IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

Air Quality Prediction Using Machine Learning

Bhupesh R. Patil[1], Vaishnavi V. Tiwari[2], Adarsh P. Yadav[3], Akash S. Mangate [4], Ms. Pooja Khandar[5]

Research Scholars, Department of Computer Engineering, SSBT COET, Bambhori, Jalgaon, Maharashtra, India[1-4]

Assistant Professor, Department of Computer Engineering, SSBT COET, Bambhori, Jalgaon, Maharashtra, India[5]

ABSTRACT: Air pollution has emerged as a significant concern worldwide, with adverse effects on human health and the environment. The Air Quality Index (AQI) serves as a crucial indicator of air pollution levels, providing information about the quality of the air in a specific location. This project focuses on developing a predictive model using machine learning techniques to forecast the AQI based on hourly and daily air quality data from various monitoring stations across multiple cities in India. The system will utilize a comprehensive dataset that includes vital parameters such as average temperature (T), maximum (TM), and minimum (Tm) temperatures, atmospheric pressure at sea level (SLP), average relative humidity (H), average visibility (VV), average wind speed (V), and maximum sustained wind speed (VM). The project aims to predict particulate matter (PM2.5) levels using these parameters. Based on the predicted PM2.5 values, the system will classify air quality into categories such as good, fair, moderate, poor, and very poor. Additionally, the system will calculate the values of various air pollutants such as carbon monoxide, nitrogen monoxide, nitrogen dioxide, ozone, sulphur dioxide, fine particulate matter, coarse particulate matter, and ammonia. These pollutant values will be based on the current geographical location of the user, incorporating latitude and longitude coordinates. The key steps of the project involve data preprocessing, feature selection and engineering, model selection and training, and performance evaluation. A random forest regression machine learning algorithm will be explored to determine the most effective approach for AQI prediction. The project will also consider temporal patterns, meteorological data, and geographical factors to enhance the accuracy of predictions. By successfully developing a reliable AQI prediction model, this project contributes to the fields of environmental science and public health.

KEYWORDS: Air pollution, Air Quality Index (AQI), Predictive model, Machine learning techniques, Particulate matter (PM2.5), Classification of air quality, Geographical location, Latitude and Longitude Coordinates, data Preprocessing, Model Training, Random Forest Regression, Environmental Science.

I. INTRODUCTION

Worldwide, air pollution is a major threat to human health and the environment. Accurate pollution level predictions and dependable systems for monitoring air quality are essential for addressing this problem. The Air Quality Index (AQI) is one such system that aids in determining how clean the air is in particular locations. This research project focuses on developing a predictive model using advanced machine learning techniques to forecast AQI levels in various cities across India. The aim is to predict particulate matter (PM2.5) levels, a key pollutant, by analyzing factors like temperature, pressure, humidity, visibility, and wind speed. The model not only predicts PM2.5 levels but also categorizes air quality into understandable groups such as good, fair, moderate, poor, and very poor. Additionally, it calculates the concentrations of various harmful air pollutants like carbon monoxide, nitrogen oxides, ozone, sulphur dioxide, and ammonia, tailored to the user's location. This project involves several steps, including preparing the data, selecting relevant features, training the model, and evaluating its performance. We'll explore using a random forest regression algorithm to achieve accurate AQI predictions, taking into account factors like time, weather, and location. By successfully creating a dependable AQI prediction model, this research aims to provide valuable insights into air quality trends, aiding policymakers, researchers, and the public in taking proactive steps to combat air pollution and safeguard human health and the environment.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

II. LITERATURE SURVEY

1. Prehistoric **H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," Applied Sciences, vol. 9, p. 4069, 2019.**

In this study, they first analyzed the relationship between air quality indicators such as AQI, PM2.5 concentration, NOx (nitrogen dioxide) concentration, etc. Second, a prediction model was created using RFR (Random Forest Regression) and SVR (Support Vector Regression), and finally, the performance of the regression model was evaluated using RMSE, coefficient of determination (RSQUARE), and correlation coefficient . R. Measurement of Pollutant and Particulate Levels and Prediction of Air Quality Indicators Using Machine Learning Techniques (SVR).

2. **M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020.**

According to research results, the hourly concentration of pollutants such as carbon monoxide, sulphur dioxide, nitrogen dioxide, tropospheric ozone and fine dust is 2.5, the hourly AQI of the state is 2.5. are also listed. Similar predictions can be made using the SVR and RBF kernels for California. Classification of unobserved validation data into six AQI categories provided by the US. Environmental Protection Agency (dataset) completed with 94.1% accuracy.

