Air Pollution Forecasting using Convolution based Long Short Term Memory Techniques

Neha Verma, Dr. Kamlesh Namdev, Dr. Kalpana Rai
M. Tech. Scholar, Professor, Head of Dept.
Department of Computer Science and Engineering
Sagar Institute of Research Technology Excellence, Bhopal
singhneha2330@gmail.com

Abstract:- During the past few years, severe air-pollution problem has garnered worldwide attention due to its effect on health and well-being of individuals. As a result, the analysis and prediction of air pollution has attracted a good deal of interest among researchers. The research areas include traditional machine learning, neural networks and deep learning. How to effectively and accurately predict air pollution becomes an important issue. In this thesis, we propose an convolution based Long Short Term Memory (CBLSTM) based on the LSTM deep learning method. In this new model, we combine local air quality monitoring station, the station in nearby industrial areas, and the stations for external pollution sources. To improve prediction accuracy, we aggregate CBLSTM models into a predictive model for early predictions based on external sources of pollution and information from nearby industrial air quality stations.

Keywords:- CBLSTM, Air Pollution, Air Quality

I. INTRODUCTION

The environmental surroundings are everything around you - the air, the land, and the rivers and oceans. The atmosphere is composed of several gases, water vapor, and dust particles. The important gases are nitrogen, oxygen-argon, carbon dioxide, neon, helium, hydrogen, ozone, etc. By volume, nitrogen is 78% and oxygen by about 21%. Together, these two gases make 99 % volume of the atmosphere, the rest 1% also important for us. Nitrogen is used by the plants for their survival. But plants cannot take nitrogen directly from the air. Bacteria that live in the soil take nitrogen from the air and change its form so that plants can use it. Air is extremely important for several cycles and makes life possible on earth (https://sciencing.com/importance-air6330367.html).

Air-Water Cycle: - Air is important for different states of water like ice and water vapor. This water cycle assures the required amount of water for every life.

Air-Carbon Cycle: - Air plays a main role in recycling Carbon Dioxide (CO2) released in the air by breathing and fossil decomposition. The plant produces energy and releases oxygen from CO2 by the process of photosynthesis. Humans and animals eat plants for the energy required for living. Decomposition of the body after their lives results in CO2 emission back into the atmosphere.

Air-Temperature Cycle: - The average earth temperature will be down to freezing point without air.

Air-Life Safety Cycle:- Earth atmosphere will save us from harmful radiation and maintain the temperature conducive for living. Air reduces the risk of space rocks reaching earth as it is vaporized in the air.

Air-Sound Cycle:- If air does not exist we could not hear a screaming jet engine nearer to your ear. Sound can be heard as the air carries sound waves from one point to other.

II. AIR POLLUTION

Currently, air pollution is recognized as a significant public pathological condition, responsible for an increasing range of health impacts which are well documented from the outcomes of extensive studies in several areas of the world. While there's little doubt that fast urbanization means we tend to are currently exposed to unhealthy concentrations and additional numerous ambient air pollutants. X-ray radiation imaging studies on the bodies of ancient mummies have detected proof of respiratory illness, emphysema, respiratory organ lump and arteriosclerosis (Zweifel et al., 2009 ; Thompson et al., 2013), while autopsies have represented in depth carbon deposits within the respiratory organ (Zimmerman et al., 1971). This, in turn, has led to a speculative link to the daily inhalation of smoke in confined areas from fuels used for heat, cooking, and lighting (Kelly and Fussel, 2015). Air pollution is the release of natural or human made harmful gases and particles into the environment (Kemp et al., 2011). Air pollution has a higher impact on society and threatens humanity's ability to survive. There was a substantial rise in the use of coal in factories and homes during the urbanization and industrial revolution. These outcomes in smog which caused dismalness and mortality in the stale climatic conditions. During the

1952 Great London smog, the tragedy of losing 4000 life in heavy pollutions highlights the connection between air pollution and human health. In this way air pollution is a developing issue in the urban locales around the globe (Ferreira et al., 2003). Air pollution is a composite of either natural or human activity gasses or particles released into the atmosphere. The number of particles released will be more harmful than the tolerable amount (Kemp et al., 2011). There are two types of pollutant sources. Natural Sources: Natural pollutants are harmful substances which are emitted by natural phenomenon. SO2, CO2, NO2, CO, and Sulphate are few natural pollutants discharged due to eruptions of volcanoes and forest wildfires.

