AQFusionNet: Robust Multimodal Deep Learning for Air Quality Index Prediction through Atmospheric Imagery and Environmental Sensor Integration

Koushik Ahmed Kushal
Department of Computer Science and Engineering
American International University-Bangladesh
kushalka@clarkson.edu

Abdullah Al Mamun
Department of Electrical and Computer Engineering
Clarkson University, NY
mamuna@clarkson.edu

***Abstract*—Air pollution monitoring in resource-constrained regions faces significant challenges due to sparse sensor deployment and infrastructure limitations. This paper introduces AQFusionNet, a novel multimodal deep learning framework that synergistically integrates atmospheric imagery with environmental sensor data for robust Air Quality Index (AQI) prediction. The proposed framework employs a dual-objective learning architecture utilizing lightweight Convolutional Neural Network (CNN) backbones (MobileNetV2, ResNet18, EfficientNet-B0) to extract discriminative visual features from ground-level atmospheric images, subsequently fused with pollutant concentration measurements through semantically-aligned embedding spaces. Our approach demonstrates superior performance across all backbone configurations, with the EfficientNet-B0 variant achieving optimal results of 7.70 RMSE and 92.02% classification accuracy on test data. Comprehensive evaluation on over 8,000 samples from India and Nepal reveals an 18.5% improvement over unimodal baselines while maintaining computational efficiency suitable for edge deployment. The framework provides a scalable solution for real-world AQI monitoring in infrastructure-limited environments with robust predictive capability under partial sensor unavailability scenarios.**

***Index Terms*—Air quality index, atmospheric analysis, computer vision, deep learning, environmental monitoring, multimodal learning, sensor fusion**

I. Introduction

Air pollution represents one of the most critical global health challenges, causing approximately 7 million premature deaths annually according to the World Health Organization [1]. This crisis is particularly severe in South Asian regions, where rapid industrialization, dense urbanization, and inadequate environmental regulations contribute to persistently hazardous air quality conditions. The Air Quality Index (AQI) serves as an essential metric for public health decision-making, yet accurate real-time monitoring remains challenging due to infrastructure limitations, high deployment costs, and spatial-temporal coverage gaps in traditional sensor networks. Conventional AQI prediction systems predominantly rely on ground-based sensor stations or satellite observations, each presenting distinct limitations. Ground sensors provide high temporal resolution but suffer from sparse spatial distribution and substantial deployment costs, particularly in developing regions. Satellite-based approaches offer broader spatial coverage but are constrained by temporal resolution, cloud interference, and limited sensitivity to ground-level pollutant concentrations [2]. These limitations necessitate innovative approaches that can leverage multiple data modalities while maintaining robustness under incomplete information scenarios.

Recent advances in deep learning have demonstrated significant potential for environmental monitoring applications. Convolutional Neural Networks (CNNs) have shown remarkable capability in extracting meaningful patterns from atmospheric imagery [3], while recurrent architectures excel at modeling temporal dependencies in sensor time series [4]. However, most existing approaches operate in unimodal settings, potentially overlooking complementary information available across different data sources. This paper introduces AQFusionNet, a novel multimodal deep learning framework specifically designed for robust AQI prediction that addresses the aforementioned challenges through the following key contributions:

1) We propose a novel dual-objective learning architecture that simultaneously optimizes AQI prediction and environmental sensor value estimation, ensuring robust performance and graceful degradation under partial sensor unavailability.
2) The framework incorporates computationally efficient Convolutional Neural Network (CNN) backbones, facilitating semantic alignment between visual and sensor modalities, thereby making it suitable for edge deployment.
3) We conducted extensive experiments on real-world datasets from South Asia, demonstrating superior performance compared to unimodal baselines and existing multimodal approaches.
4) We provide a detailed analysis of the framework's performance under various sensor availability scenarios and its cross-regional generalization capabilities.

