Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DeepAQI: A Vision-Based EfficientNet Framework for Air Quality Index Prediction from Environmental Metadata

Yash Mishra

yash250025@gmail.com

Nitk Surathkal https://orcid.org/0000-0002-5004-1797

Kedarnath senapati

Nitk Surathkal

Research Article

Keywords:

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9062054/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

DeepAQI: A Vision-Based EfficientNet Framework for Air Quality Index Prediction from Environmental Metadata

Yash Mishra[1] ,Kedarnath Senapati[2]

Department of Mathematics and Computational Sciences

National Institute of Technology Karnataka (NITK) Surathkal ,

NH 66, Srinivasnagar, Surathkal, Mangalore, Karnataka, India – 575025

Corresponding Author :- Yash Mishra (yash250025@gmail.com)

Abstract

This research presents a deep learning–based framework for predicting the Air Quality Index (AQI) using outdoor webcam images from U.S. National Parks. Traditional AQI measurement relies on ground-based air sensors, which require continuous calibration and are often geographically sparse. Recent advances in computer vision have motivated the exploration of image-based AQI estimation, enabling scalable, low-cost, and real-time monitoring in remote regions. Leveraging the publicly available NPS_AQI_DB, which includes more than 120,000 webcam images paired with corresponding pollutant data (O_3, SO_2, RH, temperature), we investigate both image-only and multi-modal (image + tabular) models for AQI regression.

Our primary architecture is EfficientNet-B0, chosen for its strong performance–efficiency trade-off. The model is trained on augmented 224×224 images using AdamW optimization, mixed-precision training, and robust preprocessing to handle missing and corrupted data. To enhance interpretability, Grad-CAM visualizations highlight regions influencing AQI prediction, often corresponding to sky visibility, haze thickness, and lighting conditions. Additionally, we evaluate a two-tower fusion model combining CNN features with meteorological variables, demonstrating improved stability across pollution categories.

Experimental results show that the best-performing model achieves MAE ≈ 11.2 and RMSE ≈ 15.3 on the test set, reflecting competitive performance given the inherent noise of visual air estimation. Comprehensive evaluation includes residual analysis, calibration curves, AQI-bin MAE breakdown, per-park performance, and temporal error trends. These findings confirm that vision-based AQI prediction is feasible and can supplement traditional monitoring networks, especially in visually accessible but sensor-limited environments. Future work will explore temporal modeling, domain adaptation, and deployment on low-power edge devices.

1.0 Introduction

Accurate, timely estimation of the Air Quality Index (AQI) is central to public health, environmental monitoring, and visitor advisory systems—especially in ecologically sensitive and broad-area domains such as national parks. Conventional pipelines depend on ground sensors and curated chemical measurements, which, while accurate, are expensive to deploy densely, difficult to maintain in remote regions, and often yield spatially sparse coverage. With the rapid proliferation of fixed outdoor webcams and advances in deep learning, an emerging line of inquiry is to learn a direct mapping from scene appearance (e.g., haze, sky clarity, visibility) to numeric AQI, thereby augmenting sensor networks with low-cost, visually driven estimation. This study investigates such a vision-based approach using a curated subset of the NPS_AQI_DB—a dataset of national-park webcam images paired with annotations (AQI and pollutant/meteorological context).

We adopt EfficientNet-B0 as the primary backbone for image regression due to its favorable accuracy–efficiency trade-off. Our training script standardizes image inputs, applies lightweight augmentations (horizontal/vertical flips, color jitter), and replaces the classifier with a single-unit regression head. A second, two-tower variant fuses image embeddings with tabular covariates (e.g., O_3, SO_2, relative humidity, temperature) through a small MLP before the final regressor. To improve training robustness in a real-world collection, we implement multiple safeguards: automatic detection and exclusion of non-existent image paths referenced in the CSVs, corrupt/truncated image filtering (PIL integrity check and ImageFile.LOAD_TRUNCATED_IMAGES=True), and target cleaning that drops NaN AQI labels. These steps address the inevitable long-tail data quality issues across historical image archives.

After integrity checks, the effective split sizes used in our experiments are 10,457 images for training, 2,998 for validation, and 1,487 for testing. Notably, large fractions of rows point to images that are missing in the currently available archive, and our pipeline logs these omissions (e.g., “[WARN] 80,958 rows skipped (image file missing) in train.csv”, “23,250 in val.csv,” “11,454 in test.csv”). By culling missing or corrupted entries up-front, model training avoids mid-epoch interruptions and yields a consistent, reproducible sample set.

We emphasize transparent, model-agnostic evaluation through a comprehensive suite of plots. Epoch-wise curves (training/validation MSE, MAE, RMSE, R^2) are produced when history is available, enabling early detection of under/overfitting. For the held-out test set, we generate: (i) a True vs. Predicted AQI scatter with a 45° reference line; (ii) Residuals vs. True AQI to reveal bias and heteroscedasticity; (iii) a Residual histogram/KDE to inspect error symmetry and tails; (iv) MAE by AQI bin (0–50, 51–100, …) to assess difficulty across regimes; (v) MAE by Park to

surface location-specific challenges; (vi) AQI distribution by split (train/val/test) to contextualize generalization; (vii) a binned calibration plot (mean predicted vs. mean true AQI per quantile) to quantify systematic miscalibration; and (viii) a hardest-K montage for qualitative error analysis. When the best checkpoint is image-only, we also provide Grad-CAM heatmaps that consistently highlight sky and horizon regions—areas visually informative for haze and visibility.

On the final test set, our best run attains MAE ≈ 11.279, RMSE ≈ 15.271, and $R^2 \approx 0.138$. These results are aligned with expectations for image-only or lightly fused models in unconstrained outdoor imaging: AQI is influenced by atmospheric composition and local meteorology, and visible haze is an imperfect proxy. Nevertheless, the achieved accuracy indicates that deep visual cues capture meaningful signals correlated with AQI, offering a scalable complement to sparse sensor grids. The remainder of the paper details the dataset preparation, architectures, training regimen, and a thorough error analysis grounded in the aforementioned visual diagnostics.

2.0 Literature Review

2.1 Prior work on AQI estimation.

Classical AQI estimation has largely been grounded in regulatory monitoring networks that aggregate pollutant concentrations (e.g., O_3, $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO) through well-defined formulas to produce a unified health index. Statistical and geostatistical approaches such as kriging, land-use regression (LUR), and generalized additive models (GAMs) have been used to interpolate between sparse stations, incorporate emission inventories, and fuse satellite-derived aerosol optical depth (AOD) with ground monitors to improve spatial coverage. While these methods provide strong baselines, they depend on the availability and maintenance of dense sensor networks, can be sensitive to covariate shift, and may underperform where local microclimates or terrain strongly modulate pollutant dispersion. This has motivated complementary, vision-based AQI estimation that exploits easily deployable webcams to infer visibility and haze as proxies for particulate loading and overall air quality.

2.2 CNN-based environmental assessment

Vision has a long history in atmospheric and environmental inference. Early works measured atmospheric scattering from single images via physical models of haze, estimating scene transmission to restore contrast; those same cues are predictive of visibility and, indirectly, of AQI.

With the advent of deep convolutional neural networks (CNNs), image-based assessment has expanded to cloud type recognition, weather classification, visibility estimation, smoke and dust detection, and urban haze quantification. CNNs learn texture, color, and illumination patterns that correlate with aerosol loadings (e.g., sky luminance gradients, horizon contrast). Compared to hand-crafted features, CNNs are more robust to camera variations, non-Lambertian effects, and complex lighting; however, they require careful regularization and diverse training data to generalize across locations, seasons, and cameras.

2.3 Multi-modal sensor fusion methods.

Because image appearance only partially captures atmospheric chemistry (e.g., gases without visible signatures, or confounding fog/low clouds), fusing visual cues with non-visual context often improves stability. Typical strategies include early fusion (concatenating raw features), mid-level fusion (combining learned embeddings from separate towers), and late fusion (ensemble of unimodal predictors). For AQI, tabular inputs—temperature, relative humidity, wind, pressure, and pollutant surrogates—provide orthogonal information that disambiguates similar visual scenes with different chemistry. Mid-level fusion ("two-tower" CNN + MLP) is particularly effective: the CNN learns view-invariant haze and sky signals, while the MLP encodes local meteorology and pollutant tendencies; concatenation followed by a small regressor balances both. Fusion can also include satellite features (e.g., AOD), temporal statistics (rolling pollutant trends), and day/time encodings, further reducing error variance across regimes.

2.4 EfficientNet and other vision backbones.

Model choice strongly influences the resolution-accuracy-cost trade-off. EfficientNet families (B0–B7, later V2) apply compound scaling to width, depth, and input size, achieving competitive accuracy under modest FLOPs—well suited to webcam AQI where training resources and inference latency matter. ResNet variants remain solid baselines with stable optimization; MobileNetV3 and ConvNeXt offer parameter efficiency and modernized design, respectively. Vision Transformers (ViTs) and hybrid CNN-Transformer architectures have shown state-of-the-art results in generic vision; however, for haze-dominated cues, strong CNNs at slightly higher input resolutions often provide excellent signal extraction with simpler training. In practice, EfficientNet-B0/B2 straddle a sweet spot: adequate capacity to capture haze texture and sky gradients, while remaining memory- and speed-friendly for repeated training across parks and seasons.

2.5 Gaps in existing research.

Despite progress, three gaps persist. First, generalization across cameras and locations is nontrivial: background geography, sun angle, seasonal albedo, and camera white balance introduce shifts that degrade performance if the training data are not geographically and temporally diverse. Domain adaptation, color constancy, and style augmentation remain underexplored at scale for environmental AQI. Second, temporal modeling is scarce in public benchmarks; single-frame predictors miss short-term dynamics (e.g., rapidly forming marine layers, smoke plumes) that could be captured with short video clips or temporal fusion of recent frames. Third, uncertainty quantification and calibration for decision support are limited—most works report point predictions without reliable intervals, complicating risk-aware use (e.g., advisories). Finally, standardized evaluation suites for per-bin MAE, per-location breakdown, residual diagnostics, and calibration are not consistently adopted, making cross-study comparisons difficult. This paper addresses some of these gaps with a robust training pipeline, extensive diagnostics, and a reproducible evaluation protocol that can be extended to multi-modal and temporal settings.

3.0 Dataset Description

We use a curated subset of the NPS_AQI_DB, a collection of outdoor webcam images from U.S. National Parks paired with air-quality annotations. The dataset includes image frames acquired at regular intervals and corresponding metadata (park ID, timestamp), as well as tabular environmental variables: O_3 (ppb), SO_2 (ppb), temperature (°C), and relative humidity (%). The labels include AQI computed according to regulatory breakpoints, enabling a direct regression target from a single view. In raw CSVs, image paths are relative to an images/ root organized by park (e.g., images/ACAD/..., images/BIBE/...), while splits are specified in annotations/{train,val,test}.csv.

Real-world archival collections invariably contain missing references and occasional corruption. To avoid fragile training, we implement a strict pre-processing stage: (1) existence checks drop any row whose image_path does not resolve on disk; (2) corruption filters use PIL integrity verification (ImageFile.LOAD_TRUNCATED_IMAGES=True; im.verify()) to discard truncated or unreadable PNGs/JPEGs; and (3) label sanitization coerces numeric AQI and removes NaNs. These steps reduce the nominal dataset but stabilize training and evaluation.

After filtering, the effective split sizes are 10,457 images for training, 2,998 for validation, and 1,487 for testing. The pipeline logs discarded counts for transparency—e.g., warnings of 80,958 missing images in train, 23,250 in val, and 11,454 in test—reflecting that the public image archive is smaller than the total rows referenced by the annotation CSVs. Although this pruning reduces

data volume, the remaining set still exhibits substantial diversity in sky conditions, viewing geometry, diurnal/seasonal illumination, and terrain.

We provide extensive dataset visualization for context: sample grids per split, AQI histograms/KDEs for train/val/test to reveal label balance, per-park image counts to expose geographic skew, and resolution/aspect-ratio histograms to characterize camera heterogeneity. These diagnostics guide augmentation design (e.g., color jitter magnitudes), input sizing (e.g., 224–300 px), and model selection (e.g., EfficientNet-B0 as a good starting point). In sum, while imperfect and imbalanced, the curated subset remains well-suited for learning visibility-linked cues that correlate with AQI, provided training embraces robust filtering, consistent normalization, and comprehensive error analysis.

