Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 73 (2024) 474–497

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

Composition and source based aerosol classification using machine learning algorithms

Annapurna S.M *, M. Anitha, Lakshmi Sutha Kumar

Department of Electronics and Communication Engineering, National Institute of Technology Puducherry, Karaikal, India

Received 27 May 2023; received in revised form 16 September 2023; accepted 30 September 2023
Available online 6 October 2023

Abstract

The aerosol particles present in the atmospheric region mainly affect the climate radiative forcing directly by scattering & absorbing the sunlight. Also, it indirectly influences the formation of clouds, precipitation and acts as a considerable uncertainty in assessing Earth's radiation budget. Determination of aerosol type is significant in characterizing the aerosol role in the atmospheric processes, feedback, and climate models. This paper proposes two aerosol classification models, one based on the source and another based on the composition, to classify the aerosols using aerosol optical properties. The source based aerosol classification method helps to identify the sources which cause pollution in a particular region. Based on the results, proper control measures can be taken to reduce pollution. The composition based aerosol classification helps to identify the nature of aerosol types, such as absorbing or non-absorbing. This classification helps to study the climate of the Kanpur region. The aerosol data is taken from AERONET (AErosol RObotic NETwork) for the period 2002–2018 for the Kanpur region. The composition based aerosol classification model uses Single Scattering Albedo (SSA), Angstrom Exponent (AE), and Fine Mode Fraction (FMF) parameters to categorize aerosols based on their composition. The source based aerosol classification model classifies the aerosols based on values of AE and Aerosol Optical Depth (AOD) and describes the source of the aerosol particles. Knowledge of aerosol sources and compositions helps execute policies or controls to reduce aerosol concentrations. Machine learning algorithms, Naïve Bayes, K Nearest Neighbor, Decision Tree, Support Vector Machine, and Random Forest are used to validate classification schemes. The performance analysis of machine learning algorithms is compared using ten different metrics, and the results are also compared with the existing aerosol classification models. The results of the classification show that the source based aerosols of the desert and arid background and the composition based aerosols of types, Mixture Absorbing, Coarse absorbing (Dust), and Black Carbon are dominant over the Kanpur region during the study period considered. The Number of non-absorbing (scattering) type aerosols are least in the study region considered during the study period at all the seasons. It is found that the Random Forest and Decision Tree models outperform the other machine learning models considered and the existing classification models in terms of accuracy (99.55 %) and other performance metrics considered.

Keywords: Aerosol classification; Aerosol source; Machine learning; Aerosol optical properties; AERONET

1. Introduction

In addition to harming plants, animals, forests, and water resources, air pollution also contributes to thinning the ozone layer. According to the WHO (World Health Organization), 93 % of children breathe air pollution that comprises Particulate Matter (PM) of less than 10, which leads to chronic disorders (Anitha and Kumar, 2023). Also,

* Corresponding author.
E-mail addresses: smannapurna@gmail.com (S.M Annapurna), anithasekaran9@gmail.com (M. Anitha), lakshmi@nitpy.ac.in (L. Sutha Kumar).

https://doi.org/10.1016/j.asr.2023.09.068
0273-1177/

the observed daily mean $PM_{2.5}$ by our Indian cities is constantly exceeding the standards of nation air quality (Mahesh et al. 2019). Aerosols are particles that are omnipresent in the atmosphere. There are a large variety of aerosol types. Also, their mixing process in the atmosphere is different. The variability in aerosol distribution and the spatiotemporal variability of its properties are proven factors for uncertainty in various climatological studies. Hence, proper knowledge of aerosol optical properties is necessary for climatological studies. These short-lived particles can be highly heterogeneous (Hamill et al. 2016). This poses difficulty in characterizing these particles, which creates difficulty in estimating the Earth's radiative budget (HaoChen et al., 2016) .

A complete understanding of aerosol optical properties can be used to identify, characterize and develop mechanisms to classify aerosol particles. Aerosols can be classified based on their particle composition and also based on their origin. Their sources greatly influence aerosol types (Lee et al. 2010). Aerosols can be naturally occurring or can be created due to some man-made activities. Knowledge about aerosol sources can facilitate the formation of rules and regulations to curb the production mechanism, improving air quality. In-situ measurements have proven excellent at identifying aerosol particles near the surface of the Earth. However, it is difficult and expensive to perform similar measures in the uplifted layers of the atmosphere. Hence, remote sensing instruments like sun photometers and spectrophotometers monitor aerosols (Siomos et al. 2020).

Recently, machine learning algorithm has been widely used to develop a model for prediction and classification purposes. It mainly uses input variables for training and testing the models. The uncertainties associated with the input variables may cause the misclassification of aerosols in the threshold-based method. In contrast, the classification based on machine learning algorithms can provide better aerosol types than threshold-based schemes. This paper uses five machine learning algorithms such as Naïve Bayes, K Nearest Neighbor, Decision Tree, Support Vector Machine, and Random Forest for composition and source based aerosol classification scheme. The literature review of related previous works is summarized next.

(Hamill et al. 2016) presented an aerosol classification scheme based on Mahalanobis distance calculation with 190 AERONET (AErosol RObotic NETwork) sites data from 1993 to 2012. For instance, it uses microphysical and optical aerosol properties such as Extinction Angstrom Exponent (EAE), Absorption Angstrom Exponent (AAE), complex refractive index, and SSA. These parameters are measured from the electromagnetic spectrum- visible region to classify aerosols in the atmosphere using a sun photometer. The categorized aerosol types are Biomass Burning, Industrial, Mixed Aerosol, Urban, Maritime, and Dust.