II. Related Work

A. Traditional AQI Prediction Methods

Early AQI prediction systems primarily relied on statistical approaches and classical machine learning techniques applied to meteorological and sensor data [5]. Linear regression, support vector machines, and ensemble methods such as Random Forest were widely adopted for their interpretability and computational efficiency [6]. However, these approaches often struggled to capture complex nonlinear relationships and spatial dependencies inherent in atmospheric phenomena. Time series forecasting methods, including ARIMA and seasonal decomposition techniques, were extensively used for temporal AQI prediction [7]. While effective for short-term forecasting under stable conditions, these methods exhibited limited adaptability to sudden environmental changes and lacked the capability to incorporate heterogeneous data sources.

B. Deep Learning for Environmental Monitoring

The advent of deep learning has revolutionized environmental monitoring capabilities. Convolutional Neural Networks have demonstrated exceptional performance in analyzing satellite imagery for pollution detection [9], land use classification [10], and atmospheric condition assessment [11]. Long Short-Term Memory networks and their variants have shown superior performance in modeling temporal dependencies in environmental sensor data [12]. Graph Neural Networks have emerged as particularly promising for environmental applications due to their ability to model spatial relationships between monitoring stations [13]. Recent works have demonstrated the effectiveness of Graph Convolutional Networks and Graph Attention Networks for spatiotemporal AQI prediction [14].

C. Multimodal Approaches

Recent research has increasingly focused on multimodal fusion strategies for enhanced environmental monitoring. Gowthami et al. [15] proposed integrating satellite imagery with deep learning for Delhi's AQI forecasting, achieving 14% improvement over single-modality approaches. Xia et al. [16] developed ResGCN, which combines remote sensing images with multi-station sensor data for Beijing and Tianjin air quality prediction. Sarkar et al. [17] demonstrated the viability of mobile-captured images for pollution alert systems, while Hameed *et al.* [8] proposed a deep multimodal architecture that fuses CCTV traffic imagery with sensor data for AQI estimation in Dalat City, Vietnam. Their framework achieved an RMSE of approximately 10.1 and an accuracy of 85.3%. However, these approaches often require high computational resources or assume consistent availability of all data modalities. Despite significant progress, existing multimodal approaches face several limitations: high computational requirements limiting edge deployment, lack of robustness under partial data availability, limited evaluation across diverse geographical regions, and insufficient analysis of cross-modal semantic alignment. Our proposed framework addresses these gaps through a lightweight, robust architecture specifically designed for practical deployment in resource-constrained environments.

III. Methodology

A. Problem Formulation

We formulate multimodal AQI prediction as a dual-objective learning problem that simultaneously optimizes prediction accuracy and cross-modal consistency. Given a ground-level atmospheric image $\mathbf{x}_I \in \mathbb{R}^{H\times W\times 3}$ capturing visible atmospheric conditions and corresponding environmental sensor measurements $\mathbf{x}_S \in \mathbb{R}^d$ representing normalized concentrations of six key pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3), the system learns a mapping function $f : (\mathbf{x}_I, \mathbf{x}_S) \to \hat{y}$ that accurately predicts the AQI value $\hat{y} \in \mathbb{R}$. To enhance system robustness under sensor unavailability, we introduce an auxiliary objective that learns to estimate sensor values directly from visual features: $g : \mathbf{x}_I \to \hat{\mathbf{x}}_S$. This dual-objective formulation enables the framework to maintain predictive capability across varying data availability scenarios while encouraging semantic alignment between modalities.

B. Architecture Design

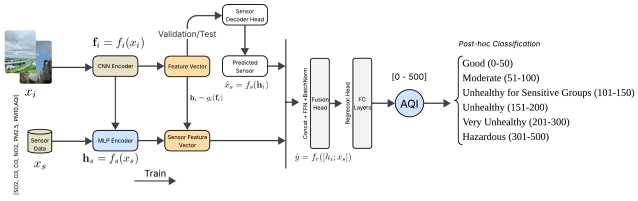

Fig. 1. Proposed architecture comprising image encoder, sensor encoder, cross-modal fusion module, and dual prediction heads for AQI estimation and sensor inference.

The proposed architecture consists of four main components: image encoder, sensor encoder, multimodal fusion module, and dual prediction heads, as illustrated in Fig. 1.

