Pollutant Level Concentration in Delhi using Deep Neural Networks: A Study

Remya Ravikumar [1], Nagesh Subbana [2], Alka Singh [3]

[1, 2, 3] Centre for Wireless Network and Applications (WNA), Amrita Vishwa Vidyapeetham, Amritapuri, India.

Email: [1] remya95ravikumar@gmail.com, [2] Snag1978@yahoo.com, [3] Alka228@gmail.com

Abstract

Air pollution is one of the major concerns that plagues the world and has consequences in different spheres of our lives. Timely information regarding air pollution is sparse and often times goes unnoticed. The air quality data obtained from orbital sensor like Sentinel 5 and 5P TROPOMI and ground sensors (Central pollution control board sensors, CPCB) provide a large amount of information about the particle pollutants present in the atmosphere. This study used the previous 24 hours data to predict the next 24 hours of air pollutant concentration level for PM2.5, NO2, CO and SO2. The region of study is Delhi, North India because this region falls in the top ten most polluted cities worldwide. Convolutional neural networks (CNN) and long short-term memories (LSTM) are two examples of deep neural networks that have demonstrated significant advantages in tackling nonlinear spatiotemporal issues. They can accurately represent temporal and spatial information and extract useful contextual elements to integrate temporal properties. Therefore, we propose a combination of CNN and LSTM models for precise air quality prediction and validation in the region for the upcoming twenty-four hours based on data acquired on the preceding twenty-four hours.

Index Terms— Air pollution, Orbital sensors, Spatiotemporal, Deep Learning, LSTM.

1. Introduction

Anything becomes a pollution only when it is detrimental to human beings, and air pollution refers to particles or objects which are present in the atmosphere which are detrimental to our livelihood. As air is not stagnant the effect of air pollution is not only restricted to the source area of pollution but rather it spreads to wider region. According to WHO article published in December 2022 around 7 million people lose their life due to causes attributed to air pollution[1], [2]. Some of the major sources of air pollution are vehicular emissions, industrial emissions, agricultural, residential and commercial energy use. These sources contribute to increase in concentration of particulate matter size 2.5 (PM2.5), ozone, Nitrogen oxides, sulphur dioxide, carbon monoxide etc., in the atmosphere which are the major causes of air pollution.

India with its huge demography and unique geography is not a stranger to air pollution. A 2019 report by WHO states that 21 out of the 30 most polluted cities in the world are located in India. According to estimates, air pollution is a contributing factor in the deaths of over two million Indians[3]. The leading contributors of air pollution in the country are vehicular emission, industrial, periodic agricultural pollutants, unrestrained emission sources, unfavourable meteorological conditions, and household pollution[4], [5]. India also stands 5th

in a study conducted recently by WHO which analysis the impact of air pollution on population[6]. The Central pollution control board (CPCB) of India, has set up monitoring stations around the country to determine the levels of pollution concentration. This forms a valuable and reliable source of data. Moreover many orbital sensors like sentinel 5P can be used to obtain the pollution details on a larger scale.

Researchers primarily take into account various machine learning and deep learning techniques like linear regression models, including linear regression, multiple linear regression, and Gaussian regression because they are relatively quick and accurate but fall short for the non-linear and unstable air quality prediction problem[7], [8]. Then, the neural network and spatiotemporal modelling have since emerged and applied in the field of air quality which did not significantly improve the accuracy[9]–[12]. The Deep Neural Network has shown major advantage in solving nonlinear problems as it can extract useful data to the greatest degree for time attributes and comprehend time dependencies well[13]. The air quality data obtained from overhead and ground sensors provide rich information about the particle pollutants present and the subsequent air quality index (AQI). Our study proposes to combine the data obtained from both of these sensors to compare fine grained properties from the coarse-grained overhead images using deep neural networks and to create an air quality data model[14]–[17].

2. Study Area

The study area in focus falls in the Indo-Gangetic Plain which comprises of the area between the banks of rivers Indus and Ganga. It is a 700 thousand km2 area encompassing northern regions of the Indian subcontinent. For this study we are mainly focusing on Delhi State 28.65195°N 77.23149°E and it falls in the top 10 most polluted cities of the world[18], [19].

