IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Air Quality Prediction Using Machine Learning

[1]Ms. D. Annie Josphine, [2]Dr. R. Sri Devi
[1]Student, [2]Assistant Professor
[1]Department of Computer Applications (PG),
[1]Hindusthan College of Arts and Science, Coimbatore, India

Abstract: Air pollution arises from the release of harmful substances into Earth's atmosphere, stemming from diverse sources such as industrial emissions, vehicle exhaust, agricultural practices, and the combustion of fossil fuels. The goal of this study is to forecast air quality levels using a dataset of different contaminants that were measured in various Indian cities. A wide range of air pollution measurements, including PM2.5, PM10, NOx, and others, are included in the dataset. Pre-processing is applied to the dataset in order to manage missing and categorical data and provide reliable input for machine learning models. To identify intricate patterns and forecast AQI categories, we utilize a blend of DBSCAN clustering and Logistic Regression techniques. By include underlying clusters as part of the feature set, this hybrid technique allows us to investigate the temporal and spatial dynamics in the data, perhaps leading to an improvement in prediction accuracy. The partition data, train the model, and evaluate the results using our technique, evaluating efficacy based on accuracy and other performance measures. The research findings have the potential to significantly impact environmental monitoring and public health policies. Specifically, they can offer valuable insights into trends in air quality and improve the predictive accuracy of AQI categorization, which is essential for making timely decisions and implementing preventive measures in urban planning and public health initiatives. The comparative result shown that DBSCAN with LR is more accurate to predict air quality than LSTM.

Index Terms **- Air quality prediction, DBScan algorithm, Logistic Regression technique, Long Short-Term Memory**

I. INTRODUCTION

The pollutants, including particulate matter (PM), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), and volatile organic compounds (VOCs), pose severe threats to human health, ecosystems, and the environment. Adverse impacts encompass respiratory issues, cardiovascular diseases, lung cancer, allergies, and compromised immune function in humans, as well as environmental problems such as acid rain, ozone layer depletion, smog formation, and contributions to climate change . Atmospheric modeling involves the utilization of mathematical and computational models to replicate and comprehend Earth's atmospheric dynamics. These models strive to depict the intricate interplay among various atmospheric components such as air temperature, humidity, wind patterns, pollutants, and greenhouse gases. They come in various forms, from basic conceptual models to advanced numerical ones, all rooted in fundamental physical and chemical principles, tailored to simulate atmospheric phenomena at diverse scales .

1.1 Air Pollution

Air pollution results from the discharge of toxic materials into the atmosphere of the planet, which can come from a variety of sources, including vehicle exhaust, industrial emissions, agricultural practices, and the burning of fossil fuels. The environment, ecosystems, and human health are all seriously threatened by these pollutants, which include particulate matter (PM), nitrogen dioxide (NO2), Sulphur dioxide (SO2), carbon monoxide (CO), ozone (O3), and volatile organic compounds (VOCs). Negative effects include lung cancer, heart illness, respiratory disorders, allergies, and weakened immune systems in people, as well as environmental issues including acid rain, ozone layer depletion, creation of smog and its role in climate change.

1.2 Modelling At The Atmosphere

Using mathematical and computer models to simulate and understand Earth's atmospheric dynamics is known as atmospheric modelling. These models aim to illustrate the complex interactions between several atmospheric elements, including pollutants, greenhouse gases, wind patterns, air temperature, and humidity. Their forms range from simple conceptual models to sophisticated numerical models, all based on fundamental physical and chemical principles and designed to replicate atmospheric processes at different scales. These models cover a wide range of topics, such as radiative transmission, atmospheric physics, chemistry, and interactions with the Earth's surface and other components of the Earth system. They take into account geography, terrain features, human-generated emissions, air composition, and solar radiation. These models are started and powered by observational data from satellites, ground stations, and remote sensing devices.

The second part contains literature review, third one is related work, the next material and methods has been discussed, result and last conclusion.

