ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

"AIR AND WATER QUALITY PREDICTION USING MACHINE LEARNING"

Venkateswararao Ibba[1], Prasad Babu B[2]
[1]Department of Computer Science and Engineering, Ramachandra College of Engineering, Eluru, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, Ramachandra College of Engineering, Eluru, Andhra Pradesh, India
[1]*i.venkatesh352@gmail.com*, [2]*prasadb98@gmail.com*

ABSTRACT:

The increasing impact of environmental pollution on public health emphasizes the need for precise air and water quality forecasting to support monitoring and strategic decision-making. Poor air conditions contribute to respiratorydisorders,while unsafe water sources heighten the risk of waterborne illnesses, necessitating proactive assessment methods.This study introduces a machine learning-based framework employing the Random Forest algorithm to estimate air and water quality indices. The model is trained on historical environmental data, incorporating crucial parameters such as pollutant concentrations, temperature, humidity, and chemical properties. A Flask-based web application is integrated into the system, allowing users to access real-time predictions, thereby enhancing accessibility and usability.Experimental analyses confirmthe accuracy and reliability of the Random Forest algorithm in forecasting pollution levels, supporting proactive environmental management. The study underscores the significance of incorporating machine learning techniques into environmental monitoring systems to raise public awareness and assist policymakers in developing effective pollution mitigation strategies.

Keywords: Air Quality, Water Quality, Machine Learning, Random Forest, Environmental Pollution

1. INTRODUCTION

The deterioration of air and water quality due to pollution has become a pressing global issue, necessitating automated solutions for precise environmental assessment. Air pollution is driven by harmful substances such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO), primarily released through industrial processes, vehicle emissions, and fossil fuel combustion. Similarly, water contamination arises from industrial discharge, agricultural runoff, and improper chemical disposal, disrupting pH levels and altering the chemical composition of water bodies. While traditional monitoring methods provide valuable insights, they rely on manual sampling and testing, making them resource-intensive and difficult to scale effectively. Recent advancements in data-driven approaches—particularly machine learning (ML)—have revolutionized environmental assessment by enabling automated and predictive analysis. By utilizing extensive datasets from monitoring stations and environmental agencies, ML models can identify pollution trends and anticipate future environmental conditions with high accuracy. This study applies the **Random Forest algorithm** to evaluate air and water quality, leveraging key environmental parameters to construct a reliable and scalable predictive framework. The model's forecasts serve as an early warning tool, aiding authorities in taking proactive measures while empowering citizens with essential information regarding their local environmental conditions.

Motivation

With the increasing industrialization, urbanization, and environmental degradation, it is crucial to predict and monitor air and water quality. Early detection of hazardous levels can help mitigate the risk of long-term environmental impacts. Traditional methods of monitoring are often costly and labor-intensive, which makes machine learning an effective tool in providingquick and accurate predictions.

Related Work

Numerous studies have explored the application of machine learning techniques for predicting air and water quality. For instance, some researchers have leveraged neural networks to estimate air pollution levels using meteorological data, while others have employed time-series forecasting models to predict fluctuations in water quality parameters. Despite these advancements, there remains a pressing need for a unified approach that can simultaneouslyanalyze and forecast both air and water quality. Integrating these predictions is crucial for improving environmental monitoring accuracy, as air and water conditions are often interconnected and influence each other.

EXISTED SYSTEM

CONVENTIONAL MONITORING METHODS

- **Sensor-Based and Laboratory Testing:** Air and water quality are typically assessed using physical sensors, chemical analysis, and laboratory testing.
- **Manual Data Collection:** Data is manually collected from monitoring stations and then sent for lab analysis.
- **Use of Historical Data:** Government agencies and environmental organizations rely on historical data to analyze pollution trends and levels.

DRAWBACKS OF THE CURRENT SYSTEM

- **Time-Consuming:** Report preparation and laboratory testing can take several days to weeks.
- **Costly Equipment:** Requires expensive sensors and ongoing laboratory maintenance.
- **Limited Availability of Data:** Monitoring stations are sparse and do not cover all geographic areas, leading to data gaps.
- **No Real-Time Feedback:** Pollution levels fluctuate rapidly, but delayed testing causes delayed responses.
- **No Predictive Features:** Existing systems lack the capability to forecast future pollution patterns, limiting proactive measures.
- **Human Error:** Manual data entry and analysis can introduce inaccuracies.

