4th International Conference on Communication and Information Processing (ICCIP-2022)
Available on: SSRN
(SSRN is an open-access online preprint community, owned by Elsevier)

Deep Learning for Air Quality Prediction after Covid-19 Pandemic based on Pollutant and Meteorological Data

Naushad Ahmad[a], Vipin Kumar[b,*]

[a,b]*Department of Computer Science and Information Technology, Mahatma Gandhi Central University, Motihari, Bihar-845401, India*

Abstract

The second wave of COIVD19 has jolted the environment and economy worldwide. Air pollution is one of the primary causes of these pandemics. Therefore, it is important to analyze the air quality index during the COVID19 pandemic. In literature, machine learning (ML) and deep learning (DL) methods have been deployed to predict PM2.5 to forecast air pollution. The central pollution control board (CPCB) of India has gathered information on pollutants such as Particulate Matter(PM) with a diameter of 2.5 microns, called PM2.5, Particulate Matter with a diameter of 10 microns or less, called PM10, Nitrogen dioxide, Sulfur dioxide, Ozone (O3), Carbon Monoxide, Temperature, Relative Humidity (R.H.), Wind Speed (W.S.), Wind Direction(W.D.), and Solar Radiation (S.R.) CPCB has recorded that Bhiwadi, Rajasthan is the world's most polluted city in 2021 $PM_{2.5}$ Ranking IQAir. In this research, the air quality of Bhiwadi has been analyzed during the COVID19 pandemic based on the above features. The features are considered from three perspectives 1) pollutant features, 2) meteorological features, and 3) Overall features. The analysis has been performed in two-phase, i.e., 1) meteorological + Pollutant and 2) meteorological+PM2.5. ExtraTreesRegressor of ML and LSTM of DL algorithms have achieved the best among other algorithms over the overall dataset (meteorological+pollutant) based on root mean squared error (RMSE) performance measures.

Keywords- Deep learning, Air quality index, Machine learning, PM2.5, COVID19

1. Introduction

The world's most challenging issue is going to be air pollution. The amount of air pollution is significantly rising every day. A single individual or nation cannot stop such global problems. It is accountable to all people on the earth. Particulate matter (PM), carbon monoxide, ozone (O3), sulphur dioxide, and nitrogen dioxide are some of the primary pollutants (Sulfur Dioxide). Several issues, including heart attacks, coughing fits, and breathing difficulties, are brought on by PM smaller than or equal to 10 micrometres in diameter. A more harmful effect is caused by carbon monoxide. It is considered a "Silent Killer".

Breathing difficulties, unconsciousness, and headaches are the most common symptoms. Primary and secondary air pollutants comprise most of the subcategories of air pollutants (Simu et al., 2015). Animals, plants, the environment, wildlife, and human health can all be negatively impacted by air pollution in both the

* Corresponding author. Tel.: +91-931-348-5512.
E-mail address: naushad13bhu@gmail.com[a]; rt.vipink@gmail.com[b].

short and long term. When it comes to effects on human health, many organs and the body's general performance might be harmed. Short-term consequences include headaches, dermatoses (skin diseases), upper respiratory infections, suffocation, and throat discomfort. Brain damage, lung cancer, concurrent Multiple Organ Dysfunction Syndrome (MODS), liver damage, respiratory illness, heart disease, and kidney damage are long-term impacts on human health(Kang et al. 2018)(Zeinalnezhad et al. 2020). The incomplete burning of coal and wood, automobile emissions, and gaseous fuels are the primary causes of carbon monoxide production. The ozone hole and ozone layer loss are reasonably well known to us. The ozone's adversaries have existed in the atmosphere since the beginning. Ozone depletion is mainly brought on by CFCs combined with ozone molecules. Cancer and negative health impacts are brought on by ozone depletion. Asthma in children is a significant source of PM and O3(Lewis et al. 2005).

Oxides of nitrogen, often known as nitrogen oxides, are very reactive gases. One of the subgroups of the same kind is NO2. The primary gaseous contaminant is NO2. Mobile sources, such as emissions from cars, trains, planes, and other moving engines that move from one location to another, are the primary source of NO2. Alternatively, put, the burning of fossil fuels. Due to the burning of rice straw, many pollutants (Repairable Suspended Particulate Matter, SO2, and NOX) are produced quickly (Chawala and Sandhu 2020). More plantations and greenery can potentially remove contaminants from the surrounding environment, which is how plants naturally absorb NO2 through their stomata.

A subset of artificial intelligence (A.I.) is called machine learning. Computer algorithms that operate in a machine learning environment without being explicitly coded. Probability theory, statistics, and applied mathematics are the sources of data-driven machine learning algorithms. They are commonly employed in computer science and mathematics (Guan and Sinnott 2018). Regarding computer science and information technology, researchers are focusing on environmental challenges. It is quite close to environmental pollution's primary sources, such as water and land contamination and air pollution. Environmental science relies heavily on it. The forecast and evolution of ecological air quality have become a hot issue of growing interest to study academics domestically and internationally based on environmental monitoring (Qin, Cen, and Guo, 2019). Researchers are quite excited to work on machine learning to estimate the degree of air pollution.

About 17 lakh (1.7 million) persons died in India in 2019 due to air pollution, which is close to 18% of all deaths. When it reviewed the information and discovered that the monthly estimates of deaths due to particulate matter (PM) in Uttar Pradesh, Maharashtra, and Bihar were all exceedingly high. There are 13432, 5223, and 4470 fatalities overall per month in the same three states. India ranked third in the world for pollution at the national level(Naqvi et al. 2021), averaging 72 ug/m3. Bangladesh and Pakistan are on top of the world air quality index list.

In this paper, the authors propose a multivariate LSTM analysis based on the Adam optimizer to predict air pollution and machine learning algorithm analysis separately. Time- series prediction model has three different types of data set used. In the time-series data set, one and most common attribute is DateTime. It downloaded hourly data from CPCB with 16 attributes, but some feature values are missing, like CO2 and VWS. All three data sets have DateTime, and PM2.5 are common. In the first data set, all pollutant and meteorological features are there. In the second data set, only meteorological elements are considered. The *first phase* of the results analyzes the ten-regression model with adjusted R- squared, R-squared, and RMSE. After analysis of the results, researchers conclude that one of the regression models is the best of all ten. In the *second phase* of the results, the authors analyze the proposed analysis LSTM model with MAPE, MSE, and RMSE performance parameters.

