Air Pollution Forecasting using Machine Learning and Deep Learning Techniques

Samriddhi Banara
2k18/IT/108(Information technology)
Delhi Technological University
Delhi, India
samriddhibanara@gmail.com

Teena Singh
2k18/IT/126(Information Technology)
Delhi Technological University
Delhi, India
twinklesingh1903@gmail.com

Yash Thenuia
2k18/IT/108 (Information Technology)
Delhi Technological University
Delhi, India
yash.thenuia21@gmail.com

Himanshu Nandanwar
Information Technology
Delhi Technological University
Delhi, India
himanshunandanwar09cm@gmail.com

Anamika Chauhan
Information Technology
Delhi Technological University
Delhi, India
anamika@dce.ac.in

***Abstract*—Industrialization, urbanization, and human activity have combined to lead to air pollution, which is a major health concern for many nations around the world. When it comes to air pollution, PM2.5, defined as particles having a diameter of less than 2.5 mm, poses a major health danger. It causes a variety of symptoms, including respiratory and cardiovascular problems. The ability to predict air pollution PM2.5 levels is therefore vital for preventing the harmful effects of air pollution. This research is being carried out for two reasons. This first objective is to identify potential sources of air pollution. As well as incorporating meteorological factors, transportation considerations shall be taken into account. Lastly, we intend to select the best model to predict air pollutant concentrations. To estimate hourly air pollution concentrations, we use machine learning and deep learning models on the acquired data.**

***Index Terms*—air pollution, machine learning, deep learning**

I. INTRODUCTION

One of the most important predictors of human health is air pollution. Industrialization, urbanization, and human activity have combined to lead to air pollution, which is a major health concern for many nations around the world. According to the World Health Organization, air pollution puts 7 million people's health at risk. It is a key risk factor for a wide range of health conditions, including asthma, heart difficulties, throat and eye disorders, bronchitis, lung cancer, and respiratory system ailments. Aside from the health risks associated with air pollution, it also poses a major danger to our planet. When it comes to air pollution, PM2.5, defined as particles having a diameter of less than 2.5 mm, poses a major health danger. It causes a variety of symptoms, including respiratory and cardiovascular problems. The ability to predict air pollution PM2.5 levels is therefore vital for preventing the harmful effects of air pollution. Pollution emissions from sources such as automobiles and industry are the root cause of the greenhouse effect, and CO2 emissions are among the most significant contributions to the phenomena. Climate change has been widely debated at global conferences and has been a hot problem for the world for the previous two decades due to growing pollution and ozone degradation. Statistical linear approaches have been used in the past to handle the problem of air pollution prediction, however these techniques are poor estimators for air pollution prediction owing to the complexity and variance in time-series data. Many machine learning strategies for dealing with complicated ways have been created during the previous 60 years.Figure 1 showcases the major sources of Air Pollution in India. [2]

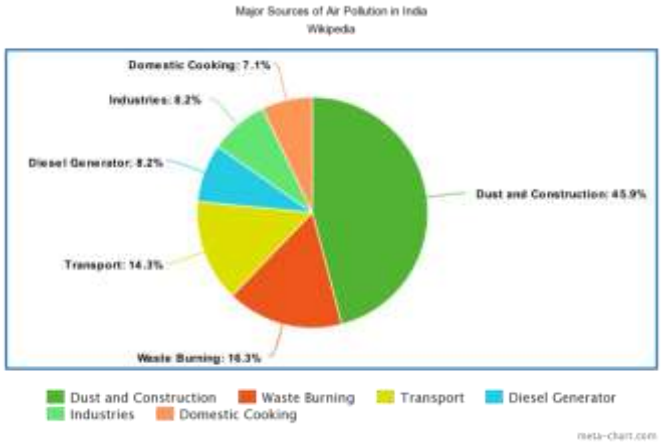

Fig. 1. Major Sources of Pollution in India

AQI is calculated using 12 parameters according to CPCB , including NO2 (Nitrogen Dioxide), NH3 (Ammonia), SO2 (Sulfur Dioxide), As (Arsenic), CO (Carbon Monoxide), O3 (Ozone), PM10, PM2.5, O3 (Ozone), Ni (Nickel) and Pb (Lead), Most of the time, the AQI is computed using the criteria pollutants (i.e. O3, CO, SO2, NO2, PM2.5 and PM10), although it is recommended to calculate the AQI using many pollutants from the list of 12 pollutants. The choice of contaminants, however, is determined by the AQI objectives,