A new aerosol classification scheme was introduced based on the FMF and SSA properties (Lee et al. 2010). It uses the dominant size mode and radiation absorptivity of atmospheric aerosol to classify particles into absorbing, non-absorbing, fine, and coarse mode aerosols for the four AERONET sites such as Alta Floresta, Beijing, GSFC, and Agoufou. The investigation shows that the aerosol types are primarily affected by their source and partly influenced by relative humidity. The urban/industrial aerosol with greater absorptivity properties is higher in Asia and Central America than in Europe and North America.

(Siomos et al. 2020) introduced a new aerosol classification procedure based on measuring a double monochromator Brewer spectrophotometer from 1998 to 2017 for Thessaloniki in Greece. This scheme uses the decision tree clustering method for aerosol classification, and Mahalanobis distance is used as a metric. The output aerosol types are UV Single Absorbing Mixtures (FNA): 64.7 %, Black Carbon Mixtures (BC): 17.4 %, Mixed: 9.8 %, and Dust Mixtures (DUST): 8.1 %. The determination is enhanced using CIMEL sunphotometer measurement as the training dataset for Brewer spectrophotometer estimation. The Mahalanobis algorithm clustering potential shows a higher typing score than manually classified types.

The study by (Christopoulos et al. 2018) focuses on using machine learning algorithms to create classifiers that automatically differentiate particles based on their chemistry and size. Single Particle Mass Spectrometry (SPMS) data produces the classifiers. The algorithms build a predictive model using a training set where the aerosol type associated with each mass spectrum is known. It primarily uses random forest for feature selection to reduce dimensionality and evaluates the trained models using the confusion matrices. The models were able to differentiate 20 unique aerosol types, as well as classify aerosols within four broader categories such as fertile soils, mineral/metallic particles, biological particles, and all other aerosols. The study achieved a classification accuracy of approximately 93 % for the broad categorization and 87 % for specific type classification.

Additionally, the trained model was applied to a "blind" mixture of aerosols with agreement found in the presence of various aerosols. It demonstrates the potential of combining online analysis techniques, such as SPMS, with machine learning to infer the behavior and origin of aerosols in laboratory and atmospheric settings. Overall, the study highlights the capabilities of random forests in capturing minor compositional differences between aerosol mass spectra, which can aid in better understanding and predicting the behavior of aerosols.

(Li et al. 2021) applied the K-means algorithm to classify global aerosol regimes based on seven primary aerosol properties simulated with the EMAC-MADE3 global aerosol model. The properties are the mass concentration of black carbon, mineral dust, sea salt, particulate organic matter, the sulfate/nitrate/ammonium system, and the aerosol number concentrations of the Aitken and accumulation modes. The algorithm identified different aerosol clusters characterized by emissions from biomass burn-

ing/biogenic sources, mineral dust, anthropogenic pollution, and their mixing. The identified groups also showed other spatial distributions and internal properties across different altitudes. Furthermore, the study proposed potential future applications of this classification scheme, including identifying model biases, supporting satellite retrieval processes, and aiding campaign planning. Overall, the K-means algorithm was a powerful tool in identifying and quantitatively defining global aerosol regimes without requiring prior classification knowledge.

(Gharibzadeh et al. 2018) They analyzed the seasonal classification of aerosol types in the atmosphere of Zanjan, Iran, using AERONET data from 2010 to 2013. It was found that AOD exhibited seasonal variations with a summer high and winter low. At the same time, the AE had higher values in fall and Winter, indicating the presence of fine particles, and lower values in spring and summer, indicating the existence of coarse particles. SSA variations revealed scattering aerosols like dust in spring, summer, and fall, while the dominance of absorbing-type aerosols in Winter was also observed. This study classified different aerosol types by analyzing various aerosol optical properties such as AOD versus AE, EAE versus SSA, EAE versus AAE, FMF AOD versus EAE, and SSA versus FMF AOD. The results showed the presence of dust and polluted dust in spring, summer, and fall, urban/industrial aerosols in all seasons, especially in fall and Winter, and mixed aerosols in all seasons over the study location, but no biomass-burning aerosols were found. Finally, the aerosol types obtained from the AERONET data were compared with CALIPSO-retrieved aerosol subtypes' profiles, which revealed the dominance of dust and polluted dust in spring, summer, and polluted dust and industrial smoke during fall and Winter. It also suggests that such research can contribute to filling the scientific gaps that exist in deep and accurate climate studies in the Middle East region.

(Choi et al. 2021) proposes a new machine-learning method for classifying aerosol types based on satellite observations. The method uses a Random Forest (RF) model trained with input variables consisting of satellite data (MODIS, TROPOMI) and a target variable of the AERONET-based aerosol type dataset. The contributions of satellite input variables to the RF-based model were quantified to determine an optimal set of input variables. The new method allows the classification of seven aerosol types: pure dust, dust-dominant mixed, pollution-dominant mixed aerosols, and pollution aerosols (strongly, moderately, weakly, and non-absorbing). The model's performance was statistically evaluated using AERONET data excluded from the model training dataset. Model accuracy for classifying the seven aerosol types was 59 %, improving to 72 % for four types (pure dust, dust-dominant mixed, strongly absorbing, and non-absorbing). The model's performance was evaluated against an earlier aerosol classification method based on the wavelength dependence of SSA and FMF values from AERONET. It demonstrates that the RF-based model is capable of satellite aerosol classification with sensitivity to the contribution of non-spherical particles. The performance of the RF-based model was better than previous threshold-based aerosol classification methods, and it can identify aerosols of up to seven types, with more aerosol types and greater accuracy than previous studies that use input variables similar to those of this study. However, the RF-based method has limitations in detailing aerosol compositions and currently does not classify the sea salt type.