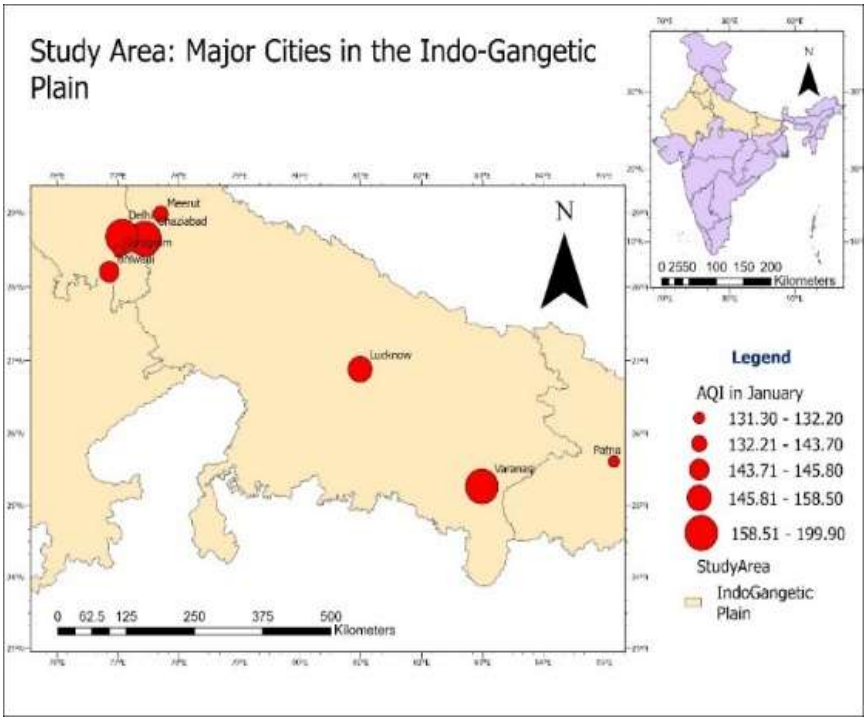


Fig 1. Major cities in the IndoGangetic plain depicting the Air quality index (AQI) as on January 2020

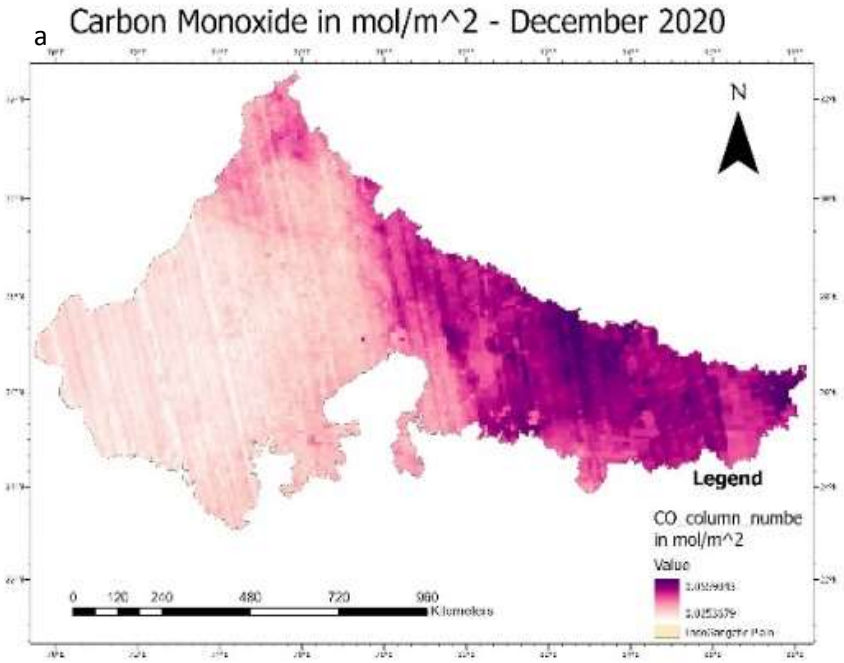
3. Data Collection

A. Remote Sensing Data

Open source remote sensing data like American Landsat and European Sentinel are important and gives free remote sensing images[20]. For this study we have taken data from Sentinel 5P TROPOMI which is an imaging spectrometer covering wavelength bands between the ultraviolet and the shortwave infrared having resolution of 7x3.5 km2. The remote sensing data from Sentinel 5P is extracted from Google earth engine for the period of two year (2020-2022) for the proposed study area. The images obtained for CO, NO2 and SO2 are shown in figure 2(a-c). The images that are obtained from sentinel are such that each band corresponds to one pollutant content. In order to obtain this image we have extracted the individual bands from the images. Once the data has been obtained it is visualised using ArcGIS Pro and the pixel values corresponding to the desired locations are extracted. The pixels corresponding to the region was extracted with help of google earth engine codes and using the same they were converted into comma separated files (csv) files for being fed into the neural network. The images obtained has pollutant concentrations in mol/m2 which were then converted into micro gram per meter square with appropriate formula after referring to sentinel 5P documentation.

B. Ground Sensor Data

The ground sensor data are obtained from central pollution control board official website. The site contains hourly pollutant level concentration of 14 pollutants which contribute to air pollution along with curated air quality index for the said hour and the prominent pollutant at that time. For our study we formulated the data by taking vales of NO2, SO2, CO and PM2.5 from the website for a period of two year (2020-2022). The same source can be used for obtaining the meteorological data for the same sensor location on a daily basis. For the purpose of our study we are currently not considering the meteorological parameters.