averaging period, data availability, monitoring frequency, and measuring techniques.The health problems grow as the AQI climbs. The AQI's goal is to enhance public awareness.

Various pollutants are assigned a single number , name, and colour. The AQI is classified as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. The atmospheric concentration levels of air pollutants and their possible health consequences determine each of these categories.Figure 2 showcases the AQI index.

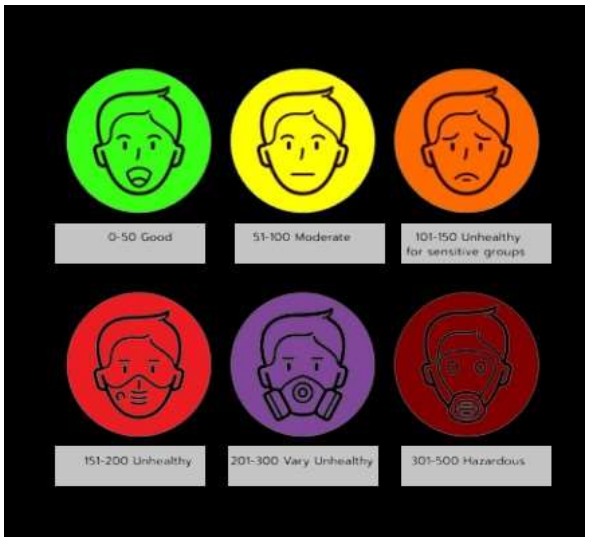

Fig. 2. AQI Index

This paper's contributions are summarized as follows:

- taking into account the latest air quality data available of India
- analyzing and comparing 3 different models to predict air quality

The rest of this work is organised as follows: The second section discusses current related studies on estimating the Air Quality Index. Section III explains the Data set in brief. Section IV describes the techniques used in this Study. Section V examines the results and Section VI concludes the research.

II. Literature Review

A variety of machine learning algorithms for predicting pollution levels have been presented in recent years. In this part, we showcase and discuss some of the most important work in the subject. The exisisting papers are summarised in Table I.

III. Dataset

We used Kaggle's "Air Quality Data in India" for this investigation (2015-2020). It is an Air Quality Index (AQI) computed from hourly and daily data collected from stations across India. This dataset consists of five files: city data, city hour data, station data, and additional data , out of which we have considered city day and city hour data only . According to the website, "Ahmedabad, Hyderabad, Amritsar, Mumbai, Chen- nai, Patna, Gurugram, Delhi, Bengaluru, Lucknow, Jaipur, Thiruvananthapuram, and other destinations are among those covered." We will focus on the "City day" and "City hour" datasets for our research, which include 16 characteristics such as City, Date, PM2.5, PM10, NO, NO2, NH3, CO, CO2, SO, O3, Benzene, Toluene, Xylene, AQI, and AQI Bucket. The total number of items in the "City day" and "City hour" datasets is roughly 29530 and 707880, respectively.

IV. Methodology

The goal of this part is to discuss the air quality dataset we utilised, how we cleaned and preprocessed the data, and what models we used to get the highest level of accuracy feasible.

A. Data Preprocessing

Like all other sensor data, pollutant data cannot be totally free of missing data and abnormal values.Hence, the data needs to be cleaned and missing values must be filled.The basic flow of data preprocessing is displayed in Figure 3

Fig. 3. Flowchart of Data - Preprocessing

B. Prediction Model

In the next part, we will compare several machine learning and deep learning models.

