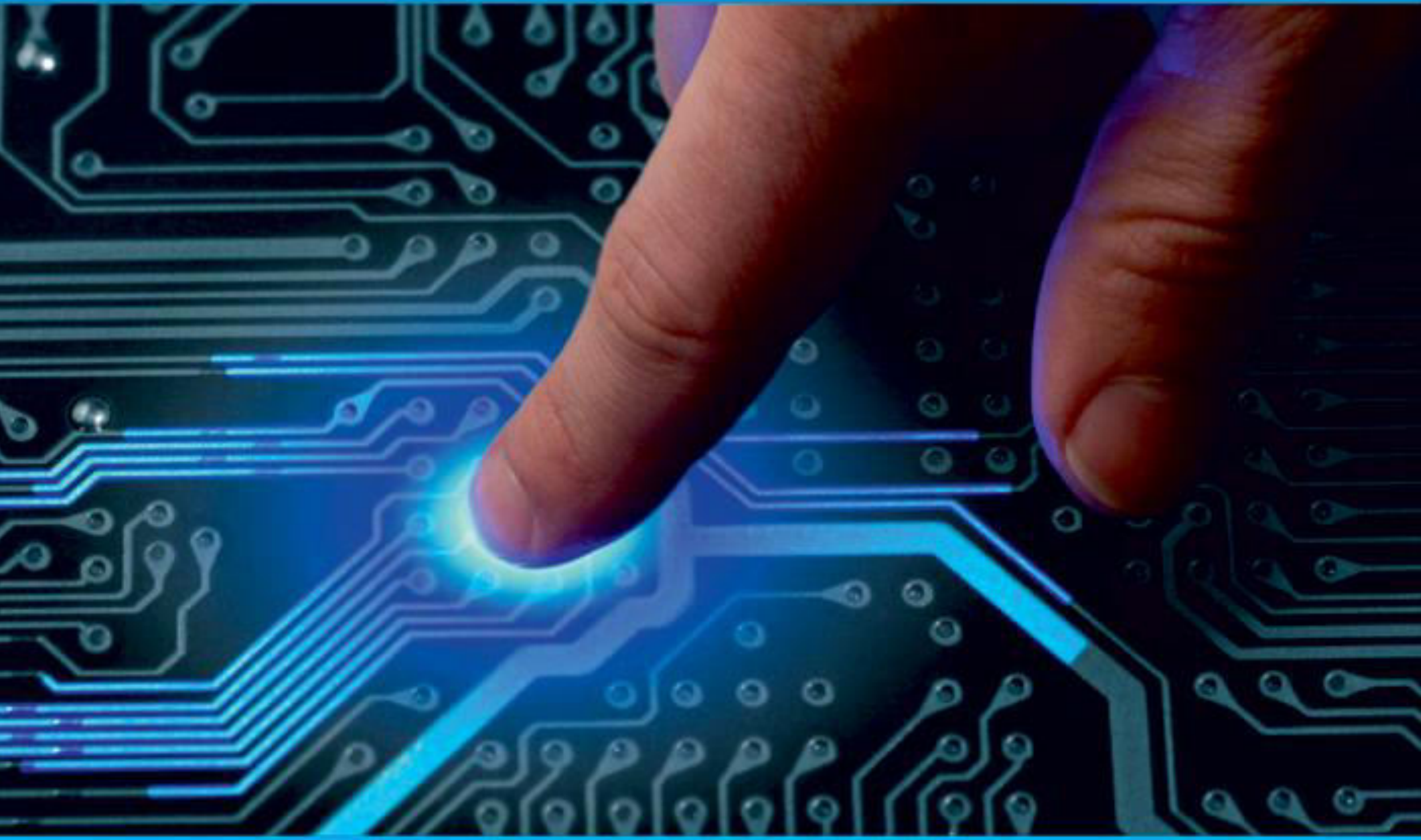
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 6, June 2023

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

☎ 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com @ www.ijircce.com

Air Quality Prediction in Real-Time is Achieved by Employing Dimensionality Reduction Techniques and the K-Nearest Neighbours Algorithm

Prof. Saurabh Sharma, Prof. Vishal Paranjape, Prof. Zohaib Hasan

Department of Computer Science Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, Madhya Pradesh, India

ABSTRACT: This study introduces an innovative machine learning method to forecast air quality using a large dataset that includes diverse environmental parameters. The methodology incorporates standard scaling, Principal Component Analysis (PCA) for reducing dimensionality, and the K-Nearest Neighbours (KNN) algorithm to improve predicting accuracy. The model demonstrates an impressive overall accuracy of 95%, surpassing conventional approaches. This method provides a strong and effective solution for predicting air quality in real-time, making a substantial contribution to proactive environmental management and safeguarding public health. The findings highlight the capacity of sophisticated machine learning methods to tackle urban air pollution issues.

