3, NH_3, and Lead (Pb). "When we have high moisture then the aerosols in the air starts to absorb water vapors and swell thereby leads to low visibility and that is how the smog are created", said by Sachin Ghude, Scientist, Indian Institute of Tropical Meteorology (IITM), which operates System of Air Quality Weather Forecasting and Research (SAFAR), so it is very important to forecast air pollutants for better life.

Many air pollutant studies involve knowledge of environmental and computer technology, which is time consuming, and many statistical methods such as multiple linear regression [1], auto regressive moving average method and generalized line regression [2] are used for air quality predictions [3]. When compared to traditional methods such as support vector machine [4] and random forest, a commonly used air pollution prediction method in environmental or atmospheric research performed better [5–7]. In making atmospheric decisions, accurate forecasting in air quality measurement is critical [8]. Air pollutants are also highly dependent on regional and seasonal fluctuations, making it difficult to anticipate Air Quality (AQ) and necessitating simultaneous monitoring of time and space.

Currently, the rising technology Artificial Intelligence (AI) is being employed in air pollution prediction, with advanced artificial intelligence approaches achieving improved results. Also AI founds to be the future promising technology that serves faster with more accuracy in short span of time without human intervention. Advanced AI creates great impact in several applications and improves people's lives by performing most typical tasks. Deep Recurrent Neural Network (DRNN) is utilized in predicting fine PM2.5 [9]. Hybrid model spatiotemporal forecasting of PM2.5 is employed by long term prediction [10] and air pollutant concentration is predicted by combining other traditional methods [11,12]. Extraction of spatiotemporal characteristics improves the air pollution prediction model [13–16]. Aggregated Long Short Term Memory (LSTM) is also employed for air quality prediction [17]. Some methods provide average air pollutant concentration and to overcome the issue LSTM with Recurrent Neural Network (RNN) and Wireless Sensor Network (WSN) is employed [18]. Bayesian model [19] and bi-directional LSTM model [20] also helps to predict air quality and found to be better compared to traditional methods.

To forecast air pollution concentrations, this research proposes a deep learning model based on LSTM. Meteorological observations are obtained from a multi-site network of monitoring stations, and missing values are rebuilt and forecast values fine-tuned to make considerable improvements. The proposed model's accuracy was improved in an experimental situation by using a real-time air pollution dataset. In addition, the suggested Deep Learning (DL) model provides accurate assessment of AQI when compared to existing methodologies, and a greater number of features were compared for air quality forecasts and accuracy in the proposed DL method, so the public is warned.

2 Methodology

The suggested method begins with the selection of a data gathering region from local and near stations, collection of data from National Air Quality Index (NAQI), Central Pollution Control Board (CPCB), Tamil Nadu Pollution Control Board (TNPCB) and KAGGLE followed by pre-processing of data such as data division, manipulating missing data and normalization. The pre-processed data is classified using LSTM to anticipate air pollution with pinpoint accuracy. The methodology's flow is depicted in Fig. 1:

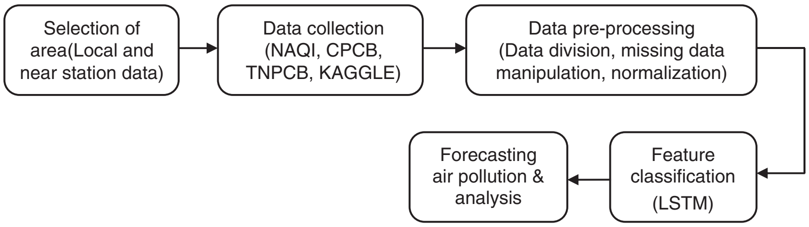

Selection of area(Local and near station data)
Data collection (NAQI, CPCB, TNPCB, KAGGLE)
Data pre-processing (Data division, missing data manipulation, normalization)
Feature classification (LSTM)
Forecasting air pollution & analysis

Figure 1: Process of methodology

2.1 Selection of Area

The research's study area was gathered from Air Quality Monitoring Stations (AQMS) in the states listed in Tab. 1. For data gathering, active stations were segregated. These study sites were chosen based on the availability of CPCB air quality data, satellite images, and the fact that the areas chosen were the most polluted, trafficked, and prone to industrial development activities. Some states have inactive AQMS and so they are identified first before selecting the sites. States having active air monitoring stations details are isolated. For the planned job, the network was trained using AQMS from various states across the country. The proposed paper focus on overall air pollution prediction of the country which can further be narrowed to particular state or city area. In our study nearly 21 states are selected for experiment.

Table 1: List of air quality monitoring stations taken for study

S. NO.	Station ID	State	TOTAL number of AQMS	Active AQMS
1	AP001	Andhra	05	01
2	AS001	Assam	01	01
3	BR001	Bihar	10	06
4	CH001	Chandigarh	01	01
5	DL001	Delhi	38	37
6	GJ001	Gujarat	06	01
7	HR001	Haryana	29	29
8	JH001	Jharkhand	01	01
9	KA001	Karnataka	20	10
10	KL001	Kerala	08	02
11	MP001	Madhya Pradesh	16	01
12	MH001	Maharashtra	22	10
13	ML001	Meghalaya	01	01
14	MZ001	Mizoram	01	01
15	OD001	Odisha	02	02
16	PB001	Punjab	08	01
17	RJ001	Rajasthan	10	03
18	TN001	Tamil Nadu	05	05
19	TG001	Telangana	06	06
20	UP001	Uttar Pradesh	26	04
21	WB001	West Bengal	14	07

2.2 Data Collection

The features to be collected from the specific site are processed once the study area has been established. It is critical to comprehend the data in order to recognize the features. As a result, self-reviewing data is required, and it is assessed for all of the chosen states or cities. Fig. 2 shows a flow diagram of the data selection process.