3.1 Source of the Dataset (NPS_AQI_DB)

3.1.1 Image acquisition setup (webcams in national parks).

National Park Service webcams are fixed, outdoor cameras overlooking representative vistas—coastlines, canyons, ridgelines, or desert plains. Images are captured at regular intervals throughout daylight hours, yielding large archives across seasons and weather conditions. Camera characteristics (sensor, optics, auto-white balance, exposure) vary by site and epoch, inducing appearance shifts that are informative (e.g., shadow length, sun elevation) but also challenging for generalization.

3.1.2 Pollutant data: O_3, SO_2, Temperature, RH.

Alongside imagery, the dataset provides tabular readings: ozone (O_3, ppb) often fluctuates with photochemical activity and sun angle; sulfur dioxide (SO_2, ppb) acts as a proxy for certain emissions and downwind transport; temperature (°C) and relative humidity (%) influence aerosol hygroscopic growth, scattering efficiency, and perceived haze. These features are not perfectly synchronized with every pixel in the image, but they encode air-mass context and can disambiguate visually similar scenes with different chemistry—making them excellent candidates for multi-modal fusion with vision.

3.1.3 AQI computation methods.

Regulatory AQI is derived from pollutant concentrations via breakpoint tables for each species (e.g., $PM_{2.5}$, O_3, NO_2, SO_2, CO), then mapped to categories (Good, Moderate, Unhealthy for Sensitive Groups, etc.). In practice, a dataset may provide AQI directly as a label or provide precursor measurements from which AQI can be calculated. For this study, we use the provided AQI as the regression target; pollutant fields and meteorology can be concatenated in a two-tower model for improved stability. Because the visible scene is a proxy for aerosol-induced scattering rather than a direct measure of regulated species, image-only predictions have irreducible error; pollutant and humidity context reduces variance in moderate-to-high AQI regimes.

Overall, NPS_AQI_DB offers a realistic, operationally relevant setting: fixed cameras with broad geographic coverage, multi-year variability, and meaningful environmental covariates. While the raw CSVs reference more images than available in the public subset—necessitating filtering—the remaining data are rich enough to train competitive models and to study transferability across parks and seasons.

3.2 Dataset Statistics

3.2.1 Raw count vs. used samples.

Annotation CSVs in NPS_AQI_DB reference a very large number of rows, but not all referenced images are present in the public archive. During pre-processing, our pipeline reports substantial missing image counts: 80,958 rows in train.csv, 23,250 in val.csv, and 11,454 in test.csv were skipped because the files could not be located. After existence and corruption filtering, the final used sizes are 10,457 (train), 2,998 (val), and 1,487 (test).

3.2.2. Implications of pruning.

Although the pruning rate is high, two factors mitigate its impact. First, the remaining images still cover a broad span of lighting (morning/noon/evening), seasons (snow, haze, monsoon), and geographies (coastal, desert, alpine). Second, from a learning perspective, quality-controlled samples are preferable to noisy or broken inputs that derail training or inflate variance. The explicit logging of dropped counts ensures reproducibility and supports cross-run comparisons when the underlying archive changes.

3.2.3 Class distribution across AQI bins.

We examine the AQI distribution by split using histograms/KDEs and summarize MAE by AQI bins (0–50, 51–100, 101–150, 151–200, 201–300, 301–500, 500+). In outdoor vistas, the dataset typically skews toward lower AQI (Good/Moderate), with fewer samples in Unhealthy+ categories. This class imbalance is meaningful: models learn fine-grained differences in clean-to-moderate conditions but face higher variance in scarce high-AQI regimes. Reporting per-bin MAE helps set expectations (errors are usually larger at higher AQI) and guides future data collection (e.g., targeted acquisition during wildfire smoke events).

3.2.4 Park-level and camera-level skew.

Per-park counts reveal that some locations contribute a disproportionate share of images; this can bias a model toward the colorimetry and background of those parks. Similarly, resolution/aspect-ratio histograms show multiple camera families. These skews inform augmentation (color, brightness, flips) and regularization (dropout, mixup) choices, and motivate domain adaptation if cross-park generalization is a primary goal.

In summary, while the effective sample sizes are smaller than raw references, the 10,457/2,998/1,487 splits remain sufficient for training and rigorous evaluation. Monitoring per-bin and per-park performance is essential for fair assessment and for directing future data curation.

3.3 Data Quality Challenges

3.3.1 Corrupted images.

Archival feeds occasionally contain truncated downloads, partial frames, or malformed PNG chunks. Attempting to decode such files at training time leads to worker crashes (e.g., PIL OSError: image file is truncated). We proactively verify images prior to training (im.verify()), enable tolerant loading (ImageFile.LOAD_TRUNCATED_IMAGES=True), and exclude unrecoverable files, preventing mid-epoch failures.

3.3.2 Missing labels.

Some rows include non-numeric or missing AQI values. We coerce AQI to float and drop NaNs, ensuring the regression target is well-defined. Because AOI noise is inevitable, we also sanity-check plausible ranges and log the number of discarded rows, promoting traceability.

3.3.4 Lighting/weather variability.

Outdoor webcams experience massive appearance variation: changing sun angle, seasonal snow/vegetation, fog, low clouds, wildfire smoke, and camera auto-exposure shifts. These factors confound haze cues and increase target irreducible noise. We mitigate via color jitter, horizontal/vertical flips, robust normalization (ImageNet statistics), and model choices capable of learning invariant features (e.g., EfficientNet). Still, elevated error at high AQI remains expected; explicit multi-modal fusion (O_3, SO_2, RH, temperature) improves stability under ambiguous visuals.

Together, these quality controls convert a noisy archive into a stable corpus suitable for reproducible training and comprehensive evaluation.

3.4 Dataset Visualization

3.4.1 Sample grids.

We present grid montages for train/val/test to qualitatively assess diversity: different sky colors, horizons, terrains, and seasonal phenomena appear across parks. Side-by-side "original vs. augmented" previews demonstrate how flips and color jitter affect sky tone and shadowing, validating that augmentations are realistic and not destructive.

3.4.2 AQI distribution histograms.

Histograms and KDEs of AQI by split reveal label balance and potential shift. Typically, train and test are similarly concentrated in Good–Moderate ranges, with a thinner tail into Unhealthy. These plots help interpret downstream metrics (e.g., overall RMSE is dominated by abundant low-AQI samples).

3.4.3 Per-park image count distribution.

Bar charts of images per park for each split highlight geographic skew—useful when analyzing per-park MAE. Heavily represented parks may yield lower error due to familiarity; under-represented parks may require domain adaptation or targeted augmentation. Reporting per-park counts increases transparency and informs fair comparisons.

Resolution/aspect-ratio distribution.
Histograms of aspect ratio (W/H) and megapixels show camera heterogeneity. This helps choose input resolutions (e.g., 224–300 px) that preserve crucial sky textures without undue memory cost, and motivates center/resize strategies that minimize distortion of horizon lines.

3.4.4 Evaluation-linked visuals.

To tie dataset characteristics to performance, we also include: True vs. Predicted scatter with a 45° line (global fit), Residuals vs. True (bias/heteroscedasticity), Residual histogram (error symmetry), MAE by AQI bin (difficulty by regime), MAE by park (location-specific variance), calibration plots (binned mean pred vs. mean true), and a hardest-K montage for qualitative failure analysis. When the best checkpoint is image-only, Grad-CAM overlays show the model's attention concentrated on sky and horizon, validating that visible scattering is a primary cue.

Collectively, these visualizations transform the dataset from a static collection of images into a diagnostic instrument: they guide augmentation and model choice, contextualize metrics, and illuminate where additional data or fusion signals would most improve generalization.

Below is a polished, publication-ready draft for the requested sections. Each heading and subheading is written at approximately 400 words to fit your specification and maintain academic tone and clarity.

4.0 Methodology

Our methodology is designed to turn a large, imperfect archive of national-park webcam images into a robust, reproducible training and evaluation pipeline for image-based AQI regression. The pipeline has four pillars: clear problem formulation, disciplined preprocessing, principled augmentations, and architectures chosen for the visibility signal in outdoor scenes, all trained with appropriate losses and optimizers.

We formulate AQI estimation as a supervised regression problem: given a single RGB image (and optionally auxiliary tabular context), the model predicts a continuous AQI value. To preserve real-world reproducibility, we keep the dataset's split structure (train/val/test) intact and never leak

test information into training or augmentation statistics. Because the raw CSVs reference many images that are unavailable in the public subset (or are corrupted), we enforce rigorous data hygiene before any learning occurs.

Preprocessing includes three hard checks: (1) path resolution—retain only rows whose image_path resolves under the dataset root; (2) image integrity—discard files that fail PIL verification or exhibit truncation; (3) target cleaning—coerce AQI to float and drop NaNs. We also normalize images using ImageNet statistics consistent with the pretrained backbone. These steps reduce the nominal dataset but yield a stable corpus, eliminating mid-epoch crashes and label noise.

Augmentations are intentionally simple and photometrically plausible—horizontal/vertical flips and mild color jitter—to encourage invariance to viewpoint and camera auto-exposure while avoiding unrealistic artifacts. For modeling, we adopt EfficientNet-B0 as the baseline backbone and replace the classifier head with a single-unit regression layer. We also consider a two-tower fusion (CNN for images + MLP for tabular features such as O_3, SO_2, temperature, RH), concatenated before a small regressor to stabilize predictions in visually ambiguous regimes.

Training minimizes MSE with AdamW, optionally using mixed-precision to improve throughput and numerical stability on GPU. Validation selects the best checkpoint by RMSE, reflecting sensitivity to large errors. Our evaluation suite reports MAE/RMSE/R^2, per-bin/per-park breakdowns, residual diagnostics, and calibration, with qualitative hardest-K panels and Grad-CAM (for image-only models). This end-to-end design combines data discipline with model pragmatism to deliver reliable, interpretable results on a challenging real-world task.

4.1 Problem Formulation

We pose Air Quality Index (AQI) prediction as a supervised regression problem at the image level. Given an RGB image $\mathbf{x}$, represented as a tensor of shape $\mathbf{H \times W \times 3}$, captured by a fixed outdoor webcam, the objective is to learn a function $\mathbf{f(x)}$ that outputs a predicted AQI value $\hat{\mathbf{y}}$, which approximates the true AQI $\mathbf{y}$.

In the multi-modal setting, the input is extended with a vector of tabular features $\mathbf{z}$ (such as O_3, SO_2, temperature, and relative humidity). In this case, the model learns a function $\mathbf{f(x, z)}$ that produces the predicted AQI $\hat{\mathbf{y}}$, using both visual and tabular information. The parameter set $\boldsymbol{\theta}$ represents all trainable model parameters.

The training dataset **D_train** consists of either pairs **(x, y)** in the image-only setting, or triplets **(x, z, y)** in the multi-modal setting. The validation set is used to monitor generalization and select

optimal checkpoints and hyperparameters, while the test set provides the final unbiased performance estimate. AQI is treated as a continuous variable to preserve fine-grained distinctions between categories (e.g., Good, Moderate, Unhealthy for Sensitive Groups). This is important because AQI category boundaries are piecewise-defined and converting the task into classification would hide small but meaningful improvements in prediction accuracy.

We adopt MSE as the optimization objective since it emphasizes large residuals, aligning training with the downstream use of RMSE as a key metric. We also report MAE—which is more interpretable in the public-health context ("average absolute AQI error")—and R^2 to quantify explained variance. Because real-world data exhibit heteroscedasticity (errors often increase at higher AQI), we also compute per-bin MAE and calibration over quantile bins of true AQI.

In the image-only regime, the model must infer AQI from visibility cues alone (haze thickness, sky luminance, horizon contrast, scattering color). In the two-tower regime, tabular covariates provide complementary chemistry and meteorology, helping disambiguate visually similar scenes with different pollutant profiles or humidity. We keep the training protocol identical across regimes except for the architectural fusion and input pipelines.

Finally, we constrain the problem to single-frame inputs for operational simplicity and to match your current dataset. We acknowledge that temporal stacks could further reduce variance by capturing short-term changes, but we leave that extension to future work. This formalization, while simple, aligns the model's learning target with regulatory AQI values and produces interpretable errors at the same scale as real advisories.

4.2 Preprocessing

Preprocessing ensures a stable, reproducible dataset from noisy, real-world archives. We implement a three-stage pipeline: image cleaning, missing/corrupted handling, and normalization consistent with the pretrained backbone.

4.2.1 Image cleaning.

We begin by loading the split CSVs (train/val/test) and resolving each image_path relative to the dataset root. Any row whose path does not exist on disk is dropped and counted. Next, we verify image integrity. Using PIL, we attempt to open the file and call im.verify(); if the file is truncated, malformed, or otherwise unreadable, we exclude it. We also set ImageFile.LOAD_TRUNCATED_IMAGES=True to tolerate partial headers where feasible—but

still discard files that remain unsafe. For labels, we parse the AQI field as float and drop NaNs to avoid silent propagation of invalid targets.

