**IRSTI 28.23.29**

**https://doi.org/10.51889/2959-5894.2026.94.2.022**

**Omojola Ayogoke Felix, L. Suleimenova[1*], U. Nessipkaliyev[1], Z. Khassenova[1]**

[1]D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan
[*]*e-mail: Lsuleimenova@edu.ektu.kz*

# DEEP LEARNING-BASED AIR QUALITY FORECASTING AND ANOMALY DETECTION USING CNN AND DBSCAN CLUSTERING

*Abstract*

Air pollution poses a serious risk to environmental sustainability and public health, especially in urban agglomerations. Accurate air quality prediction and early anomaly detection are crucial for effective environmental management and preventive medicine. This paper proposes a deep learning-based method that integrates Convolutional Neural Networks (CNNs) for air quality level prediction and DBSCAN clustering for identifying anomalous pollution patterns. A dataset spanning five years (2018–2022), comprising 2,797 samples and 13 attributes, was used to train and evaluate the models. The dataset was divided into 70% for training, 15% for validation, and 15% for testing. The CNN model was trained for 50 epochs using the Adam optimizer with a batch size of 32, employing Mean Squared Error (MSE) as the loss function. The model's performance was assessed using standard metrics, including Accuracy, Precision, Recall, and F1-score. The findings demonstrate consistent and dependable performance, achieving an overall accuracy of 76.0%, precision of 77.5%, recall of 76.2%, and an F1-score of 76.3%. The confusion matrix also indicated areas of strength and weakness, particularly the impact of false negatives on public safety. With a view to improving anomaly sensitivity, the technique proposed in this study has the potential to be used for real-time air quality monitoring and policy support.

**Keywords:** air quality forecasting, deep learning, CNN, DBSCAN clustering, anomaly detection, environmental monitoring.

Оможола Айогоке Феликс[1], Л. Сулейменова[1], У. Несипкалиев[1], З. Хасенова[1]
[1]Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан

# CNN ЖӘНЕ DBSCAN КЛАСТЕРЛЕУІН ҚОЛДАНА ОТЫРЫП, ТЕРЕҢ ОҚЫТУҒА НЕГІЗДЕЛГЕН АУА САПАСЫН БОЛЖАМДАУ ЖӘНЕ АНОМАЛИЯНЫ АНЫҚТАУ

*Аңдатпа*

Ауаның ластануы қоршаған ортаның тұрақтылығы мен қоғамдық денсаулыққа, әсіресе қалалық агломерацияларда, үлкен қауіп төндіреді. Ауа сапасын дәл болжау және аномалияны ерте анықтау қоршаған ортаны тиімді басқару және профилактикалық медицина үшін өте маңызды. Бұл мақалада ауа сапасы деңгейін болжау үшін конволюциялық нейрондық желілерді (CNN) және аномальды ластану үлгілерін анықтау үшін DBSCAN кластерлеуін біріктіретін терең оқытуға негізделген әдіс ұсынылады. Модельдерді оқыту және бағалау үшін бес жылға (2018–2022) созылған 2797 үлгіден және 13 атрибуттан тұратын деректер жиынтығы пайдаланылды. Деректер жиынтығы оқыту үшін 70%, валидация үшін 15% және тестілеу үшін 15% бөлінді. CNN моделі Adam оңтайландырғышын пайдаланып 50 қайтара оқытылды, оның партия өлшемі 32 болды, жоғалту функциясы ретінде орташа квадраттық қате (MSE) қолданылды. Модельдің өнімділігі дәлдік, оң болжамдардың дәлдігі, толықтығы және F1-шамасы сияқты стандартты көрсеткіштерді пайдаланып бағаланды, жалпы дәлдік 76,0%, оң болжамдардың дәлдігі 77,5%, толықтығы 76,2% және F1-шамасы 76,3% жетіп, тұрақты және сенімді өнімділікті көрсетті. Қателіктер матрицасы сонымен қатар күшті және әлсіз жақтарын, әсіресе жалған теріс нәтижелердің қоғамдық қауіпсіздікке әсерін көрсетті. Аномалияға сезімталдықты жақсарту мақсатында, осы зерттеуде ұсынылған бұл әдісті ауа сапасын нақты уақыт режимінде бақылауда және саясатты қолдауда қолданылуы мүмкін.

**Түйін сөздер:** ауа сапасын болжау, терең оқыту, CNN, DBSCAN кластерлеуі, аномалияны анықтау, қоршаған ортаны бақылау.

Оможола Айогоке Феликс[1], Л. Сулейменова[1], У. Несипкалиев[1], З. Хасенова[1]
[1]Восточно-Казахстанский технический университет имени Д. Серикбаева,
г.Усть-Каменогорск, Казахстан

**ПРОГНОЗИРОВАНИЕ КАЧЕСТВА ВОЗДУХА И ОБНАРУЖЕНИЕ АНОМАЛИЙ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ CNN И КЛАСТЕРИЗАЦИИ DBSCAN**