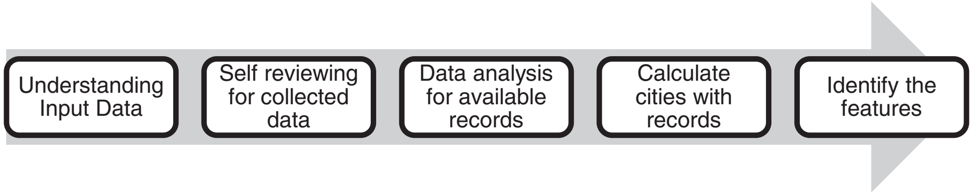

Understanding Input Data
Self reviewing for collected data
Data analysis for available records
Calculate cities with records
Identify the features

Figure 2: Process of data selection

For the available number of daily Air Quality Index data per city, about 37000 records for each station are taken on an hourly basis for the specified study areas from 2016 to 2020. The data was collected for three seasons: summer, rainy season, and winter. Before preprocessing, data collected from the KAGGLE website is rigorously scrutinized. Tab. 2 lists the features that have been identified for the proposed work. It is vital to comprehend the government-mandated averaging monitoring hours and minimal ambient concentration of air pollution levels.

Table 2: List of features taken for study

S. No.	Name of the air pollutant (features)	Symbol	Unit	Ambient concentration level of air pollutant		Monitoring time
				Industrial, residential, rural & other area	Sensitive area	
1	Particulate matter less than 2.5	$PM_{2.5}$	μg/m^3	60	60	24 h
2	Particulate matter less than 10	PM_{10}	μg/m^3	100	100	24 h
3	Nitrogen oxide	NO	μg/m^3	80	80	24 h
4	Nitrogen dioxide	NO_2	μg/m^3	80	80	24 h
5	Nitrogen oxides	NO_x	μg/m^3	80	80	24 h
6	Ammonia	NH_3	μg/m^3	400	400	24 h
7	Carbon monoxide	CO	mg/m^3	4 2	4 2	01 h 08 h
8	Sulphur dioxide	SO_2	μg/m^3	80	80	24 h
9	Ozone	O_3	μg/m^3	180 100	180 100	01 h 08 h
10	Benzene	C_6H_6	ng/m^3	5	5	08 h
11	Toluene	C_7H_8	ng/m^3	5	5	08 h
12	Xylene	C_8H_{10}	ng/m^3	5	5	08 h

Note: *AQI is measured by following units, 1. micrograms per cubic meter (μg/m^3), 2. parts per million (ppm) or parts per billion (ppb), 3. microns or micrometer.

2.3 Data Preprocessing

Once the necessary data has been gathered, it is standardized to eliminate the effects of missing numbers. Fig. 3 shows the stages involved in normalizing. Missing data is critical in preprocessing and has a significant influence on its own, thus diagnosing missing values with adequate data is critical. For these reasons, unknown values other than numbers are deleted from input data before transformation for complex numbers with a special number called Not a Number (NaN).

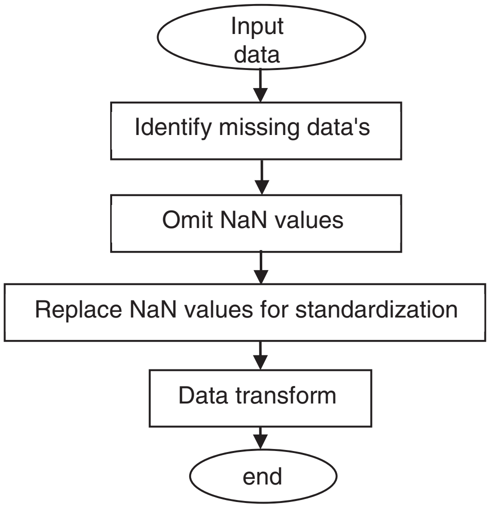


Figure 3: Input data preprocessing

2.4 Feature Classification

For training and testing purposes, we divided the input data into two portions. Nearly 70% of the 37000 records gathered are used for training, and 30% are used for testing. Ground truth parameters are collected during training, and the network is trained using the Stochastic Gradient Descent with Momentum (SGDM) optimizer. In comparison to other current algorithms, this best approach finds the model parameters that best fit the expected and actual outputs, calculates faster, and converges better with longer training time. Before training LSTM categorization, soft max is employed for activation layer during input data testing.

2.5 Forecasting Air Pollution and Analysis

Finally, the survey data is analyzed using methods from the Statistical Package for Social Sciences (SPSS). This SPSS software suite was used to conduct a detailed analysis of the data collected. The measurements done often includes mean, median, Standard Deviation (SD), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) helps to predict the performance level of classifier which enables to conclude AQI.

3 AQI Prediction Model Based on LSTM

Internal memory is used by the basic RNN to process the future variable sequence of inputs. Fig. 4 depicts the basic architecture of a basic RNN. Because the original RNN in our proposed model for training the dataset may not perform well for long-term reliance because it includes simple tanh in every repeating module, we employ LSTM, which is an expanded version of RNN, to overcome this issue.

