RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES
Social Science Journal

ADVANCING AIR QUALITY PREDICTION THROUGH ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS

P. Manjulatha[1*], V. Kaveri[1], G. Latha[1], P. Jaya Simha[1]

[1]Department of Computer Science and Engineering (AI & ML), Sree Dattha Institute of Engineering and Science, Hyderabad, Telangana, India

[*]Corresponding E-mail: manju.pidugu@sreedattha.ac.in

Abstract

In recent years, there have been substantial breakthroughs in the prediction and analysis of air quality. Previously, our dependence was significant on conventional approaches such as statistical models and simple equations. Nevertheless, these methodologies encountered difficulties in comprehending the intricate and ever-changing characteristics of air pollution. With the advancement of technology, scientists and researchers have utilized AI, machine learning, and big data analytics to enhance air quality predictions. Conversely, air pollution is a crucial worldwide problem that impacts both our surroundings and our physical and mental health. Additionally, it is associated with respiratory and cardiovascular ailments, resulting in a rise in morbidity and mortality. Precise air quality forecasts enable governments, local authorities, and citizens to promptly respond to pollution, protect public health, and optimize urban development. In order to address this urgent issue, it is imperative that we have precise air quality forecasting and examination. The development of this AI model is driven by the constraints of conventional air quality forecast approaches. It has been observed that these methods frequently lack precision and face difficulties in considering the complex components that influence air pollution. The potential of AI lies in its capacity to analyze extensive quantities of up-to-date data and detect intricate patterns, providing a possible option to improve the precision and dependability of air quality forecasts. This paper presents a novel Artificial Intelligence (AI) model that is specifically developed to accurately and efficiently predict and analyze air quality. This model seeks to address the increasing need for accurate real-time air quality data by utilizing advanced AI algorithms and data analytics approaches.

Keywords: Environmental Sustainability, Air quality prediction, Artificial Intelligence (AI), Data analytics, Environmental monitoring, Pollution control

1. Introduction

Energy consumption and its consequences are inevitable in modern age human activities. The anthropogenic sources of air pollution include emissions from industrial plants; automobiles; planes; burning of straw, coal, and kerosene; aerosol cans, etc. Various dangerous pollutants like CO, CO2, Particulate Matter (PM), NO2, SO2, O3, NH3, Pb, etc. are being released into our environment every day. Chemicals and particles constituting air pollution affect the health of humans, animals, and even plants. Air pollution can cause a multitude of serious diseases in humans, from bronchitis to heart disease, from pneumonia to lung cancer, etc. Poor air conditions lead to other contemporary environmental issues like global warming, acid rain, reduced visibility, smog, aerosol formation, climate change, and premature deaths. Scientists have realized that air pollution bears the potential to affect historical monuments adversely [1]. Vehicle emissions, atmospheric releases of power plants and factories, agriculture exhausts, etc. are responsible for increased greenhouse gases. The greenhouse gases adversely affect climate conditions and consequently, the growth of plants [2]. Emissions of inorganic carbons and greenhouse gases also affect plant-soil interactions [3]. Climatic fluctuations not

only affect humans and animals, but agricultural factors and productivity are also greatly influenced [4]. Economic losses are the allied consequences too.

The Air Quality Index (AQI), an assessment parameter is related to public health directly. higher level of AQI indicates more dangerous exposure for the human population. Therefore, the urge to predict the AQI in advance motivated the scientists to monitor and model air quality. Monitoring and predicting AQI, especially in urban areas has become a vital and challenging task with increasing motor and industrial developments. Mostly, the air quality-based studies and research works target the developing countries, although the concentration of the deadliest pollutant like PM2.5 is found to be in multiple folds in developing countries [5]. A few researchers endeavoured to undertake the study of air quality prediction for Indian cities. After going through the available literature, a strong need had been felt to fill this gap by attempting analysis and prediction of AQI for India. Various models have been exercised in the literature to predict AQI, like statistical, deterministic, physical, and Machine Learning (ML) models. The traditional techniques based on probability, and statistics are very complex and less efficient. The ML-based AQI prediction models have been proved to be more reliable and consistent. Advanced technologies and sensors made data collection easy and precise. The accurate and reliable predictions through such huge environmental data require rigorous analysis which only ML algorithms can deal with efficiently.

