How to Cite:
Praveena, S., Pandey, H., Kumar, V. P., Meenatchi, S., & Mudradi, S. K. (2022). Prediction of environment pollution by employing long short-term memory network. *International Journal of Health Sciences*, *6*(S2), 8998–9009. https://doi.org/10.53730/ijhs.v6nS2.7334

Prediction of environment pollution by employing long short-term memory network

Dr. S. Praveena
Assistant Professor, Department of Electronics & Communication Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana 500075, India
Corresponding author email: spraveena_ece@mgit.ac.in

Dr. Himani Pandey
Assistant Professor, Applied Science and Humanities, ITM (SLS) Baroda University, Vadodara, Gujarat 391510, India
Email: himanipandey@itmuniverse.ac.in

Dr. V. Pradeep Kumar
Assistant Professor, Department of Computer Science and Engineering, B V Raju Institute of Technology, Narsapur, Telangana 502313, India
Email: pradeepkumar.v@bvrit.ac.in

Dr. S Meenatchi
Associate Professor, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
Email: meenatchi.s@vit.ac.in

Dr. Shreesha Kalkoor Mudradi
Associate Professor, Department of Electronics and Communication Engineering, Sambhram Institute of Technology, Bangalore, Karnataka 560097, India
Email: shreesha133984@gmail.com

***Abstract*---**Air pollution levels have risen as an outcome of urban and industrial development in so many developing countries. People and governments all around the world are concerned about air pollution, which has a severe influence on both personal health and long-term global development. As a government, it is responsible for preventing and controlling air pollution, as well as monitoring the pollutant's impacts on human health. There are numerous computer models available, ranging from statistics to artificial intelligence. Pollution levels are still out of control in some parts of the world due to a wide range of sources and factors. Because of accurate estimates of future air pollution, the government can take necessary action. Forecasting air pollution levels based on environmental data is becoming increasingly relevant as people become more worried about global

warming and urban sustainability. For replicating the complicated linkages between these variables, advanced Deep Learning (DL) algorithms hold enormous promise. The objective of this work is to provide a high level of accurate solution to the air pollution forecasting problem. Kaggle data will be employed to train a DL model that will forecast air pollution levels. Before it can be used in the analysis, the raw data must go through the necessary pre-processing phases. This time-series data set is solved using the Long Short-Term Memory (LSTM) of the Recurrent Neural Network (RNN) approach. During the development phase, the model's performance is evaluated using training time and loss measures. The model is then used to estimate the test data. The LSTM model is evaluated using a comparison plot of actual and predicted pollution, as well as the Root Mean Square Error (RMSE).

***Keywords*---**pollution, Kaggle database, deep learning model, loss, prediction.

Introduction

Since air pollution is hazardous to human health and should be eliminated as soon as feasible in both urban and rural areas, precise air quality forecasts are essential. Although there are various sorts of pollution, the most serious is air pollution, which is crucial to human health because we get our oxygen from breathing dirty air. Scientists have begun to pay attention to air pollution as the number of automobiles with internal combustion engines and industrial and agricultural operations have increased dramatically in recent years [1]. A lot of causes can contribute to air pollution. Outside, manufacturers, companies, and vehicles pollute the air, while smokes, toxins, and fragrances can pollute the atmosphere within a home. Primary and secondary pollutants are the two main categories of pollutants that contribute to air pollution. The regular occurrence of haze caused by industrialization development has brutally driven environmental pollution to a perilous climax in previous decades. That is, it is becoming more terrible than ever. One of the most harmful pollutants is PM2 or particulate matter with a diameter of fewer than 2.5 micrometres. This type of particle is exceedingly hazardous to one's health. In addition to causing respiratory and cardiovascular problems in short-term exposures ranging from a few hours to a few weeks, the WHO estimates that almost 90% of the world's population breaths polluted air that exceeds WHO air quality limits [2,3].

