INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 11, November 2026 | ISSN: 2456-4184 | IJNRD.ORG

Deep Learning-Based Air Quality Assessment and Climate Impact Analysis Using Multimodal Environmental Images

DR. S. K. SINGH[1], AMIT KUMAR PANDEY[2], AKSHAT SHARMA [3], YUSRA KAZI[4]

[1]HOD, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India, sksingh@tcsc.edu.in

[2]Assistant Professor, Department of AIML, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India, amitpandey8089@gmail.com

[3, 4] PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India, akshatsharmaa178@gmail.com, yusra.kazi03@gmail.com

Abstract **- Air pollution is still considered as one of the major environmental problems that has a great impact on the public health, climatic change, and ecology. Among pollutants, fine particulate matter (PM2.5) is a unique and particularly harmful pollutant because of its ability to penetrate the deepest parts of the respiratory system and stay airborne longer. Ground-based monitoring stations are the main method used for traditional air quality surveillance and provide accurate air quality measurements, but they are often expensive to install and are limited to a particular geographic area. In response to such limitations, this study proposes a multimodal deep learning framework combining the environmental image and the atmospheric parameters to support the assessment of air quality through PM2.5 concentration.**

The proposed system uses PM25Vision dataset containing the environmental imagery according to the concentration of PM2.5 and a selected subset of the atmospheric parameters from NASA's Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2). As a baseline, this initial evaluation involved three image-based deep learning architectures: the EfficientNet-B0 model, ResNet-50, and Vision Transformer (ViT-B/16). Then, a multimodal architecture was designed by combining visual representations (from EfficientNet-B0) and curated atmospheric parameters (at the feature level). The integrated representation was used to classify air quality in 6 different categories according to the concentration of PM2.5.

The experimental evaluation was performed by using a detailed analysis of the confusion matrix, accuracy, precision, recall, Macro F1 and Weighted F1. Among the image-only architectures, the Vision Transformer achieved the best test accuracy of 50.06%, while the proposed multimodal

EfficientNet-B0 model outperformed the others on the overall test accuracy of 56.22% with a Weighted F1-score of 55.78%. Results show that the inclusion of the atmospheric data information results in a significant increase in the accuracy of classification compared to the methods that rely only on the information of the environment. However, the study highlights the recurring issues of class imbalance, compounding the higher pollution levels, and the issue of model overfitting, which identify suitable paths for further research in the field of multimodal environmental intelligence systems.

Keywords **-- Air Quality Assessment, PM2.5, Deep Learning, Multimodal Learning, Environmental Images, EfficientNet-B0, Vision Transformer, ResNet-50, MERRA-2 and Climate Impact Analysis.**

I. INTRODUCTION

Air pollution has become one of the most serious environmental issues globally as it has a significant impact on human health, climate and the health and balance of nature. Air pollutants have steadily increased with the rapid expansion of cities, increased industries, increased car pollution and fossil fuel burning. Fine particulate matter, or PM2.5, is one of the most harmful of these pollutants. These extremely small particles, measuring 2.5 micrometers or less, are able to penetrate deeply into the lungs and into the bloodstream. This can increase the risk for breathing difficulties, heart disease and premature death. Because of this, it's very important

to closely watch and check the levels of PM2.5 to help manage the environment and keep people's health safe.

Traditional air quality monitoring is primarily based on the fixed stations equipped with special sensors for pollutants to measure pollutant levels accurately. Although these monitoring systems provide reliable data, installation and maintenance requires significant capital and resources and therefore the cost prevents many areas from being monitored. Furthermore, the data collected at each monitoring site reflects only air quality at that site, making it difficult to obtain a comprehensive overview of the air quality in locations with limited monitoring stations. As such, there has been an interest in exploring alternative approaches that can complement existing monitoring strategies.

In recent years, the success of Artificial Intelligence, Deep Learning, and Computer Vision with complicated image data has been demonstrated in obtaining useful information. Environmental pictures display fog, air clarity, sky, clouds, and purity of the atmosphere. These features typically are related to air pollution. Because the deep learning models learn visual patterns automatically rather than manual feature extraction by humans, they are suitable for environmental images. Recently, several popular models have demonstrated impressive performance on various image classification tasks such as EfficientNet-B0, ResNet-50 and Vision Transformer (ViT-B/16), which are good choices to begin exploring how to use environmental photographs to estimate air quality.

Despite these advances, the information provided by the environmental images is not sufficient to convey all the information regarding air quality, as many factors that influence the levels of PM2.5 are not easily visible in an image, such as the type of particles, black carbon, sulfates, dust and weather conditions. This may cause a wrong classification of similar forms of pollution or severe pollution scenarios in image-only models. Additional weather data, in conjunction with observations, can be used to improve the reliability and accuracy of models and their predictions.