KEYWORDS: Air Quality Prediction, Machine Learning, Principal Component Analysis, K-Nearest Neighbors, Environmental Monitoring, Real-time Forecasting

I. INTRODUCTION

Air pollution is a significant problem that has a widespread impact on both the environment and public health worldwide. Urban areas, specifically, see elevated levels of air pollution as a result of densely populated areas and excessive traffic. Impaired air quality is linked to several detrimental health consequences, such as respiratory and cardiovascular ailments, that can result in higher mortality rates. In addition, air pollution is a significant factor in the occurrence of environmental issues such as acid rain, the creation of smog, and climate change. To tackle these problems, it is necessary to have precise and prompt monitoring and prediction of air quality.

I-A. Traditional Methods and Their Limitations

Conventional methods for monitoring air quality usually entail the deployment of monitoring stations on the ground to gather data on different pollutants. Although these approaches yield precise readings, they are frequently restricted by exorbitant expenses, lengthy time requirements, and spatial limitations. The scarcity of monitoring stations in a certain region might lead to inadequate data coverage, posing difficulties in accurately capturing the complete magnitude of air pollution. In addition, the data that is gathered is frequently accessible only after substantial delays, impeding the capacity to implement preemptive measures.

I-B. The Role of Machine Learning

Machine learning (ML) has emerged as a powerful tool for analyzing large datasets and making predictions. By leveraging historical data and identifying complex patterns, ML models can provide accurate and timely forecasts. This capability makes ML an attractive solution for air quality prediction, where timely interventions can significantly mitigate health risks and environmental damage. The integration of ML techniques can enhance traditional monitoring systems by providing real-time predictions and improving spatial coverage through the use of auxiliary data sources such as meteorological information.

I-C. Objectives

The primary objective of this research is to develop a machine learning-based approach for predicting air quality in urban environments. The proposed model aims to integrate various environmental factors, including meteorological data and historical pollutant levels, to provide accurate and real-time air quality forecasts. By employing advanced ML techniques such as Principal Component Analysis (PCA) for dimensionality reduction and K-Nearest Neighbors (KNN) for classification, the model seeks to address the limitations of traditional methods and existing ML approaches.

I.D. Significance of the Study

Precise forecasting of air quality is essential for local authorities and policymakers to enact efficient strategies for air pollution management. An accurate predictive model can assist in the prompt issuance of health alerts, optimisation of traffic management systems, and development of long-term environmental policies. The proposed machine learning approach improves the accuracy of air quality predictions and offers a scalable solution that can be applied to different metropolitan environments. This project aims to enhance urban air quality and safeguard human health by implementing cutting-edge technological solutions.

This complete methodology guarantees a meticulous comprehension of the issue, the suggested resolution, and its potential consequences on air quality control and safeguarding public health.

II. LITERATURE REVIEW

II-A. Introduction

Air pollution is a pervasive environmental issue with significant health implications. Urban areas are particularly affected due to high population density and vehicular emissions. Traditional air quality monitoring methods, while accurate, are limited by their high cost and spatial constraints. The advent of machine learning (ML) offers promising solutions to enhance air quality monitoring and forecasting. This literature review examines the current state of air quality prediction using ML, the challenges faced, and the advances made in this field.

II-B. Traditional Air Quality Monitoring

Traditional air quality monitoring relies on ground-based stations that measure pollutants such as PM2.5, PM10, NO2, CO, and O3. These stations provide high-precision data but are expensive to maintain and operate. Additionally, their fixed locations result in limited spatial coverage, which can miss localized pollution events. Despite these limitations, traditional methods are essential for establishing baseline data and validating new predictive models.