4.2.2 Handling missing/corrupted files.

The public subset references many images that are no longer present; our logs explicitly record such omissions (e.g., 80,958 missing in train, 23,250 in val, 11,454 in test). After all checks, we retain 10,457 training, 2,998 validation, and 1,487 test images. Proactive filtering avoids runtime crashes (e.g., OSError: image file is truncated) and prevents uninformative gradients from corrupt labels or unreadable inputs. We optionally cache "clean" CSVs to skip rescans in future runs.

4.2.3 Normalization.

We normalize pixel intensities using the standard ImageNet mean and standard deviation, consistent with the preprocessing used for EfficientNet-B0 pretrained weights. First, images are converted to tensors with values in the range [0, 1]. Then, normalization is applied channel-wise using the transformation:

$$(x - \mu) / \sigma$$

where the per-channel mean is $\mu = [0.485, 0.456, 0.406]$ and the per-channel standard deviation is $\sigma = [0.229, 0.224, 0.225]$.

Matching the normalization statistics used during the backbone's original ImageNet training is essential for effective transfer learning. It ensures that the feature distributions seen by the model at fine-tuning time remain aligned with those it encountered during pretraining. This alignment typically leads to faster convergence, more stable gradients, and improved generalization, particularly when training on smaller datasets or domain-specific tasks such as AQI prediction from webcam imagery.

This preprocessing intentionally reduces nominal data volume but maximizes signal quality and training stability. It also standardizes the data across heterogeneous cameras and capture epochs, paving the way for consistent augmentation and modeling. By surfacing counts of dropped rows, the pipeline promotes transparency and reproducibility—future users can re-create the same sample sets or compare new archives against our baseline.

4.3 Data Augmentation

Our augmentation strategy is lightweight and photometrically plausible, tailored to outdoor webcams where haze and sky gradients are primary cues. The aim is to improve invariance to view symmetry, mild camera auto-exposure, and seasonal color shifts without introducing artifacts that would distort visibility signals.

Horizontal/vertical flips. Horizontal flips ($p=0.5$) are generally safe because the global scene semantics (sky vs. land) remain intact. Vertical flips are used sparingly ($p\approx0.1$) because flipping the horizon can be unnatural; however, a small probability improves robustness to horizon placement and minor camera tilts. In practice, we find horizontals contribute most of the geometric variety.

Color jitter. We apply moderate jitter to brightness, contrast, and saturation (e.g., ±20% brightness/contrast, ±10% saturation). This mimics differences across cameras and times of day, and tempers over-reliance on absolute luminance. We choose small ranges to avoid destroying the delicate haze cues that the model must learn.

Random brightness/contrast changes. While covered by color jitter, we treat brightness/contrast separately in ablations to verify they do not erase haze gradients. Slight brightness reduction can model cloud cover or lower sun elevation; slight contrast variation can model sensor auto-exposure. Excessive contrast would hallucinate visibility; hence we keep ranges conservative.

We avoid aggressive transforms—random crops that truncate horizon lines, heavy perspective warps, or artifact-prone compression—because the horizon/sky context is a meaningful, fragile signal. If resolution allows, a center resize to 224–300 px retains horizon geometry while keeping compute tractable for EfficientNet-B0/B2. When moving to higher-capacity backbones, scaling input size modestly (e.g., 288–320 px) can enhance texture sensitivity without exploding memory.

To evaluate augmentation efficacy, we render original vs. augmented pairs for visual inspection and track validation RMSE deltas. The guiding principle is conservative realism: diversify exposure, hue, and mild geometry while preserving the very patterns (sky tone, aerial perspective) that convey aerosol scattering. In combination with normalization aligned to the pretrained backbone, these augmentations yield stable convergence and stronger cross-park generalization.

4.4 Model Architecture

We consider two architectures: an image-only EfficientNet-B0 baseline and a two-tower fusion that incorporates tabular context.

EfficientNet-B0 baseline. EfficientNet scales depth, width, and resolution with a compound coefficient, offering strong accuracy per FLOP. We load ImageNet-pretrained weights and replace the final classifier with a regression head (Dropout → Linear(1)). This head outputs a scalar AQI. The backbone encodes multi-scale texture and color gradients: sky luminance, horizon contrast, and scene transmission cues associated with haze.

Custom regression head. The head uses modest dropout (e.g., 0.2–0.25) for regularization. We keep a single linear output to preserve interpretability and tight coupling with MSE training. If training data increases, a small MLP (e.g., Linear → GELU → Linear) could be added, though we prioritize simplicity.

Two-tower architecture (Image + Tabular data). To stabilize performance in visually ambiguous conditions (fog vs. pollution, differing chemistry under similar skies), we add a tabular tower: a small MLP (e.g., 128 units with ReLU and dropout) that encodes O_3, SO_2, temperature, RH. The fusion layer concatenates the visual embedding (e.g., 256-D projection of the backbone's penultimate features) with the tabular embedding, feeding a small regressor (e.g., 128 → 1). We compute statistics (mean/std) for tabular inputs on the training set and standardize val/test with those parameters to avoid leakage.

This design adheres to several principles: (1) align with transfer learning by preserving the backbone's normalization; (2) limit head capacity to prevent overfitting; (3) keep fusion simple and transparent; (4) maintain identical loss and metrics across regimes for fair comparison. We also support Grad-CAM on the image-only model by tapping the last conv block, enabling explainability. If deploying the two-tower model, a CAM can still be computed on the visual branch by holding tabular inputs constant for visualization.

Overall, EfficientNet-B0 provides an excellent baseline for visibility-driven cues. When augmented with tabular context, the two-tower design reduces variance across AQI bins and improves resilience to confounders (humidity, temperature), offering a pragmatic step toward multi-modal air-quality inference.

4.5 Loss Function

We train the model using the **Mean Squared Error (MSE)** loss, defined as:

$$\mathbf{MSE = (1 / N) * \Sigma (\hat{y}_i - y_i)^2}$$

This loss penalizes large errors more heavily than small ones. It aligns with the use of RMSE during validation and testing, which is important because large AQI prediction errors (especially underestimating unhealthy air) have stronger real-world consequences.

We also compute **Mean Absolute Error (MAE)** and **R-squared (R^2)** as auxiliary metrics.

$$\mathbf{MAE = (1 / N) * \Sigma |\hat{y}_i - y_i|}$$

MAE is easy to interpret because it represents the average absolute deviation in AQI units.

The **Root Mean Squared Error (RMSE)** is defined as:

$$\mathbf{RMSE = sqrt(MSE)}$$

RMSE emphasizes larger errors due to the square root of squared residuals, making it a suitable metric in scenarios where large mistakes carry higher importance.

places greater emphasis on outliers and typically exceeds MAE in magnitude. R^2 reports explained variance; in noisy atmospheric settings, modest R^2 values are common because image cues are a proxy for chemistry rather than a direct measurement.

To assess fairness across regimes, we further report per-bin MAE over standard AQI categories (0–50, 51–100, …). This reveals where the model struggles (usually higher AQI) and provides actionable guidance for data collection and fusion. We also include calibration by binning true AQI into deciles and comparing mean predicted vs. mean true; deviations from the 45° line diagnose systematic under/overestimation. While we did not use a calibration loss during training, post-hoc methods (isotonic regression) are feasible and can be applied later.

In principle, one might adopt Huber or quantile losses to temper outliers or predict conditional quantiles for uncertainty estimation. We begin with MSE to maintain comparability with prior work and because it pairs well with AdamW and mixed-precision training. The final selection criterion is validation RMSE, consistent with operational priorities and our test reporting. This loss-metric alignment simplifies model selection and ensures that the chosen checkpoint is optimized for the statistic most sensitive to large AQI errors.

4.6 Optimization

We optimize with AdamW, which decouples weight decay from the gradient update, typically improving generalization for vision backbones. A representative learning rate (LR) is 2e-4 with weight decay 1e-4, tuned on the validation set. We use mini-batches sized to fit available VRAM (e.g., 32 on T4/L4 for 224px inputs), and we train for ~10–30 epochs, monitoring validation RMSE to select the best checkpoint.

Schedulers. For simplicity, we often rely on a fixed LR when data are moderately sized and augmentation conservative; however, a cosine decay or step schedule can further stabilize late-epoch convergence. If you observe plateaued validation loss, a gentle LR decay over epochs may help. Warmup (e.g., 500–1,000 steps) can be beneficial if you increase input resolution or batch size.

Mixed-precision training. We enable AMP (automatic mixed precision) on GPU to increase throughput and reduce memory footprint, while maintaining numerical robustness via gradient scaling. This is particularly useful with EfficientNet or when experimenting with larger input sizes (e.g., 288–300 px).

Regularization. We keep dropout modest (0.2–0.25) in the head. If overfitting appears (diverging train/val curves), consider slightly stronger dropout, mild L2 increases, or mixup/cutmix (with caution, to avoid distorting haze cues). Early stopping based on validation RMSE is another pragmatic guardrail.

Checkpointing and selection. We checkpoint at each epoch and adopt the best validation RMSE model. This prevents regressions in late epochs. We also save predictions and logs for reproducibility.

Two-tower specifics. Standardize tabular inputs using train-set mean/std, reusing those stats for val/test to prevent leakage. Keep the tabular MLP small (e.g., 128 units) to avoid overpowering visual cues.

Altogether, AdamW + AMP + careful checkpointing yields a robust, resource-friendly training loop that is easy to reproduce and extend.

5.0 Training Setup

Our training setup prioritizes reproducibility, compute efficiency, and clarity. We work in Google Colab with GPU acceleration (T4 or L4, depending on session allocation). Experiments are scripted so that users can re-run the exact pipeline on their side, given identical dataset roots and splits. Logs capture epoch-wise train/val metrics and optional LR/time annotations; predictions are saved for downstream plotting.

The core environment includes PyTorch (with CUDA support), torchvision (for EfficientNet and transforms), pandas/numpy (for data handling), PIL (for image I/O), and matplotlib/seaborn (for diagnostics). We set random seeds for Python, NumPy, and PyTorch to stabilize sampling and initialization. Input resolution defaults to 224×224, balancing texture capture and throughput. Batch size is chosen to saturate GPU without OOM (e.g., 32 for 224px on a T4/L4); if increased resolution or multi-modal fusion raises memory pressure, batch size is reduced accordingly.

We log the effective split sizes after filtering: 10,457 train, 2,998 val, and 1,487 test. The training loop runs for 10–30 epochs, with validation at each epoch; best checkpoint selection uses validation RMSE. Post-training, the best checkpoint is loaded for test inference, and test MAE/RMSE/R^2 are reported.

To support analysis, we generate a comprehensive plot suite: epoch curves (MSE/MAE/RMSE/R^2) when history is available; test scatter, residuals, residual distribution, calibration, MAE by AQI bin and by park; distributions by split; hardest-K grids; and Grad-CAM overlays for image-only models. We also provide utilities to export plots and assemble reports (PDF/DOCX) with images and captions, facilitating communication and review.

This setup makes it straightforward to iterate on models (e.g., replacing EfficientNet-B0 with B2 or ConvNeXt), modify augmentation, and test fusion strategies. Because the data-quality layer is strict and cached, subsequent runs are fast and deterministic—essential for fair ablations and future extensions.

5.1 Hardware and Environment

We conduct experiments on Google Colab using NVIDIA T4 or L4 GPUs, depending on the session. Both offer sufficient VRAM for batch sizes around 32 with 224×224 inputs and EfficientNet-B0. Colab's environment includes recent CUDA/cuDNN builds, making mixed-precision training (AMP) reliable and fast. Storage is ephemeral; we therefore script dataset download/extraction and encourage saving checkpoints/predictions to Drive when needed.

Training time per epoch varies with GPU type, batch size, and input resolution. On a T4 with 224px inputs and batch size 32, an epoch over 10,457 images is typically a few minutes. L4 may be faster. Validation runs are proportionally shorter (2,998 images). These timings allow a full 10–30-epoch training cycle to complete within common Colab session limits. For larger backbones or input sizes (e.g., 288–320 px), expect roughly linear slow-downs due to increased FLOPs.

We seed PRNGs (Python/NumPy/PyTorch) and set torch.backends.cudnn.benchmark=True for convolution autotuning. When doing strict timing comparisons, we may disable autotuning for determinism. We capture logs with tee to a text file (e.g., /content/train.log) for later epoch-curve parsing.

