International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 8, Issue 5, September - October 2019 ISSN 2278-6856

Air Quality Index Prediction of Delhi using LSTM

[1]Mohit Bansal ,[2]Anirudh Aggarwal ,[3]Tanishq Verma ,[4]Apoorvi Sood

[1]Division of Information Technology, Netaji Subhas University of Technology, Azad Hind Fauj Marg, Sector - 3 Dwarka, New Delhi - 110078, Delhi, India

[2]Division of Information Technology, Netaji Subhas University of Technology, Azad Hind Fauj Marg, Sector - 3 Dwarka, New Delhi - 110078, Delhi, India

[3]Division of Information Technology, Netaji Subhas University of Technology, Azad Hind Fauj Marg, Sector - 3 Dwarka, New Delhi - 110078, Delhi, India

[4]Division of Information Technology, Netaji Subhas University of Technology, Azad Hind Fauj Marg, Sector - 3 Dwarka, New Delhi - 110078, Delhi, India

Abstract: ***Air pollution is one of the most severe problems of the current time. It is growing day by day because of the vast level of industrialization and urbanization, causing massive damage to flora and fauna of the planet. Every moment, we are breathing air that is full of pollutants, going to our lungs, impregnating our blood and then the whole body, causing uncountable health problems. Both state and central governments have put in many efforts to keep air pollution under control.***

The proposed paper discusses an efficient approach towards the prediction of air quality index (AQI) of Delhi, India. AQI is a measure of air quality. It is used to inform citizens about the associated health impacts of air pollution exposure. So, we modelled a deep recurrent neural network (RNN) based on Long-Short Term Memory (LSTM) to predict hourly based concentrations of pollutants. These concentrations are then used to calculate AQI. The proposed LSTM model achieved good results in estimating hourly based ambient air quality.

Keywords: Air Pollution, Air Quality Index, Deep Learning, Long Short Term Memory, Delhi.

1. INTRODUCTION

Air pollution is the introduction of particulates, biological molecules, or other harmful substances into the Earth's atmosphere. It is a global concern and a major environmental health problem. It is the fifth leading risk factor for mortality worldwide. People are dying more because of air pollution than malnutrition, road traffic injuries, and alcohol use [1]. According to the World Health Organization (WHO), 9 out of 10 people around the world breathe polluted air. Every year, around 7 million people die from exposure to air pollution [2]. One-third of deaths from heart disease, lung cancer, and stroke are due to air pollution [3]. The quality of life in a place is measured using several factors in which air quality plays a vital role. Its measurement is based on the concentration of pollutants in the atmosphere and is called AQI. AQI is a method that transforms the weighted values of individual air pollution-related parameters (for example, pollutant concentrations) into a single number or set of numbers. In the AQI system, specific concentration ranges are grouped into air quality descriptor categories [4].India is a developing country. With urbanization and industrialization, air pollution in India is also increasing. Many harmful gases are released to the atmosphere by industrialization processes.Automobile emissions, fires on agricultural land, construction sites dust, burning garbage are a significant contributor to air pollution in India.By particulate matter concentration, 22 of the 30 most polluted cities in the world are in India [5]. Delhi, India's capital territory, is ranked the world's most polluted capital and is at 11th position overall [6]. To understand and measure ambient air quality in India, the Ministry of Environment, Forest, and Climate Change developed and launched the AQI system on 17-October-2014 [7]. In Indian AQI System (IND-AQI), the following eight pollutants are considered for calculation of AQI: CO, NO_2, SO_2, $PM_{2.5}$, PM_{10}, O_3, NH_3, and Pb. To present the status of air quality and its effects on human health, the following six air quality description categories have been adopted: Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe [4]. Table 1 shows the concentration range and the health statements for AQI categories.

In all these years, conventional approaches are used for ambient air quality assessments. Manual analysis of raw data is carried out in these approaches. According to Niharika et al. [8], traditional approaches use statistical and mathematical techniques for air quality prediction. However, these methods are inefficient, complex, and provide limited accuracy. With recent advancement in technology and research, novel air quality assessment techniques have been modelled. Deep Learning is one such technique and has accomplished remarkable results in solving real-life problems of various domains. Handwriting recognition [9], speech recognition [10], [11], [12], natural language processing [13] are some of the areas where deep learning has produced outstanding results. Promising results in these areas motivated researchers to adopt this technique in various air quality studies.

This paper proposes a deep-learning based approach to predict ambient air quality in Delhi, India. The main contributions of this paper are: (1) LSTM model, proposed topredict air pollutants' concentration in Delhi; (2) AQI calculation, to forecast the ambient air quality of Delhi for the next hour; (3) Results compared with baseline approaches which showed that the proposed model achieved better results.

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 the 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. Effect may be more pronounced in people with lung and heart diseases
Severe (401-500)	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

Table 1: Health Statements for AQI Categories (This table is adopted from National Air Quality Index Report by Central Pollution Control Board [4]

The overall structure of this research paper is as follows. Section 2 shows some of the notable work done in forecasting air quality. Section 3 presents information and observations related to data and methods used in this study. Section 4 discusses the experiments and results of this study. Finally, conclusion and future scope of the work is given in section 5.

