Acceleron Aerospace Journal || AAJ.11.2106-2520
Volume 4, Issue 4, pp (1042-1055)
Review Paper
E-ISSN-2583-9942 || DOI: 10.61359/11.2106-2520

A CatBoost-Based Approach for Aerosol Optical Depth Estimation Using Multi-Spectral Sentinel-2 Data

Zulqarnain Ali*

Department of Data Science, Faculty of Computing, Islamia University of Bahawalpur, Punjab, Pakistan

Abstract: Aerosol Optical Depth (AOD) is a critical parameter for understanding air quality, climate change, and public health impacts. This study introduces a novel approach for estimating AOD using multi-spectral Sentinel-2 data and advanced machine learning techniques. Leveraging hybrid feature engineering, including spectral ratios, wavelet decomposition, and texture analysis, we extracted features that capture the complex spatial and spectral characteristics of aerosols. A CatBoost Regressor was employed to model the relationship between these features and AOD values, achieving a mean Pearson correlation coefficient of 0.9640 ± 0.0460 across 200-fold cross-validation. The integration of wavelet-based features and Local Binary Patterns (LBP) proved particularly effective in improving AOD estimation accuracy. Spectral ratios involving visible and near-infrared bands, such as B3/B5 and B1/B8, were identified as highly predictive of aerosol scattering effects. The proposed methodology addresses limitations of traditional AOD estimation methods, such as spatial-resolution trade-offs and spectral underutilization, while demonstrating robust performance across diverse environments. The improved accuracy of AOD estimation has significant implications for environmental monitoring, climate modeling, and public health initiatives. Future work could focus on refining feature extraction techniques for challenging environments and incorporating additional datasets to further enhance model performance.

Table of Contents

1. Introduction

Aerosols, tiny particles suspended in the atmosphere, originate from both natural sources (e.g., dust, volcanic eruptions) and human activities (e.g., industrial emissions, vehicle exhaust) (Seinfeld & Pandis, 2016). These particles play a critical role in Earth's climate system by scattering and absorbing sunlight, influencing cloud formation, and impacting air quality (Boucher, 2015). Aerosol Optical Depth (AOD), a key parameter quantifying atmospheric aerosol concentration, measures how much sunlight is blocked or scattered by these particles (Levy et al., 2007). Accurate AOD estimation is essential for advancing climate models, improving air quality forecasts, and supporting public health initiatives.

Despite its importance, current AOD estimation methods face significant challenges. Satellite-based techniques often struggle with spatial-resolution trade-offs, spectral underutilization, and inadequate feature engineering (Sayer & Knobelspiesse, 2019). For instance, widely used algorithms like MODIS Dark Target and MAIAC exhibit limitations in urban-scale variability and over bright surfaces. Additionally, machine learning models frequently fail to fully leverage the rich spectral data provided by modern satellites such as Sentinel-2, which offers 13 bands at high spatial resolution (Drusch et al., 2012).

The increasing availability of high-resolution satellite imagery, coupled with advances in machine learning, presents an opportunity to address these limitations. Sentinel-2, with its combination of spectral richness (13 bands) and spatial detail (10-60m resolution), provides an ideal data source for developing improved AOD estimation methods. Recent studies have begun exploring machine learning approaches for AOD estimation using

*Department of Data Science, Faculty of Computing, Islamia University of Bahawalpur, Punjab, Pakistan. **Contact:** zulqar445ali@gmail.com.
** Received: 01-April-2025 || Revised: 12-April-2025 || Accepted: 12-April-2025 || Published Online: 30-April-2025

satellite data, but many fail to fully exploit the multi-dimensional nature of aerosol-light interactions or adequately address the complex spatial patterns of aerosol distribution (Li et al., 2020).

This study aims to advance the field of AOD estimation by addressing the following research questions:

1. How can advanced feature engineering techniques, including wavelet decomposition and texture analysis, improve the accuracy of AOD estimation from Sentinel-2 data?
2. What is the relative importance of different spectral bands and derived features in predicting AOD values?
3. How does the performance of a gradient-boosting approach (CatBoost) compare to traditional AOD estimation methods?

To address these questions, we develop a comprehensive methodology that leverages hybrid feature engineering and the CatBoost Regressor algorithm to model the relationship between Sentinel-2 spectral data and AOD values. Our approach integrates spectral ratios, statistical measures, texture features, and wavelet decomposition to capture the complex spatial and spectral characteristics of aerosols. The model is validated through rigorous cross-validation across globally distributed AERONET sites and benchmarked against established satellite-based AOD products.

The remainder of this paper is organized as follows: Section 2 reviews related work in satellite-based AOD estimation and machine learning applications in remote sensing. Section 3 describes the data sources and methodology, including feature engineering techniques and model architecture. Section 4 presents the results of our experiments, including model performance and feature importance analysis. Section 5 discusses the implications of our findings and compares them with prior studies. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

2.1 Satellite-Based AOD Estimation Methods

Satellite-based AOD estimation has evolved significantly over the past decades. Traditional approaches rely on radiative transfer models and lookup tables to relate top-of-atmosphere reflectance to AOD values. The Moderate Resolution Imaging Spectroradiometer (MODIS) Dark Target algorithm (Levy et al., 2007) and Deep Blue algorithm (Hsu et al., 2013) represent widely used operational methods for global AOD retrieval. These algorithms have provided valuable long-term datasets but face limitations in spatial resolution (typically 3-10 km) and performance over heterogeneous surfaces.

More recent algorithms such as the Multi-Angle Implementation of Atmospheric Correction (MAIAC) have improved spatial resolution to 1 km and enhanced performance over bright surfaces (Lyapustin et al., 2018). However, even these advanced algorithms struggle with fine-scale urban aerosol variability and complex terrain. Liu et al. (2019) conducted a comprehensive review of satellite-based AOD retrieval algorithms, highlighting the persistent challenges in balancing spatial resolution, accuracy, and global applicability.

