# Review of Air Pollution Hotspots Detection and Identifying the Source Trajectories using ML techniques

Ashwini Koshta
Department of Computer Engineering
*Modern Education Society's College of Engineering*
19, Late Principal V. K. joag, Path
Wadia College Campus Pune - 411001
Savitribai Phule Pune University
Pune, Maharashtra

Asma Bekinalkar
Department of Computer Engineering
*Modern Education Society's College of Engineering*
19, Late Principal V. K. joag, Path
Wadia College Campus Pune - 411001
Savitribai Phule Pune University
Pune, Maharashtra

Diksha Tickoo
Department of Computer Engineering
*Modern Education Society's College of Engineering*
19, Late Principal V. K. joag, Path
Wadia College Campus Pune - 411001
Savitribai Phule Pune University
Pune, Maharashtra

Liza Souza
Department of Computer Engineering
*Modern Education Society's College of Engineering*
19, Late Principal V. K. joag, Path
Wadia College Campus Pune - 411001
Savitribai Phule Pune University
Pune, Maharashtra

Prof. Shobha Raskar
Department of Computer Engineering
*Modern Education Society's College of Engineering*
19, Late Principal V. K. joag, Path
Wadia College Campus Pune - 411001
Savitribai Phule Pune University
Pune, Maharashtra

Prof. Jaya Mane
Department of Computer Engineering
*Modern Education Society's College of Engineering*
19, Late Principal V. K. joag, Path
Wadia College Campus Pune - 411001
Savitribai Phule Pune University
Pune, Maharashtra

***Abstract*— - One of our era's greatest scourges is air pollution, on account not only of its impact on climate change but also its impact on public and individual health due to increasing morbidity and mortality. Time series AQI data is collected through the CPCB sensors in different stations all over India. Classification of hotspots is done using SVM, and the time series analysis based on pollutants like PM2.5, PM10, CO, NO data samples is done using LSTM, ARIMA and SARIMA. Pollution levels of a day in the future are predicted using the said models. This review paper focuses on the various techniques used for prediction or modeling of Air Quality Index (AQI) and forecasting of future concentration levels of pollutants that may cause the air pollution so that governing bodies can take the actions to reduce the pollution.**

***Keywords—Air Pollution, PM2.5, SVM, ARIMA, SARIMA, LSTM, time-series analysis, Air quality prediction, AQI.***

## I. INTRODUCTION

The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will not evaporate into the atmosphere. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea; Particulate matter with a diameter 2.5 micrometer or less than that affects more to human health. Measures must be taken to minimize air pollution in the environment.

Because of new inventions there is a rapid increase in the development, serious population growth and increased number of vehicles will give rise to so many critical problems related to the environment such as acid rain, deforestation, air pollution, water pollution, emission of toxic materials and so on. To fulfill the needs of the growing population there is the drastic increase in industrialization that may lead to the emission of harmful gases in the atmosphere from various industries that will cause the serious air pollution problem in urban areas throughout the world. This means that the air we or people breathe is not clean air but it is polluted as so many harmful gases and particles are present in the air that adversely affect human health. The quality of air degrades due to the pollution.

In most of the urban areas the air pollution becomes a serious concern. The people should know about the air they breathe. The National Ambient Air Monitoring Network generates the data that includes the concentration of various pollutants present in the air but this data is not easily understood by the

common people. So the Central Pollution Control Board (CPCB) develops the national Air Quality Index (AQI) for the cities in India [1]. AQI gives the idea about quality of air or to what extent the air in the particular location is polluted. This means that AQI gives the actual quality of air around us in the qualitative form that is linked with various health impacts.

This paper aims to review the articles related to air pollution prediction using machine learning techniques, to make a comparison of methodologies that different authors have used and to get an overall idea about applied approaches. The usage of machine learning techniques in this area has begun to be actively developed, and many studies and observations have been done, which is conditioned by the importance of the field. The combination of all the information will help us to detect the tendency, to find out the innovations applied in the research area, which, in turn, will direct and guide us for future exploration. India ranks 3th globally among 106 countries in air pollution [2]. So it is very important to study about the air quality of India and make policies to reduce the air pollution.

### ABBREVIATIONS AND ACRONYMS

API: Application Program Interface, CPCB: Central Pollution Control Board, AQI : Air Quality Index, SVM : Support Vector Machine, ANN : Artificial Neural Network, LSTM : Long-Short-Term-Memory, ALSTM: Adaptive Long-Short-Term-Memory, ARIMA : Autoregressive Integrated Moving average, SARIMA: Seasonal Autoregressive Integrated Moving average.