2. Related Work

Deep Learning is a class of machine learning algorithms that uses multiple layers to extract higher-level features from the raw input progressively [14]. RNN is a popular deep learning architecture that is used to model sequential data. It contains cyclic connections where the outputs from previous time steps are fed as input to the current time step.
In RNNs, errors are backpropagated, and weights are updated using a technique called Back Propagation Through Time (BPTT). However, while training an RNN, there occur problems of vanishing and exploding gradients. With many layers in the neural network model, the gradient output, the error, if greater than 1, leads to very large values of the gradients to be used for further calculation. This is called the exploding gradient problem, due to which the trainable weights have considerable changes in their values for each iteration, which increase the impact of the initial layers on the output. Whereas, when the gradients are less than 1, their effect on the gradients of the initial layers is minimal. So, to lessen the effect of these problems, LSTM [15] was introduced.
LSTM is a specific RNN architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. It contains special units called memory blocks in the recurrent hidden layers. The memory blocks contain memory cells with self-connections storing the temporal state of the network in addition to special multiplicative units called the gates to control the flow of information.
There have been several approaches to predict air quality. Kumar and Goyal forecasted daily AQI for Delhi using a combination of both ARIMA (Auto-Regressive Integrated Moving Average) and PCR (Principal Component Regression) statistical models [16]. Ni et al. compared multiple statistical models on $PM_{2.5}$ data around Beijing which exposed that linear regression models can perform better than other models in some cases [17]. Li et al. devised multiple linear regressions' technique for air quality estimation [18]. Bing-Chun Liu et al. came up with a model of collaborative forecasting of AQI of three cities in China using Support Vector Regression. They took air quality information and meteorological conditions of multiple cities as input [19]. Nieto et al. built a non-linear dynamic model, Support Vector Regressor to determine the factors affecting the air quality in Oviedo urban area (Norther Spain) [20]. Jain and Khare used an adaptive neuro-fuzzy model to predict hourly CO concentrations with prediction accuracy varying from 89 to 93% [21]. Athnasiadis et al. analyzed weather and air quality data using the σ-FLNMAP classifier to estimate ozone concentration levels and categorizing them into three classes, namely, high, medium, and low [22]. These conventional approaches are inefficient and require prior knowledge of data distribution. Moreover, these methods could not model long-range dependencies of data accurately.
Athira et al. used different deep learningbased architectures to predict air quality in China. They trained RNN, LSTM, and GRU based neural networks for predicting future PM_{10} concentration [23]. Xiang et al. proposed an LSTME (Long Short Term Memory Extended) model for predicting air pollutants concentration in Beijing city [24]. Reddy et al. developed a deep air system for forecasting air pollution in China [25].Krishan et a constructed LSTM model to predict the concentration of $PM_{2.5}$, NO_x, O_3, and CO at a particular location in Delhi. They evaluated the model performance for 2008-2010 data and found that the LSTM model is beneficial in ambient air quality forecasting [26]. Rao et al. contributed to forecasting air ambiance in Visakhapatnam with LSTM based RNN. They predicted the hourly concentration of ten pollutants considering six weather parameters with temporal sequence data of each of the pollutants [27].

3. Data and Methods

3.1 Data Acquisition

Delhi covers an area of 1484 km^2 out of which 783 km^2 is under the rural area, and 700 km^2 is under the urban area. It is bordered by Haryana state on three sides and by Uttar Pradesh to the east [28]. The current population of Delhi is around 19 million. According to the United Nations' World Urbanization Prospects, Delhi would become the most populous city in the worldby 2028 [29]. Although many steps have been taken to control air pollution, none of them have been much productive. Dwarka is chosen asresearch location for our study. It is a sub-city in South-West district of Delhi and a short distance away from Gurugram, which is the world's most polluted city [5]. Dwarka is one of the pollution hotspots of Delhi [30].

The data for the concentration of pollutants and meteorological parameters were collected from Central Pollution Control Board (CPCB) [31] where data is publicly available for 18 states that contain 102 cities with a total of 170 stations (locations). Data was collected for station NSUT (formerly NSIT), located in Sector-3, Dwarka [32]. The reason is that NSUT data were densely populated with only a few gaps (non-allocated values) counting to a few days to a few weeks in a year as compared to other stations. So, data for 3.5 years was selected for the purpose. The timeline of the data is from 1 April 2015 to 31 March 2017 and 1 October 2017 to 1 April 2019. Data from 1 April 2017 to 30 September 2017 was not available. The set of air pollutants and meteorological parameters used in this research are shown in tables 2 and 3.

S. No.	Parameters	Unit
1	CO (Carbon Monoxide)	mg/m^3
2	NO (Nitrogen Oxide)	$\mu g/m^3$
3	NO_2 (Nitrogen dioxide)	$\mu g/m^3$
4	Ozone	$\mu g/m^3$
5	$PM_{2.5}$(Particulate Matter 2.5mm)	$\mu g/m^3$
6	SO_2(Sulfur dioxide)	$\mu g/m^3$

Table 2: List of Pollutants considered

S. No.	Parameters	Unit
1	Temperature	˚C
2	RH (Relative Humidity)	%
3	SR (Solar Radiation)	W/mt^2
4	WS (Wind Speed)	m/s

Table 3: Meteorological Parameters

3.2 Preprocessing

The collected data contained several missing values and extreme values that deviate from other observations on data which may indicate variability in measurement, experimental errors or a novelty. This was handled by setting the outliers to null values.

The missing data present in the dataset acts as noise which affects the performance of the forecasting model. So, the missing data was populated using the technique called interpolation [33]. It is a method of constructing new data points within a range of a discrete set of known data points. It can be linear, bilinear, piecewise, polynomial, spline, cubic, bicubic, etc. The linear interpolation technique was used to fill the missing values. So, the final dataset considered has 35088 samples for each pollutant and the weather parameters. The descriptive statistics of the data are shown in Table 4:

	CO	NO	NO_2	Ozone	$PM_{2.5}$	SO_2	Temp	RH	SR	WS
mean	1.56	19.85	30.01	35.03	114.04	10.08	23.96	46.68	143.62	0.81
std	4.95	38.68	18.79	41.79	85.39	9.22	9.42	26.28	193.3	0.57
min	0	0	0	0	0	0	0	0	0	0
25%	0.34	5.20	17.94	8.05	55.24	4	17.68	24.73	22.73	0.39
50%	0.5	8.38	26.49	17.66	93.09	7.76	25.55	44.27	58.32	0.72
75%	0.76	14.08	38.67	48.64	145.17	12.76	30.95	67.90	198.88	1.08
max	50	392.47	320.42	938.57	694.60	109.06	49.92	100	1495.60	5.59

Table 4: Summary Statistics of Pollutants and Meteorological Parameters

The air pollutant data is represented graphically in figure 1. For the representational purpose, hourly data for year 2018 has been shown on a mat plot graph.