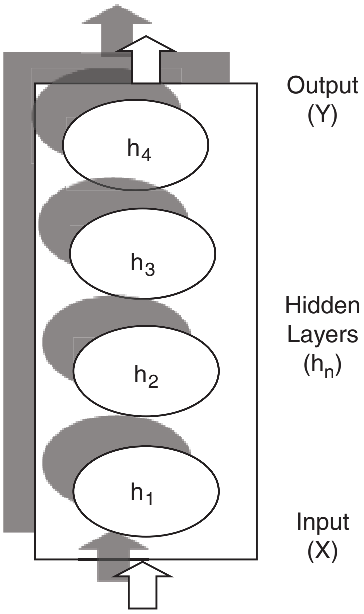


Figure 4: Basic RNN architecture

3.1 LSTM Networks

In comparison to simple RNN, the LSTM network is capable of performing long-term dependencies, which was first described by Hochreiter and Schmidhuber in 1997. It allows avoiding the long-term reliance problem. The core idea behind LSTM is as follows:

The key feature that goes horizontally through the diagram at the top is cell state. It's similar to a conveyor belt, but with a few more interactions. This cell state can be added or withdrawn based on the information and is regulated accordingly using a three-gate structure. As shown in Fig. 5 this regulation consists of a (σ) sigmoid neural net layer and a (x) point-wise multiplication operation. The main purpose of this sigmoid layer is to output values that are either zero (to signal "allow nothing through") or one (to indicate "let everything through").

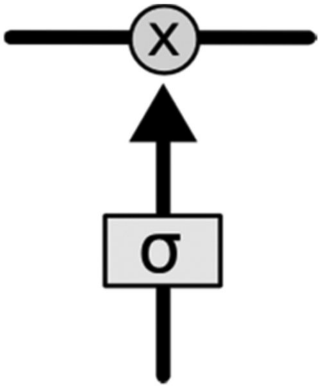


Figure 5: LSTM concept

3.2 LSTM Step by Step Process

Initially, the input data, as well as the input data concentration sequence before and after transformation, are defined. For sequence to label classification, layer array is created which includes sequence input layer, LSTM layer, fully connected layer, soft max layer and classification output layer. Sequence input layer represents total number of input features taken for study and the classes required for algorithm as decided is specified by fully connected layer. The basic block diagram of LSTM classification and regression is shown in Fig. 6:

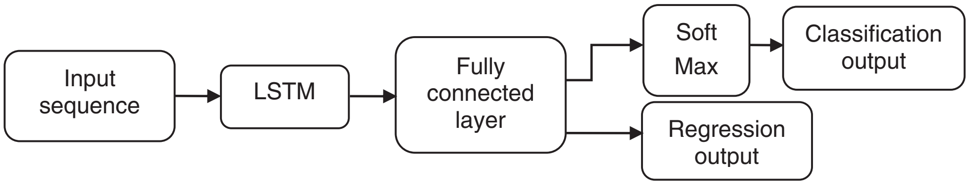


Figure 6: LSTM classification and regression

First the gender of the subject is analyzed for the given input x_t and the output value h_t. The sigmoid layer checks h_{t-1} and x_t and accordingly gives the output of number between 0 and 1 for each numbers in the cell state C_{t-1} as per Eq. (1). The new candidate value vector is created by tanh.

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

Next C_t is added to the new state followed by adding gender of the subject to the cell state as given in Eq. (2):

$$\tilde{C}_t = \tanh(W_c.[h_{t-1}, X_t] + b_c) \tag{2}$$

Now old cell state is updated C_{t-1} into cell state C_t. Later forgetting of previous information is performed by multiplying f_t with old state and adding it with C_t as shown in Eq. (3):

$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right) \tag{3}$$

Finally the output is decided from the cell state C_t.

The LSTM starts with the details of the input data that will be given to the network, and this choice is made by a sigmoid layer dubbed the “forget gate layer” (h_{t-1} and x_t), which produces a number between 00 and 11 for each cell state (C_{t-1}). The ‘input gate layer’ analyses the new information that needs to be stored in the cell state and determines which values need to be changed. The tanh layer follows the input gate, creating a vector of new added values C_{t-1}, which is then concatenated to provide an update to the cell state. We usually set the input values to tanh between −11 and 11 and multiply with the output sigmoid gate to only consider a certain section of the state [21].

4 Results and Discussion

4.1 Data Preprocessing

Data collected contains some unusual or missing data and so this impact may create side effects on the whole records and so data cleaning is very important before data preprocessing. There are numerous frequent methods to replace the missing values such as mean-median of previous or next value of current data R interpolation. Data acquisition frequently involves aberrant or missing data, which might have unforeseen repercussions for the full set of records. As a result, prior to data preparation, data cleaning is essential. Missing data is removed and relevant gaps are filled in using command tools. R-interpolation, mean-median of the previous or next value of the current data is all common ways for substituting missing values. The normalized input data is shown in Tabs. 3 and 4.