Man created (Anthropogenic) sources: Man created sources, for example, fuel burning, releases from mechanical generation forms, transportation discharges are the fundamental wellsprings of air contamination. Many pollutants are transferred by man-made sources, including hydrogen, oxygen, nitrogen, sulphur, metal composites and particulate matter. With the growing complete population and the developing economy of the scene, the planet's interest in vitality has grown dramatically. The enormous use of fossil vitality universally has also led to a rise in ecological problems that have been taken into account because of their imminent impact on human well-being and the earth (Kemp et al., 2011; Song Y et al., 2015). Air pollution is a key problem in many areas of the globe, with two critical issues: the impact on human well-being, for instance, cardiovascular diseases, and the impact on the earth, for instance, corrosive rain, environmental change, and global temperature change. (Wang J et al., 2016). The pathways for transporting and transforming matter within four categorical fields that create planet Earth (biosphere, lithosphere, hydrosphere, and atmosphere) square measure known as the Bio-geo-chemical cycle (Niharika and Rao, 2014). These cycles regulate the planet's functioning. The planet is the sun's nonparticulate receptive radiation. Thus, from the moment of its birth, the matter surrounding the planet is renovated and geographically distributed. With the rapid changes in these cycles, the planet's atmosphere is affected to blame for the existence of life on Earth that affects life on Earth reciprocally. The Department of Science and Technology employs several regulatory bodies to predict the state of the atmosphere for different places. Air quality modelling and monitoring system play a major role in all elements of pollution management and air quality, wherever forecast can play a major role. Air quality predictions provide the general public with air quality data informing them of preventive measures to avoid or limit their exposure to unhealthy levels of pollution.

Implications of Air Pollution to the Quality of Living: - Air pollution is a substantial threat to health worldwide (Burnett R, 2018; Cohen AJ et al., 2017; Forbes, 2016). As indicated by 2015 Global Burden of Disease (Balakrishnan et al., 2019) exposure to external pollution is that the world's fifth major death risk problem, accounting although air pollution is a universal problem, it is probable to cause the biggest damage in sensitive people exposed to harmful pollutants. People with chronic diseases (especially cardiorespiratory diseases), very little social support, and lack of medical facilities are most at risk from pollution. In 2015, 4.2 million fatalities and losses of 103.1 million life-years adjusted for disability (Zhu S et al., 2017). The related studies by Zhang S et al. (2016) and Fougère B et al., (2018) have shown that the related to air pollution for chronic obstructive pulmonary disease and lung cancer varies with age, and these results are biologically possible. The most worrying sign is that the incidence of chronic obstructive pulmonary disease and lung cancer is likely to be higher in older populations (aged >55 years) than younger populations (aged also liable for the depletion of the ozone layer that causes Ultra Violet rays to penetrate the Earth and acid rain that has adverse effects on trees and wildlife (Gaganjot Kaur Kang et al., 2018). Hence, regulation of air quality and its forecasting has become an important task for both developed and developing nations. As a result of increasing man-made developments, growth of pollutants concentration into the atmosphere has become inevitable and thus, depreciating air quality. Air pollutants are classified into two categories – primary pollutants and secondary pollutants. Pollutants which are generated through the process, for instance, ash from a volcanic eruption are referred to as primary pollutants. Examples – Carbon Monoxide (CO), Sulphur Dioxide (SO2). Secondary pollutants are a result of the direct or indirect reaction of primary pollutants. A prominent example of secondary pollutant includes ground level Ozone (O3). Among the six criteria pollutants, Particulate Matter 2.5 (PM2.5) is considered as one of the most pernicious (Pandey et al., 2013). With only 2.5 microns in diameter in size and light advocating them, these tiny particles tend to stay for an extended period of time in the atmosphere and cause harmful effects inside filters of the nose and throat. Growing lung cancer mortality and 4-8 percent increase in the threat of cardiopulmonary is directly associated with the upgradation of each 10- µg/m3 long-term average PM2.5 (Pope III et al., 2002). A major source of Nitrogen Dioxide (NO2) formation is emissions of power plants and automobiles which in turn leads to the formation of ground-level O3 and fine particle pollution. The mixture of O3 and NO2 is considered as a major threat to children and people suffering from lung diseases like chronic bronchitis, asthma, emphysema (Jiang et al., 2016). Prolonged exposure to a certain concentration level of O3 can also result in detrimental effects on plants, crop yield, flora, and fauna. Natural and anthropogenic emission sources of CO include forest fires, animal metabolism, IC engines and burning of carbon enriched fuels. It directly leads to the condition of subnormal oxygenation of the arterial blood and augmentation of greenhouse gases.

