IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Air Quality Predication Using Machine Learning Methods Based on Monitoring Data

K. Aruna Kumari[1], B. Chandralekha[2], B. Sirisha[3], Ch. Sumanth[4], B. Venkata Surya prasanna[5]

[1]HOD & Assistant Professor, Department of Data Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

[2,3,4,5] Students, Department of Data Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

Email id: aruna.jeshwin@gmail.com[1], bandla.chandralekha@gmail.com[2], y21cds009@gmail.com[3], ch.sumanth18@gmail.com[4], boddapativenakatasuryaprasanna2@gmail.com[5]

Abstract:

This research introduces a novel methodology for air quality prediction that addresses the limitations of traditional Air Quality Index (AQI) forecasting models by leveraging machine learning and enhanced secondary data modeling. The dataset utilized includes both forecast and actual measurements of primary pollutant concentrations and meteorological conditions, collected from monitoring stations in Jinan, China, from July 23, 2020, to July 13, 2021. A comprehensive correlation analysis identified ten key meteorological factors influencing pollutant concentrations, assessed through univariate and multivariate techniques. Performance evaluation of various machine learning algorithms revealed the Decision Tree and Random Forest models achieving high accuracies of 99%. Additionally, the K-Nearest Neighbors (KNN) classifier also demonstrated an accuracy of 99%, while Logistic Regression showed a training accuracy of 72%. These findings affirm the reliability and efficacy of machine learning techniques in enhancing air quality forecasting and underscore the importance of selecting appropriate algorithms for accurate predictions.

Keywords:

air quality; machine learning; statistical analysis; secondary modeling; prediction model.

1.Introduction

Air contamination observing has acquired consideration these days as it significantly affects the wellbeing of people just as on the biological equilibrium. Other than because of the impacts of harmful emanations on the climate, well-being, work usefulness and effectiveness of energy are additionally influenced by the air contamination. Since air contamination has caused numerous perilous consequences for people it ought to be checked persistently with the goal that it tends to be controlled adequately. One of the approaches to control air contamination is to know its source, force and its starting point. Typically, it is checked by the individual express government's current circumstance service. They keep the string of the toxin gases in the individual regions. The information introduced by the WHO is cautioning about the contamination levels in the country. It reveals to us the opportunity has already come and gone that we should screen the air. Air tracking manner to measure ambient ranges of air pollutants inside the air. Monitoring has become a major job as air pollution has been increasing day by

day. Continuous monitoring of air pollution at a place gives us the levels of pollution in that area. From the information obtained by the device gives us information about the source and intensity of the pollutants in that area. Using that information, we can take measures or make efforts to reduce the pollution level so that we can breathe in a good quality of air. Air pollution not only affects the ecological balance but also the health of humans. As the levels of gases increases in the air, those gases show a major impact on the human body and lead to hazardous effects. Air pollution also affects the seasonal rainfall too due to an increase of pollutants in the air. The rainfall is also affected. Hence, continuous monitoring of the air is necessary.

2.Related work

Predicting the concentration of air pollutants in a particular area over time is an important aspect of the field of research known as "air quality prediction [1]." The management of air pollution and its detrimental consequences on both human health and the environment depends on accurate air quality forecasting. Air quality prediction using hybrid deep learning and time series analysis, by S. Zhang et al., is one of the most current articles on air quality prediction using machine learning. It was published in Atmospheric Pollution Research in 2022 [2]. The study suggests a unique method for forecasting air quality that combines deep learning and time series analysis [3]. The authors present the idea of predicting air quality and stress the significance of good prediction for reducing the adverse effects of air pollution. The limits of conventional air quality prediction techniques are then discussed, and the necessity of a hybrid strategy that combines time series analysis and machine learning is highlighted. The proposed method is then discussed, which employs a deep neural network to extract characteristics from time series data on air quality before using a time series analysis technique to project future values [4]. Using actual air quality data from Beijing, China, the authors validate their method and show how well it can accurately predict air quality. Overall, this study emphasizes the significance of air quality forecasting and offers a viable strategy based on time series analysis and machine learning methods [5]. It also highlights the possibility of combining traditional air traffic control with machine learning. Increasing the precision and efficiency of air quality prediction techniques. Pollution prediction with the correlation of pollutants with other metrological variables for 5 cities of China was done in reference PM2.5 was the pollutant that has been used [6]. RF was found to be the best method among the four methods that have been used.PM 2.5 values have been predicted by Qin et al. using CNN and LSTM. In reference also it is done with Aggregated LSTM model. NO2, SO2, CO, and O3 levels have been classified with the Decision tree and Naive Bayes algorithm. In reference Support Vector Machines and neural networks were used to categorise the eight contaminants that affect air quality [7].