Table 3: Data normalization of input data (for first 10 records) for 6 input features

Records	$PM_{2.5}$	PM_{10}	NO	NO_2	NO_x	NH_3
1	0.3554	0.1725	−0.6568	−0.3862	−0.6529	−0.5783
2	0.2999	0.2324	−0.6650	−0.1468	−0.5655	−0.6045
3	0.4879	0.3147	−0.4204	0.0644	−0.3470	−0.4479
4	0.0454	−0.0959	−0.6055	−0.0566	−0.4973	−0.5367
5	0.1946	0.0455	−0.4832	−0.2687	−0.5105	−0.4872
6	0.1466	0.0457	−0.5079	−0.4006	−0.5752	−0.5646
7	0.2215	0.0286	−0.5129	−0.4389	−0.5932	−0.5646
8	0.5084	0.3789	−0.3696	−0.1385	−0.4070	−0.4354
9	0.4587	0.2516	−0.6783	−0.3510	−0.6516	−0.3615
10	0.4137	0.1790	−0.5386	−0.1041	−0.4793	−0.3586

Table 4: Data normalization of input data (for first 10 records) for next 6 input features

Records	CO	SO_2	O_3	C_6H_6	C_7H_8	C_8H_{10}
1	−0.5471	0.5741	4.5372	−0.2235	−0.1189	−0.4130
2	−0.5352	1.8856	4.0719	−0.2223	−0.0495	−0.4098
3	−0.5531	2.6276	3.8005	−0.2178	−0.0648	−0.4036
4	−0.5650	0.9949	5.0719	−0.2255	−0.1897	−0.4114
5	−0.5233	0.0492	3.7007	−0.2229	−0.2046	−0.4098
6	−0.5471	0.4432	4.1033	−0.2261	−0.2615	−0.4130
7	−0.5590	0.4241	4.3786	−0.2255	−0.2936	−0.4161
8	−0.5590	1.0363	4.6163	−0.2203	−0.2965	−0.4114
9	−0.5293	0.1455	3.9447	−0.2216	−0.2472	−0.4161
10	−0.5114	−0.0302	3.8294	−0.2165	−0.2611	−0.4083

4.2 Feature Validation

Statistical Validation of Extracted Features is done before classification, each piece of data that is used as an input must be evaluated for its importance. Tab. 5 shows the proposed characteristics and their accompanying metrics following validation. Number of samples (N), Standard Deviation (SD), Standard Error (SE), degree of freedom (df), Mean Square (MS), (measure of test accuracy) F1 Score, and significant are among the evaluation measures. The data was tested for the normality using **Shapiro Walik Test** and it was found that all the data was normally distributed and its significance of air pollutants is less than 0.05.

Table 5: Comparison of proposed features *vs*. metrics for evaluation

Features	N	Mean	SD	SE	df	MS	F1 Score	Sig
$PM_{2.5}$	36587	63.3840	58.0911	0.3037	36586	640.922	3.121E4	**0.0**
PM_{10}	29191	111.084	75.4643	0.4417	29190	1419.749	1.758E4	**0.0**
NO	37935	16.2983	22.2567	0.1143	37934	397.838	1.861E3	**0.0**
NO_2	37935	29.7610	23.2389	0.1193	37934	417.193	2.235E3	**0.0**
NO_X	36754	33.0347	31.8394	0.1661	36753	786.841	2.121E3	**0.0**
NH_3	27967	19.9223	16.5008	0.0987	27966	236.746	840.481	**0.0**
CO	37726	01.0531	01.6292	0.0084	37725	002.395	816.846	**0.0**
SO_2	37877	10.7421	09.9434	0.0510	37876	093.935	398.950	**0.0**
O_3	36879	33.3362	21.3029	0.1109	63878	411.616	757.173	**0.0**
C_6H_6	36059	03.9360	18.3308	0.0965	36058	335.656	008.793	**0.0**
C_7H_8	31554	09.3520	23.0653	0.1298	31553	527.082	060.004	**0.0**
C_8H_{10}	17765	02.9701	06.6244	0.0497	17764	042.887	083.475	**0.0**

One-way Analysis of Variant (ANOVA) was used to validate the input features. It can be used for further processing if the significant value is less than 0.05. Following validation, it was determined that all of the input features used in the study were significant, implying that all of the input characteristics used in the proposed study can be used for further classification using machine learning and deep learning algorithms.

4.3 Feature Classification

All of the significant features that have been validated using SPSS tools are used for classification. The corresponding sequence of air pollutant $PM_{2.5}$ for a given set of time T is defined as X, and these values are filled with record means to get $PM_{2.5}$ concentration sequence $\bar{X}$, and afterwards these data sequences are translated into supervised learning format. Because of its lengthy temporal dependency problem, the simple RNN cannot cope with large amounts of data. To overcome this, we use LSTM, which takes a lagging observation $t-1$ as an input variable and uses it to forecast the current time step T. The modified data set is represented by a $\bar{X}$, while the output variable is represented by a $\bar{Y}$. These sequences are then used to forecast individual $PM_{2.5}$ series, and the process is repeated for all of the other features in the proposed study. Finally, SPSS tools are used to compare the prediction outcomes, and the performance of the classifier is evaluated using various attributes such as root mean squared error, mean, median, standard deviation, and so on. Using these assessment markers, LSTM is found to be superior to other models in processing time series data, indicating that the current model is useful in AQI prediction. Fig. 7 depicts the steps involved in defining the LSTM algorithm prior to network training.

LSTM begins with initialization of sequence of input layers needed, fully connected layer, soft-max layer and classification layer. Then after training options are given which includes initial learning rate, (Ridge Regression) L2 regularization, drop periods, drop factors, epochs needed, batch size and SGDM. Once relevant initialization is completed then the input data's are converted to array format and later on input and ground truth are compared in activation layer. Finally the output is predicted based on the metrics such as accuracy, precision, error rate, sensitivity, specificity, F1score.