III. AIR QUALITY INDEX

An AQI is a measured metric used to record and report evenly on the air quality of various constituents in terms of human health (Feng and Yang, 2012). AQI is a daily air quality reporting index. It informs you how well we are breathing clean air. The AQI is being used as a standard for the quality of air. This AQI gives information to the people about the environment in which they are living in (Mamta and Basin, 2010). This gives the necessary information about the pollution caused due to the emissions from the industries and can act as a feedback factor so as to modify the emissions from those industries and also helps in allocation of funds to the air pollution control boards. This AQI helps us to rank the different locations in order of the pollution they have, thus highlighting the areas which are more polluted and the frequency of potential hazards. AQI helps to determine the changes in air quality that have happened over a given period of time, allowing for the prediction of air quality and control measures for pollution. This AQI has been calculated by significant levels of air pollutants. The calculation of AQI differs from nation to nation, based on the significant pollutants involved. India AQI, as per Central Pollution Control Board (CPCB) notified (http://www.cpcb.nic.in) AQI is constituted by SO2, O3, CO, PM (included PM10 and PM2.5), and NO2, Lead (Pb), Ammonia (NH3), pollutants.

Table 1: AQI Category

AQI Category	PM_{10} 24 Hr μg/m^3	PM_{25} 24 Hr μg/m^3	NO_2 24 Hr μg/m^3	O_3 24 Hr μg/m^3	CO 24 Hr μg/m^3	SO_2 24 Hr μg/m^3
Good (0-50)	0-50	0-30	0-40	0-50	0-0.1	0-40
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2	41-80
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380
Poor (210-300)	251-350	91-120	181-280	169-208	10-17	381-800
Very Poor (310-400)	351-430	121-250	281-400	209-748	17-34	801-1600
Severe (401-500)	430+	250+	400+	748+	34+	1600+

Challenges and Limitations

1. The pollutants types and levels will vary from one location to another. The sources of pollutants are also different like Natural and manmade.

2. Each pollutant will have an adverse effect on human beings.

3. The monitoring stations will give the measurements of various pollutants. More commonly the data available are 24 Hrs average.

4. All the pollutants have to consider for calculating AQI for a particular location. Most of the pollutants will vary with time. Handling a huge amount of data will be a tough task.

5. More research has been done on predicting the individual forecasting of pollutants but not on the AQI.

6. The information is supplied straight without scrutiny from the analyzers for real time AQI, so it may not be for a statutory purpose.

7. Monitoring and subsequent AQI dissemination involves various steps including the operation of sensors and analysers, their calibration, local server data acquisition, transmission via the Internet to a central database, etc. Due to multiple technical and operational elements such as lengthy power cuts and maintenance issues, monitoring station functioning may also be impacted. Given these constraints, some interruption in the continuous flow and dissemination of information may occur. However, in the event of breakdowns, immediate action is taken to restore the system to operation within a reasonable period of time.

Need for Air Quality Forecasting

Because of restricted resources and practical execution, an alternative approach to tracking air quality is needed to estimate roughly the temporal and spatial distribution of pollutants. Air Quality models are used to indicate air quality standards. It is the least expensive techniques. Regulatory officials can use this modeling as monitoring instruments to evaluate the impact of emissions on ambient air quality. This can also be used to decrease the emissions required to meet the requirements. 9 Air Quality models are generally mathematical descriptions of pollutant transport, diffusion, and chemical reactions from the sources of pollutants (Duc and Azzi, 2009). They accommodate one or a lot of mathematical formulae that include parameters that have an effect on concentrations of pollutants at varying distances downwind of emission sources. Typically, they care for sets of pollutant input data that characterize the emissions, meteorology, and topography of a section and turn out outputs that describe that basis's required air quality. Based on the vital input variable treated the models can be simple or advanced. Advanced models are essentially suited for photochemical air pollution, dispersion in complex terrain, and long-range transport of pollutants. Simpler models are suited for the prediction of particulate matter pollutants of downwind sources. Air Quality prediction has a higher degree of uncertainties than another forecast, as the forecast must diagnose in addition to standard meteorological variables. The forecasting models reduce the uncertainties by "anchoring" the forecast with prior information available and by 'adjusting' the model with additional information like input variables (past measurement values, pollutant concentrations) and meteorological parameters (Ryan, 2016). All the statistical forecasting methods must be calibrated to the effect of large scale emissions. CO, fine particulate matter with an aerodynamic diameter of less than 10 μm (PM10) and less than 25 μm (PM2.5) are the most prevalent predicted pollutants.