Air pollution is associated with 91% of premature mortality, particularly cancer, heart problems, respiratory disease, Alzheimer, and chronic obstructive. The 20% of global new born fatality causes a multitude of acute and chronic problems in childhood. A child's immune response, brain growth, the respiratory, and cardiovascular organ can all be harmed by high levels of air pollution in the mother's environment. Low birth weight and delayed growth in children have also been linked to air pollution. According to current estimates, air pollution is responsible for one out of every ten deaths in children under the age of five. Air pollution has been associated with asthma and impaired cognitive function in the

elderly [4]. A good system for predicting and measuring air pollution has tremendous implications for both human health and government policy. The method and procedure of pollution formation are extraordinarily complex owing to their irregular features in time and space [5], which have a substantial impact on the accuracy of forecasts. It is worthy of scrutiny. Moreover, because air quality information is intrinsically related to time, this is a sequence of data and has a pattern. Academics and scholars must pay close attention to time predictions given that the data is so current. This demonstrates the importance of time series analysis in a wide range of applications, like finance, health, astronomy, mining, and many others. Many works done so for in this research area are studied and a few of the works are presented below.

The paper [6] uses Recurrent Neural Network (RNN) models with LSTM units to estimate the level of PM10 particles shortly (+3 hours), as detected by sensors placed at various locations across Skopje. The models were trained using historical data from air quality measurements. To increase performance, they added temperature and humidity data to reflect the relationship between air pollution and seasonal changes in climatic conditions. The model's accuracy is compared to the PM10 concentration predicted by an Autoregressive Integrated Moving Average (ARIMA) model. The outcomes demonstrate that customized DL models frequently beat the ARIMA model, particularly when paired with past weather and polluted air information. The benefit of the proposed methods for reliable forecasts with only 0.01 MSE, may enable preventative strategies to decrease pollution levels, including such temporary closing down of big polluters or sending warnings so that inhabitants can move to a cleaner environment to prevent themself from pollution. The journal [7] presents a technique for predicting PM2.5 concentrations using RNN and LSTM. TensorFlow is used to run RNN and LSTM via a neural network developed with Keras, a high-level neural networks API written in Python. Keras is built on TensorFlow. The training data utilized in the network was provided by Taiwan's EPA (Environmental Protection Administration), which was aggregated into 20-dimension data and used for the forecasting test in 2017. They performed experiments in 66 Taiwanese locations to determine the accuracy of PM2.5 concentration forecasts for the next 4hrs. Results indicate that the suggested technique can reliably predict the concentration of PM2.5. Using the past data, they estimate the amount of PM2.5 in different Indian cities using multiple machine learning (ML) techniques in the paper [8]. The information contains temperature, wind, humidity, and levels of pollution before the selected time. They provided a comprehensive assessment of the forecasting method depending on the data and make important conclusions.

The journal [9] uses a CNN-LSTM architecture and data from smart sensor networks to anticipate current hour air pollution levels based on 24-hour pollution concentrations and a percentage of weather factors from the past hour. Before any analysis can occur, the data must be pre-processed to detect and fix any missing values. The model's performance is improved by accounting for the additional seasonal and temporal correlations of this type of data. When compared to different conventional ML approaches, the suggested DL model outperforms them on the required criteria. The accuracy of forecasting air quality requirements has been examined using a range of computing techniques such as statistics, ML, and DL. Pollutant levels are still out of control in various parts of

the world due to a variety of causes and reasons. The study [10] aims to forecast PM2.5 pollutants, using a bidirectional LSTM model. The accuracy of the proposed model is greater than that of the existing model when the following error estimation measures are compared. The mean absolute error is 7.53, and the symmetric mean absolute percentage error is 0.1664. The article [11] presents the standard variational autoencoder (VAE) and attention mechanism before building a prediction modelling approach built on a new integrated multiple directed attention (IMDA) computational intelligence framework. To evaluate the suggested predictive model, an experimental validation utilizing air quality data from four U.S. states is conducted. Six statistical factors were utilized to evaluate the accuracy of forecasts. An examination of the results reveals that IMDA variational autoencoder (IMDA-VAE) approaches perform satisfactorily in anticipating the presence of various pollutants at diverse sites. The suggested IMDA-VAE model outperformed other DL models, according to the research. They also demonstrated that when attention is introduced, the proposed model outperforms LSTM and GRU in terms of prediction accuracy.

The research tries to forecast air pollution in the future using the LSTM model. The journal is carried out in the following manner. The 1st part of the journal details the air pollution impacts and literature survey. The 2nd part shows the methodology of the research. The 3rd and 4th parts describe the data acquisition and processing steps involved to clean the data, and DL model architecture. The result obtained by the LSTM model in the train and test phase is discussed in the 5th part. Finally, the 6th part concludes the research and presents the future work.