1) Linear Regression: A linear graph is used in linear regression analysis to estimate the value of a variable. The dependent variable is the one that must be predicted. The independent variable is the variable that you use to predict the value of another variable. As a consequence, this regression method establishes a linear relationship between x (input) and y. (output). Figure 4 shows the general architecture of Linear Regression.

TABLE I
EXISTING LITERATURE SURVEY

Reference	Paper	Year	Dataset Used	Parameters Used	Method Used
[2]	Error Prediction of Air Quality at Monitoring Stations Using Random Forest in a Total Error Framework	2020	Nine AQ monitoring stations measuring the concentration of NO2 in Oslo	NO2	Random Forest
[4]	Prediction of Air Quality in Major Cities of China by Deep Learning	2020	1,615 observation stations reported hourly AQI data. covering China from 2015 to 2019	PM 2.5 PM 10 CO2 CO O3 SO2 NO2	BPNN CNN GRU LSTM BiLSTM
[5]	An Adaptive Kalman Filtering Approach to Sensing and Predicting Air Quality Index Values	2020	AQI concentrations from January 1, 2014 to December 30, 2018 in Nanjing	PM 2.5 PM 10 CO2 CO O3 SO2 NO2	Kalman Filtering Model
[6]	Sensor-Based Air Pollution Prediction Using Deep CNN-LSTM	2020	Data gathered across the Port of LA from 4 major sites for 4 years	O3, CO, SO2, NO2, PM2.5, and PM10.	1D-CNN-LSTM
[7]	CNN based Variation and Prediction Analysis of 2m Air Temperature for Different Zones of the Indian Region	2021	Temperature Readings of Ahemdabad , Comibatore , Udaipur and Balasore	O3 CO SO2 NO2 PM2.5 PM10	ARIMA CNN LSTM
[8]	Air pollutant severity prediction using Bi-directional LSTM Network	2018	numerous sites in New Delhi for forecasting up to 6 hours, 12 hours, and 24 hours in advance	O3 CO SO2 NO2 PM2.5 PM10	Bi-LSTM
[9]	Based on the n-Step Recurrent Prediction, a Sequence-to-Sequence Air Quality Predictor	2020	Beijing AQI from April 2017 to March 2018	CO NO2 O3 PM 2.5 PM 10	CNN-LSTM ANN SVM GRU
[10]	Air Pollution Hotspot Identification and Pollution Level Prediction in the City of Delhi	2020	Bitspi Air Pollution API	Humidity Air Pressure NO2 O3 SO2 CO PM 2.5 PM 10	SVM LSTM
[11]	Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things	2020	China's smog data for 74 city sites from 2014/1/1 to 2018/1/1	PM 2.5 PM 10	LSTM GRU
[12]	Temperature Prediction Using the Missing Data Refinement Model Based on a Long Short-Term Memory Neural Network	2020	Weather data released by the Meteorological Office of South Korea form 1981 - 2016	wind speed, wind direction, and humidity	LSTM

Random Forest: Random Forest Regression is a supervised learning technique that use the ensemble learning method for regression. The ensemble learning approach is a methodology that integrates predictions from numerous machine learning algorithms to generate a more accurate forecast than a single model.

A Random Forest Regression model is both strong and accurate. It generally works admirably on a wide range of problems, even those with non-linear connections. However, there is no interpretability, overfitting is possible, and we must pick the amount of trees to include in the model.

2) Convolutional Neural Network: In general, convolutional neural networks are feed-forward neural networks that process data with a grid-like topology to analyze visual images. ConvNets are another name for CNN. In convolutional neural networks, objects are detected and classified in images. Every image in CNN is represented as an array of pixel values. A convolutional neural network is made up of pixel values convolutioned. Compared to other classification algorithms, ConvNet requires much less pre-processing. Despite primitive methods that require hand-engineering of the filters, ConvNets can learn these filters/characteristics with enough

