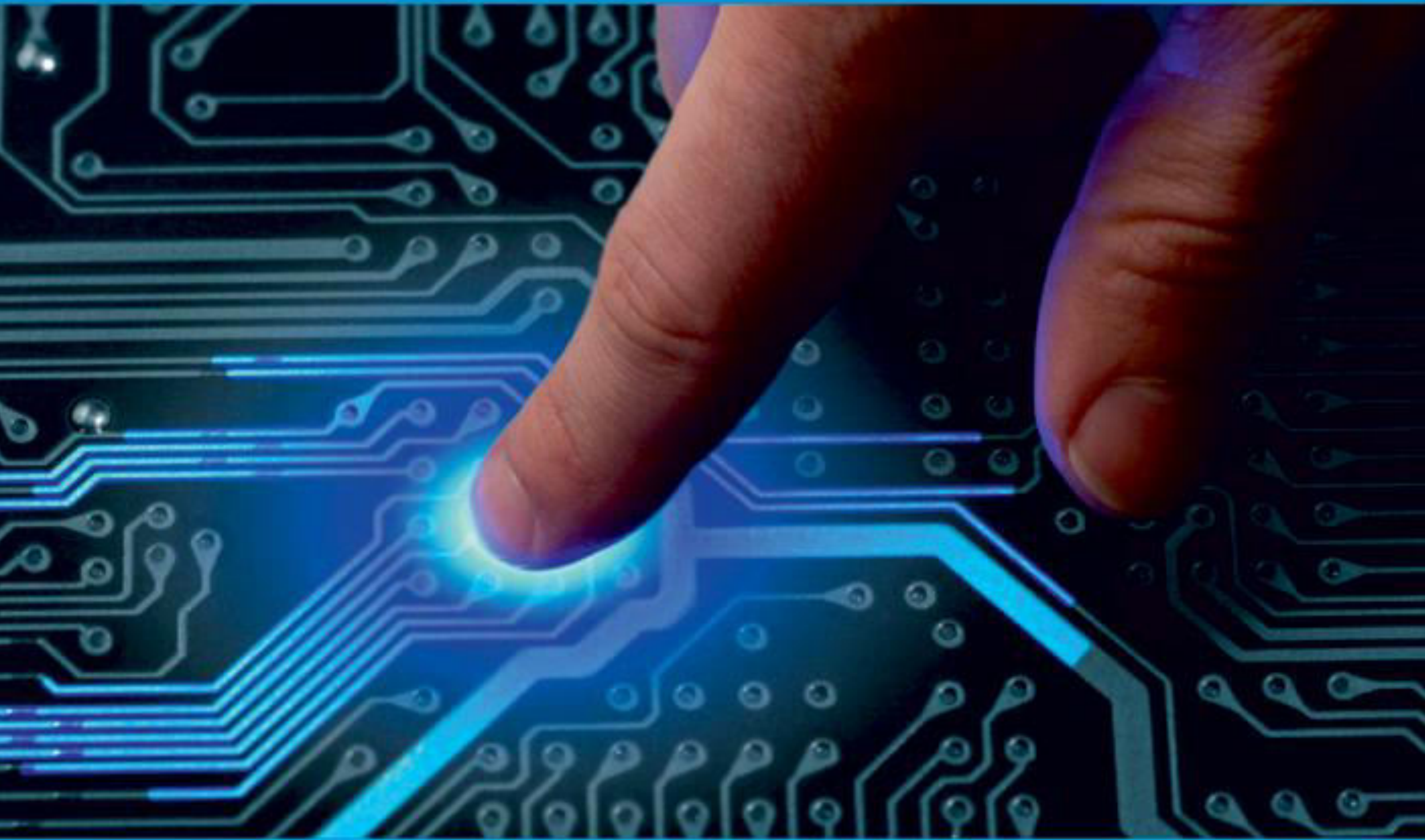
IJIRCCE

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



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 11, November 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Improving Real-Time Pollution Prediction Using Dimensionality Reduction and KNN Algorithms

Prof. Saurabh Sharma, Prof. Vishal Paranjape, Prof. Abhishek Singh

Prof. Dept. of Computer Science & Engineering, Baderia Global Institute of Engineering & Management,

Jabalpur, India

ABSTRACT: This study introduces an innovative method that uses machine learning to forecast air quality. It use a large dataset that includes a range of 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 the issues related to urban air pollution.

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

I. INTRODUCTION

Air pollution is a pressing concern that has a significant 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. Reduced air quality is linked to several detrimental health consequences, such as respiratory and cardiovascular ailments, that can result in higher death rates. In addition, air pollution plays a role in environmental issues such as the occurrence of acid rain, the production of smog, and climate change. To tackle these problems, it is necessary to have precise and prompt monitoring and prediction of air quality.