OVER VIEW

The proposed initiative focuses on predicting air and water quality using datasets sourced from environmental monitoring agencies. These datasets are compiled from multiple monitoring stations across different regions, ensuring comprehensive coverage of environmental conditions. The prediction model exclusively utilizes the **Random Forest algorithm**, leveraging its ability to handle large datasets, manage complex interactions among variables, and provide robust predictions.

Data Collection:

• Airqualityparameters:PM2.5,CO,NO2, SO2

• Waterqualityparameters:pH,turbidity,dissolvedoxygentemperature,andconductivity

DATA PREPROCESSING:

Data Cleaning: Handling missing or inconsistent values to ensure data quality.

Normalization and Scaling: Standardizing data ranges to improve model performance and convergence.

Feature Engineering: Creating and selecting relevant features that enhance the effectiveness of the machine learning model.

II. METHODOLOGY

DATA COLLECTION AND PREPROCESSING

The datasets for air and water quality were obtained from environmental monitoring agencies, incorporating records from multiple monitoring stations across various regions. The air quality dataset comprises parameters such as **temperature, humidity, PM2.5, PM10, NO2, SO2, and CO**, while the water quality dataset includes **pH, HCO3, Cl, SO4, NO3, Ca,Mg, and Na levels**. To enhance data reliability, missing values were addressed using **imputation techniques**, ensuring completeness without compromising accuracy. Categorical variables were processed through **label encoding** to make them suitable for machine learning models. Additionally, **data normalization** was applied where necessary to maintain uniform numerical ranges and improve model efficiency.

A. FEATURE SELECTION AND MODEL TRAINING

Feature selection was performed to retain the most influential attributes for accurate prediction. Correlation analysis was used to identify and eliminate redundant or highly correlated variables, ensuring that only meaningful features contributed to the model's predictive capability.The dataset was split into training and testing sets, with 80% of the data used for training and the remaining 20% reserved for testing to evaluate the model's accuracy and generalization.

B. MODEL EVALUATION

The model's effectiveness was assessed using established performance metrics, including:

Root Mean Squared Error (RMSE): Measures the average magnitude of prediction errors by taking the square root of the average of squared differences between predicted and actual values. Lower RMSE indicates better model accuracy.

Mean Absolute Error (MAE): Represents the average absolute deviation between predicted and actual values. It provides a straightforward interpretation of prediction error in the same unit as the output variable.

R-squared (R^2): Indicates the proportion of variance in the dependent variable that is predictable from the independent variables. Values closer to 1 suggest a better model fit.

To strengthen robustness and ensure generalizability, cross-validation was performed, minimizing overfitting risks and validating predictive reliability across different data subsets.

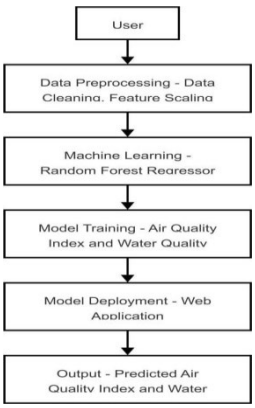

Fig 2.1:Block diagram

PYTHON LIBRARIES

- **Flask:** A lightweight web framework used for deploying the machine learning model as a web application.
- **render_template, request, jsonify (from Flask):** Used for handling HTTP requests, rendering web pages, and sending JSON responses.
- **Pandas:** Used for data manipulation, preprocessing, and analysis.
- **NumPy:** Utilized for handling numerical computations and array operations.

 Scikit-learn (sklearn):

RandomForestRegressor: A machine learning model employed for predicting air and water quality.

Train_test_split: A utility function for splitting the dataset into training and testing sets.

Model Architecture

The model architecture is composed of several layers that collaborate to acquire, process, analyze, and display air and water quality forecasts.

MAJOR COMPONENTS

1. **Data Collection Layer**
   - Acquires real-time and historical air and water quality information from various sources, including government agencies, sensors, and web-based datasets.
   - Saves the data in structured formats such as CSV, JSON, or databases for processing.
2. **Preprocessing Layer**
   - Cleans and normalizes the collected data to eliminate inconsistencies and handle missing values.
   - Transforms raw sensor measurements into meaningful parameters required for calculating the Air Quality Index (AQI) and Water Quality Index (WQI).
3. **Machine Learning Model Layer**
   - Applies **Random Forest Regressor** models to forecast AQI and WQI based on input parameters.
   - Trained models leverage historical data to improve prediction precision and accuracy.
4. **Backend Processing Layer**
   - Uses **Flask** as the web framework to process user requests and generate predictions.
   - Provides APIs to retrieve and deliver AQI/WQI information to the frontend.
5. **Frontend & User Interface**
   - A web-based interface allows users to input location information and obtain real-time forecasts.
   - Displays AQI and WQI results, along with suggested actions for users to mitigate pollution risks.