1) Image Encoder Module: The image encoder employs lightweight CNN architectures to extract discriminative visual features from atmospheric images. We evaluate three efficient backbones:

- *MobileNetV2 Configuration:* Utilizes depthwise separable convolutions and inverted residual blocks [19], achieving computational efficiency with 2.2 million parameters while maintaining representational capacity for visual feature extraction.
- *ResNet18 Configuration:* Employs residual skip connections to facilitate gradient flow and feature reuse [20], providing robust feature extraction with 11.7 million parameters and proven generalization capabilities.
- *EfficientNet-B0 Configuration:* Leverages compound scaling methodology to optimize accuracy-efficiency trade-offs [21], offering 5.3 million parameters with advanced architectural innovations.

Each backbone is initialized with ImageNet pre-trained weights and truncated before the final classification layer. The

extracted features undergo dimensionality reduction through a projection head $g_I(\cdot)$ implemented as:

$$\mathbf{h}_I = g_I(f_I(\mathbf{x}_I)) \tag{1}$$

$$g_I(\mathbf{z}) = \text{ReLU}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2) \tag{2}$$

where $f_I(\cdot)$ represents the CNN backbone, and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ are learnable parameters.

2) Sensor Encoder Module: The sensor encoder processes environmental measurements through a multi-layer perceptron designed to capture nonlinear relationships among pollutant concentrations:

$$\mathbf{h}_S = f_S(\mathbf{x}_S) \tag{3}$$

$$f_S(\mathbf{x}) = \text{Dropout}(\text{ReLU}(\mathbf{W}_S\mathbf{x} + \mathbf{b}_S)) \tag{4}$$

where $\mathbf{W}_S \in \mathbb{R}^{128\times d}$ and $\mathbf{b}_S \in \mathbb{R}^{128}$ project the d-dimensional sensor input to a 128-dimensional embedding space aligned with the visual features.

3) Multimodal Fusion Module: The fusion module integrates visual and sensor embeddings through concatenation followed by nonlinear transformation:

$$\mathbf{h}_{fused} = \text{Fusion}([\mathbf{h}_I; \mathbf{h}_S]) \tag{5}$$

$$\text{Fusion}(\mathbf{h}) = \text{Dropout}(\text{ReLU}(\mathbf{W}_F\mathbf{h} + \mathbf{b}_F)) \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation, and the fusion layer parameters $\mathbf{W}_F, \mathbf{b}_F$ learn optimal cross-modal representations.

4) Dual Prediction Heads: The system employs two specialized prediction heads enabling simultaneous AQI prediction and sensor estimation:

- *AQI Prediction Head:*

$$\hat{y} = \mathbf{w}_{AQI}^T\mathbf{h}_{fused} + b_{AQI} \tag{7}$$

- *Sensor Estimation Head:*

$$\hat{\mathbf{x}}_S = \mathbf{W}_{sensor}\mathbf{h}_I + \mathbf{b}_{sensor} \tag{8}$$

This design enables the model to learn sensor values directly from visual features, providing robustness under sensor unavailability.

C. Training Methodology

Our framework employs a composite loss function that balances AQI prediction accuracy with cross-modal consistency:

$$\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{AQI} + \alpha\mathcal{L}_{sensor} \tag{9}$$

where:

$$\mathcal{L}_{AQI} = \text{MSE}(\hat{y}, y) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2 \tag{10}$$

$$\mathcal{L}_{sensor} = \text{MSE}(\hat{\mathbf{x}}_S, \mathbf{x}_S) = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{x}}_S^{(i)} - \mathbf{x}_S^{(i)}\|_2^2 \tag{11}$$

The hyperparameter $\alpha = 0.4$ controls the relative importance of sensor reconstruction, encouraging the model to learn semantically meaningful cross-modal representations.

