**Air Quality Prediction Using Deep Learning Models**

Sashank Donavalli, Department Of Computational Intelligence, School of Computing,SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.

Harshitha Kambham, Department Of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.

Dr. M. Dhilsath Fathima, Department Of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.

**ABSTRACT**

Air Pollution is a significant concern in today’s world for both the public and the environment. Monitoring and predicting air pollutant concentrations are essential for effective pollution control and management. This study focuses on the development of a time series prediction model for multiple pollutants in the air using advanced deep-learning techniques. The time series data is taken into account, capturing temporal dependencies and seasonal patterns that are inherent in air quality datasets.

This study explores the application of various deep learning architectures, including Recurrent Neural Networks (RNN) and other hybrid models. This research highlights the effectiveness of deep learning models in predicting air quality and their potential to provide accurate and timely information for air quality monitoring and management. The findings of this study contribute to the advancement of air quality prediction systems, enabling better decision-making for public health and environmental protection.

***Keywords:*** *Long short-term memory, Recurrent neural network, Convolutional neural network, Back propagation neural network, Gated neural network, Bi-directional neural network, Time series Prediction.*

**I. INTRODUCTION**

The growing problem of urban air pollution significantly hampers the Cities' capacity to flourish sustainably and establish an ecological civilization. The lives, productivity, and health of people are all impacted by the quality of the air.

Deep learning models used in detecting air quality index in this study are RNN, LSTM, GRU, and Bi LSTM. The choice of deep learning model depends on the specific characteristics of the air quality data, the size of the dataset, and the required prediction horizon. It's important to preprocess the data properly, split it into training, validation, and test sets, and fine-tune the model's hyperparameters to achieve the best performance. The proposed model should be able to continuous monitoring and should be able to perform with any other dataset.

**II. LITERATURE REVIEW**

In research paper [1] suggests an environmental quality prediction model based on LSTM, in light of the state's increased focus on the present environment's ongoing decline in air quality. This study used the dataset taken from the Environmental Protection Agency. The paper predicts the concentration of the pollutants using parameters sulphur dioxide, carbon monoxide, nitrogen dioxide, ozone, particulate matter 2.5, particulate matter 10, wind direction, and temperature. This research paper focuses on the challenges faced in observing the quality of air in real-time. The environmental prediction model will then be introduced. Using LSTM, we finally produce an AQI prediction and examine the prediction error. According to the findings, LSTM is a good predictor of the air quality index.

In paper [3] suggests a hybrid approach to indoor air pollution prediction using deep neural networks and fuzzy logic. The paper has trained and forecasted the dataset obtained by the sensors present indoors in Shanghai from Nov 2016 and Mar 2017 using PM2.5 pollution as an example. Using CNN-LSTM and LSTM the authors suggested network FL-CNN-LSTM built on the framework PyTorch, The authors created comparison tests to test the prediction process. According to the findings, deep neural networks that have been infused with fuzzy logic can offer improved predictability and interpretability for the intended use.

In the research paper [2] to anticipate the hourly AQI, several approaches are used, including a linear model and cutting-edge methods like BILSTM, LSTM, CNN, BPNN (Back Propagation neural network), and GRU. Experiments evaluating the performance of different strategies reveal that the BiLSTM(Bidirectional Long short-term Memory) provides the greatest results.

This research [5] proposes a hybrid air quality prediction model that combines deep neural networks and K-means clustering. BiLSTM and there is a fully connected neural network to make up the deep neural network with regressive computation capability. Firstly, the research goal of this paper is to monitor the meteorological historical data provided by Qingdao City. Using the k-means clustering technique, the historical meteorological dataset is divided into four categories by a quarter.

The dataset utilized in the paper [6] is made up of hourly data on multiple air pollution types collected at multiple locations in multiple Indian cities. 16 qualities were present. Random Forest, Decision Tree, and Support Vector Machine (SVM) were the models that were used. A method for random forest categorization produces the most accurate results. With a maximum accuracy of 74%, it outperformed the other methods. The present research will benefit from these findings, which will also direct future investigations.

To make the air quality prediction easier for the general public to understand, a new comprehensive evaluation approach called the LSTM-Fuzzy algorithm is developed in the paper [7]. The experiment results show that while the accuracy of UF is often stable in comparison to the DF, the accuracy of DF typically declines as the forecast time grows.