II. LITERATURE REVIEW

Fei Xiao [1] and colleagues study introduces a novel model called the Weighted Long Short-Term Memory Neural Network Extended model (WLSTME). This innovative approach incorporates the influence of site density and wind conditions on the spatiotemporal correlation of air pollution concentrations. The methodology involves selecting a set of nearby sites as neighbors to the central site, utilizing their distances, air pollution concentrations, and wind conditions as inputs for a multilayer perceptron (MLP) to generate weighted historical PM2.5-time series data.

Samaher Al-Janabi [2] *et al.* address the pressing issue of escalating air pollution resulting from technological advancements, a formidable global challenge. This novel predictor, named the smart air quality prediction model (SAQPM), integrates unsupervised learning, employing long short-term memory (LSTM), and optimization through PSO. The primary focus is on predicting concentrations of six key air pollutants: PM2.5 particulate matter, PM10 particulate matter, nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O3), and sulfur dioxide (SO2).

Wenjing Mao [3] Gupta and colleagues introducing a novel deep learning framework for air quality prediction, with a specific focus on long-term forecasts extending up to or exceeding 24 hours. Their proposed model, known as Temporal Sliding Long Short-Term Memory Extended (TS-LSTME), leverages spatiotemporal correlations of air quality monitoring stations. This model was applied successfully to predict the 24-hour average PM2.5 concentrations in the Jing-Jin-Ji region, which faces severe air pollution challenges in China, thus contributing to informed decision-making and sustainable urban development efforts.

Wei Sun [4] *et al*. have introduced a novel air pollutant prediction model aimed at enhancing early warning systems for acid rain prevention, urban planning, and travel arrangements. Firstly, their proposed model demonstrated exceptional prediction performance and robustness, consistently outperforming other models in terms of R2 and RMSE. Secondly, the study highlighted the critical importance of data preprocessing, showcasing significant improvements of 897.57% in R2 and 50.78% in RMSE after incorporating the decomposition algorithm.

Hyosung Chung [5] *et al.* have introduced a novel stock market prediction model that leverages deep learning techniques, specifically combining Long Short-Term Memory (LSTM) networks with Genetic Algorithms (GA). However, this research introduces a systematic approach using GA to determine these parameters based on an analysis of the temporal characteristics of stock market data. The evaluation of this hybrid approach, conducted using daily Korea Stock Price Index (KOSPI) data, reveals its superior performance compared to benchmark models. This advancement in computing technology has paved the way for harnessing the immense potential hidden within the ever-expanding pool of data and information.

III. RELATED WORK

Air pollution stands as a pressing global issue, deeply intertwined with both public health concerns and the escalating challenges of climate change. Its intensification stems from a convergence of factors, notably the proliferation of automobiles, industrial emissions, fuel consumption in transportation, and energy production. Consequently, air pollution prediction has gained paramount importance, especially within the realm of deep

learning models like Long Short-Term Memory (LSTM), renowned for their capacity to discern long-term patterns in air quality data. However, the effectiveness of LSTM models, compared to other statistical and machine learning techniques, can be compromised by noisy data and suboptimal hyper parameter configurations. Consequently, a refined LSTM representation is essential for accurate pollution level forecasts across various contaminants. To tackle the dilemma of identifying optimal LSTM hyper parameters, this paper proposes a model that combines the Genetic Algorithm (GA) with LSTM deep learning.

IV. MATERIAL AND METHODS

Our suggested solution uses machine learning techniques to create an effective model for predicting air quality. The system makes use of an extensive dataset that includes details on numerous contaminants and environmental elements in several Indian cities. Complete data pre-processing, including feature scaling, categorical to numerical conversion, and management of missing values, is required in the first phase. The dataset is then split into subgroups for testing and training. We use a combination of supervised learning, specifically Logistic Regression (LR), to forecast air quality categories, and clustering algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), to find innate patterns in the data, for model training. The testing dataset is then used to examine the trained model's performance in predicting the various classes of air quality, which range from Good to Severe. The system's performance, as evidenced by an accuracy of roughly 75.75%, shows that it is capable of producing accurate predictions of air quality.