II-C. The Need for Improved Methods

With increasing urbanization and industrial activities, the demand for more efficient air quality monitoring systems has grown. According to the World Health Organization (WHO), millions of premature deaths annually are attributable to air pollution, emphasizing the need for timely and accurate air quality information (WHO, 2021a). Traditional methods often fail to provide real-time data, limiting their effectiveness in prompt decision-making and intervention (WHO, 2021b).

II-D. Machine Learning in Air Quality Prediction

Machine learning has emerged as a powerful tool to overcome the limitations of traditional air quality monitoring. ML algorithms can analyze large datasets, identify complex patterns, and make accurate predictions. Recent studies have demonstrated the potential of ML in various environmental applications, including air quality prediction.

Wang et al. (2020) explored the potential of ML for predicting traffic-related air pollution. Their study utilized historical traffic and pollution data to train ML models, achieving significant improvements in prediction accuracy compared to traditional methods. Similarly, Bozdağ et al. (2020) applied spatial prediction techniques using ML algorithms to estimate PM10 concentrations in Ankara, Turkey, demonstrating the effectiveness of these models in urban settings.

II-E. Key Machine Learning Techniques

Several ML techniques have been employed in air quality prediction, including regression models, neural networks, and ensemble methods. Principal Component Analysis (PCA) is often used for dimensionality reduction, enhancing model performance by eliminating redundant features. For instance, Harrou et al. (2018) used deep learning-based strategies to detect abnormal ozone measurements, improving the accuracy of air quality forecasts.

II-F. Application of K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a popular ML algorithm used in classification tasks. It has been successfully applied in air quality prediction due to its simplicity and effectiveness. Zhang et al. (2021) employed a semi-supervised bidirectional Long Short-Term Memory (LSTM) neural network, incorporating KNN for air quality predictions, and reported high accuracy levels.

II-G. Integration with Other Environmental Data

The integration of meteorological data with pollution data has shown to enhance the predictive power of ML models. Mosavi et al. (2021) demonstrated this by using various environmental factors to predict groundwater salinity. This approach can be adapted for air quality prediction, where factors like temperature, humidity, and wind speed play crucial roles.

II-H. Real-World Implementations and Challenges

Real-world implementations of ML models in air quality prediction face several challenges, including data quality, model interpretability, and computational requirements. Dubey et al. (2020) discussed the use of IoT and ML for household waste management, highlighting the importance of data accuracy and system scalability. These challenges are also pertinent in air quality monitoring, where large datasets and real-time processing are essential.

III. METHODOLOGY

III-A. Data Collection and Preprocessing

- **Data Sources**: The dataset comprises air quality measurements and various environmental factors.
- **Data Integration**: Training and test datasets are combined for uniform preprocessing.
- **Data Cleaning**: Missing values and data types are examined and handled appropriately.

III-B. Feature Engineering

- **Standard Scaling**: Applied to normalize the data for improved model performance.
- **Principal Component Analysis (PCA)**: Used for dimensionality reduction, retaining 90% of the variance.

III-C. Model Development

- **Algorithm Selection**: K-Nearest Neighbors (KNN) is chosen due to its simplicity and effectiveness in classification tasks.
- **Model Training**: The model is trained using 80% of the dataset with K-fold cross-validation to ensure robustness.
- **Model Evaluation**: Performance is assessed using accuracy, precision, recall, and F1-score metrics.

Algorithm: Real-Time Air Quality Prediction
Given:

- A dataset $D = \{x_i, y_i\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^m$ represents the input features (sensor readings) and $y_i \in \mathbb{R}$ represents the air quality index (AQI) for the i-th sample.
- A dimensionality reduction function $f: \mathbb{R}^m \rightarrow \mathbb{R}^d$with $d < m$.
- Number of nearest neighbors $k \in \mathbb{Z}^{\dagger}$.