PROPOSED SYSTEM

The proposed system for air quality prediction using machine learning involves several steps. The first step is to collect historical data on air quality from various sources, such as government monitoring stations and satellite data. The data is then pre- processed to remove any outliers, clean the data, and scale it to prepare it for use in machine learning models. Relevant features such as meteorological data, pollution levels, and traffic patterns are selected based on their correlation with air quality [8]. The next step is to develop machine learning models that can accurately predict air quality based on the selected features. Popular machine

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

learning algorithms used for air quality prediction include decision trees, random forests, neural networks, and support vector regression. Random forest and decision tree algorithms can handle missing data by imputing missing values, reducing the impact of missing data on the accuracy of predictions [9]. The use of random forest and decision tree algorithms also enables feature selection, which is the process of identifying the most important variables that impact air quality.

3. METHODOLOGY

Air quality prediction is an essential task that can be achieved by using machine learning (ML) algorithms. The following are the steps that can be taken to develop an ML-based air quality prediction model:

Data Collection: The first step is to collect the data required for building the air quality prediction model. This includes data on various air pollutants, weather conditions, geographical location, and other factors that impact air quality. There are several sources of data available, such as government monitoring stations, satellite imagery, and IoT devices.

Data Preprocessing: Preprocessing is the following stage after data collection. This includes cleaning the data, handling missing values, and transforming the data into a format that can be used by ML algorithms.

- Data Cleaning: This entails deleting any extraneous information, including duplicate records, irrelevant columns, and null or missing values.
- Data Integration: This step involves combining data from multiple sources, if necessary.
- Data Transformation: This involves converting the data into a suitable format for analysis, such as scaling, normalization, and feature engineering.
- Data Reduction: This step involves reducing the amount of data by sampling, feature selection, or principal component analysis (PCA).
- Data Discretization: This involves converting continuous data into discrete data by binning or bucketing.
- Data Encoding: This involves converting categorical data into numerical data, such as one-hot encoding.

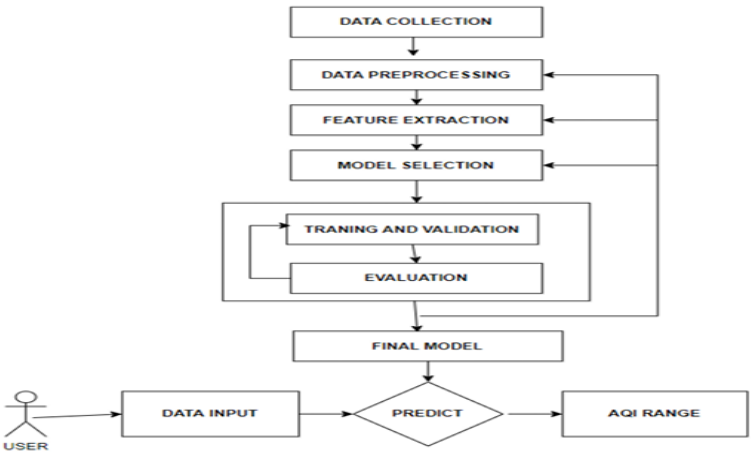


Figure: System Architecture

ISSN 2347–3657

Volume 13, Issue 2, 2025

DATA COLLECTION:

The first step in data collection is to identify the sources of data. There are several sources of air quality data, including government agencies, private organizations, and research institutions. The most reliable source of air quality data is government agencies, which collect and report data regularly. These agencies use various types of instruments to measure the levels of air pollutants in the atmosphere. Once you have identified the sources of data, the next step is to collect the relevant data. The data should include environmental factors that affect air quality, such as temperature, humidity, wind speed, and other factors. The dataset should also contain the corresponding AQI values for each data point. It is important to ensure that the data is of high quality and is collected using standardized methods to ensure consistency and accuracy.