## II. PROBLEM STATEMENT

In this paper, we have discussed methods to collect and analyze air pollution data in India to detect air pollution hotspots and predict pollution levels of a particular area for a day in the future.

Identifying Yearly Hotspots by calculating AQI from major pollutants among all States and UTs of India. An ‘air pollution hotspot’ is an area where the pollution levels of the pollutants in that area rise beyond a threshold level for consecutive days at the same time. Continuous exposure to unsafe levels of air pollution is harmful and can cause long-term health problems for people resident in that area. This type of pollution hotspots is largely noticeable in urban areas, where there are multiple sources of air pollution like heavy traffic and industries.
Air Quality Index (AQI), is used to measure the quality of air. Each pollution has an individual index and scales at different levels. The major pollutants Such as (no2, so2, rspm, spm) indexes AQI is acquired, with this individual AQI, the data can be categorized based on the limits and Visualizing Source Trajectories of pollutants over time on Indian maps.

Prediction of future concentration of PM2.5 is predicted using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model which gives the increasing value of PM2.5 in next year and provides the lowest and highest predictions (more than 100 µg/m3 ). We will also learn about different Predicting Long-Term Pollution levels with Time Series Analysis using forecasting models like ARIMA (Autoregressive Integrated Moving Average) and LSTM (Long Short-term Memory).

## III. LITERATURE SURVEY

**Prediction of air pollution using machine learning approaches on the cloud (IEEE 2018),** in this paper [10], air pollution data, specifically particulate matter of less than 2.5 micrometers (PM2.5) was collected from a variety of web-based resources and following, data cleaning analyzed with different ML models including linear regression, ANN and LSTM recurrent neural network. The NeCTAR research cloud was used for training both ANN and LSTM models. Tensorflow was used as the basis for these. It was found that LSTM performed best and was able to predict high PM2.5 values with reasonable accuracy. ANN and linear models have drawbacks in prediction of high PM2.5 values however they offered reasonable overall performance.

**Air Pollution Hotspot Identification and Pollution Level Prediction in the City of Delhi (IEEE 2020), i**n this paper [3], authors use various methods and algorithms to detect air pollution hotspots and predict pollution levels in a selected area in the city of Delhi. Time series AQI data is collected through the CPCB sensors in Delhi. Classification of hotspots is done using SVM, and the time series analysis based on pollutants like PM2.5, PM10, CO, NO data samples is done using LSTM and PROPHET. Pollution levels of a day in the future are predicted using the said models.

From the results [3], it can be seen that the algorithms which have been used, can detect the time of occurrence of the pollution hotspots, the daily and monthly seasonality, and also provides a forecast of the pollution levels of a particular area. So, if sufficient data can be collected, multiple SVM models can be used to identify hotspots across the city. The Prophet model further helps to determine the daily and monthly seasonality of the pollution levels. These results can be helpful when the pollution levels in the area are taken into consideration along with the congestion amounts while rerouting the vehicles. The forecasting properties of LSTM and Prophet are used to get a forecast for a given day from the collected data to get an idea about the pollution level of that day in the future.

**An LSTM based aggregated model for air pollution forecasting (Elsevier 2020),** in this paper [13], the authors aggregate three LSTM models into a predictive model for early predictions based on external sources of pollution and information from nearby industrial air quality stations. They

exploited the data with 17 attributes collected by Taiwan Environmental Protection Agency from 2012 to 2017 as the training data to build the ALSTM forecasting model, and then tested the model using the data collected in 2018. They conducted some experiments to compare our new ALSTM model with SVR (Support Vector Machine based Regression), GBTR (Gradient Boosted Tree Regression), LSTM, etc., in the prediction of PM2.5 for 1–8 h, and evaluated them using various assessment techniques, such as MAE, RMSE, and MAPE. The results reveal that the proposed aggregated model can effectively improve the Accuracy of prediction.

The main objectives of this research work were:

*1)* In Aggregated LSTM model, three aggregation-learning models have been used, (i) local characteristics (ii) neighborhood features (iii) and overseas characteristics. It produces three predictive characteristics for various station types. The data is generated with predicted functional data from the fully connected LSTM and the system trains data on an ongoing basis and reverse propagation adjust weights after each batch. The best results can be ultimately obtained.