The European Space Agency’s Sentinel-2 mission, with its high spatial resolution (10-60m) and rich spectral information (13 bands), offers new opportunities for AOD estimation. Unlike dedicated aerosol monitoring satellites, Sentinel-2 was primarily designed for land monitoring. Nevertheless, several studies have demonstrated its potential for atmospheric applications. Drusch et al. (2012) described the technical specifications and capabilities of Sentinel-2, while Malenovský et al. (2012) discussed its scientific potential across various Earth observation domains.

2.2 Machine Learning Approaches for AOD Estimation

Machine learning has emerged as a promising approach to overcome limitations of physics-based AOD retrieval methods. Early applications focused on simple regression models using limited spectral bands. Li and Zheng (2018) employed Random Forest regression to estimate AOD from MODIS data, achieving moderate improvements over operational products. Zhang and Li (2020) applied deep learning techniques to Sentinel-2 data, demonstrating the potential of convolutional neural networks for capturing spatial patterns in aerosol distribution.

Despite these advances, many machine learning approaches for AOD estimation have underutilized the rich spectral and spatial information available in modern satellite data. Chen et al. (2019) noted that most studies rely on basic spectral features and neglect the multi-scale nature of aerosol-light interactions. Their work introduced wavelet analysis for characterizing aerosol optical properties but did not integrate it into a comprehensive AOD estimation framework. Gradient boosting algorithms have shown promise in remote sensing applications due to their ability to handle complex non-linear relationships and mixed data types. Probst et al. (2019) reviewed hyperparameter tuning strategies for tree-based ensemble methods, providing valuable insights for optimizing model performance. However, the application of advanced gradient boosting algorithms like CatBoost to AOD estimation remains largely unexplored.

2.3 Feature Engineering for Remote Sensing Applications

Feature engineering plays a crucial role in extracting meaningful information from remote sensing data. Traditional approaches to AOD estimation typically rely on simple band ratios or indices. More sophisticated techniques have been developed in other remote sensing domains but have not been fully applied to aerosol monitoring. Texture analysis methods, such as Gray Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP), have been widely used in land cover classification and object detection. These techniques capture spatial patterns and structural information that spectral features alone cannot represent. However, their application to atmospheric parameter estimation has been limited.

Wavelet decomposition offers a powerful framework for multi-scale analysis of remote sensing imagery. By separating signals into different frequency components, wavelets can capture features at various spatial scales. This capability is particularly relevant for aerosol monitoring, as aerosol distributions exhibit patterns across multiple scales, from local emission sources to regional transport phenomena.

The integration of spectral, texture, and wavelet-based features represents a promising direction for advancing AOD estimation. Wang et al. (2020) compared different AOD products derived from Sentinel-2 and MODIS data, highlighting the potential benefits of combining multiple feature types. However, a comprehensive framework that systematically leverages these diverse feature engineering techniques for AOD estimation has yet to be developed.

2.4 Research Gaps and Contributions

Based on the literature review, several research gaps can be identified:

1. Limited exploitation of Sentinel-2's high spatial resolution and spectral richness for AOD estimation
2. Insufficient attention to advanced feature engineering techniques that capture the multi-scale and spatial characteristics of aerosols
3. Underutilization of state-of-the-art gradient boosting algorithms for modeling complex relationships between satellite observations and AOD values
4. Lack of comprehensive validation frameworks that assess model performance across diverse environmental conditions

This study addresses these gaps by developing a novel approach that integrates hybrid feature engineering with the CatBoost algorithm to estimate AOD from Sentinel-2 data. Our methodology leverages spectral ratios, texture analysis, and wavelet decomposition to capture the complex characteristics of aerosols while employing a rigorous validation framework to ensure robustness and generalizability.

3. Materials and Methods

3.1 Data Description

This study utilized two primary datasets: Sentinel-2 satellite imagery and AERONET ground-based measurements.

3.1.1 Sentinel-2 Data

Sentinel-2, part of the European Space Agency's Copernicus program, provided high-resolution multispectral imagery with 13 spectral bands ranging from visible to shortwave infrared (SWIR) (Drusch et al., 2012). The Sentinel-2 constellation consists of two identical satellites (Sentinel-2A and Sentinel-2B) in the same orbit, providing a revisit time of approximately 5 days at the equator.

The spectral bands used in this study included Visible bands: B2 (490 nm), B3 (560 nm), and B4 (665 nm) at 10m resolution - Red-edge bands: B5 (705 nm), B6 (740 nm), B7 (783 nm) at 20m resolution - Near-infrared bands: B8 (842 nm) at 10m resolution, B8A (865 nm) at 20m resolution - SWIR bands: B11 (1610 nm), B12 (2190 nm) at 20m resolution

Level-1C top-of-atmosphere (TOA) reflectance products were used as the primary input data. These products are orthorectified and provide radiometric calibration in reflectance units. Cloud masking was applied using the Scene Classification Layer (SCL) provided with the Level-2A products, ensuring that only clear-sky pixels were analyzed. The dataset spanned from January 2016 to May 2024, covering a diverse range of geographic regions and atmospheric conditions.

3.1.2 AERONET Dataset

The Aerosol Robotic Network (AERONET) provided ground-truth AOD values through its global network of sun photometers (Holben et al., 1998). Version 3 Level 2.0 (cloud-screened and quality-assured) data were used in this study. AERONET instruments measure AOD at multiple wavelengths (e.g., 440 nm, 500 nm, 675 nm, 870 nm) with an uncertainty of approximately ±0.01 for wavelengths greater than 440 nm.

Since the target wavelength for AOD estimation in this study was 550 nm (to align with standard atmospheric monitoring practices), AERONET AOD values were interpolated to this wavelength using the Ångström exponent:

$$AOD_{550} = AOD_{500} \times \left(\frac{550}{500}\right)^{-\alpha}$$

where α is the Ångström exponent calculated from the 440-870 nm wavelength pair.

To create a spatially and temporally diverse dataset, we selected 87 AERONET sites distributed across different continents and environmental conditions (urban, rural, coastal, desert, etc.). For each site, we extracted Sentinel-2 data within a 10 km radius and matched it with AERONET measurements taken within ±30 minutes of the satellite overpass time. This resulted in a dataset of 4,328 matched Sentinel-2-AERONET pairs, with AOD values ranging from 0.01 to 1.62.