*Аннотация*

Загрязнение воздуха представляет собой большую угрозу для экологической устойчивости и общественного здоровья, особенно в городских агломерациях. Точное прогнозирование качества воздуха и раннее обнаружение аномалий имеют решающее значение для эффективного управления окружающей средой и профилактической медицины. В данной статье предлагается метод на основе глубокого обучения, который интегрирует сверточные нейронные сети (CNN) для прогнозирования уровня качества воздуха и кластеризацию DBSCAN для выявления аномальных моделей загрязнения. Для обучения и оценки моделей использовался набор данных за пять лет (2018–2022), состоящий из 2797 образцов и 13 атрибутов. Набор данных был разделен на 70% для обучения, 15% для валидации и 15% для тестирования. Модель CNN обучалась с использованием оптимизатора Adam с размером пакета 32, используя среднеквадратичную ошибку (MSE) в качестве функции потерь. Производительность модели оценивалась с помощью стандартных метрик, включая точность, точность положительных предсказаний, полноту и F1-меру. Результаты демонстрируют стабильную и надежную работу, достигая общей точности 76,0%, точности положительных предсказаний 77,5%, полноты 76,2% и F1-меры 76,3%. Матрица ошибок также выявила сильные и слабые стороны, в частности, влияние ложноотрицательных результатов на общественную безопасность. Предлагаемая в данном исследовании методика, направленная на повышение чувствительности к аномалиям, имеет потенциал для использования в мониторинге качества воздуха в режиме реального времени и для поддержки разработки политики в этой области.

**Ключевые слова:** прогнозирование качества воздуха, глубокое обучение, CNN, кластеризация DBSCAN, обнаружение аномалий, мониторинг окружающей среды.

## Introduction

Air quality is one of the key factors affecting human health and the overall well-being of society. Continuous technological development and production growth inevitably lead to changes in air quality. In this regard, air quality monitoring is constantly under scrutiny by both the public and relevant government agencies. Air pollution still poses a serious threat to ecosystems, the economy, and human health worldwide. The World Health Organization (WHO) estimates that outdoor air pollution kills roughly 5.7 million lives annually [1]. According to World Bank statistics, air pollution causes more than 10,000 premature deaths and more than $10.5 billion in economic damage per year [2]. Accurate forecasts of air quality, accounting for the complex dynamics of atmospheric processes, are crucial for reducing these consequences, enabling prompt action, and guiding policy decisions. The nonlinear and complex dynamics of atmospheric processes are often overlooked by traditional air quality forecasting techniques, such as numerical and statistical models. Scientific advances in big data processing and increased computing resources make it possible to apply deep learning methods, including convolutional neural networks and clustering algorithms, to address problems in air quality monitoring and forecasting, anomaly detection, and the modeling of complex relationships among variables. The growing impact of urbanization and industrial activities has exacerbated air pollution and underscored the need for reliable air quality monitoring systems. Accurate forecasting and early detection of abnormal pollution events are critical for protecting public health and supporting environmental decision-making. In this regard, deep learning and anomaly detection methods offer promising capabilities for analyzing complex environmental data and improving monitoring effectiveness. This section proposes a deep learning-based technique that combines Convolutional Neural Networks for predicting air quality levels with a density-based clustering algorithm (DBSCAN) to recognize anomalous pollution patterns with high accuracy. For multivariate environmental data (PM2.5, PM10), gaseous pollutants, and meteorological indicators, the CNN component is used to automatically extract complex spatiotemporal characteristics. Thus, the model can pick up nonlinear relationships and subtle fluctuations that classical statistical techniques may overlook. DBSCAN does not require a priori specification of the number of clusters and can identify arbitrarily shaped clusters and noise points, making it a flexible and reliable framework for anomaly detection in heterogeneous and noisy air quality datasets.

Deep learning, a subset of machine learning, uses multi-layer neural networks to learn abstract features from unprocessed data. Its application in air quality prediction has gained momentum because it can learn intricate temporal and spatial dependencies. Convolutional Neural Networks (CNNs), initially designed for image processing, have been adapted for time-series forecasting tasks and have proved capable of learning local patterns from air quality data. For instance, deep learning has employed CNNs for pollutant concentration forecasting, achieving higher accuracy than traditional methods (Qi et al., 2020). A systematic review by Qi et al. (2020) identifies deep learning methods as the future of air quality prediction, as they are capable of handling vast amounts of data and capturing complex relationships among variables. The paper mainly focuses on a number of deep learning models, including Recurrent Neural Networks (RNNs) and hybrid networks, applied to air quality prediction problems. These models have been shown to be better at extracting temporal patterns compared to conventional statistical methods.

Anomaly detection is also critical in air quality monitoring to identify unusual occurrences, sensor failures, or data recording errors. Anomaly detection ensures the integrity of data used in forecasting and policy development. Statistical approaches used in conventional anomaly detection are unlikely to identify complex patterns in high-dimensional data. To address these challenges, unsupervised machine learning algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) have been used. DBSCAN is considered useful for identifying anomalies in air quality data because it can recognize clusters of varying densities in the presence of noise (Liu et al., 2023). By grouping data points by density, DBSCAN can identify outliers that may represent adverse events. A number of studies have pursued this approach to enhance the quality of air pollution data and improve the accuracy of subsequent analysis.

