Global aerosol-type classification using a new hybrid algorithm and Aerosol Robotic Network data

Varsha Hole[a]*, Namrata Mohane[b], Tanishq Nagdev[c], Aaditya Sawant[d], Harsh Zaveri[e]

[a]CSE Department, Sardar Patel Institute of Technology, Mumbai, India; [b]CSE Department, Sardar Patel Institute of Technology, Mumbai, India; [c]CSE Department, Sardar Patel Institute of Technology, Mumbai, India; [d]CSE Department, Sardar Patel Institute of Technology, Mumbai, India; [e]CSE Department, Sardar Patel Institute of Technology, Mumbai, India

Corresponding Author: Varsha Hole. Department of Computer Science, Sardar Patel Institute of Technology, Mumbai, India, 400 092.
Email: varsha_hole@spit.ac.in

Global aerosol-type classification using a new hybrid algorithm and Aerosol Robotic Network data

Accurate classification of global aerosol types is essential for improving the precision and efficiency of aerosol identification in atmospheric and climate studies. Traditional methods, based on clustering or empirical thresholds, often suffer from limited accuracy and scalability due to the dynamic nature of aerosol properties across regions and time. To address these limitations, a novel hybrid approach was developed that integrates a Gaussian kernel density clustering algorithm with a Random Forest classifier. An aerosol optical properties database was constructed using a Mie scattering model, leveraging data from 47 globally distributed Aerosol Robotic Network (AERONET) sites. Key features such as single-scattering albedo, aerosol optical depth, asymmetry parameter, and extinction Ångström exponent were used as model inputs. The proposed algorithm achieved a macro-precision of 93%, macro-F1-score of 93%, and overall accuracy of 93.2%. A global aerosol-type distribution map was generated, successfully characterizing the spatial distribution of dust, mixed, urban/industrial, and biomass burning aerosols. These results demonstrate that the hybrid algorithm enhances classification accuracy and enables rapid, large-scale aerosol identification, providing valuable insights for climate research, pollution monitoring, and atmospheric modeling.

Keywords: aerosol classification; machine learning; random forest; kernel density clustering; AERONET; aerosol optical properties; mie scattering; atmospheric monitoring
Subject classification codes: 86A10; 62H30

1. Introduction

1.1 Problem Statement

Accurately identifying aerosol types is a fundamental task in atmospheric science, with significant implications for understanding aerosol–climate interactions, environmental pollution, and radiative transfer modeling. Aerosols exhibit highly variable optical, physical, and chemical properties across different geographical regions and time periods, influenced by both natural sources (such as desert dust and biomass burning) and anthropogenic activities (such as urban-industrial emissions). Traditional aerosol classification methods, which rely on empirical thresholds or conventional clustering algorithms using a limited set of optical parameters, often fail to capture this complexity with high precision. Moreover, these methods are time-consuming and lack scalability, making them less effective for large-scale or real-time applications. The scope of this project is to develop a robust, scalable, and accurate machine learning-based aerosol-type classification model that can handle diverse and heterogeneous aerosol datasets, and to generate a global aerosol-type distribution map using data from multiple representative regions.

1.2 Objective

The primary objective of this project is to classify aerosol types at a global scale using a hybrid machine learning algorithm that integrates Gaussian kernel density clustering and a Random Forest classifier. The goal is to enhance the precision and efficiency of aerosol-type identification by leveraging a comprehensive database of aerosol optical properties derived from AERONET observations and Mie scattering model simulations.

1.3 Motivation

Aerosols play a critical role in modulating the Earth's climate by directly scattering and absorbing solar radiation and indirectly influencing cloud formation and properties. They also affect air quality and human health, making their accurate classification essential for climate modeling, environmental monitoring, and public health assessments. Existing classification methods are often constrained by limited feature sets, insufficient spatial coverage, and challenges in handling large datasets. By applying advanced machine learning techniques to aerosol classification, this project aims to overcome these limitations, contributing to more reliable aerosol inversion, improved pollution source identification, and enhanced understanding of aerosol-induced climatic effects.

2. Literature review

Accurate classification of aerosol types is essential for understanding aerosol–climate interactions, radiative forcing, and air quality assessment. Over the past two decades, several approaches have been developed to classify aerosol types based on their optical and microphysical properties. Traditional classification methods have primarily relied on empirical thresholds and unsupervised clustering techniques.

One widely used approach is the threshold-based classification method, which applies fixed empirical cutoffs to aerosol optical depth (AOD), single-scattering albedo (SSA), and extinction Ångström exponent (EAE) to distinguish between different aerosol types. Giles et al. proposed one such framework using AERONET data to differentiate dust, biomass burning, and urban-industrial aerosols. While thresholding methods offer interpretability and simplicity, they often fail to account for the continuous variability and regional heterogeneity of aerosol properties, leading to misclassification, particularly in mixed aerosol environments.

To address these challenges, unsupervised clustering algorithms have been applied. For example, Dubovik and King [4] used principal component analysis (PCA) and hierarchical clustering to identify aerosol groups based on multi-wavelength optical properties. Similarly, kernel density estimation techniques have been explored to estimate probability distributions of aerosol features and to identify natural clusters in high-dimensional spaces. However, clustering methods alone are sensitive to parameter choices and lack explicit classification labels, limiting their applicability for large-scale operational classification.

The emergence of machine learning algorithms, particularly supervised learning methods such as Random Forests, has opened new avenues for aerosol classification. Random Forest, introduced by Breiman, is an ensemble learning method that constructs multiple decision trees and aggregates their outputs for robust classification. It is capable of handling non-linear relationships, high-dimensional feature spaces, and multi-class problems effectively. Recent studies have successfully applied Random Forests to related tasks such as aerosol retrieval, atmospheric parameter prediction, and remote sensing classification, demonstrating superior accuracy and interpretability compared to traditional methods.