A study was conducted to classify and investigate dominant aerosol types in Asian capital cities using a satellite-based Random Forest (RF) aerosol classification model during 2018–2020, with or without AERONET observations (Choi, Lee, and Park 2021). The study identified four types of aerosols: pure dust, dust-dominated aerosols, strongly absorbing, and non-absorbing aerosols. It was found that pollution-sourced aerosols, particularly non-absorbing aerosols, were predominant in Asian capital cities. However, natural dust aerosols were observed seasonally in East Asia (March-May) and South Asia (March-August). The RF model was found to be an alternative to AERONET observations in providing spatially continuous coverage of aerosol-type information. The study suggested that the RF model can be improved by combining it with other measurement data and numerical models. The study results may help provide climatological input data in the future. Based on the literature review, the problem statement and research gap are explained next.

It is clear from the literature review that most papers classify aerosols based on source or composition. In this paper, two aerosol classification schemes are presented. The classifications used aerosol optical properties (Szkop et al. 2016). Source-based (Urban, Maritime, Desert, Biomass, and Arid) and composition-based (Mixture Absorbing, Dust (Fine Absorbing), Black Carbon Medium, Black Carbon Low, Black Carbon High, Coarse Non-Absorbing, Fine Non-Absorbing, Mixture Non-Absorbing) classification is done, and a link is established between the two. This can help understand the source of a particular aerosol and can, vice versa, determine the composition of Aerosol emitted by a specific source. Also, it is observed from the literature that the aerosol properties obtained from different measurement techniques such as spectrometers (Brewer, CIMEL, Single particle mass), satellite (TROPOMI, MODIS), and EMAC-MADE3 global aerosol model were used for aerosol classification. This paper uses four aerosol properties, AOD, AE, SSA, and FMF, acquired from globally available AERONET for classification based on machine learning algorithms. Any one of the machine learning algorithms was used to perform aerosol classification in the majority of the literature papers. This paper uses five machine learning algorithms such as Naïve Bayes (NB), K Nearest Neighbor (KNN), Support Vector Machine (SVM)), Decision Tree (DT), and Random Forest (RF) for aerosol classification and finds the most suitable machine learning algorithm for the aerosol classification. Mainly, one or two performance metrics (Accuracy, typing

score) were used to validate the performance of the machine learning algorithms in the existing papers.

In contrast, this paper uses many performance metrics such as Sensitivity/Recall, Specificity, Precision, Negative Predictive Value, False Positive Rate, False Negative Rate, False Discovery Rate, Accuracy, F1 Score, and Matthews Correlation Coefficient to identify the best machine learning algorithm for classifying aerosols in terms of source and composition based. Also, the results of this paper are compared with the existing classification schemes. The following section discusses the study area and the AERONET properties used.

2. Materials and methods

This section discusses the study area, dataset used, and methodology for aerosol classification schemes.

2.1. Region of study

Kanpur is an industrial city in India located in the central part of the Ganga basin, the largest drainage basin in the world. It is bordered by Vindhyan-Satpura ranges to the south and the north by the Himalayas, and it is enclosed by two significant rivers, Yamuna, Ganga, and their feeders. Due to the rising profitable growth and urbanization, most of the rural population is shifting to civic regions, specifically in India and, in general, in Asia. As a result of this growing population, land use patterns, as well as the density of industries in the area, have increased. This, in turn, results in increased pollution levels and high aerosol loading in the region (Choudhry et al. 2012). It is because of the dust storms frequently occurring in the IGP region (Sarvan Kumar et al., 2012). Due to its location on the Indo Gangetic Plain (IGP), the aerosol emissions from factories, automobiles, power plants, and biomass burning are the leading causes of the high aerosol loading in the Kanpur region (Raman et al. 2011). The emissions from biomass burning, particularly those that result from the burning of agricultural waste, significantly enhance the aerosol composition in the IGP region (Ojha et al. 2020). This region can be examined in detail to find out the existing aerosol load, and the source of this load can be found to control aerosol levels. The Kanpur AERONET site (longitude 80°20' E and latitude 26°26' N) is located at the Indian Institute of Technology Kanpur campus, which is 17 km away from the center of Kanpur and it is displayed in Fig. 1. This site is selected as it has too higher AOD values and has ten years of data.

2.2. AERONET

NASA developed the ground-based remote sensing system AERONET for aerosol monitoring. The sun-sky radiometer in the AERONET site measures solar radiation at eight wavelengths from 340 nm to 1020 nm (Dey et al., 2005). There are nearly 400 AERONET sites in 50 countries, and India has about 24 AERONET sites (Zheng et al. 2017).

The CIMEL sunphotometer data for Kanpur (longitude 80°20′ E and latitude 26°26′ N) from 2002 through 2018 is provided by the AERONET website (https://aeronet.gsfc.nasa.gov/cgi-bin/webtool_aod_v3), and version 3 level 2 data were considered for analysis. This site is selected because it has higher AOD values and has ten years of data.

