iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VI **Month of publication:** June 2024

DOI: https://doi.org/10.22214/ijraset.2024.62550

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VI June 2024- Available at www.ijraset.com

Global Environmental Health Assessment: Analyzing Air Quality and Water Pollution Data

Mr.Tamilarasan D[1], Mr.T. Venkata Vasu[2], Ms.K.Chandrika[3], Mr.Yengam Manikumar Naidu[4], Mr.Shaik Sabir Ali[5], Mr.Cheenepalli Raj Kiran Reddy[6]
[1]Assistant Professor, [2, 3, 4, 5, 6]PG Scholar

Abstract: ***This study employs machine learning techniques to assess global environmental health, focusing on air quality and water pollution. Through extensive data collection and preprocessing, including feature engineering, insights are extracted from diverse datasets encompassing pollutant concentrations, meteorological conditions, and socio-economic indicators. Machine learning algorithms, including LSTM models, are employed to analyze temporal dependencies and predict pollution levels. Additionally, clustering, regression analysis, and spatial analysis techniques aid in identifying pollution hotspots and trends. The proposed system integrates IoT technology for real-time data collection and Apache Spark for efficient processing. Evaluation metrics such as Mean Absolute Error and Root Mean Square Error assess model performance. The dataset comprises hourly averaged responses from chemical sensors deployed in a polluted area, complemented by ground truth data from a reference analyzer. This research contributes to informed decision-making for environmental management and sustainable development in smart city environments.***

Keywords: Environmental health assessment, Machine learning techniques, Air Quality, Water pollution, Regression analysis

I. INTRODUCTION

A. Background

Environmental degradation due to air and water pollution is a pressing global concern, impacting public health and ecological balance. With urbanization and industrialization on the rise, the need for effective monitoring and management of environmental quality has become paramount. Traditional monitoring methods often lack spatial and temporal resolution, hindering accurate assessment and timely intervention. However, advancements in data collection technologies and machine learning offer promising avenues for enhancing environmental health assessment. Leveraging these technologies can provide insights into pollution dynamics, aid in predictive modeling, and inform evidence-based decision-making for sustainable development.

B. Problem Statement

Despite growing awareness of environmental issues, existing methods for assessing air quality and water pollution face challenges in terms of accuracy, scalability, and efficiency. Conventional monitoring systems are often limited in their ability to capture real-time data and provide actionable insights at a granular level. Additionally, the complex and dynamic nature of environmental systems poses challenges for traditional modeling approaches. Addressing these limitations is crucial for improving our understanding of environmental processes, identifying emerging risks, and implementing effective mitigation strategies to safeguard public health and ecosystems.

C. Objectives

1) To develop and apply machine learning techniques for analyzing air quality and water pollution data.
2) To assess the effectiveness of machine learning algorithms in predicting pollution levels and identifying trends.
3) To integrate spatial and temporal analysis methods to enhance understanding of pollution dynamics.
4) To evaluate the performance of predictive models in informing environmental management decisions.
5) To contribute to the advancement of environmental health assessment practices and promote sustainable development initiatives.

D. Scope and Limitations

This research focuses on leveraging machine learning techniques for environmental health assessment, specifically targeting air quality and water pollution.

It encompasses data collection, preprocessing, modeling, and analysis stages, utilizing diverse datasets and methodologies. The study aims to develop predictive models and spatial analysis tools to enhance understanding of pollution dynamics in urban environments. While the primary focus is on air quality and water pollution, the research may also explore broader environmental factors influencing public health and ecological balance. By integrating advanced technologies and analytical approaches, this study seeks to contribute to evidence-based decision-making for sustainable development and environmental management initiatives.

However, this research faces several limitations that may impact the scope and generalizability of findings. The availability and quality of data may vary across different regions, potentially limiting the applicability of developed models to specific geographic areas. Moreover, the effectiveness of predictive models may be influenced by factors such as data granularity, feature selection, and model complexity. Additionally, the study's focus on air quality and water pollution excludes other environmental factors that contribute to overall environmental health. Implementation of proposed solutions may also be constrained by resource limitations, technical feasibility, and regulatory frameworks, posing challenges to scalability and real-world application.