The integration of deep learning models, such as Convolutional Neural Networks (CNNs), with clustering algorithms like DBSCAN creates a powerful framework for forecasting air quality and detecting anomalies. CNNs can be utilized to forecast pollutant levels by analyzing historical data, effectively capturing intricate temporal patterns. The residuals, or the differences between expected and actual values, can then be examined using DBSCAN to find abnormalities in the dataset.

The study tested the hypothesis that integrating CNN forecasting with DBSCAN clustering yields significant improvements in forecasting performance and more reliable anomaly detection compared to traditional statistical and standalone machine learning methods. These methods often fall short in capturing nonlinear pollutant behavior or reliably identifying abnormal pollution events. This research investigated the effectiveness of a hybrid deep learning approach that combines Convolutional Neural Networks (CNN) for forecasting with DBSCAN clustering for anomaly detection. The purpose of the study was to develop a unified model that improves prediction accuracy while simultaneously detecting pollution spikes.

*Literature Review*

Air pollution poses a major environmental and public health problem at a global scale. Accurate data anomaly detection in air quality, as well as forecasting, is crucial for mitigating pollution and informing policy guidance [3]. Statistical models have long been used to forecast air quality, but they are insufficient to capture the complex, nonlinear patterns in atmospheric observations [4]. Machine learning (ML) and deep learning (DL) methods have demonstrated their effectiveness as robust techniques for recognizing complex temporal and spatial relationships [5]. This review focused on techniques for predicting air quality, the applications of deep learning, and the use of clustering algorithms such as DBSCAN to detect anomalies. Conventional Techniques used for forecasting pollutant concentrations include Numerical models, e.g., Chemical Transport Models (CTMs), which simulate atmospheric processes to predict pollutant concentrations. Although effective, they are highly computationally intensive and require considerable domain expertise [6]. Statistical methods, including multiple linear regression (MLR) and autoregressive integrated moving average (ARIMA), have been widely used but are prone to failing to handle nonlinearity and interactions between variables [7].

Recent developments have incorporated hybrid models that combine statistical methods with machine learning approaches to enhance the precision of air quality forecasts. These hybrid methods leverage the interpretability of statistical models while benefiting from the pattern recognition capabilities inherent in machine learning techniques. Ensemble machine learning approaches like Random Forest and Gradient Boosting Machines have been used to model interactions among variables and reduce overfitting, achieving greater accuracy than traditional models across a wide array of real-world challenges [8]. Deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated exceptional performance in detecting spatiotemporal patterns of pollutants, owing to their proficiency in handling high-dimensional data [9]. Also, attention models and graph neural networks are becoming useful for analyzing inter-regional transport and the impacts of air pollution [10]. The integration of satellite imagery and IoT-connected air quality sensors has improved the resolution and breadth of data sets, enabling more accurate forecasting and anomaly detection. These advancements support real-time monitoring and inform decisions related to public health initiatives and environmental regulations [11]. As the demand for cleaner air intensifies, data-driven smart solutions will increasingly influence air quality management strategies.

*Machine Learning and Deep Learning Techniques Used for Air Quality Forecasting*

The advent of machine learning techniques, such as Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs), has improved predictive accuracy over traditional statistical models; however, their application has been constrained by the need for large datasets. Deep learning has greatly improved predictive capacity, helping model complex correlations in air quality data [5]. Deep learning employs multi-layer neural networks that were developed from machine learning for pattern recognition in high-dimensional data sets. Convolutional Neural Networks (CNNs), initially developed for image processing but now adapted for time-series forecasting, are among the most common architectures used in air quality forecasting. This enables the detection of spatial patterns in air quality data [12]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks are especially effective at capturing temporal dependencies in air pollution data and provide more precise predictions than traditional models [3]. Hybrid Models: The application of CNNs together with LSTMs has also shown promising predictive performance by capturing spatial and temporal dependencies [5]. Despite these developments, DL models remain computationally expensive and require enormous amounts of labeled data for training. In addition, their "black-box" nature limits interpretability [6].

Anomaly detection is needed for getting outlier events caused by external influences, faulty sensors, or irregular data in air quality observations. Traditional statistical approaches to anomaly detection are ineffective for high-dimensional data. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) has gained prominence because it can identify anomalies without prior knowledge of the data distribution [12]. DBSCAN separates normal observations and outliers based on data points of varying densities. DBSCAN has been used on air quality data to detect pollution spikes and sensor faults [6]. Moreover, using DBSCAN alongside deep learning algorithms can enhance the accuracy of air quality prediction by identifying anomalies in predicted values [5].

Recent studies have explored combining DBSCAN with sophisticated deep learning models to improve anomaly detection in air pollution datasets, particularly in the context of smart city initiatives [2, 13]. These hybrid models can identify unusual pollution incidents and sensor errors in real time, thereby enhancing data integrity and public response mechanisms. Also, unsupervised deep learning techniques, such as autoencoders, have proven effective in anomaly detection by learning concise representations of normal behavior and identifying deviations ([14], [15]). Such advancements are crucial for environmental monitoring systems that rely on precise, high-frequency data for effective prediction and decision-making.