- *Authors have considered the most recent world's most polluted city in 2021 PM2.5 (IQAir ranking 2022).*
- *A comparative study of ML and DL has been performed to identify the best performance of the learning model.*
- *Researchers have divided the dataset in a particular manner into pollutant and meteorological with PM2.5 concentrations.*

The paper's organization is as follows: Sect. 2 literature review with the table has explained the ML and DL model with cities collection of data and year of the published papers. Sect. Three materials and methods have four sub-sections study area and dataset, framework description, performance evaluation parameters, and hardware and software. Sect. 4 results and analysis have two sub-section description of results and analysis of results. At the last sect. Conclusions of results.

Table 1. Air quality index categories defined by EPA

AQI VALUE	HEALTH MESSAGE	AQI COLOR
0 – 0	No Data Available	BLACK
1–50	None, Best for Health	GREEN
51 – 100	Susceptible people should reduce heavy exertion	YELLOW
101 – 150	Sensitive groups should reduce heavy exertion	ORANGE
151 – 200	Only sensitive groups should avoid prolonged	RED
201 – 300	Only sensitive groups should avoid physical activity	PURPLE
301 – 500	Everyone avoids all outdoors physical activity	MAROON

2. Literature Review

In this paper (Alyousifi et al., 2020), the fuzzy time-series model is preferable when forecasting air pollution. However, it has a drawback brought on by the usage of a haphazard division of the realm of speech. This work suggests a brand-new MWFTS model based on the optimal partition approach. The proposed model's performance is assessed using the daily API data using three statistical metrics: MSE, MAPE and Theil's U statistic. To manage air pollution, the suggested model may thus be a superior alternative for forecasting air quality. In another paper (Gocheva-Ilieva et al., 2019), The robust data mining method of regression trees and classification is used to propose a standard method for creating high-quality environmental time series nonlinear models (CART). The best models are chosen using goodness-of-fit metrics like the Coefficient of determination and root-mean-square error. The findings demonstrate that CART models accurately predict roughly 90% of observed values of PM10 and match the data well. The authors wrote in(Siwek et al., 2009), The article outlines a unique strategy for precise forecasting of the PM10 concentration of the daily average. It is based on the wavelet processing of the time series indicating PM10 pollution and the use of neural networks. The use of the ensemble of predictors, integrated utilizing the blind source separation method or neural-based integration, is the major innovation of the suggested strategy.

Regarding MAPE RMSE, MAE, and errors, the numerical experiments used to estimate the daily concentration of PM10 pollution in Warsaw have demonstrated high overall prediction accuracy. In this research(Kong et al., 2021), The authors developed an ensemble method-based real-time prediction strategy for multivariate time-series data. It has tested the given model using a simulated data set to real-time forecast failures based on real-time performance log data from the server systems and air quality collected by five sensors. Traditional approaches to abnormality detection have also strongly emphasized the status of objects as either normal or abnormal depending on available data. The authors have discovered that the suggested strategy for predicting air pollution performed well and consistently for both short-term and long-term

predictions. This paper (Bui, Le, and Cha 2018) uses recurrent neural networks and long short-term memory units as a framework to optimize knowledge from time-series data of air quality and meteorological information. Finally, it looks into different setups' forecast accuracies. This study serves as a solid impetus for both to continue studying urban air quality and to assist the government in using that knowledge to implement helpful regulations(Bui, Le, and Cha 2018). Table 2 describes some of the ML and DL models of the recent year of publications. Also, describe the area of study and place where the pollutant and meteorological data were downloaded and collected by sensors. The authors describe the various machine learning and deep learning model in the field of air pollution prediction. Both machine and deep understanding have trended to predict and forecast the air quality of various cities.

Table 2. A deep comparison of various research papers with different ML models to predict air quality

S. No.	Ref. I.D.	Purpose and the Area of Study	Machine Learning / Deep Learning Model	Parameters/Data Sources	Region	Year
1	(Bhat, Manek, and Mishra, 2019)	Prediction System based for Detecting Air Pollution using Machine Learning	Linear Regression, Decision Tree Regression and R.F. Regression	Using Arduino Static Sensor Circuit (OGD) Platform India	Karnataka, India	2019
2	(Soundari, Jeslin, and Akshaya, 2019)	To Prediction and Analysis of the Indian Air Pollution using ML	Linear Regression	CPCB	Various places in India	2019
3	(Adke et al., 2019)	To use ML to Predict the Air Pollution	Multilayer Perceptron and Linear Regression	(CPCB)	Pune, India	2019
4	(Kumar, Mishra, and Singh, 2020)	To use the Machine Learning model to Predict Air Pollution	Multi-Linear Regression, Support Vector Machine and Random Forest	(CPCB)	Ghaziabad, U.P.	2019
5	(Rybarczyk and Zalakeviciute 2018)	To Predict Air quality from the Affordable Data Collections with Regression Models	Multiple Regression, Cumulative Modeling Method	EPA, U.S.	Specific zone of Quito	2018
6	(Zhu et al., 2018)	To Model Regularization and Optimization for Air Quality Prediction using Machine Learning Approach	Multi-Task Learning (MTL)	EPA, U.S.	Lewis University-Lemont Village (LU-LV) and (LMA-AV)	2018
7	(Tao et al., 2019)	Deep learning model to forecasting air pollution	Bidirectional GRU and 1D convnets	UCI ML repository	Beijing, China	2019
8	(Tripathi and Pathak 2021)	Techniques of DL for Air pollution	CNN-LSTM	CPCB	Various city	2021
9	(Gao et al., 2020)	A case study and analysis of air pollution control policies	LSTM	China National Environmental Monitoring Center	Chengdu-Chongqing, China	2020
10	(Han et al., 2020)	To air pollution forecast using Bayesian DL	Bayesian deep-learning	KDD cup for fresh air website	London, UK	2020

3. Materials and Methods

In Fig.1., twelve rows and twelve columns indicate all the same features on both axes. The diagonal (light color) is shown as $PM_{2.5}$ with $PM_{2.5}$, i.e., 1. The color of this cell is almost light; it means that all the data values of $PM_{2.5}$ columns are correlated with $PM_{2.5}$ rows values similarly. Apart from that, less than one and a negative value tells us other features are less correlated to $PM_{2.5}$. The least negative value of -0.72 shows that temp and wind speed are less correlated in data values, and the color of -0.72 is almost dark (black).