Method

In the context of the present environmental conditions, air pollution forecasting is critical. The Kaggle website is used to gather time-series data on air pollution for this study. To begin the pre-processing stage, the raw data is used. Using an encoding algorithm, the categorical values included in the data are converted to numeric values and the null values in the data are removed. After that, the data is divided into two piles: train and test, as the name implies. The Keras DL library is used to build the LSTM model. The first data set is used to train the model. The performance metric is used to enhance the LSTM model before it is released for general use. The performance metric is measured by the loss function in this project. The final LSTM model is then tested using the data from the second half of the dataset. The RMSE is used to evaluate the data's performance at this point. The complete workflow of the research on air pollution forecasting is shown in figure 1.

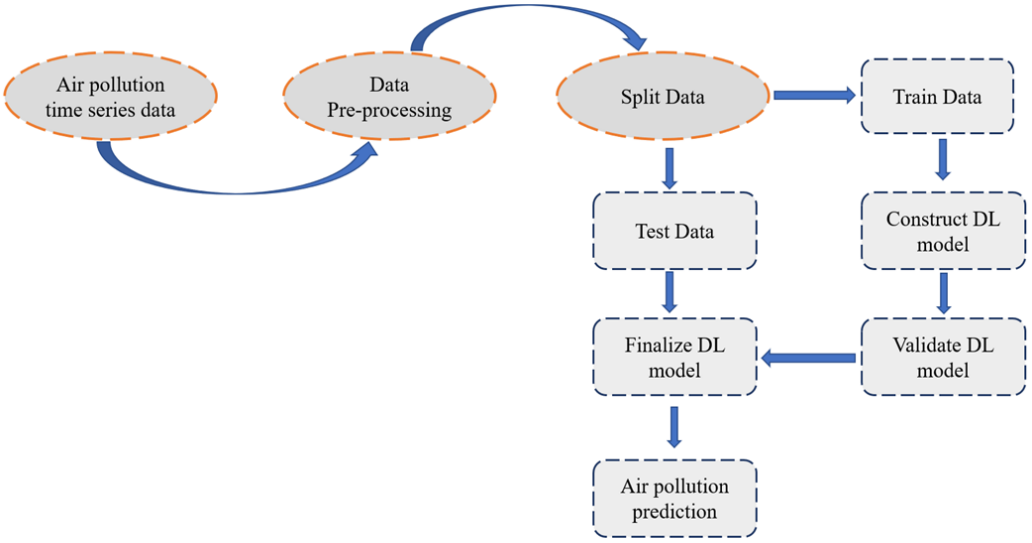

Fig. 1. Development of air pollution forecasting system

Data Acquisition and Data Processing

The dataset was collected from Kaggle and it contains data including columns for pollution, dew, temp, wind velocity, and whether it's raining or snowing. The time series predicting DL model will be used in this work to estimate pollution for the coming hours depending on pollution, dew, temperature, wind speed, snow, and rain factors. The data is depicted in Figure 2.

1 to 10 of 20000 entries Filter

index	date	pollution	dew	temp	press	wnd_dir	wnd_spd	snow	rain
0	2010-01-02 00:00:00	129.0	-16	-4.0	1020.0	SE	1.79	0	0
1	2010-01-02 01:00:00	148.0	-15	-4.0	1020.0	SE	2.68	0	0
2	2010-01-02 02:00:00	159.0	-11	-5.0	1021.0	SE	3.57	0	0
3	2010-01-02 03:00:00	181.0	-7	-5.0	1022.0	SE	5.36	1	0
4	2010-01-02 04:00:00	138.0	-7	-5.0	1022.0	SE	6.25	2	0
5	2010-01-02 05:00:00	109.0	-7	-6.0	1022.0	SE	7.14	3	0
6	2010-01-02 06:00:00	105.0	-7	-6.0	1023.0	SE	8.93	4	0
7	2010-01-02 07:00:00	124.0	-7	-5.0	1024.0	SE	10.72	0	0
8	2010-01-02 08:00:00	120.0	-8	-6.0	1024.0	SE	12.51	0	0
9	2010-01-02 09:00:00	132.0	-7	-5.0	1025.0	SE	14.3	0	0

Show 10 per page 1 2 10 100 1000 1900 1990 2000

Like what you see? Visit the data table notebook to learn more about interactive tables.