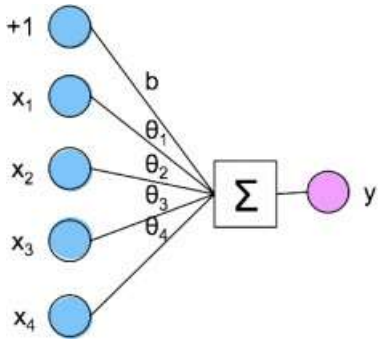

Fig. 4. The general architecture of Linear Regression

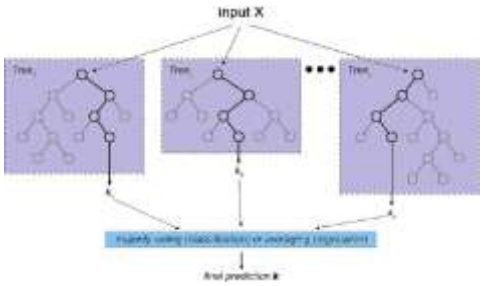

Fig. 5. The general architecture of Random Forest

training.Figure 6 shows the general architecture of ConvNet .

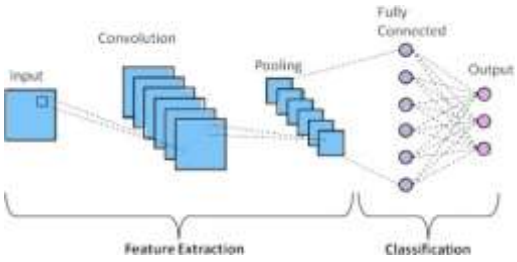

Fig. 6. The basic architecture of ConvNet

V. Results

Our proposed model's final prediction results were tested and analyzed utilizing three assessment methods: MSE (Mean Square Error) and RMSE (Root Mean Square Error). To evaluate performance, the following standard metrics are employed. The Table II and III compares the error rates of the three models we implemented.

TABLE II
PERFORMANCE ON CITY DAY

Dataset	**City Day**	
Model	***MSE***	***RMSE***
Linear Regression	1047	32.35
Random Forest	936	30.59
Convolutional Neural Network	1145	33.83

TABLE III
PERFORMANCE ON CITY HOUR

Dataset	**City Hour**	
Model	***MSE***	***RMSE***
Linear Regression	4050	63.63
Random Forest	2445	49.44
Convolutional Neural Network	1834	42.82

VI. Conclusion

To develop a solution to the air pollution prediction problem, we analysed and compared three existing air pollution prediction techniques. Linear Regression , Random Forest Regression and Convolutional Neural Network were the approaches used.Error rates have been calculated.It also implies that the lower the RMSE number, the higher the value.We found that Random Forest algorithm produced best result for city day data, giving MSE of 936. And Convolutional Neural Network algorithm produced best result for city hour data with MSE of 1834, after reviewing each of the three approaches mentioned. Based on preliminary findings, the suggested model is useful for visualising air quality.

Future study may broaden the field of investigation. We only investigated particulate matter and not environmental variables due to a limited dataset. Dealing with missing values and anticipated abnormalities in the contaminated dataset might help increase prediction precision. In continuation of this review, one can consider improving accuracy by applying Lasso Regression, RNN, Decision Tree and Multibinomial Regression.

REFERENCES

[1] "Air pollution: India," Indpaedia. http://indpaedia.com/ind/index.php/Airpollution:India (accessed Apr. 30, 2022).

[2] Lepioufle, Jean-Marie, Leif Marsteen, and Mona Johnsrud. "Error Prediction of Air Quality at Monitoring Stations Using Random Forest in a Total Error Framework." Sensors 21, no. 6 (2021): 2160. https://doi.org/10.3390/s21062160

[3]P. Chhikara, R. Tekchandani, N. Kumar, M. Guizani, and M. M. Hassan, "Federated Learning and Autonomous UAVs for Hazardous Zone Detection and AQI Prediction in IoT Environment," IEEE Internet of Things Journal, vol. 8, no. 20, pp. 15456–15467, Oct. 2021, doi: 10.1109/jiot.2021.3074523.