KNN Classifier KNN: (K-Nearest Neighbors) is a classification algorithm that determines the class of a new data point based on the classes of the k-nearest neighbors in the training set. In air quality prediction, KNN can be used to predict the AQI for a new combination of environmental factors by finding the k-nearest neighbors in the training set and determining the average AQI for those neighbors. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to determine the optimal value of k and to calculate the distances between the data points

Decision Tree Classifier

A Decision Tree is a graphical representation of all the possible outcomes of a series of decisions. In air quality prediction, Decision Tree can be used to predict the AQI for a new combination of environmental factors based on a set of decision rules. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to construct the decision tree based on the features and the AQI values.

Random Forest Classifier Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions. In air quality prediction, Random Forest can be used to predict the AQI for a new combination of environmental factors by combining the predictions of multiple decision trees. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to construct multiple decision trees using different subsets of the features and the data points.

Support Vector classifier SVC chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. The model is trained using the preprocessed dataset generated in Module 2. The training data is used to construct multiple decision trees using different subsets of the features and the data points. Regression models: Gradient Boosting Regressor Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.

Lasso Regressor Lasso regression algorithm is defined as a regularization algorithm that assists in the elimination of irrelevant parameters, thus helping in the concentration of selection and

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

regularizing the models. Lasso models can be evaluated using various metrics such as RMSE and R-Square.

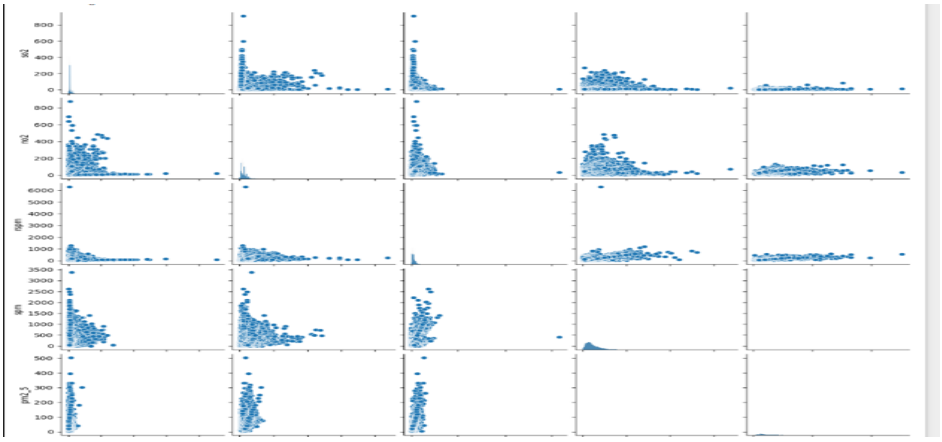
4. RESULT AND DISCUSSION

Air quality prediction using machine learning is a widely researched area due to its potential for mitigating the health and environmental effects of air pollution. Several machine learning models have been used to predict air quality, including neural networks, decision trees, and support vector machines. These models use a range of input variables, such as meteorological data, traffic data, and emission data, to predict pollutant concentrations in the atmosphere. Studies have shown that machine learning models can accurately predict air quality, with some models achieving up to 90% accuracy. However, the performance of these models is highly dependent on the quality and quantity of the input data. Models trained on incomplete or low-quality data may produce inaccurate predictions. In addition, the interpretability of machine learning models remains a challenge in air quality prediction. As machine learning models are often regarded as black boxes, it can be difficult to understand how the models arrive at their predictions. This lack of transparency can make it challenging to identify the causes of air pollution and develop effective mitigation strategies. Overall, air quality prediction using machine learning has shown promising results, but further research is needed to improve the accuracy and interpretability of these models. By combining machine learning with traditional air quality monitoring techniques, it may be possible to better understand the sources and impacts of air pollution and develop effective strategies for reducing its effects on human health and the environment.