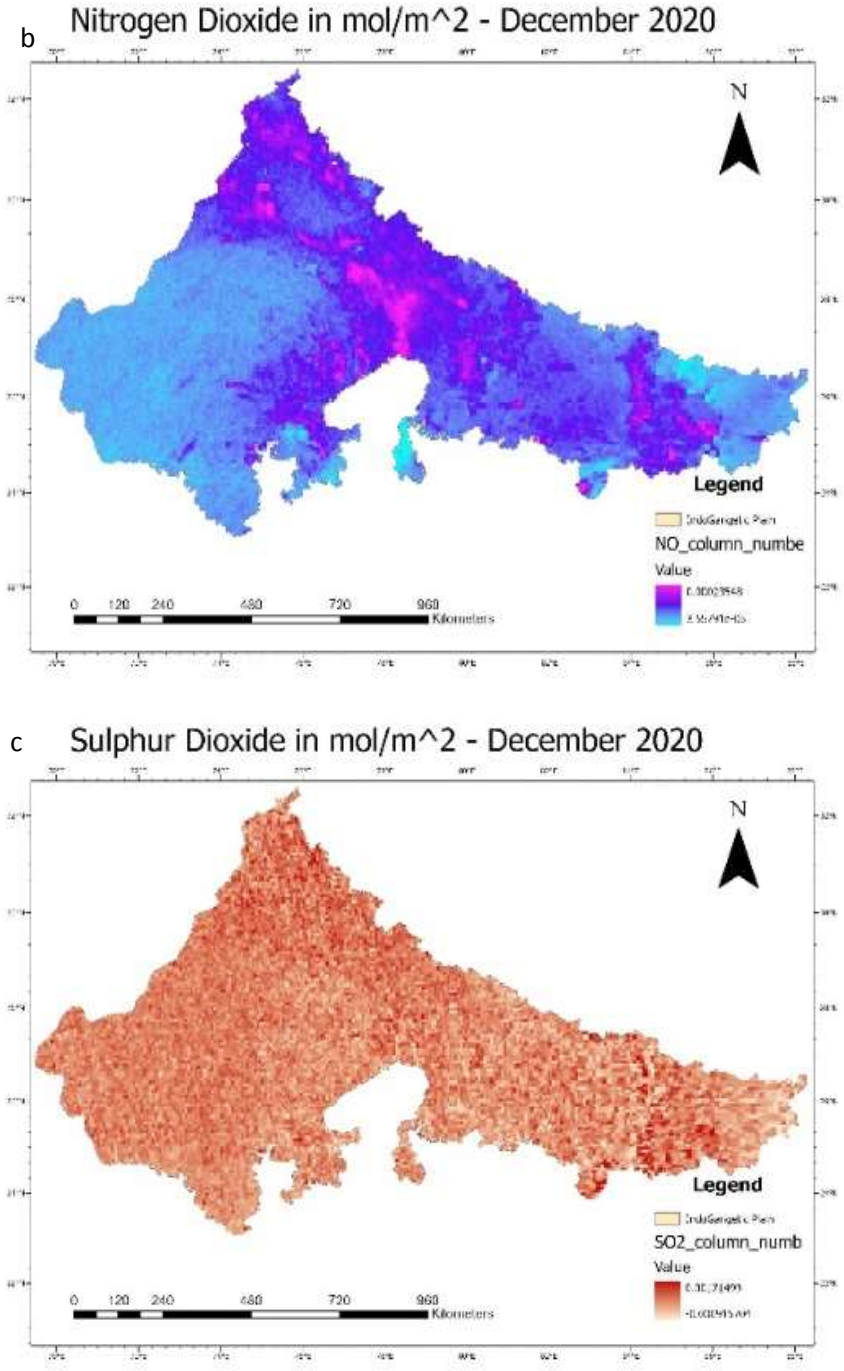


Fig 2 (a-c). The images obtained for CO, NO2 and SO2 from sentinel 5P for the year 2020

4. Methodology

The satellite images were initially converted to csv file using google earth engine which contains daily data. The ground sensor data has hourly data hence for the same day we have 24 values from ground sensor but only one value from orbital sensor. In order to make the data points same in both the cases we created 24 copies of the daily data (satellite image) for each date. Hence we considered the same value of pollutant concentration in each hour for satellite data. The missing values in the ground and orbital sensors were filled using mice algorithm. MICE stands for Multivariate Imputation by Chained Equations algorithm, a technique by which we can effortlessly impute missing values in a dataset by looking at data

from other columns and trying to estimate the best prediction for each missing value. The k-nearest neighbour algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another. Initially we applied K nearest neighbour algorithm for filling the missing data but found mice algorithm provided a better imputation efficiency. Figure 3 illustrates the methodology used in the study.

A. Models

1) CNN model: A Deep Learning technique specifically created for working with images is the convolutional neural network. It uses images as inputs, extracts and learns the attributes, then categorises the images using the learnt features. The satellite data with its daily images were initially run through a 1 dimensional convolution model with just one years (2020) data for NO2 and CO pollutant content. The 2020-2021 data were taken for training purpose and testing and validation were done using 2022 data in 1:1 ratio. The missing and NaN values were imputed using a KNN impute algorithm which considered the previous two days as input for imputing. The network takes one input parameter namely the pollutant content and is analysed against time.

