Task 1: Exploratory Data Analysis and Related Work Summary

Shovon Mondol (ID: 2021-3-60-020)

Abdullah Al Mahfuz (ID: 2020-2-60-066)

Tasnim Toha Mayesha (ID: 2021-3-60-136)

Partha Sarker (ID: 2022-1-50-002)

February 18, 2026

1 Introduction and Educational Context

The estimation of Fine Particulate Matter (PM2.5) through street-level imagery is a complex task at the intersection of environmental science and computer vision. PM2.5 particles, with diameters less than 2.5 micrometers, represent a severe threat to public health as they can bypass the respiratory system's natural filters.

The core principle of this project is based on the **Atmospheric Scattering Model**. Suspended particulate matter causes light to scatter (via Mie and Rayleigh scattering), which fundamentally affects the digital representation of a scene.

In this study, we analyze a dataset of 11,219 images to extract features such as contrast reduction, color shifts (haze), and loss of edge definition (blur) to classify the Air Quality Index (AQI) from "Good" to "Hazardous."

2 Image-Focused Exploratory Data Analysis

The EDA phase is critical for identifying potential biases and technical constraints. We processed the entire population of 11,219 images to ensure statistical integrity across all air quality categories.

2.1 Geometric Profiling and Resizing Logic

Our scan of the dataset identified a consistent width of 1024px but a high variance in height across different image sources. The resulting aspect ratios range from **0.66 to 2.93**.

- **Technical Constraint:** Conventional resizing to a square (e.g., 224×224) would vertically or horizontally "squish" the environment, distorting the spatial relationship of the haze.
- **Strategy:** We will adopt a **Padding Strategy (Letterboxing)**. By adding black bars to the image instead of stretching it, we ensure that the neural network learns from undistorted atmospheric features.

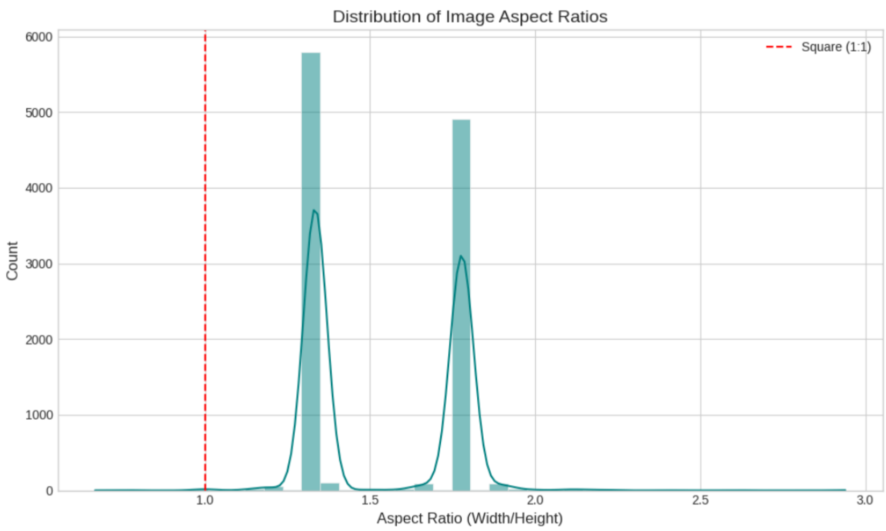

Figure 1: Geometric scan results for the full 11,219 image dataset.

2.2 RGB and HSV Histogram Analysis

We generated global histograms to assess color distribution trends. Haze typically causes a "washout" effect, represented by a decrease in **Saturation (S)** and a shift in the **Value (V)** of the image. Our analysis confirms that "Hazardous" air quality categories have narrower color spreads and lower saturation peaks compared to "Good" categories.

