Research Proposal : Multi-Modal Explainable Deep Learning for Air Pollution Classification

1 Introduction

Air pollution is a major public health concern in urban areas, with significant variations in particulate matter (PM2.5, PM10), gaseous pollutants (O_3, CO, SO_2, NO_2), and the overall Air Quality Index (AQI). Monitoring and predicting air quality using both ground-level images and sensor-based tabular data enables the development of robust, real-time systems that can provide actionable insights for policymakers and the public.

Existing solutions, such as Vision-AQ, primarily rely on convolutional neural networks (CNNs) with limited tabular inputs. These approaches often lack robustness under varying lighting, haze, and temporal conditions, and are generally "black-box" models with low interpretability.

This proposal introduces a novel dual-branch multi-modal architecture that integrates advanced image encoders, attention-based tabular processing, and explainable AI techniques. The framework aims to enhance accuracy, robustness, and interpretability for AQI classification, particularly under challenging environmental conditions.

2 Objectives

- Develop a dual-branch multi-modal network combining ground-level air pollution images and tabular sensor data.
- Introduce novel image preprocessing techniques that simulate real-world atmospheric conditions, including haze, fog, and brightness variability.
- Implement advanced image encoders such as EfficientNet-B3, ConvNeXt, and Vision Transformer to capture subtle Air Quality Index (AQI) visual cues.
- Implement attention-based tabular encoders (e.g., TabNet or Attention-based MLP) to dynamically weight pollutant features for each sample.
- Implement an attention-based fusion mechanism to adaptively combine image and tabular feature representations.
- Provide multi-modal explainability using Grad-CAM for image features and attention maps for tabular data to generate actionable insights.
- Evaluate the proposed framework on the Air Pollution Image Dataset collected from India and Nepal, consisting of approximately 12,240 images across six AQI classes.

3 Dataset

3.1 Dataset Source

The dataset used in this study is the Air Pollution Image Dataset from India and Nepal, obtained from Kaggle. It consists of a total of 12,240 ground-level air pollution images, each resized to 224×224 pixels.

3.2 Covered Cities

The dataset includes images collected from multiple urban locations across India and Nepal.

- **India**: ITO (Delhi), Dimapur (Nagaland), Spice Garden (Bengaluru), Knowledge Park III (Greater Noida), New Industrial Town (Faridabad), Borivali East (Mumbai), Oragadam (Tamil Nadu)
- **Nepal**: Biratnagar

3.3 Features

The dataset contains both visual and tabular attributes. The key features are summarized in Table 1.

Table 1 Dataset Features Description

Column	Description
Location	City or air quality monitoring station
Filename	Image file linking visual and tabular data
Year, Month, Day, Hour	Temporal attributes
AQI	Air Quality Index (aggregate pollution measure)
PM2.5	Fine particulate matter concentration ($\mu g/m^3$)
PM10	Coarse particulate matter concentration ($\mu g/m^3$)
O_3, CO, SO_2, NO_2	Optional gaseous pollutant features
AQI_Class	Categorical air quality label (6 classes: Good to Severe)

3.4 Selected Features

For multi-modal learning, only the features that exhibit strong visual correlation with atmospheric conditions were selected.

- **Used features**: AQI, PM2.5, and PM10, as these pollutants directly influence haze and visibility.
- **Ignored features**: O_3, CO, SO_2, and NO_2, as they exhibit weaker visual correlation in ground-level images.

4 Data Preprocessing

4.1 Image Preprocessing

Image preprocessing is performed using OpenCV and Albumentations to enhance robustness against real-world atmospheric variations. Table 2 summarizes the image preprocessing steps.

Table 2 Image Preprocessing Steps

Step	Details	Purpose / Novelty
Resize	224×224	Standard input size for CNNs and Vision Transformers
Normalize	ImageNet mean and std. dev.	Compatibility with pre-trained models
Augmentations	Haze, brightness, fog, blur, flip, rotation	Improves robustness to real-world atmospheric conditions (Novelty #1)

Expected Effect:

- Encourages the CNN to focus on relevant air pollution cues rather than background objects.
- Reduces misclassification in subtle or minority AQI classes such as *Very Unhealthy* and *Severe*.

4.2 Tabular Preprocessing

Tabular data preprocessing is conducted using Pandas and Scikit-learn to ensure numerical stability and temporal awareness. Table 3 summarizes the tabular preprocessing steps.

width=

Table 3 Tabular Data Preprocessing Steps

Step	Details	Purpose / Novelty
Standardization	Z-score normalization	Aligns feature scales across pollutants
Temporal Encoding	Hour, day, and month encoded as cyclic features	Captures daily and seasonal pollution trends (Novelty #3)
Optional Features	Lagged pollutant values, moving averages	Handles temporal dependencies in pollution data

Expected Effect:

- Avoids misclassification caused by time-of-day or seasonal bias.
- Enables the model to capture morning/evening pollution spikes and long-term trends.