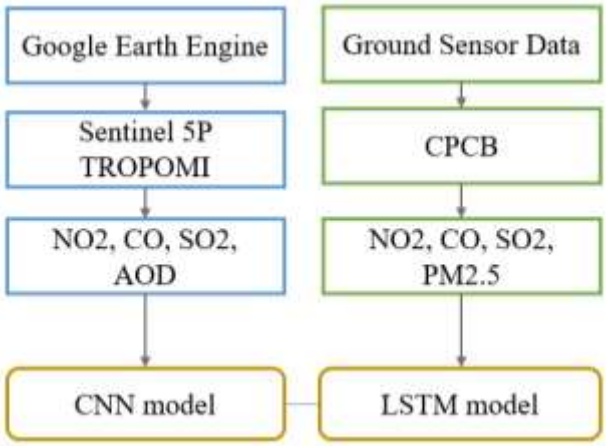


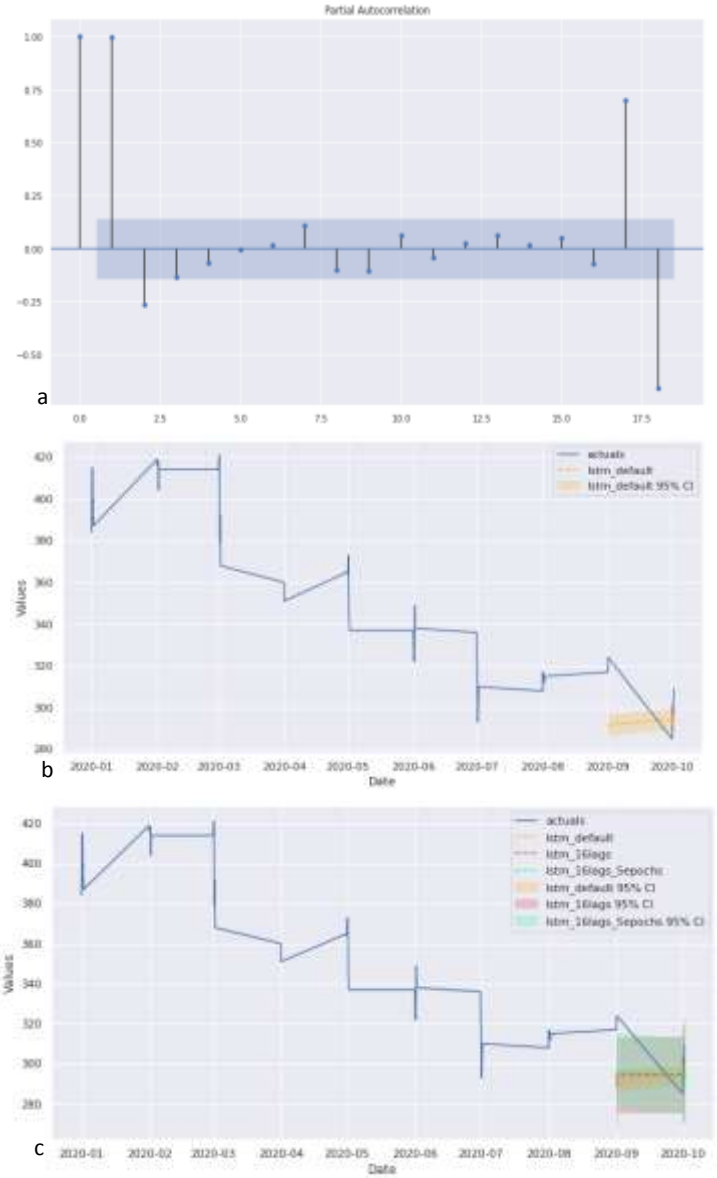
Fig 3. Methodology

Since the number of observations were less we introduced a rolling input were the first three would be taken for training and the fourth would be predicted, then the next three would include the previously predicted value. We were able to observe the trend of the pollutant concentration and the predict the pollutant content concentration for the next 24 hours.

2) LSTM model: LSTM is a type of recurrent neural network (RNN) that excels at learning long-term dependencies, particularly in issues involving sequence prediction. It has feedback connections, making it capable of processing the complete sequence of data.

The initial analysis for ground sensor data was done by considering the years (2020-2021). In this study only the pollutant concentration change with time is considered and the geographic or meteorological factors are ignored. Hence a partial auto correlation was done in order to find how the previous observations have effect on the succeeding data (Fig4a). The data was fed into an LSTM network as they are better suited for large temporal resolutions. We used a forecaster method for prediction with an LSTM estimator as the forecasting parameter. By default, this model will be run with a single input layer of 8 size, Adam optimizer, tanh

activation, a single lagged dependent-variable value to train with, a learning rate of 0.001, and no dropout. All data is scaled going into the model with a min-max scaler and un-scaled coming out figure 4b. Next we introduce a lag of 16, (lag features are target values from previous periods) this particular number was selected because while running the partial auto correlation for the pollutant concentrations it was observed that around 16 data points fall more or less in the range figure 4a. An epoch is then introduced into the network to analysis how the model works. We can see in the figure 4c that as we introduced different combinations of epochs and lags the prediction range is becoming more inclusive. It starts shifting to accommodate the peak value on both ends. Figure 4d shows the output after we introduce a lag of 24, an early stop and a patience of 5, it essentially tells the model to quit after more than 5 iterations if the result does not improve (after 5 iterations). Figure 4e shows the best LSTM model for the dataset. It can be observed that the best model incorporates all the peaks within its range.