**III. EXISTING SYSTEM**

Authors and researchers in the past have conducted various experiments with models such as machine learning models (Support Vector Machine, Random Forest) logistic Regression, logistic regression, and Naive Bias)for the prediction of Air Quality Index. These machine learning models are also used to classify the air quality as bad, satisfactory, good, and very good using machine learning techniques such as K-means clustering.

Deep learning models, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNN), have gained a lot of popularity for their ability to capture complex temporal patterns in air quality data. Many studies have recognized the importance of considering both temporal and spatial aspects of air quality. Time series analysis techniques, like autoregressive models and LSTM neural networks, have been deployed to record all the temporal dependencies that are present.

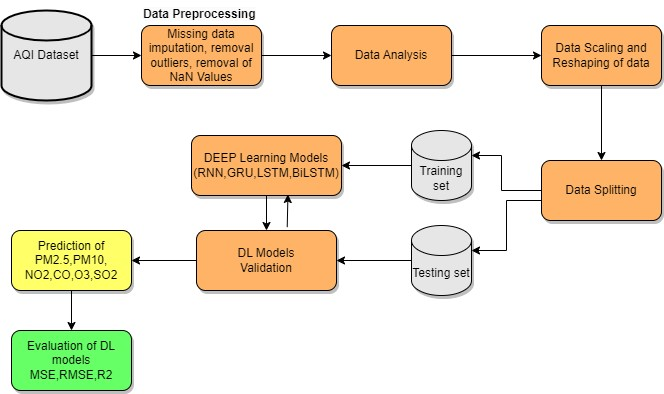
Additionally, spatial models have been developed to account for variations in air quality across different locations within a region.

Researchers have investigated the importance of various input features in predicting AQI. Many weather-related features were taken into consideration such as wind speed, wind direction, and temperature have been found to play a significant role in air quality forecasting. Geographical factors, such as proximity to pollution sources, have also been considered.

Past works have encountered challenges such as dealing with missing data, model overfitting, and the need for domain expertise in environmental science. Additionally, addressing issues related to non-stationary data and the interpretability of complex models has been a topic of concern.

**IV. PROPOSED SOLUTION**

Keeping all the points mentioned above at the forefront of the target, we have arrived at a revised architecture and solution for the project. The project that we have done contains the following modules which are described in detail below (Fig 1):



**Fig.1 Architecture of the proposed model**

The paper proposes to use four deep-learning models for the prediction. The four deep learning techniques are:

•RNN

•LSTM

•GRU

•Bi-LSTM

Using these prediction models the paper proposes to predict all six pollutants namely PM2.5, PM10, SO2, NO2, O3, and CO. Pollutant prediction is crucial for several reasons, including the fact that it plays a very big effect on the environment, the public wellbeing, and many different facets of society.

A.RNN

An Input Layer that accepts input data with the shape of (None, 1, 33), is where the model starts. This shape implies that the model has 33 characteristics and a one-time step for input sequences. Following the input layer comes a Simple RNN (Simple Recurrent Neural Network) layer with 128 units. Recurrent neural network layers called Simple RNN are utilized for sequence data. Weights and biases are among the 20,736 parameters in the Simple RNN layer. A dense layer with 128 units and the label "dense\_2" follows the Simple RNN layer. This layer contains 16,512 parameters and is fully linked. Another dense layer with the name "dense\_3" and six units are present, and the output layer has 774 parameters.

B.GRU

The model's first layer is an Input Layer that accepts input data with the shape of (None, 1, 33). This shape implies that the model anticipates 33 characteristics and one-time step for input sequences. A dense layer with 128 units follows the input layer. Weights and biases are among the 62,592 parameters in the GRU layer. Recurrent neural network layers called GRUs are employed for sequence data. A dense layer with 128 units and the name "dense\_6" follows the GRU layer. With 16,512 parameters, this layer is dense and fully connected. Another dense layer with the name "dense\_7" and six units is present, and it is most likely the output layer. There are 774 parameters in this layer.