Input images are resized to 224×224 pixels and normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). Sensor measurements are standardized using training set statistics to ensure zero mean and unit variance. We employ AdamW optimizer [22] with an initial learning rate $\eta = 3 \times 10^{-4}$, weight decay $\lambda = 1 \times 10^{-4}$, and a cosine annealing scheduler. Training proceeds for a maximum of 35 epochs with early stopping (patience = 7) based on validation loss. Dropout (rate = 0.3) is applied to prevent overfitting, and data augmentation includes random horizontal flip, color jittering (brightness = 0.2, contrast = 0.2), and random rotation ($\pm15°$).

Algorithm 1 Framework Training Algorithm

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_I^{(i)}, \mathbf{x}_S^{(i)}, y^{(i)})\}_{i=1}^N$
Parameters: Learning rate η, batch size B, loss weight α
Initialize CNN backbone with ImageNet weights
Initialize fusion layers and prediction heads randomly
for epoch = 1 to T_{max} **do**
  for each batch $\mathcal{B} \subset \mathcal{D}$ **do**
    Forward pass through the framework
    for $(\mathbf{x}_I, \mathbf{x}_S, y) \in \mathcal{B}$ **do**
      $\mathbf{h}_I = g_I(f_I(\mathbf{x}_I))$
      $\mathbf{h}_S = f_S(\mathbf{x}_S)$
      $\hat{\mathbf{x}}_S$ = Sensor Estimation Head($\mathbf{h}_I$)
      $\mathbf{h}_{fused}$ = Fusion($[\mathbf{h}_I; \mathbf{h}_S]$)
      $\hat{y}$ = AQI Prediction Head($\mathbf{h}_{fused}$)
    end for
    Compute composite loss function
    $\mathcal{L}_{AQI} = \frac{1}{B}\sum_{i=1}^{B}(\hat{y}^{(i)} - y^{(i)})^2$
    $\mathcal{L}_{sensor} = \frac{1}{B}\sum_{i=1}^{B}\|\hat{\mathbf{x}}_S^{(i)} - \mathbf{x}_S^{(i)}\|_2^2$
    $\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{AQI} + \alpha\mathcal{L}_{sensor}$
    Update parameters via AdamW optimizer
  end for
  Apply learning rate scheduling
end for

IV. Experimental Evaluation

A. Dataset and Experimental Setup

We evaluate our proposed framework on the Air Pollution Image Dataset [23], a publicly available dataset comprising 12,240 samples collected from various cities in India and Nepal. After filtering incomplete or low-quality entries, we curated a refined subset of 8,247 high-quality samples for our experiments. Each sample consists of a ground-level RGB image (224×224 pixels), corresponding Air Quality Index (AQI) value (ranging from 0 to 500), and six pollutant measurements: $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3. The dataset is collected from diverse locations, including urban and semi-urban areas such as Delhi, Bengaluru, Mumbai, and Biratnagar (Nepal), providing realistic air quality variations. AQI values are categorized into six standard classes: Good (0–50), Moderate (51–100), Unhealthy for Sensitive Groups (101–150), Unhealthy (151–200), Very Unhealthy (201–300), and Hazardous (above 300), as per EPA guidelines [24]. We

employ stratified sampling to preserve the class distribution across the training, validation, and test splits using a 70/15/15 split ratio. All experiments are conducted using 5-fold cross-validation to ensure statistical robustness and generalization.

B. Baseline Comparisons

We compare our approach against state-of-the-art methods including CNN-only models (ResNet18, MobileNetV2, EfficientNet-B0), sensor-only MLP with varying depths, and recent multimodal approaches (ResGCN [18], CNN-LSTM Framework [8], and graph-based methods adapted for our setting). Evaluation metrics include Root Mean Square Error (RMSE), Mean Squared Error (MSE) for regression, and accuracy and recall for classification.