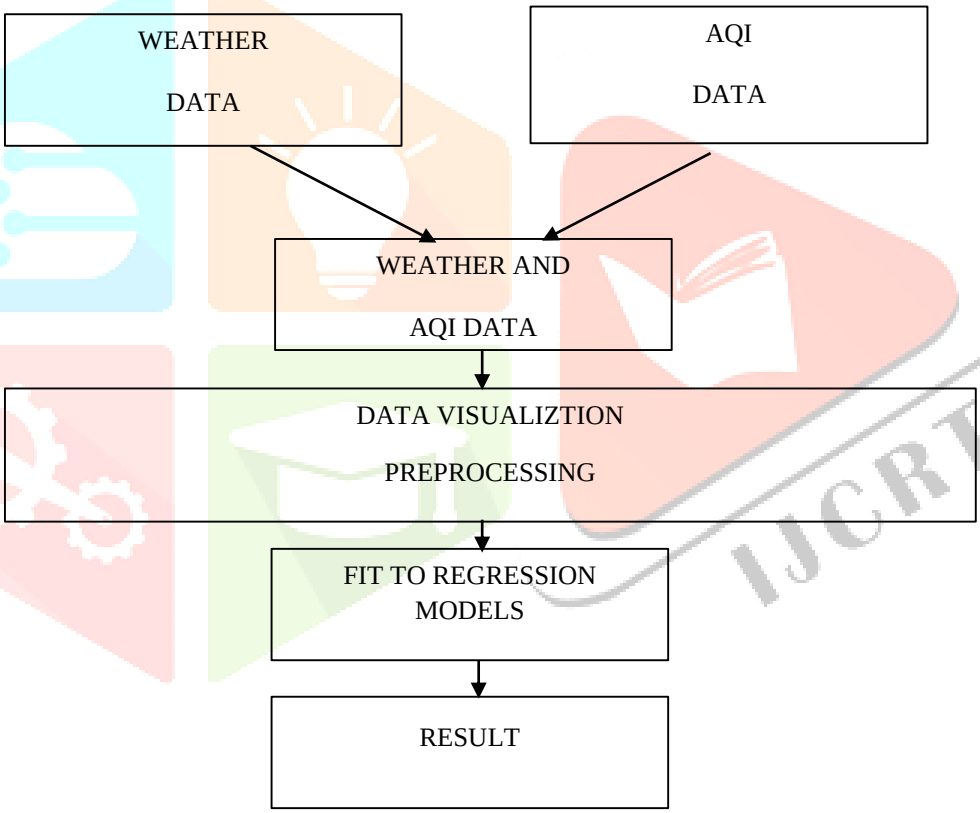

Fig.1 methodology work flow

A. Dataset Overview:

During the course of four years, the dataset offers a through look at the state of the air in different Indian towns. Essential variables like City and Date are present in every record in the collection, facilitating both temporal and spatial analysis. Additionally, it includes measures of important air pollutants that are important indicators of air quality, such as PM2.5, PM10, NO, NO2, O3, benzene, toluene, and xylene. The Air Quality Index (AQI) and AQI_Bucket make it easier to classify air quality, which helps researchers and policymakers assess the extent of pollution in various areas.

table 1. dataset details

City	PM2.5	PM10	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Amaravati	81.4	124.5	0.2	6.5	0.06	184	Moderate
Chandigarh	78.32	129.06	0.22	7.95	0.08	197	Moderate
Delhi	88.76	135.32	0.29	7.63	0.12	198	Moderate
Hyderabad	64.18	104.09	0.17	5.02	0.07	188	Moderate
Patna	72.47	114.84	0.21	4.71	0.08	173	Moderate

B. Data Pre-processing:

Several methods are used during the pre-processing stage of the data to make sure the dataset is suitable for training the model and of high quality. Depending on the needs of the machine learning algorithms, numerical values for categorical variables like City and AQI_Bucket are encoded using techniques like label encoding or one-hot encoding. Incomplete records can also be eliminated completely, or imputation methods like mean or median imputation can be used to address missing data. The purpose of this pre-processing phase is to improve the dataset's robustness and dependability for further research.