C.LSTM

The model has its first layer which is an Input layer that accepts input data with the shape of (None, 1, 33). This shape implies that the model anticipates 33 characteristics and one-time step for input sequences. A dense layer of long short-term memory (LSTM) with 128 units follows the input layer. Weights and biases are among the 82,944 parameters in the LSTM layer. Long-term dependencies in sequences can be captured by the LSTM layer, a form of recurrent neural network layer that is employed for sequence data. A dense layer with 128 units called "dense\_8" follows the dense layer. With 16,512 parameters, this layer is dense and fully connected. Another dense layer with the designation "dense\_9" and six units are present at the end, and there are 774 parameters in the output layer.

D. Bi-LSTM

The input layer accepts input data with the shape of (None, 1, 33). This shape implies that the model anticipates 33 characteristics and one-time step for input sequences. A Bidirectional layer with the label "bidirectional" follows the input layer. Sequence data is frequently processed in both forward and backward directions using bidirectional layers. There are 165,888 parameters in this bidirectional layer. An additional layer called "multi\_head\_attention" with 131,712 parameters follows the bidirectional layer. For attention-based operations and sequence processing, models like Transformers use a mechanism called multi-head attention. After the Flatten layer, there is a dense layer with 128 units. 32,896 parameters make up this dense layer. Finally, there is another dense layer with six units; this layer is the output layer and 774 parameters make up this layer.

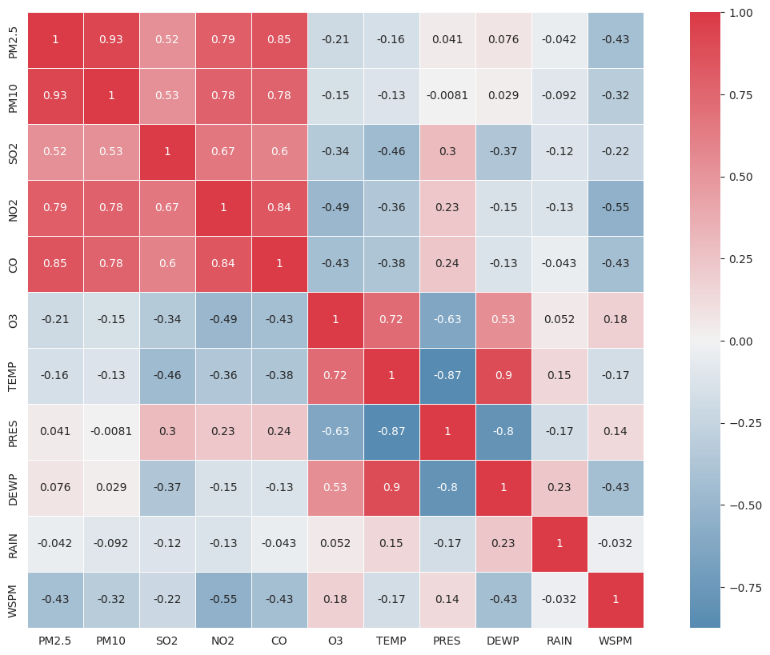
**V. DATA PREPARATION AND PREPROCESSING**

A. Data Collection: Gather historical air quality data, meteorological data, and geographical information from monitoring stations from Kaggle

B. Exploratory Data Analysis: Visualizing and summarizing the data to obtain an understanding of the dataset. The visualizations contained a summary of outliers and an analysis of Pollutants concerning station, year, day, and hour. Plotting time series data, doing statistical analyses, and spotting trends, and seasonality in the given dataset.

C. Data Preprocessing: In Data preprocessing, initial steps are taken to clean the data. All the missing values in the given dataset are replaced by the mean values. The outliers are removed to ensure data integrity and quality are maintained.

D. Data Correlation: The direct cause of the drop in air quality is pollution. Air circulation also helps to disperse toxins in the environment. There is a high correlation among the pollutants. The six pollutants namely O3, CO, NO2, SO2, PM10, and PM2.5 are used for prediction.



**Fig.2 Correlation**

**VI. TIME SERIES PREDICTION**

Time series data is essential for air quality prediction because it provides insightful information and useful applications for managing air quality. Time series data have several uses in the forecast of air quality. To protect the environment and public health, it is crucial to comprehend the temporal dynamics of air quality, locate pollution sources, and put into practice efficient remedies. In addition, early warning systems, regulatory compliance, and well-informed policy decisions are all supported by time series analysis are essential to the continuous decline in air quality, and are the reasons for negatively impacting human health.