Beer's law states that, due to the reduced solar flux effect by scattering and absorbing aerosol particles, the sun-emitted solar radiation is not equal to the sunphotometer-measured radiance (Emetere 2019). To eliminate this effect, Langley extrapolation is used to calculate AOD / AOT (Aerosol Optical Depth /Aerosol Optical Thickness) (Siomos et al. 2020). The properties such as Aerosol optical Depth at 500 nm, Fine Mode Fraction measured at 550 nm, Single Scattering Albedo calculated at 440 nm, Absorption Angstrom Exponent measured at 440–870 nm, Angstrom Exponent calculated at 440–870 nm, Refractive index (real and imaginary) at 440 nm, and Extinction Angstrom Exponent measured at 440–870 nm is used in this analysis. The explanation of all the properties is given next.

2.3. Aerosol optical properties

2.3.1. Extinction coefficient (E)

It denotes the loss of incident radiation energy due to the integrated effect of absorption and scattering. In other words, it is the fractional exhaustion of the solar radiance per unit path length (at radar frequencies, it is also called attenuation). The effect of extinction (aerosols and molecules) in the atmosphere is described by the Lambert-Beer law (Sharma et al. 2011). The following equation provides the relationship between I_o (the intensity of radiation at the source) and I (the observed intensity).

$$I = I_o e^{-El} \tag{1}$$

Where l is the length of the medium, the unit of Extinction coefficient is the inverse distance.

2.3.2. Aerosol Optical Depth (AOD)

A fundamental and most defining parameter for Aerosol is AOD. It measures desert dust, urban haze, sea salt, and smoke particles at the top of the atmosphere. The optical depth represents the total quantity of light energy removed by absorption or scattering while traveling through an air medium. Higher AOD values in a region indicate the presence or loading of more aerosols. An AOD of 0.4 would indicate a relatively hazy atmospheric environment, whereas an AOD of 0.01 would indicate an apparent atmosphere (Anitha and Kumar, 2020). As the aerosol concentration increases, the AOD value increases (Siomos et al., 2020). AOD is the most required parameter for estimating Direct Aerosol Radiative Forcing (DARF).

Fig. 1. Location of AERONET study site.

2.3.3. Single Scattering Albedo (SSA)

A factor that defines the scattering and absorbing nature of the aerosol particle is called SSA and is used as an influential phenomenon to assess the various climatic effects caused by aerosols (Siomos et al., 2020). It denotes the aerosol absorptivity property and is given by

$$SSA = \frac{Scattering}{Scattering + Absorption} \quad (2)$$

The SSA value of zero denotes the presence of absorbing nature particles, whereas a value of one denotes the rise of scattering (non-absorbing) particles (Hyvarinen et al. 2009). This study uses the SSA data calculated at 440 nm, with values ranging from 0.5548 to 0.9946. Since the data range is in higher values (close to 1), it denotes the presence of scattering particles at the region of study.

2.3.4. Refractive Index-Real Part

Refractive Index is the most complex quantity in aerosol optical properties since it holds real and imaginary parts to determine aerosols' scattering and absorbing effects (Siomos et al., 2020). The increase in the real part-refractive index denotes the increase in scattering particles in the atmospheric space (Tariq and Ali, 2015). This analysis uses the Real index of refraction data measured at 440 nm, which ranges from 1.33 – 1.57305.

2.3.5. Refractive Index-Imaginary Part

The refractive index imaginary part (IR-Img) and the real part of the refractive index are interdependent, and the linear increase of IR-Img denotes the presence of absorbing nature aerosol particles (Siomos et al. 2020). This analysis uses the imaginary index of refraction data measured at 440 nm, which ranges from 0.000502 – 0.02 0874. These low values hint at the possibility of non-absorbing particles (Tariq and Ali, 2015). This outcome produces a good agreement with the analysis of SSA data.

2.3.6. Fine Mode Fraction (FMF)

The FMF is the ratio between total AOD and the fine mode AOD measured at 550 nm wavelength (Siomos et al. 2020). Generally, these two aerosol products are

not available on the AERONET website. AERONET data provides FMF at 500 nm (Zheng et al. 2019). Therefore, in this paper, FMF at 550 nm is obtained using the interpolation method. To calculate FMF at 550 nm, a second-order polynomial fit is applied to the logarithmic scale offline mode, and total AODs are measured at 440 nm, 675 nm, 870 nm, and 1020 nm. These wavelengths are selected because they have extended data availability periods (Folonchyk et al. 2019). FMF is used as the indicator for aerosol size determination. A value of 1 denotes the dominance of fine aerosol mixtures, while zero represents the rise of coarse aerosol mixture particles. FMF data range of 0.0851 to 0.999 is used in this analysis. It is found from the Kanpur FMF data that the FMF values are equally distributed in all range values and can lead to a large variety of classification flags.

2.3.7. Angstrom Exponent (AE)

AE is an exponent that represents the AOD spectral dependence with the wavelength of incident light (Lee et al. 2010), and it is denoted by

$$\tau = \beta\lambda^{\alpha} \tag{3}$$

Where τ the AOD value at wavelength is λ, α is the Angstrom Exponent parameter, and β is acting as an equating factor called turbidity coefficient (usually indicating the AOD value at 1 μm). It measures columnar aerosol loading (Babu et al., 2013). AE indirectly calculates the aerosol size distribution. When the α value is close to zero, it denotes the domination of coarse particles, and for α near 2, fine particles are most dominant. AE can be used to calculate AOD at another wavelength using the relation.

$$\tau(\lambda) = \tau(\lambda_o)\frac{\lambda}{\lambda_o}^{\alpha} \tag{4}$$

Angstrom Exponent (AE) can be calculated using the below formula for any region:

$$AE = -\frac{\ln\frac{\tau_{AE_2}}{\tau_{AE_1}}}{\ln\frac{\lambda_2}{\lambda_1}} \tag{5}$$

Where a τ_{AEl} and τ_{AE2} are the AODs measured at wavelengths λ_1 and λ_2 (Laakso et al. 2006). From equation (5), on the logarithmic scale, AE is the negative of the first wavelength derivative versus τ or the negative of the slope. As particle size increases, the AE value decreases. The values of AE near 0 are for coarse-sized particles such as soil pollutants, while the value from 1 to 3 is for fine anthropogenic pollutants. In this analysis, the AE values measured between 440 and 870 nm, which range from 0.0169 to 1.618, are used.