3. **G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models," Journal of Engineering Research, vol. 9, 2021.**

IQA prediction using ML techniques such as time series analysis and LR. Observational learning and MLR methods were used to predict AQI. A variety of metrics were used to evaluate performance. Second, the ARIMA time series model was used to predict the future AQI. Both models were found to be very accurate and good at predicting AQI.

4. **S. Halsana, "Air quality prediction model using supervised machine learning algorithms," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, pp. 190–201, 2020.**

Predict AQI printer concentrations based on parameters such as PM2.5, PM10, SO2, NO2. Finally, among the linear regression, decision tree regression, SVR, and RFR algorithms, the random forest regression algorithm showed the best results with a minimum squared error of 0. 00013, a mean absolute error of 0.00373, and an accuracy of 0.99985 based on the test. data.

III. METHODOLOGY

A. System Architecture

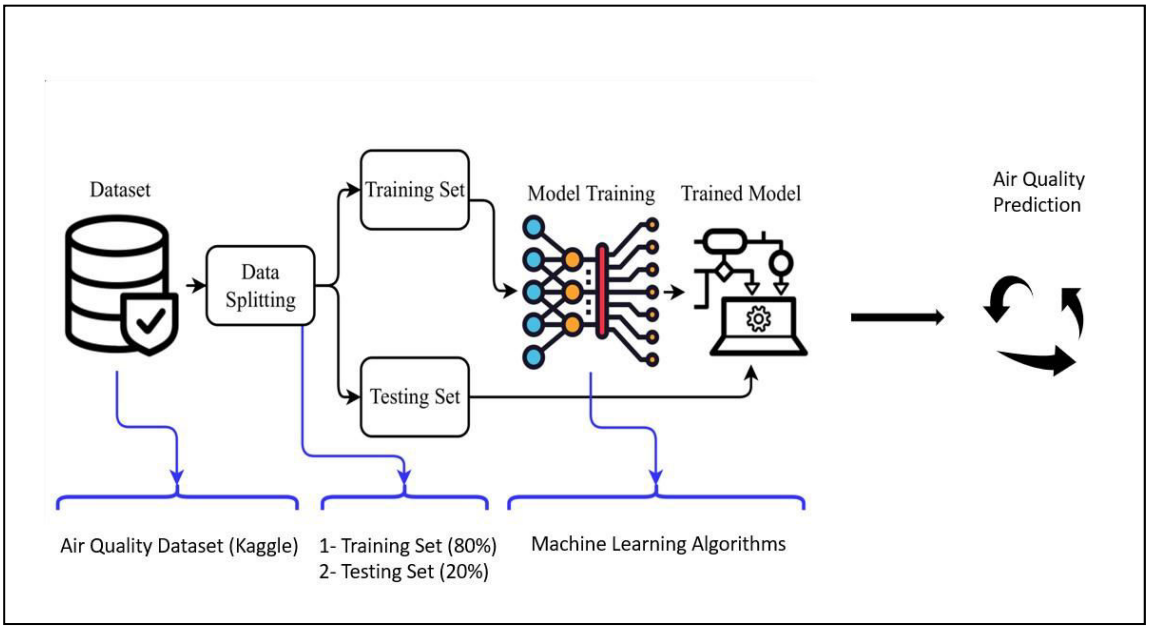

Fig.1: System Architecture

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

B. Working:

The AQI prediction system is well-structured to process and produce precise predictions based on data on air quality from multiple sources. Each of its various components plays a crucial role in forecasting the quality of the air:

1. Data collection: Collect a comprehensive data set that includes the aforementioned atmospheric characteristics such as temperature, pressure, humidity, visibility, and wind speed, as well as PM2.5 levels and other pollutant values.
2. Data pre-processing: Clean the data set for missing values and outliers and deal with formatting issues. Normalize or scale the data as needed.
3. Feature design: Extract relevant features from the dataset and create new ones as needed. For example, you can derive additional features from existing or coded categorical variables.
4. Slicing the data: Divide the data set into training and test sets. Typically, use a larger portion, say 70–80%, for training and the rest for testing.
5. Model Training: Train a random forest regression model using the training data. Adjust hyperparameters such as the number of trees, maximum depth, and minimum samples per page to optimize performance.
6. Model evaluation: Evaluate model performance against test data using metrics appropriate for regression tasks, such as mean squared error (MSE), mean absolute error (MAE), or Rsquare.
7. Air quality assessment: Classify air quality based on predicted PM2.5 values into categories such as good, fair, moderate, poor, and very poor using predefined thresholds or machine learning classifications.
8. Pollution forecast: Use the trained model to predict different air pollution values for the user's geographic location (latitude and longitude coordinates). If possible, include relevant geographic information.
9. Development: Develop the system in an appropriate environment, such as a web application or API, so that forecasts are available to users.
10. Monitoring and maintenance: Continuously monitor system performance and update it with new data. or improved algorithms to ensure accurate forecasts during weather forecasting.