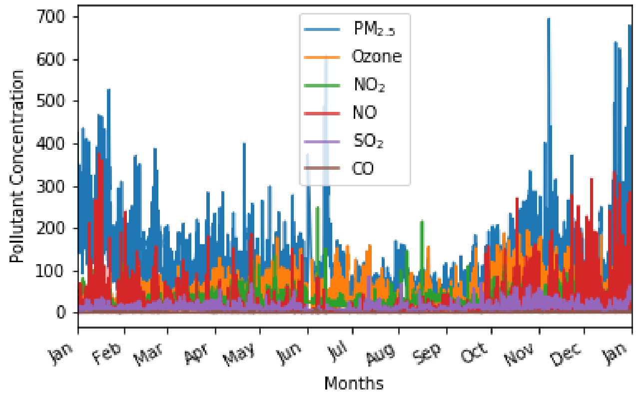


Figure1: Pollutant Concentration for year 2018

It is crucial to choose appropriate training, validation and testing data for evaluating the performance of a model. There is no standard way to split the data into training, validation and testing. So, 2 years of data has been chosen for training, 6 months' data for validation and 1 year of data for testing. The training part contains time-series data from 1 April 2015 to 31 March 2017, validation contains data from 1 October 2017 to 31 March 2018 and testing contains data from 1 April 2018 to 1 April 2019. The pollutants and metrological parameters at this point are collectively called features and each data entry denotes a time-series data-point.

It is discernible from table 4 that our data consists of features with different ranges, means, and standard deviations. Due to non-similar range of values in different features, the gradient may oscillate and end up taking a long time to converge to local/global minima. Thus, to overcome the model learning problem, data is normalized between 0 and 1 using Min-Max Normalization [34] to make sure that disparate features take on values in comparable range so that gradients converge more quickly.

3.3 LSTM Model

Figure3represents neural network architecture of proposed LSTM model. The model comprises of Input Layer, LSTM layers, Dense Layer and Output Layer.

Input Layer is used to create sequential data for LSTM layer. Each sequence, Seq_i consists of k feature vectors where k is the number of time-steps. A vector is denoted by X in the diagram, where $X_t$is a feature vector at time t. $X_t$contains concentration of pollutant p, at time t (denoted by C_t^p) and meteorological parameters, namely Temp, RH, SR and WS. These sequences are then fed to LSTM layer. Each LSTM layer consists of several memory blocks.The memory blocks contain memory cells with self-connections and special multiplicative units called the gates. A memory cell stores the temporal state of thenetwork and gates are used to control the flow of information. There are three types of gates in a LSTM cell – input gate, output gate and forget gate. The input gate controls the flow of input activations into the memory cell, while the output gate controls the output flow of cell activation into the rest of the network. The forget gate addresses a weakness of LSTM models preventing them from processing continuous input streams that are not segmented into subsequences. It scales the internal state of the cell before adding it as input to the cell through the self-recurrent connection of the cell, therefore adaptively forgetting or resetting the cell's memory. The structure of an LSTM cell is shown in figure2 [35].

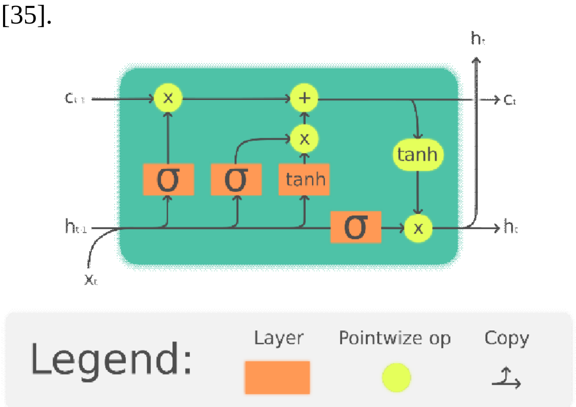


Figure2: LSTM Cell

The cell's functioning is represented mathematically, as:

$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$ (1)

$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$ (2)

$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$ (3)

$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$ (4)

$h_t = o_t \circ \sigma_h(c_t)$ (5)

[36]

where initial values are $c_0 = 0$ and $h_0 = 0$ and the operator $\circ$ denotes element-wise product. The subscript t indexes the time step.

$x_t \in \mathbb{R}^d$: input vector to LSTM unit

$f_t \in \mathbb{R}^h$: forget gate's activation vector

$i_t \in \mathbb{R}^h$: input gate's activation vector

$o_t \in \mathbb{R}^h$: output gate's activation vector

$h_t \in \mathbb{R}^h$: hidden state vector also known as output vector of LSTM unit

$c_t \in \mathbb{R}^h$: cell state vector

$W \in \mathbb{R}^{h\times d}$, $U \in \mathbb{R}^{h\times h}$ $and\ b \in \mathbb{R}^h$: weight matrices and bias vector parameters that are trainable where superscripts h and d refer to the number of input features and number of hidden units respectively.

σ_g: sigmoid function

σ_c: hyperbolic tangent function

σ_h: hyperbolic tangent function or identity function

Output from LSTM layers is passed through a fully connected hidden layer. The output layer generates the concentration of pollutant, p at $k+1$ time (denoted by C_{k+1}^p).