Despite these advancements, limited work has focused on integrating unsupervised density-based clustering with supervised machine learning classifiers for aerosol-type classification. Wei et al. proposed a hybrid algorithm combining Gaussian kernel density clustering with Random Forest classification, leveraging an extensive optical property database generated using Mie scattering simulations and AERONET observations. Their approach demonstrated improved classification performance and computational efficiency, making it suitable for global aerosol mapping.

Building upon this foundation, the present study adopts a similar hybrid framework to enhance aerosol-type classification accuracy by leveraging key optical features such as SSA, asymmetry parameter, normalized AOD, and EAE. This method addresses the limitations of previous threshold-based and purely clustering approaches, offering a scalable and accurate solution for large-scale aerosol classification.

3. Methodology

The proposed methodology integrates unsupervised clustering and supervised machine learning to achieve accurate aerosol-type classification on a global scale. The approach consists of four main components: dataset construction, preprocessing and feature extraction, hybrid model implementation, and performance evaluation. A schematic overview of the methodology is presented in Figure 1.

3.1 Dataset Construction

An aerosol optical property database was developed by combining observational data from 47 AERONET sites with simulated data from a Mie scattering model. The AERONET dataset provides multi-wavelength measurements of aerosol optical depth (AOD), single-scattering albedo (SSA), asymmetry parameter (g), and extinction Ångström exponent (EAE). The Mie scattering model was used to generate synthetic optical properties based on median size distribution parameters and refractive indices for five aerosol types: dust, mixed-coarse, mixed-fine, urban/industrial, and biomass burning.

3.2 Preprocessing and Feature Extraction

Several preprocessing steps were applied to ensure data quality and comparability:

- Missing Data Handling: Data points with incomplete or invalid optical measurements were excluded.
- Normalization: AOD values at multiple wavelengths were normalized relative to AOD at 440 nm to eliminate concentration dependence.
- Feature Selection: Based on domain knowledge and exploratory analysis, the following features were selected for classification:
    - Single-scattering albedo (SSA) at 440, 675, 870, 1020 nm
    - Asymmetry parameter (g) at 440, 675, 870, 1020 nm
    - Normalized AOD at 675, 870, 1020 nm
    - Extinction Ångström exponent (EAE440–870)

These features encapsulate key optical characteristics relevant for aerosol-type discrimination.

3.3 Hybrid Model Implementation

The classification framework integrates Gaussian kernel density clustering and a Random Forest classifier to leverage the strengths of both unsupervised and supervised learning.

- Gaussian Kernel Density Clustering: This unsupervised step identifies dominant clusters within the optical property space, generating initial labels and probabilistic boundaries between aerosol types.
- Random Forest Classifier: A supervised ensemble classifier trained on the labeled dataset refines the classification boundaries. It handles complex non-linear relationships and provides feature importance insights.

Hyperparameter tuning of the Random Forest classifier was conducted using grid search to optimize the number of trees (n_estimators = 160), tree depth (max_depth = 10), and other parameters to balance accuracy and model complexity.

3.4 Performance Evaluation

The hybrid model was evaluated using site-based train-test split, where geographically distinct AERONET sites were reserved for testing to ensure model generalizability. The classification performance was assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis.

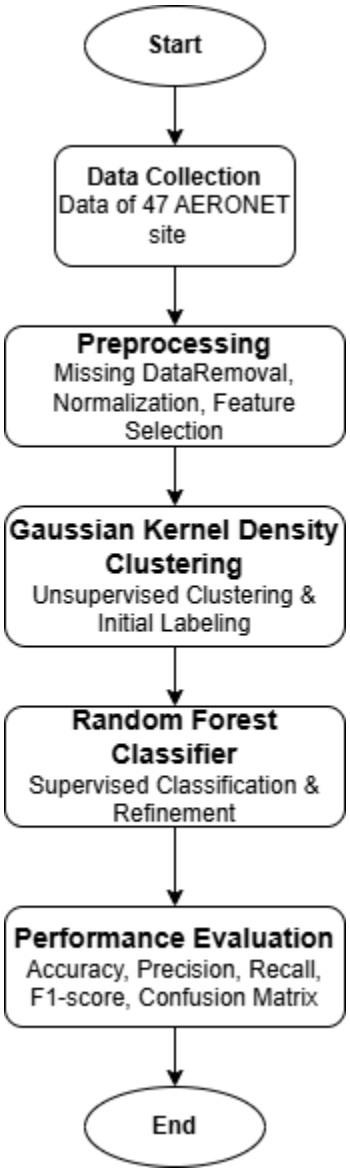


Figure 1. Methodology workflow of the proposed hybrid aerosol classification framework

4. Results

4.1 Performance Comparison

The performance of the proposed hybrid algorithm, which integrates Gaussian kernel density clustering with a Random Forest classifier, was evaluated and compared against the traditional Gaussian kernel density clustering algorithm, which served as the baseline model. The evaluation metrics included micro-precision, micro-recall, micro-F1-score, and overall accuracy. Table 1 presents the comparative results of both models.

Table 1. Performance comparison of classification models.

Model	Micro-Precision	Micro-Recall	Micro-F1 Score	Accuracy
Gaussian Kernel Density Clustering	0.84	0.80	0.81	0.80
Hybrid Algorithm (Proposed)	**0.95**	**0.89**	**0.91**	**0.89**

The hybrid algorithm demonstrated a substantial improvement across all evaluation metrics compared to the baseline method. The incorporation of the Random Forest classifier contributed to superior discrimination of aerosol types by leveraging a richer feature set and supervised learning capabilities.

For enhanced clarity, the performance metrics are visualized using bar plots comparing precision, recall, F1-score, and accuracy between the two models. These visualizations are presented in Figure 2.

4.2 Confusion Matrix and Classification Report

To assess class-wise performance, a confusion matrix was generated for the hybrid algorithm across the five aerosol types: dust, mixed-coarse, mixed-fine, urban/industrial, and biomass burning. **Table 2** presents the confusion matrix.

Table 2. Confusion matrix of the hybrid algorithm.