5 Model Architecture

5.1 Overview

The proposed framework adopts a dual-branch multi-modal architecture consisting of an image branch and a tabular branch, followed by an attention-based fusion mechanism. The overall components of the network are summarized as follows:

- **Image branch**: Extracts visual representations from ground-level air pollution images.
- **Tabular branch**: Processes sensor-based pollutant data to learn sample-specific feature importance.
- **Fusion layer**: Combines image and tabular features using dynamic attention for robust prediction.
- **Classification head**: Predicts the Air Quality Index (AQI) class across six categories.
- **Explainability module**: Provides visual and numerical interpretation of feature contributions from both modalities.

5.2 Image Branch

The image branch leverages state-of-the-art visual backbones to extract discriminative features from air pollution images. Table 4 summarizes the architecture of the image branch.

5.2.1 Backbone Options

- EfficientNet-B3
- ConvNeXt
- Vision Transformer (ViT)

5.2.2 Layer Structure

Table 4 Image Branch Architecture

Layer	Details
Backbone	Pre-trained model, frozen during initial training
Global Average Pooling	Converts feature maps into a 1D feature vector
Dense Layer	64 neurons with ReLU activation
Output	64-dimensional image feature vector

Tools: TensorFlow/Keras and PyTorch
Improvements:

- Captures global context, haze patterns, and sky color gradients more effectively than ResNet50.
- Improves feature reliability under visually ambiguous atmospheric conditions.

5.3 Tabular Branch

The tabular branch encodes pollutant sensor readings using both standard and attention-enhanced architectures. Table 5 summarizes the architecture of the tabular branch.

5.3.1 Encoder Options

- Standard MLP: Dense ($64 \rightarrow 32 \rightarrow 16$)
- Advanced encoders: TabNet or Attention-based MLP

Input Features: AQI, PM2.5, PM10, and optional temporal features

Table 5 Tabular Branch Architecture

Layer	Details
Dense Layer	64 neurons with ReLU activation
Dense Layer	32 neurons with ReLU activation
Dense Layer	16 neurons with ReLU activation
Output	16-dimensional tabular feature vector

Improvements:

- Attention mechanisms dynamically weigh pollutant importance on a per-sample basis.
- Provides interpretable feature importance for tabular sensor data.

5.4 Fusion Layer

5.4.1 Fusion Strategy

- **Baseline**: Feature concatenation
- **Proposed**: Attention-based or gated fusion mechanism

Rationale:

- Dynamically balances image and tabular modalities depending on sample characteristics.
- For instance, under clear-sky conditions with elevated PM2.5 levels, the tabular modality dominates the prediction.

Implementation:

- Compute attention weights for each modality.
- Scale image and tabular feature vectors using learned attention weights.
- Fuse the weighted features and pass them through a Dense layer with 64 neurons, ReLU activation, and Dropout (0.5).

5.5 Classification Head

Table 6 shows the classification head architecture.

Table 6 Classification Head Architecture

Layer	Details
Dense Layer	64 neurons with ReLU activation
Dropout	0.5
Dense Layer	6 neurons with Softmax activation
Output	AQI class prediction (Good to Severe)

Optional Enhancements:

- Auxiliary contrastive loss to enforce stronger class separation in the feature space.
- Uncertainty estimation to flag low-confidence predictions (Novelty #6).

6 Explainability

To enhance transparency and trust, explainability is incorporated at both the modality level and the fusion level.

- **Image modality**: Grad-CAM is employed to highlight image regions that contribute most to the AQI prediction.
- **Tabular modality**: Attention maps are used to visualize the relative importance of individual pollutant features.
- **Fusion contribution**: The learned attention weights are visualized to show the relative influence of image and tabular modalities for each sample.

Expected Impact:

- Improves trustworthiness and interpretability for policymakers and citizens.
- Helps identify model failure cases and guides future data augmentation strategies.

7 Training Strategy

7.1 Two-Stage Training

A two-stage training strategy is adopted to stabilize optimization and improve generalization.

7.1.1 Stage 1: Feature Extraction

- Freeze the image backbone to retain pre-trained representations.
- Train only the dense and fusion layers.
- Number of epochs: 15.