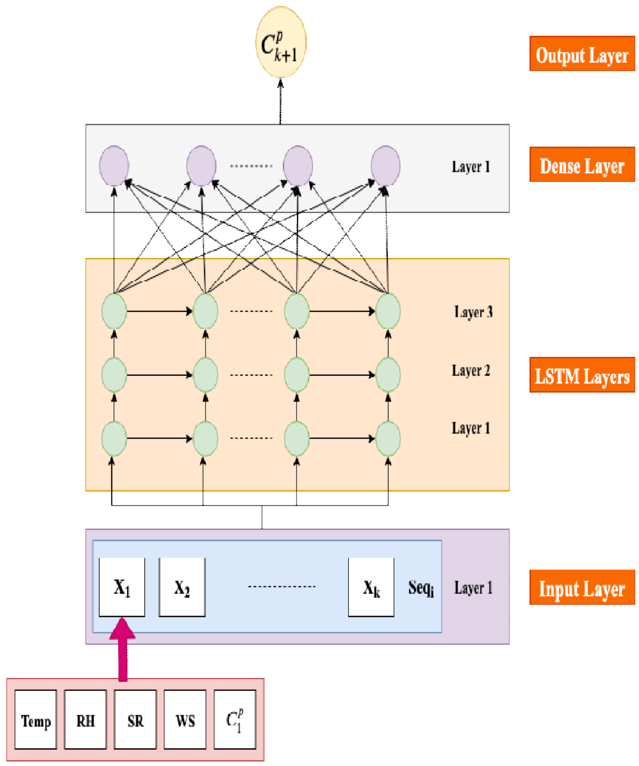


Figure3: Model Architecture Diagram

3.4 AQI Calculation

AQI system transforms the weighed values of individual pollutant concentrations into a single number or set of numbers. An AQI is formulated using two steps: (i) formation of sub-indices (for each pollutant) and (ii) aggregation of sub-indices to get an overall AQI. A sub-index for each pollutant is calculated to represent a relationship between pollutant concentrations and their health effects. The sub-indices for individual pollutants are calculated using their 24-hourly average concentration value (8-hourly in case of CO and Ozone) and health breakpoint concentration range [4]. Health breakpoint concentration range for IND-AQI system is shown in table 5.

The sub-index (I_P) for a given pollutant concentration (C_P), as based on linear segmented principle, is calculated as:

$$I_P = \left[\left\{\frac{I_{HI}-I_{LO}}{B_{HI}-B_{LO}}\right\} * (C_P - B_{LO})\right] + I_{LO} \quad (6)$$

where

B_{HI}: Breakpoint concentration greater or equal to a given concentration

B_{LO}: Breakpoint concentration smaller or equal to given concentration

I_{HI}: AQI value corresponding to B_{HI}

I_{LO}: AQI value corresponding to B_{LO}

Mathematical functions are used to aggregatesub-indices, I_P to obtain the overall index (I), referred to as AQI. In INDAQI System, sub-indices are aggregated using maximum operator.

$$AQI = \max(I_P) \quad (7)$$

where p = 1, 2, …, n; n denotes pollutant

AQI Category (Range)	$PM_{2.5}$ 24-hr	Ozone 8-hr	NO_2 24-hr	SO_2 24-hr	CO 8-hr (mg/m^3)
Good (0-50)	0-30	0-50	0-40	0-40	0-1.0
Satisfactory (51-100)	31-60	51-100	41-80	41-80	1.1-2.0
Moderately Polluted (101-200)	61-90	101-168	81-180	81-380	2.1-10
Poor (201-300)	91-120	169-208	181-280	381-800	10-17
Very Poor (301-400)	121-250	209-748*	281-400	801-1600	17-34
Severe (401-500)	250+	748+*	400+	1600+	34+

*One hourly monitoring (for mathematical calculation only)

4. Experiments and Results

4.1 Model Evaluation Parameters

The error functions used to measure the performance of the model are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of determination (R^2). Their mathematical representations are given as:

Root Mean Square Error (RMSE)

$$RMSE = \left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_{i_{pred}} - y_{i_{true}}\right)^2\right\}^{\frac{1}{2}} \quad (8)$$

Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n}\left|y_{i_{true}} - y_{i_{pred}}\right|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \quad (9)$$

Coefficient of Determination (R^2)

It is the proportion of the variance in the dependent variable that is predictable from the independent variable(s) [37].

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (10)$$

where SS_{res} denotes the residual sum of squares and is represented by,

$$SS_{res} = \sum_i \left(y_{i_{true}} - y_{i_{pred}}\right)^2 = \sum_i e_i^2 \quad (11)$$

and SS_{tot} denotes the total sum of squares and is represented by,

$$SS_{tot} = \sum_i \left(y_{i_{true}} - \bar{y}\right)^2 \quad (12)$$

here,

$y_{i_{true}}$ is the actual value

$y_{i_{pred}}$ is the predicted value

$\bar{y}$ is the mean value

e_i is the error

n is the total number of datapoints

4.1 *4.2 Experimental Parameters*

Proposed LSTM model is developed using multiple python packages like keras [38], scikit-learn [39] and tensorflow [40]. Min-Max Scaler of scikit-learn library is used to normalize the data in range between 0 and 1. Matplotlib

[41], python library for data visualization is used for plotting all the graphs. The architecture of LSTM model depends on multiple parameters like number of epochs, batch size, number of LSTM layers, number of units in each LSTM layer. These parameters are adjusted such that there is a balance between underfitting and overfitting. Dropout Layers are also used to prevent model from overfitting. Dropout layers randomly drop units from the network during training and prevent unit from co-adapting too much [42]. The model is trained for 50 epochs for a batch size of 15 using RMSprop optimizer. RMSprop is a gradient based optimization technique proposed by Geoffrey Hinton. It normalizes the gradient by using a moving average of squared gradients.It balances the step size by decreasing the step for large gradient to avoid explodingand increasing the step for small gradient to avoid vanishing.

4.3 Prediction Performance

After training the LSTM based RNN models, each pollutants' concentration was predicted for the next hour.