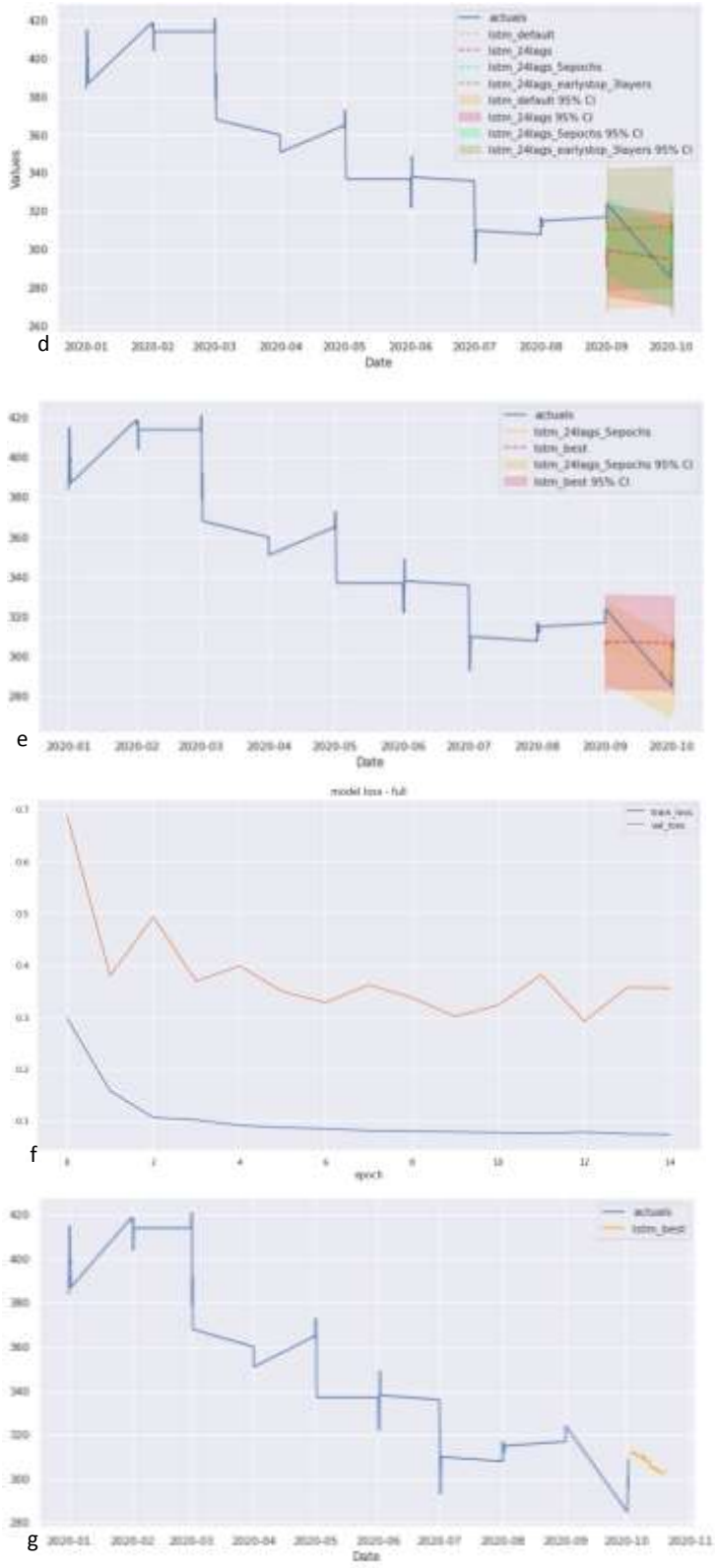


Fig. 5 (a-g). Partial autocorrelation ran on NO2; LSTM with default parameters; LSTM estimator after introducing a lag of 16; LSTM with different combination of parameters; Best LSTM parameters which includes most of the peak values; Model loss; The prediction using LSTM estimator for the next 24 hours

5. Results and Discussion

A one dimensional convolution model was run for the satellite image for the region for the period of two years. The 2020-2021 data was taken as the training set and the testing was done on 2022 image. It can be observed from fig 6(a-c) that there is a significant increase in the pollutant concentration level in the winter months (November-March) compared to summer or monsoon. It could be because of the increase in agricultural pollutant content as a result of large scale stubble burning. Moreover the northern plain is subjected to higher vehicular traffic because of the dense pollution. From fig 8(a-i) we can see the pollutant level change that happened within the span of two years. It can be observed that in the year 2020 the pollution level for all the three pollutants (SO2, NO2 and CO) is lesser compared to 2021 and 2022. Moreover there seems to be steady increase in the pollution levels with each succeeding year. We have observed the pollutant levels in the year 2019 and found that the 2022 levels are on par with the 2019 levels. The decrease in concentration of pollutants in 2020 can be created to the stringent lockdown measures in the country due to Covid19. So we concluded that the decrease is the exception and the pollutants levels are on the rise year on year. There is difference in pollutants levels from region to region with urban areas showing more pollution than rural. The urban/rural angle is an ongoing part of this project but is not discussed in this paper.

The dataset selected was for Delhi region for a period of two year (2020-2022) and which were formulated using CPCB data. The training set was taken as 2020-21 data and the testing was done on 2022 data.

The input for our LSTM model was the previous 24 hours pollutant concentration level of NO2, SO2, CO and the output obtained from 1D CNN for the same time period. The output was the predicted sequence of the pollutants contents in the next 24 hours. Here we have validated the model by trying to fit it within the preceding 24 hours data. The figure 4f shows the model loss as we can observe that the model has relatively less validation loss and training loss which implies that our model is correct. Figure 4g shows the prediction for the next 24 hours, it follows the trend which has been observed in the previous dates. A multiple linear regression model was also ran to check if it could provide a better result and for the next 24 hours and it was found that LSTM model gives a better prediction.