[4]C. Zhan, S. Li, J. Li, Y. Guo, Q. Wen, and W. Wen, "Prediction of air quality in major cities of China by deep learning," Nov. 2020. Accessed: Apr. 30, 2022. [Online]. Available: http://dx.doi.org/10.1109/cis52066.2020.00023

[5]J. Chen, K. Chen, C. Ding, G. Wang, Q. Liu, and X. Liu, "An Adaptive Kalman Filtering Approach to Sensing and Predicting Air Quality Index Values," IEEE Access, vol. 8, pp. 4265–4272, 2020, doi: 10.1109/access.2019.2963416.

[6]K. Nagrecha et al., "Sensor-Based Air Pollution Prediction Using Deep CNN-LSTM," Dec. 2020. Accessed: Apr. 30, 2022. [Online]. Available: http://dx.doi.org/10.1109/csci51800.2020.00127

[7]S. Patel, J. Patel, and U. Tyagi, "CNN based Variation and Prediction Analysis of 2m Air Temperature for Different Zones of the Indian Region," Apr. 2021. Accessed: Apr. 30, 2022. [Online]. Available: http://dx.doi.org/10.1109/iccmc51019.2021.9418316

[8]I. Verma, R. Ahuja, H. Meisheri, and L. Dey, "Air Pollutant Severity Prediction Using Bi-Directional LSTM Network," Dec. 2018. Accessed: Apr. 30, 2022. [Online]. Available: http://dx.doi.org/10.1109/wi.2018.00-19

[9]B. Liu et al., "A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction," IEEE Access, vol. 7, pp. 43331–43345, 2019, doi: 10.1109/access.2019.2908081.

[10]S. Sur, R. Ghosal, and R. Mondal, "Air Pollution Hotspot Identification and Pollution Level Prediction in the City of Delhi," Sep. 2020. Accessed: Apr. 30, 2022. [Online]. Available: http://dx.doi.org/10.1109/icce50343.2020.9290698

[11]B. Wang, W. Kong, H. Guan, and N. N. Xiong, "Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things," IEEE Access, vol. 7, pp. 69524–69534, 2019, doi: 10.1109/access.2019.2917277.

[12][2]Park, Kim, Lee, Kim, Song, and Kim, "Temperature Prediction Using the Missing Data Refinement Model Based on a Long Short-Term Memory Neural Network," Atmosphere, vol. 10, no. 11, p. 718, Nov. 2019, doi: 10.3390/atmos10110718.

[13]S. Lee, Y.-S. Lee, and Y. Son, "Forecasting Daily Temperatures with Different Time Interval Data Using Deep Neural Networks," Applied Sciences, vol. 10, no. 5, p. 1609, Feb. 2020, doi: 10.3390/app10051609.

[14]Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner, "Data-Driven Air Quality Characterization for Urban Environments: A Case Study," IEEE Access, vol. 6, pp. 77996–78006, 2018, doi: 10.1109/access.2018.2884647.

[15]K. K. Rani Samal, K. Sathya Babu, A. Acharya, and S. K. Das, "Long Term Forecasting of Ambient Air Quality Using Deep Learning Approach," Dec. 2020. Accessed: Apr. 30, 2022. [Online]. Available: http://dx.doi.org/10.1109/indicon49873.2020.9342529

[16]E. Sharma, R. C. Deo, R. Prasad, A. V. Parisi, and N. Raj, "Deep Air Quality Forecasts: Suspended Particulate Matter Modeling With Convolutional Neural and Long Short-Term Memory Networks," IEEE Access, vol. 8, pp. 209503–209516, 2020, doi: 10.1109/access.2020.3039002.

[17]Y. Rybarczyk and R. Zalakeviciute, Regression Models to Predict Air Pollution from Affordable Data Collections. 09 2018

[18]U.Ramani, R.Nithya, S.Sathieshkumar, T.Santhoshkumar . "Automatic Weather Monitoring Analysis for Renewable Energy System" .Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020).