*2)* Building a deep neural network model with Tensorflow and Keras and provide predictive data on the real-time PM2.5 concentration over the coming 8 h.

*3)* Three sub-neural LSTM networks are established: local features, neighboring station features of industrial regions, and chimney and other features from abroad, to generate three different types of feature data. A wider sub-neural network is established to learn the characteristics of air pollution sources from different sources through the aggregated model combined with the LSTM neural network.

**Air Pollution Prediction Using Machine Learning Supervised Learning Approach (IJSTR 2020),** in this paper [12], the Air pollutants information is retrieved from the sensors which are processed in a unified schema and stored as a dataset. This dataset is preprocessed with different functionalities such as normalization, attribute selection and discretization. Once the dataset is ready, it is splitted into training dataset and test dataset. And further Supervised Machine Learning Algorithms are applied on the training dataset. The obtained results are matched with the testing dataset and results are analyzed. The air pollution prediction using Supervised Machine Learning approach considers four machine learning algorithms such as LR, SVM, DT, and RF. Logistic Regression (LR), Decision Tree, Support Vector Machine (SVM), Random Forest were used to predict PM2.5 values. This research has shown that Random Forest gives 0.84 deviations from actual results, Decision Tree gives 1.34, SVM has 3.89 and Linear Regression proves to give 6.01 deviations from actual results. Thus, compared to all other algorithms Random forest gives better results.

**Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach (IEEE 2021**), In this [14], authors proposed model in view of the current problems in the study of atmospheric PM2.5 in Lahore City (combined with the research on the analysis of related meteorological trajectories), this study uses theoretical knowledge such as environmental science, atmospheric chemistry, meteorology and geochemistry, from September 2014 to 2019. During this period, samples of PM2.5 in Lahore city were collected, and the pollution characteristics of different pollutants with PM2.5 and PM10 were preliminarily analyzed. At the same time, online trajectory sources were used to analyze and quantify the contribution of different pollution sources to ions in atmospheric particles, combined with HYSPLIT backward airflow.

**Machine Learning algorithms for air pollutants forecasting (IEEE 2020),** In this paper [11], the authors presented several Machine Learning algorithms, the possible software that can be used for them and the applications used in the field of air quality. Based on the research in the field, they proposed SVR, ARIMA and LSTM, 3 Machine Learning models, which can be used to predict air pollution. These algorithms have been tested using time-series for PM10 and PM2.5 particles. It was demonstrated that SVR and ARIMA algorithms are the most suitable in forecasting the air pollutants concentrations, because the correlation coefficient was 0.966 and 0.921 respectively for PM10 concentration. The results showed that SVR and ARIMA algorithms are the most suitable in forecasting air pollutant concentrations.

## IV. DISCUSSION

The atmospheric environment system is a system with both complexity and variability, a huge amount of monitoring data has been accumulated in the past few decades, a traditional prediction models are difficult to capture effective information from a large amount of historical monitoring data, which leads to unsatisfactory prediction results. Therefore, in the study of air quality forecasting, the establishment of an effective air pollution forecasting model has become are liable tool to reduce the negative impact of environmental pollution on health and to formulate more complete prevention policies. In recent years, deep learning methods have been widely used in various time series forecasting problems. Among them, the time series models demonstrated its powerful time series processing capabilities. However, the use of attention mechanism to predict the air quality research has also nothing.

This paper proposes to build a time series forecasting model using LSTM, ARIMA and SARIMA and apply it to the air quality forecasting field. The forecasting properties of LSTM are used to get a forecast for a given day from the collected data to get an idea about the pollution level of that day in the future. With its strong nonlinear processing ability and noise

tolerance ability, it can realize the efficient forecast of air quality. From the results of research paper read it can be seen that the algorithms which have been used, can detect the time of occurrence of the pollution hotspots, the daily and monthly seasonality, and also provides a forecast of the pollution levels of a particular area. So, if sufficient data can be collected, multiple SVM models can be used to identify hotspots across the city. It has been studied that the impact of outdoor air pollution on the burden of disease in cities around the world is significant.