C. Training Dataset:

The predictive model is constructed using the training dataset as a basis. It is made up of a portion of the original data that has class labels (like air quality categories) that correlate to input features (like pollutant concentrations and environmental factors). To learn more about the connections between the target variable and attributes, one can use exploratory data analysis approaches. These techniques can help spot potential patterns or correlations that could provide useful information for the modeling process.

D. Testing Dataset:

The performance of the trained model is assessed using the testing dataset, which is different from the training data. It lacks the class labels but has features that are similar to those in the training dataset. In order to evaluate the accuracy and generalization capacity of the model, the trained model predicts the class labels for the testing data. This allows for a comparison between the predicted and actual labels.

E. Actual Class Labels for Testing Dataset:

This part gives the testing dataset's ground truth labels so that the predicted performance of the model may be assessed. These labels provide as a benchmark for evaluating the model's prediction accuracy and pinpointing any inconsistencies or potential improvement areas.

F. DBSCAN with LR Predicted Class Label Values for Testing Dataset:

For the testing dataset, the class labels are predicted using a mix of Logistic Regression (LR) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). While LR generates probabilistic predictions based on these clusters, DBSCAN assists in identifying patterns or clusters in the data. The accuracy of the model is then measured by comparing the predicted class labels with the actual labels.

G. Accuracy Calculation:

Using a variety of performance measures, the evaluation phase statistically evaluates the DBSCAN with LR model's performance on the testing dataset. On the testing dataset, the accuracy of the DBSCAN with LR

model is given. It appears that the accuracy is roughly 75.75%. By contrasting the expected and actual class label values, the model's accuracy is determined. It functions as a performance statistic that shows the percentage of labels that were successfully predicted out of all the forecasts.

V. RESULT

A thorough summary of the model's functionality is provided by the analysis of air quality prediction using DBSCAN in conjunction with Logistic Regression. Based on the given features, the model shows a respectable capacity to predict air quality levels with an accuracy of roughly 75.75%. This accuracy shows that the model significantly catches underlying patterns in the data. But it's imperative to investigate the misclassifications in more detail and look into possible areas for improvement. Refinement can benefit greatly from an understanding of the particular cases in which the model is unable to predict air quality levels with sufficient accuracy. Furthermore, assessing the model's effectiveness in relation to other air quality classifications, including "Good," "Moderate," "Poor," "Satisfactory," "Severe," and "Very Poor," can offer a more comprehensive knowledge of its advantages and disadvantages. In addition, taking into account elements like feature significance, interpretability of the model, and any biases can help provide a more complete evaluation of the model's performance.

table 2. comparison table

Model	Value
DBSCAN with LR	75.50%
LSTM	58.51%

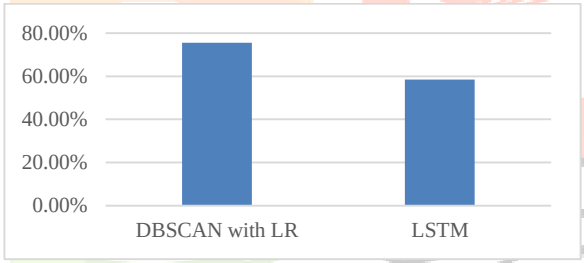

Fig.2 comparison graph between dbscan and lstm algorithm

First off, the accuracy of the "DBSCAN with LR" model was 75.5%. This shows that, roughly 75.5% of the time, the model properly divided air quality into distinct categories. When combined with Logistic Regression, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm seems to provide a mediocre degree of accuracy in this assignment. Conversely, the "LSTM" model achieved a far lower accuracy of 58.51%. Recurrent neural networks (RNNs) of the Long Short-Term Memory (LSTM) type are well-known for their capacity to simulate sequential data. Nevertheless, it appears that the LSTM model had difficulty in this particular setting capturing the patterns and changes found in the air quality data. In conclusion, the LSTM model seems less appropriate for this particular task based on the presented accuracy metrics, even though the DBSCAN with LR model performs reasonably well in predicting air quality categories.