With proper initialization of training options the network is trained for classification. Defining LSTM layers includes input sequence (fully connected layer), LSTM 120 (soft-max) and LSTM 60 (classification layer). Initialization of learning rate (0.1), L2 regularization (0.0001), schedule (piecewise), drop factor (0.1), drop period (100), maximum epochs (500), mini batch size (128), and shuffling for every epoch plots are some of the learning rate of training options.

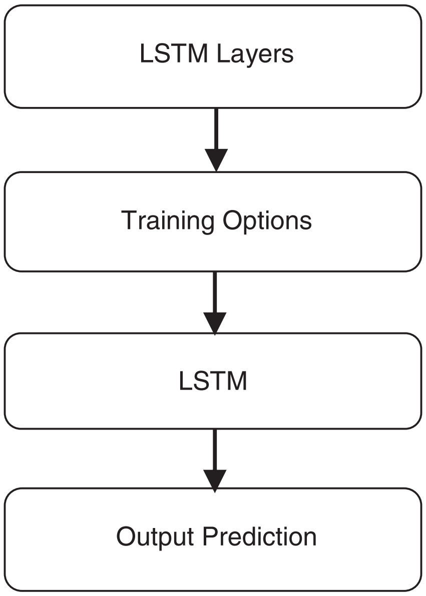


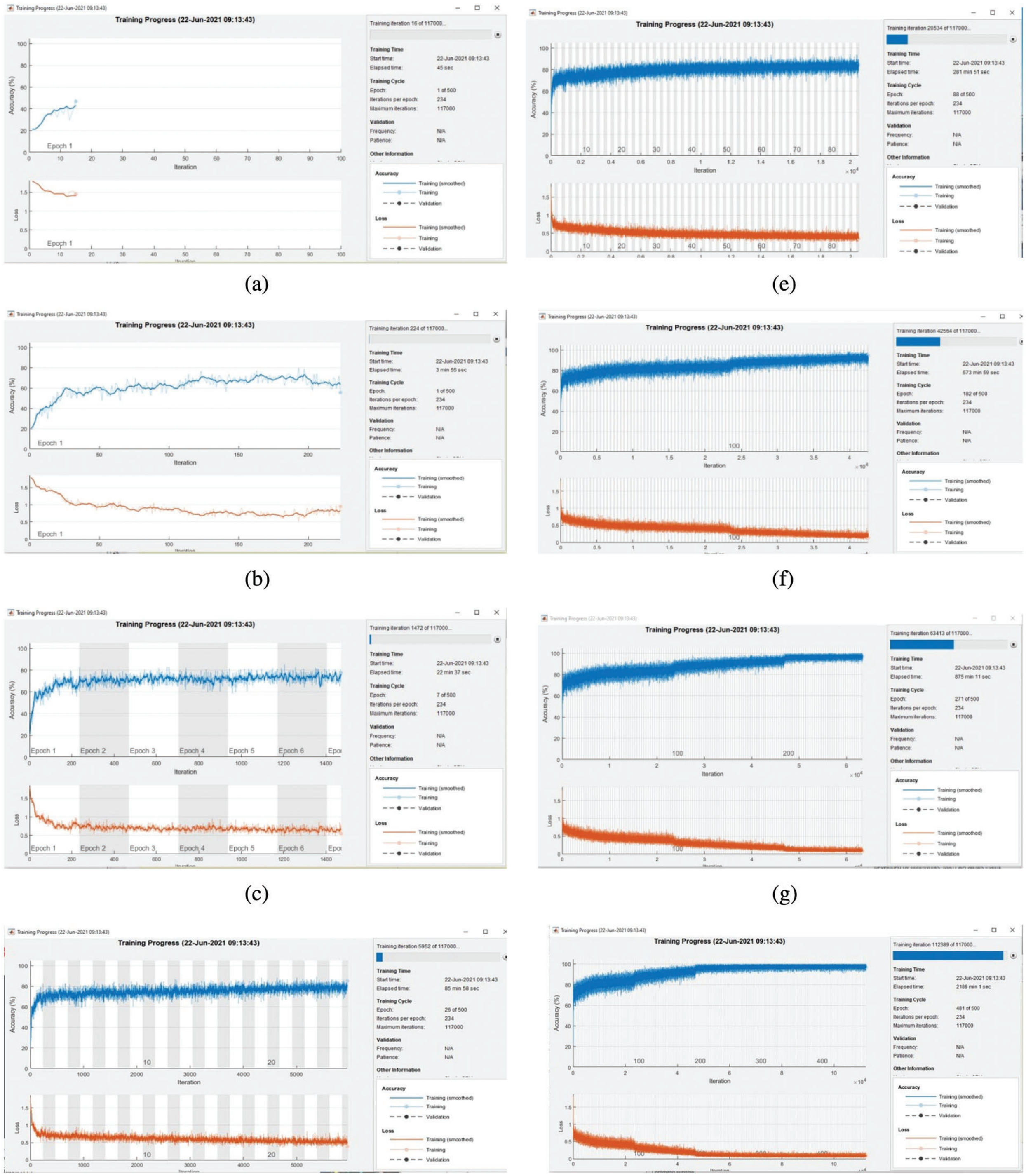
Figure 7: LSTM processing steps before network training

Each epoch is trained using 200 iterations, and it was discovered that the mini-batch loss and iterations are inversely proportional, with the batch loss reducing as the number of iterations grows. Tab. 6 shows the network's initial stage of training.