[19]D. Birant, "Comparison of decision tree algorithms for predicting potential air pollutant emissions with data mining models," Journal of Environmental Informatics, vol. 17, pp. 46–53, 03 2011

[20] Rijal N, Gutta RT, Cao T, Lin J, Bo Q, Zhang J. Ensemble of deep neural networks for estimating particulate matter from images. In: 2018 IEEE 3rd international conference on image, vision and computing (ICIVC). IEEE; 2018, p. 733–8.

[21] Zhang L, Li D, Guo Q. Deep learning from spatio-temporal data using orthogonal regularizaion residual cnn for air prediction. IEEE Access. 2020;8:66037–47.

[22] Li J, Jin M, Li H. Exploring spatial influence of remotely sensed pm2.5 concentration using a developed deep convo-lutional neural network model. Int J Environ Res Public Health. 2019;16(3):454.

[23] SamirLemes, Air quality index (AQI) – comparative study and assessment of an appropriate model for Bamp;H, in Conference: 12th Scientific/Research Symposium with International Participation "Metallic and Non-metallic Materials" MNM 2018, Vlas˘ic´, Bosnia and Herzegovina, 2018.

[24] Harsh N. Shah, Zishan Khan, Abbas Ali Merchant, MoinMoghal, AamirShaikh, PritiRane, IOT based air pollution monitoring system, Int. J. Sci. Eng. Res. 9 (2) (2018), ISSN 2229-5518.

[25]S. Muthukumar, W. Sherine Mary, et al., IoT based air pollution monitoring and control system, in Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018) IEEE Xplore Compliant Part Number: CFP18N67-ART, ISBN: 978-1-5386-2456-2.

[26]David Brooks, An Arduino-Based System for Measuring Airborne Particle Concentrations, Institute for Earth Science Research and Education, HTTPS:// www.instesre.org/PM/PMMeasurements.htm.

[27]K. Yamunathangam, P VarunaPritheka, IoT enabled air pollution monitoring and awareness creation system, Int. J. Recent Technol. Eng. (IJRTE) 7 (4S) (2018)

[28][12]Somayya Madakam, R. Ramaswamy, "Internet of Things", Journal of Computer and Communications, 2015, 3, 164-173

[29] Anjaiah Guthi, "Implementation of an Efficient Noise and Air Pollution Monitoring System Using Internet of Things (IoT)", Vol. 5, Issue 7, July 2016 ISO 3297:2007 Certified.

[30] Palghat Yaswanth Sai "An IOT Based Automated Noise and Air Pollution Monitoring System", Vol. 6, Issue 3, March 2017

[31] Uppugunduru Anil Kumar, G Keerthi et-al "IOT BASED NOISE AND AIR POLLUTION MONITORING SYSTEM USING RASPBERRY PI", Vol. 5, Issue 3, March 2017

[32]W. Mao, W. Wang, L. Jiao, S. Zhao, and A. Liu, "Modeling air quality prediction using a deep learning approach: Method optimization and evaluation," Sustainable Cities and Society, vol. 65, p. 102567, Feb. 2021, doi: 10.1016/j.scs.2020.102567.

[33]Nandanwar, Himanshu, and Anamika Chauhan. "IOT based Smart Environment Monitoring Systems: A Key To Smart and Clean Urban Living Spaces." In 2021 Asian Conference on Innovation in Technology (ASIANCON), pp. 1-9. IEEE, 2021

[34]H. Nandanwar, A. Chauhan, D. Pahl and H. Meena, "A Survey of Application of ML and Data Mining Techniques for Smart Irrigation System," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 205-212, doi: 10.1109/ICIRCA48905.2020.9183088.

[35]H. Meena, H. Nandanwar, D. Pahl and A. Chauhan, "IoT based perceptive monitoring and controlling an automated irrigation system," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225455.

[36] Agrawal, Sachin Kumar and Kapil Sharma. "5G millimeter wave (mmWave) communications." 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (2016): 3630-3634.

[37] Kapil Sharma, Sandeep Tayal, "Indian Smart City Ranking Model " International Journal of Recent Technology and Engineering, Volume 8, Year 2019, Pages 4820-4832 DOI:10.35940/ijrte.B2472.078219