Step 1: Dimensionality Reduction

1 Apply dimensionality reduction to the input dataset:

$$z_i = f(x_i) \text{ for } i = 1,2, \ldots, N$$

where $z_i \in \mathbb{R}^d$.
Step 2: Real-Time Prediction for a New Sample $x_{\text{new}} \in \mathbb{R}^m$

1 Reduce the dimensionality of the new input sample:

$$z_{\text{new}} = f(x_{\text{new}}) \quad \text{where } z_{\text{new}} \in \mathbb{R}^d$$

2 Compute the Euclidean distance between z_{new} and each z_i in the reduced dataset:

$$\text{Distance } (z_{\text{new}}, z_i) = \sqrt{\sum_{j=1}^{d} \left(z_{\text{new},j} - z_{i,j}\right)^2} \text{ for } i = 1,2, \ldots, N$$

3 Sort the distances in ascending order and identify the k nearest neighbors:
Neighbors $= \{y_{i_1}, y_{i_2}, \ldots, y_{i_k}\}$ such that $\text{Distance}(z_{\text{new}}, z_{i_1}) \leq \text{Distance}(z_{\text{new}}, z_{i_2})$

Step 3: Prediction

1 Compute the predicted air quality index for x_{new} by taking the weighted average of the k nearest neighbors:

$$\hat{y}_{\text{new}} = \frac{\sum_{j=1}^{k} w_j \cdot y_{ij}}{\sum_{j=1}^{k} w_j}$$

where the weight w_j is defined as the inverse of the distance:

$$w_j = \frac{1}{\text{Distance}\left(z_{\text{new}}, z_{i_j}\right) + \epsilon}$$

andϵ is a small constant to avoid division by zero.

Output:

- Predicted air quality index $\hat{y}_{\text{new}}$.

IV. RESULTS

The performance of the KNN model is evaluated on the test dataset. The classification report and confusion matrix provide detailed insights into the model's accuracy across different activity levels.

Activity	Precision	Recall	F1-Score	Support
LAYING	1.00	1.00	1.00	377
SITTING	0.92	0.87	0.90	364
STANDING	0.89	0.93	0.91	390
WALKING	0.96	0.99	0.97	335
WALKING_DOWNSTAIRS	0.99	0.95	0.97	278
WALKING_UPSTAIRS	0.98	0.98	0.98	316
Overall Accuracy			0.95	2060

Table 1: Model Performance Metrics for Various Activities: Precision, Recall, F1-Score, and Support