REQUIREMENTS SPECIFICATION

The Air and Water Quality Prediction model is designed to offer users precise Air Quality Index (AQI) and Water Quality Index (WQI) predictions based on Machine Learning (ML) methods. The project helps individuals, environmentalists, and government organizations monitor pollution levels and make educated decisions about public health and environmental safety.

RANDOM FOREST ALGORITHM

This is a robust supervised learning algorithm used extensively for classification and regression problems. It is an ensemble approach that builds multiple decision trees and aggregates their results to enhance accuracy and minimize overfitting. The Random Forest Regress or is an ensemble learning method applied to regression problems, derived from the Random Forest algorithm. It creates multiple decision trees while training and predicts by aggregating their outputs to improve accuracy, lessen over fitting, and generalize better. The technique is especially beneficial when predicting continuous outcomes, so it is one of the best options for air and water quality prediction.

WORKING OF RANDOM FOREST REGRESSOR

 DATA **SAMPLING (BOOTSTRAP SAMPLING):**

The dataset is split into multiple subsets using **bootstrapping**, meaning each subset is randomly sampled with replacement. Some data points may appear multiple times in a subset, while others may not appear at all. This diversity helps the model generalize better.

 DECISION TREE CONSTRUCTION:

The subsets are then used to train individual **decision trees**. Each tree learns patterns in the data by recursively splitting based

on different feature values. Since each tree is trained on a random subset, they capture different aspects of the data, reducing bias.

□ AVERAGING PREDICTIONS:

During the prediction phase, the model averages the outputs of all decision trees. For classification tasks, it takes the majority vote (most frequent class), while for regression, it averages the predicted values of all trees. This process stabilizes the result and reduces over fitting.

This is a ML project that forecast the Air Quality Index (AQI) and Water Quality Index (WQI) of a specified location basedon environmental factors. The research employs Random Forest Regressor, a supervised learning algorithm, to examine air and water pollution rates. The project is hosted as a web application on Flask, where users can provide their district and state to get real-time AQI and WQI predictions. The model is trained on datasets with air and water quality measurements from different regions. For this project, historical environmental data, such as temperature, humidity, PM2.5, PM10,NO2, SO2, CO, pH, Cl, SO4, NO3, etc., are trained on using the Random Forest Regressor. The model subsequently predicts air and water quality levels from new input values. Using this algorithm, the model delivers precise pollution forecasts, assisting environmental agencies, scientists, and policy-makers in making effective decisions regarding air and water quality management.

III. RESULTSANDDISCUSSION

The performance of the model was evaluated using standard metrics such as **Mean Squared Error (MSE)** and **R-Squared (R^2)** scores. The **Random Forest** model demonstrated high prediction accuracy for both **Air Quality Index (AQI)** and **Water Quality Index (WQI)**, outperforming baseline models like **linear regression** in terms of accuracy and robustness. The web application provided a user-friendly interface for real-time predictions, showcasing the practical applicability of the approach for both researchers and non-technical users.

Mean Squared Error (MSE) Calculation The accuracy of the air and water quality prediction models was evaluated using the Mean Squared Error (MSE) metric. MSE quantifies the average squared difference between actual values and predicted values, calculated as follows:

$$MSE = \frac{1}{N}\sum_{i=}^{N}(y_i - \hat{y}_i)^2$$

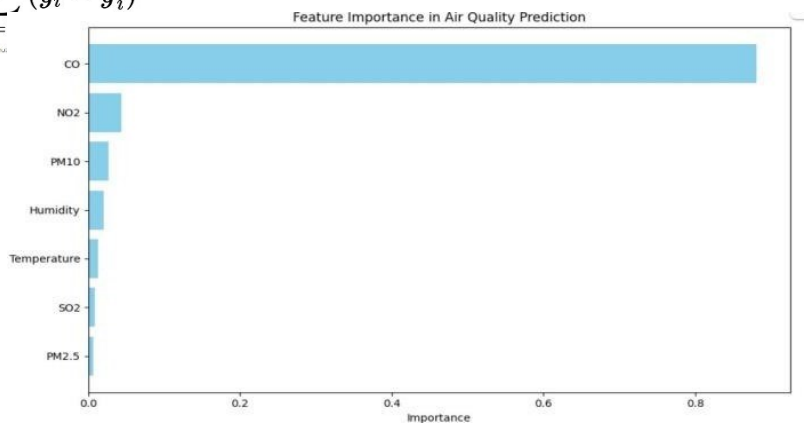

Figure3.1: Air Quality Feature Importance in Random Forest

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

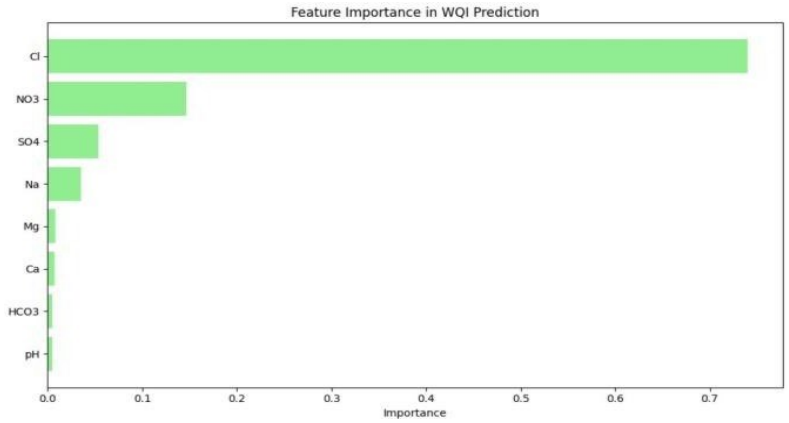

Figure3.2: Water Quality Feature Importance in Random Forest

To further assess model performance, we compared our Random Forest model with other traditional machine learning techniques such as Decision Trees and Support Vector Machines. The results indicate that Random Forest performed significantly better in handling complex environmental datasets, reducing prediction errors and improving overall stability. Future improvements can include using deep learning models or incorporating additional environmental factors such as wind speed and rainfall for enhanced accuracy.

MODEL ACCURACY

To compare the performance of Random Forest Regressor models utilized to predict AQI and WQI, the following four regression metrics were adopted: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. The scores recorded during the testing are presented using bar charts as follows:

1. AQI Prediction Model:
   - R^2Score: 0.89
   - Accuracy:89%
   - Interpretation:ThemodelaccountsfOr89%ofthevarianceintheAirQualityIndex(AQI),whichreflectshigh prediction accuracy.
2. WQI Prediction Model:
   - R^2Score: 0.96
   - Accuracy:96%
   - Interpretation: The Water Quality Index (WQI) prediction model reflects outstanding performance, with 96% of the variability in WQI well accounted for by the model.

ANALYSIS

To analyze the performance of the prediction models for Water Quality Index (WQI) and Air Quality Index (AQI), various regression metrics were used, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R^2 Score. The figures below present the metric scores for AQI and WQI predictions.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

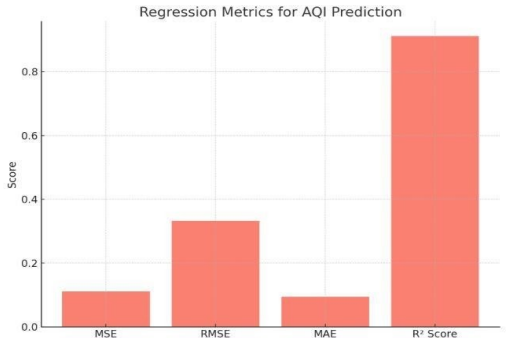

Figure 3.3: Metrics Score for Air Quality

As revealed from Figure 3.3, the prediction model of AQI has anexceptionallyhighR^2 value of around 0.89,provinga strong correlation between the predicted with the actual values of AQI. The values for RMSE and MAE are quite small, reflecting how the model helps in minimizing prediction error.