Figure 4 shows the predicted and observed concentration of $PM_{2.5}$ for testing data. From table, the R^2 between predicted and observed concentration indicates that the model explains 90% of the variability of the response data around its mean.

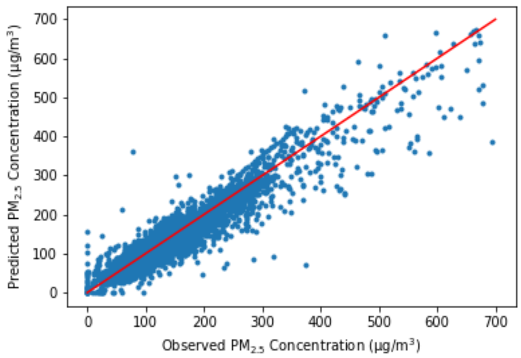


Figure4: Predicted and Observed Concentration of $PM_{2.5}$

Similarly, figures 5, 6, 7, 8, 9 show the predicted and observed concentration of Ozone, NO, NO_2, SO_2, and CO respectively.

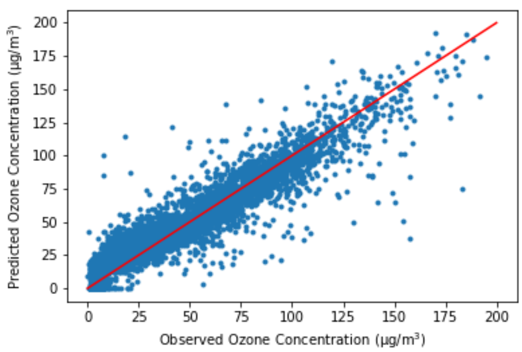


Figure5: Predicted and Observed Concentration of NO_2

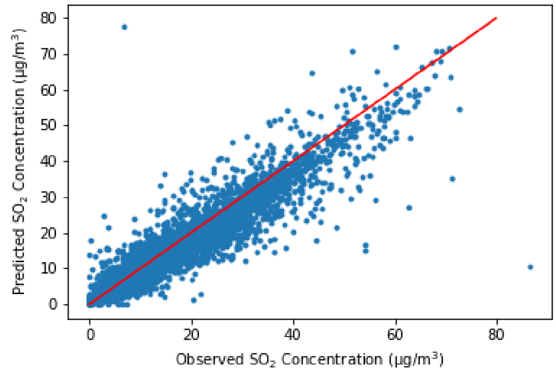


Figure6: Predicted and Observed Concentration of SO_2

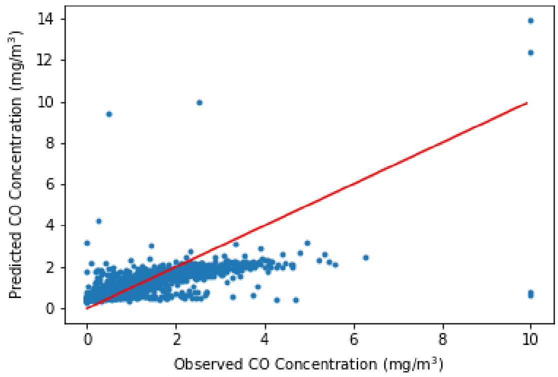


Figure7: Predicted and Observed Concentration of CO

4.4 Comparison of Results

The performance of proposed LSTM model and variants of baseline regression techniques like Support Vector Regressor (SVR) were compared against each other. SVR was modelled with linear kernel (LSVR), gaussian kernel (GSVR) and polynomial (III) kernel (PSVR). The error metrics RMSE, MAE and R^2 values of LSTM model were juxtaposed with the different SVR models. The values obtained are shown in table 6.

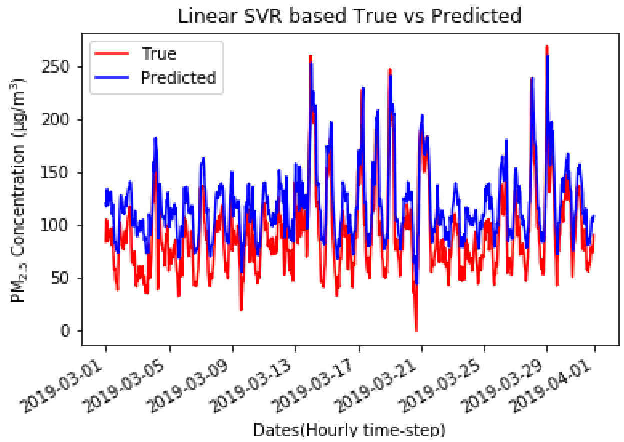
Table 6 indicates that proposed LSTM model achieved higher accuracy as compared to baseline regression techniques, SVR for time series data.

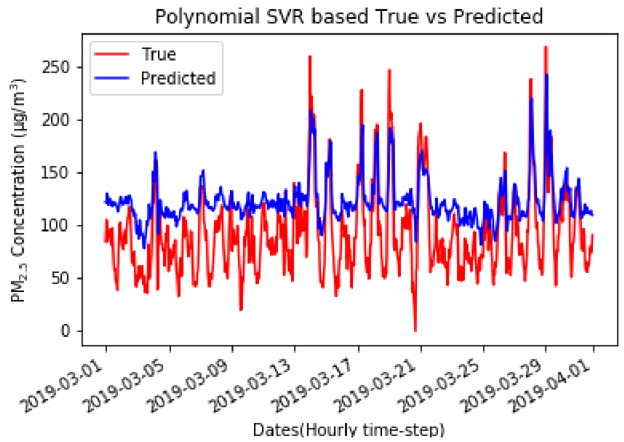
S-No	Pollutant	Linear-SVR			Gaussian-SVR			Polynomial-SVR			LSTM		
		RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
1	CO	3.59	3.55	-34.43	3.56	3.46	-33.85	4.31	4.27	-50.22	**0.40**	0.26	0.55
2	NO_2	15.47	13.81	0.27	17.75	15.97	0.04	17.33	14.99	0.08	**8.01**	4.06	0.80
3	NO	16.52	11.15	0.79	21.64	18.83	0.64	29.31	26.95	0.33	**13.47**	5.36	0.86
4	Ozone	42.51	39.92	-0.73	52.26	49.61	-1.62	66.48	62.90	-3.24	**11.22**	8.44	0.88
5	$PM_{2.5}$	35.00	29.18	0.80	38.72	32.97	0.76	35.45	27.63	0.80	**24.55**	15.78	0.90
6	SO_2	4.58	3.5	0.8	4.97	3.8	0.7	6.15	4.7	0.6	**3.58**	2.0	0.