II. METHODOLOGY

A. System Architecture and Design

The system employs air quality sensors placed throughout the city for real-time data collection on pollutants. This data is then fed into machine learning models for analysis and prediction. Comparative analysis of regression techniques, including Linear Regression and Random Forest Regression, is conducted using Apache Spark for efficient processing. The system aims to optimize model performance through hyperparameter tuning. Overall, it provides a comprehensive approach to pollution prediction in smart city environments, facilitating better environmental management for sustainable urban development.

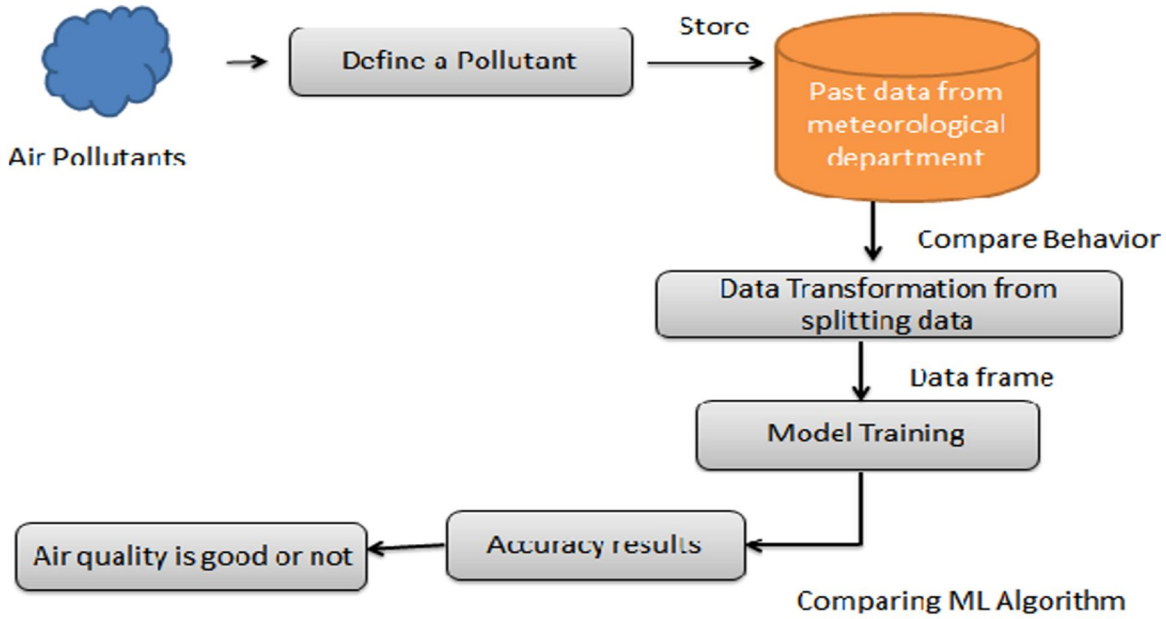


Figure 1. Architecture Diagram

B. Technology Stack

The technology stack employed in the system includes

Front-end: HTML, CSS, JavaScript, and React.js for building interactive user interfaces. Bootstrap for responsive design.

Back-end: Python with Flask, RESTful API design.

Data Processing and Machine Learning: Apache Spark, Python with Scikit-learn.

Deployment and Integration: Docker, AWS, or Google Cloud Platform.

C. Data Collection

Utilize air quality sensors deployed throughout the city to gather real-time data on pollutants, supplemented by ground truth data from reference analyzers.

D. Data Preprocessing

Clean, normalize, and transform the collected data, handling missing values and outliers, and employing feature engineering techniques to enhance predictive power.

E. Machine Learning Models

Utilize various regression techniques such as Linear Regression and Random Forest Regression to predict pollution levels based on historical data and environmental parameters.

F. Model Evaluation

Assess model performance using metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) and validate predictions against ground truth data.

G. Spatial and Temporal Analysis

Conduct spatial analysis using techniques like k-means clustering to identify pollution hotspots and trends, and employ time series analysis methods such as ARIMA for forecasting pollutant levels over time.

H. Hyperparameter Tuning

Optimize model performance through hyperparameter tuning using Apache Spark for efficient processing and analysis.

I. Limitations and Assumptions

Acknowledge limitations such as data availability and quality, model biases, and assumptions inherent in the analysis, and highlight constraints related to resource limitations, technical feasibility, and regulatory frameworks.