3.1. Study Area and Dataset

The authors have taken the most polluted city in the world according to IQAir report 2021. Bhiwadi, India top most polluted city ranking based on annual average $PM_{2.5}$ concentration with historical data (2017-2021)("World's Most Polluted Cities (Historical Data 2017-2021)" n.d.). IQAir is a Switzerland (swiss) air quality technology company. The data has been collected from the CPCB("CENTRAL CONTROL ROOM FOR AIR QUALITY MANAGEMENT - DELHI NCR" n.d.), India, under the Ministry of Environment, Forest and Climate Change. The duration of the data set is from January 1, 2021, to June 19, 2022, and the average period is 1 hour. The location of the data set is a single station RIICO Ind. Area III, Bhiwadi - RSPCB, Rajasthan. The total number of instances is 12833, each with sixteen attributes like From Date, To Date, $PM_{2.5}$, PM_{10}, NO, NO_2, SO_2, CO, CO_2, Ozone, Temp, R.H., W.S., W.D., S.R., and VWS.

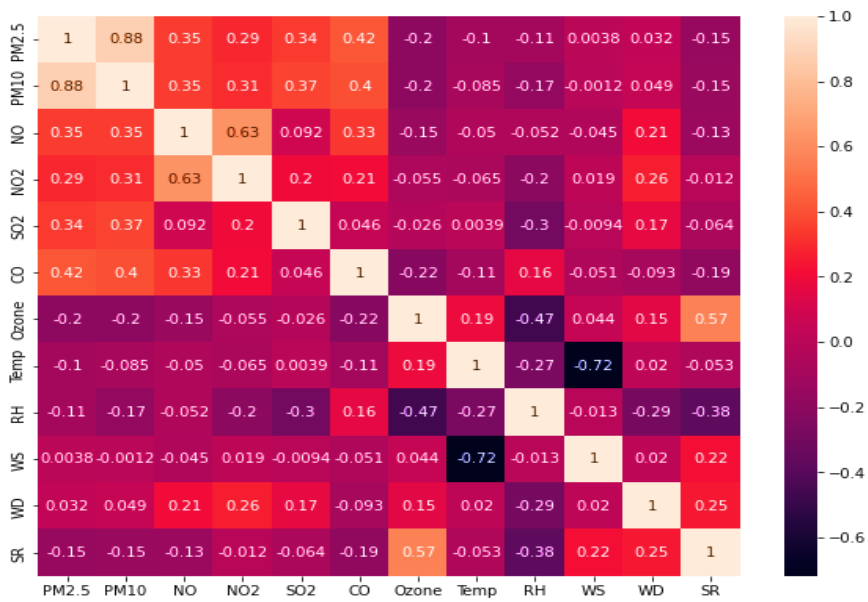

Fig. 1. Correlation heat map of interrelated feature

- **Pollutant feature:** PM_{10}, $PM_{2.5}$, NO, NO_2, SO_2, CO and Ozone. Each parameter has $\mu g/m^3$ or Micrograms per Cubic Meter of Air
- **Meteorological feature:** Other meteorological data used Temp, R.H., W.S., W.D., and S.R.
- **Predictive data feature:** Authors are enthusiastic about predicting the PM2.5 concentrations.

- Through overall exploratory data analysis, two features VWS and CO_2 data, are missing a lot. Researchers remove both features, and some other features are missing value none, and the mean of the features replaces Nan or NaN (Not a Number).

Table three complete describes the pattern of the data set of the twelve features with index count, mean, standard deviation, min of the feature, 25% percentile, 50% percentile, 75% percentile, and maximum value of the corresponding feature. Index of the count, the total number of instances of the single feature looks like 12833 for all. The mean index tells us the average of the feature, and the mean value of all the features is different. Index min & max are minimum and maximum values of the single instance of the data set of the corresponding features.

Table 3. The statistical description of the dataset

index	$PM_{2.5}$	PM_{10}	NO	NO_2	SO_2	CO	Ozone	Temp	RH	WS	WD	SR
count	12833	12833	12833	12833	12833	12833	12833	12833	12833	12833	12833	12833
mean	115.0475	235.1854	35.26587	51.3326	29.86724	0.929373	25.95185	24.37852	51.72864	4.644441	194.2326	125.4076
std	72.0679	133.0301	37.36619	33.69521	28.1683	0.513274	24.32432	18.1841	22.89393	13.03859	81.86997	169.5478
min	0.04	0.1	0.08	0.05	0.02	0	0.03	-50	6.21	0.14	0.14	3.33
25%	62.53	140.18	11.98	27.2	11.92	0.63	7.42	20.38	33.34	0.57	130.69	6.65
50%	105.45	219.83	22.57	43.24	21.08	0.85	17.52	28.53	51.72	0.87	194.23	15.78
75%	149.69	298.93	41.98	63.42	36.05	1.11	36.82	33.55	69.1	1.46	272.59	215.85
max	670.92	976.7	427.95	472.86	199.76	8.58	184.38	48.43	98.97	50.98	358.13	769.74

3.2. Description of Framework

Long-short-term memory is a Recurrent Neural Network used in deep learning. LSTM model is used for specific time series forecasting problems. Many types of LSTM models are used in time series forecasting and prediction. Univariate and multivariate LSTM has one common attribute: date and time.

Step 1: Collect the data from the CPCB Ministry of Environment, Forest and Climate Change with the pollutant and meteorological data.

Step 2: Preprocessing the whole dataset

- **Step 2.1:** The collected data unluckily inserted many missing values. Remove the Nan or NaN (Not a Number) and None value from the dataset and fill the cells with the mean of the feature.
- **Step 2.2:** Label encoding for more than one feature in the range of 0 to n-1, where n is the total number of features. In this data, the full features are twelve available. Then the encoding range is 0 to 11.
- **Step 2.3:** MinMaxScaler to the range of 0 to 1; if the features have a negative value, then the range will be -1 to 1. MinMaxScaler can convert the highest value of the feature into one and the lowest value of the feature into 0.

Step 3: Splitting the data set with 70-15-15. Seventy percent for training data, 15 per cent for validation, and 15 per cent for testing the data.

Step 4: Applying the prediction LSTM model to the dataset

- **Step 4.1:** Adam optimizer (Bock, Goppold, and Weiß 2018) for each parameter W_i

$$S_t = \alpha_1 \times S_{t-1} - (1 - \alpha_1) \times G_t \quad (1)$$

whereas S_t : Exponential average of gradients by the side of W_i

$$V_t = \alpha_2 \times V_{t-1} - (1 - \alpha_2) \times G_1^2 \quad (2)$$

whereas V_t: EV of squares of the gradients by the side of W_i

$$\Delta W_t = -\eta \times S_t / \sqrt{(V_t + \varepsilon)} \times G_t \quad (3)$$

$$W_{t+1} = W_t + \Delta W_t \quad (4)$$

where η : Initial learning rate, G_t: Gradient at time t by the side of W_i, α_1 α_2 : Hyper parameters

- **Step 4.2:** Tuning with epoch/batch size. Taking epoch size in list containing [50,100,150,200,250].

Step 5: Evaluate the model's performance with RMSE, MSE and MAPE are three evaluation parameters. In each epoch, the researcher collects the results of each epoch of all three evaluation parameters in an excel sheet.

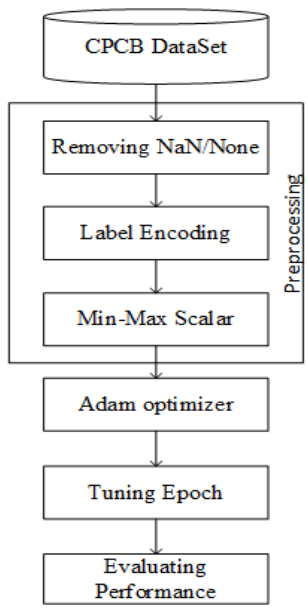

Fig. 2. Framework for deep learning regression analysis to get the AQI.

3.3. Performance evaluation parameters

The researcher evaluated the ML and DL prediction model using MSE, RMSE, MAPE, adjusted- $R^{2,}$ and R^2 (Coefficient of determent). Mean Absolute Error tells the arrays of the actual and the predicted values of air pollution parameter concentrations. In other words, mean absolute error is the average magnitude of absolute differences between N predicted vectors $S = \{X_1, X_2, X_3, \ldots, X_n\}$ and $S = \{Y_1, Y_2, Y_3, \ldots, Y_n\}$
Mean Absolute Error (MAE) (Nath et al. 2021) is calculated as shown in eq.5:

$$\frac{1}{N}\sum_{i=1}^{N} |\text{Yi} - \text{Xi}| \quad (5)$$

The range of RMSE and MAE is from 0 to infinite. The lower RMSE is, the better the model fits your data set. Hence Root Mean Square Error is the computer taking the standard deviation of the prediction error values on the air pollution dataset (see eq.6).

$$RMSE = \sqrt{MSE} \tag{6}$$

Root Mean Square Error(Arnaudo, Farasin, and Rossi 2020) or Root Square Deviation is calculated as shown in eq.7:

$$\sqrt{\sum_{i=1}^{n} \frac{(P_i - O_i)^2}{n}} \tag{7}$$

where P_i is the prediction value for the i^{th} observation in air pollution data set and O_i is the observed value for the i^{th} observation in air pollution data set along with n is the sample size. The higher the $R2$ or Coefficient of determination, the better the model fits your data set. The $r - range$ squared from 0 to 1 or from 0% to 100%. If the $r - squared$ values are more significant, the 0.75 is substantial or good.

3.4. Hardware and software required

This section lists and briefly describes the hardware and software needed to carry out the preprocessing and analysis procedures in the section above. Using a scientific Python 3.6 environment, all preliminaries—including data gathering, aggregation, and cleaning—were completed with the aid of the pandas(McKinney 2010) and NumPy(Oliphant 2007) libraries, as well as the stats models package(Seabold and Perktold 2010) for time series analysis and matplotlib for data visualization. It used the well-known Scikit-learn toolkit(Barupal and Fiehn 2019), which provides broad functionality in many machine learning fields, from data preparation to metrics computation, for the data imputation phase and the subsequent trials. It used the Jupyter Notebook IDE (Integrated Development Environment**),** Keras, and TensorFlow, a sizable library with a wide range of capabilities, as a deep learning alternative. The LSTM model's implementation and training specifically made use of the latter. All the experiments, analyses and procedures briefly described were performed on a Windows workstation with 64 G.B. of RAM, Intel® Xeon® Silver 4210 CPU @2.20Ghz 2.1 Hz, and another windows laptop has 6 G.B. of RAM core i3, 5th generation.

4. Results and analysis

4.1. Description of results

In Fig 4 x-axis represents machine learning models, and the y-axis represents the RMSE value. The Fig 5 x-axis denotes the machine learning models, and the y-axis denotes a range of the Coefficient of determination. In Fig 6 x-axis represents machine learning models, the y-axis represents the RMSE value, and in fig 7 x-axis denotes the machine learning models. The y-axis indicates a range of the Coefficient of determination. In fig 8 x-axis represent machine learning models, the y-axis represents the RMSE value, and in Fig 9 x-axis denotes the machine learning models. The y-axis indicates a range of the Coefficient of determination. In fig 10 x-axis represents the different types of data set, the y-axis denotes the RMSE value, in Fig 11 x-axis denotes the different types of data set, and the y-axis denotes the range of adjusted r-squared and r-squared. In Fig 12 x-axis symbolizes the epoch size of all features dataset, and the y-axis indicates the MAPE value; in Fig 13 x-axis represents the epoch size, and the y-axis denotes the error range of MSE and RMSE. Fig 14 x-axis denotes the epoch size of all features dataset, the y-axis represents the MAPE value, Fig 15 x-axis denotes the epoch size, and the y-axis indicates the error range of MSE and RMSE. In Fig 16 x-axis denotes the epoch size of all features dataset, and the y-axis represents the MAPE value; in Fig 17 x-axis denotes the epoch size,

and the y-axis indicates the error range of MSE and RMSE. Fig 18 denotes all features data set results of deep learning and machine learning RMSE. Fig 19 represents the only meteorological data with $PM_{2.5}$ results of deep learning and machine learning RMSE. Fig 20 denotes the only pollutant data set results of deep learning and machine learning RMSE. Form Fig 4 to Fig 9, the ML model description are as follows: 1- ExtraTreesRegressior, 2- LGBMRegressor, 3- RandomForestRegressor, 4- HistGradientBoostingRegressor, 5- XGBRegressor, 6- BaggingRegreesor, 7- GradientBoostingRegressor, 8- MLPRegressor, 9- TransforemdTargetRegressor and 10- LinearRegression

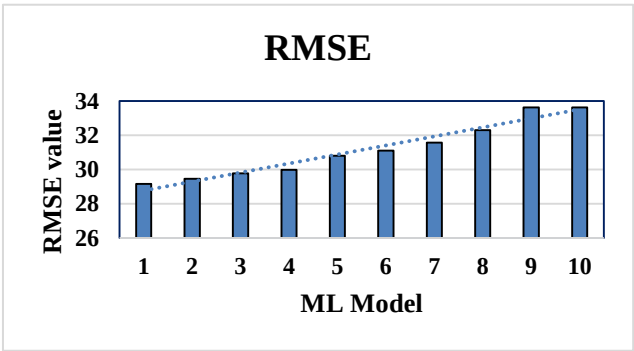

Fig: 4 Predicting PM2.5 using all features

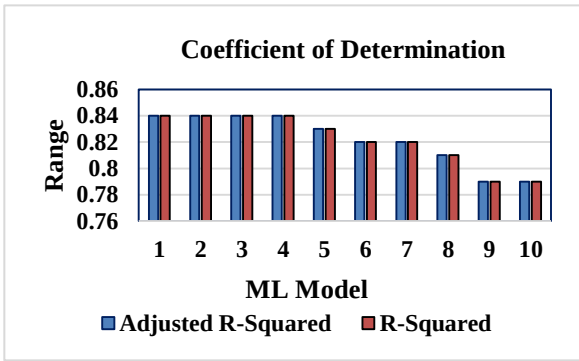

Fig: 5 Predicting PM2.5 using all features

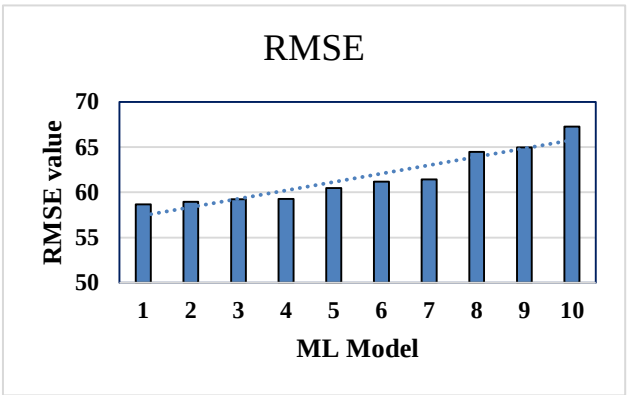

Fig: 6 Predicting PM2.5 using meteorological data

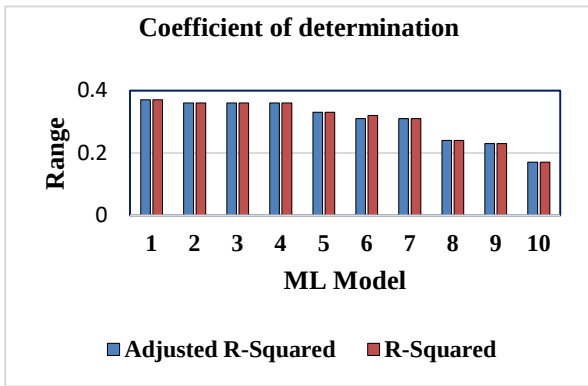

Fig:7 Predicting PM2.5 using meteorological

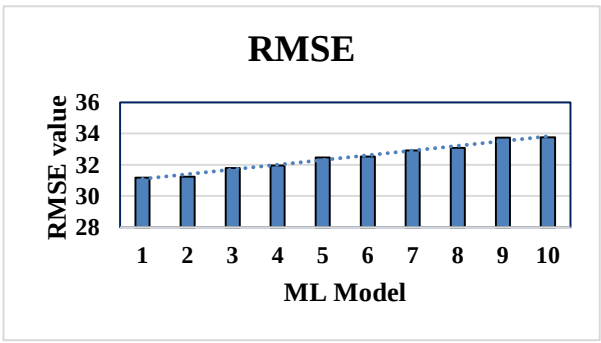

Fig: 8 Predicting PM2.5 using only pollutant

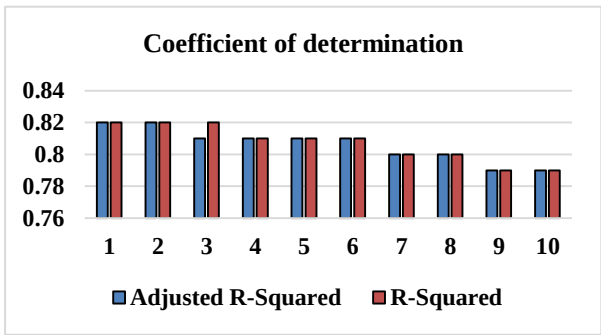

Fig: 9 Predicting PM2.5 using only pollutant

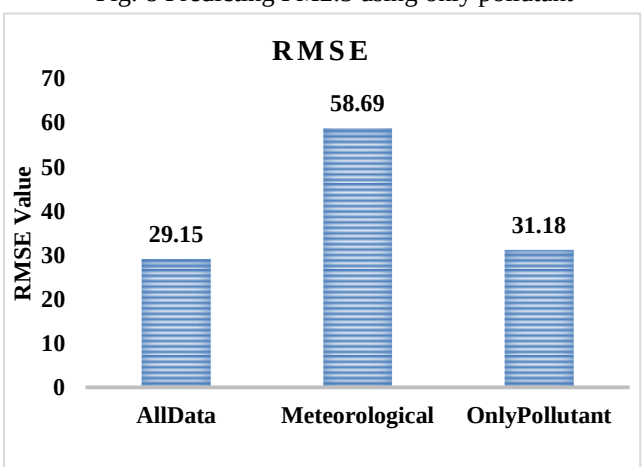

Fig: 10 Compression of ExtraTreesRegressor

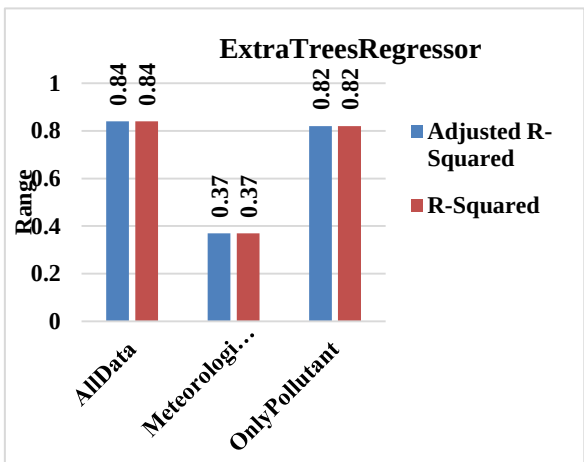

Fig: 11 Compression of ExtraTreesRegressor

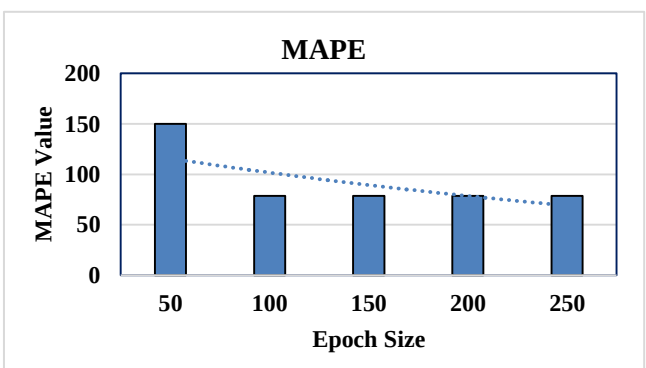

Fig: 12 Predicting air pollution using all features

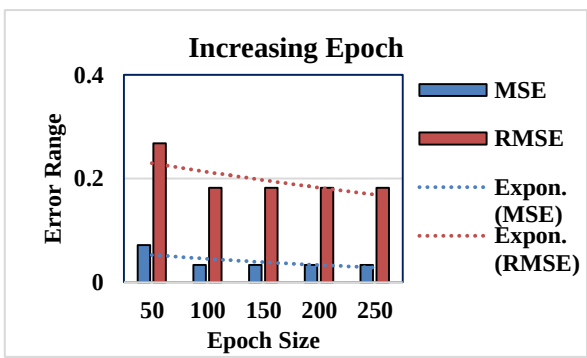

Fig:13 Predicting air pollution using all features

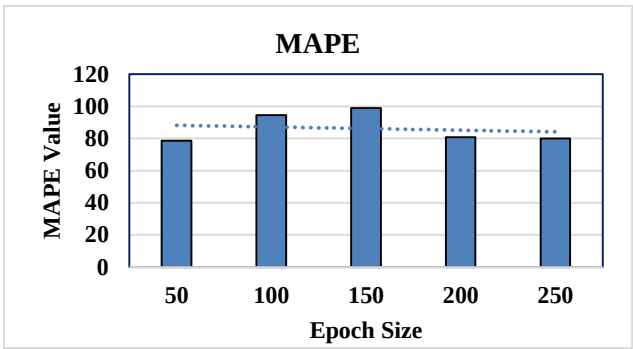

Fig: 14 Predicting air pollution meteorological

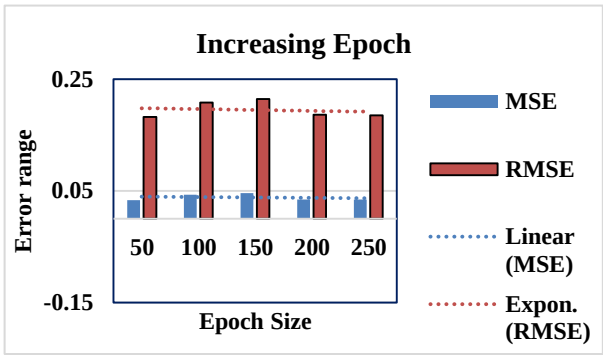

Fig:15 Predicting air pollution meteorological

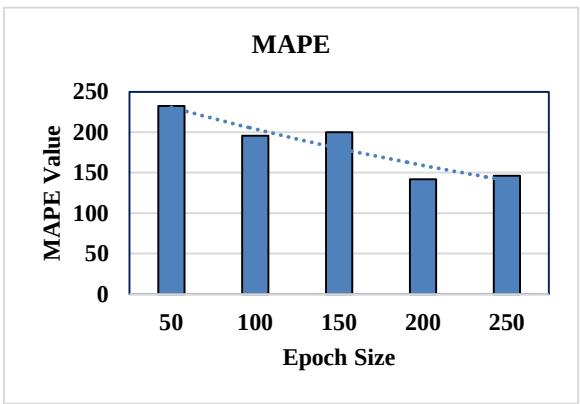

Fig: 16 Predicting air pollution using pollutant

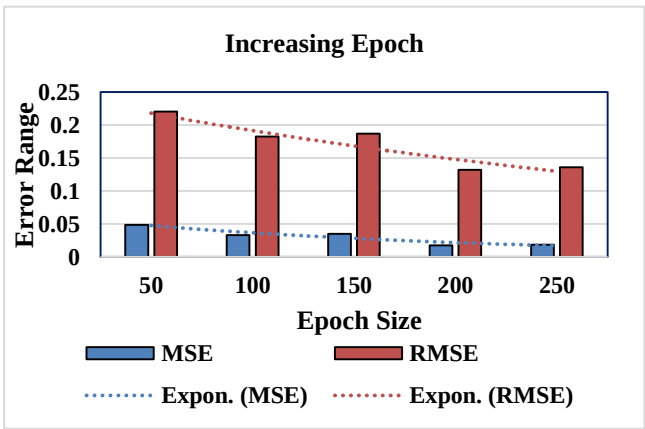

Fig:17 Predicting air pollution using pollutant

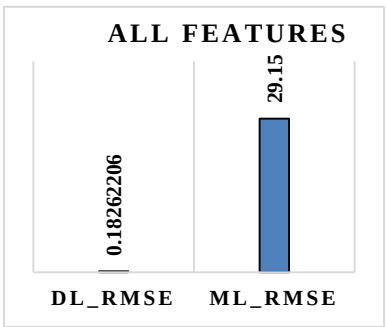

Fig: 18 All features

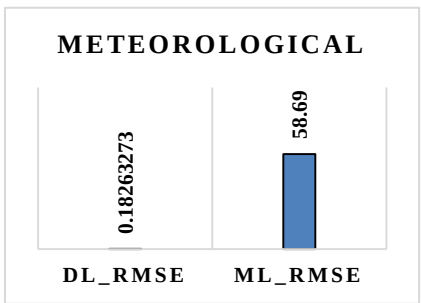

Fig: 19 Meteorological

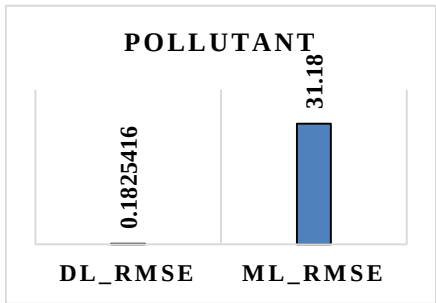

Fig: 20 Pollutant

4.2. Analysis of results

- *Performance analysis of ML algorithm*

The authors have divided the analysis of the data set into three parts; the first part contains all the dataset holes called (D_1). The second part contains only meteorological reached (D_2), and the last chunk has only pollutant features (D_3). The first phase of the analysis involved Fig 4 to Fig 11. Fig 4 and Fig 5 results of the D_1 data set. In Fig 4, The less RMSE error range from 29 to 30; four ML models fall into that slot ExtraTreesRegressor, LGBMRegressor, RandomForestRegressor, and HistGradientBoostingRegressor. But the least error among them all is ExtraTreesRegressor. In Fig 5, adjusted r-squared and r-squared values are the same, but the best deal of the Coefficient of determination is 0.84 or 84% which is as substantial. Fig 6 and Fig 7 results of the D_2 data set. In Fig 6, The lower RMSE error range from 58 to 60; four ML models fall into that slot ExtraTreesRegressor, LGBMRegressor, RandomForestRegressor, and HistGradientBoostingRegressor. But the least error among them all is ExtraTreesRegressor. In Fig 7, adjusted r-squared and r-squared values have the same fluctuation, but the best value of the Coefficient of determination is 0.37 or 37% which is a week. Fig 8 and Fig 9 results of the D_3 data set. In Fig 8, The less RMSE error range from 31 to 31.5; two ML models fall into that slot ExtraTreesRegressor and RandomForestRegressor. But the least error has ExtraTreesRegressor. In Fig 9, adjusted r-squared and r-squared values have the same fluctuation, but the best value of the Coefficient of determination is 0.82 or 82%, that is a week. In conclusion, Fig 10 and Fig 11 it is showing clearly our D_1 dataset has less error and a better coefficient of determination.

- *Performance analysis of DL algorithm:*

The second phase of the analysis involved Fig 12 to Fig 17. Fig 12 and Fig 13 results of the D_1 data set. In Fig 12, The less MAPE error got the 100-epoch size. The trending exponential line has less slope. In Fig 13, after 100 epochs, RMSE and MSE trending exponential lines are almost flat. Fig 14 and Fig 15 results of the D_2 data set. In Fig 14, The less MAPE error got after 150 epoch size. The trending exponential line has no slope. In Fig 15, after 50 epochs, RMSE and MSE trending exponential lines are fully almost flat. Fig 16 and Fig 17 results of the D_3 data set. In Fig 16, The less MAPE error got after 200 epoch size. The trending exponential line has declined. In Fig 17, at the 250 epoch, RMSE and MSE errors are almost significantly less, and trending exponential lines decline at 250 epoch size. The authors concluded the MAPE value in all three data set is D_1 but less RMSE value in the D_3 dataset.

- *Comparative study of ML and DL algorithm:*

The last phase of the analysis involved Fig 18 to Fig 20. In Fig 18, all features or D_1 dataset DL RMSE value is 0.18262206 and ML RMSE value is 29.15, DL RMSE value is very less to ML RMSE value. In Fig 19, meteorological features or D_2 dataset DL RMSE value is 0.18263273 and ML RMSE value is 58.69, DL RMSE value is very less to ML RMSE value. In Fig 20, pollutant features or D_3 dataset DL RMSE value is 0.1825416 and ML RMSE value is 31.18, DL RMSE value is very less to ML RMSE value.

- Conclusion of overall analysis:

The best performance of the model for the D_1 data set to RMSE value for the ExtraTreesRegressor for machine learning. The best model for the D_3 data set to RMSE value for the LSTM prediction for deep learning. We finally conclude two data set has good work in machine learning and deep learning. Data sets have all features good for machine learning ExtraTreesRegressor model for RMSE and only pollutant data for deep learning LSTM model. The deep learning LSTM model has fewer errors in predicting air quality.