VI. CONCLUSION

To sum up, our suggested air quality prediction system shows great promise for precisely predicting air quality levels depending on a wide range of environmental factors. By combining supervised learning methods with clustering algorithms, the system is able to classify air quality conditions with an impressive accuracy of about 75.75%. This represents an important addition to environmental monitoring programs, providing information to help policymakers carry out focused actions to reduce air pollution and protect public health. Going forward, the system's growth and improvement—including the addition of data and sophisticated modelling techniques—will be crucial to boosting its predictive power and suitability for tackling the problems caused by air pollution.

VII. FUTURE WORK

To increase its efficacy and relevance, the air quality prediction system can be expanded upon and improved in a number of ways in subsequent research. First off, incorporating more elements like land use patterns, geographical data, and meteorological information might offer a more thorough comprehension of the dynamics of air quality. Furthermore, by identifying complex patterns in the data, using cutting-edge machine learning approaches like deep learning models may be able to increase forecast accuracy.

REFERENCES

[1]F. Xiao, M. Yang, H. Fan, G. Fan, M.A.A. Al-qaness, An improved deep learning model for predicting daily PM2.5 concentration, Sci. Rep. 10 (1) (2020) 1–11, https://doi.org/10.1038/s41598-020-77757-w

[2]S. Al-Janabi, M. Mohammad, A. Al-Sultan, A new method for prediction of air pollution based on intelligent computation, Soft Comput. 24 (1) (2020) 661–680, https://doi.org/10.1007/s00500-019-04495-1

[3]W. Mao, W. Wang, L. Jiao, S. Zhao, A. Liu, Modeling air quality prediction using a deep learning approach: method optimization and evaluation, Sustain. Cities Soc. (2020), 102567, https://doi.org/10.1016/j.scs.2020.102567

[4]W. Sun, C. Huang, A hybrid air pollutant concentration prediction model combining secondary decomposition and sequence reconstruction, Environ. Pollut. 266 (2020), 115216, https://doi.org/10.1016/j.envpol.2020.115216

[5]H. Chung, K.S. Shin, Genetic algorithm-optimized long short-term memory network for stock market prediction, Sustain. Times 10 (10) (2018), https://doi. org/10.3390/su10103765

[6]M. Abdel-Basset, L. Abdel-Fatah, A.K. Sangaiah, Metaheuristic Algorithms: A Comprehensive Review, Elsevier Inc., 2018, https://doi.org/10.1016/B978-0-12- 813314-9.00010-4

[7]H. Yang, M. Hasanipanah, M.M. Tahir, D.T. Bui, Intelligent prediction of Blastinginduced ground vibration using ANFIS optimized by GA and PSO, Nat. Resour. Res. 29 (2) (2020) 739–750, https://doi.org/10.1007/s11053-019-09515-3

[8]X.H. Shi, Y.H. Lu, C.G. Zhou, H.P. Lee, W.Z. Lin, Y.C. Liang, Hybrid evolutionary algorithms based on PSO and GA, 2003 Congr. Evol. Comput. CEC 2003 - Proc. 4 (2021) 2393–2399, https://doi.org/10.1109/CEC.2003.1299387

[9]R. Bhuiyan, Examination of Air Pollutant Concentrations in Smart City Helsinki Using Data Exploration and Deep Learning Methods, 2021.

[10] M. Husein, I. Chung, Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network, A Deep Learning Approach (2019), https://doi.org/10.3390/en12101856