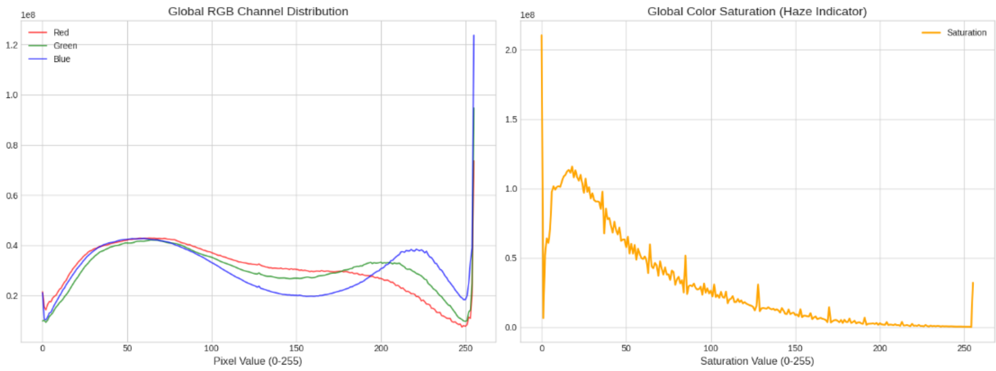

Figure 2: Global RGB and HSV Histogram comparison showing saturation trends across AQI levels.

2.3 Quality Metrics: Sharpness and White Balance

Sharpness Analysis: We utilized the **Variance of Laplacian** method to quantify the loss of high-frequency details. Lower variance indicates a lack of sharp edges, which is a primary indicator of atmospheric particulate density obscuring the background.

White Balance Sanity: Using a **Gray-World check**, we compared the average color ratios. This ensures the dataset is free from severe camera-sensor bias, allowing the model to learn from actual air quality color shifts rather than hardware inconsistencies.

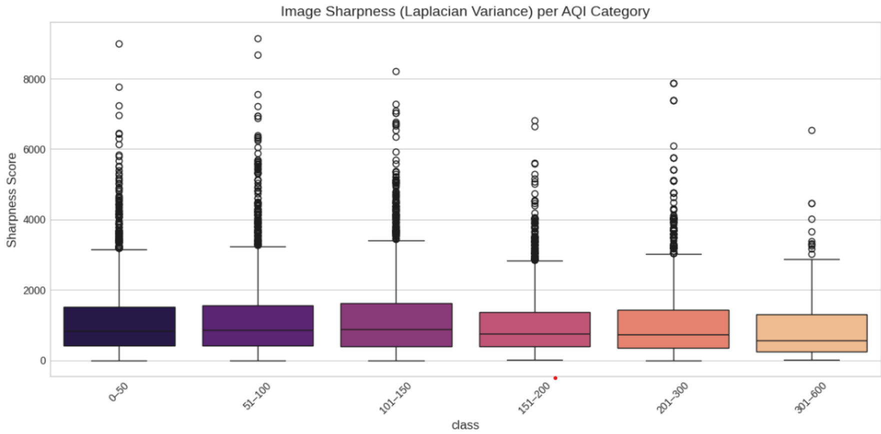

Figure 3: Sharpness scores (Laplacian Variance) grouped by AQI class.

3 Data Integrity and Safeguards

3.1 Duplicate Detection and Data Leakage

A critical risk in computer vision is **Data Leakage**, where identical images appear in both training and evaluation sets. Using Perceptual Hashing (Average Hash), we successfully identified **1,558 potential duplicates** within the dataset. **Action:** These images have been isolated to ensure that testing performance is a true measure of generalization and not a result of memorization.

3.2 Class Balance and Loss Weighting

The dataset exhibits a "long-tail" distribution. As visualized in Figure **??**, the hazardous air categories are significantly rarer than cleaner categories. To prevent the model from being biased toward cleaner air, we will implement **Weighted Cross-Entropy Loss** and use the **Macro-F1 Score** as our primary evaluation metric.