7.1.2 Stage 2: Fine-Tuning

- Partially unfreeze the image backbone.
- Reduce the learning rate to 1×10^{-5}.
- Number of epochs: 10.

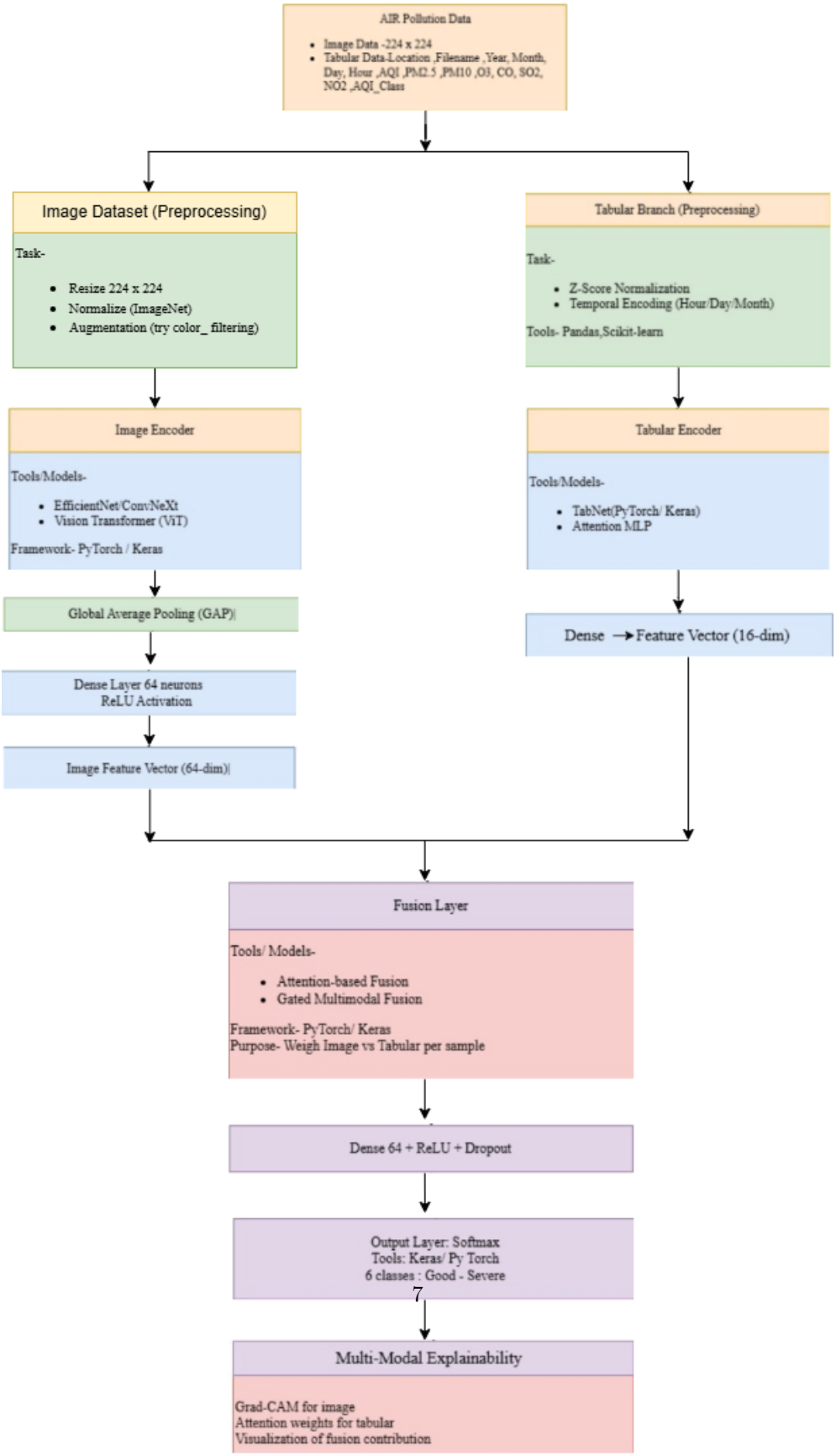

Fig. 1 Model architecture

7.2 Overfitting Control

The following regularization strategies are applied to mitigate overfitting:

- Early stopping based on validation loss.
- Model checkpointing to retain the best-performing weights.
- Dropout layers in dense and fusion blocks.

7.3 Error Reduction Strategies

Table 7 shows the error reduction strategies and their effect.