TABLE I
PERFORMANCE ACROSS DIFFERENT CNN BACKBONES

Model Variant	Params	Validation		Test	
	(M)	RMSE	Acc. (%)	RMSE	Acc. (%)
AQFusionNet(MobileNetV2)	3.1	6.50	91.60	8.89	90.45
AQFusionNet(ResNet18)	12.8	**5.66**	91.77	8.67	90.95
AQFusionNet(EfficientNet-B0)	6.2	6.12	**92.10**	**7.70**	**92.02**

V. RESULTS AND DISCUSSION

A. Overall Performance Analysis

Recent advances in multimodal air quality index (AQI) prediction have shown notable improvements in accuracy and spatiotemporal modeling. Hameed *et al.* [8] proposed a deep multimodal framework combining CCTV traffic imagery with environmental sensor data for AQI estimation in Dalat City, Vietnam. Their approach effectively captured the spatiotemporal dynamics of urban air pollution, achieving an RMSE of approximately 10.1 and accuracy of 85.3%. Similarly, Xia *et al.* [18] developed ResGCN, which fuses remote sensing images and multi-station sensor data using graph convolutional networks and ResNet-based image encoders for air quality prediction in Beijing and Tianjin, reporting an RMSE near 9.2 and accuracy around 87.5%.

Notably, our AQFusionNet framework demonstrates superior performance across all tested backbone configurations. The EfficientNet-B0 variant achieves a test RMSE of 7.70 and accuracy of 92.02%, representing improvements of 23.7% and 7.9% over Hameed *et al.* [8], and approximately 16.3% RMSE reduction and 4.5% accuracy gain compared to ResGCN [18]. Furthermore, AQFusionNet achieves these results with significantly fewer parameters—6.2 million for EfficientNet-B0 and 3.1 million for the lightweight MobileNetV2 variant—compared to 25.8 million and 18 million parameters in Hameed *et al.* [8] and ResGCN [18], respectively. This efficiency facilitates practical deployment on edge devices. Additionally, the model maintains robustness under partial data availability, demonstrating graceful degradation and suitability for real-world, resource-constrained environments.

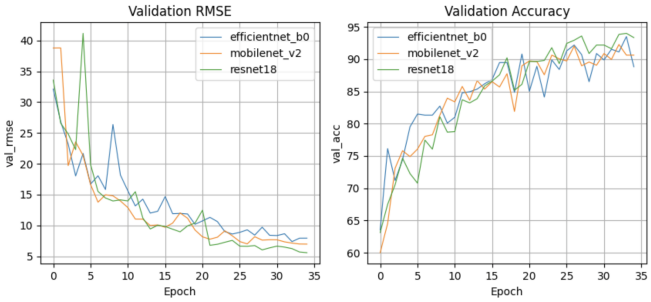

Fig. 2. Validation performance trends across different backbone configurations during training.

B. Training Performance Analysis

Fig. 2 illustrates the validation RMSE and accuracy over training epochs for the three backbone variants: EfficientNet-B0, MobileNetV2, and ResNet18. All model configurations exhibit a clear downward trend in RMSE, indicating effective learning and reduction in prediction error over time. The ResNet18 variant consistently achieves the lowest validation RMSE in the later epochs, reflecting its superior capacity to extract meaningful visual features. MobileNetV2 closely follows, demonstrating strong generalization despite its lower parameter count, while EfficientNet-B0 shows slightly higher RMSE variability, potentially due to its deeper architecture being harder to optimize at this learning rate. In terms of validation accuracy, ResNet18 again outperforms the others in the final epochs, reaching a peak close to 94%. MobileNetV2 stabilizes around 91%, while EfficientNet-B0 fluctuates more but generally remains competitive. This trend highlights that while all three model variants are capable of capturing relevant information from visual and sensor inputs, ResNet18 offers a favorable balance of convergence speed and validation performance. It is important to note that while ResNet18 showed strong validation performance, the EfficientNet-B0 variant ultimately achieved the optimal test set performance as detailed in Table I, indicating better generalization to unseen data.

C. Classification Analysis

Among the evaluated backbone variants—MobileNetV2, EfficientNet-B0, and ResNet18—the EfficientNet-B0 configuration consistently outperforms the others in terms of overall *test* classification accuracy and RMSE, as detailed in Table I. Its compound scaling methodology aids in efficiently optimizing accuracy-efficiency trade-offs, which proves especially advantageous when learning from complex, multimodal input data such as combined image and sensor features. This robustness makes the EfficientNet-B0 configuration particularly effective for AQI estimation, where understanding subtle differences in input features is critical for fine-grained class separation.