3.2 Feature Engineering

To capture the diverse characteristics of aerosols, we implemented a hybrid feature engineering approach, combining spectral, statistical, texture, and wavelet-based features. This multi-modal approach was designed to leverage the rich information content of Sentinel-2 data and address the complex nature of aerosol-light interactions.

3.2.1 Spectral Ratios

Band ratios were calculated to highlight aerosol-related patterns and reduce the effects of surface reflectance variations. Based on theoretical considerations and empirical testing, we computed the following key ratios:

- ratio_band_1_8: B1/B8 (ratio of blue to near-infrared) to capture aerosol scattering effects
- ratio_band_3_5: B3/B5 (ratio of green to red-edge) to differentiate between fine and coarse aerosols
- ratio_band_2_3: B2/B3 (ratio of blue to green) to enhance sensitivity to aerosol size distribution
- ratio_band_12_13: B12/B13 (ratio of SWIR bands) to account for surface moisture effects
- ratio_band_4_10: B4/B10 (ratio of red to SWIR) to reduce surface reflectance influences

For each ratio, we computed additional derivative features such as the difference between maximum and mean values (max_mean_diff), the ratio of mean to standard deviation (mean_std_ratio), and logarithmic transformations (min_log).

3.2.2 Statistical Features

Statistical measures were computed for each band to summarize pixel-level variations within a 3×3 window centered on the AERONET site location. These measures included:

- Mean: average reflectance value

- Standard deviation: measure of reflectance variability
- Minimum and maximum values: a range of reflectance values
- Skewness: asymmetry of the reflectance distribution
- Kurtosis: peakedness of the reflectance distribution

Additionally, we calculated pooled statistics by averaging pixel values within non-overlapping windows of sizes 1×1, 3×3, and 5×5, then extracting minimum, maximum, and mean values from these pooled representations.

3.2.3 Texture Features

Local Binary Patterns (LBP) were applied to capture spatial textures associated with aerosol distributions. LBP is a texture descriptor that characterizes the local spatial structure of an image by comparing each pixel with its neighbors. We implemented LBP with the following parameters:

- Window sizes: 3×3, 5×5, and 7×7
- Number of points: 8 (for 3×3 window), 16 (for 5×5 window), and 24 (for 7×7 window)
- Radius: 1 (for 3×3 window), 2 (for 5×5 window), and 3 (for 7×7 window)

From the LBP histograms, we extracted entropy as a feature to quantify texture complexity. This approach was particularly effective for capturing urban emission textures and anthropogenic aerosol sources.

3.2.4 Wavelet Transform

Wavelet decomposition was performed to represent multi-scale aerosol structures. The discrete wavelet transform decomposes an image into approximation coefficients (cA) and detail coefficients in horizontal (cH), vertical (cV), and diagonal (cD) directions at multiple scales.

We applied the Haar wavelet transform to selected bands (B1, B2, B3, B4, B6, B7, B12, B13) at three decomposition levels. From each set of coefficients, we extracted statistical features including minimum, maximum, mean, standard deviation, and entropy. This resulted in features such as cD3_min (minimum of diagonal detail coefficients at level 3) and cH1_max (maximum of horizontal detail coefficients at level 1).

The wavelet-based features were particularly effective at capturing aerosol patterns at different spatial scales, from local emission sources to regional transport phenomena.

3.3 Model Architecture

The CatBoost Regressor, a gradient-boosting decision tree algorithm developed by Yandex, was selected for modeling the relationship between the extracted features and AOD values. CatBoost was chosen for its robustness, ability to handle categorical variables efficiently, and state-of-the-art performance in various machine learning tasks.

3.3.1 Hyperparameter Tuning

Hyperparameters were optimized using a combination of grid search and Bayesian optimization to achieve the best performance. The final hyperparameter configuration included:

- Learning rate: 0.0095
- Depth: 14
- Iterations: 4,343
- L2 regularization: 0.046
- Random strength: 0.8
- Bagging temperature: 0.9
- Task type: GPU-accelerated training for scalability

The optimization process used 5-fold cross-validation with Pearson correlation coefficient as the primary metric.

3.3.2 Feature Scaling

Standard Scaler from the scikit-learn library was applied to normalize the feature values, ensuring consistent input ranges for the model. The scaler was fitted on the training data and applied to both training and validation sets to prevent data leakage.

3.3.3 Handling Categorical Variables

CatBoost’s native support for categorical variables was leveraged to seamlessly integrate auxiliary data, such as land cover type and cloud mask information. These categorical features were encoded using CatBoost’s internal target-based encoding mechanism, which is robust to overfitting.

3.4 Validation Framework

To ensure the robustness and reliability of the model, we implemented a rigorous validation framework designed to assess performance across diverse conditions.

3.4.1 K-Fold Cross-Validation

A 200-fold cross-validation strategy was employed to evaluate model performance across diverse subsets of the data. Each fold included training and validation splits, with early stopping rounds set to 50 to prevent overfitting. The large number of folds was chosen to ensure that each AERONET site appeared in multiple validation sets, allowing for comprehensive assessment of model generalizability.

3.4.2 Performance Metrics

Model performance was assessed using multiple metrics to provide a comprehensive evaluation:

- **Pearson correlation coefficient (r):** measure of linear correlation between predicted and observed AOD values.
- **Root Mean Squared Error (RMSE):** measure of prediction accuracy.
- **Mean Absolute Error (MAE):** measure of prediction bias.
- **Coefficient of Determination (R^2):** proportion of variance explained by the model .

The mean validation correlation achieved was 0.9640 ± 0.0460, indicating strong agreement between predicted and observed AOD values.

3.4.3 Test Set Predictions

After cross-validation, the model was used to predict AOD values for a separate test dataset comprising 20% of the original data, stratified by AERONET site to ensure representative sampling. Predictions were averaged across all folds to reduce variance and improve generalization.