**Research methodology**

The technique employed in the present study focuses on achieving accurate air quality prediction and anomaly detection through deep learning methods, using Convolutional Neural Networks (CNNs) and the DBSCAN clustering algorithm. The block diagram for the proposed study is shown in Figure 1.

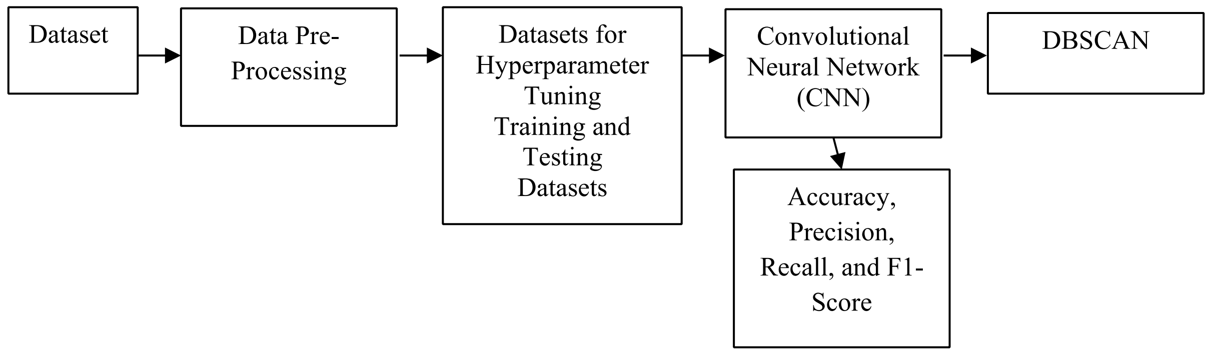


*Figure 1. Block Diagram for the Proposed Study*

*Dataset Description.* This study uses a multivariate environmental dataset for the period from January 2018 to December 2022. The dataset includes air quality variables (PM2.5, PM10, CO, $NO_2$, $SO_2$, and $O_3$) along with meteorological parameters such as temperature, humidity, wind speed, barometric pressure, dew point, rainfall, and solar radiation. Meteorological observations were obtained by the NOAA Aviation Meteorological Center and integrated with air quality measurements to improve forecast accuracy. The final dataset contains 2,797 records and 13 attributes.

*Dataset Preprocessing.* To maintain uniformity in the distribution of continuous features, missing values were substituted with the mean of each specific feature. Additionally, given that the dataset includes time-series data, missing values were imputed using the most recent available value to preserve existing trends. Data records with excessive missing data, above a specified limit, were removed to prevent distortion of the dataset. The dataset contains atmospheric pollutant concentrations (PM2.5, PM10, CO, NO2, SO2, and O3) along with meteorological variables. To ensure reproducibility, the source of environmental observations and the data preprocessing algorithm must be described in detail.

*Data Normalization.* To ensure uniformity in input features, Min-Max Scaling was applied, transforming the values to a [0,1] range using the formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

where:

$X$ is the original feature value,

$X_{max}$ , $X_{min}$ are the minimum and maximum values of the feature, respectively.

The CNN model was chosen as one of the most effective models for detecting spatial patterns in air pollution data. Its architecture includes an input layer to normalize air-quality characteristics and provides a uniform scale to improve model convergence. A convolutional layer for extracting significant spatial features using cores that help recognize patterns in the distribution of pollutants and meteorological variables. Pooling layers reduce dimensionality while preserving essential features, thereby increasing computational efficiency and generalizing ability. The fully connected layer integrates the extracted features to produce the final prediction, enabling the model to capture complex relationships in the data. The output layer generates air quality forecasts, providing expected pollution levels. The detailed architecture of the CNN-Based Air Quality Forecasting Model is described in Figure 2. The architecture consists of two one-dimensional convolutional (Conv1D) layers. The first convolutional layer contains 32 filters with a 3x3 kernel and a ReLU activation, enabling the network to extract local features from the input sequence. The second convolutional layer contains 64 filters with the same kernel size and activation function, allowing the network to learn more complex tasks.