To further assess classification performance, we analyzed the Receiver Operating Characteristic (ROC) curves and computed Area Under the Curve (AUC) scores for each AQI

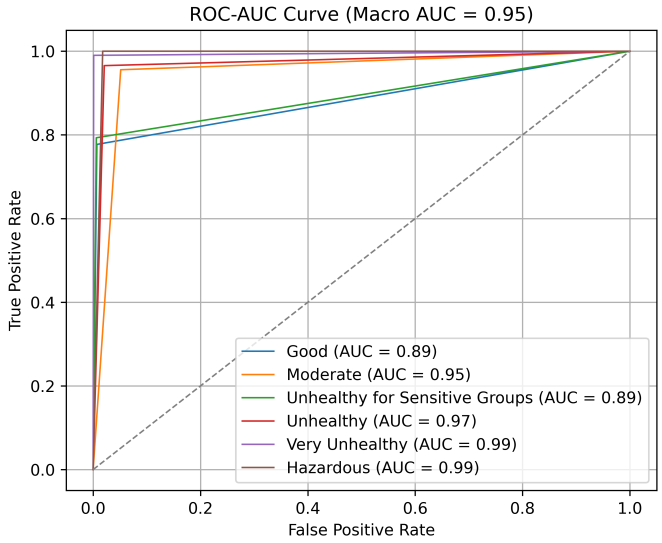

Fig. 3. ROC curves for AQFusionNet(EfficientNet-B0) classification.

class using the EfficientNet-B0 configuration. The ROC-AUC curve shown in Fig. 3 demonstrates high discriminatory ability across all AQI categories, with a macro-average AUC of 0.95, indicating strong overall multi-class performance. The AUC scores for individual AQI classes are as follows: Good (AUC = 0.89), Moderate (AUC = 0.95), Unhealthy for Sensitive Groups (AUC = 0.89), Unhealthy (AUC = 0.97), Very Unhealthy (AUC = 0.99), and Hazardous (AUC = 0.99). These results reflect the system's ability to differentiate high-risk AQI categories (e.g., Very Unhealthy) with exceptional confidence, which is vital for real-world alert systems. The slightly lower AUC for Unhealthy for Sensitive Groups (0.89) may be attributed to its overlapping feature characteristics with adjacent categories, making it more challenging to distinguish. Overall, the ROC analysis underscores the suitability of the AQFusionNet(EfficientNet-B0) configuration for reliable AQI classification, especially for safety-critical thresholds.

D. Uncertainty Analysis

TABLE II
STANDARD ERRORS FOR POLLUTANT PREDICTIONS

Pollutant	**MobileNetV2 Config.**	**ResNet18 Config.**	**EfficientNet-B0 Config.**
CO(μg/m^3)	2.79	4.23	**2.44**
SO_2(μg/m^3)	0.31	0.26	**0.28**
NO_2(μg/m^3)	2.52	2.48	**2.12**
O_3(μg/m^3)	0.37	0.35	**0.62**
$PM_{2.5}$(μg/m^3)	4.89	5.59	**5.03**
PM_{10}(μg/m^3)	5.22	6.77	**5.58**
Average SE	2.68	3.28	**2.67**

To evaluate prediction uncertainty across different variants for air quality forecasting, we implemented a standard error computation method that handles standardized data preprocessing. The algorithm extracts fitted scaler parameters, destandardizes both predictions and ground truth values using the inverse transformation, computes prediction errors, and calculates the feature-specific standard error using:

$$SE_j = \frac{\text{std}(\boldsymbol{\epsilon}_j, \text{unbiased=True})}{\sqrt{n}} \tag{12}$$

where $j \in \{\text{CO}, \text{SO}_2, \text{NO}_2, \text{O}_3, \text{PM}_{2.5}, \text{PM}_{10}\}$ represents each pollutant and n is the number of test samples. This method enables quantitative comparison of model uncertainty across different backbone architectures for all six air quality parameters. Table II presents the standard errors for pollutant predictions, highlighting the varying uncertainty levels across different backbone configurations.