2.3.8. Absorption Angstrom Exponent (AAE)

It is calculated from the direct slope of absorbing aerosol optical thickness to the wavelength function (Siomos et al. 2020). It uses aerosol absorption optical thickness (AAOT) calculated at 870 and 440 nm for finding the AAE (Cazorla et al. 2013).

$$AAE = -\frac{log(AAOT(870nm)) - log(AAOT(440nm))}{log((870nm)) - log((440nm))} \tag{6}$$

Where AAOT can be calculated using

$$AAOT(\lambda) = (1 - SSA(\lambda)) * AOD(\lambda) \tag{7}$$

The AAE data measured at 440 nm-870 nm, which ranges from values 0.403025——3.85915, is used in this analysis.

2.3.9. Extinction Angstrom Exponent (EAE)

It represents the slope between wavelength function and extinction optical thickness (Extinction = scattering + absorption) (Siomos et al. 2020). EAE is computed based on the parameter aerosol extinction optical thickness (EOT) measured at 440 nm, and 870 nm (Cazorla et al. 2013)

$$EAE = -\frac{log(EOT(870nm)) - log(EOT(440nm))}{log((870nm)) - log((440nm))} \tag{8}$$

This analysis uses the EAE data measured at 440 nm and 870 nm, which ranges from 0.150052 – 2.121572. Both the Angstrom exponents are qualitative properties. Negative AE values are often acquired when measured in the near-infrared and visible (VIS) spectrum. AE is an essential tool for determining the AOD in the shortwave spectral region and also acts as an essential indicator for the size of the aerosol particle. If AE $\leq$ 1, it denotes the size distribution caused by coarse mode aerosols with radii $\geq$ 0.5 μm (e.g., dust and sea salt), and the values of AE $\geq$ 2 represent the presence of fine mode aerosols (radii $\leq$ 0.5 μm) such as particles arises due to urban pollution and biomass burning.

3. Aerosol classification methods

Aerosols are widespread in the atmosphere. Their source, composition, and chemical properties and reactions in human and plant bodies differ. These particles affect animal health and the Earth's atmosphere. Aerosols also cause visibility degradation in cities and debase the work of archaeological monuments. Therefore, proper classification of aerosols is necessary as this can be useful in backtracking the source of aerosols and in implementing rules and regulations to limit further emissions. The appropriate composition of aerosol can help take precautions for human health. Aerosol classification is done using aerosol parameters.

In this paper, the simulations are done using Matlab and Python programming. The 3D plots are plotted using the R programming language. It is found from the previous literatures that, aerosol parameters such as AE, FMF, AOD, and SSA are most significant as compared to other features like RI (real and imaginary), EAE, and AAE. Therefore, a classification model based on a threshold using these critical features assign classification flags to the data. Table 1 shows the aerosol optical properties of AERONET data from 2002 to 2018 at Kanpur.

Table 1
Aerosol optical properties of AERONET data from 2002 to 2018 at Kanpur.

Year	Value	AOD	SSA	FMF	AE
2002	Average	0.606	0.899	0.633	0.931
	Minimum	0.375	0.846	0.292	0.224
	Maximum	0.92	0.938	0.958	1.4
2003	Average	0.649	0.914	0.671	0.957
	Minimum	0.416	0.896	0.212	0.159
	Maximum	1.105	0.94	0.948	1.364
2004	Average	0.646	0.886	0.613	0.847
	Minimum	0.327	0.876	0.225	0.171
	Maximum	0.958	0.906	0.954	1.357
2005	Average	0.584	0.92	0.614	0.936
	Minimum	0.456	0.896	0.291	0.357
	Maximum	0.767	0.957	0.938	1.374
2006	Average	0.625	0.91	0.571	0.943
	Minimum	0.317	0.88	0.221	0.237
	Maximum	0.996	0.928	0.956	1.374
2007	Average	0.702	0.901	0.884	1.363
	Minimum	0.657	0.897	0.783	1.333
	Maximum	0.734	0.907	0.946	1.394
2008	Average	0.729	0.911	0.696	1.037
	Minimum	0.434	0.897	0.318	0.434
	Maximum	1.056	0.925	0.963	1.398
2009	Average	0.595	0.919	0.604	0.918
	Minimum	0.413	0.903	0.307	0.406
	Maximum	0.858	0.954	0.917	1.324
2010	Average	0.676	0.921	0.669	0.974
	Minimum	0.424	0.889	0.249	0.244
	Maximum	0.889	0.945	0.943	1.338
2011	Average	0.677	0.918	0.722	1.104
	Minimum	0.431	0.888	0.366	0.536
	Maximum	1.065	0.937	0.946	1.433
2012	Average	0.672	0.93	0.611	0.908
	Minimum	0.54	0.902	0.229	0.255
	Maximum	0.922	0.976	0.936	1.348
2013	Average	0.678	0.906	0.759	1.174
	Minimum	0.331	0.883	0.404	0.614
	Maximum	1.021	0.919	0.954	1.415
2014	Average	0.666	0.926	0.688	1.012
	Minimum	0.429	0.897	0.456	0.699
	Maximum	0.999	0.975	0.92	1.336
2015	Average	0.667	0.934	0.708	1.053
	Minimum	0.388	0.898	0.415	0.637
	Maximum	0.976	0.97	0.942	1.332
2016	Average	0.692	0.934	0.727	1.069
	Minimum	0.447	0.914	0.345	0.491
	Maximum	0.944	0.967	0.914	1.322
2017	Average	0.672	0.926	0.772	1.138
	Minimum	0.377	0.911	0.397	0.578
	Maximum	1.125	0.961	0.958	1.508
2018	Average	0.719	0.933	0.667	1.006
	Minimum	0.428	0.92	0.426	0.621
	Maximum	0.989	0.961	0.953	1.354