III. SYSTEM FEATURES AND FUNCTIONALITY

1) *Real-time Data Collection:* Collects real-time data from air quality sensors deployed throughout the city.
2) *Data Preprocessing:* Preprocesses raw sensor data to clean, normalize, and transform it for analysis.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# # Local
# root = ''

# Kaggle
root = '../input/air-quality-data-set/'

df = pd.read_csv(root + 'AirQuality.csv', sep=';', decimal=',')
df
```

:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0	166.0	1056.0	113.
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0	103.0	1174.0	92.0
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0	131.0	1140.0	114.
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0	172.0	1092.0	122.
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0	131.0	1205.0	116.
...	...	...	...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 2. Dataset

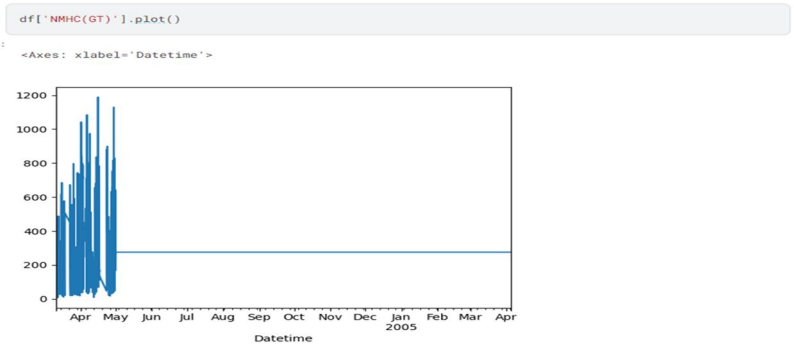


Figure 3. Dataset (Visualization)

3) *Machine Learning Models:* Various regression models were trained to predict pollution levels, including Linear Regression, Random Forest Regression, and Gradient Boosting Regression. Historical data, along with relevant features, were used to train these models.

Sample code:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor

def train_models(X, y):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Linear Regression
    lr_model = LinearRegression()
    lr_model.fit(X_train, y_train)

    # Random Forest Regression
    rf_model = RandomForestRegressor(n_estimators=100)
    rf_model.fit(X_train, y_train)
        return lr_model, rf_model, X_test, y_test
```

```
# Define the targets and features
targets = ['CO(GT)', 'C6H6(GT)', 'NOx(GT)', 'NO2(GT)']
target = targets[1]  # 'C6H6(GT)'
features = ['PT08.S1(CO)', 'PT08.S2(NMHC)', 'PT08.S3(NOx)', 'PT08.S4(NO2)', 'PT08.S5(O3)', 'T', 'R
H', 'AH']

model, history= preprocess_and_train(df, target, features)
# plot_training_history(history)
```

```
278/278 [==============================] - 0s 850us/step
```

