

# Physics-Constrained Lightweight Neural Networks for Calibrated Smog-Level Classification

BMVC 2025 Submission # ??

## Abstract

Air pollution remains a pressing environmental and public health concern, with fine particulate matter ( $PM_{2.5}$ ) linked to millions of premature deaths annually. While ground-based sensors provide accurate air quality measurements, their high cost and sparse coverage limit large-scale deployment. As an alternative, recent studies have explored image-based approaches, using street-level photographs to infer pollution severity. However, most existing methods employ large deep networks that are computationally expensive, overlook physical principles of haze formation, and provide poorly calibrated predictions.

This work presents a lightweight, physics-informed, and calibrated framework for smog-level classification from street-level imagery. The backbone model is MobileNetV3-Small, a compact convolutional network designed for efficient CPU inference. To improve recognition of minority pollution categories, physics-informed regularization is introduced by embedding atmospheric scattering cues—image contrast, sharpness, and dark channel priors—into the loss function. Probability reliability is enhanced through test-time augmentation and temperature scaling, while interpretability is supported by Grad-CAM visualizations and monotonic physics-feature trends.

Experiments on the Smartphone-Based Air Pollution Image Dataset (SAPID) demonstrate that the proposed Physics v2 model achieves 86.5% test accuracy and a macro-F1 score of 0.835, outperforming the baseline MobileNetV3-Small (81.1%, 0.756 macro-F1). The model operates in real time on CPU hardware at over 100 FPS, with an Expected Calibration Error (ECE) of 0.071. These results highlight the potential of integrating lightweight architectures, physics priors, and calibration techniques to deliver accurate, interpretable, and deployable computer vision systems for low-cost urban air quality monitoring.

## 1 Introduction

Air pollution has emerged as one of the most significant environmental health risks worldwide. According to the World Health Organization (WHO), approximately seven million premature deaths occur annually due to exposure to fine particulate matter and other air pollutants [? ]. Long-term exposure to  $PM_{2.5}$  and  $PM_{10}$  has been linked to cardiovascular diseases, respiratory illnesses, and impaired lung development in children [? ]. Rapid urbanization and industrialization have intensified air quality concerns in low- and middle-income countries, where reliable monitoring infrastructure is often limited.

Traditional approaches to monitoring air quality rely on ground-based stations equipped with high-precision sensors. Although these stations provide accurate measurements of pollutant concentrations, their high installation and maintenance costs restrict dense deployment. Consequently, many regions suffer from sparse coverage, limiting real-time and fine-grained air quality assessments [? ]. This gap motivates the exploration of alternative, low-cost methods that leverage ubiquitous sensing modalities.

In recent years, the proliferation of smartphones and public cameras has enabled large-scale collection of street-level imagery, which captures visible effects of air pollution such as haze, smog, and reduced visibility. These images provide a complementary data source for air quality estimation and have sparked research in computer vision-based environmental monitoring. Early methods exploited handcrafted features, including edge sharpness and color attenuation, to estimate haze concentration. With the advent of deep learning, convolutional neural networks (CNNs) have been applied to regress particulate matter levels directly from urban images [? ]. While these approaches show promise, several challenges remain unresolved.

First, many existing vision-based models rely on large-scale networks such as ResNet or VGG, which achieve high predictive performance but require GPUs for efficient inference. This dependence hinders deployment on low-power devices in resource-constrained environments. Second, most methods treat the problem as a purely data-driven task, often overlooking the physical principles of atmospheric scattering that govern the visibility of objects under haze and smog. Third, rare but critical classes—such as “Unhealthy for Sensitive Groups” or “Very Unhealthy”—are underrepresented in existing datasets, causing imbalanced performance across categories. Finally, deep models often suffer from overconfidence in predictions, underscoring the importance of calibration for trustworthy decision-making in environmental applications [? ].

To address these limitations, recent work has shifted towards lightweight and interpretable architectures. MobileNet families, particularly MobileNetV3, have been shown to deliver strong accuracy-efficiency trade-offs for image classification tasks [? ]. At the same time, physics-informed learning paradigms have gained attention, where domain knowledge is incorporated into loss functions or architectures to regularize learning. In the context of air pollution, physical cues such as image contrast, sharpness, and dark channel priors can provide valuable inductive biases aligned with the visibility degradation caused by haze. Furthermore, calibration strategies such as temperature scaling have demonstrated effectiveness in aligning predicted probabilities with empirical accuracies, improving trustworthiness in decision-critical tasks [? ]. Interpretability methods, including Grad-CAM [? ], further support the transparency of deep models by highlighting the regions of an image that most influence predictions.

Against this backdrop, this study investigates a lightweight, physics-informed, and calibrated deep learning framework for smog-level classification using street-level images. The approach is based on MobileNetV3-Small, augmented with physics-based regularization and temperature scaling, and is designed to operate efficiently on CPUs without the need for dedicated GPUs. The model is evaluated on the Smartphone-based Air Pollution Image Dataset (SAPID), which contains five air quality categories ranging from Good to Very Unhealthy. Experimental results demonstrate that the integration of physics priors improves recognition of minority classes, while calibration ensures reliable confidence estimation. The framework thus contributes to the development of deployable, interpretable, and accurate computer vision solutions for air quality monitoring.

**Key Contributions**

This study makes the following contributions:

- A lightweight smog-level classification framework based on MobileNetV3-Small, enabling real-time deployment on CPU devices.
- A physics-informed regularization strategy that leverages atmospheric scattering priors (contrast, sharpness, dark channel) to improve classification, particularly for minority classes.
- A calibrated prediction pipeline using test-time augmentation and temperature scaling, achieving reliable confidence estimates with low Expected Calibration Error (ECE).
- Comprehensive evaluation on SAPID with accuracy, macro-F1, calibration, interpretability, and efficiency analyses.

**2 Related Work**

**2.1 Lightweight CNNs for Resource-Constrained Vision**

Compact convolutional networks have been central to on-device perception. MobileNetV2 introduced inverted residuals with linear bottlenecks to preserve representational power while reducing multiplications via depthwise separable convolutions [? ]. MobileNetV3 further combined hardware-aware neural architecture search with novel activation (h-swish) and squeeze-and-excitation, yielding state-of-the-art mobile accuracy under tight latency constraints in both “Large” and “Small” variants [? ]. These families remain strong baselines for CPU-only deployment where low parameter count, FLOPs, and cache-friendly operators are decisive.

**2.2 Haze and Air-Quality Estimation from Images**

Early work in adverse-weather vision formalized the atmospheric scattering model, showing that observed radiance combines attenuated scene radiance with airlight and that visibility degrades with particle density [? ]. The dark channel prior (DCP) provided a seminal single-image dehazing heuristic tightly coupled to that physics [? ]. Subsequent surveys and benchmarks compared dehazing families across synthetic/real data (e.g., RESIDE) and highlighted trade-offs among priors and learning-based models [? ].

Image-based air-quality estimation has evolved from hand-crafted visibility cues and support vector regression to deep CNN regressors/classifiers. Recent studies report PM<sub>2.5</sub>/AQI estimation from street or surveillance imagery, including daytime–nighttime modeling and sequence learning (CNN–LSTM) to leverage temporal continuity [? ]. Practical deployments have also been explored for real-time regression from camera feeds [? ]. Complementary high-resolution satellite–vision pipelines and hybrid sensing systems continue to expand coverage and robustness.

**2.3 Physics-Informed Machine Learning (PIML)**

Physics-informed neural networks (PINNs) embed governing equations (e.g., PDE residuals) as soft constraints during training, improving sample efficiency and physical fidelity [? ].

In vision under haze/smog, physically motivated image cues—contrast/visibility, sharpness (e.g., Laplacian variance), and dark-channel statistics—are often used as priors or regularizers aligned with the scattering model. Recent works exploit haze-level indicators derived from DCP to guide learning and data cleaning, further bridging data-driven models with interpretable physics [? ].

## 2.4 Calibration and Post-hoc Temperature Scaling

Modern deep classifiers are frequently miscalibrated—overconfident in incorrect predictions. Guo *et al.* provided a systematic analysis and showed that simple temperature scaling, fitted on a validation set, markedly improves probability calibration without affecting accuracy [? ]. In safety- and policy-relevant environmental applications, calibrated confidence estimates can be as important as raw accuracy for thresholding and decision support.

## 2.5 Datasets for Image-based Air Quality and Haze

For dehazing and visibility estimation, RESIDE established a comprehensive benchmark with synthetic and real subsets to evaluate single-image dehazing methods [? ]. For image–pollution learning with paired measurements, HVAQ contributed a high-resolution dataset linking time-synchronized images to PM<sub>2.5</sub>, PM<sub>10</sub>, temperature, and humidity across multiple cities [? ]. For smartphone imagery, the Smartphone-Based Air Pollution Image Dataset (SAPID) organizes street-level photos into five EPA AQI categories, enabling lightweight classification studies directly from consumer images [? ].

## 2.6 Domain Shift, Cross-Region Generalization, and Test-Time Adaptation

A practical obstacle for image-based air-quality estimation is *dataset shift*: deployment imagery may differ from training data due to camera optics, geographic locale, season, illumination, and co-occurring weather (e.g., fog versus smog), which alters the joint distribution of pixels and labels [? ]. Classical visual domain adaptation addresses such covariate shift by aligning source and target representations, using either discrepancy minimization (e.g., CORAL alignment of second-order statistics) [? ] or adversarial learning that encourages domain-invariant features (e.g., domain-adversarial training) [? ]. Surveys of deep domain adaptation document substantial performance gains from these strategies when the target domain lacks labels [? ? ].

More recently, *test-time adaptation* dispenses with access to source data during deployment and adapts a model online using only the incoming unlabeled stream. Entropy-minimization at test time (TENT) updates normalization and affine parameters to reduce predictive uncertainty under shift, showing improvements across image classification benchmarks without retraining from scratch [? ]. Source-free adaptation approaches such as SHOT leverage only a source-trained hypothesis and self-supervised objectives to adapt to the target distribution [? ]. In environmental vision, where cameras and cities vary widely and labeled data are scarce, such lightweight adaptation is attractive for maintaining robustness under operational shifts while respecting data-governance constraints (e.g., when source data cannot be moved).

For air-quality classification specifically, domain shift manifests as systematic biases across cities and acquisition devices, and as seasonal drifts in aerosol composition that affect

color channels and haze thickness. Methods that are both compact (for edge/CPU deployment) and amenable to on-the-fly adaptation are therefore relevant complements to physics-informed inductive biases.

## 2.7 Survey Synthesis and Research Gap

Across lightweight CNNs, image-based haze/AQ estimation, physics-informed learning, calibration, and domain adaptation, several gaps remain salient for deployment in resource-constrained urban monitoring:

- **Deployment realism:** Many models assume GPU availability and large backbones; fewer studies report accuracy–latency trade-offs for CPU-only inference suitable for smartphones or embedded devices (cf. MobileNet families) [? ].
- **Physics alignment:** Learning pipelines often remain purely data-driven; the literature shows fewer examples where atmospheric scattering cues (contrast, sharpness, dark-channel statistics) are explicitly enforced during training in a way that improves minority-class recognition.
- **Probability reliability:** Image-based AQ estimation systems rarely report calibration metrics (e.g., ECE); yet calibrated confidence is critical for thresholding and alerting in environmental applications [? ].
- **Low-data imbalance:** Datasets such as SAPID exhibit severe class imbalance, with under-represented high-severity categories (e.g., USG, Very Unhealthy). Robust learning under such imbalance, coupled with interpretability (e.g., Grad-CAM), remains under-explored at the edge.
- **Shift robustness:** Cross-region generalization and test-time adaptation are underutilized in AQ-from-images, despite clear distribution shifts across cameras, cities, and seasons [? ? ? ].

These observations indicate a research gap for a *lightweight, physics-informed, and calibrated* air-quality classifier that performs on CPU-class hardware, improves minority-class recognition, and remains robust under realistic distribution shifts via simple test-time procedures. Such a design directly addresses operational constraints in low-cost, scalable urban monitoring.

## 3 Research Methodology

### 3.1 Dataset

The Smartphone-Based Air Pollution Image Dataset (SAPID) [? ] is used as the primary benchmark. It consists of 492 street-level images annotated with five air quality categories following the US Environmental Protection Agency (EPA) Air Quality Index (AQI): Good, Moderate, Unhealthy for Sensitive Groups (USG), Unhealthy, and Very Unhealthy. As shown earlier (Table ??), the dataset is highly imbalanced, with only 32 samples in the USG class and 40 in Very Unhealthy, in contrast to 188 in the Moderate class. Images are resized to  $224 \times 224$  and normalized using ImageNet mean and standard deviation. Data

augmentation includes random flips, rotations, and color jittering to improve generalization [? ].

### 3.2 Baseline Model: MobileNetV3-Small

The baseline classifier is based on MobileNetV3-Small [? ], which uses depthwise separable convolutions, inverted residuals, and squeeze-and-excitation blocks with h-swish activations. For an input image  $x \in \mathbb{R}^{3 \times H \times W}$ , the network extracts features  $f(x) \in \mathbb{R}^d$ , which are passed to a fully connected classifier:

$$z = Wf(x) + b, \quad \hat{y} = \text{softmax}(z),$$

where  $z \in \mathbb{R}^C$  are the logits,  $C = 5$  is the number of AQI classes, and  $\hat{y}$  is the predicted probability distribution. The baseline is trained using class-weighted cross-entropy loss:

$$\mathcal{L}_{CE} = - \sum_{i=1}^C w_i y_i \log \hat{y}_i,$$

where  $y$  is the one-hot ground truth vector and  $w_i$  are class weights to address dataset imbalance.

### 3.3 Physics-Informed Model

To integrate atmospheric scattering priors [? ? ], three physics-inspired features are extracted from each input image:

1. **Contrast** ( $C(x)$ ): Standard deviation of pixel intensities or global contrast tensor, expected to decrease with pollution severity.
2. **Sharpness** ( $S(x)$ ): Laplacian variance measuring edge clarity, also decreasing under haze.
3. **Dark Channel Prior** ( $D(x)$ ): Defined as

$$D(x) = \min_{c \in \{R, G, B\}} \left( \min_{u \in \Omega(x)} I_c(u) \right),$$

where  $\Omega(x)$  is a local patch around pixel  $x$ , and  $I_c(u)$  is the intensity in color channel  $c$ . The DCP increases under heavier smog.

Let  $\phi(x) = [C(x), S(x), D(x)]$  be the physics feature vector. The model enforces monotonic consistency by penalizing violations of the expected order across pollution categories. For two samples  $(x_i, y_i)$  and  $(x_j, y_j)$  with  $y_i < y_j$  (less polluted vs. more polluted), the physics-informed ranking loss is defined as:

$$\mathcal{L}_{physics} = \sum_{i,j} \left( \max(0, (C(x_j) - C(x_i))) + \max(0, (S(x_j) - S(x_i))) + \max(0, (D(x_i) - D(x_j))) \right)$$

This enforces decreasing contrast and sharpness, and increasing dark channel prior, with rising pollution severity.

The total objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{phys} \mathcal{L}_{physics},$$

where  $\lambda_{phys} = 0.2$  balances classification and physics constraints.

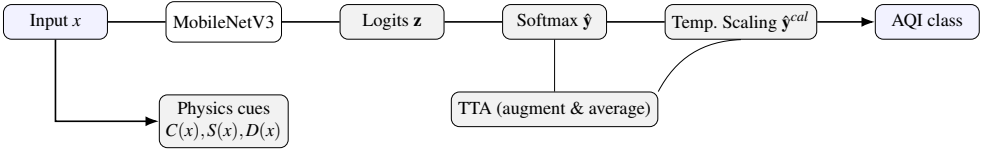


Figure 1: Architecture overview (*what* components exist): MobileNetV3 backbone with a physics-cues branch (used at training time) and inference-time calibration and TTA. No training logic is shown here.

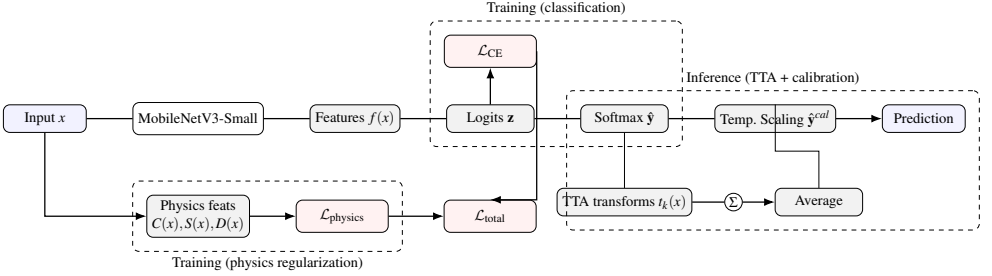


Figure 2: End-to-end pipeline (*how* it runs): training optimizes  $\mathcal{L}_{CE}$  and  $\mathcal{L}_{physics}$  to form  $\mathcal{L}_{total}$ ; inference applies TTA and temperature scaling for calibrated predictions.

### 3.4 Calibration and Robustness

Neural networks are often miscalibrated, producing overconfident predictions [? ]. Logits  $z$  are calibrated using temperature scaling:

$$\hat{y}_i^{cal} = \frac{\exp(z_i/T)}{\sum_{j=1}^C \exp(z_j/T)},$$

where  $T > 0$  is a learned temperature parameter optimized on the validation set. A perfectly calibrated model satisfies

$$P(Y = y \mid \hat{p}) = \hat{p}, \quad \forall \hat{p} \in [0, 1].$$

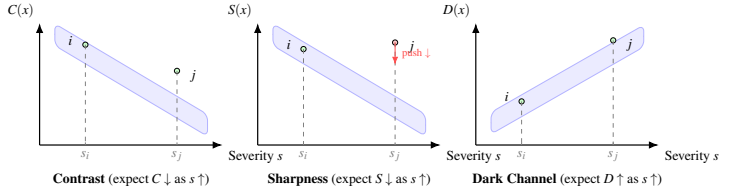
To improve robustness, test-time augmentation (TTA) generates  $K$  transformations  $\{t_k(x)\}_{k=1}^K$  for each input, and predictions are averaged:

$$\hat{y}_{TTA} = \frac{1}{K} \sum_{k=1}^K \hat{y}(t_k(x)).$$

Physics loss mechanics for pairwise ranking is depicted in Figure ?? . For two samples with severities  $s_i < s_j$ , contrast  $C$  and sharpness  $S$  are expected to decrease, while dark channel  $D$  is expected to increase with severity. A hinge margin  $\delta$  penalizes violations, forming the physics-informed component of the total objective.

### 3.5 Training Setup

The models are trained in PyTorch using the AdamW optimizer with weight decay  $10^{-4}$  and an initial learning rate of  $3 \times 10^{-4}$ , following a cosine decay schedule. A batch size of 32 is



Pairwise ranking loss for  $s_i < s_j$ :  
 $\mathcal{L}_{\text{phys}} = \max(0, C(x_j) - C(x_i) + \delta) + \max(0, S(x_j) - S(x_i) + \delta) + \max(0, D(x_i) - D(x_j) + \delta)$ .  
 Margin  $\delta > 0$  enforces monotonic ordering; total objective  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{phys}} \mathcal{L}_{\text{phys}}$ .

Figure 3: Physics-informed pairwise ranking: contrast  $C$  and sharpness  $S$  decrease with severity, while dark channel  $D$  increases. Violations incur a hinge penalty (margin  $\delta$ ).

used. Early stopping is triggered if validation macro-F1 does not improve for six consecutive epochs. To address dataset imbalance, class-weighted cross-entropy with label smoothing ( $\epsilon = 0.05$ ) is employed. Gradient clipping ( $\ell_2$  norm capped at 1.0) prevents instability.

### 3.6 Evaluation Metrics

Performance is evaluated using accuracy and macro-F1, the latter being particularly important under imbalanced data distributions. In addition, per-class precision, recall, and average precision (AP) are reported. Reliability of predicted probabilities is measured using Expected Calibration Error (ECE) [? ]. Model efficiency is evaluated in terms of parameter count, file size, CPU latency (milliseconds per image), and frames per second (FPS). Interpretability is assessed qualitatively with Grad-CAM [? ] and quantitatively with physics-feature alignment trends.

## 4 Experimental Setup

### 4.1 Hardware and Software Environment

All experiments were conducted in Google Colab with access to an NVIDIA Tesla T4 GPU (16 GB) when available and CPU-only mode otherwise. The model architecture and training routines were implemented in PyTorch 2.1, using the Torchvision model zoo for MobileNetV3-Small initialization [? ]. Training and evaluation pipelines were executed under Python 3.12, with supporting libraries including NumPy, Pandas, and scikit-learn. Grad-CAM visualizations were generated using the TorchCAM library.

### 4.2 Data Splits

The SAPID dataset [? ] was split into training (70%), validation (15%), and test (15%) sets, ensuring class-stratified sampling to preserve distribution across splits. The validation set was used for hyperparameter tuning, early stopping, and calibration (temperature scaling). The final test set was reserved strictly for performance reporting.



### 4.3 Preprocessing and Augmentation

All images were resized to  $224 \times 224$  pixels and normalized with ImageNet mean and standard deviation values. To increase robustness and mitigate overfitting, training augmentations included:

- Random horizontal flips and random rotations ( $\pm 15^\circ$ ).
- Random brightness, contrast, and saturation jitter.
- Random cropping and scaling.

For inference, standard resizing was applied, and test-time augmentation (TTA) was used to generate multiple crops and scales per image, with averaged predictions.

### 4.4 Training Configuration

Models were trained using the AdamW optimizer with an initial learning rate of  $3 \times 10^{-4}$  and cosine decay scheduling. A batch size of 32 and early stopping with patience of six epochs were employed. Gradient clipping with a maximum  $\ell_2$  norm of 1.0 stabilized optimization. To address dataset imbalance, class-weighted cross-entropy loss was combined with label smoothing ( $\epsilon = 0.05$ ). For the physics-informed variant, a regularization weight of  $\lambda_{phys} = 0.2$  was used for the physics-based loss component.

### 4.5 Evaluation Protocol

Performance was assessed using overall accuracy and macro-F1 score, the latter being critical for imbalanced datasets. Class-wise precision, recall, and average precision (AP) were also reported. Calibration quality was evaluated via Expected Calibration Error (ECE) following [? ]. Model interpretability was analyzed through Grad-CAM heatmaps [? ] and physics-feature trend plots (contrast, sharpness, dark channel prior). Efficiency metrics included parameter count, model file size, CPU latency, and frames per second (FPS).

### 4.6 Baselines and Comparisons

The following configurations were compared:

1. **Baseline:** MobileNetV3-Small with standard cross-entropy training.
2. **Physics v1:** Initial physics-informed model with basic regularization.
3. **Physics v2:** Improved physics-informed model with tuned  $\lambda_{phys}$  and ranking constraints.
4. **Physics v2 + Calibration:** Incorporating TTA and temperature scaling for reliability.
5. **Ensemble:** Weighted averaging of baseline and physics-informed predictions.

This setup ensures a fair ablation study and highlights the contributions of physics-informed regularization and calibration.

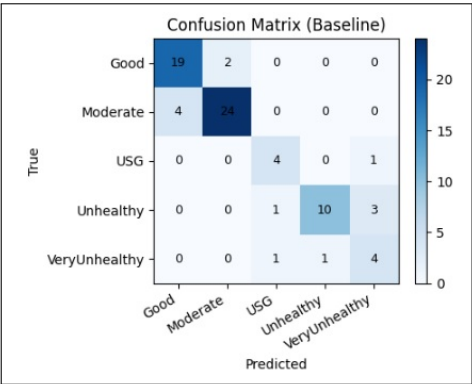
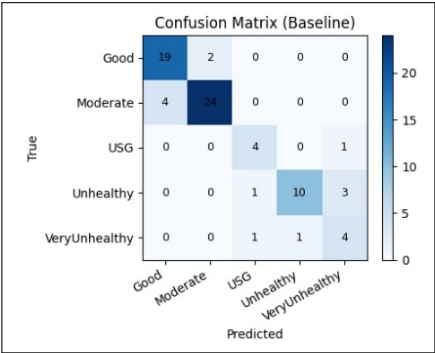
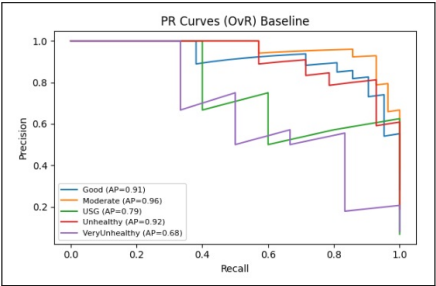


Figure 4: Class distribution of SAPID dataset across five AQI categories: Good, Moderate, USG, Unhealthy, Very Unhealthy.



(a) Confusion Matrix (Baseline)



(b) PR Curves (Baseline)

Figure 5: Baseline results: (a) confusion matrix shows strong performance on major classes but confusion in minority categories; (b) PR curves reveal lower average precision for “Very Unhealthy”.

## 5 Results and Discussion

### 5.1 Dataset Distribution

Figure ?? shows the SAPID dataset distribution, highlighting significant class imbalance. Minority categories (USG and Very Unhealthy) contain fewer than 50 samples each, motivating the use of class weighting and physics-informed regularization.

### 5.2 Baseline Performance

The baseline MobileNetV3-Small achieved 81.1% test accuracy and 0.756 macro-F1 with TTA. Figure ?? presents the confusion matrix and one-vs-rest precision-recall (PR) curves. While performance was strong for “Good” and “Moderate”, the model misclassified minority categories (USG and Very Unhealthy).

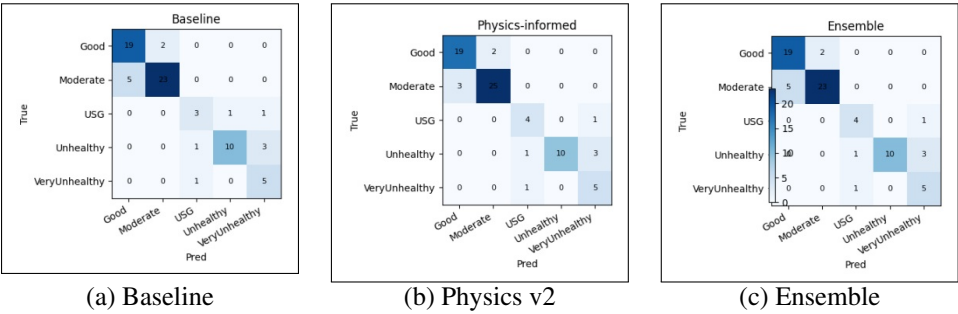


Figure 6: Confusion matrices comparing Baseline, Physics v2, and Ensemble models. Physics-informed regularization improves rare-class recognition, particularly USG and Very Unhealthy.

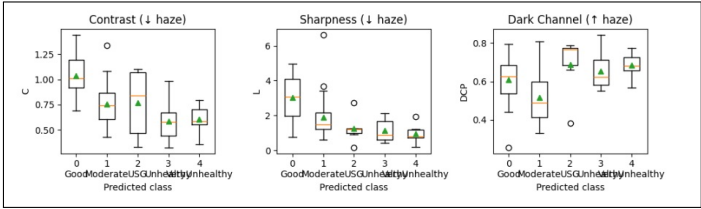


Figure 7: Boxplots of physics features (contrast, sharpness, dark channel prior) across predicted classes. Monotonic trends confirm physical interpretability of the model.

### 5.3 Physics-Informed Model Improvements

The Physics v2 model improved recognition of minority categories by enforcing monotonic trends in haze-sensitive features. It achieved 85.1% accuracy and 0.804 macro-F1 with TTA. With calibration, performance further improved to 86.5% accuracy and 0.835 macro-F1. Figure ?? compares confusion matrices of baseline, physics-informed, and ensemble models.

### 5.4 Physics-Feature Trends

To validate interpretability, physics features (contrast, sharpness, dark channel prior) were analyzed against predicted classes. As shown in Figure ??, contrast and sharpness decrease with increasing pollution severity, while the dark channel prior increases. These trends are consistent with atmospheric scattering theory.

### 5.5 Grad-CAM Interpretability

Figure ?? presents Grad-CAM visualizations for sample test images. The model consistently attends to haze-heavy regions of the sky and building outlines, confirming that decisions align with haze-relevant image regions rather than spurious background features.

### 5.6 Calibration and Reliability

Temperature scaling improved probability calibration. Figure ?? shows the reliability diagram of Physics v2 after calibration, with an Expected Calibration Error (ECE) of 0.071. This demonstrates alignment between predicted confidence and observed accuracy.

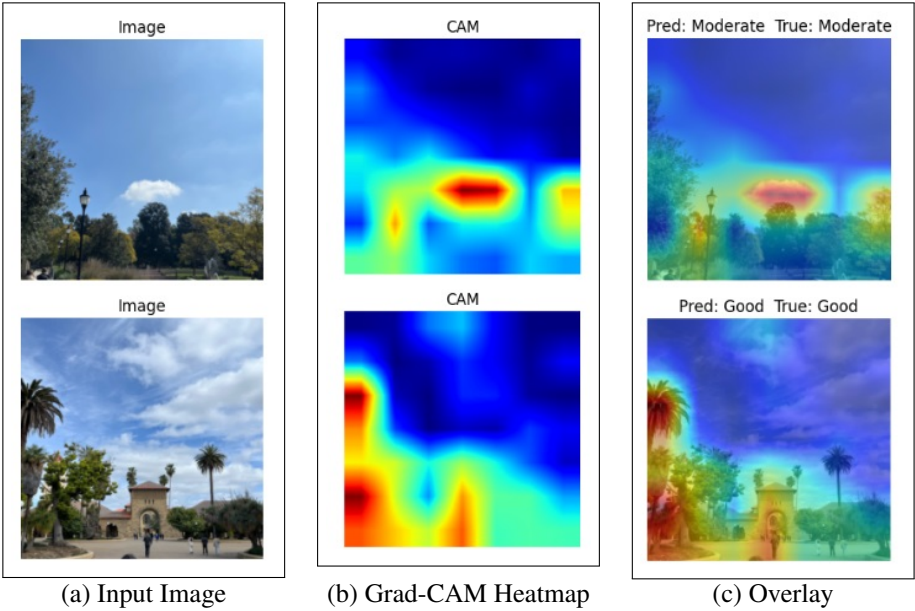


Figure 8: Grad-CAM visualizations: the model focuses on haze-dense regions, providing interpretability in smog-level classification.

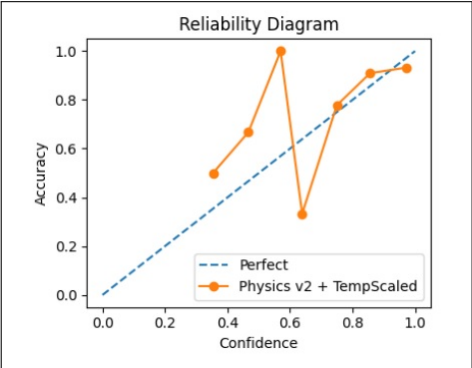


Figure 9: Reliability diagram for Physics v2 + temperature scaling. The model achieves ECE=0.071, indicating well-calibrated confidence.

Model	Params (M)	File Size (MB)	CPU Latency (ms)	FPS
Baseline (v3-small)	1.52	6.23	9.16	109.2
Physics v2 (v3-small)	1.52	6.23	8.99	111.2

Table 1: Efficiency comparison between baseline and Physics v2 models. Both are lightweight and real-time on CPU hardware.

Model	Test Accuracy	Macro-F1
Baseline (TTA)	0.811	0.756
Physics v2 (TTA)	0.851	0.804
Physics v2 + TTA + Temp Scaling	<b>0.865</b>	<b>0.835</b>
Ensemble (equal weights, TTA)	0.824	0.788

Table 2: Final results summary on SAPID test set. Physics v2 with calibration achieves the best accuracy and macro-F1.

### 5.7 Efficiency Analysis

Table ?? compares efficiency metrics. Both baseline and physics-informed models remain lightweight, with ~1.5M parameters, ~6 MB file size, and CPU inference speeds exceeding 100 FPS. Physics-informed modifications incur negligible additional cost.

### 5.8 Final Results Summary

Table ?? reports the final comparison of baseline, physics-informed, and calibrated models. The Physics v2 model with TTA and temperature scaling achieved the best performance with 86.5% accuracy and 0.835 macro-F1.

## 6 Conclusion and Future Work

This study presented a lightweight, physics-informed, and calibrated vision framework for smog-level classification from street-level imagery. Using MobileNetV3-Small as the backbone, the approach demonstrated that integrating atmospheric scattering priors—specifically contrast, sharpness, and dark channel statistics—into the loss function improved recognition of minority air quality classes in the imbalanced SAPID dataset. The final Physics v2 model, combined with test-time augmentation and temperature scaling, achieved 86.5% test accuracy and a macro-F1 score of 0.835, surpassing the baseline MobileNetV3-Small performance. Importantly, these gains were obtained without compromising computational efficiency, with both baseline and physics-informed models operating in real time on CPU hardware at over 100 FPS. Interpretability was further supported by Grad-CAM analyses and monotonic physics-feature trends, while calibration reduced Expected Calibration Error to 0.071, ensuring trustworthy confidence estimates.

The findings underscore three key contributions to vision-based air quality estimation: (1) the feasibility of CPU-deployable lightweight models for real-world environmental monitoring, (2) the value of embedding physics-informed constraints to improve robustness under data imbalance, and (3) the necessity of calibration for reliability in decision-critical tasks.

Future Research Directions

While the results are promising, several avenues remain open for further research:

- **Domain Adaptation:** Extending the model to address dataset shift across geographic regions, seasons, and camera devices through source-free or test-time adaptation techniques [? ? ].
- **Multimodal Fusion:** Incorporating meteorological and satellite data alongside street-level images to enhance generalization and reduce ambiguity under variable lighting and weather conditions.
- **Semi-supervised and Weakly-supervised Learning:** Leveraging unlabeled data or noisy labels to overcome scarcity in high-pollution categories.
- **Edge Deployment Studies:** Evaluating the framework on embedded platforms such as smartphones, Raspberry Pi, and Jetson Nano to validate practical scalability.
- **Explainability Beyond Grad-CAM:** Employing advanced attribution methods and physics-driven interpretability frameworks to further enhance transparency and trust in AI-assisted environmental monitoring.

By bridging efficiency, interpretability, and calibration, the proposed approach contributes to the growing literature on vision-based environmental intelligence and provides a foundation for scalable, low-cost air quality monitoring in real-world deployments.