Table 1 and Fig. 2 show the year-to-year variation of aerosol properties such as AOD, AE, SSA, and FMF with minimum, average, and maximum values during the study period. Fig. 2(a) depicts the minimum, average, and maximum variation of AOD and AE values from 2002 to 2018. The minimum, average, and maximum AOD values range are 0.317 to 0.657, 0.584 to 0.729, and 0.734 to 1.125, respectively. Similarly, the minimum, average, and maximum AE ranges are as follows: 0.159 to 1.333, 0.847 to 1.363, and 1.322 to 1.508. During 2007, the AOD minimum, AE minimum, and AE average values were very high compared to other years. AOD is an indicator of aerosol concentration in a region. If the AOD value is higher in a place, it indicates a higher aerosol concentration in that region. Therefore, the AOD value is used to confirm and quantify the presence of aerosols. The maximum AOD value is from 0.734 to 1.125, representing the presence of more pollutant particles over the entire study period. AE acts as a quantitative indicator. It signifies the size of an aerosol particle. AE value near zero indicates the presence of coarse particles, whereas near 2 denotes the presence of fine aerosol particles. The overall AE range is from 0.159 to 1.508. So, it indicates the Mixture of fine and coarse mode aerosols present in the Kanpur region.

Fig. 2(b) shows the year-to-year variation of SSA and FMF parameters during the study period. The minimum, average, and maximum SSA values range is 0.846 to 0.92, 0.886 to 0.934, and 0.906 to 0.976, respectively. Similarly, the minimum, average, and maximum FMF ranges are 0.212 to 0.783, 0.571 to 0.884, and 0.914 to 0.963. Similar to Fig. 2(a) plot, this SSA and FMF variation figure also shows that during 2007 FMF (average, minimum) variation was very high as compared to other years. The value of SSA and FMF always lies between 0 and 1. The SSA value of 0 indicates the presence of absorbing particles, whereas the value 1 denotes scattering particles.

Similarly, the FMF value of 0 represents the presence of coarse-sized particles, whereas fine-sized particles are always associated with the FMF value of 1. It can be observed from Fig. 2(b) that the SSA ranges from 0.8 to 1, and it denotes the availability of scattering aerosol particles in the Kanpur region. The overall range of FMF is from 0.212 to 0.963 and suggests the existence of mixed (fine and coarse) particles in the study region, like AE.

Generally, India has four seasons. The seasons are Winter (Dec-Jan-Feb), Pre-Monsoon (Summer) (Mar-Apr-May), Monsoon (Jun-Jul-Aug-Sep), and Post-Monsoon (Oct-Nov). The minimum AOD, AE, SSA, and FMF values mainly occurred from December to July. Summer season holds many of the minimum values for all the years. Similarly, the parameters such as AOD, AE, SSA, and FMF have their maximum values from June to February. Post-Monsoon and Winter seasons show most of the maximum importance for all the years. The FMF and AE values reach their minimum values at the same months in all the years except 2007 and 2014. Similarly, the maximum values for FMF and AE always occurred in similar and adjacent months for most of the years. Detailed explanations of these two aerosol classification schemes are given next.

3.1. Classification based on particle composition

Because of the changes in the sources, aerosols have distinct optical and physical properties depending on the season and region. FMF denotes the fraction of fine mode AOD with the size of aerosol particles smaller than 1 μm.

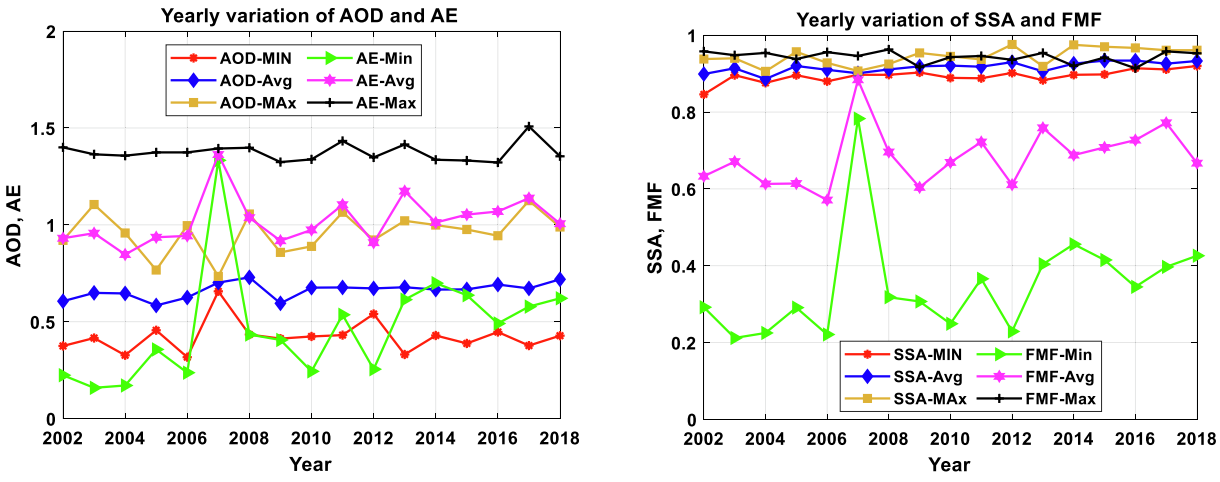


Fig. 2. Year-to-year variation of aerosol properties.