IV. PROPSOED METHODOLOGY

Machine Learning (ML) is the data handling frameworks, which are built and executed to show the human cerebrum. The main object of the ML analysis is to develop a process of computing device for modeling the brain to execute various process of computing tasks at a faster rate than the traditional systems. ML execute various tasks such pattern matching and classification, optimization function and data clustering. These errands are exceptionally troublesome for conventional PCs, which are quicker in algorithmic procedure of registering undertakings and exact number juggling tasks. ML gangs substantial number of exceedingly interconnected preparing components called hubs or unit or neuron, which more often than not work in parallel and are arranged in standard models [8]. Every neuron is associated with the other by an association interface. Every association interface is related with weights, which contain data about the info flag. This data is required by neuron net to take care of a specific issue. ML, s aggregate conduct is portrayed by their capacity to learn, review and sum up preparing examples or information like that of human mind.

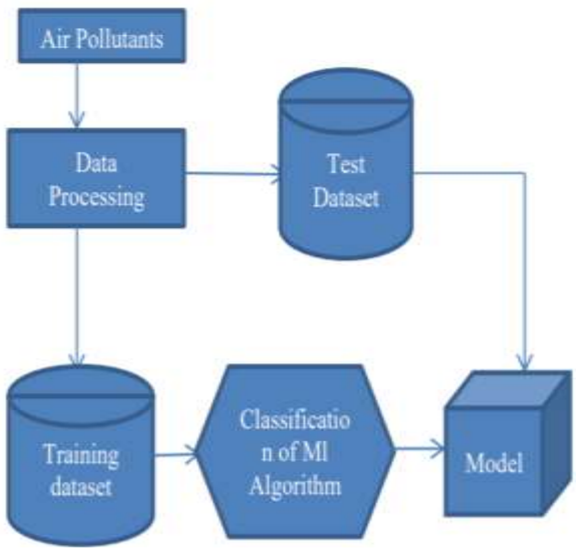


Fig. 1: Basic Diagram of Air Pollution Prediction

Research Methodology

The model consists of three parts. In the first part, the one-dimensional convolutional neural networks (convnets) performs local feature learning and dimensionality reduction on five input variables, the original data is processed by convolution and pooling to form low-dimensional feature sequences.

1-D CNNs fuse the feature extraction and feature classification processes into a single learning body. They can learn to optimize the features during the training phase directly from the raw input.

Since 1-D CNN neurons are sparsely-connected with tied weights, CNNs can process large inputs with a great computational efficiency compared to the conventional fully-connected Multi-Layer Perceptrons (MLP) networks. 1-D CNNs are immune to small transformations in the input data including translation, scaling, skewing and distortion. 1-D CNNs can adapt to different input sizes.

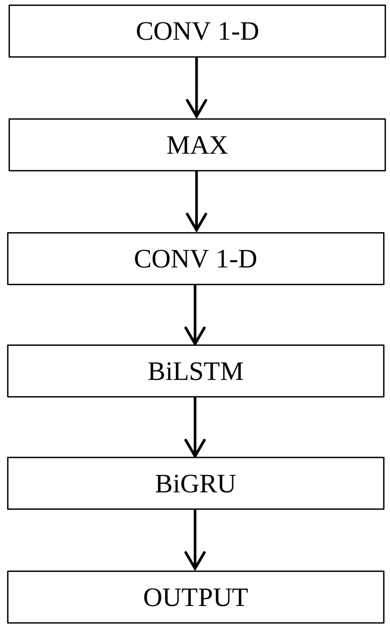


Fig. 2: Block Diagram of Proposed Methodology

In the second part, the bidirectional LSTM (BiLSTM) performs predictive features generated by each sub-model. The basic concept of the aggregated LSTM model is similar to that of voting. Instead of gathering the predictive features generated by each sub-model, this aggregated model, with all predictive features merged, will provide the final predicted value. The model consists of two categories: mixing and aggregation-learning. The category of mixing trains an independent model to produce predictive features first, and then uses these predictive features to train a model to combine it into a predicted final value. On the other hand, the category of aggregation-learning is responsible for synchronizing the predictive features from the sub-models provide a final composite dataset.