The model accuracy was considered with root mean square error for both MLR and LSTM models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

Where RMSE = Root mean squared error

i = Variable (Pollutant concentrations)

N = Number of observations

x_i = Predicted value

$\hat{x}_i$ = Actual value.

The RMSE values were 17.45 for LSTM model and 20.31 for MLR model.

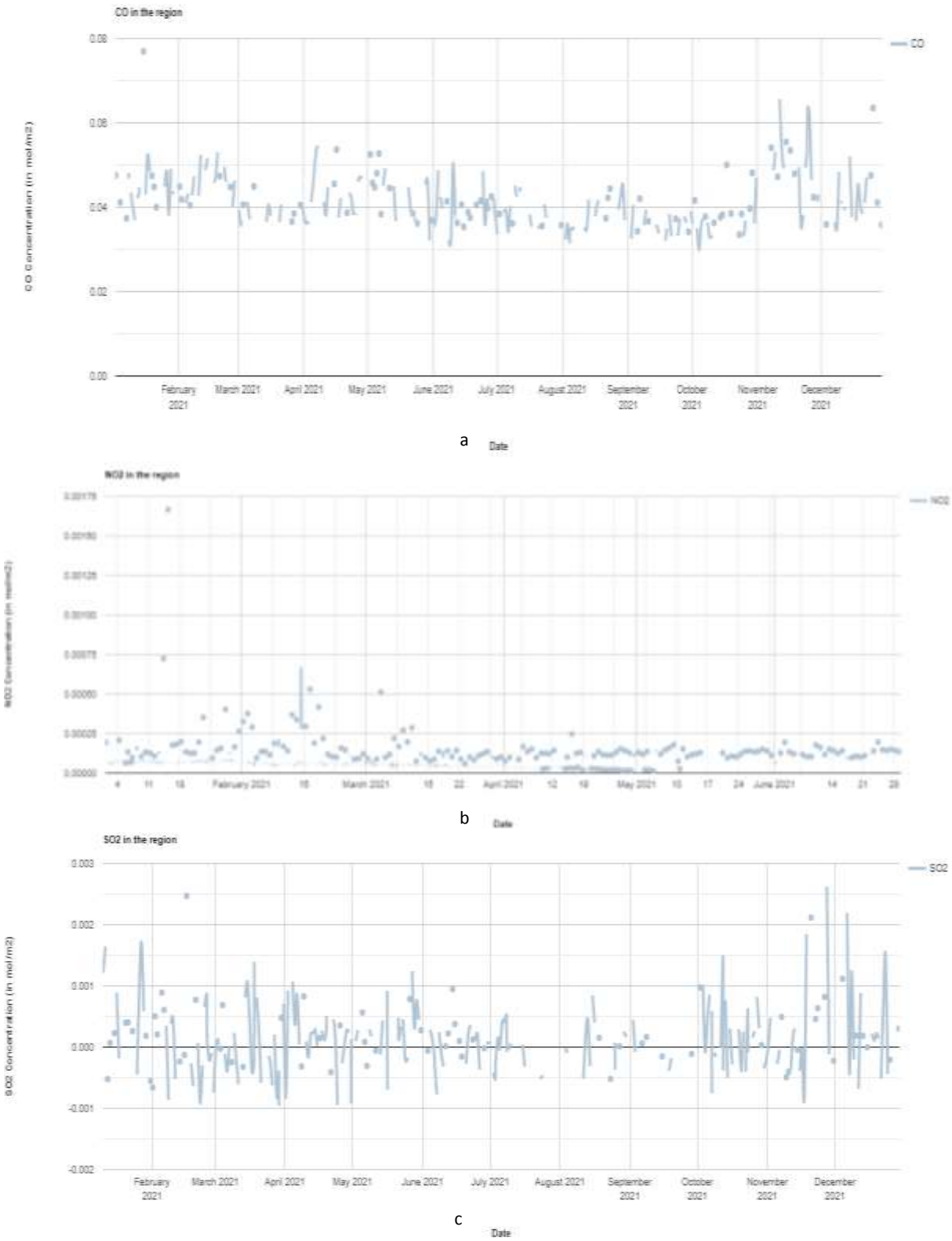


Fig. 6 (a-c). The three graphs depicts the variations in the pollutant level concentrations of CO, NO2 and SO2 for the year 2021. The monthly variation shows an increase in pollutant concentration for the months from November-march (coinciding with the winter months in India).