Warning: total number of rows (43800) exceeds max_rows (20000). Limiting to first (20000) rows.

Fig. 2. Air pollution dataset

We couldn't use the data we had gathered since it wasn't ready. To make anything useful, we must first prepare it. First, the date-time information was condensed into a single date-time. The identification of null values has been completed. Missing data refers to a variable's data value that is not stored for some reason (or missing values). A study's conclusions can be significantly influenced by a lack of complete data [12]. As detailed in the article [13], numerous research has been

conducted on handling incomplete information, challenges connected to incomplete information, and approaches to prevent or reduce incomplete information. A brief examination suggests that the few readings are not available. The 'NA' elements inside the database have been tagged as 0 since they could pose issues in the future. We eliminated the item labelled "no". Ultimately, the 'NA' elements were replaced with '0' entries, as well as the rows for the 'NA' elements were eliminated. Figure 3 depicts a line plot with 8 subplots exhibiting five years of data for each input and output variable. This image facilitates comprehension of data behaviour.

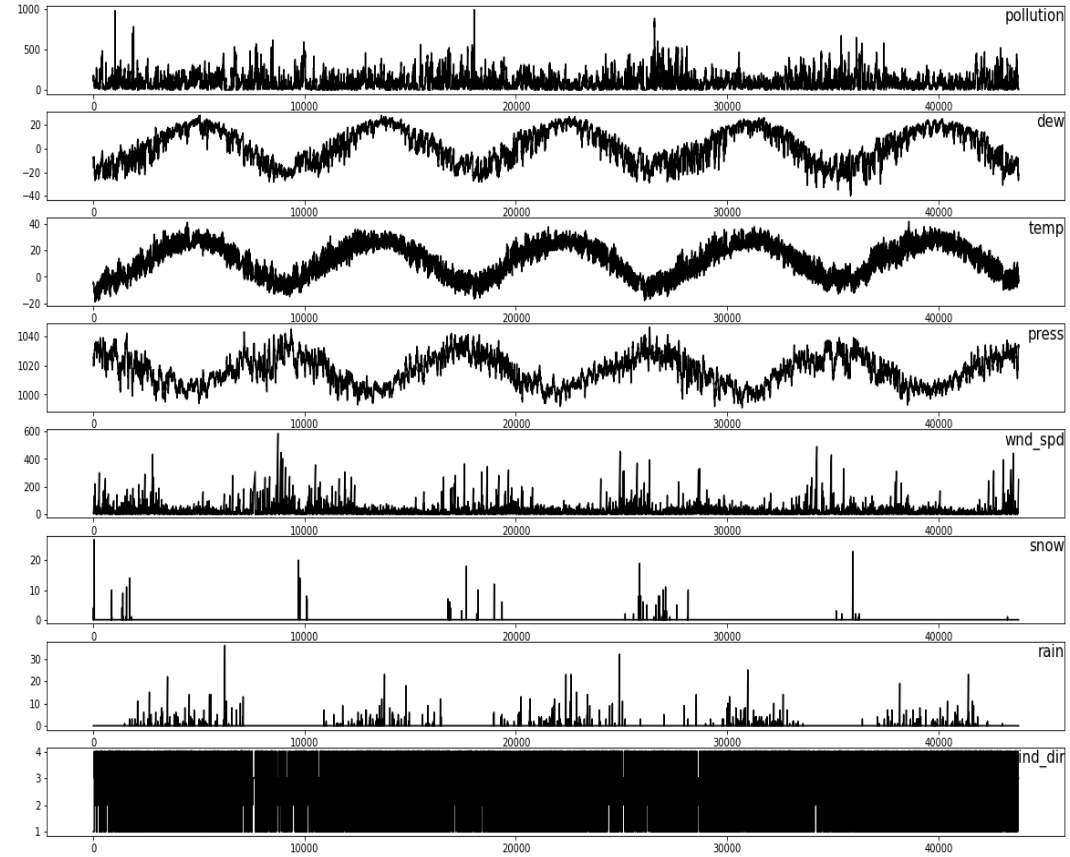

Fig. 3. Plotting of air pollution dataset

Secondly, the pollution data had to be prepared. The input parameters were normalized for the supervised training. It predicts pollution measurements concerning current time (t) depending on the present climatic factors and the preceding time (t-1) pollution measurements. Loaded data and then used label encoding to encode the dataset's wind speed attribute (integer encoding). A dataset's features may include one or more labels in numeric or text format. Humans will have an easier time analysing the facts in this manner, but computers will struggle [14].