In order to overcome this, this research proposes a deep learning system using images of the environment from the PM25Vision dataset and some weather data from NASA's MERRA-2. The detailed information of the environmental image is extracted using EfficientNet-B0, while the specific factors of the atmosphere are processed in a separate numerical feature section. The features extracted from both these data sets are merged to provide a unified set of features for predicting air quality at 6 different levels, using PM2.5. In addition, three deep learning models are adopted for a point of comparison: EfficientNet-B0, ResNet-50, and Vision Transformer (ViT-B/16) models are only image-based models, to serve as a baseline and serve as a comparison for the multimodal learning models.

The method emphasizes the careful data setup, and fairness of model testing. The PM25Vision data was cleaned to remove inconsistencies and duplicate data and 8,190 unique environmental images were collected from 2,528 monitoring stations. Data was divided according to the stations and a strategy was implemented to prevent the leakage of information. So that images from the same station are not included in multiple groups such as training, validation, and testing. To get a complete picture of the model's performance, Accuracy, Precision, Recall, Macro F1 score, Weighted F1 score, and Confusion Matrix were used, particularly in cases where there was a class imbalance. Experimental results demonstrate that the incorporation of atmospheric information, together with other data types, via a technique that integrates features from multiple data sets yields improved classification than non-atmospheric-only models. It also emphasises the continuing problems of class imbalance, learning too much from the training data, having poor generalisation, and problems in predicting accurately when pollution is high.

II. LITERATURE REVIEW

The current studies focused on machine learning and deep learning methods for air quality prediction and classification. Madokoro and Nix [1] explored the use of mobile devices and environment data in the prediction of particulate matter in a multimodal approach. The air quality pattern and relationship were further investigated by several researchers under different perspectives. For example, Wu et al. [2] proposed a hybrid DVMD Informer-CNN-LSTM model for capturing the complex air quality patterns, and Wang et al. [3] introduced a VMD-GAT-BiLSTM model for analysing the spatial and temporal relationship in air quality forecasting.

There has also been an interest in using a multi-modal method integrating multiple environmental data sources. Xia et al. [4] fused the multi-station time-series data and remote sensing images for air quality prediction. Hameed et al., [5] investigated the multimodal prediction of urban air quality information with the help of traffic and environment. Likewise, Kushal and Al Mamun [6] have suggested an Imagery-Sensor Fusion based model for prediction of Air Quality Index. To classify air pollution in smart cities, Mehmood et al. [7] proposed an explainable multimodal classification method named Vision-AQ.

The remote-sensing and spatiotemporal approaches have been explored in other studies. Kalaiselvi et al. [8] proposed a hybrid deep learning model with multiple RS data sources. Bui et al. [9] suggested a new method, called STAR, for the multimodal spatiotemporal air quality prediction, and likewise, the use of satellite imagery and artificial intelligence (AI) in multimodal predictions to forecast air quality was investigated by the authors of the paper [10].

While these studies provide proof of the potential for deep learning and multimodal learning, many of them rely on measurements from sensors, time-series data, traffic data, or satellite data. The potential to explore the use of a mixture of environmental images at the ground with atmospheric reanalysis data for PM2.5 classification still exists.

Thus this study is based on PM25Vision dataset and NASA's MERRA-2 atmospheric data. EfficientNet-B0, ResNet-50 and Vision Transformer (ViT-B/16) are tested on image-only basis. Features of the visual and selected atmospheric features are then fused together at the feature level to create a multimodal EfficientNet-B0 model. Performance is assessed with a station based data contingency strategy for unseen monitoring stations.

III. METHODOLOGY

The proposed steps of this approach are: preparing the data; preparing the data for the model; preparing the model; synthesizing features across various kinds of data; testing the model. The entire working process of the system proposed is shown in figure 1.

A. Dataset

For this study, the PM25Vision dataset is used, consisting of environmental pictures associated with PM2.5 data collected from different monitoring stations. Removing duplications and inconsistent records, the final data set contained 8,190 unique environmental pictures collected from 2,528 monitoring stations. Each picture was grouped into one of six different groups based on the PM2.5 levels: 0 to 50, 51 to 100, 101 to 150, 151 to 200, 201 to 300, and 301 to 600 $\mu g/m^3$.

B. Data Preprocessing

To make sure the model can be trained properly, any repeated records and samples that didn't match were taken out of the dataset. Images within the same monitoring station were clustered together by splitting the data based on the station, in order to minimise the risk of information leaking between the training, validation and testing sets. The final dataset was divided into 5,921 images for training, 1,248 images for validation and 1,021 for testing. The statistics used to make the atmospheric variables in NASA MERRA-2 standard were derived from the training data set only and applied to the validation and testing data sets.

C. Image-based Deep Learning Models

The three advanced deep learning models were evaluated for air quality image classification. The three advanced deep learning models were evaluated for image classification for the purpose of air quality classification.

- EfficientNet-B0
- ResNet-50
- Vision Transformer (ViT-B/16)

These models were trained with characteristics of the atmosphere such as haze, visibility, cloud cover and others directly from images of the environment. The performances were utilized as a basis to make comparisons with the new multimodal framework.

D. Proposed Multimodal Framework

The model integrates visual and atmospheric data; features are combined at the feature level. EfficientNet-B0 creates a 1280-dimensional vector that represents visual features. At the same time, another neural network takes eight chosen MERRA-2 atmospheric variables and turns them into a 128-dimensional numerical feature representation. These representations are fused to form a 1408 dimensional multimodal feature vector that goes through fully connected layers to classify the air quality based on the 6 PM2.5 levels.

E. Performance Evaluation

The models developed were validated with a number of metrics such as Accuracy, Precision, Recall, Macro F1-score, Weighted F1-score, and Confusion Matrix analysis. As the class sizes are not equal, it was important to examine Macro F1-score, Weighted F1 score and overall classification accuracy in addition to the overall classification accuracy.

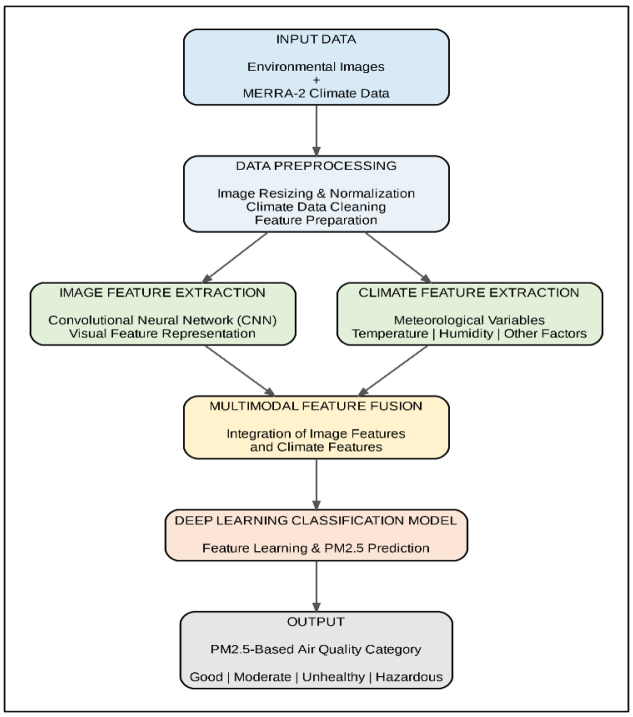


Fig. 1 Architecture

IV. RESULTS AND DISCUSSION

That was compared with the performance of three image-based deep learning models in the experimental evaluation. The best image-only models were ViT-B/16 with test accuracy 50.06%, ResNet-50 (49.62%) and EfficientNet-B0 (48.28%). The findings show that the environmental images can be used to estimate the PM2.5 categories with useful visual information for classification, but with moderate performance due to similarity of PM2.5 levels.

The inclusion of the atmospheric information from NASA MERRA-2 has provided an improvement in air quality classification, with the proposed multimodal EfficientNet-B0 model showing the best performance with a test accuracy of 56.22% and a Weighted F1-score of 55.78%. The results showed that, as compared with the best image-only model, the accuracy of the multimodal framework was enhanced by around 6.16 percentage points, which demonstrated the effectiveness of multimodal feature fusion.

The proposed multimodal model achieved the weighted F1-score of 55.78% and macro F1-score of 45.47% apart from accuracy. The score difference between the two reflects the better performance on classes with higher numbers of samples as opposed to the less represented pollution classes. This illustrates the need for assessing air quality classification with various performance measures, especially in a dataset with an unequal distribution of air pollution classes.

Performance analysis also showed that pollution categories which were lower and moderate had better classification rate than severe pollution categories. The number of samples for the higher PM2.5 classes was small, which led to a lower accuracy of the classification and more mixed-up in between pollution levels. The confusion matrix revealed that a number of samples in the 301–600 µg/m³ category were misclassified as 201–300 µg/m³, due to the difficulty in visually distinguishing between the similar conditions of high pollution.

Results of the training also showed signs of overfitting, though the multimodal model was performing better. Validation performance was relatively low, but accuracy was significantly higher in later epochs, which indicates that the model does not appear to have generalized well to unseen monitoring stations. Additionally, the experiments with class weighted training did not show improvement in Macro F1-score, which means that class weighting was not alone enough to overcome the dataset imbalance.

Table I. Performance Comparison of Deep Learning Models

Model	Test Accuracy	Remarks
EfficientNet-B0	48.28%	Baseline CNN model
ResNet-50	49.62%	Improved feature extraction
Vision Transformer (ViT-B/16)	50.06%	Best image-only model
Proposed Multimodal EfficientNet-B0	56.22%	Best overall performance

V. CONCLUSION AND FUTURE WORK

In this study, a multimodal deep learning framework was developed to assess air quality through PM2.5, combining images from the environment with selected atmospheric variables from NASA MERRA-2. Experimental results showed that using the atmospheric information significantly enhanced the classification accuracy compared to the deep learning models only using image data. The multimodal EfficientNet-B0 model proposed here obtained the best overall classification accuracy and showed the potential of fusing heterogeneous environmental information for air quality assessment.

However, the study revealed that there were some challenges: Class imbalance, lower classification accuracy for cases involving severe pollution levels, and model overfitting. The results suggest that multimodal

learning can be a viable approach for environmental monitoring and emphasize the necessity of acquiring more balanced datasets and enhancing the generalization methods.

Further studies could include collecting more samples for each class of pollution, exploring more sophisticated multimodal fusion methods, using stronger regularization and augmentation methods, adding more meteorological parameters, and expanding the framework to continuous prediction of PM2.5 and multi pollutant air quality assessment. This enhancement may further improve the reliability and applicability of deep learning-based environmental monitoring systems.

VI. REFERENCES

1. H. Madokoro and S. Nix, "Multimodal Particulate Matter Prediction: Enabling Scalable and High-Precision Air Quality Monitoring Using Mobile Devices and Deep Learning Models," *Sensors*, vol. 25, no. 13, Art. no. 4053, 2025, doi: 10.3390/s25134053.

2. Y. Wu, C. Qian, and H. Huang, "Enhanced Air Quality Prediction Using a Coupled DVMD Informer-CNN-LSTM Model Optimized with Dung Beetle Algorithm," *Entropy*, vol. 26, no. 7, Art. no. 534, 2024, doi: 10.3390/e26070534.

3. X. Wang, S. Zhang, Y. Chen, L. He, Y. Ren, Z. Zhang, J. Li, and S. Zhang, "Air Quality Forecasting Using a Spatiotemporal Hybrid Deep Learning Model Based on VMD-GAT-BiLSTM," *Scientific Reports*, vol. 14, Art. no. 17841, 2024, doi: 10.1038/s41598-024-68874-x.

4. H. Xia, X. Chen, Z. Wang, X. Chen, and F. Dong, "A Multi-Modal Deep-Learning Air Quality Prediction Method Based on Multi-Station Time-Series Data and Remote-Sensing Images: Case Study of Beijing and Tianjin," *Entropy*, vol. 26, no. 1, Art. no. 91, 2024, doi: 10.3390/e26010091.

5. S. Hameed, A. Islam, K. Ahmad, S. B. Belhaouari, J. Qadir, and A. Al-Fuqaha, "Deep Learning Based Multimodal Urban Air Quality Prediction and Traffic Analytics," *Scientific Reports*, vol. 13, no. 1, Art. no. 22181, 2023, doi: 10.1038/s41598-023-49296-7.

6. K. A. Kushal and A. Al Mamun, "AQFusionNet: Multimodal Deep Learning for Air Quality Index Prediction with Imagery and Sensor Data," *arXiv preprint arXiv:2509.00353*, 2025.

7. F. Mehmood, S. Ur Rehman, and A. Choi, "Vision-AQ: Explainable Multi-Modal Deep Learning for Air Pollution Classification in Smart Cities," *Mathematics*, vol. 13, no. 18, Art. no. 3017, 2025, doi: 10.3390/math13183017.

8. S. Kalaiselvi, V. Anitha, V. Manimaran, and T. S. Lawrence, "Air Quality Prediction Using Multi-Source Remote Sensing Data Integration with Hybrid Deep Learning Framework," *Scientific Reports*, vol. 16, Art. no. 2688, 2026, doi: 10.1038/s41598-025-32466-0.

9. T.-C. Bui, J. Kim, T. Kang, D. Lee, J. Choi, I. Yang, K. Jung, and S. K. Cha, "STAR: Spatio-Temporal Prediction of Air Quality Using a Multimodal Approach," in *Intelligent Systems and Applications*, Springer, 2021, pp. 389–406, doi: 10.1007/978-3-030-55187-2_31.

10. A. Rowley and O. Karakuş, "Predicting Air Quality via Multimodal AI and Satellite Imagery," *Remote Sensing of Environment*, vol. 293, Art. no. 113609, 2023, doi: 10.1016/j.rse.2023.113609.