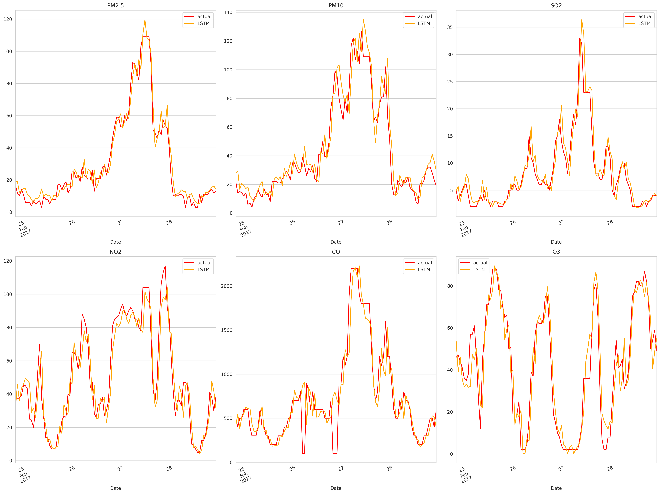
In this paper, the time-series prediction of pollution prediction by the four models RNN (Fig 3), GRU(Fig 4), LSTM(Fig 5), and Bi-LSTM (Fig 6) is shown below:



**Fig.3. Time series prediction by RNN**



**Fig.4. Time series prediction by GRU**



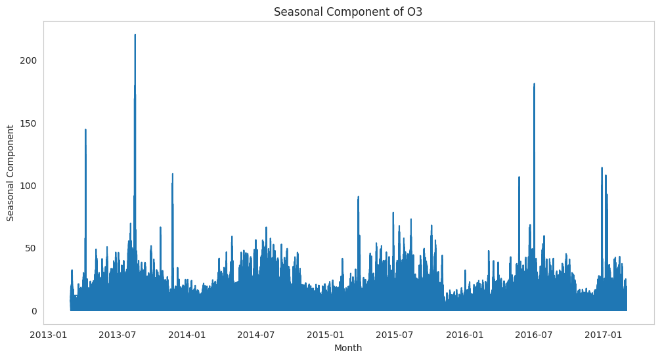
**Fig.5. Time series prediction by LSTM**



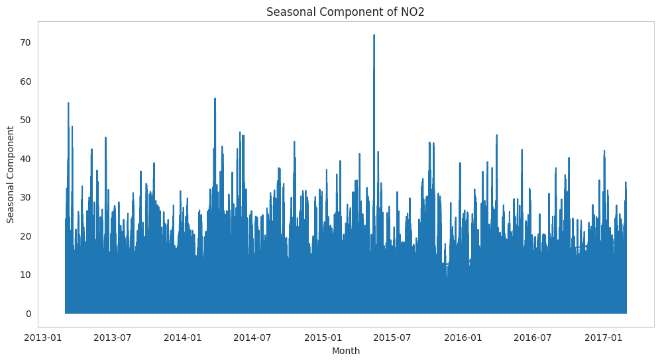
**Fig.6. Time series prediction by Bi-LSTM**

**VI. SEASONALITY TREADS**

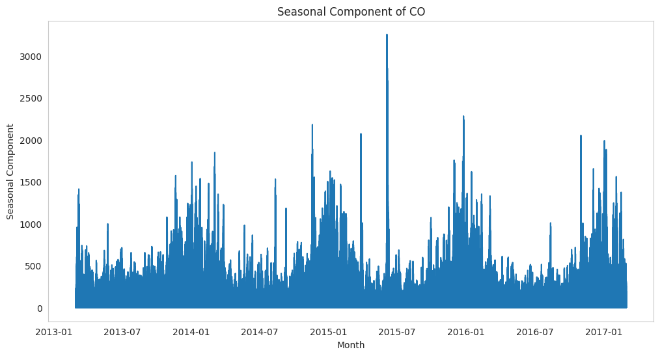
Seasonal patterns in air quality refer to regular fluctuations in the concentrations of pollutants that occur according to a specific timetable all year. These patterns have important ramifications for the forecast and management of air quality and are impacted by various factors, such as weather, human activity, and natural phenomena.