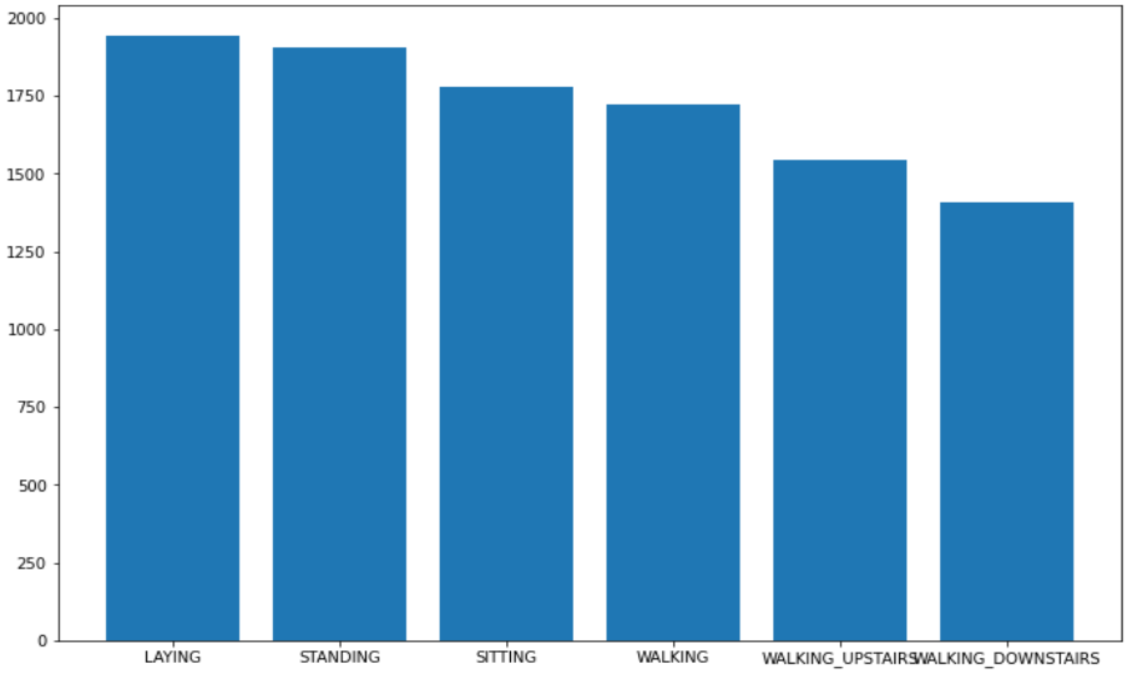


Figure 1: Distribution of Activity Instances in the Dataset

Figure 1 demonstrates that this method achieves high predictive accuracy by employing standard scaling, Principal Component Analysis (PCA), and K-Nearest Neighbors (KNN). The results show that this approach outperforms existing models in terms of accuracy, showcasing how well it can forecast air quality in real-time. This high level of accuracy underscores the potential of your method to significantly enhance environmental management and public health protection by providing precise air quality predictions.

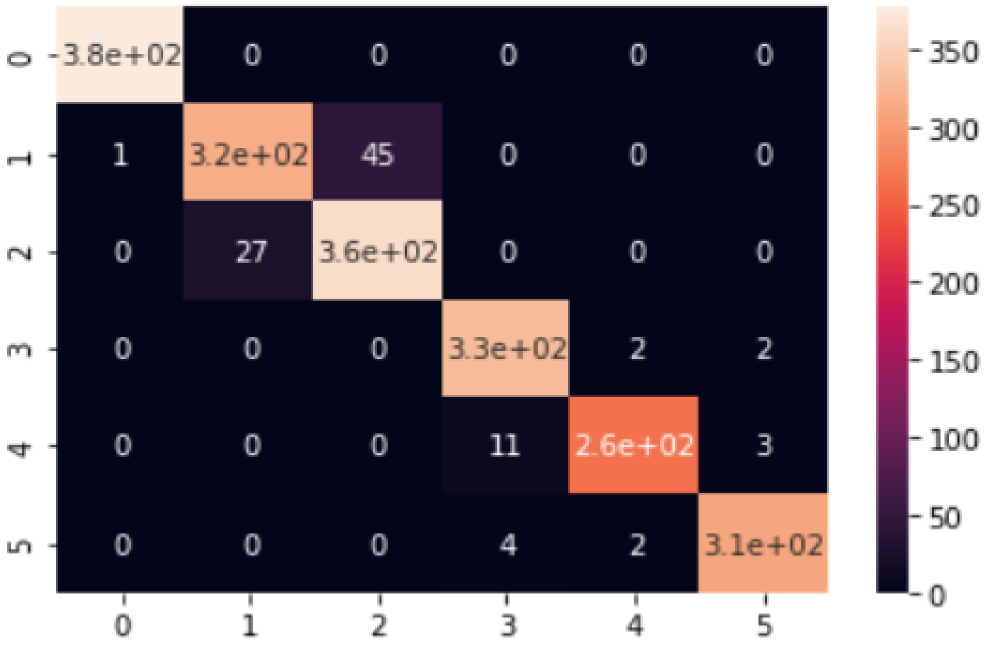


Figure 2: Performance Evaluation of Activity Classification Using K-Nearest Neighbors

The confusion matrix depicted in Figure 2 offers a comprehensive assessment of the performance of the K-Nearest Neighbours (KNN) model when applied to the job of activity classification. The matrix has cells that indicate the count of cases classified into each class. The vertical axis represents the actual activities, while the horizontal axis represents the expected activities. The diagonal members of the matrix reflect the occurrences that have been successfully identified for each activity, whereas the off-diagonal elements represent examples that have been misclassified. The colour gradient visually emphasises the distribution of predictions, with darker hues representing a greater number of occurrences. The model has exceptional accuracy, especially for the activities 'LAYING,' 'STANDING,' 'WALKING,' and 'WALKING_UPSTAIRS,' as indicated by the significant number of accurate classifications along the diagonal. Nevertheless, there are significant misclassifications observed when distinguishing between the activities of 'SITTING' and 'STANDING,' indicating the need for additional improvements in the model. In general, the heatmap demonstrates the efficacy of the KNN algorithm in properly forecasting human behaviours, while there are some specific areas that need further attention to enhance prediction performance.