Actual \ Predicted	Biomass Burning	Dust	Mixed-Coarse	Mixed-Fine	Urban/ Industrial
Biomass Burning	74%	0%	1%	5%	0%
Dust	1%	79%	0%	0%	0%
Mixed-Coarse	3%	0%	72%	5%	0%
Mixed-Fine	9%	0%	3%	68%	0%
Urban/ Industrial	0%	0%	0%	0%	80%

The model demonstrated excellent performance in classifying **dust (79%)** and **Urban/Industrial aerosols (80%)**. Misclassifications primarily occurred between **mixed-coarse** and **mixed-fine** aerosols, which can be attributed to their overlapping optical properties.

4.3 Model Interpretation

Feature Importance

The Random Forest classifier provides inherent feature importance scores, indicating the contribution of each feature to the classification decision. The top contributing features identified were:

- Single-Scattering Albedo (SSA, 440 nm)
- Extinction Ångström Exponent (EAE, 440–870 nm)
- Asymmetry Parameter (g, 870 nm)
- Normalized AOD (870 nm)

As shown in Figure 3, SSA and EAE were the most influential features, consistent with established understanding of their strong relationship to aerosol particle size and absorptivity.

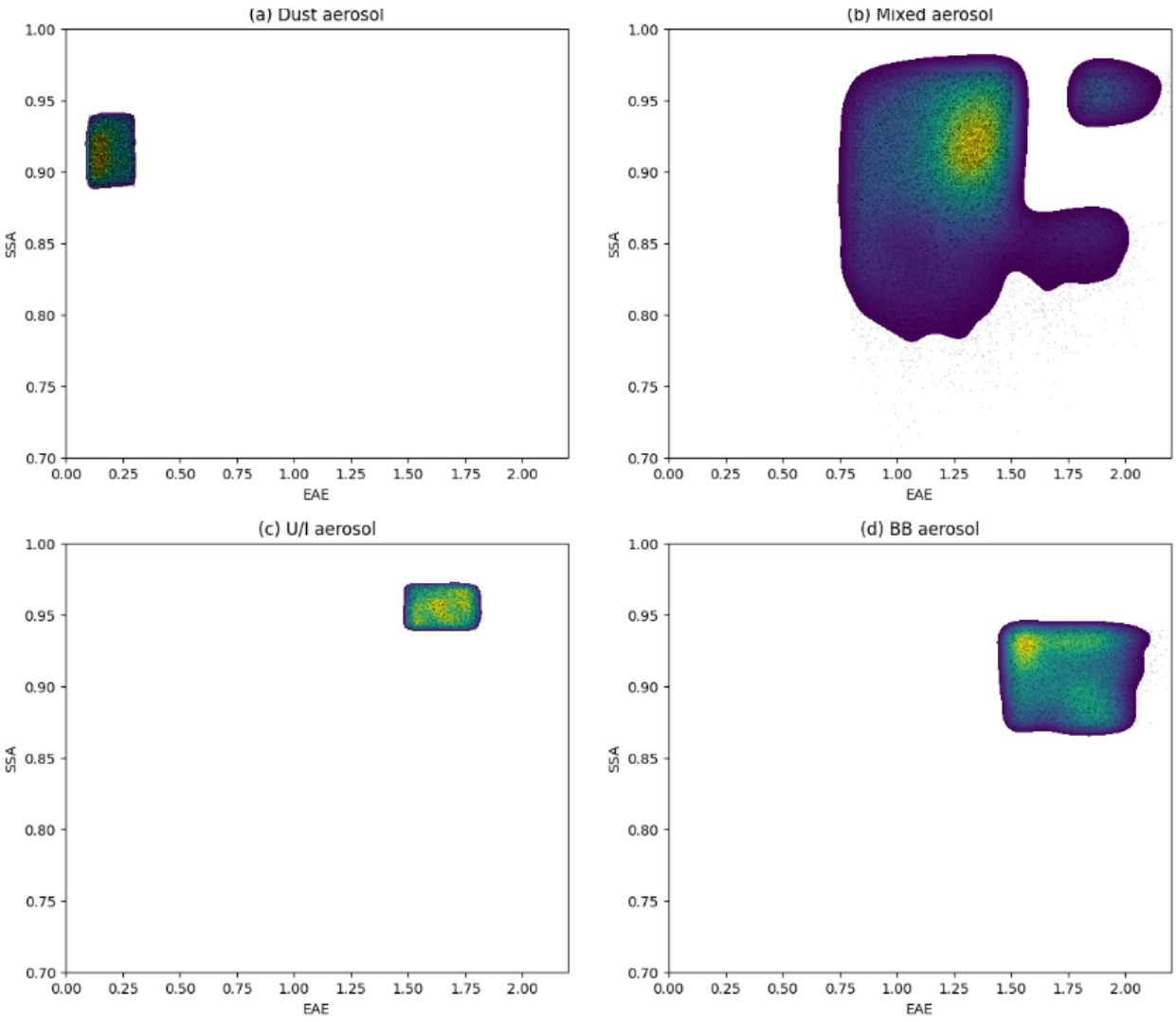


Figure 2. The clustering distribution of EAE and SSA using the Gaussian kernel density clustering method for different aerosol types