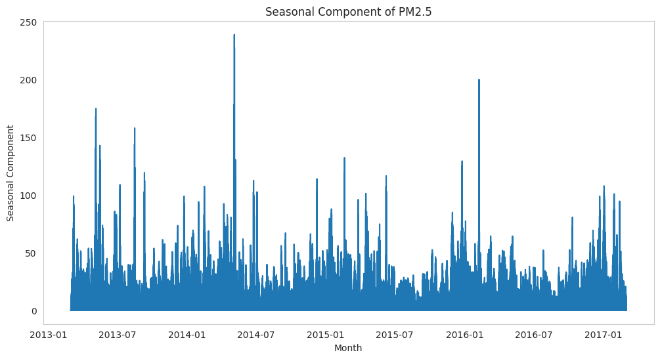
**Fig.7. Seasonal Component of O3**



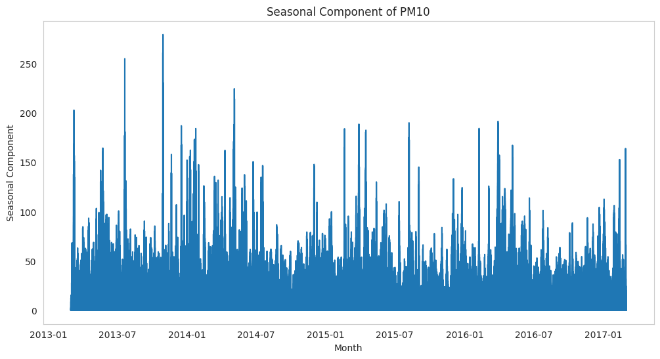
**Fig.8. Seasonal Component of NO3**



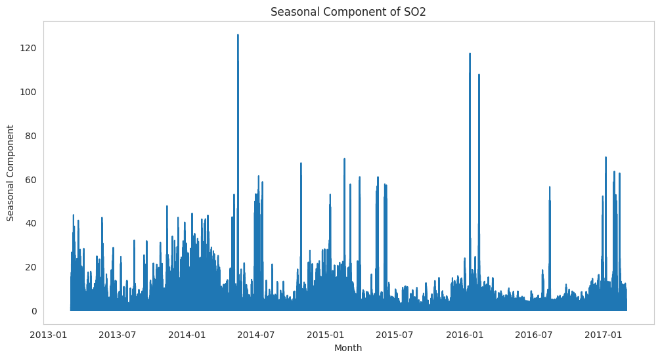
**Fig.9. Seasonal Component of CO**



**Fig.10. Seasonal Component of PM2.5**



**Fig.11. Seasonal Component of PM10**



**Fig.12. Seasonal Component of SO2**

**VI. RESULTS**

The model is tested by the three evaluation parameters:

* R2(R-Squared)
* MAE(Mean Absolute Error)
* RMSE (Root Mean Squared Error)

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | MAE | R2 |
| RNN | 102.4 | 32.139 | 0.856 |
| GRU | 101.393 | 31.690 | 0.858 |
| LSTM | 102.745 | 31.978 | 0.859 |
| Bi LSTM | 101.1 | 31.605 | 0.861 |

**VI. CONCLUSION**

This time series prediction is a crucial and multidimensional endeavour with significant implications for public health, environmental conservation, and policymaking. This complex task requires the integration of deep learning techniques and real-time monitoring.

In our proposed approach, the prediction of the concentrations of major pollutants are taken into consideration where the model entertains to be trained and tested on real-time data. In the future, the model can developed for Health Impact Assessment where the model can assess the potential threat and alert the population regarding the threat in the environment.

**VII. CHALLENGES**

Predicting air quality is a crucial task with several obstacles. The intricate interaction of several variables that affect air quality, such as weather, emissions from different sources, and chemical reactions in the atmosphere, is one of the main challenges. Precise forecasting is a difficult endeavor since these interactions are dynamic and complex. Furthermore, since real-time and spatially distributed data are crucial for predictive model performance, obtaining and maintaining high-quality data can be quite difficult. The complexity is further increased by the introduction of new pollutants, such as volatile organic compounds and fine particulate matter, and the requirement for fine-grained spatial projections.