5. Conclusions

The proposed research has focused on analysing the most pollutant place (Bhiwadi) of India, where the COVID19 pandemic period has been considered. The data has been collected from the CPCB with pollutants and meteorological features. The dataset has been analyzed from three perspectives pollutant data, meteorological, and a combination of both. The subsets of the dataset have been studied and learned from ML and DL algorithms to predict the air quality index. The ExtraTreesRegressor for machine learning has performed the best among ML algorithms, and the LSTM model performs the best in deep learning models. In conclusion, the DL model has better performance than ML algorithms.

Future Research: More cities in India may be utilized to see a more effective prediction pattern corresponding to the pollutant and meteorological features. The pollutant feature set and meteorological elements set may consider as two views of the data, which may deploy the Multi-view learning to enhance the prediction of the air quality index.

References

Adke, Rohit, Suyog Bachhav, Akash Bambale, and Bhushan Wawre. 2019. "Air Pollution Prediction Using Machine Learning," 332–34.

Alyousifi, Yousif, Mahmod Othman, Ibrahima Faye, Rajalingam Sokkalingam, and Petronio C. L. Silva. 2020. "Markov Weighted Fuzzy Time-Series Model Based on an Optimum Partition Method for Forecasting Air Pollution." *International Journal of Fuzzy Systems* 22 (5): 1468–86. https://doi.org/10.1007/s40815-020-00841-w.

Arnaudo, Edoardo, Alessandro Farasin, and Claudio Rossi. 2020. "A Comparative Analysis for Air Quality Estimation from Traffic and Meteorological Data." *Applied Sciences* 10 (13): 4587. https://doi.org/10.3390/app10134587.

Barupal, Dinesh Kumar, and Oliver Fiehn. 2019. "Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach." *Environmental Health Perspectives* 127 (9): 097008. https://doi.org/10.1289/EHP4713.

Bhat, Ajit, Asha S Manek, and Pranay Mishra. 2019. "Machine Learning Based Prediction System for Detecting Air Pollution" 8 (09): 155–59.

Bock, Sebastian, Josef Goppold, and Martin Weiß. 2018. "An Improvement of the Convergence Proof of the ADAM-Optimizer," April, 1–5. http://arxiv.org/abs/1804.10587.

Bui, Tien-Cuong, Van-Duc Le, and Sang-Kyun Cha. 2018. "A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM," April. https://doi.org/https://doi.org/10.48550/arXiv.1804.07891.