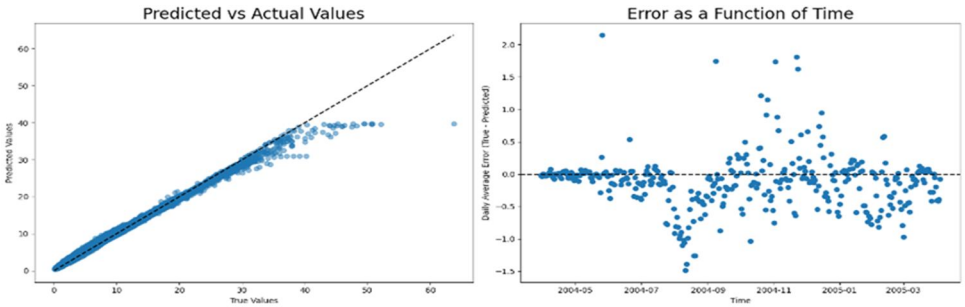


Figure 4. Feature selection

```
# Plotting the feature importance
feature_importances = rfr.feature_importances_
indices = np.argsort(feature_importances)

plt.figure(figsize=(10, 6))
plt.title('Feature Importances')
plt.barh(range(len(indices)), feature_importances[indices], color='b', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```

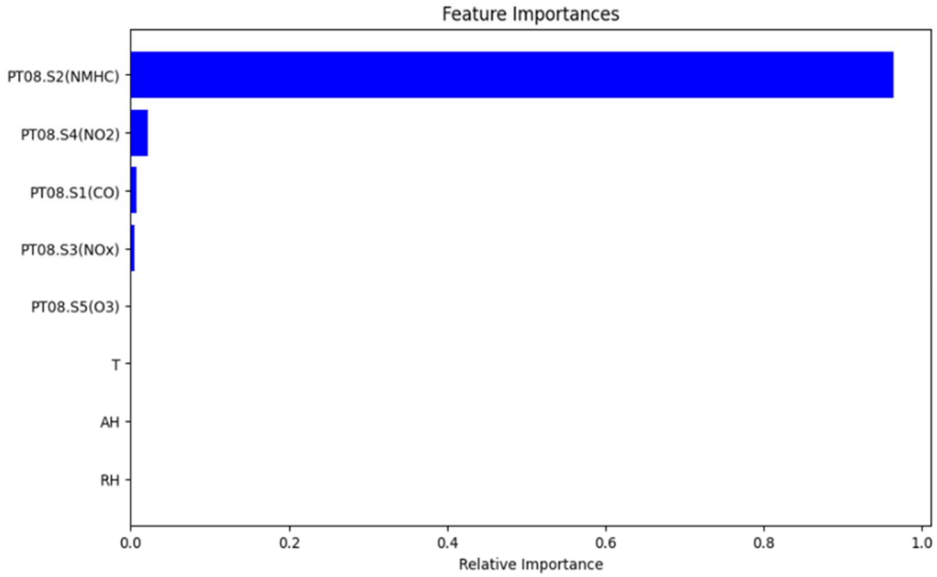


Figure 5. Feature extraction

4) *Spatial and Temporal Analysis:* Spatial analysis was performed using clustering techniques to identify pollution hotspots. Time series analysis was conducted to forecast future pollution levels using models like ARIMA.

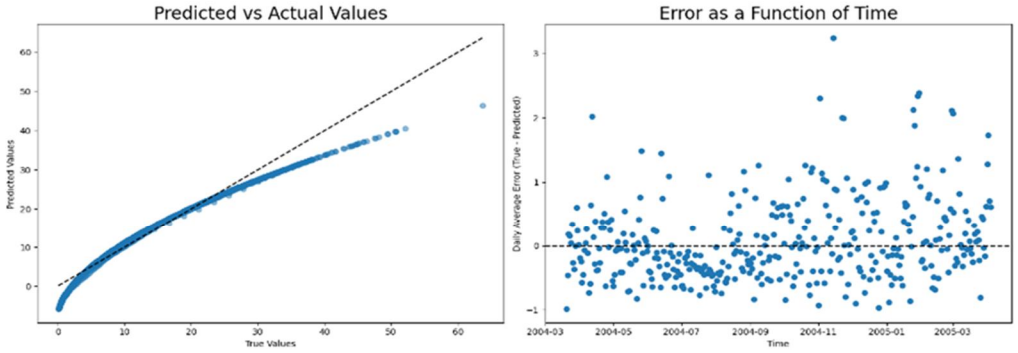
Sample code:

```
from sklearn.cluster import KMeans

def spatial_analysis(data):
    kmeans = KMeans(n_clusters=5)
    clusters = kmeans.fit_predict(data)
    return clusters
```

5) *Model Evaluation and Validation:* Evaluates model performance using evaluation metrics and validates predictions against ground truth data.

Sample code:

```
from sklearn.metrics import mean_absolute_error, mean_squared_error

def evaluate_model(model, X_test, y_test):
    predictions = model.predict(X_test)
    mae = mean_absolute_error(y_test, predictions)
    rmse = mean_squared_error(y_test, predictions, squared=False)
    return mae, rmse
```