Table 6: Network training of epoch 1 to 3

Epoch	Iteration	Time elapsed (hh:mm:ss)	Mini-batch accuracy	Mini-batch loss	Base learning rate
1	1	00:00:04	21.09%	1.7951	0.1000
	50	00:00:44	57.03%	0.9743	0.1000
	100	00:01:14	61.72%	0.9100	0.1000
	150	00:01:45	65.63%	0.7632	0.1000
	200	00:02:15	65.63%	0.8067	0.1000
2	250	00:02:44	71.88%	0.7116	0.1000
	300	00:03:14	64.84%	0.7304	0.1000
	350	00:03:44	64.06%	0.8772	0.1000
	400	00:04:13	64.06%	0.8426	0.1000
	450	00:04:42	67.97%	0.6993	0.1000
3	500	00:05:17	67.97%	0.7930	0.1000
	550	00:05:58	77.34%	0.6215	0.1000
	600	00:06:29	68.75%	0.7011	0.1000
	650	00:07:01	74.22%	0.6566	0.1000
	655	00:07:04	74.22%	0.6626	0.1000

The accuracy and other characteristics are assessed for various epochs after the network has been trained for the above configurations constructed according to the suggested LSTM model. Fig. 8a through Fig. 8h illustrate the relevant network training plots.



Training iteration 16 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 45 sec
Training Cycle
Epoch: 1 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(a)

Training iteration 224 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 3 min 55 sec
Training Cycle
Epoch: 1 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(b)

Training iteration 1472 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 22 min 37 sec
Training Cycle
Epoch: 7 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(c)

Training iteration 5952 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 85 min 58 sec
Training Cycle
Epoch: 26 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(d)

Training iteration 20534 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 281 min 51 sec
Training Cycle
Epoch: 88 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(e)

Training iteration 42564 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 573 min 59 sec
Training Cycle
Epoch: 182 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(f)

Training iteration 63413 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 875 min 11 sec
Training Cycle
Epoch: 271 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(g)

Training iteration 112389 of 117000...
Training Time
Start time: 22-Jun-2021 09:13:43
Elapsed time: 2189 min 1 sec
Training Cycle
Epoch: 481 of 500
Iterations per epoch: 234
Maximum iterations: 117000
Validation
Frequency: N/A
Patience: N/A
Other Information

(h)

Figure 8: (a) Accuracy *vs*. iteration at 45 s (b) Accuracy *vs*. iteration at 3 min (c) Accuracy *vs*. iteration at 22 min (d) Accuracy *vs*. iteration at 85 min (e) Accuracy *vs*. iteration at 281 min (f) Accuracy *vs*. iteration at 573 min 59 (g) Accuracy *vs*. iteration at 875 min (h) Accuracy *vs*. iteration at 2900

Over or under fitting can create to classification issues in a training network, hence regularisation is crucial. In machine learning, regularisation is used to solve this problem, and in deep learning, dropout regularisation is used to prevent over-fitting and under-fitting by removing random neurons from hidden layers. For large data sets hold-out validation works good compared to cross-out validation.

In general, having too many epochs might lead to the model overfitting the training data. It signifies that the model memorises rather than learns the data. The accuracy of validation data is checked for each epoch or iteration to see if it over-fits or not. The number of epoch determines how the network's weights are changed. As the number of epochs grows, so do the number of times the neural network's weights are modified, and the border shifts from underfitting to optimal to overfitting.

For better performance, training data is shuffled for every epochs. As CPU is the available source mini batch size can be implemented that represents short sequences. Once all the desired configuration is inserted the network starts training. For every epoch and iterations the accuracy level and corresponding error rate is plotted. During run time the behaviour of network is analyzed by its accuracy level and error rate.

4.4 Algorithm Analysis

A 64-bit operating system AMD A4-5000 APU with Radeon (TM) HD graphics with 1.50 GHz and 8:00 GB RAM is utilized in conjunction with MATLAB 2019a for modeling, processing, comparisons and visualizing the experimental numbers and findings through various deep learning algorithms such as Support Vector Machine (SVM), Neural Network (NN), K-Nearest Neighbor (KNN), Naive Bayes (NB), Ensemble (EN) and LSTM.

4.4.1 LSTM Performance for Various Input Features

The performance of the LSTM classifier is examined using a variety of methods, one of which is shown in Tab. 7. The 12 input features of the planned study are compared to various computations in this section. When compared to other features, the error rate of PM_{10} was determined to be lower.

Table 7: Input features *vs*. computations

Input features	Accuracy	Error rate	Sensitivity	Specificity	Precision	F_1 score
$PM_{2.5}$	0.7930	0.2070	0.5283	0.9432	0.5703	0.5300
PM_{10}	0.8085	0.1915	0.6772	0.9480	0.6150	0.6370
NO_x	0.6575	0.3425	0.2430	0.8940	0.2150	0.2280
NO	0.5220	0.4780	0.2075	0.8646	0.7311	0.1852
NO_2	0.6400	0.3600	0.2495	0.8875	0.2909	0.2464
NH_3	0.5600	0.4400	0.1895	0.6854	0.1807	0.1684
CO	0.6285	0.3715	0.2381	0.8901	0.2078	0.2208
SO_2	0.6110	0.3890	0.2135	0.8702	0.2061	0.1986
O_3	0.6055	0.3945	0.2343	0.8870	0.2053	0.2143
C_6H_6	0.5485	0.4515	0.2056	0.8650	0.1801	0.1914
C_7H_8	0.5230	0.4770	0.1771	0.8445	0.1543	0.1574
C_8H_{10}	0.3660	0.6340	0.1427	0.7340	0.1285	0.1230

4.4.2 LSTM Performance for Various Computations

Fig. 9 shows how the accuracy level of each feature is assessed. For each of the 12 input features, various other metrics like as error rate, sensitivity, specificity, accuracy, and F1score were determined individually. It was found that accuracy is high for PM_{10} and low for Xylene.