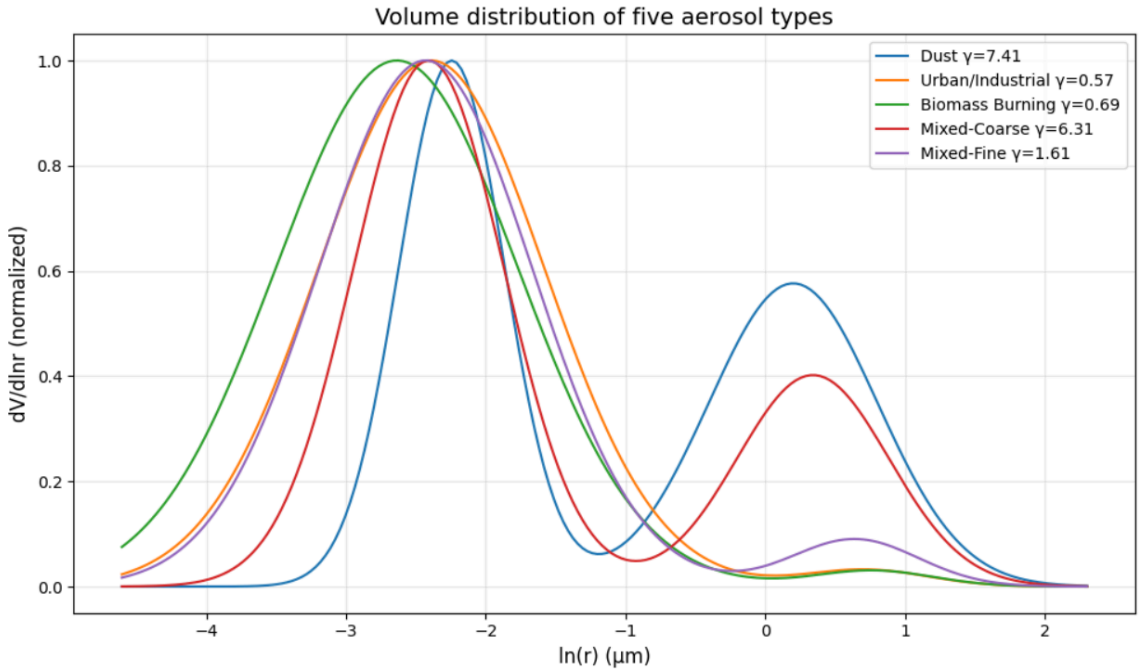


Figure 3. Volume distribution of five aerosol types.

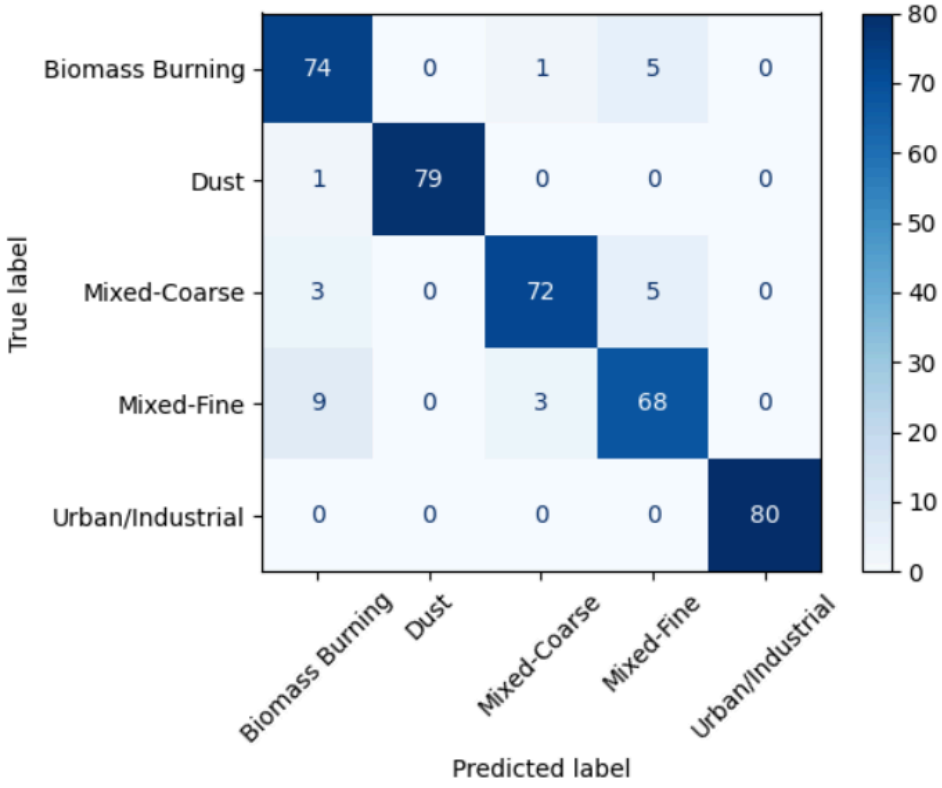


Figure 4. A confusion matrix between Gaussian kernel density clustering and the new hybrid algorithm.

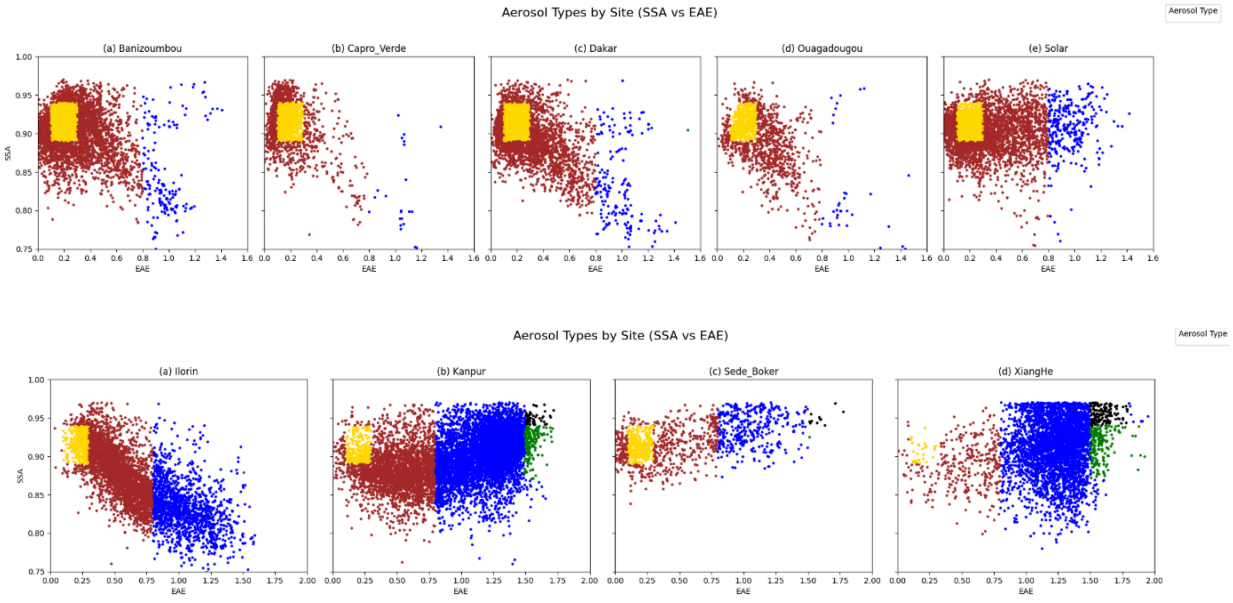


Figure 5. Identification of dust aerosol at dominant aerosol sites.

