PM 2.5 Estimation using Deep Learning methods

Angad Singh
Computer Science & Engineering
Thapar Institute of Engineering & Technology
Patiala, India
asingh36_be22@thapar.edu

Tarun Kumar
Computer Science & Engineering
CDAC, Mohali
Mohali, India
tarunkumar2992@gmail.comw

Gurmohan Singh
Computer Science & Engineering
CDAC, Mohali
Mohali, India
gurmohan@cdac.in

***Abstract*— Air pollution is a major health and environmental concern, especially in urban areas. Among various pollutants, PM2.5 is known to have the most harmful effects on human health. In this study, deep learning has been used to predict PM2.5 levels from sky images. The model is based on the VGG19 architecture with added residual connections to improve learning stability and performance. The network takes a 224×224 pixel sky image as input and predicts the PM2.5 concentration as a continuous value. The model was trained on a custom image dataset collected from regions in India and Nepal available on Kaggle. During evaluation, the R^2 score was measured to check how well the predicted values matched the actual PM2.5 levels. Results show that the model successfully learns visual patterns such as haze, cloud density, and lighting — all of correlate with pollution levels. Since expensive air quality sensors are not available in many rural areas, this model can serve as a low-cost alternative. With just a mobile phone camera, users can capture a photo of the sky and receive an estimate of air pollution in real time.**

Keywords*— *Air Pollution, PM 2.5 Prediction, Deep Learning, Sky Image Analysis, VGG19, CNN, Image-Based Regression

I. INTRODUCTION

Air pollution is a growing environmental and health challenge across the globe, especially in developing and densely populated countries. It refers to the presence of harmful substances in the atmosphere, including chemicals, toxic gases, and particulate matter. These pollutants originate from various anthropogenic sources such as vehicle emissions, industrial activities, open burning, and construction dust. Their continued accumulation in the air degrades air quality, reduces visibility, harms ecosystems, and poses serious risks to human health.

One of the most hazardous pollutants is **particulate matter (PM)**, particularly **PM2.5**, which refers to particles with a diameter of 2.5 microns or less. Due to their small size, PM2.5 particles can easily bypass the body's natural defenses and penetrate deep into the lungs and even enter the bloodstream. Long-term exposure to elevated PM2.5 levels has been linked to various respiratory and cardiovascular diseases, including asthma, bronchitis, heart attacks, and even premature death. According to the World Health Organization (WHO), air pollution contributes to millions of deaths every year and has become a leading cause of global health issues.

To control and mitigate the effects of air pollution, it is essential to monitor air quality levels accurately and in real time. Traditionally, this is done using **ground-based air quality monitoring stations** equipped with sophisticated sensors that measure various pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3. In India, such systems are deployed by agencies like the Central Pollution Control Board (CPCB). These stations provide reliable and high-precision data, but they come with significant limitations. Their installation, operation, and maintenance are expensive and complex, limiting their deployment to major urban centers. As a result, vast regions, particularly rural and remote areas, remain **unmonitored** or **under-monitored**, leading to significant gaps in pollution data.

To address these challenges, alternative methods have been explored. One approach involves the use of **low-cost sensors (LCS)** like Airveda or PurpleAir, which are portable and more affordable than government-grade equipment. However, these devices suffer from reduced accuracy, are sensitive to environmental conditions like humidity and temperature, and often require regular calibration to maintain reliability. Another emerging solution is the use of **artificial intelligence (AI)** and **computer vision techniques** to estimate air quality based on visual information, such as images of the sky.

In this work, a **deep learning-based method** to estimate **PM2.5 concentration** using only **images of the sky** has been proposed, which can be easily captured using a smartphone or webcam. This approach aims to provide a **low-cost**, **scalable**, and **accessible** alternative to traditional air quality monitoring systems, especially in regions where sensor-based infrastructure is unavailable or unreliable.

A custom convolutional neural network (CNN) model based on the VGG19 architecture, which is widely used in image classification tasks, was developed. To improve its learning capability and training stability, the model includes residual connections. The input to the model is a sky image resized to **224×224 pixels**, and the output is a **continuous PM2.5 value**. The model was trained on a publicly available dataset of sky images collected from various cities and towns in **India and Nepal**. Each image is associated with a corresponding PM2.5 label.

This model learns to identify visual patterns that correlate with air pollution, such as **haze**, **cloud density**, **light scattering**, and **overall sky color**. These features help the model make informed predictions about the level of particulate matter in the atmosphere.

By doing so, the system enables users to estimate air quality without needing any specialized hardware — only a camera and an internet connection.

II. LITERATURE REVIEW

A. Traditional Air Quality Monitoring Systems

Conventional systems rely on ground-based stations operated by the Central Pollution Control Board (CPCB) to measure pollutants such as PM2.5, PM10, NO_2, SO_2, and O_3. These systems provide highly reliable data; however, the cost of installation and maintenance limits their deployment mostly to urban areas, leaving rural and remote regions largely unmonitored [1]

B. Low-Cost Sensor Networks

To address these coverage gaps, low-cost sensors (LCS) such as Airveda and PurpleAir have been explored as portable and affordable alternatives. While they improve scalability, their readings are often affected by environmental factors like temperature and humidity. Hasenfratz et al. [2] and Morawska et al. [3] emphasize the trade-off between accuracy and accessibility, recommending periodic calibration and algorithmic correction for improved reliability.

C. Early Deep Learning and Non-Image Approaches

Initial studies used traditional machine learning models and basic convolutional neural networks (CNNs) for PM2.5 estimation from structured environmental data. However, these methods lacked robustness in real-time settings and struggled to interpret visual features effectively. Rouniyar et al. [4] introduced image-based estimation using sky photos, laying the groundwork for more advanced deep learning approaches.

III. METHODOLOGY

This section shows how the deep learning model was trained to predict PM2.5 levels using sky images. Sky images were taken in different regions – Bengaluru, Delhi, Dimapur, Faridabad, Greater Noida, Mumbai, Tamil Nadu, and Biratnagar. The images were then resized to 224x224 pixels and normalized. Then, the model was designed using a neural network based on VGG19 and added some improvements throughout to make it more stable. Finally, the model was trained and tested how well it could predict pollution levels based on new and testing images.

A. Dataset Collection

The data used was taken from a publicly available dataset from Kaggle that contains sky images labelled with corresponding PM2.5 values. The images were taken from different cities in India and Nepal, covering a wide range of pollution levels and weather conditions. Each image is paired with a PM2.5, PM10, O3 and AQI reading, which was measured by nearby air quality monitoring stations at the time the photo was captured. In total, the dataset includes 12,240 images, yielding enough data to train and test a deep learning model effectively.

B. Data Preprocessing

Before training the model, several preprocessing steps were applied to the image dataset. All images were resized to 224×224 pixels to match the input dimensions required by the VGG19 architecture.

Next, pixel values were normalized to a range of 0 to 1 by dividing each value by 255, which helps improve training stability and model convergence.

The dataset was then split into three parts: 7,834 images for training, 1,960 for validation, and 2,449 for testing.

The training set was used to fit the model, the validation set helped in tuning hyperparameters and preventing overfitting, and the test set was used to evaluate the model's final performance on unseen data.

C. Model Architecture

The deep learning model is based on the popular VGG19 architecture, which is good at analyzing images. Improvements were introduced by adding residual connections, which help the model learn faster and more accurately, especially as it gets deeper.

The model has five main blocks of layers:

- The first two blocks each have two convolution layers that detect patterns like edges and textures.
- The next three blocks each have four convolution layers, which learn more complex features.
- After each block, a residual shortcut is used that adds the input back to the output, helping the model avoid common training problems.
- A max pooling layer was added after every block to reduce the size of the image gradually.

Once all the blocks are done, the layers are flattened and the output is passed through two dense (fully connected) layers with 1024 units each. These layers help the model combine everything it has learned so far.

Finally, the model gives three outputs:

- One predicts the PM2.5 level.
- One predicts the PM10 level.
- One predicts the overall AQI.

Each output uses a single neuron and linear activation, since real numbers are being predicted (not categories). The model is trained using the Adam optimizer (a popular training method) with a small learning rate (0.0001), and Mean Absolute Error (MAE) to measure how far off our predictions are.

The full model design is shown in **Fig. 1** below:

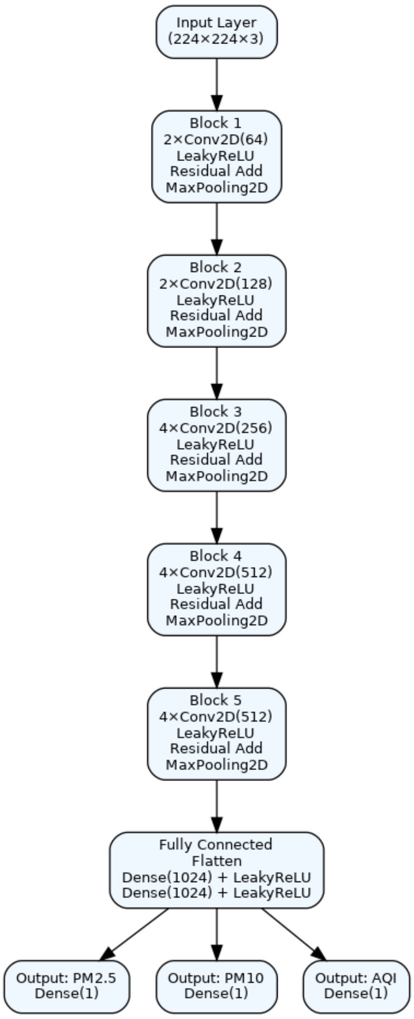

Fig 1. Custom Model Architecture of VGG19

D. *Model Training and Evaluation*

The model was trained using 150 epochs and a batch size of 16. To prevent overfitting, two callbacks were used:

- **EarlyStopping:** Stopped the training if the model did not improve for 10 validation epochs.
- **ModelCheckpoint:** Saved the best model weights during validation loss.

The model was evaluated on a separate test set during Mean Absolute Error (MAE) and R^2 score as a performance matrix:

- The model achieved an MAE of 12.54, meaning the average prediction error was around 12.54 µg/m³.
- The R^2 score was 0.9001, which means the model could explain over 90.01% of the variation in actual PM2.5 levels.

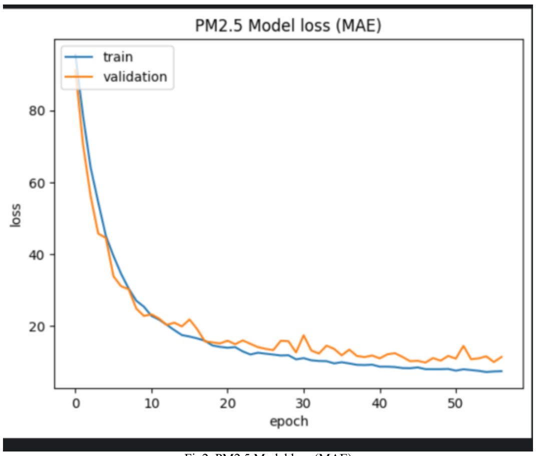

Fig2. PM2.5 Model loss (MAE)

IV. Results And Discussion

To make the system more practical and accessible, the trained model was deployed using **Streamlit**, an open-source Python library for building interactive web applications. The deployed interface allows users to either upload an image or capture one using a webcam, and immediately obtain an estimate of the PM2.5 concentration. The source code is available for review and adaptation. This deployment demonstrates the model's real-world usability and makes it possible to perform air quality estimation in real time using just a mobile camera – no sensors required. Such a solution can benefit rural communities, travellers, or anyone lacking access to traditional monitoring tools.

V. Conclusion

A deep learning-based approach was proposed to estimate PM2.5 air pollution levels using only sky images. The model was built on a modified VGG19 architecture incorporating residual connections and was trained on a publicly available dataset of sky images collected from various regions in India and Nepal. It demonstrated strong performance, achieving an R^2 score of **0.9001**, indicating high accuracy in predicting PM2.5 concentrations.

To enhance usability and accessibility, the model was deployed using **Streamlit**, enabling real-time predictions from uploaded or captured sky images. This solution offers a low-cost, user-friendly alternative to traditional air quality monitoring systems, particularly in areas lacking sensor-based infrastructure.

Future extensions may include integrating additional environmental factors such as temperature, humidity, and time of day, or combining image data with satellite observations to further improve prediction accuracy. The model thus presents strong potential as a scalable and portable tool for air quality estimation.

References

[1] Central Pollution Control Board (CPCB), "National Air Quality Monitoring Programme," [Online]. Available: https://cpcb.nic.in/

[2] A. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory Air Pollution Monitoring Using Smartphones," *Proc. Mobile Sensing*, Toronto, Canada, 2012.

[3] L. Morawska et al., "Applications of Low-Cost Sensing Technologies for Air Quality Monitoring and Exposure Assessment: How Far Have They Gone?" *Environment International*, vol. 116, pp. 286–299, 2018.

[4] L. Zhang, X. Chen, and L. Zhao, "Estimating PM2.5 concentrations using deep CNNs on satellite data," *Atmospheric Environment*, vol. 190, pp. 35–45, 2018.

[5] A. Rouniyar, S. Utomo, J. A., and P.-A. Hsiung, "Air Pollution Image Dataset from India and Nepal," *Kaggle*, 2023. [Online]. Available: https://doi.org/10.34740/KAGGLE/DS/3152196