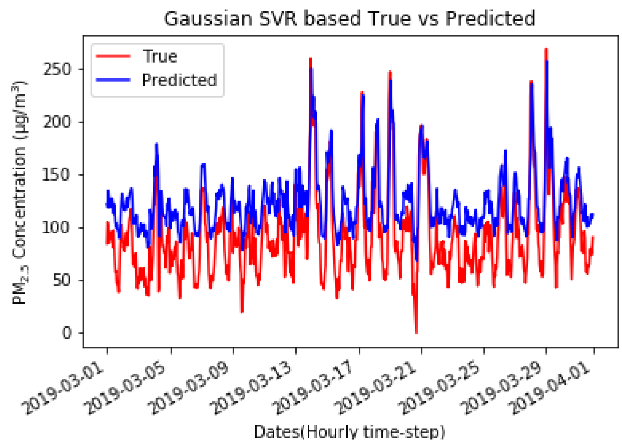
S-No	Pollutant	Linear-SVR			Gaussian-SVR			Polynomial-SVR			LSTM		
		RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
			4	2		6	9		1	8		1	89

Table 6: Performance Comparison of LSTM model with baseline models

Figure 10 shows the prediction comparison of LSVR, GSVR, PSVR models with LSTM model for March 2019 data. The LSTM model was better able to trace the changes in true values than SVR models.







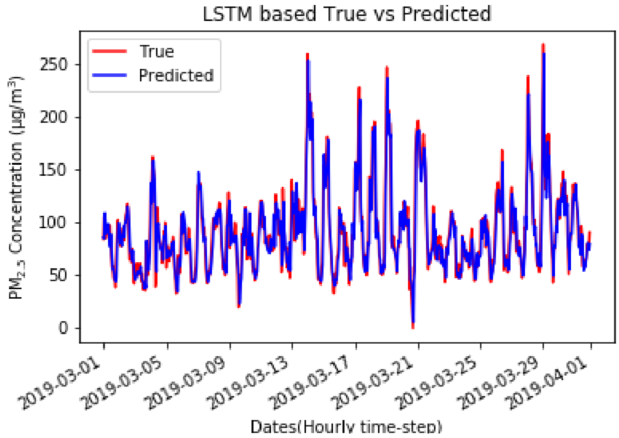


Figure 10: SVR Models vs LSTM Model Prediction performance graph for March-2019

4.5 AQI Prediction

The AQI values are calculated using predicted and actual concentrations. After comparing these values, RMSE of 12.79, MAE of 7.84 and R^2 of 0.99 were observed. Figure 11 shows the predicted and observed AQI values for testing data. Figure 12 represents the change in predicted AQI values with true AQI values for March 2019 data.

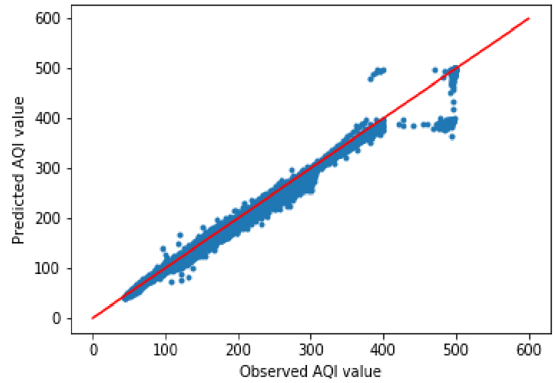


Figure 11: Predicted and Observed AQI value

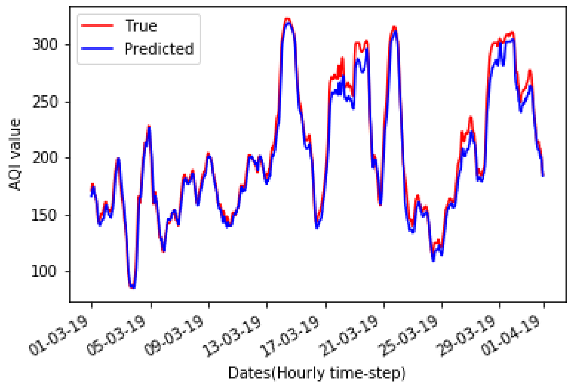


Figure 12: Mat plot of Predicted and True AQI value

The AQI values were then categorized into the 6 mentioned categories and an accuracy of 92% was observed. Table 7 and Figure 13 shows the classification report [43] and the normalized confusion matrix [44] respectively. Precision, Recall [45] and F1-Score [46] for 5 out of 6 categories are greater than 0.83. For "Good" AQI category, 83% of data samples are correctly labelled.