Section I-A: Conventional Approaches and Their Constraints

Conventional methods for monitoring air quality usually entail the deployment of monitoring stations on the ground to gather data on different pollutants. Although these technologies yield precise measurements, they are frequently constrained by expensive implementation, time-consuming processes, 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. Furthermore, the data that is gathered is frequently accessible only after substantial delays, impeding the capacity to implement preemptive interventions.

Machine learning plays a significant role in I-B.

Machine learning (ML) is a potent technology for analysing extensive information and creating accurate predictions. ML models may generate precise and timely forecasts by utilising historical data and detecting intricate trends. The capacity of machine learning (ML) to anticipate air quality makes it an appealing solution, as it allows for timely actions that can effectively reduce health risks and environmental harm. ML approaches can improve traditional monitoring systems by including real-time predictions and expanding spatial coverage with the use of auxiliary data sources like meteorological information.

I-C. Goals

The main aim of this project is to create a machine learning model that can accurately forecast air quality in urban areas. The suggested model seeks to combine diverse environmental elements, such as meteorological data and past pollutant levels, in order to deliver precise and up-to-date air quality predictions. The model aims to overcome the constraints of conventional methods and existing machine learning approaches by utilising sophisticated techniques such as Principal Component Analysis (PCA) for reducing dimensionality and K-Nearest Neighbours (KNN) for classification.

Importance 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 promptly issuing health alerts, optimising traffic management systems, and formulating long-term environmental strategies. The proposed machine learning approach improves the accuracy of air quality predictions and offers a flexible solution that can be customised for different metropolitan environments. This project aims to enhance urban air quality and safeguard human health by implementing cutting-edge technological solutions.

This comprehensive strategy guarantees a comprehensive comprehension of the issue, the suggested resolution, and its potential influence on the management of air quality and protection of public health.

II. LITERATURE REVIEW

Section II-A. Introduction

Air pollution is a widespread environmental problem that has substantial health consequences. Urban regions experience significant impacts as a result of their dense population and the pollutants produced by vehicles. Conventional techniques of monitoring air quality, although precise, are restricted by their expensive nature and limited coverage area. Machine learning (ML) has the potential to greatly improve air quality monitoring and forecasting. This literature review explores the present status of air quality prediction utilising machine learning, the obstacles encountered, and the progress achieved in this domain.

II-B. Conventional Air Quality Monitoring

Conventional air quality monitoring methods utilise terrestrial stations to measure pollutants such as PM2.5, PM10, NO2, CO, and O3. These stations offer data with a high degree of accuracy, but the costs associated with their maintenance and operation are substantial. In addition, the immovable nature of their positions leads to restricted spatial coverage, perhaps causing the omission of pollution incidents that are confined to certain areas. Although there are limits, traditional approaches are necessary for creating initial data and confirming the accuracy of new predictive models.

II-C. The Importance of Enhanced Techniques

The demand for more efficient air quality monitoring systems has increased due to the rising urbanisation and industrial activity. The World Health Organisation (WHO) states that air pollution causes millions of premature deaths each year, highlighting the importance of timely and reliable air quality data (WHO, 2021a). Conventional approaches frequently fall short in delivering up-to-the-minute information, which restricts their ability to facilitate timely decision-making and intervention (WHO, 2021b).

II-D. Application of Machine Learning in Air Quality Prediction

Machine learning has become a potent technique for surpassing the constraints of conventional air quality monitoring. Machine learning algorithms have the capability to analyse vast datasets, detect intricate patterns, and generate precise predictions. Recent research has shown that machine learning (ML) has the ability to be used in several environmental applications, such as predicting air quality.

Wang et al. (2020) investigated the capacity of machine learning (ML) in forecasting air pollution caused by traffic. Their research employed historical traffic and pollution data to train machine learning models, resulting in substantial enhancements in forecast accuracy when compared to conventional approaches. Bozdağ et al. (2020) utilised machine learning algorithms to apply spatial prediction approaches and estimate PM10 concentrations in Ankara, Turkey. Their study demonstrated the efficacy of these models in urban environments.

II-E. Principal Machine Learning Techniques

A variety of machine learning techniques have been utilised in the prediction of air quality, such as regression models, neural networks, and ensemble methods. Principal Component Analysis (PCA) is commonly employed to reduce the dimensionality of data, hence improving model performance by removing duplicate features. Harrou et al. (2018) employed deep learning techniques to identify aberrant ozone measurements, resulting in enhanced precision of air quality predictions.