Therefore, to improve the air pollution situation, it is necessary to enhance the residents awareness of environmental protection, expand the area of urban green spaces, and reduce pollution emissions mainly from the industry and transportation industries. Industry needs to strengthen the structural adjustment of heavily polluting industries, strengthen source control, and promote the rationalization of industrial structure; In the transportation industry, priority is given to the development of public transportation to control the disorderly growth of the number of cars. Through data and analysis, we can see that the more developed areas are more polluted than the less developed areas, we cannot take the path of pollution first and then treat it and we must look at economic and environmental issues from a development perspective. Sustainable development is growth that meets contemporary people's needs without undermining future generations capacity to meet their needs. The current study is limited to the cities in India but in the future we will expand it to all the cities in different countries.

## V. ADVANTAGES

There are numerous benefits of air pollution monitoring system, that includes:

1. The data collected from air quality monitoring helps us assess impacts caused by poor air quality on public health.

2. Air quality data helps us determine if an area is meeting the air quality standards devised by CPCB, WHO or OSHA.

3. The data collected from air quality monitoring would primarily help us identify polluted areas, the level of pollution and air quality level.

4. Air quality monitoring would assist in determining if air pollution control programmes devised in a locality are working efficiently or not.

5. Air quality data helps us understand the mortality rate of any location due to air pollution. We can also access and compare the short term and long term diseases/disorders which are a result of air pollution.

6. Based upon the data collected control measures can be devised for protection of the environment and health of all living organisms.

## VI. CONCLUSION

The purpose of this review paper is to know in detail about the Air Quality Index (AQI) as AQI tells whether the air around us is polluted or not. It is important to know about AQI because unless and until the people know the worst impacts or hazards of air pollution they will not become that much aware about the air pollution and try to reduce it. As per this review most of the researchers worked on AQI and pollutants concentration level forecasting that will give the actual idea about AQI. LSTM, SVR, ARIMA and SARIMA are the choices of many researchers for the prediction of AQI and air pollutants concentration. High accuracies achieved with these machine learning algorithms explains it all why these algorithms should be preferred over traditional approaches.

## ACKNOWLEDGMENT

We would like to thank the Department of Computer Engineering, Modern Education Society's College of Engineering (Wadia Campus), Pune, for all the support provided for this research. We'd like to express our gratitude to Prof. S.S Raskar for sharing their pearls of knowledge with us during the research process.

## REFERENCES

[1] https://app.cpcbccr.com/ccr_docs/FINAL-REPORT_AQI_.pdf

[2] https://www.iqair.com/us/india

[3] Soumyadeep Sur, Rohit Ghoshal and Ritwik Mondal "Air Pollution Hotspot Identification and Pollution Level Prediction in the City of Delhi," 2020 IEEE International Conference for Convergence in Engineering.

[4] https://iopscience.iop.org/article/10.1088/1742-6596/1917/1/012029

[5] Athanasiadis, Ioannis N., et al. ``Applying machine learning techniques on air quality data for real-time decision support." First international NAISO symposium on information technologies in environmental engineering (ITEE 2003), Gdansk, Poland. 2003.

[6] https://en.wikipedia.org/wiki/Particulates

[7]http://aaqr.org/article/download?articleId=334&path=/files/article/334/10_AAQR-15-03-OA-0193_405-416.pdf

[8]https://www.hindawi.com/journals/jece/2017/5106045/abs/

[9] https://ieeexplore.ieee.org/abstract/document/7892954

[10] Ziyue Guan, Richard O. Sinnott, "Prediction of air pollution using machine learning approaches on the cloud", 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)

[11] Marius Dobrea, AndreeaBădicu, Marina Barbu, OanaȘubea, MihaelaBălănescu, GerogeSuciu, Andrei Bîrdici, OanaOrza and CiprianDobre, "Machine Learning algorithms for air pollutants forecasting," 2020 IEEE 26th International Symposium for Design and Technology in Electronic Packaging (SIITME).

[12] Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar, "Air Pollution Prediction Using Machine Learning Supervised Learning Approach", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 04, APRIL 2020.

[13] Yue-Shan Changa, Hsin-Ta Chiaob, Satheesh Abimannanc, Yo-Ping Huangd, Yi-Ting Tsaia, Kuan-Ming Lina, "An LSTM based aggregated model for air pollution forecasting", Elsevier 2020.

[14] UZAIR ASLAM BHATTI 1 , YUHUAN YAN2 , MINGQUAN ZHOU2,3, SAJID ALI4 , AAMIR HUSSAIN5 , HUO QINGSONG2 , ZHAOYUAN YU1 , AND LINWANG YUAN1, "Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach", IEEE 2021.