As a result, we use encoding to make a computer interpret these labels. Label Encoder assigns a numerical value to each label and replaces this value in the dataset. It is useful when the labels have different priorities. After all dataset characteristics were normalized, the data was turned into a supervised technique. The weather variables for the anticipated hour were then eliminated. Incorporating label encoding for wind speed parameters, employing differencing and seasonal adaptation, keeping every series constant, and supplying over 1 hour of incoming time sequence data can make this data processing more flexible.

LSTM

RNNs such as LSTM was proposed for the first time in 1980 [15]. RNNs are a form of potent artificial neural network that is likely employed for time-series prediction issues. RNNs can recall previous data and make a decision based on that. Furthermore, RNNs are susceptible to disappearing and inflating gradients, resulting in abnormally sluggish or halted model development. In 1997, LSTMs were created in response to these challenges. LSTMs possess extended recollections and therefore can train on inputs with considerable time lags between them. As in sequence learning challenges, remembering earlier data is vital. Long-term LSTM learning resolves the gradient fade as well as explosion issues. Due to the memory module, the LSTM can retain inaccurate data and maintain gradient movement. As a solution, the fading issue has been addressed, and many step sequencing could be utilized to acquire fresh information. LSTM surpasses other RNN architectures since it eliminates the issue of vanishing gradients. A sequence of memory gateways, including the forget, input, and output distinguishes the LSTM from RNN. The less critical stuff is deleted during learning and allocated space inside the memory module for more recent and pertinent information. This is the problem that forgets gateway is intended to address. The forget gateway deletes or maintains data by increasing the memory module value by a positive number within 0 and 1. If a value must be maintained through multiple memory module steps, an input gateway is introduced to the LSTM architecture. And the output gateway is employed to resolve the issue of multiple memories competing with one another. As once the memory gateway is understood, the LSTM model may be expressed.

$$i_t = \sigma(W^{it}\bar{x}_t + W^{it}h_{t-1}) \quad [1]$$
$$o_t = \sigma(W^{io}\bar{x}_t + W^{ho}h_{t-1}) \quad [2]$$
$$f_t = \sigma(W^{if}\bar{x}_t + W^{hf}h_{t-1}) \quad [3]$$

Where,
i_t−→ input
o_t−→ output
f_t−→ forget

An LSTM consists of three gateways: an input gateway that selects either to receive new data, a forget gateway that erases useless information, as well as an output gateway that specifies which data to output. The operation of these three mathematical gateways depended upon this sigmoid function, which works between 0 and 1. These three sigmoid gates are depicted in figure 4. A horizontal line flowing through the cell represents its state.