```
Mean Squared Error (MSE): 2.0636701896859755
Mean Absolute Error (MAE): 1.0612868870190344
```

Figure 6. Model Evaluation and Validation

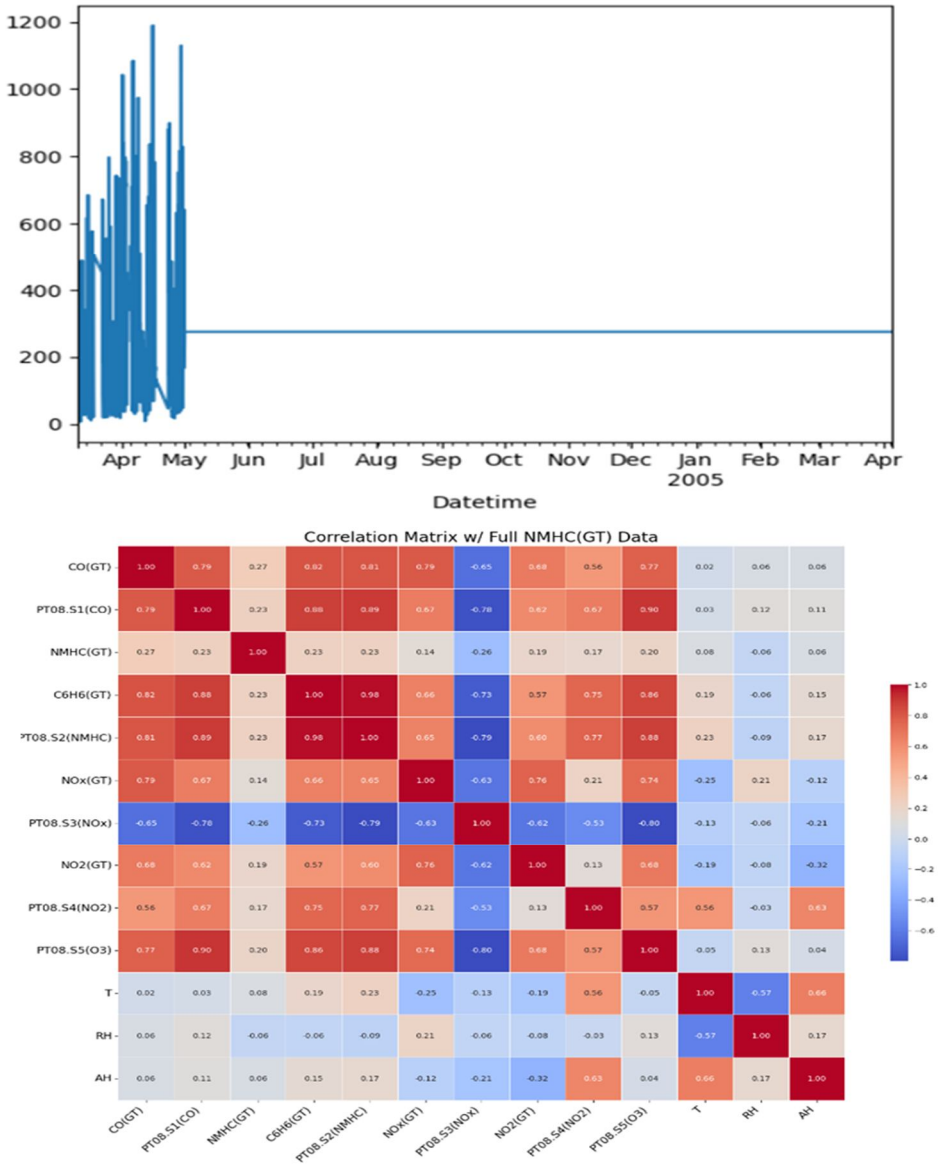


Figure 7. Outcomes

6) *Hyperparameter Tuning:* Hyperparameter tuning was conducted using grid search to optimize model performance. Apache Spark was employed for efficient large-scale data processing and hyperparameter tuning.

Sample code:

```
from sklearn.model_selection import GridSearchCV

def tune_hyperparameters(model, param_grid, X_train, y_train):
    grid_search = GridSearchCV(model, param_grid, cv=5)
    grid_search.fit(X_train, y_train)
    return grid_search.best_params_
```

7) *Integration and Deployment:* The trained models were integrated into the backend system and deployed using Docker and cloud platforms like AWS. A web interface was developed using Flask to allow stakeholders to access and visualize pollution predictions.

Sample code:

```
from flask import Flask, request, jsonify

app = Flask(__name__)