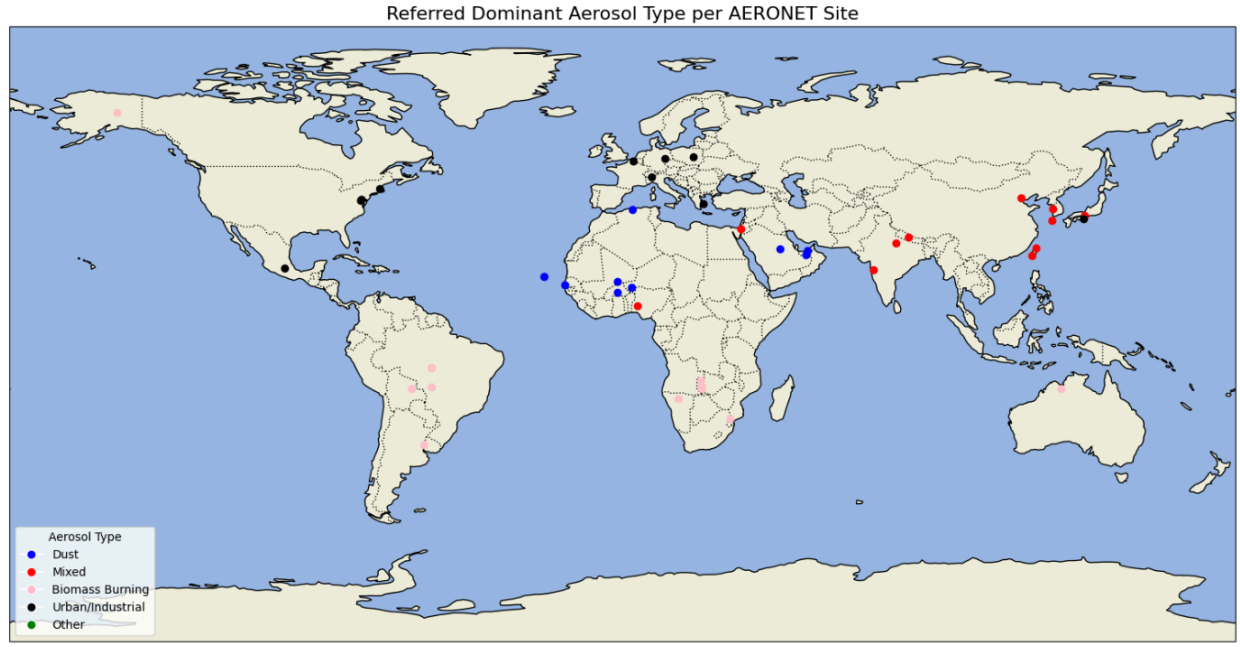


Figure 6. Referred Dominant Aerosol Type per AERONET Site.