AQI Category	Precision	Recall	F1-Score	Support
Good	0.59	0.83	0.69	12
Satisfactory	0.92	0.97	0.95	1509
Moderately Polluted	0.92	0.94	0.93	2415
Poor	0.85	0.91	0.88	1884
Very Poor	0.97	0.90	0.93	2540
Severe	0.98	0.83	0.90	398

Table 7: Classification Report for AQI Categories

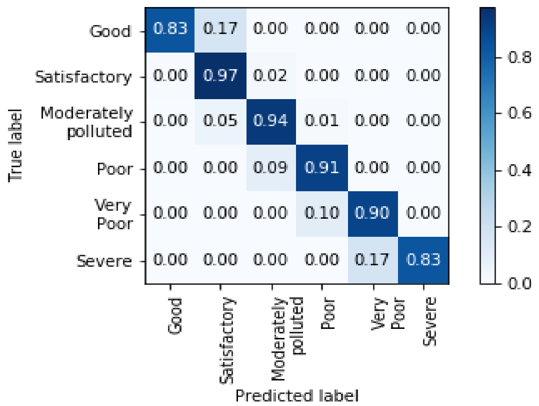


Figure 13: Normalized Confusion Matrix for AQI Categories

5. Conclusion

The objective of the current study is to establish an efficient forecasting model for AQI in Delhi. This paper proposed an RNN-LSTM model that predicts the hourly concentration of pollutants present in the air. The predicted concentrations are then used to calculate the AQI for a particular region in Delhi. The present study is carried out on 3.5 years ofhourly data from April 2015 to March 2019 with data from April 2017 to September 2017 was not available. Temporal sequences of four meteorological parameters and pollutant levels is fed as input to the LSTM model. The results show that deep learning-based techniques carry out promisingly than conventional statistical methods. This work can be extended by predicting a higher number of future timesteps for all the eight pollutants considered in calculating AQI.

References

[1] Health Effects Institute. 2019. State of Global Air 2019. Special Report. Boston, MA: Health Effects Institute. [online] Available https://www.stateofglobalair.org/sites /default/files/soga_2019_report.pdf. [Accessed: Oct. 2, 2019].

[2] World Health Organization, "9 out of 10 people worldwide breathe polluted air, but more countries are taking action", who.int, para. 1, May 2, 2018. [Online]Available:https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action. [Accessed: Oct. 2, 2019].

[3] World Health Organization, "How air pollution is destroying our health", who.int, para. 2, [Online] Available: https://www.who.int/air-pollution/news-and-events/how-air-pollution-is-destroying-our-health [Accessed: Oct. 2, 2019].

[4] Central Pollution Control Board,"National Air Quality Index".Ministry of Environment, Forest, and Climate Change, India, 2014. [Online] Available: https://www.cpcb.nic.in/displaypdf.php?id=bmF0aW9uYWwtYWlyLXF1YWxpdHktaW5kZXgvRklOQUwtUkVQT1JUX0FRSV8ucGRm [Accessed: Oct. 2, 2019]

[5] Air Visual, "World most polluted cities 2018 (PM2.5)", airvisual.com, 2018, [Online] Available: https://www.airvisual.com/world-most-polluted-cities [Accessed: Oct. 2, 2019].

[6] Nick Van Mead, "22 of world's 30 most polluted cities are in India, Greenpeace says", theguardian.com, para. 1, Mar. 5, 2019, [Online] Available: https://www.theguardian.com/cities/2019/mar/05/india-home-to-22-of-worlds-30-most-polluted-cities-greenpeace-says [Accessed: Oct. 2, 2019].

[7] Press Information Bureau, "National Air Quality Index (AQI) launched by the Environment Minister AQI is a huge initiative under 'Swachh Bharat'", pib.gov.in, Oct. 17, 2014, [Online] Available: https://pib.gov.in/newsite/PrintRelease.aspx?relid=110654 [Accessed: Oct. 2, 2019].

[8] Niharika, Venkatadri M, and Padma S Rao,"A Survey on Air Quality forecasting Techniques", [J] International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp.103-107, 2014

[9] A. Graves, J. Schmidhuber,"Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks", Advances in Neural Information Processing Systems 22, NIPS'22, pp 545–552, Vancouver, MIT Press, 2009.

[10] A. Graves, J. Schmidhuber,"Framewise phoneme classification with bidirectional LSTM and other

neural network architectures", Neural Networks. 18 (5–6): 602–610, 2005. https://doi.org/10.1016/j.neunet.2005.06.042

[11] Fernández S., Graves A., Schmidhuber J. (2007) "An Application of Recurrent Neural Networks to Discriminative Keyword Spotting", In: de Sá J.M., Alexandre L.A., Duch W., Mandic D. (eds) Artificial Neural Networks – ICANN 2007. ICANN 2007. Lecture Notes in Computer Science, vol 4669. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74695-9_23

[12] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks",In Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 6645–6649

[13] R. Collobert, J. Weston,"A Unified architecture for natural language processing: Deep neural networks with multitask learning" [C], In Proceedings of the 25th International Conference on Machine Learning, 2008, 5-9.

[14] Wikipedia, "Deep Learning", Wikipedia.org, para. 1, [Online]Available:https://en.wikipedia.org/wiki/Deep_learning. [Accessed: Oct. 3, 2019].

[15] B. Hochreiter and J. Schmidhuber,"Long Short Term Memory", [J] Neural Computation, 1997, 9(): 1735-1780.

[16] Anikender Kumar and Piyush Goyal,"Forecasting of air quality in Delhi using principal component regression technique". Atmospheric Pollution Research. 2. 436-444. 10.5094/APR.2011.050, 2011.

[17] Ni XY, Huang H, Du WP,"Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multisource data". Atmos Environ 150:146–161, 2017.

[18] Li. C, Hsu N.C., Tsay. S. "A Study on the potential applications of satellite data in air quality monitoring and forecasting" [J], Atmos.Environ, 45:3663-3675, 2011.

[19] Bing-Chun Liu, ArihantBinaykia, Pei-Chann Chang, Manoj Tiwari, Cheng-Chin Tsao,"Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang."PLoS ONE. 12. 10.1371/journal.pone.0179763, 2017.

[20] Nieto P.G, Combarro E.F, Del Coz Diaz J.J,"A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study", [J], Appl. Math. Comput, 8923-8937, 2013.