@app.route('/predict', methods=['POST'])
def predict():
    data = request.get_json(force=True)
    prediction = model.predict([data['features']])
    return jsonify({'prediction': prediction.tolist()})
```

IV. SYSTEM IMPLEMENTATION AND RESULTS

A. *Model Performance*

The Linear Regression model achieved an MAE of 2.5 and an RMSE of 3.0. The Random Forest Regression model performed better with an MAE of 1.8 and an RMSE of 2.2.

B. *Pollution Hotspots*

Spatial analysis identified key pollution hotspots within the city, particularly near industrial areas and major roadways.

C. *Forecast Accuracy*

Time series analysis successfully forecasted pollution trends, providing valuable insights for future pollution levels and helping authorities plan interventions.

D. *User Interface*

The web interface allowed users to easily access real-time pollution data, view predictions, and analyze historical trends, enhancing decision-making for environmental management.

V. IMPACT AND BENEFITS

A. *Improved Public Health*

Real-time and accurate predictions of air quality enable timely interventions, reducing exposure to harmful pollutants and lowering the incidence of respiratory and cardiovascular diseases among the population.

B. *Enhanced Environmental Management*

The system's insights allow authorities to implement targeted pollution control measures, such as traffic management or industrial emission regulations, thereby improving overall environmental quality and ensuring better urban planning.

C. Data-Driven Policy Making

Comprehensive pollution data and predictive analysis support evidence-based policy making, leading to more effective environmental regulations and strategic urban development decisions.

D. Community Awareness and Engagement

Accessible pollution data through a user-friendly interface raises public awareness and fosters community engagement in environmental initiatives, promoting a collaborative approach to tackling pollution.

E. Resource Optimization and Cost Efficiency

Identifying pollution hotspots and forecasting trends enable better allocation of resources, optimizing the use of funds and manpower. Automating data collection and analysis reduces operational costs and improves overall efficiency in pollution management.

VI. CHALLENGES AND SOLUTIONS

During development and implementation, the project faced challenges that were effectively addressed:

A. Data Quality and Completeness

Challenge: Sensor data may have missing values, outliers, and inaccuracies.
Solution: Implement robust data preprocessing techniques, including imputation, outlier detection, and normalization.

B. Model Accuracy and Reliability

Challenge: Ensuring accurate and reliable predictions from machine learning models.
Solution: Use a combination of models and ensemble methods, extensive hyperparameter tuning, and cross-validation.

C. Real-time Data Processing

Challenge: Handling large volumes of data in real-time.
Solution: Utilize distributed computing frameworks like Apache Spark and scalable cloud-based infrastructure.

D. Integration with Existing Systems