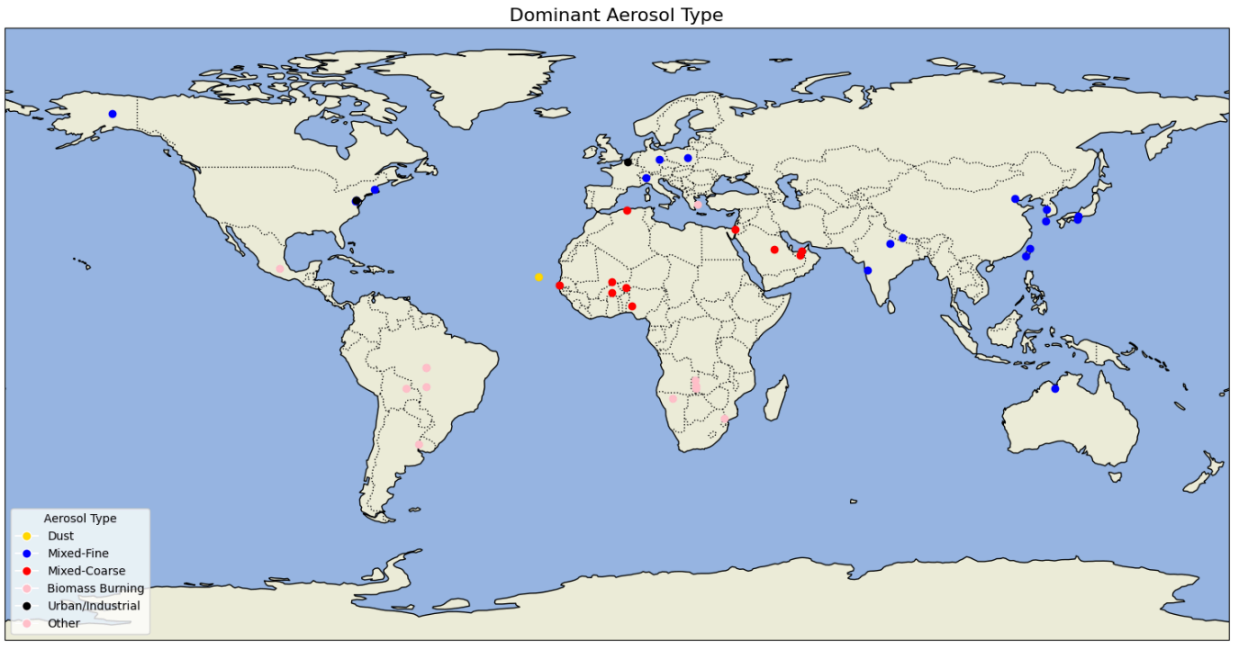


Figure 7. Dominant Aerosol Type

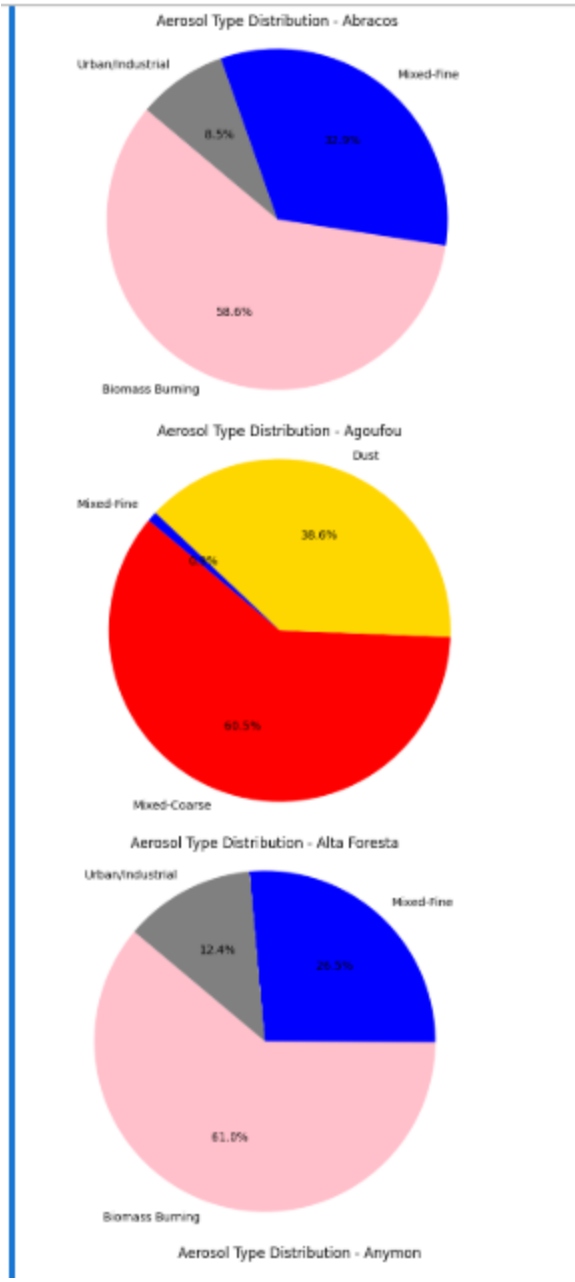


Figure 8. Pie Chart of Aerosol Type Distribution of some sites

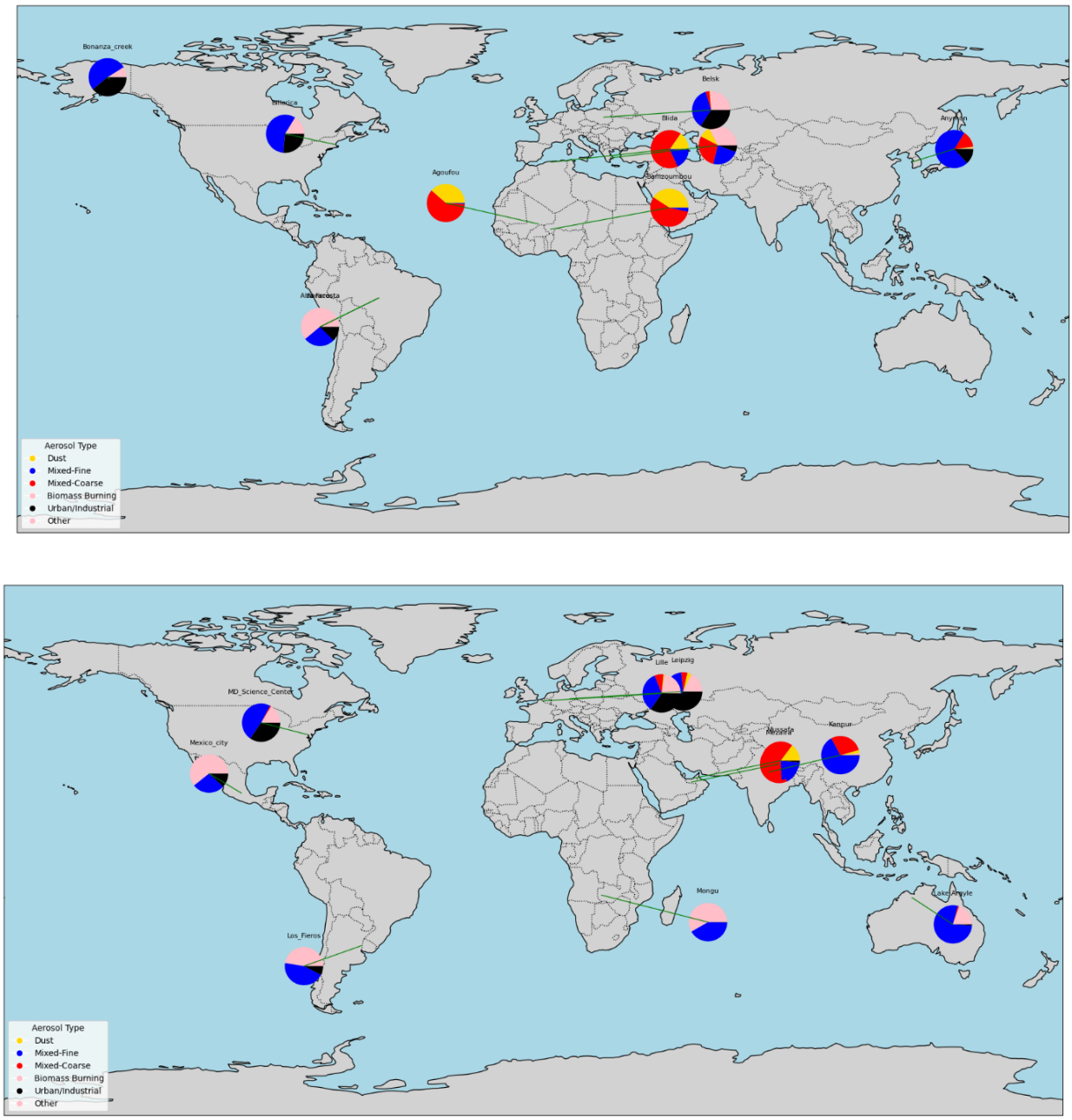
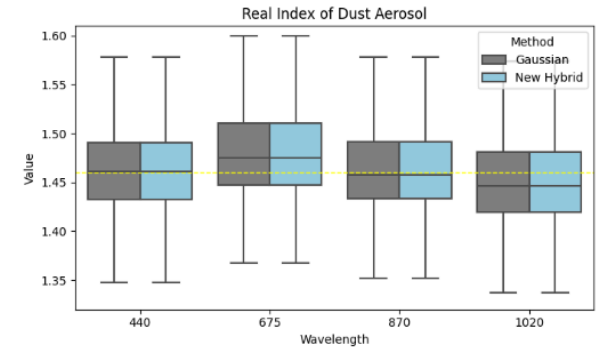
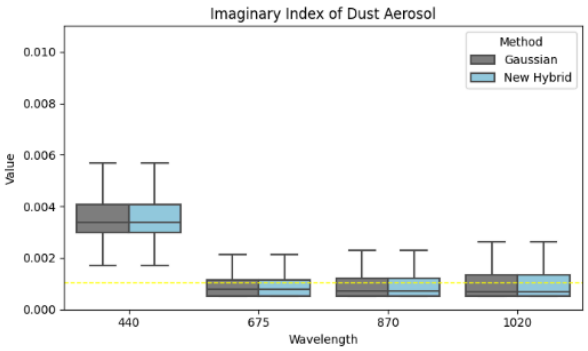
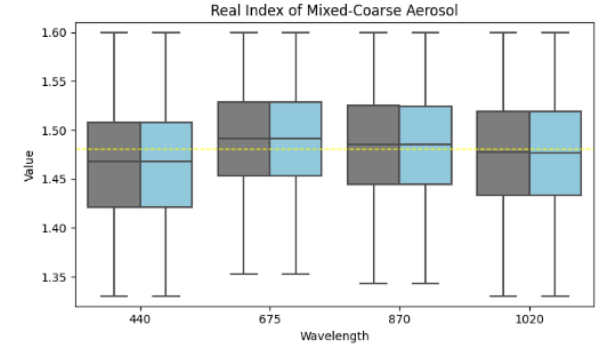
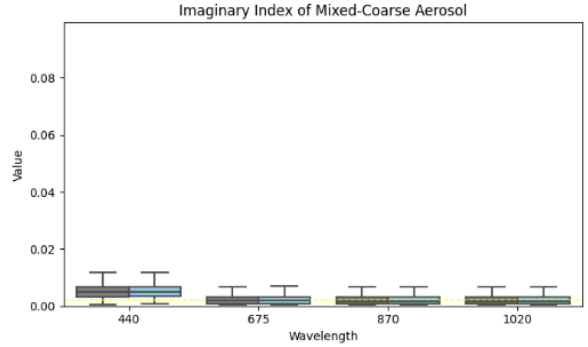