2. Literature Survey

In [6], Gopalakrishnan (2021) combined Google's Street view data and ML to predict air quality at different places in Oakland city, California. He targeted the places where the data were unavailable. The author developed a web application to predict air quality for any location in the city neighborhood. Sanjeev [7] studied a dataset that included the concentration of pollutants and meteorological factors. The author analyzed and predicted the air quality and claimed that the Random Forest (RF) classifier performed the best as it is less prone to over-fitting. Castelli et al. [8] endeavoured to forecast air quality in California in terms of pollutants and particulate levels through the Support Vector Regression (SVR) ML algorithm. The authors claimed to develop a novel method to model hourly atmospheric pollution. Doreswamy et al. [9] investigated ML predictive models for forecasting PM concentration in the air. The authors studied six years of air quality monitoring data in Taiwan and applied existing models. They claimed that predicted values and actual values were very close to each other.

In [10], Liang et al. studied the performances of six ML classifiers to predict the AQI of Taiwan based on 11 years of data. The authors reported that Adaptive Boosting (AdaBoost) and Stacking Ensemble are most suitable for air quality prediction, but the forecasting performance varies over different geographical regions. Madan et al. [11] compared twenty different literary works over pollutants studied, ML algorithms applied, and their respective performances. The authors found that many works incorporated meteorological data such as humidity, wind speed, and temperature to predict pollution levels more accurately. They found that the Neural Network (NN) and boosting models outperformed the other eminent ML algorithms. Madhuri et al. [12] mentioned that wind speed, wind direction, humidity, and temperature played a significant role in the concentration of air pollutants. The authors employed supervised ML techniques to predict the AQI and found that the RF algorithm exhibited the least classification errors. Monisri et al. [13] collected air pollution data from various sources and endeavoured to develop a mixed model for predicting air quality. The authors claimed that the proposed model aims to help people in small towns to analyze and predict air quality.

Patil et al. [14] presented some literary works on various ML techniques for AQI modeling and forecasting. The authors found that Artificial Neural Network (ANN), Linear Regression (LR), and Logistic Regression (LogR) models were exploited by most of the scholars for AQI prediction. Bhalgat et al. [15] applied the ML technique to predict the concentration of SO2 in the environment of

Maharashtra, India. The authors concluded that being highly polluted, some cities of this Indian province require grave attention. The authors mentioned that their model was not capable of exhibiting expected outputs. Mahalingam et al. [16] developed a model to predict the AQI of smart cities and tested it in Delhi, India. The authors reported that the medium Gaussian Support Vector Machine (SVM) exhibited maximum accuracy. The authors claim that their model can be used in other smart cities too. Soundari et al. [17] developed a model based on NNs to predict the AQI of India. The authors claimed that their proposed model could predict the AQI of the whole county, of any province, or of any geographical region when the past data on concentration of pollutants were available. Sweileh et al. [18] came up with a very interesting study about the analysis of global peer-reviewed literature about air pollution and respiratory health. The authors extracted 3635 documents from the Scopus database published between 1990 and 2017. They observed that there was a substantial increase in publications from 2007 to 2017. The authors reported active countries, institutions, journals, authors, international collaborations in the realm and concluded that research works on air pollution and respiratory health had been receiving a lot of attention. They suggested securing public opinions about mitigation of outdoor air pollution and investment in green technologies.