The Angstrom Exponent (AE) uses the exponent of power-law representing the dependence of AOD with wavelength. Using AE and FMF, it is easy to find the dominant-sized aerosol particles. SSA values at Kanpur help to differentiate between non-absorbing and absorbing aerosols. A composition-based aerosol classification was done by (Mohan et al., 2021) using SSA and FMF data. This study analysis uses level 2 inversion products with hourly averaged data from AERONET at Kanpur from 2002 to 2018, including FMF, AE, and SSA data. Following the classification scheme in (Lee et al. 2010), aerosol particles are categorized into eight types, as in Table 2. Fine absorbing and coarse absorbing aerosols are considered Black Carbon (BC) and dust, respectively. Based on (Lee et al. 2010), the definition for each compositional type are given below:

- Coarse non-absorbing: particle radius larger than 2.5 μm and smaller than 10 μm in diameter, which scatter radiation.
- Coarse absorbing: particle radius larger than 2.5 μm and smaller than 10 μm in diameter, which absorbs radiation.
- Mixed non-absorbing: a mixture of various types of aerosols, usually dominated by dust and scatter visible radiation.
- Mixed absorbing: a mixture of various types of aerosols, usually dominated by dust, and absorb visible radiation.
- Fine non-absorbing: Particles of 2.5 μm in diameter and smaller, and these particles scatter visible radiation.
- Black carbon high, medium, and low: these three classes depend on the level of dark carbon particles contained in the particle.

Fig. 3 illustrates the aerosol classification scheme performed in this paper using AE, FMF, and SSA from AERONET at Kanpur. Following this classification scheme, aerosols are classified into eight types, as shown in Table 2, and Fig. 4 shows the corresponding results. The results of the composition-based classification scheme, as explained in Fig. 4, are provided in Table 4, along with the source-based classification scheme, which is described next.

3.2. Classification based on particle sources

Aerosols in the atmosphere are a complex and dynamic mixture of solid and liquid particles from natural and man-made sources. Regular foundation aerosols can be found without human intervention, but anthropogenic sources dominate the urban airborne environment. Primary particles are constantly ejected into the atmosphere in both cir-

Table 2
Threshold of aerosol classification based on aerosol composition.

	SSA	FMF	AE
Coarse non-absorbing	> 0.95	≤ 0.4	≤ 0.6
Coarse absorbing	≤ 0.95	≤ 0.4	≤ 0.6
Mixed non-absorbing	> 0.95	$0.4 \leq$ FMF < 0.6	$0.6 \leq$ AE < 1.2
Mixed absorbing	≤ 0.95	$0.4 \leq$ FMF < 0.6	$0.6 \leq$ AE < 1.2
Fine non-absorbing	> 0.95	> 0.6	> 1.2
Black carbon high	≤ 0.85	> 0.6	> 1.2
Black carbon medium	$0.85 \leq$ SSA < 0.9	> 0.6	> 1.2
Black carbon low	$0.9 \leq$ SSA < 0.95	> 0.6	> 1.2

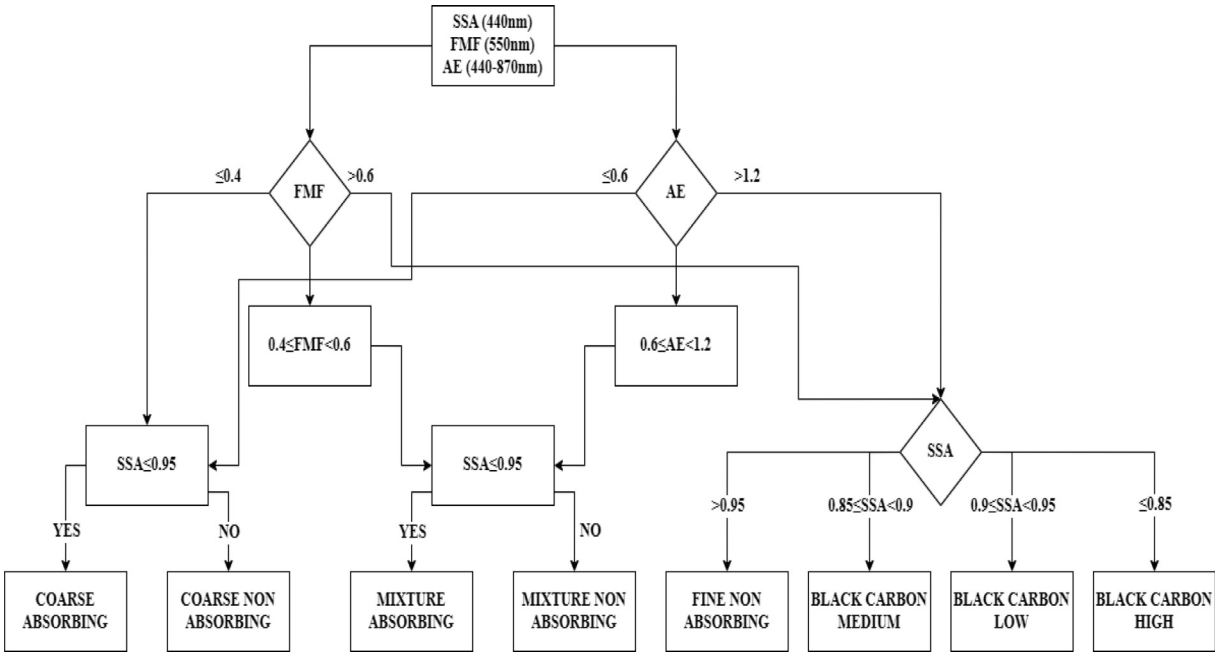


Fig. 3. Flow chart of aerosol classification based on aerosol composition.