Third, the feature sequences is fed into the bidirectional GRU (BiGRU) neural networks, which reset gate and update gate constantly adjust their parameters in a large amount of training, so that it can learn the time dependence relationship between the information extracted from the convolutional neural networks. The layer contains only one neuron without any activation function, generating the predicted value of the PM2.5 concentration. Theoretically, the innovation of this method is the combination of the local feature extraction ability and lightness of convnets with the time series prediction ability of GRU by using 1D convnet as a preprocessing step before a GRU. On the other hand, by processing a sequence both way, a bidirectional GRU is able to catch patterns that may have been overlooked by a one-direction GRU.

Model Architecture Information
Model Summary
Model: "sequential"

Table II: Model Summary

Layer (type)	Output Shape	Param#
conv1d (Conv1D)	(None, 4, 32)	160
max_pooling1d (MaxPooling1D)	(None, 4, 32)	0
conv1d_1 (Conv1D)	(None, 4, 32)	1056
bidirectional (Bidirectional)	(None, 4, 80)	17760
bidirectional_1 (Bidirectional)	(None, 4, 80)	38720
bidirectional_2 (Bidirectional)	(None, 80)	29280
dense (Dense)	(None, 1)	81
Total params: 87,057 Trainable params: 87,057 Non-trainable params: 0		

Table 3: Hyper Parameters

Hyper Parameters	Value
Convolution layer filters	32
Convolution layer kernel_size	1
Convolution layer activation function	tanh
Convolution layer padding	Same
Pooling layer pool_size	1
Pooling layer padding	Same
Number of hidden units in BiLSTM layer	80
Number of hidden units in BiGRU layer	80
LSTM and GRU layer activation function	tanh
Time_step	4
Batch_size	64
Learning rate	0.001
Optimizer	Adam
Loss function	mean_absolute_error
Epochs	50

V. SIMULATION RESULT

Step 1:
Collect the dataset which wasn't so hard to find as it is publicly available from UCI Machine Learning Repository Link https://archive.ics.uci.edu/ml/machine-learning-databases/00381/.

Attribute Information:

- No: row number
- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- pm2.5: PM2.5 concentration (ug/m^3)
- DEWP: Dew Point (â„ƒ)
- TEMP: Temperature (â„ƒ)
- PRES: Pressure (hPa)
- cbwd: Combined wind direction
- Iws: Cumulated wind speed (m/s)
- Is: Cumulated hours of snow
- Ir: Cumulated hours of rain

Step 2: Cleaning the data, this is the task we should focus on very much. A cleaned data will perform much better than a noise full data (Nan values).

Step 3: There are 8 features important for the forecast: PM2.5, dew point, temperature, pressure, wind direction, wind speed and the cumulative number of hours of snow and rain. Hence, dropping other features. Feature selection is the next task to execute selecting 5 useful features out of everything that was given in the raw dataset.

Table 4: Selected features

	Pm 2.5	DEWP	cbwd	Iws
24	129.0	-16	SE	1.79
25	148.0	-15	SE	2.68
26	159.0	-11	SE	3.57
27	181.0	-7	SE	5.36
28	138.0	-7	SE	6.25