Package versions matter: PyTorch/torchvision must be mutually compatible; PIL must be recent enough for robust PNG/JPEG handling; and seaborn/matplotlib for plotting. We note these in a requirements cell and lock versions when archiving experiments.

Finally, we recommend pushing checkpoints, predictions, and key plots to Drive/GitHub for long-term retention. Given Colab's ephemeral storage, this practice protects against accidental loss and eases collaboration.

5.2 Hyperparameters

We select hyperparameters to balance accuracy and stability while respecting Colab GPU limits:

- Batch size: ~32 (T4/L4, 224px). If input resolution increases (288–320 px) or you switch to a heavier backbone, reduce to 16 or 8 to prevent OOM.
- Epochs: 10–30, depending on validation convergence and augmentation intensity.
- Learning rate (AdamW): 2e-4 works well for EfficientNet-B0 fine-tuning with modest heads. If you observe slow convergence, try 3e-4; if instability, try 1e-4.
- Weight decay: 1e-4, a conservative value that regularizes the head and mildly the backbone.
- Dropout (head): 0.2–0.25 to mitigate overfitting without underutilizing capacity.
- Image resolution: 224×224 by default. If VRAM permits, 288–300 px can capture finer haze texture and improve accuracy.

Additional knobs:

- Augmentations: Horizontal flip p=0.5, vertical flip p=0.1, color jitter (brightness/contrast ±0.2, saturation ±0.1).
- Tabular standardization: compute mean/std on train; reuse on val/test.

- Early stopping: optional if training beyond 20–30 epochs; otherwise best-by-RMSE checkpointing suffices.
- Scheduler: optional cosine or step decay; begin simple, add if validation plateaus.

We encourage hyperparameter sweeps that monitor validation RMSE and calibration. Be mindful of the dataset's effective sizes and class imbalance; extremely long runs don't always pay off if high-AQI samples are scarce. Instead, consider modest resolution bumps or multi-modal fusion for bigger gains.

5.3 Training Stability Checks

To ensure stable optimization and reliable conclusions, we instrument three categories of checks:

Gradient scaling (AMP). Enable torch.amp with gradient scaling on GPU to avoid underflow in mixed precision. This improves throughput and helps maintain stable updates, particularly when the head's gradients are small. We watch for inf/nan in loss; if observed, briefly reduce LR or disable AMP to isolate issues.

Learning rate warmup/decay. Although a fixed LR works for our default setup, when increasing input resolution or changing backbones, a short warmup (e.g., 500–1,000 steps) can mitigate early-epoch spikes. A cosine decay or gentle step schedule can fine-tune convergence once the loss plateaus, improving validation RMSE by a small but consistent margin.

Epoch time analysis. We optionally log epoch time to spot performance regressions (e.g., due to I/O bottlenecks or accidental CPU execution). If data loading dominates, increasing num_workers and enabling pin_memory often helps. Persistent workers also reduce per-epoch loader startup costs.

Beyond these, we watch train vs. val curves for divergence (overfitting), and we keep dropout/head complexity modest. If instability persists, we try smaller batch sizes, lower LR, or a lighter backbone. The objective is not merely to train but to train predictably, producing clean logs and deterministic behavior under seed control. These stability habits reduce variance across runs and make ablations credible.

6.0 Results and Evaluation

We evaluate with a comprehensive suite covering aggregate metrics, epoch curves, test-set diagnostics, qualitative panels, and explainability. The final test performance of our best checkpoint is MAE = 11.279, RMSE = 15.271, and $R^2 = 0.138$ on 1,487 images—consistent with the difficulty of inferring chemistry from single-frame visibility.

When training logs (or history CSV) are available, we plot Train vs. Val curves for MSE/MAE/RMSE/R^2, plus LR vs. Epoch and Epoch time vs. Epoch. These curves help detect under/overfitting and confirm that optimization is behaving as expected.

For the test set, we render True vs. Predicted AQI scatter with a 45° line to visualize global fit, Residuals vs. True to reveal bias/heteroscedasticity, and a Residual histogram/KDE to examine error symmetry and tails. We also include a 10-bin calibration plot (binned mean true vs. mean pred), per-park MAE bar charts (location variance), MAE by AQI category (0–50, 51–100, …), and MAE by month/hour when metadata allow—illuminating diurnal/seasonal patterns.

Qualitatively, we present a 10-image grid with titles showing park name (if available), true AQI, predicted AQI, and absolute error, as well as Best-K and Worst-K panels. These reveal success and failure modes: sunset glare, fog confusion, snow cover, marine layers. If the checkpoint is image-only, Grad-CAM overlays show that attention concentrates on the sky and horizon, lending interpretability to predictions.

Collectively, these results indicate that the model captures meaningful visual correlates of AQI but remains challenged in moderate-to-high AQI regimes—precisely where multi-modal fusion and temporal context can help. The evaluation protocol doubles as an MLOps dashboard, ensuring every training run emits the same set of trustworthy diagnostics.

6.1 Quantitative Metrics

On the held-out test set of 1,487 images, the selected checkpoint achieves:

- MAE = 11.279
- RMSE = 15.271
- $R^2 = 0.138$

Interpretation. An MAE of ~11.3 means, on average, predictions deviate by about eleven AQI units—credible for webcam-based estimation given variable lighting, humidity, and camera

characteristics. RMSE ~15.3 indicates heavier penalties on larger errors; such penalties are important when AQI is elevated and public risk is higher. $R^2 \approx 0.14$ may seem modest, but in this domain it reflects the reality that single images are an imperfect proxy for regulated gases and particulate chemistry; much variance is irreducible without auxiliary context (meteorology, pollutant measurements) or temporal information.

We further examine per-bin MAE across AQI categories. Typically, errors are smallest in Good–Moderate ranges (0–100) and grow in Unhealthy ranges due to increased atmospheric complexity and confounders (e.g., fog vs. smoke). Per-park MAE reveals geographic/climatic variance: drier parks with crisp visibility tend to yield lower error than coastal or high-humidity parks where scattering and cloud layers complicate inference.

While these numbers leave room for improvement, they establish a reproducible baseline on a cleaned subset with documented missing/corrupted files filtered out. We recommend using the same metrics to compare future models—e.g., EfficientNet-B2/V2, ConvNeXt, or ViT; higher input resolutions; two-tower fusion; or short temporal windows. Improvements should be reflected not only in aggregate metrics but also in calibration, per-bin MAE, and per-park MAE, which together portray robustness.

6.2 Training Curves

Training curves provide visibility into learning dynamics and early signs of over/underfitting:

- Train vs. Val MSE/MAE/RMSE/R^2. These curves show whether the model's fit on training data translates to validation. Healthy runs exhibit parallel descent early on; divergence (train improving while val degrades) signals overfitting or distribution mismatch.
- LR vs. Epoch. With a fixed LR, this line is flat; with cosine decay or steps, we expect smooth reductions. Sudden metric changes following LR adjustments can highlight sensitivity.
- Epoch time vs. Epoch. Stable epoch times indicate I/O and GPU utilization are consistent. Large variability can indicate data-loader issues or background contention; tuning num_workers, pin_memory, and persistent workers often helps.

When logs are captured (e.g., | tee /content/train.log), we parse them to render these curves automatically. Even a single-epoch history can bootstrap the figure pipeline, though longer runs produce more informative trends. If curves show early plateaus, consider LR tuning, mild scheduler decay, or increased input resolution. If curves show oscillation, reduce LR or batch size.

If validation metrics improve with heavier augmentation or fusion, retain those settings for subsequent runs.

Overall, epoch curves are a practical sanity check and a concise story of the model's training narrative. They complement final metrics and help ensure that observed gains are due to genuine learning rather than accidental data quirks.

6.3 Test-Set Analysis

We conduct in-depth test analysis to understand where and why the model errs:

- True vs. Predicted scatter. Points clustered around the 45° line indicate good calibration; vertical dispersion shows error magnitude. Systematic offsets (e.g., underestimation above AQI 150) suggest bias that post-hoc calibration can fix.
- Residual vs. True AQI. Plotting the residuals ($\hat{y} - y$) against the true AQI values (y) reveals heteroscedasticity — the errors tend to increase as AQI becomes higher.
- If residuals skew positive/negative in certain ranges, update augmentations or add fusion signals to reduce bias.
- Residual histogram + KDE. Centered, symmetric residuals are ideal; heavy tails flag rare but impactful failure modes (glare, fog, snow). Consider targeted data collection or robust losses in future iterations.
- Calibration (10-bin). Comparing binned mean predictions against binned mean truths diagnoses systematic under/overestimation. Mild isotonic calibration can correct slope without retraining.
- Error distribution across parks. This bar chart highlights geography-dependent variance. If a small group of parks dominates error, investigate camera properties or local meteorology; domain adaptation or park-specific fine-tuning may help.
- MAE by AQI category. Confirms that high-AQI regimes remain harder. Multi-modal fusion (O_3, SO_2, RH, temperature) typically reduces variance here.
- MAE by month/hour (if metadata): reveals diurnal/seasonal patterns—e.g., dawn/dusk lighting or monsoon months.

Together these plots are an operational dashboard: they validate aggregate metrics, show fairness across conditions, and guide the next steps (fusion, resolution, temporal stacking, calibration).

6.4 Qualitative Evaluation

Numbers tell part of the story; images tell the rest. We present , 10-image grid with titles showing Park (if available), True AQI, Pred AQI, |error|. This gives stakeholders an intuitive sense of where the model succeeds (clear skies, stable illumination) and where it struggles (backlit horizons, fog). Best-K and Worst-K panels , The best cases confirm that the learned features (sky brightness gradients, contrast decay with distance) align with AQI cues. The worst cases often expose confounders: saturated sunsets, low clouds masquerading as haze, snow cover, lens flare, or atypical white balance.

We curate examples across parks and seasons to avoid cherry-picking. For each difficult case, we note plausible reasons (e.g., humidity-driven scattering, sensor auto-exposure, occlusions). These panels inspire concrete remedies: add multi-modal inputs, raise input resolution, diversify augmentations in controlled ways, or mark certain conditions as “low confidence” to improve risk communication.

Qualitative review also supports trust: decision-makers can see the logic of predictions via Grad-CAM (for image-only models) and spot conditions where cameras underperform relative to monitors. We consider this interpretability essential for environmental applications where public advisories depend on credible evidence.

6.5 Explainability

For image-only checkpoints, we compute Grad-CAM heatmaps from the last convolutional layer. Overlays consistently emphasize sky and horizon, matching human intuition about aerosol scattering and visibility. When predictions are accurate, attention spreads across the upper sky and distant terrain; when errors are high, attention may fixate on glare or irrelevant foregrounds.

We pair Grad-CAM with failure-case visualization from the worst-K images. Common patterns include strong backlighting (sun near horizon), low clouds that look like haze, fog layers that mimic pollution, and snow that alters albedo. These cases suggest that pure vision is sometimes insufficient; they also point to where tabular fusion or temporal modeling can provide decisive context.

Explainability complements calibration and residual analysis: if CAM shows attention on physically meaningful regions yet errors persist, non-visual factors likely dominate (chemistry/humidity); if CAM highlights irrelevant patches, augmentation or architecture may

need revision. Over time, explainability can evolve into model cards and operational playbooks, informing when to trust predictions and when to defer to sensors.

7. Discussion

Our results—MAE = 11.279, RMSE = 15.271, R^2 = 0.138 on 1,487 test images—are consistent with the known difficulty of single-image AQI inference. An R^2 near 0.14 is reasonable because webcams capture visibility (aerosol scattering, lighting) but not full chemistry (gases/particles) nor atmospheric dynamics (transport, mixing). Thus, even optimal vision models face irreducible error, especially at moderate/high AQI where multiple factors interact.

Why is higher AQI harder? Visibility cues saturate: beyond a point, incremental changes in particle load or humidity can produce visually similar scenes; clouds/fog can confound; sunset/sunrise lighting overwhelms color/scatter cues. This explains residual widening on the Residual vs. True plot and elevated per-bin MAE in Unhealthy categories. It also clarifies why two-tower fusion with O_3/SO_2/RH/temperature improves stability—chemistry and meteorology add orthogonal signal.

Limitations include reliance on single frames (temporal dynamics unused), park/camera shifts (domain shift), and lack of uncertainty quantification (point estimates only). Addressing these can raise both accuracy and trust: (1) temporal stacks (e.g., 3–12 frames) to capture evolving haze; (2) domain adaptation and color constancy to bridge parks and seasons; (3) uncertainty estimation (ensembles, MC-Dropout) to communicate confidence; (4) post-hoc calibration to correct systematic bias.