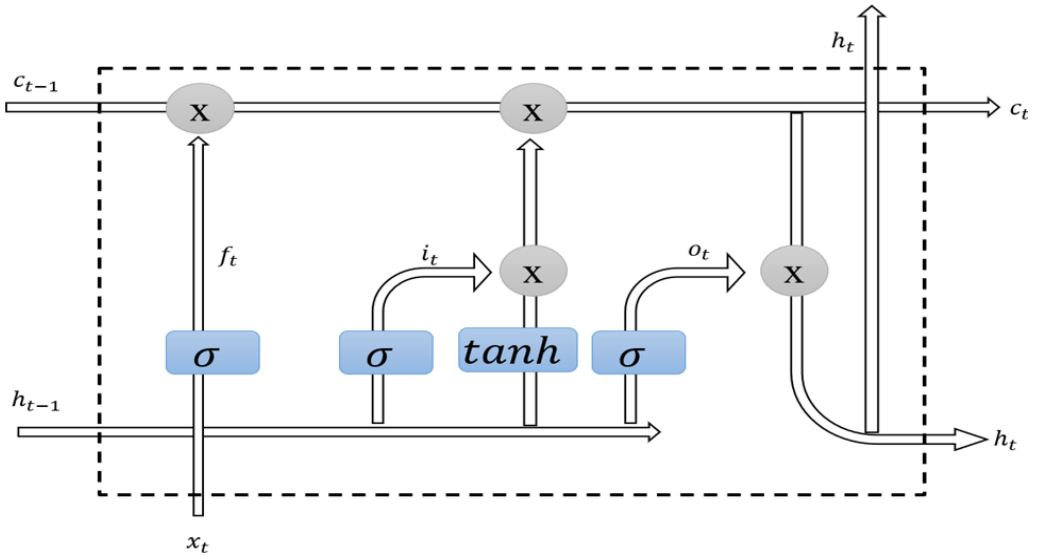

Fig. 4. LSTM architecture

Result and Discussion

This part discusses the result obtained by training and testing data. The collected data from Kaggle is used in this research. After collecting the raw data, the pre-processing steps like missing value elimination in all samples of data, and categorical data conversion in the wind column. The pure data is obtained after this stage. This data is given to the LSTM model for training purposes. The training is done for 40 epochs. And the loss function, batch size, and optimizer chosen are RMSE, 72, and ADAM. Then the loss result obtained in the training phase is plotted in figure 5. Figure 5 comprises two different colour plots such as blue, and cement. The blue plot represents the loss value obtained in the training stage and the cement represents the loss value obtained in the validation stage. After the first three epochs, the loss value will be lower than 0.001. The loss value becomes constant after the 7th epoch, the constant loss value of the validation stage is 0.0005 and for training is 0.0008.

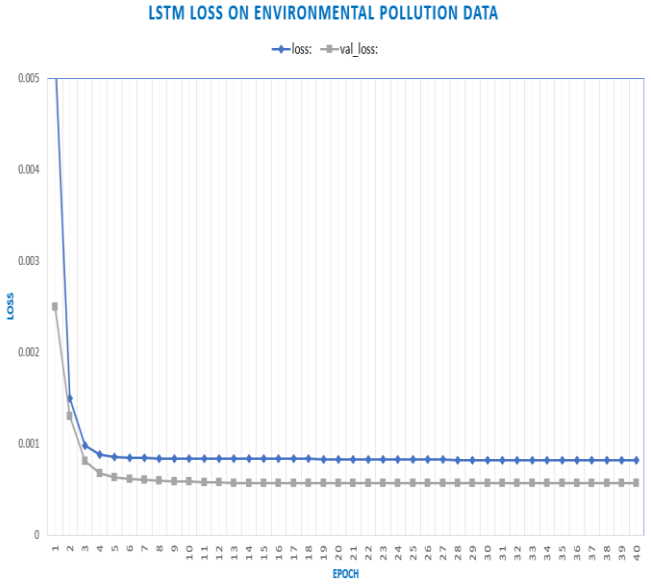

Fig. 5. Loss of LSTM model in the training phase

The loss obtained from two phases of training and validation shows that the model is well enough to forecast air pollution.

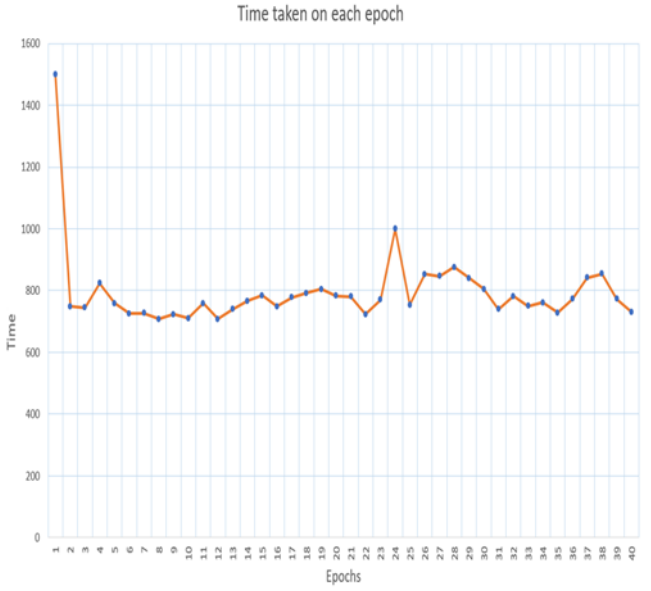

Fig. 6. Time is taken to train the LSTM model

Time is also an important parameter for the DL model. If the time taken to run the model is too long, then the implementation of the model in the real-world faces so many difficulties. So, the time required to train and validate the model is analysed and plotted using the line graph which is shown in figure 6. The figure shows the time taken by the LSTM model in each epoch. The x and y-axis of the

plot take the epochs and time in milliseconds. From the figure, it is clear that the time taken for both the train and to validate the model is only milliseconds. So, this criterion also satisfies the LSTM model.

After getting satisfactory results in the training stage. The mode is taken for testing. The input parameters of the test data are passed to the LSTM model and allow the model to predict the air pollution in ppm concerning the hours. The model has done the work very effectively. The proof for this statement is the comparison plot shown in figure 7. The comparison chart is composed of two plots such as actual air pollution in each hour and is denoted by green colour, next to the predicted air pollution and it is represented by blue colour. The figure clearly shows that the LSTM model accurately predicts future air pollution. To prove the model quality mathematically RMSE is employed. The root square of the mean square of all errors in the RMSE. It may readily indicate the accuracy of the predicted outcome. The calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad [4]$$

The RMSE value obtained by the model is 2.67.

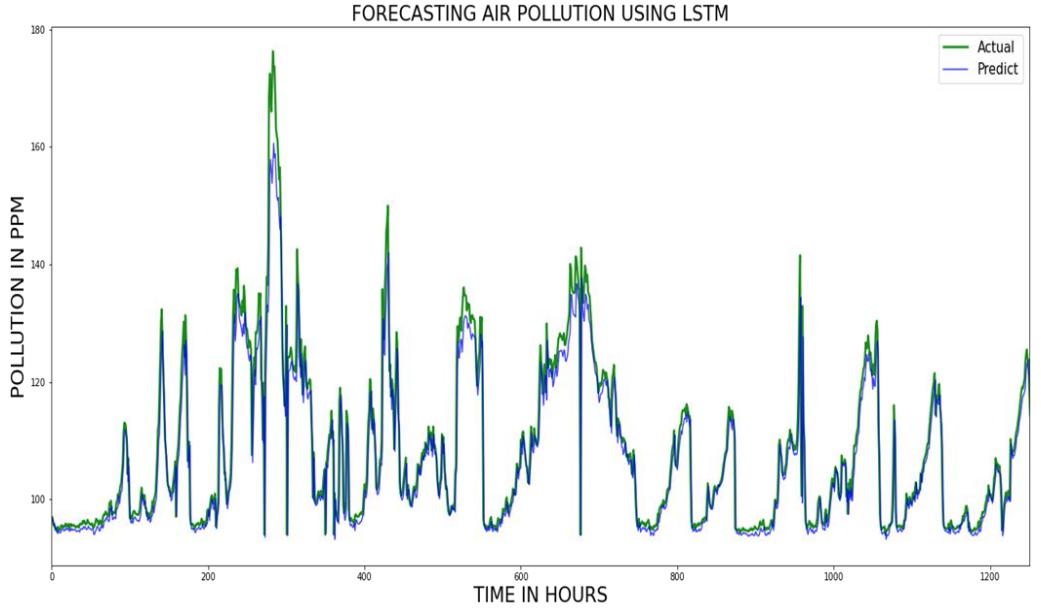

Fig. 7. Actual versus predicted air pollution by LSTM

Conclusion

The environment includes everyone and everything in our immediate surroundings. Human activities and natural calamities are contaminating the environment, and one of the worst offenders is air pollution. The number of pollution particles in the atmosphere is affected by wind speed, direction, relative humidity, and temperature. When there is a lot of humidity in the air, we feel

hotter because our perspiration cannot evaporate. Urbanization contributes significantly to air pollution as transportation infrastructure improves. Another major source of air pollution is industrialization. It is critical to take action to reduce environmental air pollution. Forecasting is essential for understanding and reducing air pollution. Prediction of air pollution has previously been attempted using traditional methods such as mathematical and statistical, but these are time-consuming and lead to inaccuracy. Because of technological developments, it is now quite straightforward to access data on pollution levels of the previous year from Google. Data for training and testing were segregated. Using the LSTM frameworks employed in this paper, air pollution can be anticipated in advance. The model learns from training data and confirms its results with test data. Different metrics, including RMSE, loss, and time, are used to evaluate the development.