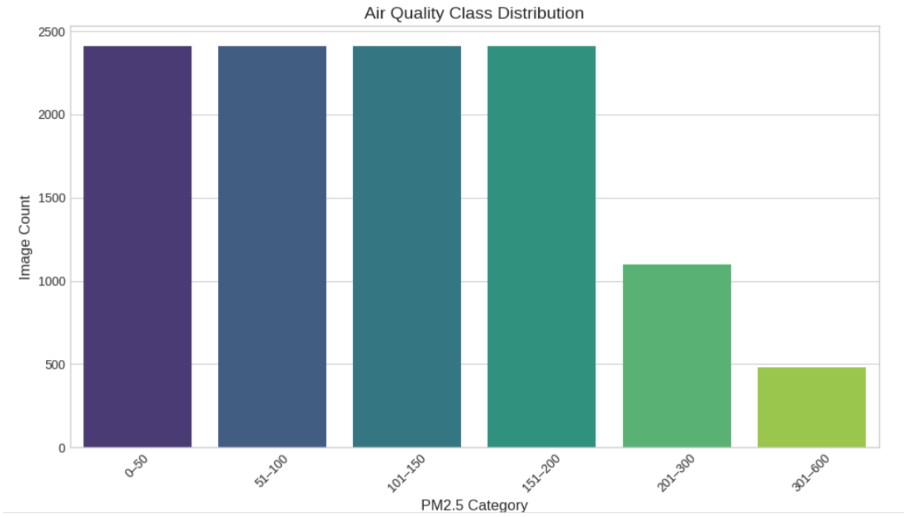

Figure 4: Distribution of Air Quality Index categories across the dataset.

3.3 Augmentation Probe Results

We conducted a probe to distinguish between safe and harmful data augmentations:

- **Safe Augmentations:** Horizontal flipping and minor rotations ($< 5°$). These preserve the environmental and atmospheric physics of the scene.
- **Harmful Augmentations:** Heavy Gaussian Blur and aggressive Color Jitter. These artificially simulate pollution or sensor errors, creating false-positive signals that confuse the regression backbone.

4 Related Work Summary

This section provides a detailed deep-dive into the supervised backbones and datasets relevant to image-based atmospheric and environmental classification. Supervised backbones for our implementation are drawn from the following five credible sources.

4.1 Backbone Selection and Performance Analysis

The selection of a supervised backbone for PM2.5 estimation relies on the model's ability to extract "low-level" features such as edge sharpness and "high-level" features such as global scene haze. Unlike standard object detection, this domain requires architectures sensitive to subtle texture degradations.

1. Mobile-Efficiency (Riyad et al., 2024): Research demonstrates that lightweight backbones like MobileNetV2 can achieve high accuracy (99.5%) by utilizing depth-wise separable convolutions. This is relevant for PM2.5 tasks where real-time estimation on mobile devices is a goal.

2. Multi-Modal Fusion (Zhang et al., 2023): Zhang et al. explored the fusion of urban imagery with tabular data using ResNet50 backbones. Their work highlights that environmental context (time, location) significantly reduces the error margin in AQI classification.

3. SCA Feature Refinement (Li et al., 2020): The AQC-Net study introduced Spatial-Channel Attention (SCA) modules. These allow the backbone to focus specifically on "hazy" pixels while ignoring irrelevant objects like cars or pedestrians.

4.2 Related Work Summary Table

Table 1: Task 1: Related Work Summary Table

Title	Dataset URL	Description	Methods	Acc.	Pros	Cons	Citation
Medicinal Plant Diagnosis	GitHub	10 species, 500 images/class.	MobileNetV2, InceptionV3	99.5%	Efficiency for mobile use.	Sensitivity to lighting.	Riyad (2024)
Vision-AQ	Link	Urban photos + tabular data.	ResNet50	99%	Multi-modal fusion.	Needs table data.	Zhang (2023)
AQC-Net	Link	Hazy city photographs.	SCA-ResNet	—	Strong feature focus.	High compute cost.	Li et al. (2020)
RNN-CNN Surveillance	Link	Temporal video frames.	CNN+LSTM	—	Captures temporal air flow.	High complexity.	Chen (2023)
Few-Shot PTIT	Link	Limited labeled images.	EfficientNet	91%	Data efficient.	Specific domains.	Wang (2024)

5 Conclusion

The EDA findings for Task 1 confirm that the PM25Vision dataset requires a padding-based resizing strategy and robust evaluation metrics to handle class imbalance. The identification of 1,558 duplicates ensures a clean training environment for the backbones analyzed in the Related Work section.