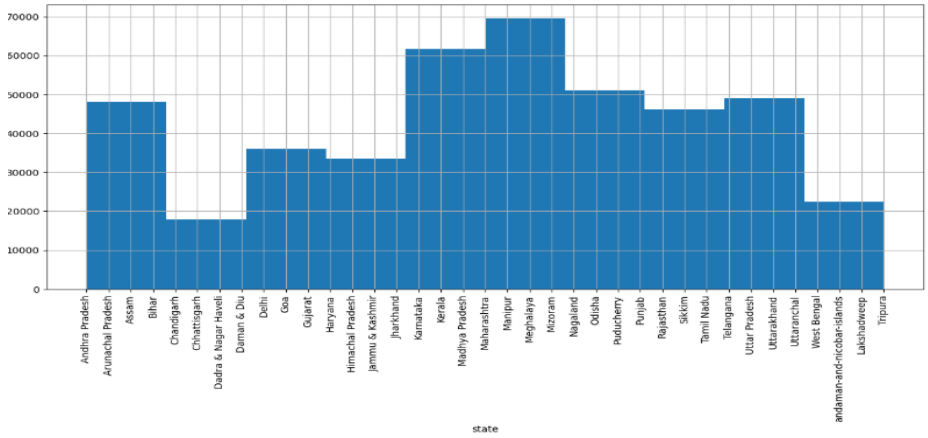
	so2	no2	rspm	spm	pm2_5
count	401096.000000	419509.000000	395520.000000	198355.000000	9314.000000
mean	10.829414	25.809623	108.832784	220.783480	40.791467
std	11.177187	18.503086	74.872430	151.395457	30.832525
min	0.000000	0.000000	0.000000	0.000000	3.000000
25%	5.000000	14.000000	56.000000	111.000000	24.000000
50%	8.000000	22.000000	90.000000	187.000000	32.000000
75%	13.700000	32.200000	142.000000	296.000000	46.000000
max	909.000000	876.000000	6307.033333	3380.000000	504.000000

```
df.nunique()
```

```
stn_code                       803
sampling_date                 5485
state                           37
location                       304
agency                          64
type                            10
so2                           4197
no2                           6864
rspm                          6065
spm                           6658
location_monitoring_station    991
pm2_5                          433
date                          5067
dtype: int64
```



International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025



CONCLUSION

In conclusion, air quality prediction using machine learning has shown potential for accurately predicting air pollution concentrations. However, the performance of these models is highly dependent on the quality and quantity of the input data, and the interpretability of these models remains a challenge. Therefore, further research is needed to enhance the accuracy and interpretability of machine learning models for air quality prediction. Future research could focus on improving the quality and quantity of input data used for air quality prediction. This could involve incorporating more sources of data, such as satellite imagery or data from low-cost air quality sensors. Additionally, research could focus on developing methods to account for uncertainty in input data, which could improve the accuracy of machine learning models. Another important area for future research is improving the interpretability of machine learning models. This could involve developing methods for explaining the predictions made by machine learning models, such as feature importance analysis or local interpretability methods. Finally, machine learning models for air quality prediction could be integrated with traditional air quality monitoring techniques to provide a more comprehensive understanding of air pollution. This could involve combining machine learning models with ground-based air quality monitoring stations or integrating them with mobile air quality monitoring platforms, such as drones or vehicles

REFERENCES

1. Temesegan Walelign Ayele, Rutvik Mehta, “Air pollution monitoring and prediction using IoT”, Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018
2. Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, Muhammad Nabeel Asghar, “Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities”, IEEE Access (Volume: 7), 2019
3. Yi-Ting Tsai, Yu-Ren Zeng, Yue-Shan Chang, “Air Pollution Forecasting Using RNN with LSTM”, 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, 2018

ISSN 2347–3657

Volume 13, Issue 2, 2025

4. Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy, “Air Quality Prediction Of Data Log By Machine Learning”, 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020
5. Shengdong Du, Tianrui Li, Yan Yang, Shi-Jinn Horng, “Deep Air Quality Forecasting Using Hybrid Deep Learning Framework”, Transactions on Knowledge and Data Engineering (Volume: 33, Issue: 6), 2021
6. Ke Gu, Junfei Qiao, Weisi Lin, “Recurrent Air Quality Predictor Based on Meteorology- and PollutionRelated Factors”, IEEE Transactions on Industrial Informatics (Volume: 14, Issue: 9), 2018
7. Bo Liu, Shuo Yan, Jianqiang Li, Guangzhi Qu, Yong Li, Jianlei Lang, Rentao Gu, “A Sequence-toSequence Air Quality Predictor Based on the n-Step Recurrent Prediction”, IEEE Access (Volume: 7), 2019
8. Baowei Wang, Weiwen Kong, Hui Guan, Neal N. Xiong, “Air Quality Forecasting Based on Gated Recurrent Long Short-Term Memory Model in Internet of Things”, IEEE Access (Volume: 7), 2019
9. Yuanni Wang, Tao Kong, “Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid”, IEEE Access (Volume: 7), 2019