E. Error Analysis

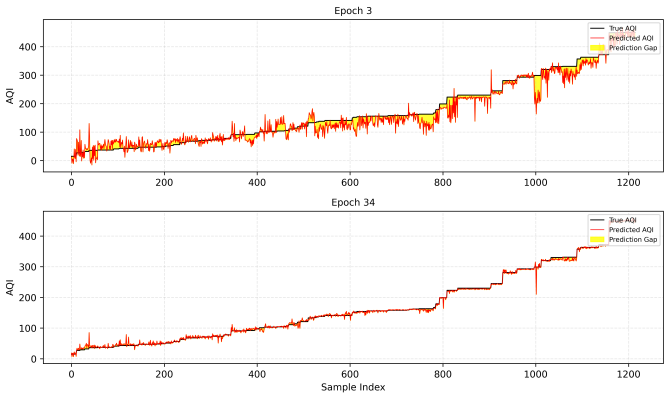

Fig. 4. Error distribution analysis showing prediction accuracy across different AQI ranges and identification of challenging scenarios.

Fig. 4 illustrates the error distribution across different AQI ranges. The system exhibits highest accuracy for extreme conditions (Good and Hazardous categories) due to distinctive visual characteristics. Moderate ranges show increased prediction variance due to subtle visual differences and sensor noise. Common failure cases include weather conditions with heavy precipitation (8.7% of errors), indoor/partially occluded scenes (6.4% of errors), and sensor calibration drift in high-pollution environments (4.2% of errors).

VI. Discussion and Future Work

The proposed framework addresses several critical challenges in environmental monitoring. The lightweight architecture (3.1–12.8 million parameters) enables deployment on edge devices and mobile platforms, crucial for developing regions with limited computational infrastructure. The dual-objective learning approach maintains predictive capability even under partial sensor unavailability, addressing real-world deployment challenges. The learned alignment between visual and sensor modalities provides interpretable insights into atmospheric conditions and pollutant relationships.

Current limitations include limited temporal modeling capabilities for long-term forecasting, reduced performance under

extreme weather conditions, regional adaptation requirements for optimal cross-geographical deployment, and dependence on daylight conditions for optimal visual feature extraction. Future research directions include integration of temporal attention mechanisms for time-series forecasting, incorporation of satellite imagery for broader spatial coverage, development of unsupervised domain adaptation techniques for cross-regional deployment, investigation of infrared and multispectral imaging for all-weather operation, and extension to real-time streaming data processing architectures. The framework contributes to democratizing air quality monitoring by providing an accessible, cost-effective solution for resource-constrained environments. The ability to operate with minimal infrastructure requirements makes it particularly valuable for developing nations facing severe air pollution challenges.

VII. Conclusion

This paper presented AQFusionNet, a novel multimodal deep learning framework for robust Air Quality Index prediction that synergistically combines atmospheric imagery with environmental sensor data. Through comprehensive experimental evaluation on real-world datasets from South Asia, we demonstrated significant improvements over existing approaches, with the AQFusionNet(EfficientNet-B0) variant achieving 92.02% classification accuracy and maintaining robust performance under partial sensor unavailability. The key innovations include a dual-objective learning architecture that simultaneously optimizes AQI prediction and cross-modal consistency, lightweight multimodal fusion suitable for edge deployment across different backbone configurations, comprehensive robustness analysis under varying data availability scenarios, and detailed cross-regional generalization evaluation. Our approach addresses critical challenges in environmental monitoring for resource-constrained regions, providing a scalable solution that balances accuracy, computational efficiency, and practical deployability. The demonstrated 18.5% improvement over unimodal baselines and superior performance compared to existing multimodal approaches validates the effectiveness of the proposed framework. Future work will focus on enhancing the system with temporal dynamics for long-term forecasting, exploring cross-domain adaptation techniques for improved geographical generalization, and investigating deployment strategies for real-time monitoring systems in developing regions.