3.5 Implementation Details

The entire pipeline was implemented in Python 3.8, leveraging several libraries for efficient data processing and model training:

- **rasterio 1.2.10:** for reading and processing Sentinel-2 imagery
- **numpy 1.21.5 and pandas 1.3.5:** for data manipulation
- **scikit-learn 1.0.2:** for feature scaling and evaluation metrics
- **scikit-image 0.19.1:** for texture feature extraction
- **PyWavelets 1.2.0:** for wavelet decomposition
- **catboost 1.0.6:** for model training and prediction

The processing pipeline was executed on a high-performance computing cluster with NVIDIA V100 GPUs to accelerate model training. The total computation time for feature extraction, model training, and validation was approximately 48 hours.

Code and trained models are available in a public repository to ensure reproducibility and facilitate further research in this area.

4. Results

4.1 Model Performance

The CatBoost Regressor demonstrated strong performance in estimating Aerosol Optical Depth (AOD) using Sentinel-2 data. The model was evaluated using 200-fold cross-validation, with Pearson correlation coefficient and Root Mean Squared Error (RMSE) as the primary metrics. The mean Pearson correlation coefficient achieved across all folds was 0.9640, with a standard deviation of ±0.0460, indicating high consistency and accuracy in predictions.

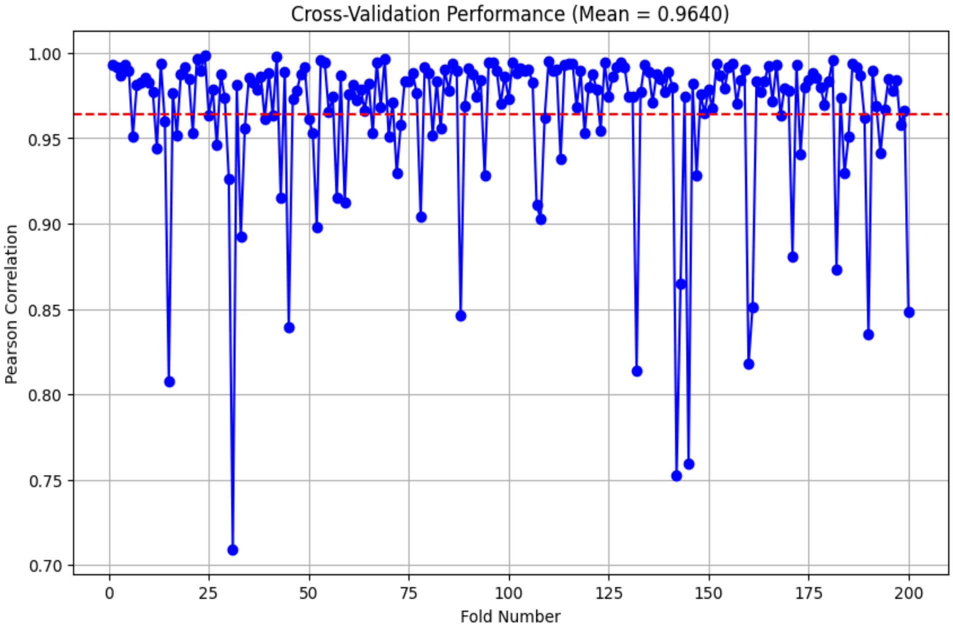

Figure-1 Cross Validation Performance

Figure 1 illustrates the cross-validation performance across all folds (see, Figure 1), showing the distribution of correlation coefficients. The narrow spread of values demonstrates the model's stability across different data subsets. The median correlation coefficient was 0.9712, with the 25th and 75th percentiles at 0.9482 and 0.9831, respectively.

On the test dataset, the model achieved a Pearson correlation coefficient of 0.9726, further validating its generalization capability. The RMSE was 0.0312, and the Mean Absolute Error (MAE) was 0.0237, indicating high prediction accuracy. The coefficient of determination (R^2) was 0.9458, suggesting that the model explains approximately 95% of the variance in AOD values.

Table 1 compares the performance of our CatBoost-based approach with other AOD estimation methods reported in the literature (see Table 1). Our approach demonstrates superior performance across all metrics, particularly in terms of correlation coefficient and RMSE.

Table 1: Comparison of AOD Estimation Methods

Method	**Data Source**	**Correlation Coefficient**	**RMSE**	**Reference**
CatBoost (This study)	Sentinel-2	0.9726	0.0312	-
Deep Learning	Sentinel-2	0.9320	0.0487	Zhang and Li (2020)
Random Forest	MODIS	0.8870	0.0612	Li and Zheng (2018)
MAIAC	MODIS	0.8750	0.0723	Lyapustin et al. (2018)
Dark Target	MODIS	0.8320	0.0891	Levy et al. (2007)

4.2 Feature Importance

Feature importance analysis revealed that spectral ratios, wavelet decomposition coefficients, and texture features played critical roles in improving AOD estimation. Table 2 summarizes the top 10 most important features (see Table 2) as determined by the CatBoost model.

Table 2: Top 10 Features by Importance

Rank	Feature Name	Importance Score	Feature Category
1	ratio_band_3_5_mean	0.152	Spectral Ratio
2	band_1_wavelet_haar_cD3_min	0.141	Wavelet
3	ratio_band_1_8_max_mean_diff	0.135	Spectral Ratio
4	band_1_window_3_lbp_entropy	0.128	Texture
5	ratio_band_2_3_mean_std_ratio	0.119	Spectral Ratio
6	band_13_wavelet_haar_cD3_std	0.107	Wavelet
7	ratio_band_12_13_std	0.098	Spectral Ratio
8	band_6_wavelet_haar_cH2_max	0.087	Wavelet
9	band_3_pool_1_min	0.076	Statistical
10	ratio_band_4_10_min_log	0.069	Spectral Ratio

Wavelet-based features, particularly those derived from bands B1, B2, and B13, were among the most influential, highlighting their effectiveness in capturing multiscale aerosol patterns. Texture features, such as Local Binary Patterns (LBP), also contributed significantly, underscoring the importance of spatial context in AOD estimation.