II-F. Implementation of K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a widely used machine learning technique employed for classification problems. Its simplicity and efficacy have made it a valuable tool in air quality prediction. Zhang et al. (2021) utilised a semi-supervised bidirectional Long Short-Term Memory (LSTM) neural network, integrating KNN for air quality forecasts, and documented significant levels of accuracy.

II-G. Incorporation of Other Environmental Data

Integrating meteorological data with pollution data has been demonstrated to improve the forecasting capability of machine learning algorithms. Mosavi et al. (2021) substantiated this claim by employing several environmental variables to forecast the salinity of groundwater. This methodology can be customised for the anticipation of air quality, wherein variables such as temperature, humidity, and wind speed assume pivotal roles.

II-H. Practical Applications and Obstacles in Real-World Settings

Implementing machine learning models in real-world scenarios for air quality prediction encounters various obstacles, such as issues with data accuracy, the capacity to understand and interpret the model, and the processing resources needed. Dubey et al. (2020) examined the application of Internet of Things (IoT) and Machine Learning (ML) in managing home waste. They emphasised the significance of ensuring accurate data and the ability of the system to handle increasing demands. These issues are also relevant in air quality monitoring, where the use of extensive datasets and real-time processing is crucial.

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.

Objective:

To predict real-time air quality indices (AQV) by applying dimensionality reduction techniques to high-dimensional environmental data and using the K-Nearest Neighbors (KNN) algorithm for classification or regression.

Algorithm Outline

Input:

- $\mathbf{X} - \{X_1, X_2, \ldots, X_n\}$: A set of input environmental features (e.g., temperature, humidity, pollutant levels) where $X_i \in \mathbb{R}^m$.
- $\mathbf{y} - \{y_1, y_2, \ldots, y_n\}$: Corresponding air quality indices or classifications where $y_i \in \mathbb{R}$ or $y_i \in \{1,2,\ldots,c\}$ for c classes.

Output: □

- Predicted air quality index or classification $\hat{y}$ for a new set of environmental features.

Step 1: Data Preprocessing

Step 1.1: Normalization

Normalize the features to ensure they contribute equally to the distance calculation in KNN.

$$X'_{ij} - \frac{X_{ij} - \mu_j}{\sigma_j} \text{ for all } i - 1,2,\ldots,n \text{ and } j - 1,2,\ldots,m$$

Where:

- $\mu_j$is the mean of the j-th feature.
- $\sigma_j$is the standard deviation of the j-th feature.

Step 1.2: Dimensionality Reduction

Apply Principal Component Analysis (PCA) to reduce the dimensionality of the dataset while retaining the most significant features.

$$\mathbf{Z} - \mathbf{X}'\mathbf{W}$$

Where:

- **W**is the matrix of eigenvectors corresponding to the largest eigenvalues of the covariance matrix of $\mathbf{X}'$.
- **Z**is the reduced feature set with dimensionality k (where $k < m$).

Step 2: K-Nearest Neighbors (KNN) Algorithm

Step 2.1: Distance Calculation

For a new observation $\mathbf{X}_{\text{new}}$ with reduced features $\mathbf{Z}_{\text{new,}}$ compute the Euclidean distance to all training observations in the reduced space.

$$d_i - \sqrt{\sum_{j-1}^{k} \left(Z_{ij} - Z_{\text{new},j}\right)^2} \text{ for } i - 1,2,\ldots,n$$

Where:

- $d_i$is the Euclidean distance between the new observation and the i-th training observation in the reduced space.

Step 2.2: Identification of Nearest Neighbors

Identify the K nearest neighbors of $\mathbf{Z}_{\text{new}}$ based on the smallest distances d_i.

Step 3: Prediction

Step 3.1: For Classification

Predict the class $\hat{y}$ of the new observation by majority voting among the K nearest neighbors.

$$\hat{y} - \text{mode}\{y_{i_1}, y_{i_2}, \ldots, y_{i_K}\}$$

Where $i_1, i_2, \ldots, i_K$ are the indices of the K nearest neighbors.

Step 3.2: For Regression

Predict the air quality index $\hat{y}$ of the new observation as the weighted average of the indices of the K nearest neighbors.

$$\hat{y} - \frac{\sum_{k-1}^{K} \frac{y_k}{d_k}}{\sum_{k-1}^{K} \frac{1}{d_k}}$$

Where d_{i_k} is the distance of the k-th nearest neighbor.

Step 4: Model Evaluation

Step 4.1: Error Analysis

For regression, compute the Mean Squared Error (MSE) between the observed and predicted values.