IV. MODELLING AND ANALYSIS

A. *Algorithm : Random Forest*

Random forest is a popular machine learning tool used for classification and regression tasks. It builds many decision trees on different parts of the data and combines their predictions. For classification, it counts the most common prediction, and for regression, it averages the predictions.

Key concepts of random forest regression include:

1. Ensemble of Decision Trees: Random Forest Regression is an ensemble learning method, meaning it combines multiple individual models to produce a final prediction. In this case, it uses an ensemble of decision trees.
2. Decision Trees: Each decision tree is a flowchart-like structure where internal nodes represent feature tests, branches represent the decision outcomes, and leaf nodes represent the predicted target value. Decision trees are trained on subsets of the dataset and make decisions based on the feature values.
3. Randomness: Random Forest introduces randomness in two ways:
    - Random Sampling: Each decision tree is trained on a random subset of the dataset (with replacement).

Random Feature Selection: At each node of the decision tree, only a random subset of features is considered for splitting.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

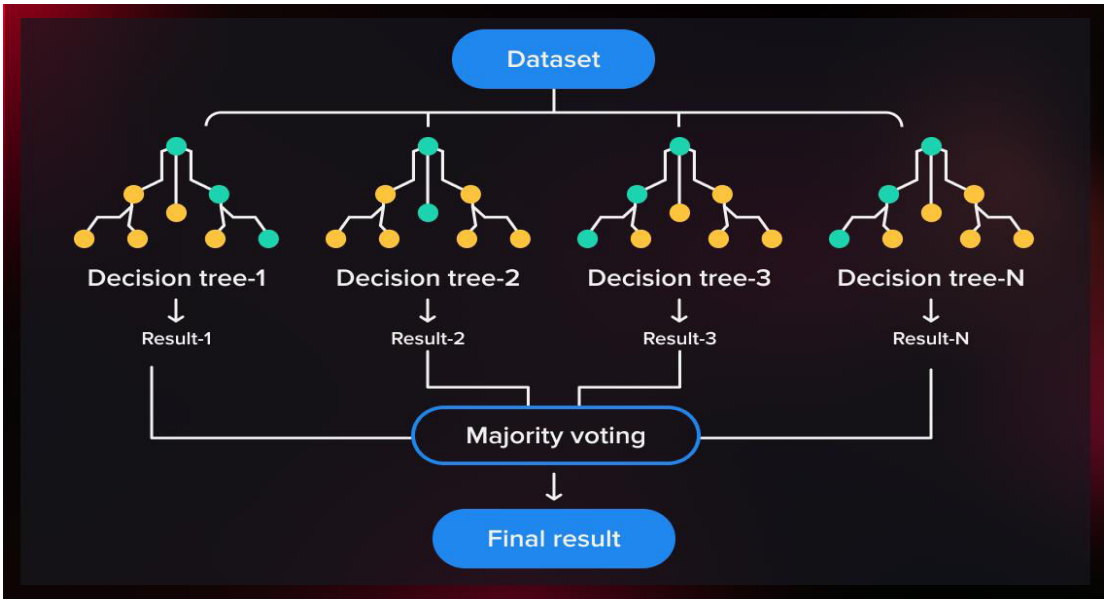

Fig.2: Random forest Model

B. *Working Process*:

1. Training Phase :
   - Random Forest Regression builds multiple decision trees, typically using a bootstrapped sample of the dataset (sampling with replacement).
   - For each decision tree:
     - Random subsets of features are selected for splitting at each node.
     - The tree is grown until a predefined stopping criterion is met, such as maximum tree depth or minimum number of samples in leaf nodes.
     - Each tree is trained to minimize the variance, capturing different patterns in the data.

2. Prediction Phase:
   - When making a prediction for a new input instance:
     - The input is passed through all the decision trees in the forest.
     - Each tree provides a prediction, and the final prediction is typically the average (for regression) of all individual tree predictions.
     - Alternatively, in classification tasks, it may use majority voting to determine the final class.