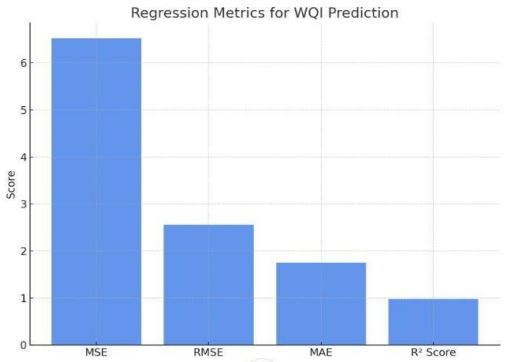

Figure 3.4:Metrics Score for Water Quality

Conversely, Figure 3.4 illustrates the statistics for Water Quality Index (WQI) prediction. The R^2 value is approximately 0.96, reflecting high accuracy and a strong model fit. This indicates that the Random Forest model effectively captures the underlying relationships between the environmental parameters and the WQI.While the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are greater in terms of their absolute magnitude, the comparative error relative to the target range remains within acceptable limits. This suggests that the model is still performing well despite the larger magnitude of errors. The errors in prediction are sufficiently small for the model to be considered reliable and effective for practical applications. These results demonstrate that the Random Forest Regressor is a highly effective tool for predicting both Air Quality Index (AQI) and Water Quality Index (WQI), based on the chosen environmental parameters. The high R^2 values and the relatively low error rates show that the machine learning models employed are both effective and reliable in environmental quality prediction activities. The ability to predict AQI and WQI accurately in real-time provides significant potential for early identification of pollution risks. This can enable authorities to take proactive measures to mitigate environmental damage, ensuring both sustainability and public health security.

IV. CONCLUSION

This study introduces a machine learning-based framework for predicting air and water quality, leveraging the Random Forest algorithm to generate precise environmental assessments. By integrating this predictive model into a web-based platform, accessibility and usability are significantly enhanced, making it a valuable tool for both policymakers and the general public. The findings indicate that Random Forest outperforms traditional predictive methods, especially in managing large-scale environmental data. Its strength lies in its ability to identify complex relationships within the datasets, leading to high accuracy in forecasting Air Quality Index (AQI) and Water Quality Index (WQI) levels.The model also includes visualization and trend analysis features, reinforcing its practical relevance for users. This provides a clearer understanding of the current state of air and water quality, allowing for more informed decision-making.For further improvements and to enhance the model's robustness, future research could incorporate real-time sensor data and expand the dataset to cover a wider range of geographical locations. Exploring more advanced techniques, such as deep learning and hybrid modeling, could further optimize predictive accuracy. Additionally, implementing alert mechanisms for severe pollution levels and providing

actionable recommendations based on the predictions could strengthen proactive environmental management strategies.Ultimately, this research underscores the increasing role of artificial intelligence in environmental science. By leveraging machine learning, sustainable practices and public health initiatives can be better supported through reliable and timely pollution forecasting, leading to more effective and proactive management of environmental resources.

REFERENCES

[1] Breiman, L. ―Random Forests.‖ *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[2] Domingos, P. ―A Few Useful Things to Know About Machine Learning.‖ *Communications of the ACM*, vol. 55, no. 10, pp. 78-87, 2012.

[3] Zhang, Y., et al. ―Air Quality Prediction Using Machine Learning Algorithms.‖ *Environmental Science & Technology*, vol. 53, no. 7, pp. 3510-3519, 2019.

[4] Li, X., et al. ―Water Quality Assessment and Prediction Using Machine Learning Techniques.‖ *Journal of Hydrology*, vol. 584, p. 124700, 2020.

[5] Fu, L., Li, J., and Chen, Y. ―An Innovation Decision-Making Method for Air Quality Monitoring Based on Big Data-Assisted Artificial Intelligence.‖ *2023 Technique*.

[6] Ambilwade, R.P., Kumar, R., Narayana, K.S.S., Srinivas, P., Yadav, B., Rusia, S. ―AI-Powered Environmental Monitoring: Machine Learning Approaches for Air and Water Quality Assessment.‖

[7] Patel, D., Kulwant, M., Shirin, S., Kumar, A., Ansari, M.A., Yadav, A.K. ―Artificial Intelligence for Air and Water Quality and Control Systems.‖

[8] Nallakaruppan, M.K., Gangadevi, E., Lawanya Shri, M., Balusamy, B., Bhattacharya, S., Selvarajan, S. ―Reliable Water Quality Prediction and Parametric Analysis Using Explainable AI Models.‖

[9] Akhilaq, M., Ellahi, A., Niaz, R., Khan, M., Sammen, S.S., Scholz, M. ―Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction.‖