3. Proposed Methodology

Air quality prediction using IoT sensor data is a critical application that leverages technology to monitor, assess, and forecast air quality conditions in various environments as shown in Figure 1. This process involves collecting real-time data from a network of IoT sensors deployed in different locations, analyzing this data, and using it to make predictions about air quality. In the process of collecting and managing data from IoT sensors, the information gathered is carefully stored within a centralized database or cloud-based platform. This data is marked with timestamps, providing details about the location of the sensors, the specific type of sensors used, and the actual measurements recorded. This meticulous record-keeping ensures that we have a comprehensive dataset to work with. Prior to delving into data analysis, there is a crucial step known as data preprocessing. During this phase, the data undergoes a series of operations aimed at refining it for further analysis. These operations include addressing missing data points, handling outliers, and, if necessary, converting data into standardized formats. This step ensures that the data is in its best possible condition for accurate analysis. Once the data is pre-processed, the next step involves feature engineering. Here, we extract and create relevant features from the raw sensor data.

Moving forward, this research employs Deep learning and statistical models to scrutinize the data and construct predictive models. These models draw insights from historical data, which is often used to train them. A range of algorithms, such as regression, time series analysis, or neural networks, can be applied in this context. The primary objective is to build models capable of forecasting future air quality conditions based on both historical patterns and the latest sensor readings. With the trained models in place, the proposed model equipped to make real-time predictions about upcoming air quality conditions, leveraging the most recent sensor readings. These predictions encompass a variety of valuable information, including AQI values, pollutant concentrations, and air quality forecasts tailored to specific time intervals, such as hourly or daily periods.

To ensure that the air quality information is accessible and comprehensible, it is often visualized through intuitive mediums like dashboards, maps, or graphs. This facilitates easy understanding for the general public, environmental agencies, and policymakers. Additionally, mechanisms are put in place to issue alerts and warnings in cases where air quality levels exceed safety thresholds. The insights drawn from air quality predictions can guide important actions and mitigation strategies. For instance, these predictions can inform decisions related to traffic management, industrial emissions control, and the issuance of health advisories to safeguard public well-being.

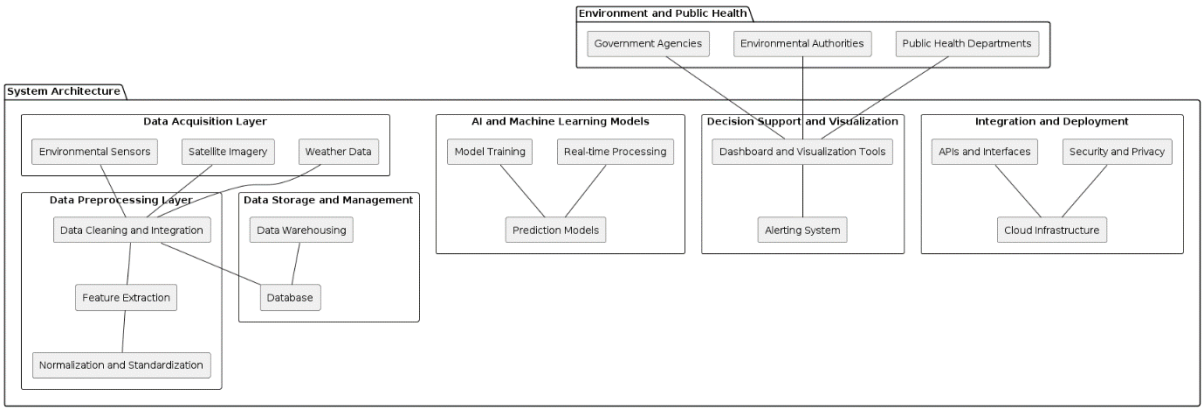

Figure 1. System Architecture

3.1 ANN Classifier

Although today the Perceptron is widely recognized as an algorithm, it was initially intended as an image recognition machine. It gets its name from performing the human-like function of perception, seeing, and recognizing images. Interest has been centered on the idea of a machine which would be capable of conceptualizing inputs impinging directly from the physical environment of light, sound, temperature, etc. — the "phenomenal world" with which we are all familiar — rather than requiring the intervention of a human agent to digest and code the necessary information. Rosenblatt's perceptron machine relied on a basic unit of computation, the neuron. Just like in previous models, each neuron has a cell that receives a series of pairs of inputs and weights. The major difference in Rosenblatt's model is that inputs are combined in a weighted sum and, if the weighted sum exceeds a predefined threshold, the neuron fires and produces an output.

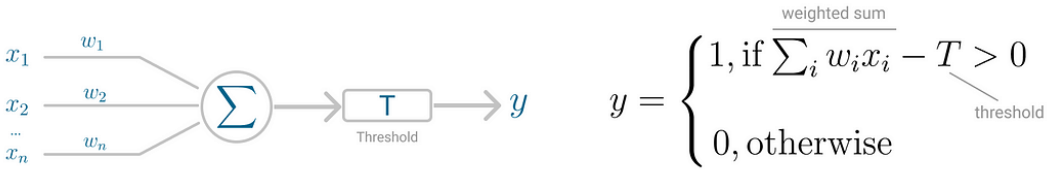

Figure 2: Perceptron neuron model (left) and threshold logic (right).

Threshold *T* represents the activation function. If the weighted sum of the inputs is greater than zero the neuron outputs the value 1, otherwise the output value is zero. With this discrete output, controlled by the activation function, the perceptron can be used as a binary classification model, defining a linear decision boundary.

It finds the separating hyperplane that minimizes the distance between misclassified points and the decision boundary. The perceptron loss function is defined as below:

$$\underbrace{D(w,c)}_{\text{distance}} = -\sum_{i \in M} \overbrace{y_i}^{\text{output}} (x_i w_i + c)$$

misclassified observations

To minimize this distance, perceptron uses stochastic gradient descent (SGD) as the optimization function. If the data is linearly separable, it is guaranteed that SGD will converge in a finite number of steps. The last piece that Perceptron needs is the activation function, the function that determines if the neuron will fire or not. Initial Perceptron models used sigmoid function, and just by looking at its shape,

it makes a lot of sense! The sigmoid function maps any real input to a value that is either 0 or 1 and encodes a non-linear function. The neuron can receive negative numbers as input, and it will still be able to produce an output that is either 0 or 1.

But, if you look at Deep Learning papers and algorithms from the last decade, you'll see the most of them use the Rectified Linear Unit (ReLU) as the neuron's activation function. The reason why ReLU became more adopted is that it allows better optimization using SGD, more efficient computation and is scale-invariant, meaning, its characteristics are not affected by the scale of the input. The neuron receives inputs and picks an initial set of weights random. These are combined in weighted sum and then ReLU, the activation function, determines the value of the output.

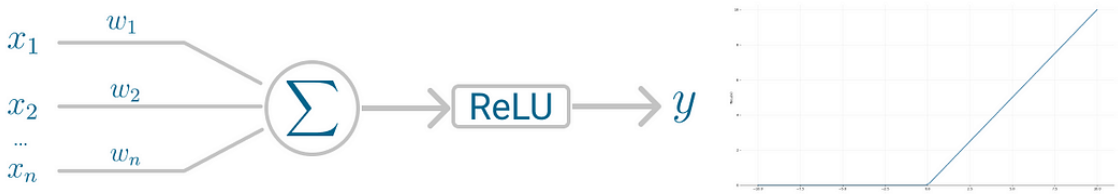

Figure 3: Perceptron neuron model (left) and activation function (right).

Perceptron uses SGD to find, or you might say learn, the set of weight that minimizes the distance between the misclassified points and the decision boundary. Once SGD converges, the dataset is separated into two regions by a linear hyperplane. Although it was said the Perceptron could represent any circuit and logic, the biggest criticism was that it couldn't represent the XOR gate, exclusive OR, where the gate only returns 1 if the inputs are different. This was proved almost a decade later and highlights the fact that Perceptron, with only one neuron, can't be applied to non-linear data.

4. Results and discussion

Figure 4 is a visualization for the classification of air quality based on the calculated AQI values. The classification likely involves different categories such as "good," "moderate," "poor," "unhealthy," "very unhealthy," and "hazardous." These categories indicate the level of pollution and associated health risks.

	state	SOi	Noi	Rpi	SPMi	AQI
0	Andhra Pradesh	6.000	21.750	0.0	0.0	21.750
1	Andhra Pradesh	3.875	8.750	0.0	0.0	8.750
2	Andhra Pradesh	7.750	35.625	0.0	0.0	35.625
3	Andhra Pradesh	7.875	18.375	0.0	0.0	18.375
4	Andhra Pradesh	5.875	9.375	0.0	0.0	9.375

Figure 4: Header of Air Quality Index calculated from every data value.

	state	location	type	so2	no2	rspm	spm	pm2_5	SOi	Noi	Rpi	SPMi	AQI	AQI_Range
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	0.0	0.0	0.0	6.000	21.750	0.0	0.0	21.750	Good
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	0.0	0.0	0.0	3.875	8.750	0.0	0.0	8.750	Good
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	0.0	0.0	0.0	7.750	35.625	0.0	0.0	35.625	Good
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	0.0	0.0	0.0	7.875	18.375	0.0	0.0	18.375	Good
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	0.0	0.0	0.0	5.875	9.375	0.0	0.0	9.375	Good

Good	219643
Poor	93272
Moderate	56571
Unhealthy	31733
Hazardous	18700
Very unhealthy	15823

Figure 5: Obtained classification of air quality as good, moderate, poor, unhealthy, very unhealthy, and Hazardous.

Table 1 provides a comparison of two different machine learning models used for air quality prediction based on two evaluation metrics: Root Mean Squared Error (RMSE) and R-squared (R^2) score. RMSE (Root Mean Squared Error): The RMSE is a metric used to measure the average magnitude of the errors between predicted values and actual (observed) values. It quantifies how well the predictions align with the actual data. A lower RMSE value indicates better predictive performance, as it means the model's predictions are closer to the actual values. From Table 1:

— For the "LR" model, the RMSE is 13.67.

— For the "ANN Model" model, the RMSE is 0.

A lower RMSE for the ANN Model suggests that it has smaller prediction errors compared to the LR model. R^2-score (Coefficient of Determination): The R^2 score is a statistical measure that represents the proportion of the variance in the dependent variable that's explained by the independent variables in a regression model. It ranges from 0 to 1, where higher values indicate that the model's predictions closely match the actual data. An R^2 score of 1 indicates a perfect fit. From Table 1:

— For the "LR" model, the R^2 score is 0.9847.

— For the "ANN" model, the R^2 score is 0.999999.

The R^2 scores for both models are quite high, indicating that they both provide excellent fits to the data. However, the ANN Model score of 0.999 suggests an almost perfect fit, meaning that it captures the variability in the data extremely well. Finally, the ANN Model outperforms the LR model in terms of both RMSE and R^2 score, indicating its superior predictive capability and ability to explain the variance in air quality data.

Table 1: Comparison of models.