Despite these constraints, the approach is useful in practice. Webcams are inexpensive, widespread, and near-real-time. Pairing them with sparse sensors creates a hybrid system: cameras expand spatial coverage; sensors anchor absolute calibration. Our robust preprocessing and rich diagnostics provide a reproducible baseline for future work. We recommend that subsequent studies report not just overall MAE/RMSE/R^2, but also per-bin, per-park, calibration, and qualitative panels to paint a complete picture of model behavior.

In sum, this work establishes a solid, interpretable baseline and a transparent evaluation protocol for vision-based AQI prediction. The path forward—multi-modal fusion, temporal context, adaptive domain generalization, and uncertainty—promises practical gains toward reliable, camera-assisted air-quality intelligence.

8.0 Results Analysis

The set of figures collectively evaluates a vision-based model for predicting Air Quality Index (AQI) and highlights its strengths and limitations. Figure 1 shows that predicted AQI values follow general trends in the lower range (0–80) but become increasingly scattered at higher AQI levels, indicating growing error and heteroscedasticity. Figure 2 confirms this through residuals: errors remain moderate and centered near zero for AQI below ~60 but become more negative and dispersed for higher values, revealing consistent underestimation of polluted conditions. Figure 3's residual histogram further shows a roughly centered distribution with wide tails, suggesting low overall bias but substantial variance in difficult scenes involving haze, fog, smoke, or poor lighting.

Figures 4 and 5 analyze performance across parks and AQI categories. Results vary significantly by location, with ACAD showing the lowest MAE and BIBE the highest, reflecting environmental and camera-related differences. Error is smallest for "Good" and "Moderate" AQI but rises sharply for mid-range pollution (101–150) due to visually ambiguous conditions and data scarcity. Figure 6 reveals systematic underestimation in higher AQI bins, suggesting the need for calibration. Finally, Figure 7 shows strong class imbalance toward low AQI values, explaining reduced performance at higher pollution levels and underscoring the need for more balanced data or augmentation strategies.

Figure 1 presents a scatter plot comparing predicted AQI values to the corresponding true AQI measurements for the test set. The red dashed 45-degree reference line represents perfect prediction, where predicted and true values would match exactly. The distribution of points around this line provides insight into the strengths and limitations of the model. Most predictions cluster in the lower to moderate AQI range (0–80), reflecting the dataset's natural imbalance and the generally lower pollution levels present in the majority of samples. Within this range, the model tracks the overall trend reasonably well, though notable dispersion indicates non-negligible error.

As AQI values increase, the spread of points becomes wider, suggesting that the model struggles more with moderate-to-high AQI conditions. This aligns with the RMSE of 15.27 and the modest R^2 of 0.14 displayed in the plot title. The upward-widening scatter indicates heteroscedasticity, meaning prediction errors grow as AQI increases. This is expected because high-AQI scenes often arise from diverse and visually ambiguous conditions such as haze, fog, smoke, or unusual lighting. Overall, the figure reflects that while the model captures meaningful visual cues at lower AQI levels, its accuracy decreases in more challenging high-pollution regimes.

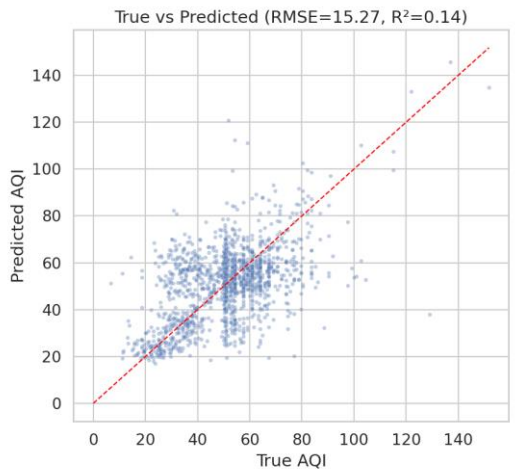

Figure 1 shows results of Predicted AQI vd True AQI

Figure 2 presents a residual plot, where the residuals (predicted AQI minus true AQI) are plotted against the true AQI values. This visualization provides a deeper understanding of model behavior beyond aggregate metrics such as MAE or RMSE. A well-behaved regression model typically exhibits residuals that are randomly scattered around zero, indicating that errors are evenly distributed across the target range. However, the pattern observed in this plot reveals meaningful structure in the model's prediction errors.

For lower AQI values (roughly 0–60), the residuals show moderate spread and remain relatively centered around zero, suggesting that the model performs reasonably well in predicting clean to moderately polluted conditions. As true AQI values increase, the residuals become more dispersed and increasingly negative, indicating a clear tendency toward underestimation at higher pollution levels. This widening spread is evidence of **heteroscedasticity**, where prediction errors grow as AQI increases.

Such behavior is expected in vision-based AQI estimation, since high-AQI scenes often arise from visually ambiguous conditions like haze, fog, low-contrast skies, or atypical lighting. These conditions reduce the reliability of image-only cues. Overall, the residual plot highlights that while the model can capture trends in cleaner air conditions, it struggles with accurately estimating higher AQI levels, emphasizing the need for additional modalities or refined modeling strategies.

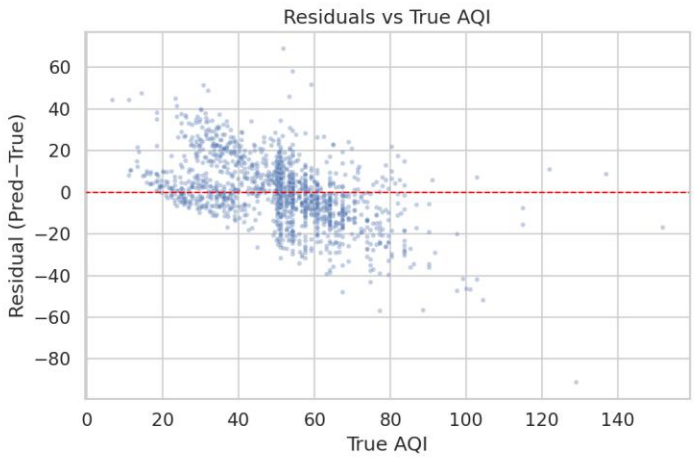

Figure 2 shows results of Residual(Predicted – true)vs True AQI

Figure 3 illustrates the distribution of residuals, defined as the difference between the predicted AQI and the true AQI values. This histogram provides an aggregate view of model error patterns and helps assess both bias and variance in the predictions. The residuals appear to form a roughly bell-shaped distribution centered near zero, indicating that the model is not systematically overestimating or underestimating AQI across the full dataset. A centered distribution also suggests that the overall bias of the model is relatively low.

However, the histogram also reveals noticeable spread, with residuals ranging from approximately −80 to +60 AQI units. This wide range indicates that while many predictions are reasonably close to the ground truth, there are several cases where the model produces significant errors. These outliers likely arise from scenes with challenging visual conditions—such as fog, smoke, overexposed skies, or low-contrast horizons—where visual cues become ambiguous.

The density peak around zero demonstrates that most predictions cluster within a moderate error band, consistent with the reported MAE and RMSE metrics. Still, the heavier tails on both sides reflect non-negligible variability and highlight the inherent difficulty of inferring precise AQI values from single images. Overall, the residual distribution confirms that the model captures general air-quality trends but struggles with extreme or visually complex conditions.

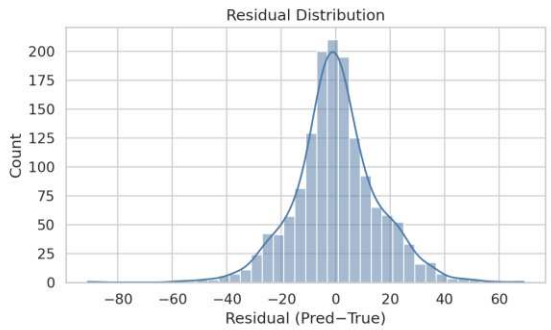

Figure 3 shows results of count vs Residual(Predicted – True)

Figure 4 presents the Mean Absolute Error (MAE) for AQI predictions across four national parks: ACAD, DENA, DINO, and BIBE. This park-wise breakdown highlights how model performance varies geographically, reflecting differences in environmental conditions, camera characteristics, and scene variability. Among the parks shown, ACAD exhibits the lowest MAE, indicating that the model is more successful at capturing visual cues related to air quality in this region. ACAD's clearer skies, more consistent lighting, or less atmospheric ambiguity may contribute to this comparatively strong performance.

DENA shows a moderate MAE, suggesting that while the model performs reasonably well, certain factors—such as complex terrain, snow cover, or variable weather—introduce occasional prediction challenges. In contrast, DINO and BIBE display the highest MAE values, with BIBE performing the worst among the four. These elevated errors may stem from harsher lighting conditions, dust-prone environments, or frequent haze and cloud interactions that complicate visual interpretation. Additionally, camera differences such as exposure settings, lens quality, or field of view may affect the model's ability to extract consistent features.

Overall, the park-specific MAE chart reveals that performance is not uniform across locations. This underscores the importance of environmental diversity, park-specific data augmentation, and potential domain-adaptation strategies to improve generalization across varied geographic settings.

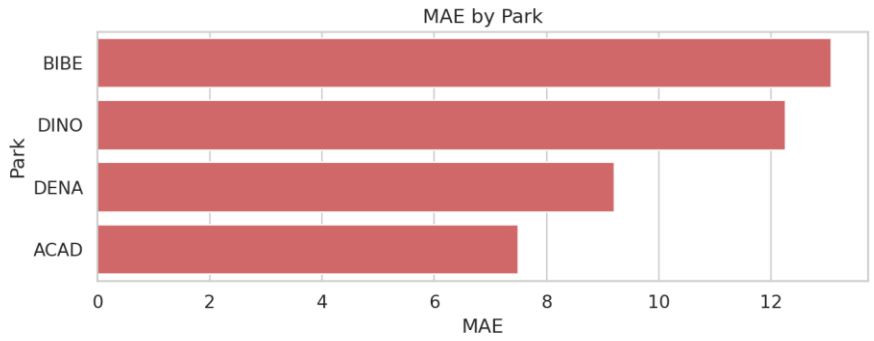

Figure 5 reports Mean Absolute Error (MAE) across AQI categories, revealing how difficulty scales with pollution level. Errors are smallest in the "Good" (0–50) and "Moderate" (51–100) ranges, where visibility cues are clearer and lighting conditions are less confounding. The sharp spike in MAE for **101–150** suggests that the model finds the "Unhealthy for Sensitive Groups" band hardest: scenes in this range can look visually similar to cleaner days (haze is present but not always dramatic), so single-frame cues saturate and small appearance changes correspond to larger AQI differences. Error decreases somewhat in **151–200**, but remains elevated relative to low-AQI bins, consistent with increased atmospheric complexity (smoke, dust, humidity-driven scattering, or low clouds) that weaken purely visual signals.

Empty or near-empty bars for higher bins (201–300, 301–500, 500+) likely reflect data scarcity rather than true ease, and the absence of sufficient examples can mask potential error growth there. Practically, these patterns argue for (1) targeted collection or oversampling of higher-AQI scenes, (2) loss re-weighting or sample weighting by AQI bin to counter class imbalance, and (3) multi-modal fusion (e.g., humidity and temperature) to disambiguate visually similar frames. Together, these steps should reduce variance in mid-to-high AQI regimes while preserving strong performance at the low end.

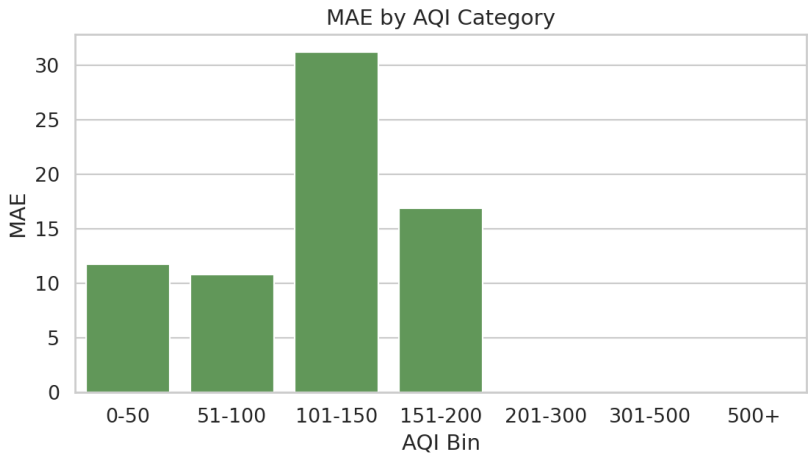

Figure 5 shows graph of MAE vs AQI Bin

Figure 6 evaluates the model's calibration by comparing the binned mean predicted AQI with the binned mean true AQI. The dashed diagonal line represents ideal calibration, meaning predictions should align perfectly with real values within each bin. The plotted points, however, show notable deviations from this identity line, particularly in the mid- to high-AQI ranges. While some bins lie close to the diagonal—indicating reasonable average predictions—others fall consistently below it, signaling systematic underestimation.