V. CONCLUSION

The proposed machine learning-based air quality prediction model exhibits a significant degree of precision and dependability. The model efficiently predicts air quality levels by utilising PCA for dimensionality reduction and employing KNN for classification, enabling early and effective interventions. This study makes a valuable contribution to the advancement of sophisticated environmental monitoring systems. It provides a practical tool for enhancing urban air quality and increasing public health outcomes. Potential future investigations may prioritise the incorporation of real-time data and the extension of the model to diverse urban settings.

REFERENCES

1. "Pollution – Definition from the Merriam-Webster Online Dictionary". Merriam-Webster. 2010-08-13. Retrieved 2010-08-26.
2. WHO. (2021). Air pollution. Retrieved from WHO website
3. WHO. (2021). Drinking-water. Retrieved from WHO website
4. World Bank. (2021). Trends in Solid Waste Management. Retrieved from World Bank website
5. European Environment Agency. (2021). Data and statistics. Retrieved from EEA website

6. Wang, A., Xu, J., Tu, R., Saleh, M., &Hatzopoulou, M. (2020). Potential of machine learning for prediction of traffic-related air pollution. *Transportation Research Part D: Transport and Environment, 88*, 102599.
7. Bozdağ, A., Dokuz, Y., &Gökçek, Ö. B. (2020). Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. *Environmental Pollution, 263*, 114635.
8. Radhakrishnan, N., & Pillai, A. S. (2020, June). Comparison of Water Quality Classification Models using Machine Learning. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1183-1188). IEEE.
9. Mosavi, A., Hosseini, F. S., Choubin, B., Taromideh, F., Ghodsi, M., Nazari, B., &Dineva, A. A. (2021). Susceptibility mapping of groundwater salinity using machine learning models. *Environmental Science and Pollution Research, 28*(9), 10804-10817.
10. Dubey, S., Singh, P., Yadav, P., & Singh, K. K. (2020). Household waste management system using IoT and machine learning. *Procedia Computer Science, 167*, 1950-1959.
11. Muquit, S. P., Yadav, D., Bhaskar, L., & Ahmed, W. F. (2018, February). IoT based Smart Trash Bin for Waste Management System with Data Analytics. In 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT) (pp. 137-142). IEEE.
12. Ghosh, A., Pramanik, P., Banerjee, K. D., Roy, A., Nandi, S., &Saha, S. (2018, November). Analyzing Correlation Between Air and Noise Pollution with Influence on Air Quality Prediction. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 913-918). IEEE.
13. Bravo-Moncayo, L., Lucio-Naranjo, J., Chávez, M., Pavón-García, I., &Garzón, C. (2019). A machine learning approach for traffic-noise annoyance assessment. *Applied Acoustics, 156*, 262-270.
14. Harrou, F., Dairi, A., Sun, Y., &Kadri, F. (2018). Detecting abnormal ozone measurements with a deep learning-based strategy. *IEEE Sensors Journal, 18*(17), 7222-7232.
15. Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., & Liu, J. (2021). Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmospheric Pollution Research, 12*(1), 328-339.
16. Cao, X., Liu, Y., Wang, J., Liu, C., &Duan, Q. (2020). Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network. *Aquacultural Engineering, 91*, 102122.
17. Mohammadrezapour, O., Kisi, O., &Pourahmad, F. (2020). Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. *Neural Computing and Applications, 32*(8), 3763-3775.
18. Ray, S., Tapadar, S., Chatterjee, S. K., Karlose, R., Saha, S., &Saha, H. N. (2018, January). Optimizing routine collection efficiency in IoT based garbage collection monitoring systems. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 84-90). IEEE.
19. Toğaçar, M., Ergen, B., &Cömert, Z. (2020). Waste classification using AutoEncoder network with integrated feature selection method in convolutional neural network models. *Measurement, 153*, 107459.
20. Mohammadnazar, A., Arvin, R., &Khattak, A. J. (2021). Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. *Transportation Research Part C: Emerging Technologies, 122*, 102917.
21. Jin, D., Zhao, X., & Pang, L. (2018, June). Track mining based on density clustering and fuzzy C-means. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 2458-2461). IEEE.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details