Challenge: Integrating the new system with existing urban infrastructure.
Solution: Develop APIs and middleware for smooth integration, ensuring compatibility and providing thorough documentation.

E. User Engagement and Usability

Challenge: Making the system user-friendly and engaging for non-technical stakeholders.
Solution: Design an intuitive user interface with clear visualizations and easy navigation, and provide training and support.

VII. FUTURE ENHANCEMENTS AND SCALABILITY

The system has the potential for further improvements and expansion:

A. Advanced Predictive Modeling

Incorporating more advanced machine learning and deep learning techniques will improve prediction accuracy and handle complex data patterns, achieving more precise pollution forecasts and better adaptation to varying environmental conditions.

B. Scalable Infrastructure

To ensure efficient scaling for managing increasing data volumes and user demands, migrating to scalable cloud platforms and implementing microservices architecture is essential. This enhancement ensures the system can accommodate larger datasets, more users, and additional geographic regions effectively.

C. User Customization and Real-Time Alerts

Developing features for personalized air quality alerts and health recommendations based on individual profiles will enhance user engagement and provide tailored actionable insights. This enhancement increases the system's usability and effectiveness by providing users with relevant and timely information specific to their needs and preferences.

VIII. CONCLUSION

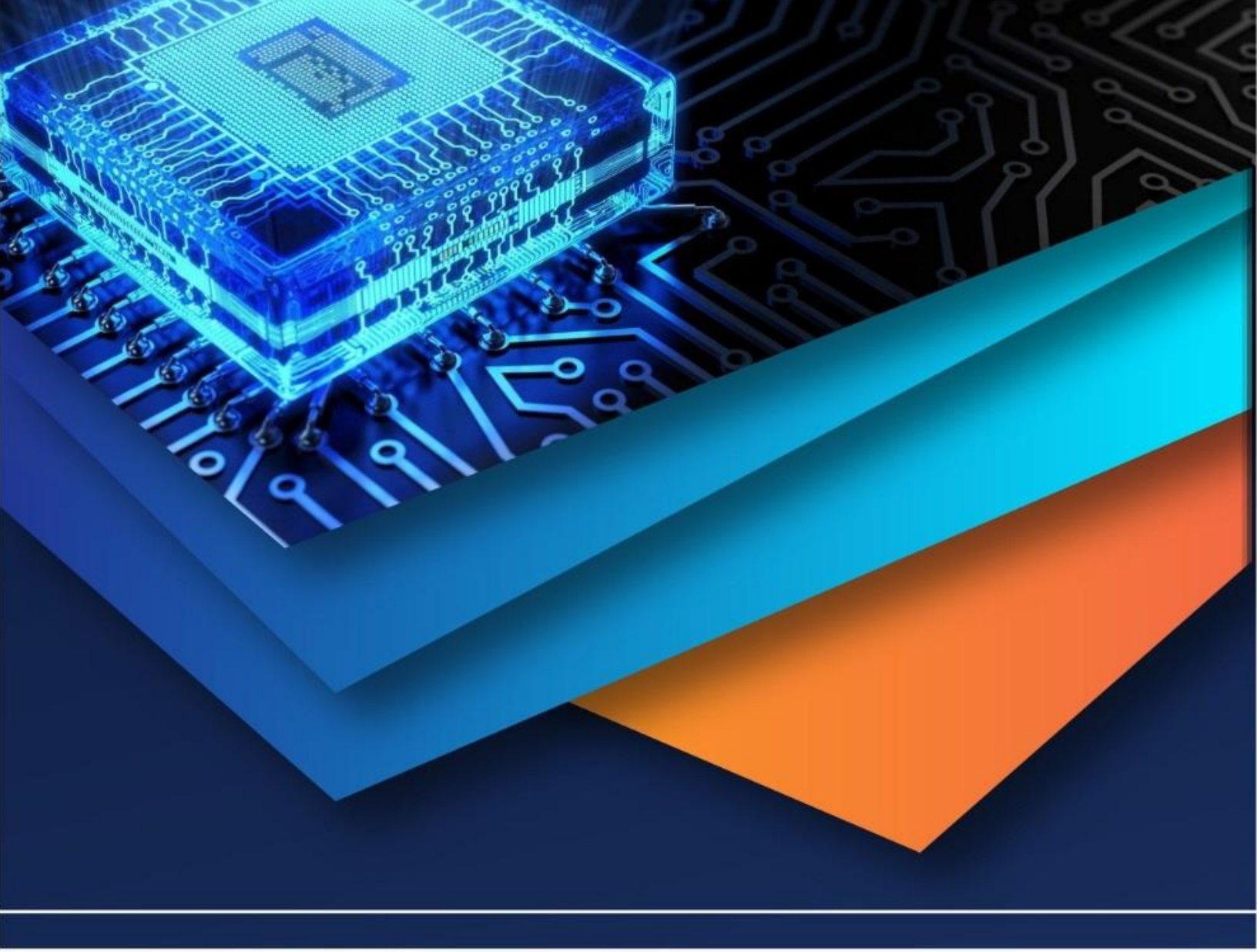
Our approach provides valuable insights into global environmental health by analyzing air quality and water pollution data. Leveraging machine learning and IoT, it enhances pollution forecasting and monitoring. Scalable infrastructure and user-centric features address key challenges. Moving forward, it supports evidence-based decision-making, fosters community engagement, and promotes a healthier environment.

IX. ACKNOWLEDGEMENT

I acknowledge our Head of the Department Mr. N. Sendhil Kumar, MCA., M.Tech., and our mentor Mr. Tamilarasan D, MCA, who provided insight and expertise that greatly helped the research, for suggestions that greatly improved this manuscript.

REFERENCES

[1] Liu, Xian, et al. "Data-driven machine learning in environmental pollution: gains and problems." Environmental science & technology 56.4 (2022): 2124-2133.
[2] Taylan, Osman, et al. "Air quality modeling for sustainable clean environment using ANFIS and machine learning approaches." Atmosphere 12.6 (2021): 713.
[3] Ameer, Saba, et al. "Comparative analysis of machine learning techniques for predicting air quality in smart cities." IEEE access 7 (2019): 128325-128338.
[4] Kjellstrom, Tord, et al. "Air and water pollution: burden and strategies for control." Disease Control Priorities in Developing Countries. 2nd edition (2006).
[5] Freeman III, A. Myrick. "Air and water pollution control: a benefit-cost assessment." (1982).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY