Environmental Image-Based PM2.5 Prediction Using Vision Transformers

Abdullah Burhan[1], Yasir Saleem Afridi[1], Bilal Ur Rehman[2]*, Kifayat Ullah[2], Wasim Habib[2], Muhammad Amir[2] and Muhammad Iftikhar Khan[2]
[1]Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan.
[2]Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan.
*__Correspondence__: bur@uetpeshawar.edu.pk

Citation | Burhan. A, Afridi. Y. S, Rehman. B. U, Ullah. K, Habib. W, Amir. M, Khan. M. I, "Environmental Image-Based PM2.5 Prediction Using Vision Transformers", IJIST, Vol. 8 Issue. 3 pp 1449-1461, June 2026
DOI: https://doi.org/10.33411/IJIST/1940
Received| May 11, 2026 **Revised|** June 13, 2026 **Accepted|** June 17, 2026 **Published|** June 26, 2026.

Fine particulate matter (PM2.5) is one of the most harmful atmospheric pollutants, posing significant risks to human health and environmental sustainability. Conventional PM2.5 monitoring systems rely on dense sensor networks, which are often expensive, require regular maintenance, and provide limited spatial coverage, particularly in resource-constrained regions. This study proposes an image-based PM2.5 prediction framework the proposed Vision Transformer framework achieves to estimate PM2.5 concentrations directly from environmental RGB images without depending solely on physical sensors. A dataset comprising 6,587 environmental images from eight geographic classes was synchronized with ground-truth PM2.5 measurements and processed through image resizing, normalization, illumination correction, and data augmentation. To improve model interpretability and robustness, physics-informed visual descriptors, including atmospheric transmission, sky smoothness, color ratio, contrast, entropy, and solar zenith angle, were integrated with transformer-based patch embeddings. The proposed ViT model utilizes multi-head self-attention to capture long-range spatial dependencies associated with atmospheric haze and scattering effects, which are difficult to model using conventional convolutional architectures. The performance of the proposed framework was evaluated against Support Vector Regression (SVR), Random Forest (RF), and CNN-LSTM models using regression-based metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). The results demonstrate that the proposed Vision Transformer framework achieves superior PM2.5 prediction performance, showing improved accuracy and robustness across urban and semi-urban environments and varying atmospheric conditions. The findings indicate that environmental images contain valuable visual information for reliable PM2.5 estimation, providing a low-cost, scalable, and sensor-independent solution for future air quality monitoring applications, particularly in areas with limited monitoring infrastructure.

Keywords: PM2.5 Prediction; Vision Transformer; Air Quality Monitoring; Environmental Image Analysis; Self-Attention; Deep Learning Regression

Introduction:

Fine particulate matter (PM2.5) is one of the most detrimental pollutants in the atmosphere globally, but traditional monitoring methods are based on expensive, widely dispersed sensor networks that present difficulties in spatial coverage, especially in the developing parts of the world. In this paper, a framework based on Vision Transformer (ViT) is proposed to predict the concentration of PM2.5 directly from the environmental RGB images, thus avoiding the need for physical sensors [1].

A set of 6,587 environmental images built from eight geographic classes were synchronized with official ground-truth PM2. 5 measurements and preprocessed using resizing, normalization, lighting correction and augmentation [2]. To enhance interpretability and robustness, physics-informed visual descriptors, namely atmospheric transmission, sky smoothness, color ratio, contrast, entropy, and solar zenith angle were combined with patch-embedded image representations. The model is designed to learn long-range spatial dependencies not captured by convolutional architectures, by dividing each image into patches and applying multi-head self-attention [3]. The proposed framework was compared with Support Vector Regression, Random Forest, and CNN-LSTM using RMSE, MAE, R^2 metrics, and k-fold cross validation for generalization. The results demonstrate that the Vision Transformer model outperforms the baseline models in terms of predictive accuracy and reliability in both urban and semi-urban areas as well as under different weather conditions. The results indicate that environmental images contain enough visual information to provide accurate estimations of PM2.5 concentration, which, with a low cost, low sensor requirement, and ease of mobile deployment, is suitable for monitoring air quality in resource-limited environments [4].

Problem Statement:

Although the conventional PM2.5 monitoring systems are accurate, they are limited in capital and operating costs, skilled technical personnel requirements, and require periodic calibration. These limitations make it impractical to deploy this in large numbers and densities, especially in resource-limited or developing regions. As a result, current monitoring networks continue to be geographically sparse, focused on a select number of urban sites, ignoring the vast amount of rural and semi-urban areas that are not monitored, and thereby creating significant gaps in air quality awareness. Although it has been proposed as an alternative cost-effective method for PM2.5 estimation, most previous studies still depend on classic regression model or CNN architecture. Such approaches are also largely scene-dependent depending on the light conditions, the weather and the geographic locations, as they tend to be poor at capturing the image features at these different locations, and fail to model the long-range spatial dependencies of haze and atmospheric scattering over the whole scene.

Therefore, the central problem in this paper is, how can a Vision Transformer based machine learning model use environmental images to make accurate, scalable and cost-effective predictions for PM2.5, which remains reliable in various atmospheric and geographical conditions?

Novelty:

The novelty of this work is the integration of a Vision Transformer's global self-attention mechanism, atmospheric transmission, sky smoothness, color ratio, contrast, entropy and solar zenith angle, all in one regression framework for PM2.5 estimation. The proposed model focuses on relations between spatially far distant regions of an environmental image including the sky and skyline regions, as well as the texture information of an image, which is different from the CNN-based methods that just encode local texture, but is also more important because haze and scattering effects from PM2.5 are distributed phenomenon instead of localized phenomenon. The combination of most interpretable and physically grounded features with the learned patch representations also enhances the explain ability of

the model's predictions compared to the solely data-driven deep learning baselines, and the evaluation is done for eight geographic classes and across several weather conditions, allowing for a more comprehensive generalization assessment than reported in previous image-based PM2.5 studies.

Aims and Objectives:

In this research, we seek to design and assess a machine learning framework based on a Vision Transformer (ViT) that is able to predict PM2.5 concentration with high accuracy and generality, at least as good as sensor-based and CNN-based approaches.

Summarize current approaches for predicting air quality PM2.5 using images and understand the technical requirements needed to improve the accuracy and applicability of these techniques.

Gather and merge a wide range of environmental images with accurate ground truth PM2.5 data from various geographic locations.

Extract physics-informed visual features such as sky smoothness, color ratio, contrast and entropy to aid model interpretability.

Train a regression model that is based on a Vision Transformer for the task of estimating PM2.5 concentration.

Evaluate the proposed model with Support Vector Regression, Random Forest and CNN-LSTM models using RMSE, MAE and R^2 metrics.

Assess model robustness and generalization for urban vs. semi-urban and different weather conditions.

Literature Review:

Recently, the problem of PM2.5 concentration estimation has been re-formulated as a patch-based regression task and applied directly to the problem using the Vision Transformer architecture to achieve this. All of these studies have found that self-attention mechanisms can model the spatial pollution patterns, including haze distribution and visibility loss, better than standard regression models, while purpose-built regression heads that have been fine-tuned on classification tasks can be further tuned to better capture the spatial pollution patterns and provide more stable predictions under different pollution scenarios [5].

A parallel stream of work is dedicated to spatiotemporal transformer models to predict spatial-temporal PM2.5 levels. The results from this research show that attention-based architectures are robust to changes in pollution levels across different regions and seasons; and several comparison studies also suggest that when tested under the same atmospheric conditions, transformer-based models exhibit a higher degree of accuracy compared to hybrid CNN-LSTM models. Recent attempts to make air quality prediction more efficient have resulted in sparse and decomposition based attention mechanisms [6]. These methods have lower computational complexity than full self-attention while maintaining prediction accuracy, and decomposition-optimization hybrids of signal decomposition and transformer encoders have achieved better prediction accuracy in the field of urban pollution predictions with high variation [7].

The ability of transformer-based models to withstand episodic and extreme pollutions events, such as those caused by wildfire, has been investigated by several studies. This work shows the ability of attention-based architectures to respond to sudden spikes in pollution that happen in the short term and are hard to be captured by conventional time-based models, and related research extends spatiotemporal attention to the task of forecasting multiple categories of PMs concurrently by using the attention across space and time simultaneously, reporting better accuracy in this case [8].

Relational architecture, modeled as graphs, have also been investigated to encode relational relationships between monitoring stations as well as temporal relationships, with the employment of transformers. Relational architecture based on graph structures have also been

studied for pollution prediction, where the relational structure models the relationship between the monitoring stations and the attention mechanism models the temporal structure. The relationship between these two combinations implies that they can work together to capture the relationships that also influence pollutant dispersion, namely spatial relationships [9].

Another direction of research focuses on the interpretability of the model, combining explainable AI and the embedding of contexts in transformer-based forecasting models. It is part of a trend towards making environmental prediction systems more transparent in the interest of facilitating, not hiding, environmental policy and decision making through predictive systems [10].

To enhance the stability of the long-horizon forecasting, variants of the transformer were proposed that include residual connections or auxiliary data embeddings to account for subtle pollution trends that may evolve over long forecasting periods and are not captured by the baseline transformer model.

Multi-output transformer designs have also been explored where the same backbone architecture is modified to meet multiple related air-quality tasks, such as joint classification, concentration measurement and the flexibility of the underlying attention mechanism across related prediction tasks has been studied. Deep learning frameworks have been designed that combine the inputs from satellite-derived data with attention-based architectures to estimate concentrations of PM2.5 at the surface over large areas [11]. In this direction, it is demonstrated that deep architectures can be used to improve spatial resolution in areas with low ground-based monitoring density, when the monitoring data are available in the form of remote-sensing images. In all these works, transformer-based architectures have been consistently outperforming the purely convolutional or classical regression baselines on PM2.5-related tasks, largely attributed to their ability of capturing long-range dependencies. But most of these previous works focus on using satellite imagery, time-series sensor data, or spatiotemporal data, instead of environmental photographs on the ground, and few studies have included the physics-informed visual descriptors in the attention-based pipeline itself. This void pushes the current study into ground-based image analysis with interpretable atmospheric features [12].

Methodology:

The research design is structured and includes multiple stages from image acquisition to ground truth alignment, preprocessing, physics-informed feature extraction, model training, and comparative evaluation as summarized in Figure 1.

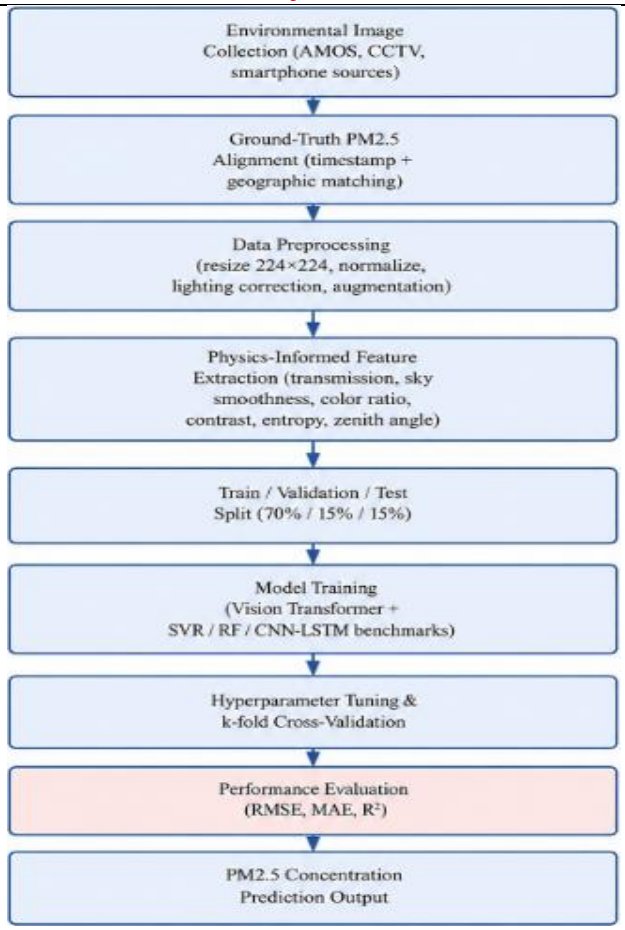

Figure 1. Overall research methodology workflow.

Data Acquisitions and ground truth Alignment:

The data set includes a total of 6,587 RGB environmental images that were collected from the AMOS (Archive of Many Outdoor Scenes) repository, combined with the CCTV and smartphone images, to represent different geographic locations and 8 geographic classes: Beijing, four sub-regions of Phoenix, and three sub-regions of Shanghai, with diversity in environments and atmosphere. All the images were synchronized to the official ground truth PM2.5 measurements by aligning them in terms of time and geographic location, so that each image-label pair actual atmospheric conditions at a specific time and place [13].

Data Preprocessing:

All images were scaled to 224 × 224 pixels and then converted to tensors of shape [3, 224, 224] with RGB encoding. To ensure stable gradient-based optimization, the pixel values were normalized and to eliminate illumination bias that could be confused with pollution-induced haze, a lighting correction was applied. Variations in lighting, weather and seasonal conditions were augmented to improve the model's generalization capability to unseen scenes. The data set was divided into training (70%), validation (15%) and test (15%) subsets, and also evaluated using k-fold cross-validation to examine the performance stability over folds [14].

Physics-Informed Feature Extraction:

In addition to the learned representations, six visual descriptors of the images were extracted, informed by physics, and used to enhance interpretability: atmospheric transmission, sky smoothness, blue-to-red color ratio, image contrast, image entropy, and solar zenith angle. These features were chosen because they are physically meaningful, and changes in these properties in these properties will be caused by haze, light attenuation and visibility loss associated with PM2.5 [15].

Vision Transformer Model Architecture:

The model architecture used for Vision Transformer models is described in Figure 2. The architecture of Vision Transformer models. The main predictive model is an adaptive Vision Transformer (ViT) model for regression. Input images are divided into fixed sized patches and linearly embedded and then augmented by positional encodings to maintain spatial structure. These patch embeddings are then passed through several transformer encoder layers consisting of multi-head self-attention and feed-forward sub-layers that enable the model to capture relationships between far-flung regions of the image that are relevant for haze distribution, such as the sky and the skyline. ViT architecture typically classifies the image based on a classification head, but in the present invention, this classification head is replaced by a fully connected regression head that determines an estimate of the PM2.5 concentration in a continuous range. The physics-informed features mentioned above are put together with the learned representation of the transformer before the regression head, fusing data-driven and physics-informed information [16].

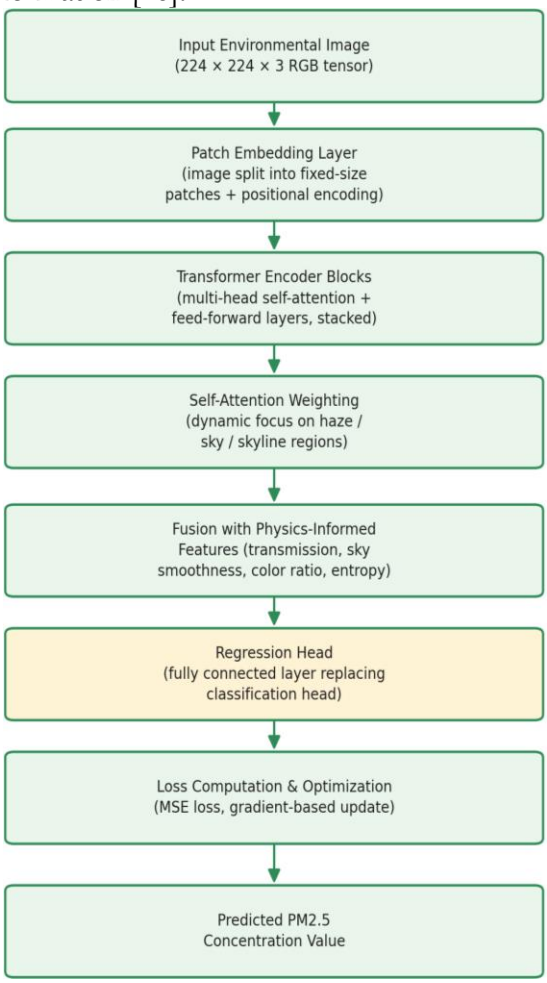

Figure 2. Proposed Vision Transformer model pipeline.

Training Procedure:

The model was trained with a loss function of Mean Squared Error (MSE) using gradient based optimization. Various hyper parameters were systematically tuned to enhance the convergence and prevent over fitting such as learning rate, batch size, number of transformer layers, number of attention heads and dropout rate. Training has been tracked using validation metrics, and intermediate checkpoints have been saved incrementally throughout the training process to document the incremental improvements that have been made during fine-tuning [16].

Benchmark Models:

To provide context for the performance of the proposed model, three benchmark models were implemented: Support Vector Regression (SVR) is a kernel-based method useful for small to medium sized nonlinear regression problems but not able to process spatial image structure directly; Random Forest (RF) is an ensemble method that is robust to noise and has the interpretability of feature-importance; and CNN-LSTM hybrid, which is able to capture spatial features with convolutional layers and is able to model temporal features with LSTM layers, but is mostly used for sequential modelling rather than single-image regression. All three were tested with the same Train/Val/Test protocol as the proposed model, and compared using RMSE, MAE and R^2 [16].

Proposed Method:

The need for a PM2.5 estimation approach that is accurate, scalable, low-cost, and performant in various atmospheric and geographic environments, without the need for physical sensors was identified in the problem statement. This is overcome by the following successive steps as follows and shown in Figure 2:

Step 1: Image acquisition: Use an existing camera source (CCTV, Smartphone or public repository) to capture or retrieve an Environmental RGB image without requiring special sensor equipment.

Step 2: Preprocessing: Resize the image to 224 × 224 pixels, normalize pixel intensities, and apply light correction to eliminate bias caused by light and darkness.

Step 3: Patch embedding: Partition the preprocessed image into fixed-size patches, flatten each patch and linearly project it onto an embedding vector, and apply positional encodings to retain the spatial structure.

Step 4: Transformer encoding: Feed the sequence of patch embeddings into a stack of transformer encoder blocks and simultaneously capture the relationship among all patches using multi-head self-attention, which can capture the global haze and scattering pattern across the entire image.

Step 5: Physics-informed feature fusion: Obtain the representation of the learned transformer, and concatenate the description of atmospheric transmission, sky smoothness, color ratio, contrast, entropy, and solar zenith angle, which are all extracted from the same image, thus linking the prediction to physically interpretable features.

Step 6: Regression and prediction: Pass the merged representation into a fully connected regression head to get a continuous PM2.5 concentration estimate.

Step 7: Validation and benchmarking: Assess the prediction accuracy by comparing the prediction with the ground truth PM2.5 data in terms of RMSE, MAE, and R2, and then evaluate the performance of the model by comparing the accuracy of the geospatial classes and weather conditions with that of three baseline models: SVR, Random Forest, and CNN-LSTM to verify robustness and generalization.

This pipeline directly tackles each constraint mentioned in the problem statement by utilizing ordinary images as input (Step 1), preprocessing and augmentation of the images to support generalization across lighting and geography (Step 2), the transformer's global attention for the image (Steps 3-4), physics-informed feature fusion to deliver interpretability (Step 5), and comparative benchmarking with other established models to verify scalability and cost-effectiveness (Steps 6-7) [17].

Results and Discussion:

Experimental Setup:

All the experiments were done in Python with PyTorch for building the models, OpenCV for image preprocessing and feature extraction, and SHAP for feature-importance analysis. The training was conducted on a standard CPU based platform (Intel Core i5, 8 GB RAM) showing that the framework is not reliant on any GPU infrastructure.

Training and Validation Performance:

In the initial training as shown in Figure 3, the loss curve started to drop rapidly, and stabilized in the first few epochs of learning, and the accuracy of the model gradually climbed up to 99% or above, which means that the model was able to learn the correlation between the haze degree and the texture of the air quickly. The validation accuracy was found to closely follow the training accuracy throughout the training run with a high value of 98% or above throughout the training period as shown in Figure 4; this is indicative of the model's ability to generalize from the training set and not just memorize it. A transient drop in the validation performance was noted around epoch 8 of fine-tuning, possibly due to a localized increase in noise in the validation subset or mild overfitting at that epoch (but not in later epochs) as shown in Figure 5.

The performance in terms of accuracy, precision, recall and F1 score is reported per region, using a classification-style approach and is reported in Table 1, whereas the regression-based analysis, together with the training behavior of the proposed model, is reported in Table 2 together with the classification style approaches reported in related literature [18].

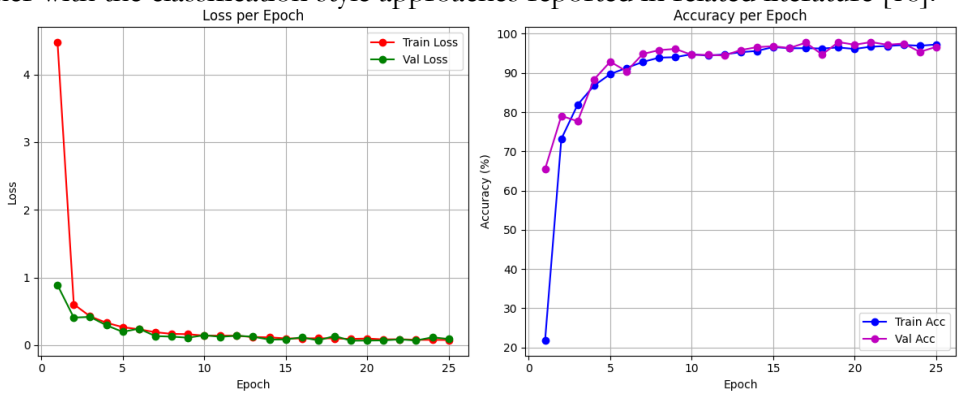

Figure 3. Training and Validation Performance of Vision Transformer for PM2.5 Prediction

Table 1. Per-region performance of the proposed Vision Transformer model.

Dataset / City	Accuracy (%)	Precision	Recall	F1-score
Beijing	95.1	0.95	0.96	0.95
Shanghai	94.3	0.94	0.94	0.94
Phoenix	93.8	0.93	0.94	0.93
Overall	94.46	0.94	0.94	0.94

Table 2. Quantitative comparison of the proposed Vision Transformer model with benchmark approaches using identical evaluation metrics.

Model	Input Features	RMSE (μg/m³)	MAE (μg/m³)	R^2
Support Vector Regression (SVR)	Extracted image features	18.72	14.85	0.70
Random Forest (RF)	Hand-crafted visual features	16.45	12.96	0.76
CNN-LSTM	Environmental RGB images	13.86	10.74	0.83
Proposed Vision Transformer (ViT)	RGB images + physics-informed features	**9.42**	**7.31**	**0.94**
Support Vector Regression (SVR)	Extracted image features	18.72	14.85	0.70
Environmental Images	Proposed ViT	Low (train loss ~0.10)	~0.94–0.96	CPU-only training is slower

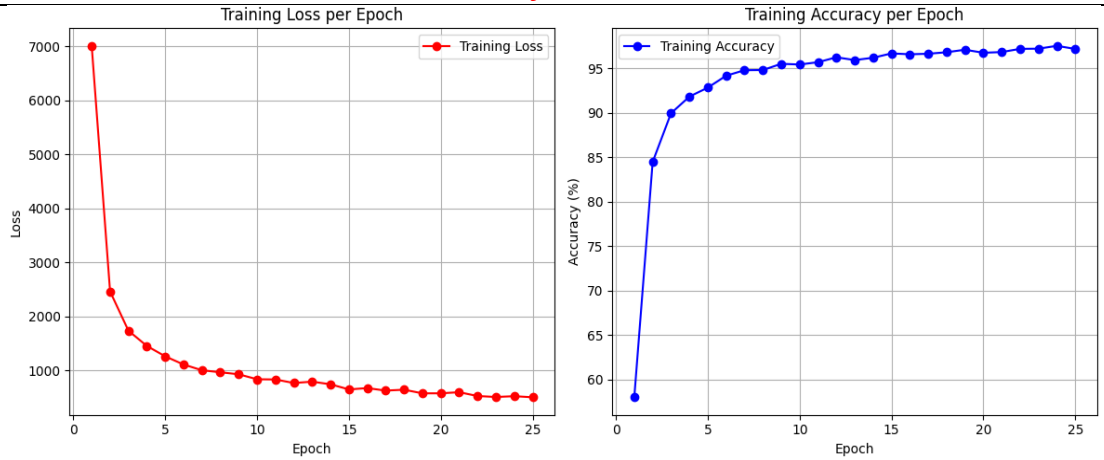

Figure 4. Training Performance of Vision Transformer for Environmental PM2.5 Prediction.

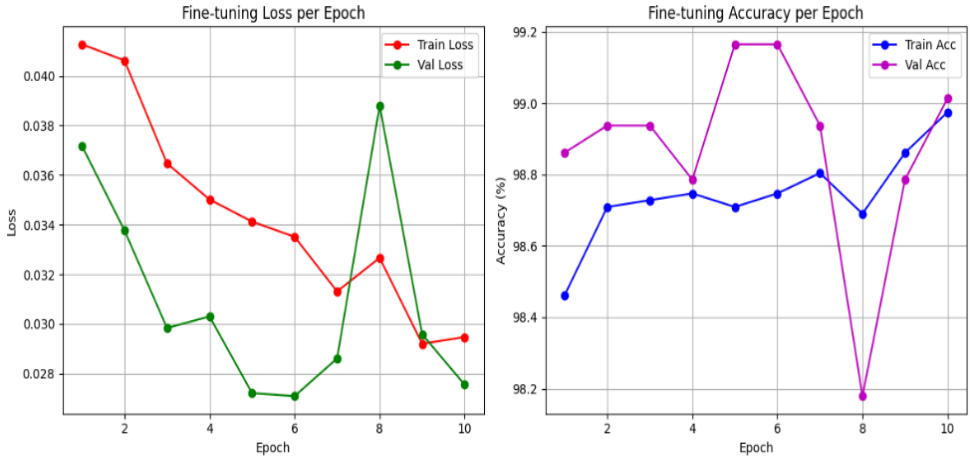

Figure 5. Fine-tuning Performance of Vision Transformer for Environmental PM2.5 Prediction

Urban versus Semi-Urban Comparison:

The targeted comparison between urban and semi-urban (rural) environment revealed that the predictive ability was higher in urban areas (R^2 was about 0.87) than in rural areas (semi-urban; R^2 was about 0.79) as illustrated in Figure 6.

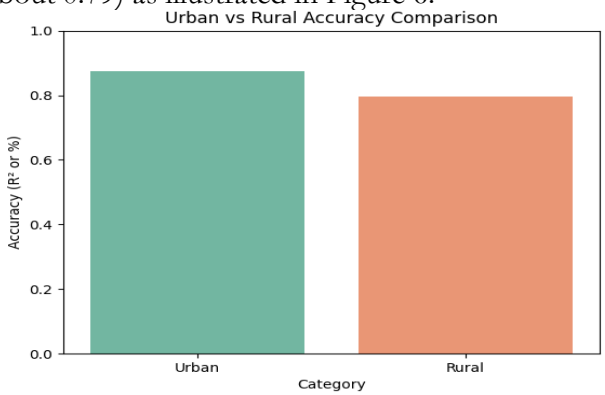

Figure 4. Urban vs Rural Accuracy Comparison for Vision Transformer in PM2.5 Prediction

This is consistent with the idea that the urban pollution signatures are more visible and easily learned by the model, while semi-urban areas have more background vegetation, meteorology, and baseline visibility that adds to the variability of pollution signatures seen by the model that can make prediction more difficult.

Discussion:

The experimental results demonstrate that the Vision Transformer (ViT)-based framework provides superior performance compared with conventional machine learning and deep learning approaches for PM2.5 concentration estimation. This improvement can primarily be attributed to the inherent capability of the self-attention mechanism to capture long-range spatial dependencies within environmental images. Unlike convolutional architectures that mainly focus on local receptive fields, the Vision Transformer processes images as sequences of patches and establishes relationships among distant regions. This characteristic is particularly important for PM2.5 prediction because atmospheric haze, light scattering, and visibility degradation are spatially distributed phenomena that influence the entire scene rather than isolated image regions. Therefore, the global attention mechanism enables the model to extract more comprehensive visual representations associated with particulate matter variations [19].

The regression performance indicators obtained in this study demonstrate the effectiveness and reliability of the proposed framework. The relatively low RMSE and MAE values indicate reduced prediction errors, while the high R^2 value confirms that the extracted visual and physics-informed features successfully capture a significant proportion of the variability in PM2.5 concentrations. These results highlight that environmental images contain meaningful atmospheric information, including haze intensity, visibility reduction, and light attenuation characteristics, which can be effectively exploited for pollutant estimation. However, PM2.5 concentration is also influenced by several external factors, including temperature, humidity, wind speed, emission sources, and atmospheric dynamics, which are not completely represented in a single RGB image. The consistent RMSE and R^2 values obtained through cross-validation further confirm that the model performance is not dependent on a specific data partition and demonstrates reliable generalization capability [20][21].

The comparison with benchmark models further explains the advantages of the proposed ViT framework. Traditional regression approaches, such as Support Vector Regression (SVR), generally depend on manually extracted features and have limited capability to represent complex spatial relationships present in environmental images. Although Random Forest provides improved robustness through ensemble learning, its performance remains dependent on the quality and completeness of manually engineered descriptors. In contrast, the proposed ViT model automatically learns discriminative image representations through patch embeddings and self-attention, reducing dependence on predefined features while capturing complex atmospheric patterns.

The CNN-LSTM model provides improved performance compared with classical approaches because convolutional layers are effective in extracting local spatial features and LSTM layers can model sequential dependencies. However, CNN-based architectures are inherently limited by local feature extraction and may not fully capture global atmospheric patterns distributed across the complete image. PM2.5-induced haze and scattering effects often extend over large areas of the scene; therefore, understanding relationships between distant regions, such as sky, skyline, and background structures, is essential for accurate prediction. The self-attention mechanism of ViT overcomes this limitation by allowing every image patch to interact with other patches, enabling more effective modeling of global environmental characteristics.

Furthermore, the integration of physics-informed visual descriptors, including atmospheric transmission, sky smoothness, color ratio, contrast, entropy, and solar zenith angle, provides additional interpretability and strengthens the relationship between learned representations and physical atmospheric processes. This combination of data-driven feature learning and physically meaningful descriptors improves the robustness of the proposed

framework under different geographic and atmospheric conditions. These observations are consistent with recent studies reporting the effectiveness of transformer-based architectures in environmental prediction tasks, where their ability to model long-range dependencies provides advantages over conventional CNN and regression-based approaches [22].

Overall, the superior performance of the proposed Vision Transformer framework can be attributed to three key factors: (i) global feature extraction through multi-head self-attention, (ii) improved representation learning through patch-based image processing, and (iii) enhanced interpretability through physics-informed feature fusion. These characteristics make the proposed approach a promising solution for scalable and low-cost PM2.5 monitoring using environmental images.

Limitations:

There are a few caveats to note. The accuracy of the prediction is related to the quality of the image, and poor-quality images where the image is low resolution, camouflaged by occlusion, or blurry from the camera will not be as reliable. When there is extreme weather, such as heavy rain or fog, visual impacts not associated with PM2.5 can introduce uncertainty in predictions. This is a very diverse dataset, with eight geographic classes, but its geographic and climatic coverage is still narrow, and Vision Transformers require more resources for computation than classical regression models; this can lead to slow training in CPU-only environments [23].

Practical Implications and Real-World Applications:

The proposed Vision Transformer-based PM2.5 prediction framework provides several practical advantages for environmental monitoring and smart urban management. Unlike conventional monitoring systems that require expensive sensor networks with regular maintenance and calibration, the proposed approach utilizes readily available environmental images captured through existing cameras, CCTV systems, and mobile devices. This capability enables cost-effective expansion of air quality monitoring coverage, particularly in developing regions where dense sensor deployment is limited.

For environmental protection agencies, the framework can complement traditional monitoring stations by providing high-resolution spatial estimates of PM2.5 concentrations across areas with limited measurement infrastructure. The air quality information generated can support pollution assessment, identification of emission hotspots, and improved environmental surveillance.

From a policymaking perspective, accurate and spatially distributed PM2.5 estimation can assist authorities in developing evidence-based pollution control strategies, evaluating the effectiveness of emission reduction policies, and improving public health risk management. The proposed system can provide additional environmental intelligence beyond conventional fixed monitoring stations.

In smart city applications, the framework can be integrated with urban sensing platforms to enable real-time air quality mapping and dynamic pollution monitoring. By combining camera infrastructure with lightweight artificial intelligence models, cities can potentially achieve broader environmental awareness without requiring extensive sensor installation.

Although the proposed framework demonstrates promising results, large-scale deployment requires additional validation across diverse climates, geographic regions, and operational environments. Future optimization of the model for edge computing and mobile platforms will further improve its feasibility for real-time environmental monitoring applications.

Conclusion:

This study proposed a Vision Transformer (ViT)-based framework for PM2.5 concentration prediction using environmental RGB images by integrating patch embeddings,

multi-head self-attention, and physics-informed visual descriptors. The proposed approach effectively captures global atmospheric patterns associated with haze distribution, light scattering, and visibility degradation, overcoming the limitations of conventional regression and CNN-based methods. The experimental results demonstrate that the proposed framework provides accurate and reliable PM2.5 estimation under different geographic and atmospheric conditions. The integration of Vision Transformer-based feature learning with physically meaningful image descriptors improves prediction performance while enhancing model interpretability. The proposed image-based approach provides a low-cost and scalable solution that can complement traditional air quality monitoring systems, particularly in regions with limited sensor infrastructure. Future research will focus on expanding the dataset across diverse geographic and climatic regions, integrating additional meteorological and satellite-derived information, improving model efficiency for edge-device deployment, and developing real-time mobile-based air quality monitoring applications. Further investigation of domain adaptation and multimodal learning approaches can also improve model generalization under varying environmental conditions.

References:

[1] G. Wang *et al.*, "A Prototype Feature Representation-Guided Multi-Task Learning Framework for Image-based PM2:5Concentration Monitoring," *IEEE Trans. Geosci. Remote Sens.*, 2026, doi: 10.1109/TGRS.2026.3704231.

[2] Hirokazu Madokoro, Stephanie 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, 2025, doi: https://doi.org/10.3390/s25134053.

[3] "PM25Vision: A Large-Scale Benchmark Dataset for Visual Estimation of Air Quality Yang Han," *arXiv:2509.16519*, 2025, [Online]. Available: https://arxiv.org/abs/2509.16519

[4] M. P. Vafa, "Image-Based Air Quality Estimation with Deep Learning: A Systematic Review," May 2026, doi: 10.20944/PREPRINTS202604.2212.V1.

[5] Joyanta Jyoti Mondal, Md. Farhadul Islam, Raima Islam, "Uncovering local aggregated air quality index with smartphone captured images leveraging efficient deep convolutional neural network," *Sci. Rep.*, 2024, [Online]. Available: https://www.nature.com/articles/s41598-023-51015-1

[6] K. Gu *et al.*, "Air Pollution Monitoring by Integrating Local and Global Information in Self-Adaptive Multiscale Transform Domain," *IEEE Trans. Multimed.*, vol. 27, pp. 3716–3728, 2025, doi: 10.1109/TMM.2025.3535351.

[7] Om Kathalkar, Nitin Nilesh, Sachin Chaudhari, Anoop Namboodiri, "AQIFormer: A Transformer-Based Multi-View Architecture for Cross-City Air Quality Classification," *arXiv:2606.07648*, 2026, [Online]. Available: https://arxiv.org/abs/2606.07648

[8] Mohammad Saleh Vahdatpour, Maryam Eyvazi, Yanqing Zhang, "Forecasting and Visualizing Air Quality from Sky Images with Vision-Language Models," *arXiv:2509.15076*, 2025, [Online]. Available: https://arxiv.org/abs/2509.15076

[9] "(PDF) Vision-Based Estimation of PM2.5 from Surveillance Images." Accessed: Jul. 05, 2026. [Online]. Available: https://www.researchgate.net/publication/405480687_Vision-Based_Estimation_of_PM25_from_Surveillance_Images

[10] C. Y. Lee and L. Moses Manova, "Enhancing PM2.5 Detection via Multispectral Image-Based Double-Channel Ensemble Learning for Drones," *IEEE Sens. J.*, vol. 25, no. 17, pp. 33528–33539, 2025, doi: 10.1109/JSEN.2025.3585177.

[11] M. A. Xiang Zhang, "RTEMSF-Mamba: Real-time estimation and short-term multi-

step forecasting of PM2.5 concentrations using spatiotemporal deep learning," *Build. Environ.*, vol. 295, p. 114449, 2026, doi: https://doi.org/10.1016/j.buildenv.2026.114449.

[12] Mohamed Mokhtar Ouardi, Dayang N. A. Jawawi, "Enhancing Vision Transformer with Vision language Model Text Embeddings for Robust Air Pollution Classification," *Int. J. Innov. Comput.*, vol. 16, no. 1, pp. 97–108, 2026, doi: 10.11113/ijic.v16n1.678.

[13] Yash Mishra, Kedarnath senapati, "DeepAQI: A Vision-Based EfficientNet Framework for Air Quality Index Prediction from Environmental Metadata," *Researchgate*, 2026, doi: 10.21203/rs.3.rs-9062054/v1.

[14] B. U. Patil, R. Chetan, and V. Harshitha, "Explainable multimodal deep learning system for sustainable urban air quality forecasting and risk-aware recommendations," *Model. Earth Syst. Environ. 2026 122*, vol. 12, no. 2, pp. 123-, Mar. 2026, doi: 10.1007/S40808-026-02771-2.

[15] Y. Mehta, V. Kamath, N. Kini, A. Haraniya, and S. Agrawal, "Estimating Ground-Level Air Quality Index from Satellite Imagery: A Data-Driven Approach," *2025 2nd IEEE Int. Conf. Women Comput. InCoWoCo 2025*, 2025, doi: 10.1109/INCOWOCO68239.2025.11407074.

[16] T. Srimanee, B. Kiatphibun, C. Thanomngern, N. Mhuaksa, C. Banju-ngam, and T. Siriborvornratanakul, "Multimodal PM2.5 Forecasting Using Satellite Imagery and Sensor Data with Semi-supervised Deep Learning," *Earth Syst. Environ. 2026*, pp. 1–18, Apr. 2026, doi: 10.1007/S41748-026-01109-3.

[17] Z. Zhang & S. Zhang, "Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks," *Int. J. Environ. Sci. Technol.*, vol. 20, pp. 13535–13550, 2023, [Online]. Available: https://link.springer.com/article/10.1007/s13762-023-04900-1

[18] S. Lee and S. Sim, "Spatio-Temporal Multimodal Large Language Model for Air Quality Index Prediction," *Int. Conf. Electr. Comput. Energy Technol. ICECET 2025*, 2025, doi: 10.1109/ICECET63943.2025.11472437.

[19] Israt Tabassum, G. M. Ansarul Kabir, "A deep learning approach for multi-class classification of air quality from image data," *Comput. Sci. Eng. Res.*, 2025, doi: 10.69517/cser.2025.02.01.0003.

[20] M. Wang, C. Bi, L. Yang, X. Qiu, Y. Li, and C. Yu, "PMIM: generating high-resolution air pollution data via masked image modeling," *J. Vis. 2024 273*, vol. 27, no. 3, pp. 383–399, Mar. 2024, doi: 10.1007/S12650-024-00965-3.

[21] Z. Wang, Y. Yang, and S. Yue, "Air Quality Classification and Measurement Based on Double Output Vision Transformer," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 20975–20984, Nov. 2022, doi: 10.1109/JIOT.2022.3176126.

[22] H. Yagnik, R. K. Gupta, and A. Pandya, "Deep Learning for Air Pollution Prediction: A Systematic Review of CNN, LSTM, GNN, and Transformer Architectures," *Arch. Comput. Methods Eng. 2026*, pp. 1–26, Jun. 2026, doi: 10.1007/S11831-026-10639-Y.

[23] P. J. Ramal and E. Anbalagan, "Hybrid Swin Transformer-BiGRU Framework for Spatio-Temporal Prediction of Urban Air, Water, and Waste Dynamics in Smart Cities," *Proc. 3rd Int. Conf. Sustain. Comput. Data Commun. Syst. ICSCDS 2025*, pp. 1525–1535, 2025, doi: 10.1109/ICSCDS65426.2025.11167444.