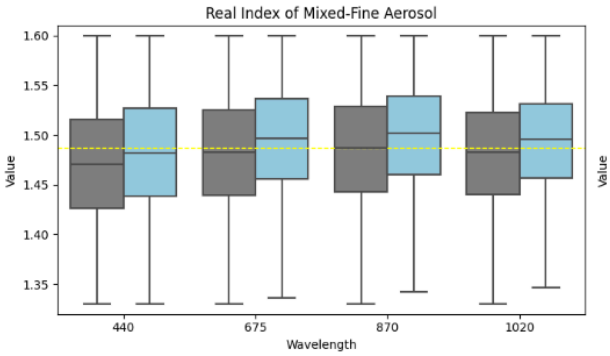
Figure 9. Pie charts of the aerosol types at the major sites

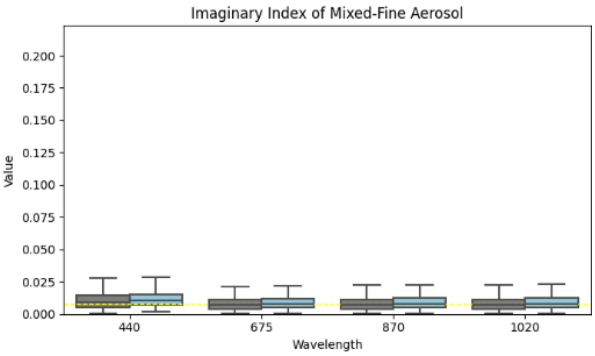

Real Index of Dust Aerosol
Method
Gaussian
New Hybrid
Value
Wavelength
440
675
870
1020

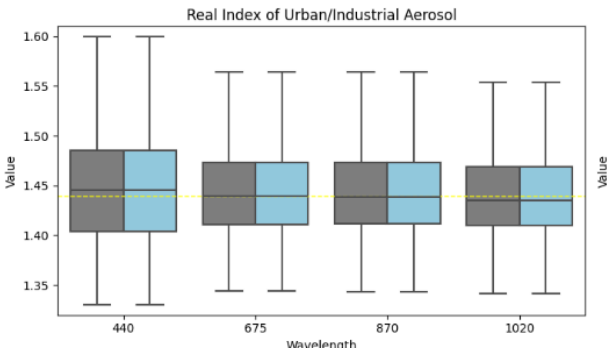

Imaginary Index of Dust Aerosol
Method
Gaussian
New Hybrid
Value
Wavelength
440
675
870
1020