In the future, hardware and the Internet of Things (IoT) could be employed to put this work into practice. An IoT-enabled sensor network captures environmental data, which may later be freely shared globally. Using real-time data and IoT, the government can assess pollution levels in any city. Because of this, the government can take proactive measures like limiting driving hours, shutting down factories temporarily, and sending out public service announcements. This method contributes to the country's attempts to clear the environment of harmful pollutants.

References

1. U. A. Hvidtfeldt, M. Ketzel, M. Sørensen, O. Hertel, J. Khan, J. Brandt, and O. Raaschou-Nielsen, “Evaluation of the danish airgis air pollution modeling system against measured concentrations of pm2.5, pm10, and black carbon,” Environmental Epidemiology, vol. 2, no. 2, p. e014, 2018, 10.1097/EE9.0000000000000014
2. Nada Osseiran, Christian Lindmeier, “9 out of 10 people worldwide breathe polluted air, but more countries are taking action”, 2018. https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action
3. Pöschl U. “Atmospheric aerosols: composition, transformation, climate and health effects”. Angewandte Chemie Int Ed., Vol. 44, No. 46, pp. 7520–40., 2005, https://doi.org/10.1002/anie.200501122.
4. Samayan Bhattacharya, Sk Shahnawaz, “Using Machine Learning to Predict Air Quality Index in New Delhi”, 2021, https://arxiv.org/pdf/2112.05753
5. Lu D, Mao W, Xiao W, Zhang L. “Non-linear response of pm2.5 pollution to land use change in China”. Remote Sens. Vol. 13, No. 9, pp. 1612, 2021, 10.3390/rs13091612
6. M. Arsov et al., "Short-term air pollution forecasting based on environmental factors and DL models," 2020 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 15-22, 2020, 10.15439/2020F211.
7. Y. -T. Tsai, Y. -R. Zeng and Y. -S. Chang, "Air Pollution Forecasting Using RNN with LSTM," *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and*

Technology Congress(DASC/PiCom/DataCom/CyberSciTech), pp. 1074-1079, 2018, 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00178.

8. J. Mohith, D. Kulshrestha and K. R. Jothi, "A Comprehensive Analysis of Machine Learning Methods for Air Pollution Forecasting," *2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)*, pp. 15-19, 2021, 10.1109/ICITech50181.2021.9590113.
9. L. Jovova and K. Trivodaliev, "Air Pollution Forecasting Using CNN-LSTM DL Model," *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO),* pp. 1091-1096, 2021, 10.23919/MIPRO52101.2021.9596860.
10. S. Jeya and L. Sankari, "Air Pollution Prediction by DL Model," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS),* pp. 736-741, 2020, 10.1109/ICICCS48265.2020.9120932.
11. A. Dairi, F. Harrou, S. Khadraoui and Y. Sun, "Integrated Multiple Directed Attention-Based DL for Improved Air Pollution Forecasting," in *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-15, 2021, Art no. 3520815, 10.1109/TIM.2021.3091511.
12. Graham JW. “Missing data analysis: making it work in the real world.” Annu Rev Psychol. Vol. 60, pp. 549–576, 2009, 10.1146/annurev.psych.58.110405.085530.
13. O'Neill RT, Temple R. “The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it”. Clin Pharmacol Ther. Vol. 91, pp. 550–554, 2012, 10.1038/clpt.2011.340.
14. Neha Sharma, Harsh Vardhan Bhandari, Narendra Singh Yadav, Harsh Vardhan Jonathan Shroff, “Optimization of IDS using Filter-Based Feature Selection and Machine Learning Algorithms”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Vol. 10 Issue-2, 2020, 10.35940/ijitee.B8278.1210220
15. Werbos PJ. Generalization of backpropagation with application to a recurrent gas market model. Neural Netw, Vol. 1, No. 4, pp. 339–56, 1988, 10.1016/0893-6080(88)90007-X