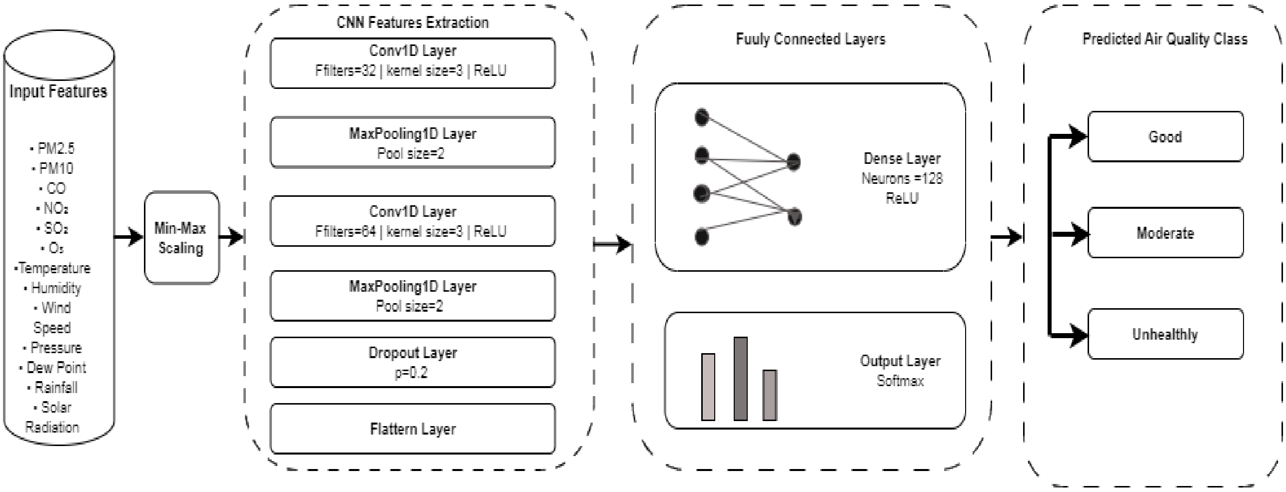


*Figure 2. Detailed architecture of the CNN-Based Air Quality Forecasting Model*

*Anomaly Detection Using DBSCAN*

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was used for anomaly detection to identify unusual spikes in pollution. DBSCAN groups points based on density and labels outliers. The dataset processing procedure based on the DBSCAN algorithm is presented in Figure 3.

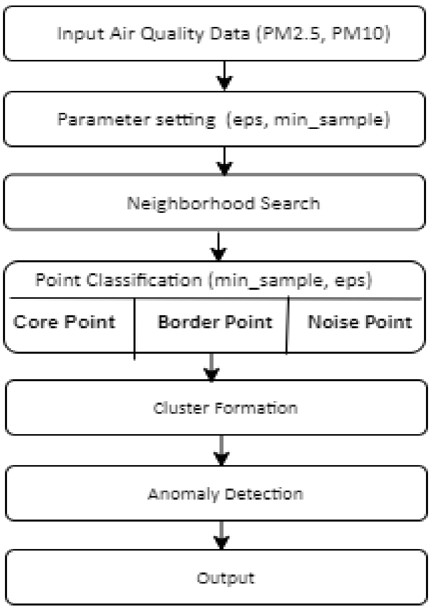


*Figure 3. DBSCAN algorithm for anomaly detection*

For anomaly detection, DBSCAN was configured with parameters eps = 0.20 and min_sample = 5. These values were chosen experimentally based on the density distribution of normalized PM2.5 and PM10 observations and were found to effectively discriminate between normal air-quality readings and anomalous pollution events.

Figure 3 illustrates the DBSCAN-based framework used to identify anomalous air pollution events from observed PM2.5 and PM10 concentrations. The algorithm first defines the neighborhood radius (eps) and the minimum number of neighboring points (min_sample), then classifies observations as core, border, or noise points based on local density. Density-connected observations are grouped into clusters that represent normal air-quality patterns, while isolated noise points are identified as anomalous pollution events.

*Model Training and Evaluation*

*Splitting the Data.* The dataset is divided into three separate sets to achieve efficient model training and testing. To train the model and identify patterns in the data, the training set, which comprises about 70% of the entire dataset, is used. To help avoid overfitting, the validation set, which comprises 15% of the dataset, is used to tune hyperparameters and evaluate the model's performance during training. The test set, which is the remaining 15% of the dataset, is reserved for the final evaluation of the model's predictive power on unseen data. Instead of just memorizing the training data, this structured split ensures that the model can generalize to new inputs.

*Performance Metrics*

(i) Accuracy is the ratio of times that the model correctly predicts whether or not there is pollution. The model gets the right prediction most of the time when accuracy is high, but it can be a poor measure of performance if the data set is biased (e.g., far more "no pollution" instances than "pollution" instances).

$$Accuracy = \frac{(TP+TN)}{(\mathrm{TP+TN+FP+FN})} \tag{2}$$

(ii) Precision informs us about how well the model performs at being accurate when it reports that there is pollution. Precision is high when the model reports there is pollution; it is frequently accurate. This is useful when false alarms (false positives) need to be prevented, e.g., when unnecessary environmental alerts could cause panic.

$$Precision = \frac{TP}{(\mathrm{TP+FP})} \tag{3}$$

(iii) Recall (Sensitivity) = ability of the model to detect real pollution cases. Good recall prevents most pollution cases from being undetected, so there will be fewer chances to miss high-hazard levels of pollution. Important in applications involving public health, where an incorrect failure to detect a true case of pollution (false negative) can have catastrophic consequences.

$$Recall = \frac{TP}{(\mathrm{TP+FN})} \tag{4}$$

(iv) F1-Score is evenly weighted between precision and recall, which provides an aggregate measure that is easy to use when false negatives and false positives are both at stake. A high F1-score indicates that the model performs well at correctly identifying pollution, with not too many false positives or false negatives.

$$F1 - Score = \frac{2*Precision*Recall}{(\mathrm{Precision+Recall})} \tag{5}$$

*Confusion Matrix*

The confusion matrix is used for computing Precision, Recall, F1-score, and Accuracy. It compares actual values with predicted values using the following: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as shown in Table 1. Since the dataset is about air quality, we categorized air quality into classes based on pollutant levels: Good, Moderate, and