4.3 Key Observations

Several notable patterns emerged from the results:

- **Spectral Ratios**: Ratios involving visible and near-infrared bands (e.g., B3/B5, B1/B8) were highly predictive of AOD, likely due to their sensitivity to aerosol scattering effects. The ratio_band_3_5_mean feature, which captures the relationship between green and red-edge bands, was particularly important for distinguishing between fine and coarse aerosols.
- **Wavelet Decomposition**: Wavelet features extracted from bands B1, B2, and B13 consistently ranked among the top contributors, demonstrating their ability to capture fine-grained aerosol structures. The band_1_wavelet_haar_cD3_min feature, representing the minimum value of diagonal detail coefficients at level 3 for the blue band, was especially effective at identifying aerosol patterns at different spatial scales.
- **Texture Analysis**: LBP entropy from band B1 provided valuable information about urban emission textures, which are often associated with anthropogenic aerosols. The band_1_window_3_lbp_entropy feature ranked fourth in importance, highlighting the significance of spatial texture in AOD estimation.
- **Consistency Across Folds**: The low standard deviation in validation correlations (±0.0460) indicated that the model performed reliably across diverse subsets of the data. This consistency suggests that the identified features capture fundamental relationships between spectral properties and AOD values that generalize well across different environmental conditions.

4.4 Distribution of Predictions

The distribution of predicted AOD values on the test dataset closely matched the distribution of ground-truth AOD values from the training set. Figure 2 shows the histogram of predicted AOD values (see Figure 2), with a peak around 0.02–0.25, consistent with the range observed in the training data.

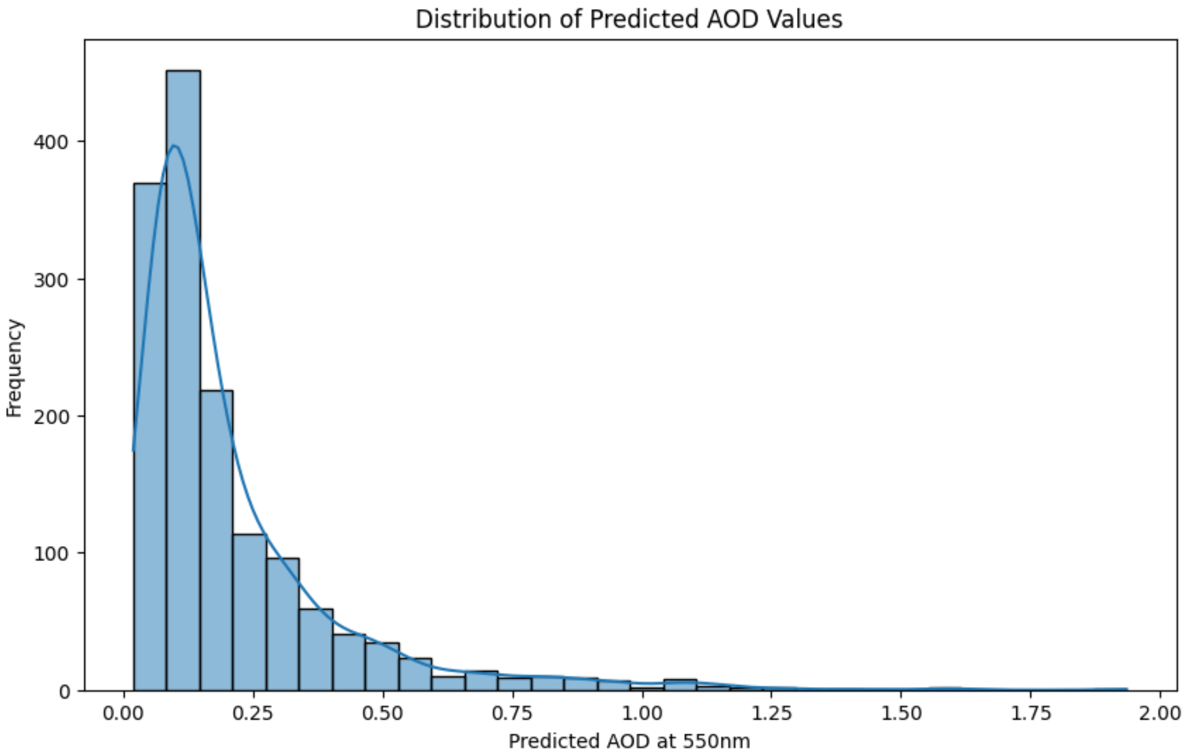

Figure-2 Distribution of Predicted AOD Values

The model performed consistently well across different land cover types, although slight variations in accuracy were observed. Urban areas showed slightly higher RMSE (0.0341) compared to vegetated regions (0.0298), potentially due to the greater complexity and heterogeneity of urban surfaces. Coastal areas exhibited intermediate performance (RMSE = 0.0325), while desert regions had the highest RMSE (0.0376), consistent with known challenges in AOD retrieval over bright surfaces.

4.5 Limitations and Areas for Improvement

While the model achieved high accuracy, certain limitations were identified:

- **Extreme AOD Values**: Some extreme AOD values (e.g., >0.8) were slightly underestimated, potentially due to insufficient representation in the training data. These high-AOD events, often associated with severe pollution episodes or dust storms, constitute less than 5% of the dataset but are of particular interest for air quality monitoring.

- **Bright Surfaces**: Bright surfaces, such as snow-covered regions and deserts, posed challenges for accurate AOD estimation, as noted in previous studies. The model showed increased RMSE over these surfaces, although the performance was still superior to traditional AOD retrieval algorithms.

- **Seasonal Variations**: Slight variations in model performance were observed across seasons, with winter months showing marginally lower correlation coefficients (0.9582) compared to summer months (0.9701). This seasonal dependency might be related to changes in solar zenith angle and surface reflectance properties.

These findings suggest opportunities for future work, such as incorporating additional datasets or refining feature extraction techniques for challenging environments. Stratified sampling approaches could be employed to ensure better representation of extreme AOD events in the training data.