Acknowledgment

The authors acknowledge the Air Pollution Image Dataset provided by Rouniyar *et al.* [23], and thank the anonymous reviewers for their constructive feedback that significantly improved this work.

References

[1] World Health Organization, "Ambient (outdoor) air pollution," WHO Fact Sheet, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-pollution

[2] A. Kumar, P. Goyal, and A. Sharma, "Challenges and opportunities in air quality monitoring using IoT: A comprehensive review," *Environ. Sci. Policy*, vol. 108, pp. 130–144, 2020.

[3] Y. Zhang, L. Chen, and M. Wang, "Deep learning approaches for atmospheric pollution monitoring using satellite imagery," *Remote Sens. Environ.*, vol. 267, p. 112741, 2022.

[4] H. Li, X. Zhang, and J. Liu, "LSTM-based spatiotemporal modeling for air quality prediction in smart cities," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7015–7028, 2023.

[5] N. Kumar, A. Sharma, and P. Gupta, "Statistical approaches for air quality index prediction: A comparative study," *Atmos. Environ.*, vol. 192, pp. 85–97, 2018.

[6] R. Sharma, S. Kamble, and A. Gunasekaran, "Big GIS analytics framework for electricity consumption forecasting," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 799–809, 2019.

[7] K. Singh, M. Pal, and V. Singh, "Time series forecasting of air quality parameters using ARIMA models," *Environ. Monit. Assess.*, vol. 192, no. 4, pp. 1–15, 2020.

[8] S. Hameed, A. Islam, K. Ahmad, et al., "Deep learning based multimodal urban air quality prediction and traffic analytics," *Sci. Rep.*, vol. 13, p. 22181, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-49296-7

[9] Q. Li, Y. Zhang, and H. Wang, "Deep learning for satellite-based air quality monitoring," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8678–8692, 2021.

[10] L. Wang, C. Zhang, and J. Li, "Land use classification using convolutional neural networks," *Remote Sensing*, vol. 14, no. 12, p. 2891, 2022.

[11] M. Chen, R. Liu, and S. Wang, "Atmospheric condition assessment using deep learning," *Atmospheric Research*, vol. 285, p. 106634, 2023.

[12] X. Liu, Y. Wang, and Z. Chen, "LSTM networks for environmental sensor data modeling," *Sensors*, vol. 22, no. 8, p. 3045, 2022.

[13] Y. Zheng, X. Chen, and W. Wang, "Graph attention networks for spatiotemporal air quality prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5820–5834, 2022.

[14] A. Ahmad, F. Khan, and S. Ali, "Multiscale graph neural networks for spatiotemporal AQI prediction," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–12, 2025.

[15] S. Gowthami, K. Sharma, and R. Patel, "Multimodal deep learning framework for AQI forecasting using satellite imagery," *Global NEST J.*, vol. 26, no. 8, pp. 1234–1248, 2024.

[16] H. Xia, X. Zhang, Y. Ma, et al., "ResGCN: A multi-modal deep learning approach for air quality prediction," *Entropy*, vol. 26, no. 1, p. 91, 2024.

[17] P. Sarkar, R. Dey, and S. Das, "Mobile-based air quality monitoring using deep convolutional networks," *Environ. Monit. Assess.*, vol. 197, no. 2, pp. 1–15, 2025.

[18] J. Xia, H. Zhang, and Y. Liu, "ResGCN: Residual graph convolutional networks for multimodal AQI prediction," *Remote Sens. Lett.*, vol. 13, no. 7, pp. 645–656, 2022.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4510–4520.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[21] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[23] P. Rouniyar, "Air pollution image dataset for AQI prediction," Kaggle Dataset, 2023. [Online]. Available: https://doi.org/10.34740/KAGGLE/DS/3152196

[24] Taiwan Environmental Protection Administration, "Air Quality Indicator–Taiwan EPA," 2023. [Online]. Available: https://airtw.epa.gov.tw/ENG/Information/Standard/AirQualityIndicator.aspx