"Central Control Room for Air Quality Management - Delhi Ncr." n.d. Accessed June 19, 2022. https://app.cpcbccr.com/ccr/#/caaqm-dashboard/caaqm-landing/caaqm-comparison-data.

Chawala, Pratika, and H A S Sandhu. 2020. "Stubble Burn Area Estimation And Its Impact on Ambient Air Quality Of Patiala and Ludhiana District, Punjab, India." *Heliyon* 6 (July 2019): e03095. https://doi.org/10.1016/j.heliyon.2019.e03095.

Gao, Hao, Weixin Yang, Jiawei Wang, and Xiaoyun Zheng. 2020. "Analysis of the Effectiveness of Air Pollution Control Policies Based on Historical Evaluation and Deep Learning Forecast: A Case Study of Chengdu-Chongqing Region in China." *Sustainability* 13 (1): 206. https://doi.org/10.3390/su13010206.

Gocheva-Ilieva, Snezhana Georgieva, Desislava Stoyanova Voynikova, Maya Plamenova Stoimenova, Atanas Valev Ivanov, and Iliycho Petkov Iliev. 2019. "Regression Trees Modeling of Time Series for Air Pollution Analysis and Forecasting." *Neural Computing and Applications* 31 (12): 9023–39. https://doi.org/10.1007/s00521-019-04432-1.

Guan, Ziyue, and Richard O. Sinnott. 2018. "Prediction of Air Pollution Through Machine Learning Approaches on the Cloud," 51–60. https://doi.org/10.1109/BDCAT.2018.00015.

Han, Yang, Jacqueline C.K. Lam, Victor OK Li, and Qi Zhang. 2020. "A Domain-Specific Bayesian Deep-Learning Approach for Air Pollution Forecast." *IEEE Transactions on Big Data* 7790 (c): 1–1. https://doi.org/10.1109/TBDATA.2020.3005368.

Kang, Gaganjot Kaur, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. 2018. "Air Quality Prediction: Big Data

And Machine Learning Approaches," no. August 2020. https://doi.org/10.18178/ijesd.2018.9.1.1066.

Kong, Taewoon, Dongguen Choi, Geonseok Lee, and Kichun Lee. 2021. "Air Pollution Prediction Using an Ensemble of Dynamic Transfer Models for Multivariate Time Series." *Sustainability* 13 (3): 1367. https://doi.org/10.3390/su13031367.

Kumar, Saurabh, Shweta Mishra, and Sunil Kumar Singh. 2020. "Heliyon A Machine Learning-Based Model to Estimate PM2 .5 Concentration Levels in Delhi' s Atmosphere." *Heliyon* 6 (November): e05618. https://doi.org/10.1016/j.heliyon.2020.e05618.

Lewis, Toby C, Thomas G Robins, J Timothy Dvonch, Gerald J Keeler, Fuyuen Y Yip, Graciela B Mentz, Xihong Lin, et al. 2005. "Air Pollution-Associated Changes in Lung Funtion Among Asthmatic Children In Detroit," no. 8: 1068–75. https://doi.org/10.1289/ehp.7533.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 1:56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Naqvi, Hasan Raja, Manali Datta, Guneet Mutreja, Masood Ahsan Siddiqui, Daraksha Fatima Naqvi, and Afsar Raza Naqvi. 2021. "Improved Air Quality and Associated Mortalitites in India under Covid-19 Lockdown." *Environmental Pollution* 268 (2): 115691. https://doi.org/10.1016/j.envpol.2020.115691.

Nath, Pritthijit, Pratik Saha, Asif Iqbal Middya, and Sarbani Roy. 2021. "Long-Term Time-Series Pollution Forecast Using Statistical and Deep Learning Methods." *Neural Computing and Applications* 33 (19): 12551–70. https://doi.org/10.1007/s00521-021-05901-2.

Oliphant, Travis E. 2007. "Python for Scientific Computing." *Computing in Science & Engineering* 9 (3): 10–20. https://doi.org/10.1109/MCSE.2007.58.

Qin, Zepeng, Chen Cen, and Xu Guo. 2019. "Prediction of Air Quality Based on Knn-Lstm." https://doi.org/10.1088/1742-6596/1237/4/042030.

Rybarczyk, Yves, and Rasa Zalakeviciute. 2018. "Regression Models to Predict Air Pollution from Affordable Data Collections." *Machine Learning - Advanced Techniques and Emerging Applications*. https://doi.org/10.5772/intechopen.71848.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference*, 92–96. https://doi.org/10.25080/Majora-92bf1922-011.

Simu, Shreyas, Varsha Turkar, Rohit Martires, Vranda Asolkar, Swizel Monteiro, Vaylon Fernandes, and Vassant Salgaoncar. 2015. "Air Pollution Prediction using Machine Learning," 231–36. https://doi.org/10.1109/IBSSC51096.2020.9332184.

Siwek, Krzysztof, Stanislaw Osowski, Konrad Garanty, and Mieczyslaw Sowinski. 2009. "Ensemble of Predictors for Forecasting the PM10 Pollution." *15th International Symposium on Theoretical Electrical Engineering, ISTET 2009*, 318–22.

Soundari, A. Gnana, J. Gnana Jeslin, and A. C. Akshaya. 2019. "Indian Air Quality Prediction And Analysis Using Machine Learning" 14 (11): 181–86.

Tao, Qing, Fang Liu, Yong Li, and Denis Sidorov. 2019. "Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU." *IEEE Access* P.P. (c): 1. https://doi.org/10.1109/ACCESS.2019.2921578.

Tripathi, Kshitij, and Pooja Pathak. 2021. "Deep Learning Techniques for Air Pollution." In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 1013–20. IEEE. https://doi.org/10.1109/ICCCIS51004.2021.9397130.

"World's Most Polluted Cities (Historical Data 2017-2021)." n.d. Accessed April 12, 2022. https://www.iqair.com/in-en/world-most-polluted-cities.

Zeinalnezhad, Masoomeh, Abdoulmohammad Gholamzadeh Chofreh, Feybi Ariani Goni, and Jiří Jaromír Klemeš. 2020. "Air Pollution Prediction using Semi-Experimental Regression Model and Adaptive Neuro-Fuzzy Inference System." *Journal of Cleaner Production* 261. https://doi.org/10.1016/j.jclepro.2020.121218.

Zhu, Dixian, Changjie Cai, Tianbao Yang, and Xun Zhou. 2018. "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization." *Big Data and Cognitive Computing* 2 (1): 1–15. https://doi.org/10.3390/bdcc2010005.