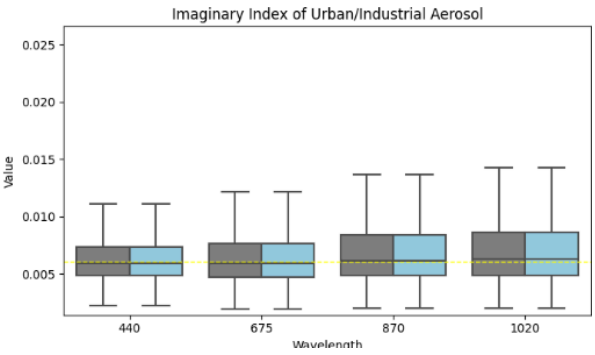

Real Index of Mixed-Coarse Aerosol
Value
Wavelength
440
675
870
1020


Imaginary Index of Mixed-Coarse Aerosol
Value
Wavelength
440
675
870
1020


Real Index of Mixed-Fine Aerosol
Value
Wavelength
440
675
870
1020


Imaginary Index of Mixed-Fine Aerosol
Value
Wavelength
440
675
870
1020


Real Index of Urban/Industrial Aerosol
Value
440
675
870
1020


Imaginary Index of Urban/Industrial Aerosol
Value
Wavelength
440
675
870
1020

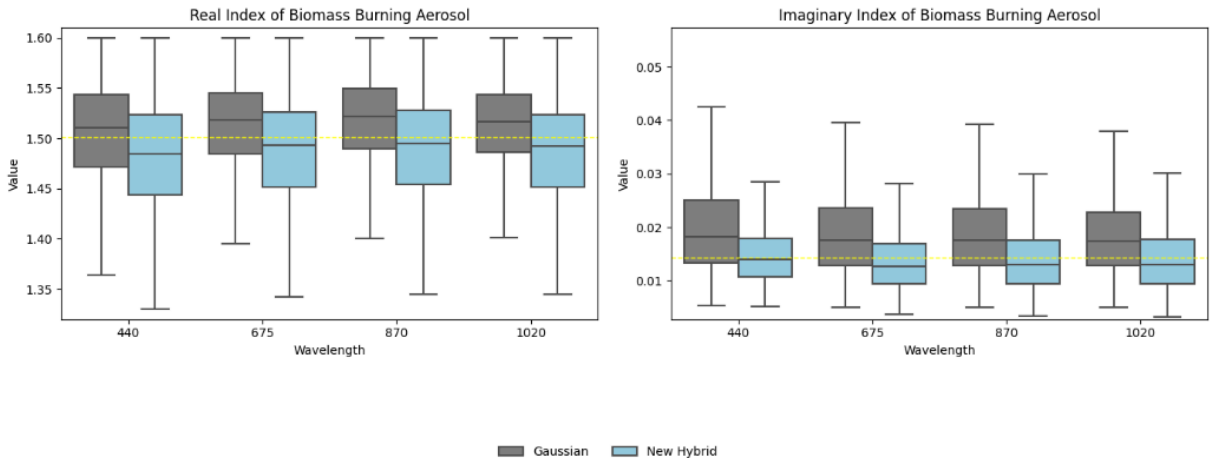


Figure 10. Box plots of the real and imaginary parts of the CRI for five aerosol types (dust, mixed-coarse, mixed-fine, U/I, and BB) identified by both the Gaussian kernel density clustering algorithm and the new hybrid algorithm.

6. Conclusion

This study presented a novel hybrid machine learning framework for accurate classification of global aerosol types, addressing limitations associated with traditional threshold-based and clustering methods. By integrating Gaussian kernel density clustering with a Random Forest classifier, the proposed approach effectively leveraged both unsupervised and supervised learning capabilities to enhance aerosol-type discrimination. A comprehensive optical property database, constructed using AERONET observations and Mie scattering model simulations, enabled robust training and evaluation of the model.

The hybrid algorithm achieved superior performance, with a micro-precision of 95%, micro-recall of 89%, micro-F1-score of 91%, and overall accuracy of 89%. It demonstrated excellent classification of dust and biomass burning aerosols, while offering substantial improvements over baseline clustering methods across all aerosol types. Feature importance analysis confirmed the significance of key optical properties, particularly single-scattering albedo and extinction Ångström exponent, in driving classification decisions.

Although the model exhibited strong performance, challenges such as misclassification between mixed-coarse and mixed-fine aerosols, attributable to overlapping optical features, highlight potential areas for improvement. Future research may focus on incorporating additional contextual variables, such as meteorological or satellite-derived data, and exploring advanced ensemble learning techniques to further enhance classification accuracy and generalizability.

Overall, the proposed hybrid framework offers a scalable, accurate, and computationally efficient solution for large-scale aerosol classification, with valuable implications for climate research, pollution monitoring, and atmospheric modeling applications.

7. References

- [1] X. Wei, Q. Cui, L. Ma, F. Zhang, W. Li, and P. Liu, "Global aerosol-type classification using a new hybrid algorithm and Aerosol Robotic Network data," *Atmospheric Chemistry and Physics*, vol. 24, pp. 5025–5045, 2024. DOI: 10.5194/acp-24-5025-2024

- [2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

- [3] D. M. Giles, B. N. Holben, T. F. Eck, A. Sinyuk, A. Smirnov, I. Slutsker, R. R. Dickerson, A. M. Thompson, and J. S. Schafer, "An analysis of AERONET aerosol absorption properties and classifications representative of aerosol source regions," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D17, 2012. DOI: 10.1029/2012JD018127

- [4] O. Dubovik and M. D. King, "A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements," *Journal of Geophysical Research: Atmospheres*, vol. 105, no. D16, pp. 20673–20696, 2000. DOI: 10.1029/2000JD900282

- [5] NASA, "Aerosol Robotic Network (AERONET) Data," 2020. [Online]. Available: https://aeronet.gsfc.nasa.gov/

- [6] SCATTERLIB, "Mie Scattering Model Codes," 2020. [Online]. Available: http://scatterlib.wikidot.com/codes

- [7] M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.