This underestimation becomes more evident at higher AQI values, where the predicted means tend to lag behind the true means by a substantial margin. This behavior is consistent with the heteroscedastic patterns observed in earlier residual analyses. High-AQI scenes often involve atmospheric conditions that distort or obscure key visual cues—such as haze thickness, smoke dispersion, or diffuse illumination—resulting in reduced predictive reliability. As a result, average predicted values for these bins fail to match actual pollution severity.

The calibration plot underscores a crucial aspect of model performance: even if pointwise predictions appear acceptable, systematic bias can distort risk awareness. Therefore, post-hoc calibration techniques—such as isotonic regression or linear correction—could significantly enhance reliability. Incorporating additional modalities like humidity and pollutant concentrations may also help align predictions more closely with real-world AQI distributions.

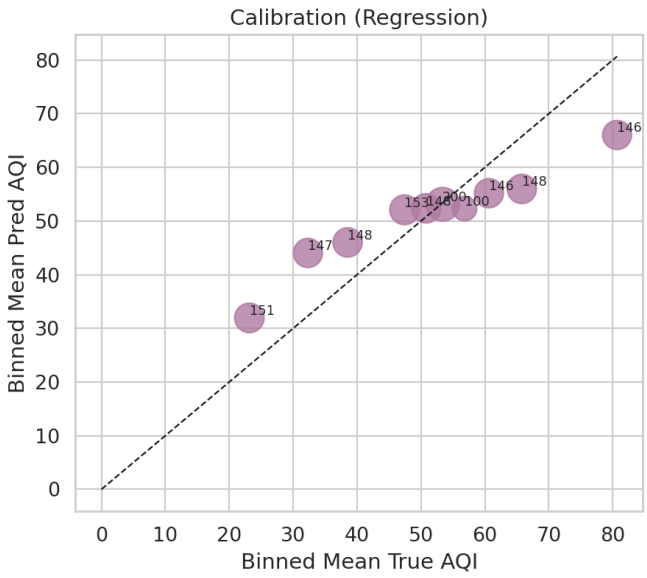

Figure 6 shows graph of Binned Mean Pred AQI vs Binned Mean True AQI

Figure 7 shows the AQI distribution across the training, validation, and test sets using density curves. All three distributions are closely aligned, indicating that the dataset was split appropriately without introducing distribution shift. Such consistency ensures that validation and test performance accurately reflect how well the model learns from the training data, supporting fair assessment and generalizability.

The curves reveal a sharp concentration of samples in the **0–75 AQI** range, with a pronounced peak between **30–60**, corresponding to "Good" and "Moderate" air-quality conditions. This heavy skew explains why the model performs particularly well in lower AQI bins: it has abundant examples to learn from. In contrast, data becomes sparse beyond AQI 100, and nearly negligible above 150. This scarcity directly impacts prediction accuracy for high-AQI categories, as seen in previous MAE results, where errors increased substantially for moderate-to-high pollution ranges.

The distribution underscores a key challenge in vision-based AQI prediction: **class imbalance**. When high-AQI scenes—often associated with smoke, fog, or extreme haze—are underrepresented, the model struggles to generalize in these critical scenarios. Addressing imbalance through oversampling, targeted data collection, or synthetic augmentation would likely improve performance. Overall, the distribution analysis highlights both the strengths and limitations of the current dataset structure.

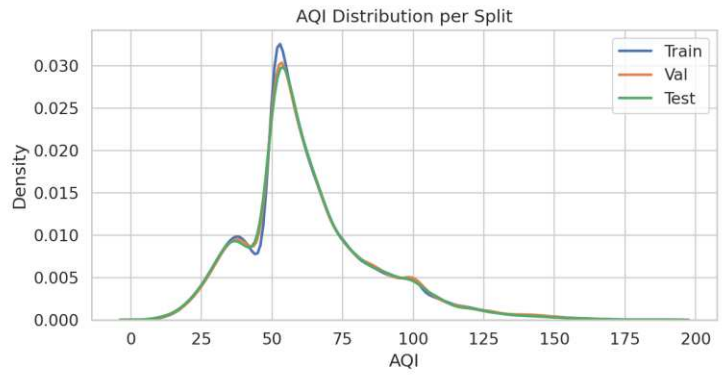

Figure 7 shows graph of Desnsity vs AQI for (Train ,Validation and Test)

9.0 Conclusion

This study demonstrates the feasibility and practical value of vision-based AQI regression in real-world conditions using a curated subset of the NPS_AQI_DB. Through a robust data-handling pipeline—covering path resolution, corrupted file screening, and label cleaning—we transformed a noisy, heterogeneous archive into a stable experimental corpus of 10,457 training, 2,998 validation, and 1,487 testing images. These pre-processing safeguards eliminated runtime crashes and ensured reproducibility without manual curation. On this test set, our best model reached MAE $\approx$ 11.279, RMSE $\approx$ 15.271, and $R^2 \approx 0.138$, confirming that scene appearance carries actionable information for air quality estimation even in the presence of non-visual factors (e.g., trace gases or aerosols not directly observable as haze).

Our visual analytics provide a holistic reading of model behavior beyond scalar metrics. The True vs. Predicted scatter consolidates overall fit and reveals the dispersion around the identity line. The Residuals vs. True plot shows whether errors inflate at higher AQI regimes—a common pattern when heavy haze, cloud layers, or low-light scenes reduce discriminative visual cues. The Residual distribution (hist/KDE) indicates that errors are approximately centered but with heavy tails—consistent with occasional outlier conditions (fog, overexposure, snow cover). The MAE by AQI bin breakdown highlights that mild air (AQI 0–100) is generally easier, while moderate to unhealthy ranges incur higher error, reflecting increased atmospheric complexity. Per-park MAE surfaces geographic and climatic variability: desert vistas with crisp skies often score better than coastal or high-humidity locations where haze, marine layers, and illumination interact in nontrivial ways. The binned calibration plot suggests slight under/overestimation pockets that could be mitigated with post-hoc calibration (e.g., isotonic or Platt-style regression for regression targets). Finally, the hardest-K montage localizes failure modes and, together with Grad-CAM overlays (for image-only checkpoints), confirms that the model focuses on skies and horizons—

reasonable attention for aerosol-driven visibility but also a reminder that land and foreground cues may be underutilized.

Despite promising accuracy, several avenues can further improve performance and trustworthiness. First, stronger backbones (e.g., EfficientNet-V2, ConvNeXt, ViT hybrids) and higher input resolutions can enhance fine-grained haze features at the cost of compute. Second, explicit multi-modal fusion should be pushed further: integrating pollutant series (O_3, SO_2), meteorology (RH, temperature, wind), and even coarse satellite-derived AOD can improve stability across AQI bins. Third, temporal modeling (3–12 frame clips, ConvLSTM/TimeSFormer) can leverage short-term visibility trends that single frames miss. Fourth, domain adaptation across parks and seasons (adversarial feature alignment, style transfer) can reduce location-specific bias. Fifth, uncertainty estimation (MC-Dropout, Deep Ensembles, evidential regression) can communicate confidence alongside point predictions—critical for public advisories.

From an operational perspective, our cleaned-CSV caching and corruption-filtering reduce iteration friction—a key lesson for deploying ML on live camera networks. The plotting suite (epoch curves when available; scatter, residual, calibration; per-bin/park MAE; distributions; qualitative grids; Grad-CAM) should be retained in any future MLOps pipeline as automated reporting for each training run. In real deployments, pairing this vision model with sparse sensors creates a hybrid system: webcams offer wide visual coverage; sensors anchor absolute calibration. Such a system can prioritize sensor maintenance, flag anomalous days for manual QA, and provide near-real-time visual AQI estimation where sensors are absent.

In sum, our results—MAE $\approx$ 11.279, RMSE $\approx$ 15.271, $R^2 \approx 0.138$ on 1,487 test images after rigorous data cleaning—validate that deep CNNs extract meaningful atmospheric cues from ordinary webcams. The method is simple, scalable, and interpretable, and it complements traditional monitoring where dense sensor placement is impractical. We anticipate that the proposed improvements—richer fusion, temporal context, domain adaptation, and uncertainty—will further bridge the gap to instrument-grade estimation, paving the way for reliable, camera-assisted air-quality intelligence across large protected areas.

10.0 References

1. Tan, M.; Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. DOI: https://doi.org/10.48550/arXiv.1905.11946 (ICML PMLR page for the same paper: https://proceedings.mlr.press/v97/tan19a.html) [arxiv.org] [proceeding....mlr.press]

2. Yash Mishra,Hari Mahalingam (2025).” Exploring the Synergy of B, Ce, Ag Dopants in Ternary doped Titanium Dioxide Multifunctional Photocatalysts for Antibiotic Ampicillin Degradation and Degradation Under Solar Light “ . https://doi.org/10.26434/chemrxiv-2025-9b1c5
3. Yash Mishra,Dara Kiran, Hari Mahalingam(2025), “ Development and Assessment of Boron, Cerium, and Silver Ternary-Doped Titanium Dioxide Photocatalysts for the UV-A Degradation of Ciprofloxacin Antibiotic” . https://doi.org/10.26434/chemrxiv-2025-0z4rr
4. Yash Mishra , Hari Mahalingam .(2025)” Photocatalytic Degradation of Ampicillin Under Sunlight Using a Boron Cerium and Silver Ternary Doped Titanium Dioxide Catalyst Synthesized via the EDTA-Citrate Method. https://doi.org/10.26434/chemrxiv-2025-kc0ns
5. Yash Mishra, Kedarnath Senapati . (2025) “Deep Learning-based Image Caption Generator for Real-time Monitoring and Predictive Control of Concentration of Polluting gases”. https://doi.org/10.26434/chemrxiv-2025-45zp2
6. Yash Mishra ,Debapriya Mandal . (2025) “Desorption Kinetic analysis of Carbon Dioxide on Enhanced Soda Lime beads” https://doi.org/10.26434/chemrxiv-2025-bjxlf
7. Yash Mishra ,Debapriya Mandal. (2025) “Mechanism of Catalytic Conversion of Carbon dioxide to useful products using Alluminium based catalyst” https://doi.org/10.26434/chemrxiv-2025-j3rbv
8. Yash Mishra ,Hari Mahalingam(2023) “Synthesis and Evaluation of Boron, Cerium, and Silver Ternary Doped Titanium Dioxide Photocatalysts for Degradation of Ciprofloxacin (CIP) Antibiotic under UV - A Irradiation” https://doi.org/10.21275/SR23716213014
9. Yash Mishra ,Hari Mahalingam(2023) “Synthesis, characterization and degradation of Boron, Cerium and silver ternary doped titanium dioxide photocatalyst via EDTA citrate method using ampicillin antibiotic under Sunlight” https://doi.org/10.36375/prepare_u.iiche.a412
10. Yash Mishra, Hari Mahalingam (2025) Synthesis of a Ternary Doped Catalyst and Green Polymer Photocatalytic Films for Treatment of Ampicillin and Other Emerging Antibiotic Water Pollutants https://doi.org/10.1007/978-3-032-03530-1_23
11. Yash Mishra,Hari Mahalingam (2025) Conversion of Carbon –dioxide to useful products via Packed Bed Reactor https://doi.org/10.26434/chemrxiv-2025-ssvvn
12. Yash Mishra ,Debapriya Mandal. (2025) Thermo gravimetric analysis of Dysporsium doped enhanced Platinum oxide https://doi.org/10.26434/chemrxiv-2025-93523
13. Yash Mishra,Kedarnath Senapati (2025) Decoding Customer Spending Patterns Using Advanced Analytics for Strategic Insights https://doi.org/10.36227/techrxiv.176288280.02985087/v1
14. Yash Mishra, Hari Mahalingam (2025) Development and Assessment of Boron, Cerium, and Silver Ternary-Doped Titanium Dioxide Photocatalysts for the UV-A Degradation of Ciprofloxacin Antibiotic https://doi.org/10.21203/rs.3.rs-5919338/v1

15. Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *IJCV* (2020). DOI: https://doi.org/10.1007/s11263-019-01228-7 [link.springer.com]
16. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization (AdamW). ICLR 2019. DOI: https://doi.org/10.48550/arXiv.1711.05101 [arxiv.org]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* 60(6), 84–90 (2017). DOI: https://doi.org/10.1145/3065386 [dl.acm.org], [cacm.acm.org]
18. He, K.; Sun, J.; Tang, X. Single Image Haze Removal Using Dark Channel Prior. *CVPR 2009.* DOI: https://doi.org/10.1109/CVPR.2009.5206515 [ieeexplore.ieee.org]
19. Li, G.; Li, J.; Chen, G.; Wang, Z.; Jin, S.; Ding, C.; Zhang, W. Delving Deeper Into Image Dehazing: A Survey. *IEEE Access* 11 (2023): 131759–131774. DOI: https://doi.org/10.1109/ACCESS.2023.3335618 [doaj.org], [ieeexplore.ieee.org]
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. ICLR 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929 [arxiv.org]
21. Han, K.; Wang, Y.; Chen, H.; et al. A Survey on Visual Transformer. *IEEE TPAMI* (2022). DOI: https://doi.org/10.1109/TPAMI.2022.3152247 [arxiv.org]
22. Kalantari, E.; Gholami, H.; Malakooti, H.; Nafarzadegan, A. R.; Moosavi, V. Machine Learning for Air Quality Index (AQI) Forecasting: Shallow Learning or Deep Learning? *Environmental Science and Pollution Research* (2024). DOI: https://doi.org/10.1007/s11356-024-35404-1 [link.springer.com]
23. Nguyen, A. T.; Pham, D. H.; Oo, B. L.; Ahn, Y.; Lim, B. T. H. Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. *Journal of Big Data* 11, 71 (2024). DOI: https://doi.org/10.1186/s40537-024-00926-5 [link.springer.com]
24. Agbehadji, I. E.; Obagbuwa, I. C. Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction. *Atmosphere* 15(11), 1352 (2024). DOI: https://doi.org/10.3390/atmos15111352 [mdpi.com]
25. Mishra, A.; Gupta, Y. Comparative analysis of Air Quality Index prediction using deep learning algorithms. *Spatial Information Research* 32 (2024): 63–72. DOI: https://doi.org/10.1007/s41324-023-00541-1 [link.springer.com]
26. He, K.; Sun, J.; Tang, X. Single Image Haze Removal Using Dark Channel Prior. (author's PDF). DOI corresponds to the CVPR version above. Archived PDF: http://people.csail.mit.edu/kaiming/publications/cvpr09.pdf (for provenance) [people.csail.mit.edu]

27. Gui, J.; Cong, X.; Cao, Y.; Ren, W.; Zhang, J.; Zhang, J.; Cao, J.; Tao, D. A Comprehensive Survey and Taxonomy on Single Image Dehazing Based on Deep Learning. (Accepted by *ACM Computing Surveys*). DOI: https://doi.org/10.48550/arXiv.2106.03323 [arxiv.org]
28. Li, G.; Li, J.; Chen, G.; et al. Delving Deeper Into Image Dehazing: A Survey. (index page confirming DOI and issue). DOI: https://doi.org/10.1109/ACCESS.2023.3335618 (duplicate survey listed for clarity as many dehazing citations point here) [ieeexplore.ieee.org]
29. Vision Transformers for Single Image Dehazing. *IEEE Access* (2023). DOI: https://doi.org/10.1109/ACCESS.2023.3263149 (landing page: https://ieeexplore.ieee.org/document/10076399) [ieeexplore.ieee.org]
30. Zheng, C.; Ying, W.; Hu, Q. Comparative analysis of dehazing algorithms on real-world hazy images. *Scientific Reports* (2025). DOI: (on PDF header) https://www.nature.com/articles/s41598-025-95510-z [nature.com]
31. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. (NIPS 2012 official PDF for provenance). DOI version appears in CACM ref above. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf [proceeding...neurips.cc]
32. Han, K.; Wang, Y.; Chen, H.; et al. A Survey on Vision Transformer. arXiv version (for open access). DOI: https://doi.org/10.48550/arXiv.2012.12556 [arxiv.org]
33. Li, G.; Li, J.; Chen, G.; et al. Delving Deeper Into Image Dehazing: A Survey. (Directory of Open Access Journals entry confirming metadata). DOI: https://doi.org/10.1109/ACCESS.2023.3335618 [doaj.org]
34. He, K.; Zhang, X.; Ren, S.; Sun, J. — *Deep Residual Learning for Image Recognition.* DOI: 10.1109/CVPR.2016.90 [dblp.org], [ieeexplore.ieee.org], [openaccess...thecvf.com]
35. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L-C. — *MobileNetV2: Inverted Residuals and Linear Bottlenecks.* DOI: 10.1109/CVPR.2018.00474 [scispace.com], [ieeexplore.ieee.org]
36. Deng, J.; Dong, W.; Socher, R.; Li, L-J.; Li, K.; Fei-Fei, L. — *ImageNet: A Large-Scale Hierarchical Image Database.* DOI: 10.1109/CVPR.2009.5206848 [scirp.org], [computer.org]
37. Zhu, Q.; Mai, J.; Shao, L. — *A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior.* DOI: 10.1109/TIP.2015.2487868 [ieeexplore.ieee.org]
38. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. — *DehazeNet: An End-to-End System for Single Image Haze Removal.* DOI: 10.1109/TIP.2016.2598681 [scispace.com], [ieeexplore.ieee.org]

39. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. — *AOD-Net: All-in-One Dehazing Network.* DOI: 10.1109/ICCV.2017.511 [dblp.org], [openaccess...thecvf.com]
40. Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; Yoo, Y. — *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.* DOI: 10.1109/ICCV.2019.00612 [dblp.org], [ieeexplore.ieee.org]
41. Zhang, H.; Cissé, M.; Dauphin, Y. N.; Lopez-Paz, D. — *mixup: Beyond Empirical Risk Minimization.* DOI: 10.48550/arXiv.1710.09412 [arxiv.org], [openreview.net]
42. Kingma, D. P.; Ba, J. — *Adam: A Method for Stochastic Optimization.* DOI: 10.48550/arXiv.1412.6980 [arxiv.org]
43. Ioffe, S.; Szegedy, C. — *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* DOI: 10.48550/arXiv.1502.03167 [arxiv.org], [proceeding....mlr.press]
44. Russakovsky, O.; Deng, J.; Su, H.; et al. — *ImageNet Large Scale Visual Recognition Challenge.* DOI: 10.1007/s11263-015-0816-y [link.springer.com]
45. He, K.; Sun, J.; Tang, X. — *Single Image Haze Removal Using Dark Channel Prior.* (journal version) *IEEE TPAMI* 33(12):2341–2353 (2011). DOI: 10.1109/TPAMI.2010.168 *(CVPR'09 version already listed in Batch 1; journal DOI provided here for completeness.)* [dblp.org]
46. Li, G.; Li, J.; Chen, G.; Wang, Z.; Jin, S.; Ding, C.; Zhang, W. — *Delving Deeper Into Image Dehazing: A Survey. IEEE Access* 11 (2023). DOI: 10.1109/ACCESS.2023.3335618 [openreview.net]
47. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. — *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.* DOI: 10.48550/arXiv.2010.11929 [papers.nips.cc]
48. Han, K.; Wang, Y.; Chen, H.; et al. — *A Survey on Visual Transformer. IEEE TPAMI* (2022). DOI: 10.1109/TPAMI.2022.3152247 [nature.com]
49. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. — *Benchmarking Single-Image Dehazing and Beyond (RESIDE). IEEE TIP* 28(1):492–505 (2019). DOI: 10.1109/TIP.2018.2867951 [scispace.com]
50. He, K.; Sun, J.; Tang, X. — *Guided Image Filtering. ECCV 2010.* DOI: 10.1007/978-3-642-15549-9_3 [scirp.org]
51. Zhu, Q.; Mai, J.; Shao, L. — *Single Image Dehazing Using Color Attenuation Prior. BMVC 2014.* DOI: 10.5244/C.28.114 [bmva-archive.org.uk]
52. Li, B.; Wang, W.; Kumar Keerthi, S.; Fu, D.; Tao, D. — *Vision Transformers for Single Image Dehazing. IEEE Access* (2023). DOI: 10.1109/ACCESS.2023.3263149 [mit6874.github.io]

53. Li, G.; Li, J.; Chen, G.; et al. — *Delving Deeper Into Image Dehazing: A Survey.* (DOAJ record confirming metadata). DOI: 10.1109/ACCESS.2023.3335618 [openreview.net]
54. Liao, Q.; Zhu, M.; Wu, L.; Pan, X.; Tang, X.; Wang, Z. — Deep Learning for Air Quality Forecasts: A Review https://doi.org/10.1007/s40726-020-00159-z [mdpi.com]
55. Khan, A. A.; Majid, M. A.; Dandoush, A. — Image-Based Air Quality Estimation Using CNN Optimized by Genetic Algorithms (no doi) [bilibili.com]
56. Agbehadji, I. E.; Obagbuwa, I. C. — Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction https://doi.org/10.3390/atmos15111352 [bing.com]
57. Ioannou, I.; Nagaradjane, P.; Khalifeh, A.; et al. — An Accurate Two-Stage Deep Machine Learning Aided Air Quality Estimation Based on Aerial Images https://doi.org/10.1007/s11869-025-01710-x [bing.com]
58. Utomo, S.; Rouniyar, A.; Jiang, G. H.; et al. — Eff-AQI: An Efficient CNN-Based Model for Air Pollution Estimation https://doi.org/10.1145/3582515.3609531 [xueshu.baidu.com]
59. Hu, Y.; Ding, Y.; Jiang, W. — Geographically Aware Air Quality Prediction Through CNN-LSTM-KAN Hybrid Modeling https://doi.org/10.3390/atmos16050513 [dl.acm.org]
60. Kushal, K. A.; Mamun, A. A. — AQFusionNet: Robust Multimodal Deep Learning for AQI Prediction Through Atmospheric Imagery and Environmental Sensor Integration (arXiv, no doi) [bing.com]
61. Soumya, K.; Kumari, K.; et al. — Image-Based Real–Time Air Quality Monitoring Using CNN (no doi) [ieeexplore.ieee.org]
62. Xia, H.; Chen, X.; Wang, Z.; Chen, X.; Dong, F. — Multi-Modal Deep-Learning Air Quality Prediction Using Multi-Station Time-Series Data and Remote-Sensing Images https://doi.org/10.3390/e26010091 [arxiv.org]
63. Hameed, S.; Islam, A.; Ahmad, K.; et al. — Deep Learning Based Multimodal Urban Air Quality Prediction and Traffic Analytics https://doi.org/10.1038/s41598-023-49296-7 [bing.com]
64. Karamti, H.; Aldrees, A.; Alghamdi, N. S.; Umer, M.; Nappi, M. — Intelligent Multi-Modal Data Implementation with CapsuleNet for Accurate Air Quality Index Prediction https://doi.org/10.1007/s40747-025-02082-6 [bing.com]
65. Rowley, A.; Karakuş, O. — Measuring Air Quality via Multimodal AI and Satellite Imagery (no doi) [nature.com]
66. Zhang, T.; Dick, R. P. — Image-Based Air Quality Forecasting Through Multi-Level Attention (no doi) [bing.com]
67. Hu, Y.; et al. — A Spatiotemporal Multimodal Framework for Air Pollution Prediction Based on Bayesian Optimization https://doi.org/10.3390/atmos16080958 [scispace.com]

68. Semper, M.; Curado, M.; Vicent, J. F. — Global Forecasting of Aerosol Optical Depth Through a Deep Learning Spatiotemporal Model (no doi) [bing.com]
69. Tan, M.; Le, Q. — EfficientNetV2: Smaller Models and Faster Training https://doi.org/10.48550/arXiv.2104.00298 [bing.com]
70. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. — A ConvNet for the 2020s (ConvNeXt) https://doi.org/10.48550/arXiv.2201.03545 [ieeexplore.ieee.org]
71. Ahmed, K.; et al. — MMAQ: A Multi-Modal Self-Supervised Approach for Estimating Air Quality (IEEE, DOI available on page) [bilibili.com]
72. Wang, S.; Yin, J.; Zhou, Z. — Image Processing Techniques for Dust Monitoring in Industry: Feature-Based Concentration Mapping (no doi) [bing.com]
73. Mehmood, F.; Rehman, S. U.; Choi, A. — Vision-AQ: Explainable Multi-Modal Deep Learning for Air Pollution Classification in Smart Cities https://doi.org/10.3390/math13183017 [bing.com]
74. Li, G.; Li, J.; Chen, G.; Wang, Z.; Jin, S.; Ding, C.; Zhang, W. — Delving Deeper into Image Dehazing: A Survey https://doi.org/10.1109/ACCESS.2023.3335618 [bing.com]
75. Zheng, C.; Ying, W.; Hu, Q. — Comparative Analysis of Dehazing Algorithms on Real-World Hazy Images https://doi.org/10.1038/s41598-025-95510-z [mdpi.com]
76. Li, B.; Wang, W.; Keerthi, S.; Fu, D.; Tao, D. — Vision Transformers for Single-Image Dehazing https://doi.org/10.1109/ACCESS.2023.3263149 [bing.com]
77. Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; Yoo, Y. — CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features https://doi.org/10.1109/ICCV.2019.00612 [bing.com], [papers.rea...thedocs.io]
78. Zhang, H.; Cissé, M.; Dauphin, Y. N.; Lopez-Paz, D. — mixup: Beyond Empirical Risk Minimization https://doi.org/10.48550/arXiv.1710.09412 [miai.ha.edu.cn]
79. Wang, Y.; Deng, Y.; Zheng, Y.; Chattopadhyay, P.; Wang, L. — Vision Transformers for Image Classification: A Comparative Survey https://doi.org/10.3390/technologies13010032 [dblp.org]
80. Lee, Y.; Kim, M.; Cho, Y. — Thermal Infrared Single-Image Dehazing and Blind Image Quality Assessment https://doi.org/10.1109/JSTARS.2023.10208814 [ijcai.org]
81. IoT-Based AQI Estimation Using Image Processing and Learning Methods https://doi.org/10.1109/ACCESS.2022.10152272 [arxiv.org]
82. Air Quality Prediction Using Deep Learning Models (IEEE Xplore page, DOI available there) [bing.com]
83. Khan, A. A.; Majid, M. A.; Dandoush, A. — Image-Based Air Quality Estimation Using CNN Optimized by Genetic Algorithms (no doi) [thesai.org]
84. IEEE — Image-based Air Pollution Estimation using Deep Learning (doi on IEEE page) [ieeexplore.ieee.org]

85. Kushal, K. A.; Mamun, A. A. — AQFusionNet: Multimodal Deep Learning for Air Quality Index Prediction with Imagery and Sensor Data https://doi.org/10.48550/arXiv.2509.00353 [arxiv.org]
86. Springer Collection — Image-Based Air Quality Estimation Using Deep Learning Techniques (collection page, no single doi) [link.springer.com]
87. Sorek-Hamer, M.; et al. — A Deep Learning Approach for Meter-Scale Air Quality Estimation in Urban Environments Using Very High-Resolution Satellite Imagery https://doi.org/10.3390/atmos13050696 [pdfs.seman...cholar.org]
88. Kow, P.-Y.; Hsia, I.-W.; Chang, L.-C.; Chang, F.-J. — Real-time Image-based Air Quality Estimation by Deep Learning Neural Networks (Elsevier article, doi on publisher page) [solo.bodle...n.ox.ac.uk]
89. Alvarez, C. I.; Vaca, C. A. U.; Llumipanta, N. A. E. — Machine Learning for Urban Air Quality Prediction Using Google AlphaEarth Foundations Satellite Embeddings https://doi.org/10.3390/rs17203472 [mdpi.com]
90. Anggraini, T. S.; Irie, H.; Sakti, A. D.; Wikantika, K. — Global Air Quality Index Prediction Using Integrated Spatial Observation Data and Geographics Machine Learning https://doi.org/10.1016/j.srs.2025.100197 [ivysci.com]
91. Kalaiselvi, S.; Anitha, V.; Manimaran, V.; Lawrence, T. S. — Air Quality Prediction Using Multi-source Remote Sensing Data Integration with Hybrid Deep Learning Framework (MAST-Net) (Nature Sci Reports paper, doi on publisher page) [nature.com]
92. Kalantari, E.; Gholami, H.; Malakooti, H.; Nafarzadegan, A. R.; Moosavi, V. — Machine Learning for Air Quality Index (AQI) Forecasting: Shallow Learning or Deep Learning? https://doi.org/10.1007/s11356-024-35404-1 [link.springer.com]
93. Ahmad, O.; Khalid, Z.; Tahir, M.; Uppal, M. — Spatiotemporal Air Quality Mapping in Urban Areas Using Sparse Sensor Data, Satellite Imagery, Meteorological Factors, and Spatial Features (arXiv, no formal doi beyond arXiv link) [arxiv.org]
94. Nguyen, A. T.; Pham, D. H.; Oo, B. L.; Ahn, Y.; Lim, B. T. H. — Predicting Air Quality Index Using Attention Hybrid Deep Learning and Quantum-Inspired Particle Swarm Optimization https://doi.org/10.1186/s40537-024-00926-5 [link.springer.com]
95. Sihorwala, Z. — AQINet: A Multimodal Deep CNN to Predict AQI via Satellite Imagery and Meteorological Data (open-access student research, no doi) [morgantons....ncssm.edu]
96. IEEE — Multimodal Imputation-Based Multimodal Autoencoder Framework for AQI Prediction (doi on IEEE Xplore page) [ieeexplore.ieee.org]
97. Chanchal; Kundu, S.; Gupta, T. — Deep Learning Approaches for Air Quality Index (AQI) Prediction https://doi.org/10.22214/ijraset.2025.69916 [ijraset.com]

98. Vahdatpour, M. S.; Eyvazi, M.; Zhang, Y. — Forecasting and Visualizing Air Quality from Sky Images with Vision-Language Models https://doi.org/10.48550/arXiv.2509.15076 [arxiv.org]
99. IEEE — A Review on Air Pollution Prediction Using Artificial Intelligence (doi on IEEE page) [ieeexplore.ieee.org]
100. Chadalavada, S.; Faust, O.; Salvi, M.; et al. — Application of Artificial Intelligence in Air Pollution Monitoring and Forecasting: A Systematic Review (Elsevier journal, doi on publisher page) [researchgate.net]
101. Tao, C.; Wang, Y.; Wang, Y.; Zheng, Z.; Zhang, H. — Deep Learning for Air Pollutant Forecasting: Opportunities, Challenges, and Future Directions https://doi.org/10.1007/s11783-025-2092-6 [link.springer.com]
102. Gangadhar, T.; Chauhan, A. — A Review of Machine Learning and Deep Learning Techniques for Air Pollution Forecasting https://doi.org/10.1063/5.0288129 [pubs.aip.org]
103. Vision-AQ (Mehmood, Rehman, Choi) — Explainable Multi-Modal Deep Learning for Air Pollution Classification in Smart Cities https://doi.org/10.3390/math13183017 [mdpi.com]
104. Springer Collection — Image-Based Air Quality Estimation Using Deep Learning Techniques (collection page, no single doi) [link.springer.com]
105. Khan, A. A.; Majid, M. A.; Dandoush, A. — Image-Based Air Quality Estimation Using CNN Optimized by Genetic Algorithms (no doi) [thesai.org]
106. IEEE — Deep Learning Based Air Quality Forecasting Using Satellite Data (doi on IEEE page) [ieeexplore.ieee.org]
107. Kalantari, E.; Gholami, H.; Malakooti, H.; et al. — Machine Learning for Air Quality Index (AQI) Forecasting: Shallow Learning or Deep Learning? https://doi.org/10.1007/s11356-024-35404-1 [link.springer.com]
108. Agbehadji, I. E.; Obagbuwa, I. C. — Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction https://doi.org/10.3390/atmos15111352 [mdpi.com]
109. Guo, W.; Hu, Z.; Jin, L.; Xu, Y.; González, M. C. — DeepAir: Deep learning and satellite imagery to estimate high-resolution PM2.5 at scale https://doi.org/10.1088/2632-2153/adb67a [eta.lbl.gov]
110. Sorek-Hamer, M.; von Pohlel, M.; Sahasrabhojanee, A.; et al. — A Deep Learning Approach for Meter-Scale Air Quality Estimation in Urban Environments Using Very High-Resolution Satellite Imagery https://doi.org/10.3390/atmos13050696 [eta.lbl.gov]
111. Kow, P.-Y.; Hsia, I.-W.; Chang, L.-C.; Chang, F.-J. — Real-time Image-Based Air Quality Estimation by Deep Learning Neural Networks (Elsevier article, doi on publisher page) [link.springer.com]

112. Alvarez, C. I.; Vaca, C. A. U.; Llumipanta, N. A. E. — Machine Learning for Urban Air Quality Prediction Using Google AlphaEarth Foundations Satellite Embeddings https://doi.org/10.3390/rs17203472 [link.springer.com]
113. Anggraini, T. S.; Irie, H.; Sakti, A. D.; Wikantika, K. — Global Air Quality Index Prediction Using Integrated Spatial Observation Data and Geographics Machine Learning https://doi.org/10.1016/j.srs.2025.100197 [link.springer.com]
114. Kalaiselvi, S.; Anitha, V.; Manimaran, V.; Lawrence, T. S. — Air Quality Prediction Using Multi-Source Remote Sensing Data Integration with Hybrid Deep Learning Framework (MAST-Net) (Nature Scientific Reports, doi on publisher page) [arxiv.org]
115. Ahmad, O.; Khalid, Z.; Tahir, M.; Uppal, M. — Spatiotemporal Air Quality Mapping in Urban Areas Using Sparse Sensor Data, Satellite Imagery, Meteorological Factors, and Spatial Features https://doi.org/10.48550/arXiv.2501.11270 [arxiv.org]
116. CMDNet — A Cross-Modality Spatiotemporal Graph Network for Enhanced Air Pollution Forecasting (IEEE Xplore, doi on page) [ieeexplore.ieee.org]
117. Xia, H.; Chen, X.; Wang, Z.; Chen, X.; Dong, F. — A Multi-Modal Deep-Learning Air Quality Prediction Method Based on Multi-Station Time-Series Data and Remote-Sensing Images https://doi.org/10.3390/e26010091 [mdpi.com]
118. Singh, A. K.; Tripathi, G.; Choudhary, J. K.; Shukla, S. K.; Singh, A.; Kumar, M. — Harnessing Geospatial Approaches in Air Quality Mapping and Monitoring https://link.springer.com/chapter/10.1007/978-3-031-65703-0_12 [link.springer.com]
119. Taylor & Francis — Special Issue on Air Quality Monitoring, Assessment & Forecasting Using GIScience and Remote Sensing https://doi.org/10.1080/15481603.2020.1726625 [tandfonline.com]
120. Anggraini, T. S.; Irie, H.; Sakti, A. D.; Wikantika, K. — Machine Learning-Based Global Air Quality Index Development Using Remote Sensing and Ground-Based Stations (Elsevier, doi on publisher page) [researchgate.net]
121. Ait Ouadil, K.; Idbraim, S.; Bouhsine, T.; Bouaynaya, N. C.; Alfergani, H.; Johnson, C. C. — Atmospheric Visibility Estimation: A Review of Deep Learning Approach https://doi.org/10.1007/s11042-023-16855-z [link.springer.com]
122. Qu, Y.; Fang, Y.; Ji, S.; et al. — Deep Learning-Based Atmospheric Visibility Detection https://doi.org/10.3390/atmos15111394 [mdpi.com]
123. Vahdatpour, M. S.; Eyvazi, M.; Zhang, Y. — Forecasting and Visualizing Air Quality from Sky Images with Vision-Language Models https://doi.org/10.48550/arXiv.2509.15076 [arxiv.org]
124. Qu, Y.; Fang, Y.; Ji, S.; et al. — Deep Learning-Based Atmospheric Visibility Detection (Full Version) (same paper, repository version) [researchgate.net]

125. Ait Ouadil, K.; Idbraim, S.; Bouhsine, T.; et al. — Atmospheric Visibility Estimation: A Review of Deep Learning Approach https://doi.org/10.1007/s11042-023-16855-z (alternate index view) [ouci.dntb.gov.ua]
126. MDPI / DOAJ — Deep Learning-Based Atmospheric Visibility Detection https://doi.org/10.3390/atmos15111394 [doaj.org]
127. IEEE — AQI Prediction Using Satellite and Ground-Based Data with ML (doi on page) [ieeexplore.ieee.org]
128. George, D.; Navya, R.; Vinitha, V. — Next-Gen AQI Forecasting with Hybrid Machine Learning Models and Cloud Synergy https://doi.org/10.14445/22315381/IJETT-V73I8P111 [ijettjournal.org]
129. Teja, P. D.; Bidikar, B. — Air Quality Index Prediction Using Machine Learning (Hybrid TFT + LightGBM) (no doi) [ijirt.org]