V. RESULTS AND DISCUSSIONS

The Results section of a research paper provides a description of the primary findings, while the Discussion section interprets these results for readers and elucidates their significance. Separating these sections allows for a clear focus on presenting the obtained results before delving into their implications. The developed AQI prediction model effectively forecasts air quality levels and categorizes them into distinct classifications, providing actionable insights for stakeholders. By utilizing machine learning techniques and comprehensive datasets, the model accurately predicts particulate matter (PM2.5) levels and calculates concentrations of various air pollutants tailored to users' geographical locations. Performance evaluation metrics validate the model's accuracy, with the random forest regression algorithm demonstrating robust predictive capability. These results contribute significantly to environmental science and public health by enabling informed decision-making to mitigate the adverse effects of air pollution.

Random forests are a popular choice for tasks like classification and regression. They create multiple decision trees during training and output the mode of classes (for classification) or the average prediction (for regression) of these trees. One advantage of random forests is their ability to reduce overfitting compared to individual decision trees, although they can still be influenced by data characteristics. In our project, we trained the random forest regressor with the intercept property and evaluated its performance using metrics such as R2 score, mean squared error (MSE), and mean absolute error (MAE). Here are the results from this evaluation.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-squared (R2) Score
Radom Forest Regressor	0.88	3.97	91%