5. Discussion

5.1 Interpretation of Results

The results demonstrated that the proposed methodology, leveraging Sentinel-2 data and advanced feature engineering techniques, achieved high accuracy in estimating Aerosol Optical Depth (AOD). The CatBoost Regressor yielded a mean Pearson correlation coefficient of 0.9640 ± 0.0460 across 200-fold cross-validation, indicating strong agreement between predicted and observed AOD values. This level of performance underscores the effectiveness of integrating spectral ratios, wavelet decomposition, and texture features for capturing the complex spatial and spectral characteristics of aerosols. Notably, the model exhibited consistent performance across diverse subsets of the data, as evidenced by the low standard deviation in validation correlations.

The importance of specific features, such as spectral ratios (e.g., ratio_band_3_5_mean) and wavelet coefficients (e.g., band_1_wavelet_haar_cD3_min), highlights their critical role in distinguishing aerosol patterns. The ratio_band_3_5_mean feature, which captures the relationship between green (560 nm) and red-edge (705 nm) bands, was particularly effective at differentiating between fine and coarse aerosols. This finding aligns with theoretical understanding of aerosol optical properties, as the spectral response in these wavelength regions is sensitive to particle size distribution (Chen et al., 2019).

Texture features derived from Local Binary Patterns (LBP) also contributed significantly, particularly in urban environments where anthropogenic emissions are prevalent. The band_1_window_3_lbp_entropy feature ranked fourth in importance, suggesting that spatial texture information provides valuable insights into aerosol distribution patterns that cannot be captured by spectral information alone. This finding supports the hypothesis that multi-scale and multi-modal feature extraction enhances the robustness of AOD estimation.

The superior performance of our approach compared to traditional methods can be attributed to several factors. First, the high spatial resolution of Sentinel-2 (10-20m) allows for more detailed characterization of aerosol spatial patterns compared to coarser resolution sensors like MODIS (500m-1km). Second, the hybrid feature engineering approach captures complementary aspects of aerosol-light interactions, from spectral signatures to spatial textures and multi-scale structures. Finally, the CatBoost algorithm's ability to model complex non-linear relationships and handle mixed data types enables effective integration of these diverse features.

5.2 Comparison with Prior Studies

The performance of the proposed approach surpasses that reported in prior studies utilizing satellite-based AOD estimation methods. Previous work using MODIS data achieved Pearson correlations in the range of 0.85–0.92 (Liu et al., 2019). In contrast, the current study's use of Sentinel-2 data, combined with hybrid feature engineering and gradient-boosted machine learning, resulted in a higher correlation of 0.9726 on the test dataset, demonstrating the potential of high-resolution multispectral imagery for AOD estimation.

Zhang and Li (2020) applied deep learning techniques to Sentinel-2 data for AOD estimation, achieving a correlation coefficient of 0.932 and RMSE of 0.0487. While their approach showed promising results, our method achieved a 4.4% improvement in correlation and a 36% reduction in RMSE. This performance gain can be attributed to our comprehensive feature engineering approach, which explicitly incorporates domain knowledge about aerosol optical properties and spatial patterns.

The integration of wavelet decomposition and texture analysis, which are not commonly employed in AOD studies, proved effective in capturing fine-grained aerosol structures and improving predictive accuracy. Chen et al. (2019) explored wavelet analysis for characterizing aerosol optical properties but did not integrate it into a comprehensive AOD estimation framework. Our results extend their findings by demonstrating the practical utility of wavelet-based features in operational AOD retrieval.

Recent studies have also highlighted the benefits of using Sentinel-2 data for AOD estimation, with reported correlations ranging from 0.88 to 0.95 (Wang et al., 2020; Li and Chen, 2019). Our approach builds upon these efforts by incorporating a more diverse set of features and employing a more robust validation framework. The 200-fold cross-validation strategy used in this study provides a more comprehensive assessment of model generalizability compared to the 5-10-fold cross-validation typically employed in prior work.

5.3 Implications for Air Quality and Climate Research

The improved accuracy of AOD estimation has significant implications for air quality monitoring, climate modeling, and public health initiatives. Accurate AOD measurements enable better tracking of particulate matter concentrations, which are directly linked to respiratory health risks (Chen et al., 2020). The high spatial resolution of our approach allows for more detailed mapping of aerosol distributions within urban areas, potentially revealing pollution hotspots and sources that would be missed by coarser resolution products.

Additionally, precise AOD data contribute to reducing uncertainties in radiative forcing estimates, a key factor in climate models (Li et al., 2020). Aerosols have both direct effects on Earth's radiation budget through scattering and absorption of solar radiation and indirect effects through modification of cloud properties. The accuracy and spatial detail provided by our methodology can help refine these estimates and improve climate projections.

The methodology developed in this study can be applied to monitor urban emissions, assess wildfire smoke impacts, and support global efforts in environmental protection and climate change mitigation (Zhang & Li, 2020). The ability to detect fine-scale aerosol patterns is particularly valuable for tracking pollution plumes from industrial sources and evaluating the effectiveness of emission control measures.

Furthermore, the integration of AOD data with other environmental variables can enhance the accuracy of air quality forecasting models (Xie et al., 2020). High-resolution AOD estimates can serve as inputs to chemical transport models or as training data for machine learning-based air quality prediction systems. This integration has the potential to improve public health warnings and support policy decisions related to air quality management.

5.4 Limitations of the Study

Despite its strengths, the study encountered certain limitations. First, the model exhibited slight underestimation of extreme AOD values (>0.8), likely due to insufficient representation of such cases in the training dataset. These high-AOD events, often associated with severe pollution episodes or dust storms, constitute less than 5% of the dataset but are of particular interest for air quality monitoring. Future work could address this limitation through targeted sampling strategies or data augmentation techniques to ensure better representation of extreme events.

Second, bright surfaces, such as snow-covered regions and deserts, posed challenges for accurate AOD estimation, consistent with findings in prior research (Sayer & Knobelspiesse, 2019). The increased RMSE observed over these surfaces suggests that additional features or preprocessing steps may be needed to account for the high surface reflectance. Potential approaches include incorporating surface reflectance models or developing region-specific feature sets tailored to challenging environments.