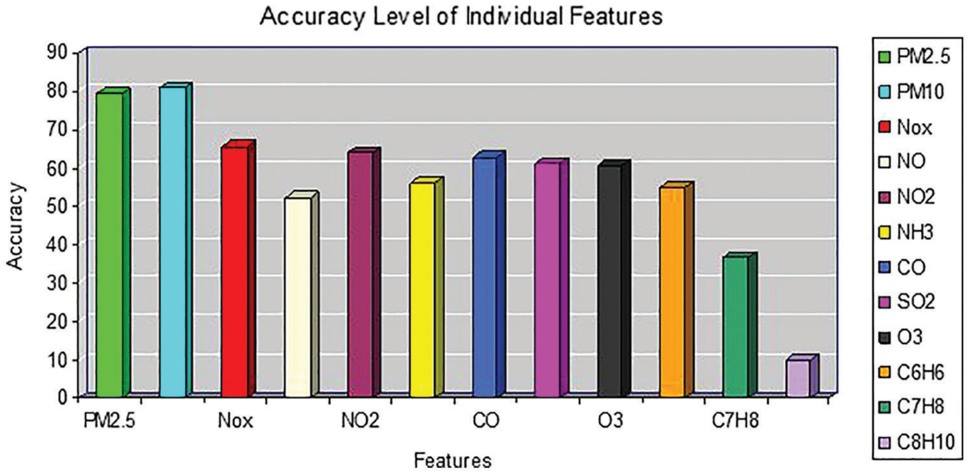


Figure 9: Accuracy level comparison of all input features

4.4.3 Algorithm Comparison with Proposed Work

Fig. 10 depicts the accuracy level of several approaches used, with the LSTM method proving to be the most accurate. Six different algorithms were taken for comparison for the same set of inputs.

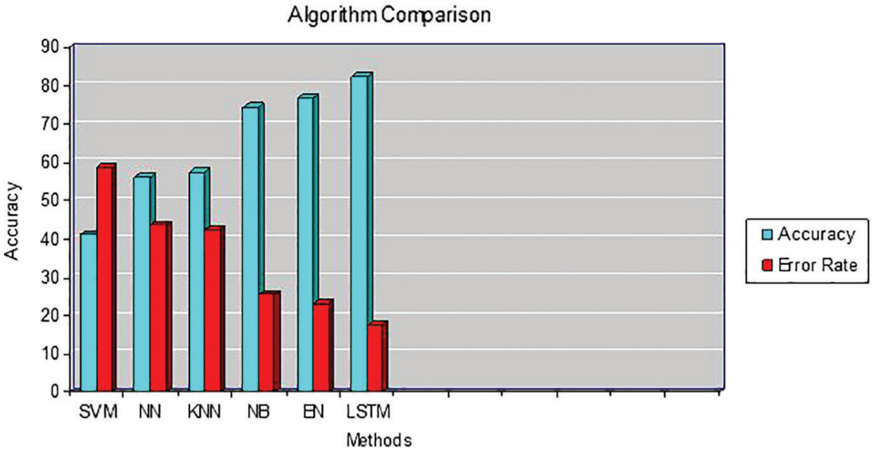


Figure 10: Accuracy level for various algorithms

As a result, various measures were examined using the LSTM approach, as shown in Tab. 8. The error rate found is minimum for the proposed LSTM method and subsequently accuracy is better compared to other methods.

Table 8: Algorithm comparison of accuracy, sensitivity, specificity, precision and F1 score

Algorithm	Accuracy	Error rate	Sensitivity	Specificity	Precision	F1 score
SVM (Cubic)	41.25%	0.5875	0.3997	0.8890	0.7013	0.3983
NN (Bilayered)	56.20%	0.4380	0.3941	0.8850	0.7330	0.4511
KNN (Weighted)	57.54%	0.4246	0.3976	0.8880	0.7576	0.4546
NB (Optimizable)	74.48%	0.2552	0.7103	0.9416	0.6097	0.6264
EN (Boosted)	76.80%	0.232	0.6800	0.9427	0.7901	0.7232
LSTM (Standard)	**82.40%**	**0.1760**	**0.6137**	**0.9512**	**0.5759**	**0.5843**

4.5 Limitations of Study and Future Work

The study's shortcoming was that the computation time was prolonged. The proposed study also has the disadvantage of not monitoring air pollutant concentrations in conjunction with other AQMS around or adjacent to it. Normally, both physical and chemical features of aerosols are used to predict air quality, however biological components and qualities are limited in this case. To increase the measurement level, future work can be expanded by including more air contaminants and additional data such as satellite images and industrial emissions into the atmosphere. To further understand the consequences of air pollution and human action, the article can be expanded by looking at specific states in relation to the current pandemic, as well as the situation before and after lockdown. Also, harmful air pollutants can be projected in advance for specific sites such as homes or roads, and the same can be combined with Internet of Things (IoT) and updated in real time in cloud computing for the benefit of people.