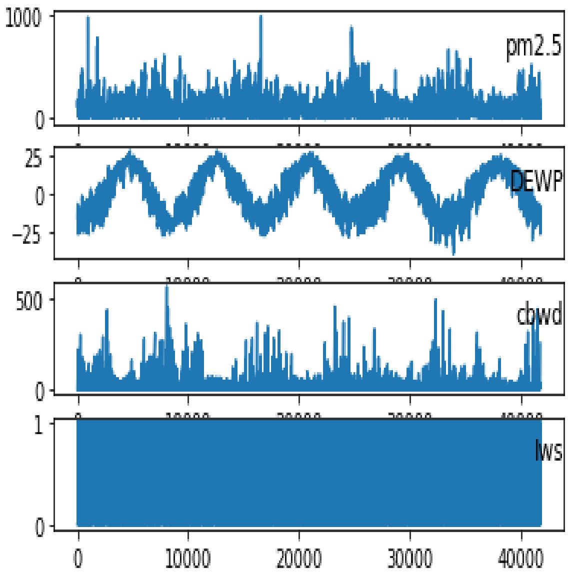
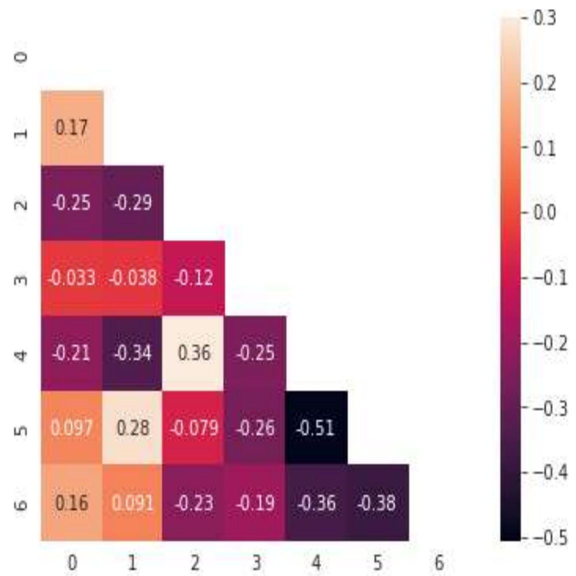


Fig. 3: Graph for Selected Features

Step 4: Encoding and Correlation the data is very important although our data is in numerical data but the wind direction, so it was necessary to perform.

Table 5: Encoding Data

	Pm 2.5	DEWP	Iws	NE	NW	SE	cv
24	129.0	-16	1.79	0	0	1	0
25	148.0	-15	2.68	0	0	1	0
26	159.0	-11	3.57	0	0	1	0
27	181.0	-7	5.36	0	0	1	0
28	138.0	-7	6.25	0	0	1	0



Step 5: We must sort the data according to the time. The dataset is transformed into a supervised learning problem and also normalize using MinMaxScaler(0,1).

Table 6: data according to the time

	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
4	129.0	1.79	0.0	1.0	138.0
5	148.0	2.68	0.0	1.0	109.0
6	159.0	3.57	0.0	1.0	105.0
7	181.0	5.36	0.0	1.0	124.0
8	138.0	6.25	0.0	1.0	120.0

Step 6: Reshaping the data into [samples, time steps, features], and splitting into train and test.

Table 7: Feature

train_X.shape	y.shape	test_X.shape	test_y.shape
(4320, 4, 4)	(4320,)	(24, 4, 4)	(24,)