Unhealthy. PM2.5 is a common indicator of air quality; therefore, to predict air quality classes based on PM2.5 levels, we use the following classes:

$$Good\ (0): PM2.5 \leq 12\, mg/m^3$$
$$Moderate\ (1): 12 < PM2.5 \leq 35\, mg/m^3$$
$$Unhealthy\ for\ Sensitive\ Groups\ (2): 35 < PM2.5 \leq 55\, mg/m^3$$
$$Unhealthy\ (3): 55 < PM2.5 \leq 150\, mg/m^3$$
$$Very\ Unhealthy\ (4): PM2.5 > 150\, mg/m^3$$

**Results of the study**

From Table 1 and the confusion matrix shown in Figure 4, we can derive the following insights:

(i) True Positives (TP = 589): These instances represent cases where the model accurately identified poor air quality as poor. The high TP count indicates that the model is good at noticing harmful air conditions.

(ii) True Negatives (TN = 171): In these cases, the model rightly predicted good air quality when it was really good, showing its effectiveness in identifying favorable air conditions.

(iii) False Positives (FP = 57) (Type I Error): This refers to instances where the model incorrectly indicated poor air quality, despite the actual conditions being good. A high FP rate would cause the system to issue unnecessary air pollution alerts, reducing people's reliance on the system.

(iv) False Negatives (FN = 183) (Type II Error): The model incorrectly predicted good air quality, whereas in reality it was bad. This is a serious issue, as it suggests the model failed to detect harmful air quality, which could have led to significant health risks.

*Table 1. Confusion Matrix Table for Air Quality Classification*

| *Actual \ Predicted* | *Negative (0)* | *Positive (1)* |
|---|---|---|
| Negative (0) | 171 (True Negative - TN) | 57 (False Positive - FP) |
| Positive (1) | 183 (False Negative - FN) | 589 (True Positive - TP) |

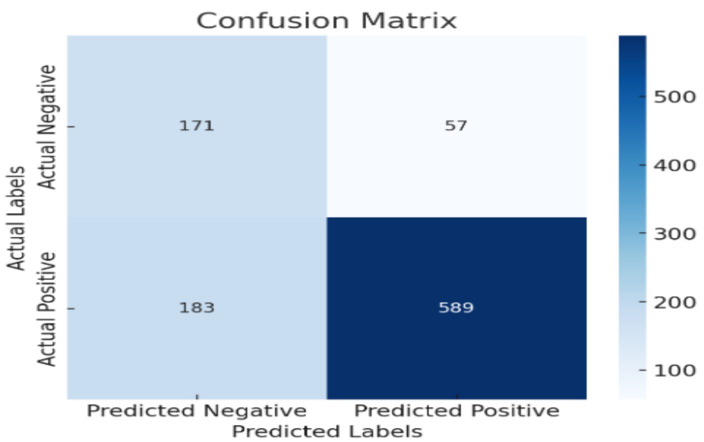


*Figure 4. Confusion Matrix Diagram for Air Quality Classification*

The bar plot comparing Accuracy, Precision, Recall, and F1-Score for the air quality classification model is shown in Figure 4. The bar plot shows that (i) Accuracy (76.0%): Percent of correctly classified instances of all cases. (ii) Precision (77.5%): Percent of correctly predicted poor air quality cases (TP) of all predicted poor air quality cases (TP + FP). High precision results in fewer false alarms. (iii) Recall (76.2%): The percentage of actual poor air quality instances that were accurately predicted. A lower recall indicates that the model fails to capture some dangerous air conditions. (iv) F1-Score (76.3%): The trade-off between precision and recall, indicating the model's overall performance (Figure 5).

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was applied to identify anomalous air pollution patterns using observed values of PM2.5 and PM10 concentrations.

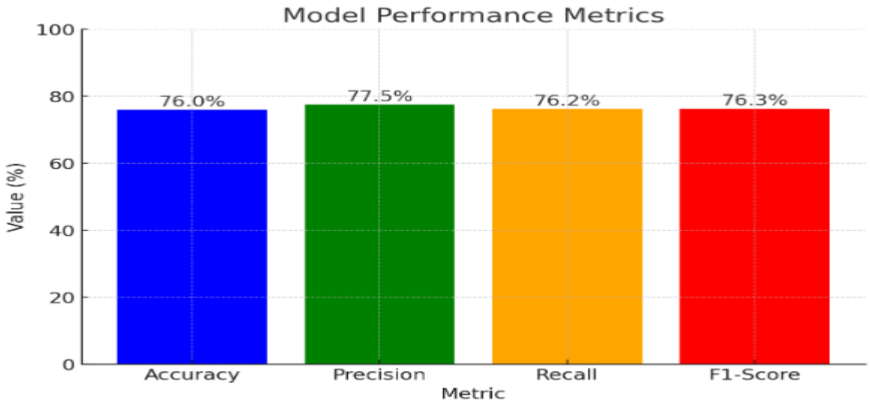


*Figure 5. Bar Chart showing Performance Metrics for Air Quality Classification*

In the scatter plot in Figure 5, each point represents a PM2.5 vs. PM10 measurement from the dataset. Clusters are assigned different colors to indicate normal groupings of air quality data by density. The points marked as -1 (usually shown in a unique color like grey or black) is considered outliers or anomalies by DBSCAN. These outlier data points represent atypical pollution levels, such as sudden increases in PM2.5 and PM10, that significantly deviate from typical air quality patterns. These irregularities are essential, as they may indicate pollution incidents such as industrial emissions, traffic jams, or environmental threats that conventional forecasting models might overlook [16]. Therefore, by integrating DBSCAN with CNN forecasting, this study establishes a more comprehensive framework for identifying both anticipated and unforeseen changes in air quality, thereby facilitating prompt alerts and policy interventions (Figure 6).