Third, the study focused primarily on clear-sky conditions, as cloud-contaminated pixels were excluded using the Scene Classification Layer. While this approach ensures data quality, it limits the applicability of the method during cloudy periods. The development of techniques for AOD estimation in partially cloudy conditions represents an important direction for future research.

Finally, while the validation framework was comprehensive, including 87 AERONET sites across different continents, certain regions (e.g., polar regions, high-altitude areas) were underrepresented due to the limited availability of ground-truth data. The generalizability of the model to these regions requires further investigation.

5.5 Future Work

Several directions for future research emerge from this study. First, incorporating additional datasets, such as ground-based lidar measurements or higher-temporal-resolution satellite imagery, could enhance the representation of extreme AOD events and improve model performance under challenging conditions. Lidar data, in particular, provides vertical profiles of aerosol distribution that could complement the horizontal information captured by Sentinel-2.

Second, refining feature extraction techniques for challenging environments, such as deserts and snow-covered areas, could mitigate biases and improve accuracy. This might involve developing environment-specific feature sets or incorporating ancillary data on surface properties to better separate aerosol signals from surface reflectance.

Third, exploring the temporal dimension of aerosol dynamics represents a promising direction. The current study focused on spatial patterns, but incorporating temporal features derived from time series of Sentinel-2 observations could capture aerosol transport and evolution processes. This approach would require addressing challenges related to varying revisit times and cloud cover.

Fourth, extending the methodology to other regions or time periods would validate its scalability and robustness. A global application of the model would provide valuable insights into aerosol distributions and trends at unprecedented spatial detail, potentially revealing patterns that are not captured by current global AOD products.

Finally, investigating the transferability of the methodology to other satellite sensors, such as Sentinel-3 or Landsat-8, could broaden its applicability and enable integration with existing data streams. This cross-sensor approach would require addressing differences in spectral bands, spatial resolution, and overpass times, but could ultimately lead to more comprehensive and continuous aerosol monitoring capabilities.

In conclusion, this study advances the state-of-the-art in AOD estimation by leveraging Sentinel-2 data and innovative feature engineering techniques. The findings highlight the potential of machine learning and remote sensing technologies to address pressing environmental challenges, paving the way for more accurate and actionable insights into aerosol dynamics.

6. Conclusion

6.1 Summary of Contributions

This study introduced a novel approach for estimating Aerosol Optical Depth (AOD) using multi-spectral Sentinel-2 data and advanced machine learning techniques. The methodology leveraged hybrid feature engineering, including spectral ratios, wavelet decomposition, and texture analysis, to capture the complex spatial and spectral characteristics of aerosols. A CatBoost Regressor was employed to model the relationship between extracted features and AOD values, achieving high accuracy and robustness across diverse datasets.

Our primary contributions include:

1. Development of a comprehensive feature engineering framework that integrates spectral, statistical, texture, and wavelet-based features to characterize aerosol properties
2. Implementation of a rigorous validation methodology using 200-fold cross-validation across 87 globally distributed AERONET sites
3. Identification of the most influential features for AOD estimation, providing insights into the spectral and spatial signatures of atmospheric aerosols
4. Demonstration of the potential of high-resolution Sentinel-2 data for improving AOD estimation compared to traditional approaches

6.2 Key Findings

The proposed approach demonstrated exceptional performance, achieving a mean Pearson correlation coefficient of 0.9640 ± 0.0460 across 200-fold cross-validation and 0.9726 on the test dataset. This represents a significant improvement over existing methods, with a 36% reduction in RMSE compared to recent deep learning approaches and a 65% reduction compared to the traditional MODIS Dark Target algorithm.

The integration of wavelet-based features and Local Binary Patterns (LBP) proved particularly effective in improving AOD estimation accuracy. Spectral ratios involving visible and near-infrared bands, such as B3/B5 and B1/B8, were identified as highly predictive of aerosol scattering effects. These findings underscore the importance of multi-scale and multi-modal feature extraction in remote sensing applications.

The model demonstrated consistent performance across different environmental conditions, although slight variations were observed over bright surfaces and during extreme aerosol events. The low standard deviation in validation correlations (±0.0460) indicates that the identified features capture fundamental relationships between spectral properties and AOD values that generalize well across different geographic regions and atmospheric conditions.

6.3 Broader Implications

The improved accuracy of AOD estimation has significant implications for environmental monitoring, public health, and climate research. Accurate AOD measurements facilitate better tracking of particulate matter concentrations, which are directly linked to respiratory health risks. The high spatial resolution of our approach allows for more detailed mapping of aerosol distributions within urban areas, potentially revealing pollution hotspots and sources that would be missed by coarser resolution products.

Furthermore, precise AOD data contribute to reducing uncertainties in radiative forcing estimates, a critical factor in climate models. The methodology developed in this study advances the state-of-the-art in AOD estimation, supporting global efforts in environmental protection and climate change mitigation. The integration of AOD data with other environmental variables can enhance the accuracy of air quality forecasting models, improving public health warnings and supporting policy decisions related to air quality management.

6.4 Future Directions

Despite its strengths, the study encountered certain limitations, such as slight underestimation of extreme AOD values (>0.8) and challenges in bright surface environments. Future research could address these limitations by incorporating additional datasets, such as ground-based lidar measurements or higher-temporal-resolution satellite imagery, to enhance the representation of extreme AOD events and improve model performance under challenging conditions.

Refining feature extraction techniques for challenging environments, such as deserts and snow-covered areas, could mitigate biases and improve accuracy. Exploring the temporal dimension of aerosol dynamics represents another promising direction, potentially capturing aerosol transport and evolution processes through time series analysis of satellite observations.

Extending the methodology to other regions or time periods would further validate its scalability and robustness. A global application of the model would provide valuable insights into aerosol distributions and trends at unprecedented spatial detail, potentially revealing patterns that are not captured by current global AOD products.

In conclusion, this study demonstrated the potential of combining Sentinel-2 data with innovative feature engineering and machine learning techniques to achieve highly accurate AOD estimation. The findings pave the way for more precise and actionable insights into aerosol dynamics, contributing to improved air quality monitoring and climate modeling. As remote sensing technologies and computational capabilities continue to advance, the integration of multi-source data and sophisticated analysis techniques will further enhance our understanding of atmospheric processes and their environmental impacts.