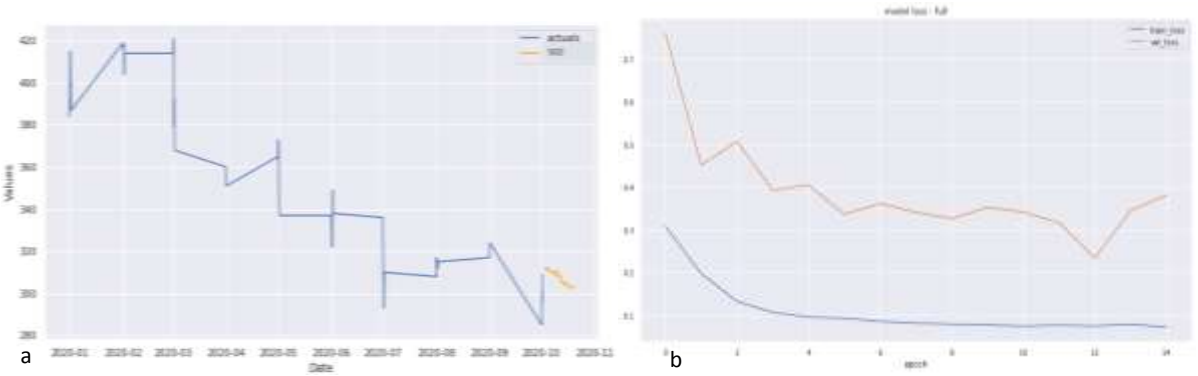


Fig. 7 (a, b). The prediction using CNN for the next 24 hours for pollutant data obtained from satellite images; Model loss curve

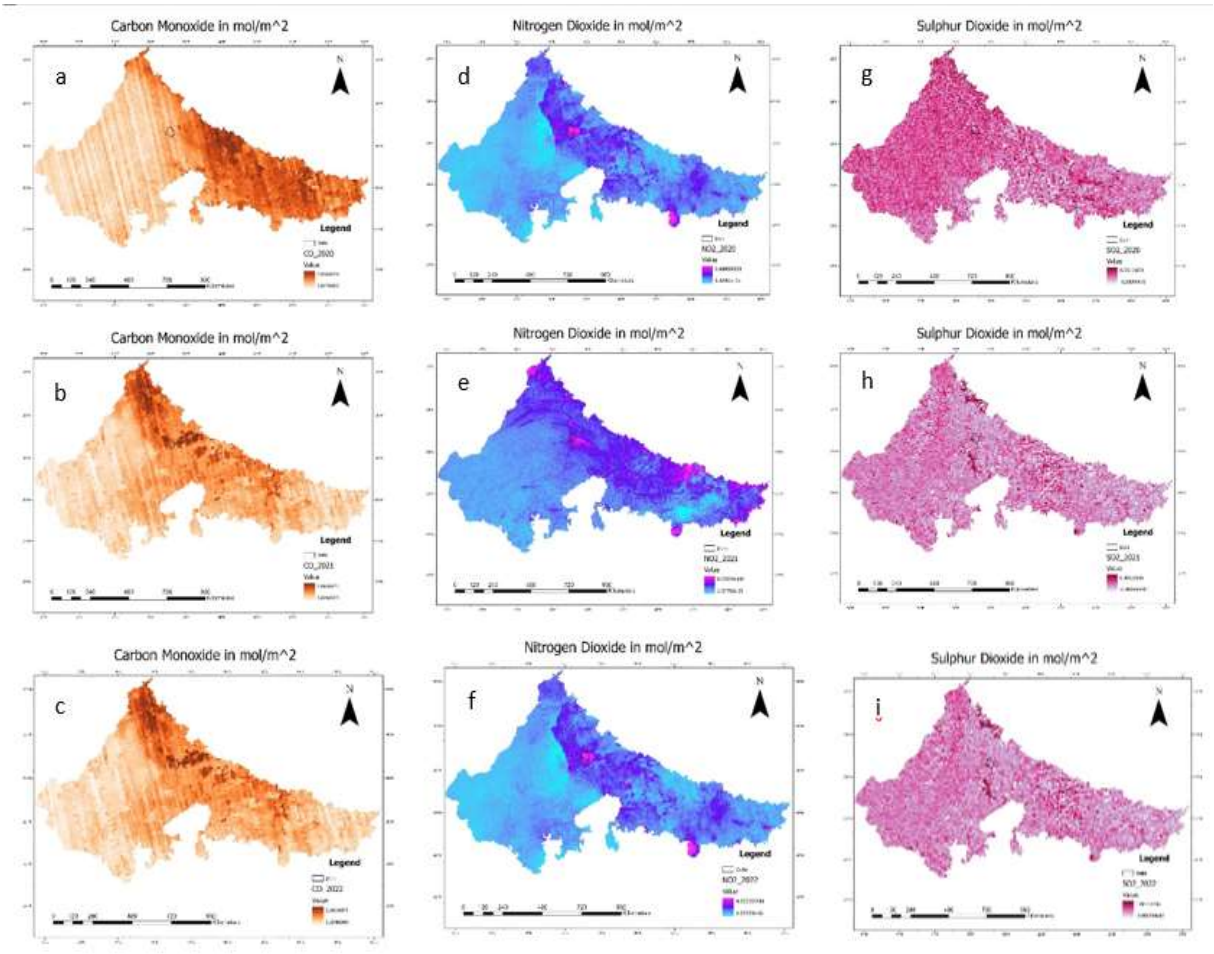


Fig 8 (a-i). The pollutant concentration levels for CO, NO2 and SO2 from the years 2020-2022. (a-c) shows the yearly variation observed for CO, (d-f) for NO2 and (g-i) for SO2