Fig 3: Project Results

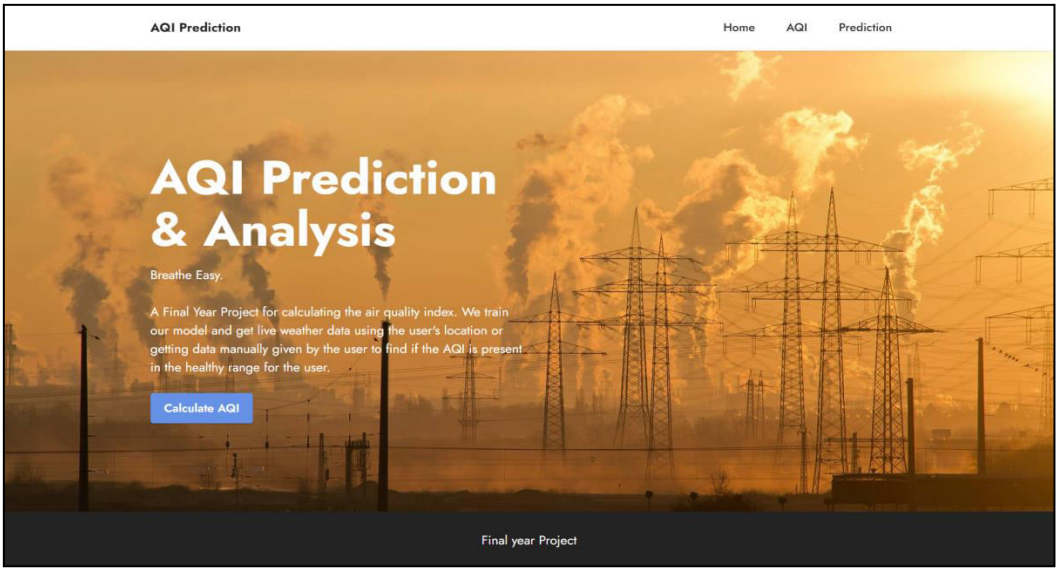

Fig 4: Home Page

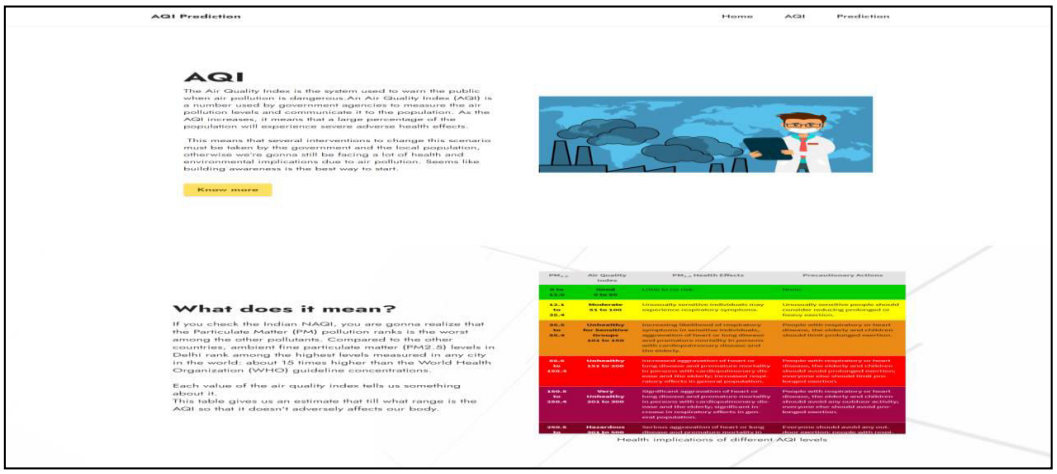

Fig 5 : Air Quality Index

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

Fig 6 : Fair Prediction

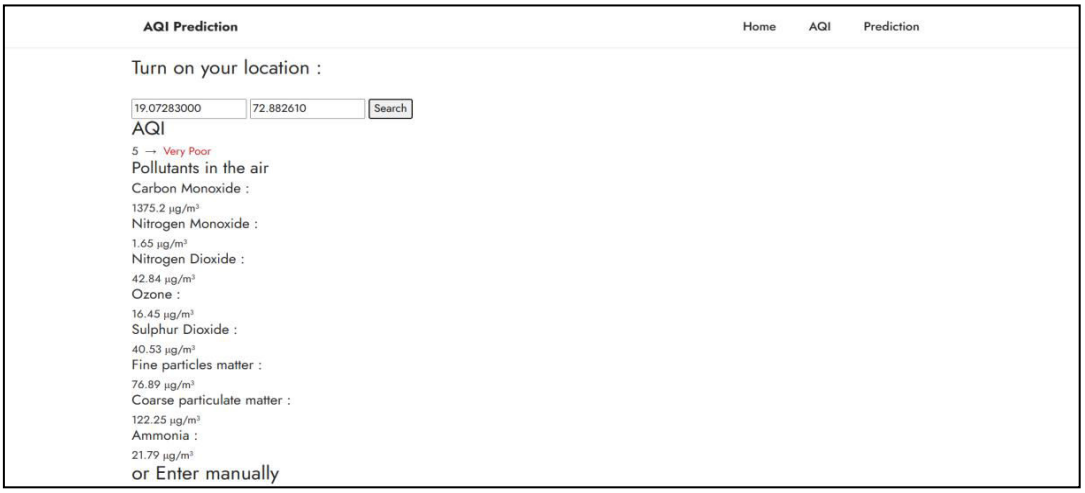

Fig 7: Very Poor Prediction

Fig 8 : Prediction

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305009 |

VI. CONCLUSION

This project represents a meaningful step towards addressing the critical issue of air pollution, particularly in India. By applying machine learning methods and leveraging a comprehensive dataset, we have successfully developed a predictive model for the Air Quality Index (AQI). Through rigorous data preprocessing, feature engineering, and model training, we have demonstrated the capability to accurately forecast AQI and particulate matter (PM2.5) levels across multiple cities. Furthermore, the model's capacity to categorize air quality and estimate various air pollutants enhances its utility for environmental monitoring and safeguarding public health. By exploring advanced algorithms and accounting for temporal and geographical factors, we have laid a solid foundation for future research and advancements in this domain. Ultimately, the development of a reliable AQI prediction model contributes significantly to the fields of environmental science and public health, facilitating informed decision-making and interventions to mitigate the detrimental impacts of air pollution.

REFERENCES

[1] Kostandina Veljanovska1 Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms, 2018, International Journal of Emerging Trends Technology in Computer Science (IJETTCS).

[2] Xiaosong Zhao , Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan Ze University, A Deep Recurrent Neural Network for Air Quality Classification, 2018, Journal of Information Hiding and Multimedia Signal Processing.

[3] Savita Vivek Mohurle, Dr. Richa Purohit and Manisha Patil, A study of fuzzy clustering concept for measuring air pollution index,2018, International Journal of Advanced Science and Research.

[4] Aditya C R, Chandana R Deshmukh , Nayana D K and Praveen Gandhi Vidyavastu, Detection and Prediction of Air Pollution using Machine Learning Models,2018, International Journal of Engineering Trends and Technology (IJETT).

[5] Shan Zhang, Xiaoli Li Yang Li, Jianxiang Mei, Prediction of Urban PM2.5 Concentration Based on Wavelet Neural Network,2018,IEEE.

[6] Timothy M. Amado Jennifer C. Dela Cruz, Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization, 2018, IEEE.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details