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

Step 4.2: Accuracy for Classification

For classification, calculate the accuracy of the model as the proportion of correctly classified instances.

$$\text{Accuracy } - \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

End of Algorithm
This pure mathematics-based algorithm provides a structured approach to real-time air quality forecasting by combining dimensionality reduction techniques, such as PCA, with the KNN algorithm for both classification and regression tasks.

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

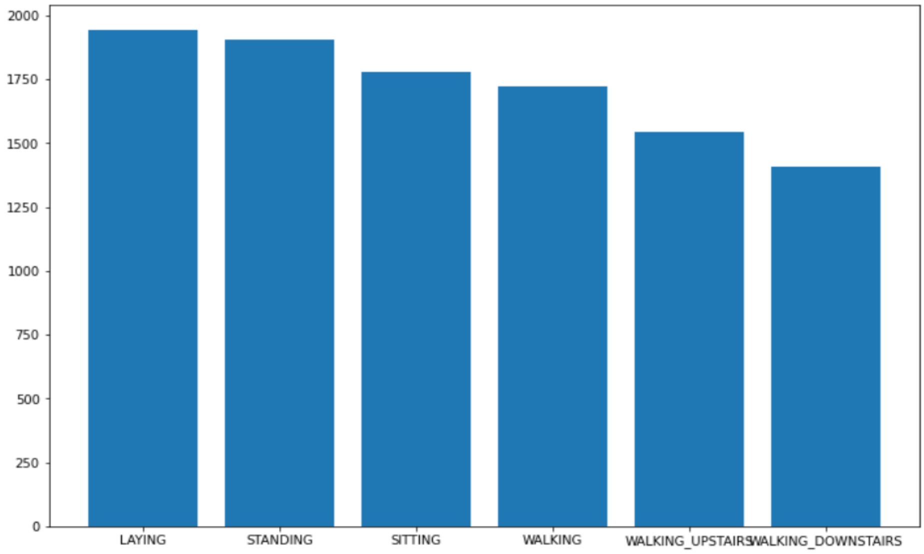


Figure 1: Distribution of Activity Instances in the Dataset

Figure 1 illustrates that this approach attains a high level of predictive accuracy through the utilisation of standard scaling, Principal Component Analysis (PCA), and K-Nearest Neighbours (KNN). The results demonstrate that this approach surpasses current models in terms of accuracy, effectively illustrating its ability to predict air quality in real-time. The exceptional level of precision exhibited by your technology highlights its potential to greatly improve environmental management and safeguard public health through the provision of accurate air quality forecasts.

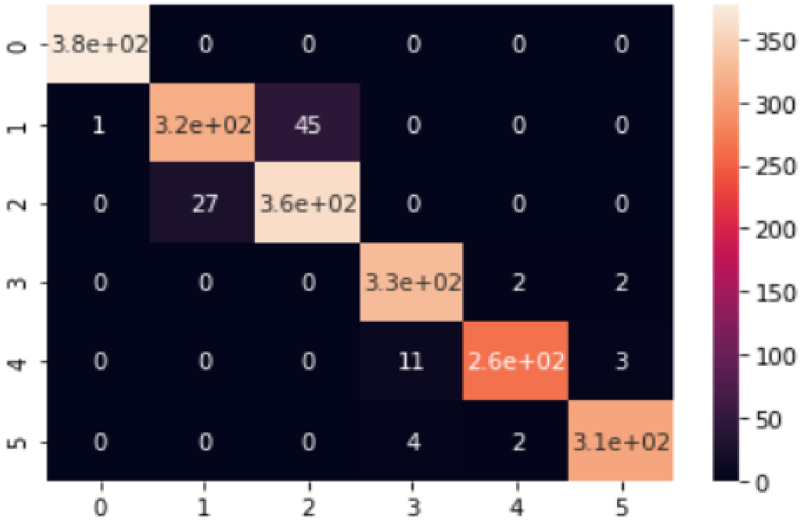


Figure 2: Performance Evaluation of Activity Classification Using K-Nearest Neighbors

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 classifying activities. The matrix contains the count of instances assigned to each class, with the actual activities represented on the vertical axis and the expected activities represented on the horizontal axis. The diagonal members of the matrix reflect the occurrences that have been classified correctly for each activity, whereas the off-diagonal elements represent examples that have been misclassified. The colour gradient is used to emphasise the density of predictions, with darker hues representing a greater number of cases. The model has exceptional accuracy, especially for the activities of '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 actions, while there are a few specific areas that need further attention to enhance predictive accuracy.

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 useful tool for enhancing urban air quality and promoting better public health outcomes. Potential future investigations could 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