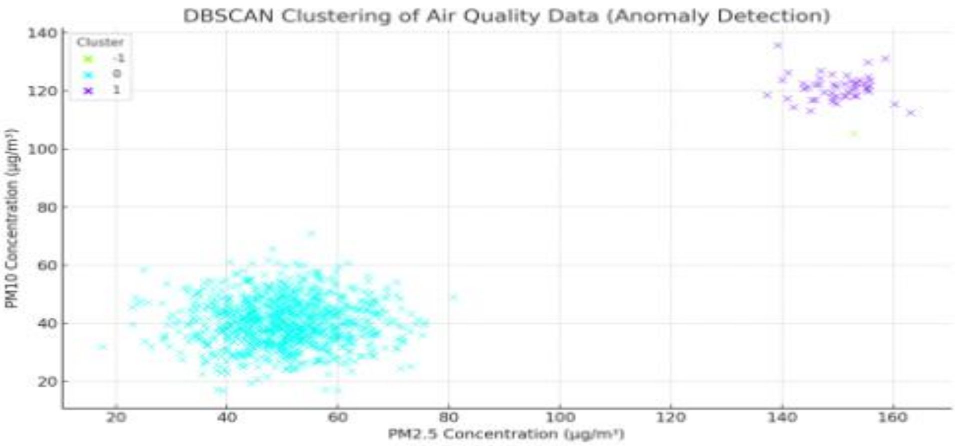


*Figure 6. DBSCAN Clustering for Anomaly Detection in Air Quality*

**Discussion**

The results demonstrate that the proposed CNN-based framework effectively learns complex nonlinear relationships between air pollutant concentrations and meteorological variables. The balanced performance across all evaluation metrics indicates that the model can reliably distinguish between different air quality conditions without exhibiting a strong bias toward either positive or negative classifications. Such behavior is particularly important in environmental monitoring applications, where both missed pollution events and excessive false alarms can adversely affect decision-making. The confusion matrix analysis further highlights the practical applicability of the proposed approach. Although some pollution events went undetected, the model successfully identified most hazardous air quality conditions. This suggests that convolutional neural networks can extract meaningful patterns from multivariate environmental data and provide robust predictions even in the presence of variability and uncertainty inherent in atmospheric observations. These findings are consistent with previous studies reporting the effectiveness of deep learning models for air quality forecasting and environmental monitoring. The integration of DBSCAN clustering complements the forecasting component by enabling the identification of unusual pollution patterns that may not follow normal temporal trends. Unlike supervised prediction models that primarily focus on forecasting accuracy, the proposed framework adds an additional layer of analysis through anomaly detection. This capability is particularly valuable for identifying unexpected pollution episodes, sensor malfunctions, or abrupt environmental changes that require immediate attention.

Scientific Significance of the Results. The scientific contribution of this study lies in integrating CNN-based air quality forecasting and DBSCAN-based anomaly detection into a unified analytical framework. While many existing studies focus exclusively on pollutant prediction, the proposed approach simultaneously addresses forecasting and anomaly identification, providing a more comprehensive understanding of air quality dynamics. The obtained results confirm that convolutional neural networks can effectively capture nonlinear relationships between pollutant concentrations and meteorological variables, whereas DBSCAN can identify unusual pollution events without requiring predefined class labels. This combination improves the capability of environmental monitoring systems to detect both expected pollution trends and unexpected hazardous events, thereby contributing to the development of intelligent decision-support tools for environmental management and smart-city applications.

**Conclusion**

This study presented a deep learning–based framework for air quality forecasting and anomaly detection using Convolutional Neural Networks (CNN) and DBSCAN clustering. The experimental results demonstrate that the proposed model achieves balanced and reliable classification performance, with an accuracy of 76.0%, precision of 77.5%, recall of 76.2%, and an F1-score of 76.3%. Unlike conventional air quality forecasting approaches that focus primarily on prediction accuracy, the proposed framework combines forecasting and anomaly detection capabilities within a single analytical workflow. This integration enables not only the prediction of future air quality conditions but also the identification of unusual pollution events, providing a more comprehensive assessment of environmental conditions and supporting proactive environmental management.