Table 7 Error Reduction Strategies and Their Effects

Component	Strategy	Effect
Image	Haze, fog, and brightness augmentation	CNN focuses on atmospheric cues
Image Encoder	EfficientNet / ViT	Captures subtle global features
Tabular Encoder	TabNet / Attention-based MLP	Highlights dominant pollutants
Fusion Layer	Attention-based fusion	Resolves conflicts between modalities
Temporal Features	Cyclic encoding	Reduces morning/evening misclassification
Training Loss	Class-weighted loss	Improves minority class detection
Uncertainty Module	Auxiliary output head	Flags borderline cases for review

8 Implementation Tools

Table 8 shows the implementation tools and libraries.

Table 8 Implementation Tools and Libraries

Component	Tools / Libraries
Image Preprocessing	OpenCV, Albumentations
Tabular Preprocessing	Pandas, Scikit-learn StandardScaler
Image Encoder	TensorFlow/Keras or PyTorch (EfficientNet, ConvNeXt, ViT)
Tabular Encoder	PyTorch TabNet or custom Attention-based MLP
Fusion Layer	PyTorch / TensorFlow attention modules
Classification	Keras/PyTorch Dense layers with Softmax
Explainability	Grad-CAM and attention visualization libraries
Training	PyTorch Lightning or Keras training loops
Evaluation	Scikit-learn metrics (Accuracy, Weighted F1-score)

9 Expected Results

The proposed multi-modal AQI classification framework is expected to achieve the following outcomes:

- **Accuracy**: Approximately 99% (comparable to the Vision-AQ baseline)
- **Weighted F1-score**: Approximately 0.99
- **Minority class performance**: Robust classification of subtle AQI classes such as *Very Unhealthy* and *Severe*
- **Interpretability**: Visual explanations for both image and tabular contributions
- **Deployment readiness**: A real-world capable pipeline able to handle varying haze, lighting, and temporal conditions

10 Novel Contributions

The key contributions of this work are summarized as follows:

- **Robust Multi-Modal Preprocessing**: Haze/fog augmentation for images and temporal cyclic encoding for tabular data
- **Advanced Image Encoders**: Use of EfficientNet-B3, ConvNeXt, and Vision Transformer (ViT) to capture subtle atmospheric cues
- **Attention-Based Tabular Encoding**: Dynamic weighting of pollutants per sample
- **Adaptive Fusion Layer**: Resolves conflicts between image and sensor modalities
- **Explainability Across Modalities**: Grad-CAM for images, attention maps for tabular features, and visualization of fusion contribution
- **Real-World Readiness**: Handles time-of-day, seasonal trends, and low-confidence predictions

Key Point: The novelty lies in robustness, interpretability, and generalization, rather than solely numeric accuracy.

11 Conclusion

This research proposal presents a novel, multi-modal, and explainable deep learning framework for AQI classification using ground-level images and sensor data. The approach emphasizes:

- **Robustness**: Effective under varying environmental and temporal conditions
- **Interpretability**: Provides actionable insights for policymakers and citizens
- **Minority class reliability**: Accurate prediction for Severe AQI levels

The proposed pipeline is developer-ready, with detailed guidance on tools, layer structures, and training strategies, providing a strong foundation for real-world deployment and further research.

12 Reference paper

The dataset we used, obtained from Kaggle, has also been employed in [1]. The concept of the multi-modal architecture was inspired by [2]. Additionally, we referred to [3] and [4] for further insights and guidance.

References

[1] Mehmood, F., Rehman, S.U., Choi, A.: Vision-aq: Explainable multi-modal deep learning for air pollution classification in smart cities. Mathematics **13**(18) (2025) https://doi.org/10.3390/math13183017

[2] Ahsan, S., Hossain, E., Sharif, O., Das, A., Hoque, M.M., Dewan, M.: A multimodal framework to detect target aware aggression in memes. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2487–2500. Association for Computational Linguistics, St. Julian's, Malta (2024). https://doi.org/10.18653/v1/2024.eacl-long.153 . https://aclanthology.org/2024.eacl-long.153/

[3] Chakroborty, B., Rudra, K., Chakrabarty, D., Naskar, S., Das, S.: Decoding spatiotemporal dynamics of air pollution and its underlying drivers in kolkata metropolitan area through integrated field investigation and geospatial analysis. Discover Cities **2**(1), 77 (2025) https://doi.org/10.1007/s44327-025-00118-7

[4] Karamti, H., Aldrees, A., Alghamdi, N., Umer, M., Nappi, M.: Intelligent multi-modal data implementation with capsulenet for accurate air quality index prediction. Complex Intelligent Systems **11** (2025) https://doi.org/10.1007/s40747-025-02082-6