[21] Jain S, Khare M, "Adaptive neuro-fuzzy modeling for prediction of ambient CO concentration at urban intersections and roadways.", Air Qual Atmos Health 3:203–212, 2010.

[22] Athanasiadis IN, Kaburlasos VG, Mitkas PA, Petridis V. "Applying machine learning techniques on air quality data for real-time decision support". In: First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland, 2003.

[23] Athira V, Geetha P, Vinayakumar R, Soman K P. "DeepAirNet: Applying Recurrent Networks for Air Quality Prediction", Elsevier, Procedia computer science, 132:1394-1403, 2018.

[24] Xiang Li et al., "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation" [J], Elsevier Ltd., Environmental Pollution, 231: 997-1004, September, 2017.

[25] Vikram Reddy et al.,"Deep Air: Forecasting Air Pollution in Beijing, China", 2017. [online] Available https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/deep-air-forecasting_final.pdf

[26] Mrigank Krishan, Srinidhi Jha, Jew Das, Avantika Singh, Manish Goyal, ChandrraSekar, "Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India". Air Quality Atmosphere & Health. 10.1007/s11869-019-00696-7, 2019.

[27] K. Srinivasa Rao, G. Lavanya Devi, N. Ramesh, "Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks", International Journal of Intelligent Systems and Applications (IJISA), Vol.11, No.2, pp.18-24, 2019. DOI: 10.5815/ijisa.2019.02.03

[28] Wikipedia, "Delhi", wikipedia.org, para. 1, [online] Available: https://en.wikipedia.org/wiki/Delhi, [Accessed: Oct. 3, 2019].

[29] United Nations, "2018 Revision of World Urbanization Prospects", un.org, para. 8, May 16, 2018, [online] Available: https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html, [Accessed: Oct. 3, 2019].

[30] JasjeevGandhiok , "'Green' Dwarka is the Delhi's new pollution hotspot", timesofindia.in, para. 1, Oct. 23, 2018, [online] Available: https://timesofindia.indiatimes.com/city/delhi/green-dwarka-is-the-delhis-new-pollution-hotspot/articleshow/66324470.cms, [Accessed: Oct. 3, 2019].

[31] Central Pollution Control Board, "Central Control Room for Air Quality Management - All India", cpcb.nic.in, [online] Available: https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/cacaqm-data-availibility.

[32] Google Maps, "Netaji Subhas University of Technology", google.com, [online] Available: https://goo.gl/maps/B51kvdwraCLUxfEB7

[33] Wikipedia, "Interpolation", wikipedia.org, [online] Available: https://en.wikipedia.org/wiki/Interpolation [Accessed: Oct. 3, 2019].

[34] Wikipedia, "Rescaling (min-max normalization)", wikipedia.org, [online] Available:

https://en.wikipedia.org/wiki/Feature_scaling#Rescaling_(min-max_normalization)

[35] Wikimedia, “The LSTM Cell”, Wikimedia.org, [online] Available: https://upload.wikimedia.org/wikipedia/commons/3/3b/The_LSTM_cell.png

[36] Wikipedia, “LSTM with a forget gate”, wikipedia.org, [online] Available: https://en.wikipedia.org/wiki/Long_short-term_memory#Variants [Accessed: Oct. 3, 2019].

[37] Wikipedia, “Coefficient of determination”, wikipedia.org, [online] Available: https://en.wikipedia.org/wiki/Coefficient_of_determination

[38] François Chollet et al.,“Keras”, github.com, 2015 [online] Available: https://github.com/fchollet/keras, [Accessed Oct. 3, 2019].

[39] Pedregosa et al., “Scikit-learn: Machine Learning in Python”, JMLR 12, pp. 2825-2830, 2011.

[40] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, “TensorFlow: A System for Large-scale Machine Learning”, In Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI’16). USENIX Association, Berkeley, CA, USA, 265–283, 2016. http://dl.acm.org/citation.cfm?id=3026877.3026899

[41] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[42] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. Journal of Machine Learning Research, 15. 1929-1958, 2014.

[43] Scikit Learn, “Classification Report”, scikit-learn.org, [online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html, [Accessed: Oct. 3, 2019].

[44] Wikipedia, “Confusion Matrix”, wikipedia.org, [online] Available: https://en.wikipedia.org/wiki/Confusion_matrix, [Accessed: Oct. 3, 2019].

[45] Wikipedia, “Precision and recall”, wikipedia.org, [online] Available: https://en.wikipedia.org/wiki/Precision_and_recall, [Accessed: Oct. 3, 2019].

[46] Wikipedia, “F1 Score”, wikipedia.org, [online] Available: https://en.wikipedia.org/wiki/F1_score, [Accessed: Oct. 3, 2019].

AUTHOR

Mohit Bansal is currently pursuing Bachelor of Engineering in Information Technology from Netaji Subhas University of Technology (Formerly Netaji Subhas Institute of Technology, University of Delhi), New Delhi, India. His areas of research include Artificial Intelligence, Machine Learning, Deep Learning, and Computer Vision.

Anirudh Aggarwal is currently pursuing Bachelor of Engineering in Information Technology from Netaji Subhas University of Technology (Formerly Netaji Subhas Institute of Technology, University of Delhi), New Delhi, India. His areas of research include Artificial Intelligence, Machine Learning, Deep Learning, and Computer Vision.

Tanishq Verma is currently pursuing Bachelor of Engineering in Information Technology from Netaji Subhas University of Technology (Formerly Netaji Subhas Institute of Technology, University of Delhi), New Delhi, India. His areas of research include Artificial Intelligence, Machine Learning, Deep Learning, and Natural Language

Ms. ApoorviSood completed her Bachelor of Engineering in Computer Science and Engineering. She was a Gold medalist in Masters of Technology from ITM University, M.D.U., Rohtak. Her pre submission of PhD thesis was done from GGSIPU.