Model name	RMSE	R^2-score
LR	13.67	0.9847
ANN	1.16	0.999

5. Conclusion

In the realm of air quality prediction, the use of Artificial Neural Networks (ANN) has been pivotal in providing valuable insights and forecasts. ANN models demonstrate significant advantages over traditional Linear Regression (LR) models, particularly in handling complex relationships and mitigating overfitting. While LR models are straightforward and interpretable, they often struggle to capture the nuances of complex, non-linear interactions within air quality data. On the other hand, ANN models, especially when leveraging deep learning techniques, exhibit superior performance by

automatically learning intricate relationships and interactions between various air quality parameters. The flexibility of ANN models allows them to adapt to diverse data patterns, making them well-suited for capturing the complex dynamics inherent in real-world air quality scenarios. Additionally, ANN models, particularly deep neural networks, can effectively handle large datasets with high dimensionality, which is often the case in air quality prediction tasks. Moreover, the ability of ANN models to generalize well to new data and their robustness against outliers further contribute to their superior performance. While ANN models have proven to be a formidable choice for air quality prediction, the field of air quality forecasting continues to evolve. Further exploration of feature engineering techniques, including the creation of novel features and the incorporation of additional environmental and meteorological data, can enhance the performance of ANN models. Additionally, the development of hybrid models that combine the strengths of ANN with other machine learning techniques, such as Random Forests or Gradient Boosting Machines, can lead to even more accurate and robust predictions.

References

[1] Rogers CD (2019) Pollution's impact on historical monuments pollution's impact on historical monuments. SCIENCING. https://sciencing.com/about-6372037-pollution-s-impact-historical-monum ents.html

[2] Fahad S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan, V (2021a) Plant growth regulators for climate-smart agriculture (1st ed.). CRC Press. https://doi.org/10.1201/9781003109013

[3] Fahad, S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021b) Sustainable soil and land management and climate change (1st ed.). CRC Press. https://doi.org/10.1201/9781003108894

[4] Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021) Climate change and plants: biodiversity, growth and interactions (S. Fahad, Ed.) (1st ed.). CRC Press. https://doi.org/10.1201/9781003108931

[5] Rybarczyk Y, Zalakeviciute R (2021) Assessing the COVID-19 impact on air quality: a machine learning approach. Geophys Res Lett. https://doi.org/10.1029/2020GL091202

[6] Gopalakrishnan V (2021) Hyperlocal air quality prediction using machine learning. Towards data science. https://towardsdatascie nce.com/hyperlocal-air-quality-prediction-using-machine-learn ing-ed3a661b9a71.

[7] Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538.

[8] Castelli M, Clemente FM, Popovi˘c A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/ 2020/8049504.

[9] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

[10] Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learning-based prediction of air quality. Appl Sci 10(9151):1–17. https:// doi.org/10.3390/app10249151

[11] Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145. https://doi.org/10.1109/ ICACCCN51052.2020.9362912

[12] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res 9(4):118–123.

[13] Monisri PR, Vikas RK, Rohit NK, Varma MC, Chaithanya BN (2020) Prediction and analysis of air quality using machine learning. Int J Adv Sci Technol 29(5):6934–6943 Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J Adv Comput Technol 9(9):3831–3840.

[14] Patil RM, Dinde HT, Powar SK (2020) A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms 5(8):1148–1152.

[15] Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. Int J Comput Appl Technol Res 8(9):367–370. https://doi.org/10.7753/IJCATR0809.1006

[16] Mahalingam U, Elangovan K, Dobhal H, Valliappa C, Shrestha S, Kedam G (2019) A machine learning model for air quality prediction for smart cities. In: 2019 international conference on wireless communications signal processing and networking (WiSPNET). IEEE 452–457. https://doi.org/10.1109/WiSPNET45539.2019. 9032734.

[17] Soundari AG, Jeslin JG, Akshaya AC (2019) Indian air quality prediction and analysis using machine learning. Int J Appl Eng Res 14(11):181–186.

[18] Sweileh WM, Al-Jabi SW, Zyoud SH, Sawalha AF (2018) Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900–2017). Multidiscip Respiratory Med. https://doi.org/10.1186/s40248-018-0128-5.

[19] Deshpande T (2021) India Has 9 Of World's 10 most-polluted cities, but few air quality monitors. India spend. https://www.india spend.com/pollution/india-has-9-of-worlds-10-most-pollutedcities-but-few-air-quality-monitors-792521