Forecasting air quality is complicated and faced with a number of issues due to geographic factors. Mountains, valleys, and bodies of water are examples of different topography that can have a big impact on local wind patterns and air pollution dispersal. It can be challenging to generalize predictions across wider regions due to the possibility of microclimates with unique air quality characteristics resulting from these topography variances. Forecasting can also be made more difficult by the proximity of emission sources, such as industrial sites or urban areas, to topographical features, which can result in the creation of localized pollution hotspots. Topography can have an impact on meteorological conditions, which can change air stability and mixing and affect how pollutants are transported and dispersed.

**REFERENCES**

1. C. Zhan, S. Li, J. Li, Y. Guo, Q. Wen and W. Wen, "Prediction of Air Quality in Major Cities of China by Deep Learning," 2020 16th International Conference on Computational Intelligence and Security (CIS), Guangxi, China, 2020, pp. 68-72, doi: 10.1109/CIS52066.2020.00023
2. C. Zhan, S. Li, J. Li, Y. Guo, Q. Wen and W. Wen, "Prediction of Air Quality in Major Cities of China by Deep Learning," 2020 16th International Conference on Computational Intelligence and Security (CIS), Guangxi, China, 2020, pp. 68-72, doi: 10.1109/CIS52066.2020.00023.
3. R. Bao, Y. Zhou, and W. Jiang, "FL-CNN-LSTM: Indoor Air Quality Prediction Using Fuzzy Logic and CNN-LSTM Model," 2022 2nd International Conference on Electrical Engineering and Control Science (IC2ECS), Nanjing, China, 2022, pp. 986-989, doi: 10.1109/IC2ECS57645.2022.10088050.
4. ] K. Manikandan, R. Rajkumar, P. Jayanthi, S. Jothimani, R. Venkatesh and S. Govinda Rao, "An Intelligent Bayesian Optimization with Stacked BiLSTM Model for Air Quality Index Prediction," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 1699-1704, doi: 10.1109/ICIRCA57980.2023.10220650.
5. D. Ao, Z. Cui and D. Gu, "Hybrid model of Air Quality Prediction Using K-Means Clustering and Deep Neural Network," 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 8416-8421, doi: 10.23919/ChiCC.2019.8865861.
6. ] K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara and B. Kuhaneswaran, "Comparative Analysis of Machine Learning Algorithms for Predicting Air Quality Index," 2021 From Innovation To Impact (FITI), Colombo, Sri Lanka, 2021, pp. 1-5, doi: 10.1109/FITI54902.2021.9833033.
7. J. -C. Cheng and H. -C. Peng, "Air Quality Forecast and Evaluation Based on Long Short-Term Memory Network and Fuzzy Algorithm," 2021 IEEE 4th International Conference on Knowledge Innovation and Invention (ICKII), Taichung, Taiwan, 2021, pp. 88-91, doi: 10.1109/ICKII51822.2021.9574707.
8. W. Wang and S. Yang, "Research on Air Quality Forecasting Based on Big Data and Neural Network," 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2020, pp. 180-184, doi: 10.1109/ICCNEA50255.2020.00045.
9. W. Wang and S. Yang, "Research on Air Quality Forecasting Based on Big Data and Neural Network," 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2020, pp. 180-184, doi: 10.1109/ICCNEA50255.2020.00045.
10. C. R. K, N. R. K, B. P. K and P. S. Rajendran, "The Prediction of Quality of the Air Using Supervised Learning," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICCES51350.2021.9488983.
11. A. Akanksha, N. Maurya, M. Jain and S. Arya, "Prediction And Analysis of Air Pollution Using Machine Learning Algorithms," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-6, doi: 10.1109/CONIT59222.2023.10205615.
12. A. Catovic, E. Kadusic, C. Ruland, N. Zivic and N. Hadzajlic, "Air pollution prediction and warning system using IoT and machine learning," 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 2022, pp. 1-4, doi: 10.1109/ICECCME55909.2022.9987957.
13. S. Samadi and A. K. Kumawat, "IoT Enabled Low-Cost Real-Time Remote Air Quality Monitoring and Forecasting System," 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 2023, pp. 253-258, doi: 10.1109/DICCT56244.2023.10110229.

*.*