6. Conclusion

Air pollution poses a major concern to people's lives. Over two million Indians are said to lose their life to causes attributed to air pollution. Most of this pollution comes from industries closely followed by vehicular pollution, unrestrained emission sources, periodic agricultural pollutants, and household pollution[21]. The study explored the fusion of data obtained from ground (CPCB) and orbital sensors to better estimate the air quality parameters like PM2.5, NO2, CO and SO2. The combination of these improved parameters will subsequently enhance the pollutant prediction. In this study, we developed a Long short term

memory based neural network based air quality model. The model was able to correctly predict the pollutant concentration for the next 24 hours and the prediction was validated using the Convolution neural network (CNN) as well as with the previous observations. Deep Neural Networks like Convolutional Neural Network (CNN) and Long short-term memory (LSTM) have shown major advantage in solving nonlinear spatio-temporal problems. They can extract valuable contextual features to combine temporal attributes, and model temporal and spatial dependencies accurately. In future works we propose a hybrid CNN-LSTM model which will combine the spatial attributes of orbital sensors using a CNN and temporal aspects of ground sensors using an LSTM, thereby creating a more dynamic and accurate pollutant prediction. Using the combination of these improved parameters we will subsequently determine an enhance air quality index (AQI) approximation. New research that could compliment the ongoing work will be by introducing meteorological parameters and also to analyse the impact on health of the population by using the air quality model created.

Reference

[1] WHO, "Ambient (outdoor) air pollution," Dec. 19, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] "IQAir." [Online]. Available: https://www.iqair.com/about-iqair

[3] K. Balakrishnan et al., "The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017," Lancet Planet. Health, vol. 3, no. 1, pp. e26–e39, Jan. 2019, doi: 10.1016/S2542-5196(18)30261-4.

[4] M. D. Adams and P. S. Kanaroglou, "Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models," J. Environ. Manage., vol. 168, pp. 133–141, Mar. 2016, doi: 10.1016/j.jenvman.2015.12.012.

[5] B. R. Gurjar et al., "Human health risks in megacities due to air pollution," Atmos. Environ., vol. 44, no. 36, pp. 4606–4613, Nov. 2010, doi: 10.1016/j.atmosenv.2010.08.011.

[6] "World Health Organization (WHO)." [Online]. Available: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ambient-air-pollution

[7] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," Complexity, vol. 2020, pp. 1–23, Aug. 2020, doi: 10.1155/2020/8049504.

[8] the Department of Computer Engineering, San Jose State University, USA, G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," Int. J. Environ. Sci. Dev., vol. 9, no. 1, pp. 8–16, 2018, doi: 10.18178/ijesd.2018.9.1.1066.

[9] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," Sustain. Cities Soc., vol. 67, p. 102720, Apr. 2021, doi: 10.1016/j.scs.2021.102720.

[10] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," Environ. Sci. Pollut. Res., vol. 23, no. 22, pp. 22408–22417, Nov. 2016, doi: 10.1007/s11356-016-7812-9.

[11] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep Learning for Air Quality Forecasts: a Review," Curr. Pollut. Rep., vol. 6, no. 4, pp. 399–409, Dec. 2020, doi: 10.1007/s40726-020-00159-z.

[12] Q. Zhang, F. Fu, and R. Tian, "A deep learning and image-based model for air quality estimation," Sci. Total Environ., vol. 724, p. 138178, Jul. 2020, doi: 10.1016/j.scitotenv.2020.138178.

[13] K. Tripathi and P. Pathak, "Deep Learning Techniques for Air Pollution," in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India: IEEE, Feb. 2021, pp. 1013–1020. doi: 10.1109/ICCCIS51004.2021.9397130.

[14] F. Hamami and I. A. Dahlan, "Univariate Time Series Data Forecasting of Air Pollution using LSTM Neural Network," in 2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS), Lombok, Indonesia: IEEE, Oct. 2020, pp. 1–5. doi: 10.1109/ICADEIS49811.2020.9277393.

[15] M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal, and C. Sekar, "Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India," Air Qual. Atmosphere Health, vol. 12, no. 8, pp. 899–908, Aug. 2019, doi: 10.1007/s11869-019-00696-7.

[16] R. Navares and J. L. Aznarte, "Predicting air quality with deep learning LSTM: Towards comprehensive models," Ecol. Inform., vol. 55, p. 101019, Jan. 2020, doi: 10.1016/j.ecoinf.2019.101019.

[17] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, and B. Zhang, "A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM $_{2.5}$ Concentration," IEEE Access, vol. 7, pp. 20050–20059, 2019, doi: 10.1109/ACCESS.2019.2897028.

[18] IQAir, "Air quality and pollution city ranking," Switzerland. [Online]. Available: https://www.iqair.com/world-air-quality-ranking

[19] IQAir, "World's most polluted cities (historical data 2017-2022)." [Online]. Available: https://www.iqair.com/world-most-polluted-cities

[20] S. K. Saha, P. S. Pang, and D. Bhattacharyya, Eds., Smart Technologies in Data Science and Communication: Proceedings of SMART-DSC 2021, vol. 210. in Lecture Notes in Networks and Systems, vol. 210. Singapore: Springer Singapore, 2021. doi: 10.1007/978-981-16-1773-7.

[21] R. Ravikumar, N. Subbana, A. Singh, and R. Vargas Maretto, "Assessment of Impact of air pollution using deep learning-based Air Quality data Model ," oral, other, Feb. 2023. doi: 10.5194/egusphere-egu23-728.