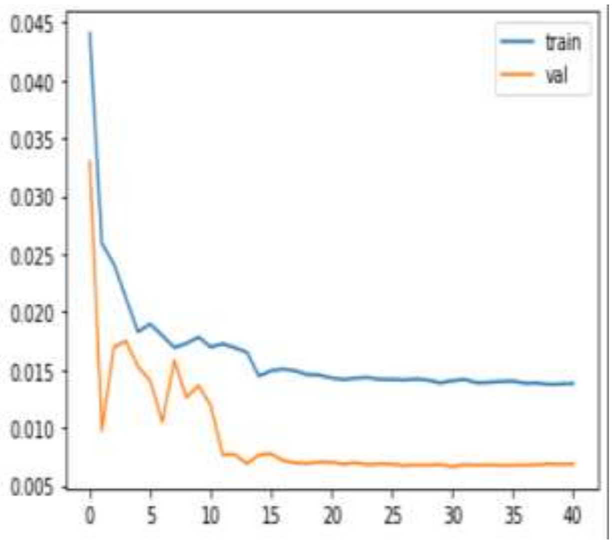


Fig. 4: Graphical Representation of Loss

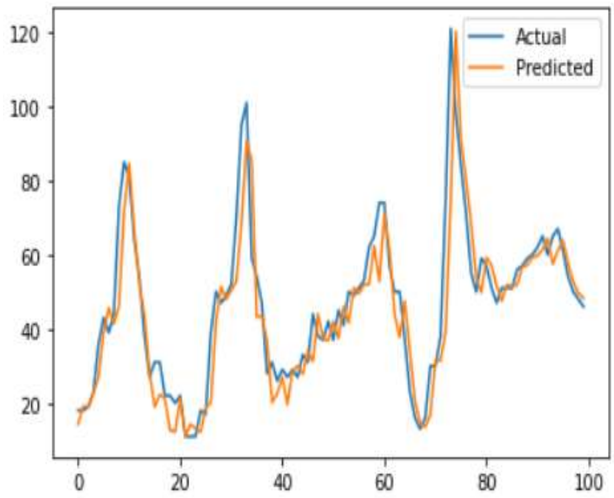


Fig. 5: Graphical Representation of Prediction

VI. CONCLUSION

An integrated database of monitoring station air quality and meteorological data can be used for validating the proposed models. The aim of this thesis is to use CBLSTM to investigate a dataset of air pollutants records for the Indian meteorological sector. It is more difficult to determine air quality. This research work will attempt to reduce the risk factor associated with forecasting the Air Quality Index (AQI) of India to a safe human level in order to save a significant amount of meteorological time and resources, as well as to predict whether the air quality is bad or good. In addition, compared with the other benchmark models, the accuracy of the CBLSTM model is significantly improved, which shows that the convnets can help the GRU to obtain better prediction performance, because convnets uses its local feature learning ability and subsampling ability to obtain a sequence pattern that is more conducive to GRU processing.

VII. FUTURE SCOPE

Generally, the prediction of air quality primarily emphasizes particulate pollutants and gaseous pollutants. Another pollutant in the air that cannot be ignored is aerosol pollutants. In the study of atmospheric aerosols related to haze pollution, most of them discuss the physical and chemical properties of aerosols, and the research on their biological components and properties is very limited. In the succeeding work, we will conduct research on the above research directions, and the present prediction method about the value of each pollution index in the air will play a significant role in the subsequent research work. The reason for this is that the source and spatiotemporal distribution of bio-aerosols in the air are complicated, and are largely affected by other pollutants in the air.

REFERENCES

[1] Chakradhar Reddy K, Nagarjuna Reddy K, Brahmaji Prasad K and P.Selvi Rajendran, "The Prediction of Quality of the Air Using Supervised Learning", 6th International Conference on Communication and Electronics Systems (ICCES), IEEE 2021.

[2] Jovan Kalajdjieski, Eftim Zdravevski , Roberto Corizzo, Petre Lameski, Slobodan Kalajdziski, Ivan Miguel Pires, Nuno M. Garcia and Vladimir Trajkovik, "Air Pollution Prediction with Multi-Modal Data and Deep Neural Networks", Remote Sensing, 2020.

[3] Zhang, Y.; Guo, L.; Wang, Z.; Yu, Y.; Liu, X.; Xu, F. Intelligent Ship Detection in Remote Sensing Images Based on Multi-Layer Convolutional Feature Fusion. Remote Sens. 2020, 12, 3316.

[4] Bai, L., Wang, J., Ma, X., and Lu, H. (2018), 'Air pollution forecasts: An overview', International journal of environmental research and public health 15(4), 780.

[5] Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J. and Sabde, Y. (2019), 'The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017'. The Lancet Planetary Health 3(1), e26-e39.

[6] Cai, S., Wang, Y., Zhao, B., Wang, S., Chang, X., and Hao, J. (2017), 'The impact of the "air pollution prevention and control action plan" on PM2. 5 concentrations in JingJin-Ji region during 2012–2020', Science of the Total Environment 580, 197-209.

[7] Cascio, W. E., and Long, T. C. (2018), 'Ambient Air Quality and Cardiovascular Health Translation of Environmental Research for Public Health and Clinical Care', North Carolina medical journal 79(5), 306-312.

[8] Huang, M., Zhang, T., Wang, J., & Zhu, L. (2015, September). A new air quality forecasting model using data mining and artificial neural network. In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 259-262).

[9] Huang, M., Zhang, T., Wang, J., and Zhu, L. (2015), A new air quality forecasting model using data mining and artificial neural network. In '2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)', pp. 259-262.

[10] Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., and Xie, G. (2018), 'Air quality prediction: Big data and machine learning approaches', International Journal of Environmental Science and Development, 9(1), 8-16.

[11] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002), 'An efficient k-means clustering algorithm: Analysis and implementation', IEEE Transactions on Pattern Analysis and Machine Intelligence, (7), 881-892.

[12] Kelly, F. J., & Fussell, J. C. (2015). Air pollution and public health: emerging hazards and improved understanding of risk. Environmental geochemistry and health, 37(4), 631-649.

[13] Kemp, A. C., Horton, B. P., Donnelly, J. P., Mann, M. E., Vermeer, M., and Rahmstorf, S. (2011), 'Climate related sea-level variations over the past two millennia', Proceedings of the National Academy of Sciences, 108(27), 11017-11022.