5 Conclusions

Based on historical air pollutant concentration, meteorological and time stamp data, this study provides an LSTM algorithm for predicting air pollutants in various sites. For predicting 12 major air contaminants, fine-grained air quality data is taken from active AQMS in 21 states across the country, India. Using the same dataset, six other models, including the proposed LSTM model, are evaluated, and the trials show that the suggested LSTM outperforms other techniques. By classifying air quality data and calculating dirty pixels using an LSTM classifier, the suggested work assists in obtaining specific information and permits precise knowledge of current pollutant levels in real environments of many sites. The classifier outputs the air pollutant level with higher compilation and efficiency than earlier approaches by comparing ground readings and data obtained from specific areas through private agencies, as well as suitable network training. The proposed approach delivers the best accuracy 82.4 percent of air pollution measurements for approximately 12 major air pollutants, according to the findings. This air quality measurement aids the environmental board in notifying the public and diverting traffic to low-polluting routes or areas, as well as taking appropriate measures such as tree planting, by anticipating air pollutants in advance.

Acknowledgement: The authors would like to thank Central Pollution Control Board of India (https://app.cpcbccr.com/AQI India/), for gathering information.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. P. K. Tai, L. J. Mickley and D. J. Jacob, "Correlations between fine Particulate Matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change," *Atmospheric Environment*, vol. 44, no. 32, pp. 3976–3984, 2010.

[2] H. Zhang, S. Zhang, P. Wang, Y. Qin and H. Wang, "Forecasting of particulate matter time series using wavelet analysis and wavelet-ARMA/ARIMA model in Taiyuan, China," *Journal of the Air & Waste Management Association*, vol. 67, no. 7, pp. 776–788, 2017.

[3] C. H. M. Tong, S. H. L. Yim, D. Rothenberg, C. Wang, C. Y. Lin *et al.,* "Assessing the impacts of seasonal and vertical atmospheric conditions on air quality over the Pearl River Delta region," *Atmospheric Environment*, vol. 180, pp. 69–78, 2018.

[4] W. Sun and J. Sun, "Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm," *Journal of Environmental Management*, vol. 188, pp. 144–152, 2017.

[5] G. Chen, S. Li, L. D. Knibbs, N. A. S. Hamm, W. Cao *et al.,* "A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information," *Science of the Total Environment*, vol. 636, pp. 52–60, 2018.

[6] X. Hu, J. H. Belle, X. Meng, A. Wildani, L. Waller *et al.,* "Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach," *Environmental Science & Technology*, vol. 51, no. 12, pp. 6936–6944, 2017.

[7] K. Huang, Q. Xiao, X. Meng, G. Geng, Y. Wang *et al.,* "Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain," *Environmental Pollution*, vol. 242, pp. 675–683, 2018.

[8] Y. Zhang and Z. Li, "Remote sensing of atmospheric fine Particulate Matter ($PM_{2.5}$) mass concentration near the ground from satellite observation," *Remote Sensing of Environment*, vol. 160, pp. 252–262, 2015.

[9] B. T. Ong, K. Sugiura and K. Zettsu, "Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$," *Neural Computing and Applications*, vol. 27, no. 6, pp. 1553–1566, 2016.

[10] Y. Qi, Q. Li, H. Karimian and D. Liu, "A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory," *Science of the Total Environment*, vol. 664, pp. 1–10, 2019.

[11] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu *et al.,* "Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation," *Environmental Pollution*, vol. 231, pp. 997–1004, 2017.

[12] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao *et al.,* "A novel combined prediction scheme based on CNN and LSTM for urban $PM_{2.5}$ concentration," *IEEE Access*, vol. 7, pp. 20050–20059, 2019.

[13] D. Seng, Q. Zhang, X. Zhang, G. Chen and X. Chen, "Spatiotemporal prediction of air quality based on LSTM neural network," *Alexandria Engineering Journal*, vol. 60, no. 2, pp. 2021–2032, 2021.

[14] C. Wen, S. Liu, X. Yao, L. Peng, X. Li *et al.,* "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction," *Science of the Total Environment*, vol. 654, pp. 1091–1099, 2019.

[15] J. Ma, Y. Ding, V. J. L. Gan, C. Lin and Z. Wan, "Spatiotemporal prediction of $PM_{2.5}$ concentrations at different time granularities using IDW-BLSTM," *IEEE Access*, vol. 7, pp. 107897–107907, 2019.

[16] P. W. Soh, J. W. Chang and J. W. Huang, "Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations," *IEEE Access*, vol. 6, pp. 38186–38199, 2018.

[17] Y. S. Chang, H. T. Chiao, S. Abimannan, Y. P. Huang, Y. T. Tsai *et al.,* "An LSTM-based aggregated model for air pollution forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451–1463, 2020.

[18] S. V. Belavadi, S. Rajagopal, R. Ranjani and R. Mohan, "Air quality forecasting using LSTM RNN and wireless sensor networks," *Procedia Computer Science*, vol. 170, pp. 241–248, 2020.

[19] Y. Han, J. C. Lam, V. O. Li and D. Reiner, "A Bayesian LSTM model to evaluate the effects of air pollution control regulations in Beijing, China," *Environmental Science and Policy*, vol. 115, no. 11, pp. 26–34, 2021.

[20] L. Zhang, P. Liu, L. Zhao, G. Wang, W. Zhang *et al.,* "Air quality predictions with a semi-supervised bidirectional LSTM neural network," *Atmospheric Pollution Research*, vol. 12, no. 1, pp. 328–339, 2021.

[21] C. Olah, "Understanding LSTM Networks," 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.