Integrating CNN forecasting with DBSCAN clustering effectively captures complex nonlinear relationships in air quality data and provides a reliable balance between detecting hazardous pollution events and limiting false alarms. The analysis of the confusion matrix indicates high system reliability, ensuring effective detection of poor air quality while minimizing false-positive classifications. The balance between high sensitivity and low false alarms is essential for real-world air quality monitoring systems. Skipping pollution detections can pose a serious threat to public health, while an excessive number of warnings can reduce user confidence and the system's overall practical value. The results confirm the applicability of the proposed approach to early warning and decision support in environmental management. In addition, the use of DBSCAN clustering demonstrates the model's ability to identify abnormal pollution patterns without prior assumptions about the data distribution. Allocating clusters with increased concentrations of PM2.5 and PM10 enables effective detection of unusual and extreme pollution episodes that may not be accurately identified by controlled classification methods alone. The integration of unsupervised clustering with a deep learning model increases the system's overall stability, especially under noise, emissions, and unsteadiness in sensory data. In general, the results of the study confirm the expediency of combining deep learning-based forecasting with unsupervised analysis methods for a comprehensive assessment of air quality. The proposed architecture forms the basis for building intelligent, scalable, and reliable monitoring systems for applications in the infrastructure of "smart cities", environmental management, and anomaly detection.

*Limitations of the Study and Future Work*

Although the proposed CNN–DBSCAN system demonstrates high performance, it still has some drawbacks: it generates a significant number of false positives. However, improving pollution detection remains a priority. Future work can build on this study by increasing the number of air pollutant sources and weather variables to better understand and predict non-standard air quality changes. Using adaptive clustering methods in real time can help the system respond faster to changing urban conditions and reduce missed detections, making air quality monitoring more reliable for environmental management.

*References*


*[1] World Health Organization. (2021). Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease. Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.*

*[2] Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: A review. Current Pollution Reports, 6(2), 1 – 11. https://doi.org/10.1007/s40726-020-00159-z*

*[3] Rollo, F., Bachechi, C., & Po, L. (2023). Anomaly detection and repairing for improving air quality monitoring. Sensors, 23(2), 640. https://doi.org/10.3390/s23020640*

*[4] Haehnel, P., Marecek, J., Monteil, J., & O'Donncha, F. (2018). Using Deep Learning to Extend the Range of Air-Pollution Monitoring and Forecasting. arXiv preprint arXiv:1810.09425.*

*[5] Xu, R., Wang, D., Li, J., Wan, H., Shen, S., & Guo, X. (2023). A hybrid deep learning model for air quality prediction based on the time–frequency domain relationship. Atmosphere, 14(2), 405. https://doi.org/10.3390/atmos14020405*

*[6] Murad, A., Kraemer, F. A., Bach, K., & Taylor, G. (2021). Probabilistic deep learning to quantify uncertainty in air quality forecasting. Sensors, 21(23), 8009. https://doi.org/10.3390/s21238009*

*[7] Hameed, S., Islam, A., Ahmad, K., et al. (2023). Deep learning based multimodal urban air quality prediction and traffic analytics. Scientific Reports, 13, 22181. https://doi.org/10.1038/s41598-023-49296-7*

*[8] Chen, S., Zhao, Q., Chen, Z., Jin, Y., & Zhang, C. (2025). A Hybrid Air Quality Prediction Model Integrating KL-PV-CBGRU: Case Studies of Shijiazhuang and Beijing. Atmosphere, 16(8), 965. https://doi.org/10.3390/atmos16080965*

*[9] Iskandaryan, D., Ramos, F., & Trilles, S. (2023). Graph neural network for air quality prediction: A case study in Madrid. IEEE Access, 11, 2729–2742. https://doi.org/10.1109/ACCESS.2023.3234214*

*[10] Kavitha, B. (2024). Air quality monitoring system using machine learning and IoT. International Journal of Innovative Research in Information Security, 10(3), 369–379. https://doi.org/10.26562/ijiris.2024.v1003.40*

*[11] Li, J., Izakian, H., Pedrycz, W., & Jamal, I. (2025). Clustering-based anomaly detection in multivariate time series data. arXiv. https://doi.org/10.48550/arXiv.2511.08072*

*[12] Chauhan, R., Kaur, H., & Alankar, B. (2021). Air quality forecast using convolutional neural network for sustainable development in urban environments. Sustainable Cities and Society, 75, 103239. https://doi.org/10.1016/j.scs.2021.103239*

*[13] Du, Shengdong, et al. (2021). Deep Air Quality Forecasting Using Hybrid Deep Learning Framework. IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, Jun. 2021, pp. 2412-2424. https://doi.org/10.1109/tkde.2019.2954510.*

*[14] Mojtahedi, F. F., Yousefpour, N., Chow, S. H., & Cassidy, M. (2025). Deep learning for time series forecasting: Review and applications in geotechnics and geosciences. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-025-10244-5*

*[15] Mao, Y., Fang, A., Chen, Z., & Zheng, Y. (2023). Deep learning for air quality prediction: A comparative study. In Proceedings of the 2023 5th International Academic Exchange Conference on Science and Technology Innovation (IAECST). IEEE. https://doi.org/10.1109/IAECST60924.2023.10503030*

*[16] Parhusipa, H. A., Trihandaru, S., Susanto, B., Kurniawan, J. D., Heriadi, A. H., Santosa, P. P., & Sardjono, Y. (2024). Density-based spatial clustering of applications with noise (DBSCAN) and principal component analysis (PCA) for anomaly detection. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 15(2), 75. https://doi.org/10.24843/LKJITI.2024.v15.i02.p01*