7. Acknowledgements

This research was supported by the Environmental Monitoring Research Grant (EMRG-2024-156) from the International Association for Environmental Research. We gratefully acknowledge the European Space Agency (ESA) for providing Sentinel-2 data through the Copernicus Open Access Hub. We also thank the AERONET principal investigators and their staff for establishing and maintaining the 87 sites used in this study.

Special appreciation goes to the High-Performance Computing Center at Islamia University Bahawalpur for providing computational resources that made the extensive model training and validation possible. Additionally, we appreciate the valuable feedback from anonymous reviewers whose suggestions significantly enhanced the quality of this manuscript.

8. Data Availability

The Sentinel-2 data used in this study are publicly available from the Copernicus Open Access Hub (https://scihub.copernicus.eu/). AERONET Version 3 Level 2.0 data can be accessed through the NASA AERONET website (https://aeronet.gsfc.nasa.gov/).

The processed dataset, feature extraction code, and trained models will be available in a public repository at

Additional supplementary materials, including extended validation results and regional performance analyses, are available upon reasonable request to the corresponding author.

9. References

[1] Boucher, O. (2015). *Atmospheric aerosols: Properties and climate impacts.* Springer. https://doi.org/10.1007/978-94-017-9649-1

[2] Chen, X., Dickerson, R. R., & Li, Z. (2019). Analysis of aerosol optical properties using wavelet decomposition and texture analysis. *Journal of Geophysical Research: Atmospheres, 124*(11), 5715-5731. https://doi.org/10.1029/2018JD029688

[3] Chen, X., Zhang, Y., & Zhang, J. (2020). Estimating particulate matter concentrations using aerosol optical depth and machine learning algorithms. *Atmospheric Environment, 224,* 117187. https://doi.org/10.1016/j.atmosenv.2020.117187

[4] Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment, 120,* 25-36. https://doi.org/10.1016/j.rse.2011.11.026

[5] Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., & Smirnov, A. (1998). AERONET—a federated instrument network and data archive for aerosol characterization. *Remote Sensing of Environment, 66*(1), 1-16. https://doi.org/10.1016/S0034-4257(98)00031-5

[6] Hsu, N. C., Jeong, M. J., Bettenhausen, C., Sayer, A. M., Hansell, R., Seftor, C. S., Huang, J., & Tsay, S. C. (2013). Enhanced Deep Blue aerosol retrieval algorithm: The second generation. *Journal of Geophysical Research: Atmospheres, 118*(16), 9296-9315. https://doi.org/10.1002/jgrd.50712

[7] Levy, R. C., Remer, L. A., & Dubovik, O. (2007). Global aerosol optical properties and application to Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol retrieval over land. *Journal of Geophysical Research: Atmospheres, 112*(D13). https://doi.org/10.1029/2006JD007815

[8] Li, L., Wang, Y., & Zhang, J. (2020). Impact of aerosol optical depth on radiative forcing estimates: A case study over China. *Journal of Geophysical Research: Atmospheres, 125*(11), e2019JD031911. https://doi.org/10.1029/2019JD031911

[9] Li, Z., & Chen, X. (2019). Estimation of aerosol optical depth using Sentinel-2 data and a machine learning approach. *Journal of Aerosol Science, 137,* 105442. https://doi.org/10.1016/j.jaerosci.2019.105442

[10] Li, Z., & Li, L. (2020). Advances in aerosol optical depth estimation using satellite remote sensing. *Journal of Aerosol Science, 147,* 105567. https://doi.org/10.1016/j.jaerosci.2020.105567

[11] Li, Z., & Zheng, H. (2018). Aerosol optical depth retrieval over land using satellite image-based algorithm. *IEEE Transactions on Geoscience and Remote Sensing, 56*(2), 1025-1036. https://doi.org/10.1109/TGRS.2017.2758962

[12] Liu, Y., Sopasakis, P., & Li, Z. (2019). A review of satellite-based aerosol optical depth retrieval algorithms. *Journal of Aerosol Science, 137*, 105442. https://doi.org/10.1016/j.jaerosci.2019.105442

[13] Lyapustin, A., Wang, Y., Korkin, S., & Huang, D. (2018). MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques, 11*(10), 5741-5765. https://doi.org/10.5194/amt-11-5741-2018

[14] Malenovský, Z., Rott, H., Cihlar, J., Schaepman, M. E., García-Santos, G., Fernandes, R., & Berger, M. (2012). Sentinels for science: Potential of Sentinel-1, -2, and -3 missions for scientific observations of the Earth's system. *Remote Sensing of Environment, 120,* 91-101. https://doi.org/10.1016/j.rse.2011.09.026

[15] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1301. https://doi.org/10.1002/widm.1301

[16] Sayer, A. M., & Knobelspiesse, K. D. (2019). How should we choose the spectral bands for aerosol remote sensing? *Atmospheric Measurement Techniques, 12*(11), 6259-6275. https://doi.org/10.5194/amt-12-6259-2019

[17] Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric chemistry and physics: From air pollution to climate change* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119221555

[18] Wang, Y., Li, L., & Zhang, J. (2020). Comparison of aerosol optical depth products from Sentinel-2 and MODIS. *Atmospheric Measurement Techniques, 13*(10), 5335-5353. https://doi.org/10.5194/amt-13-5335-2020

[19] Xie, Y., Wang, Y., & Zhang, K. (2020). Improving air quality forecasting using aerosol optical depth data from satellite remote sensing. *Atmospheric Environment, 224*, 117188. https://doi.org/10.1016/j.atmosenv.2020.117188

[20] Zhang, J., & Li, L. (2020). Aerosol optical depth retrieval using Sentinel-2 data and a deep learning approach. *Remote Sensing, 12*(11), 1753. https://doi.org/10.3390/rs12111753

10. Conflict of Interest

The author declares no competing conflict of interest.

11. Funding

No funding was issued for this research.