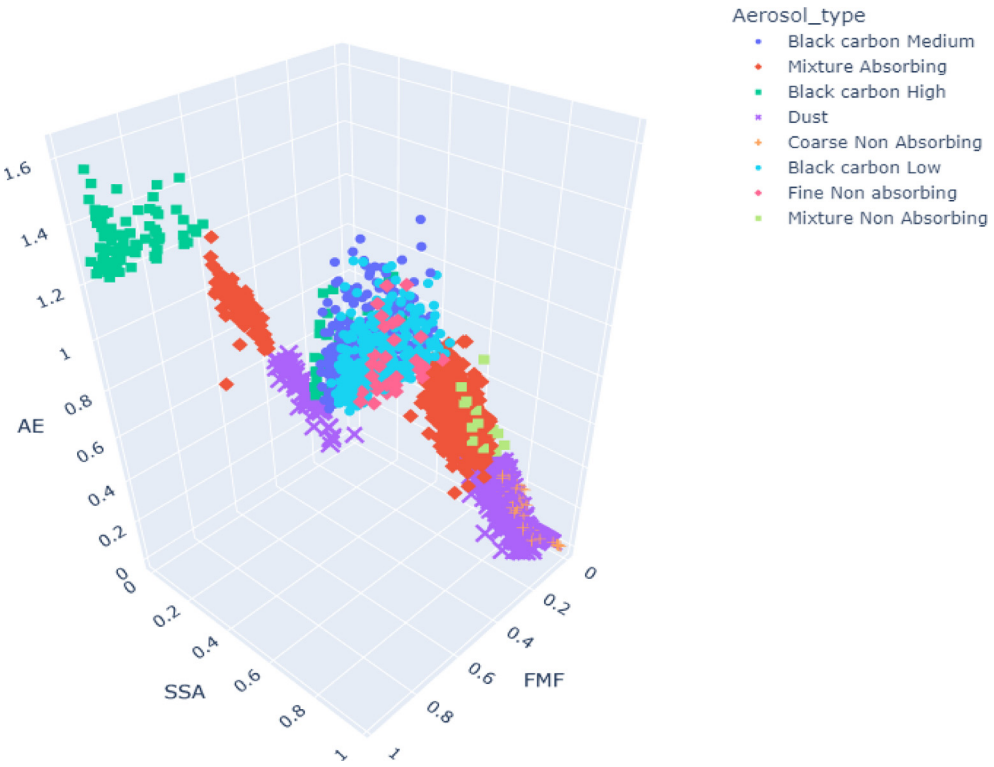


Fig. 4. Composition based aerosol type determination in Kanpur Region for 2002–2018.

cumstances, and chemical reactions produce secondary particles. The air we breathe has an impact on our life. It impacts global weather, local climate, visibility, and personal well-being. Particulate matter from soil dust and volcanic aerosol emissions account for most natural aerosols. The big particles are usually the result of direct anthropogenic pollution. Due to their heavy weight, such particles fall out near the source. The other sources create fine particles with a low mass. Those particles can stay in the air for several days, long enough to travel across continents. Combustion is one of the anthropogenic sources of such particles (Payra et al. 2021).

A threshold-based method has been applied to the data for Kanpur from 2002 to 2018. This is the reference dataset. The threshold values are used for AOD and AE. Five aerosol types are deduced following the classification scheme (Maciszewska et al. 2010). The method is explained in Table 3. AOD value signifies the presence of aerosols and can be used to quantify the concentration of aerosols. AE value is a quantitative indicator. It specifies the size of the aerosol particle. These two parameters can be used to identify the source of the Aerosol. The classified aerosol types are Urban, Maritime, Desert, Biomass, and Arid. The definition for these aerosol types are provided below.

- Maritime aerosols: Originate from sources associated with the marine environment. These aerosols can include sea salt, organic matter from marine organisms, and other materials emitted by the ocean.
- Urban aerosols: Particles present in the atmosphere of cities and metropolitan areas. They arise from a combination of anthropogenic activities, including combustion processes, industrial emissions, construction activities, and traffic-related sources.
- Desert aerosols: Originate from arid and semi-arid regions, such as deserts. They primarily consist of mineral dust, which is composed of fine particles of soil and sand lifted into the atmosphere by wind erosion.
- Arid aerosols: Particles found in regions characterized by low precipitation and limited vegetation cover. These aerosols are primarily composed of mineral dust, but can also include particles from other sources specific to arid climates.
- Biomass aerosols: Particles formed from the combustion or processing of organic materials, such as wood, crop residues, and plant matter. They can be produced through both natural wildfires and human activities like biomass burning for energy or agricultural purposes.

The aerosol types obtained by using the composition type classification scheme are Fine Non-Absorbing Mixtures, Coarse absorbing (Dust), Mixture absorbing, Low Black Carbon Mixtures, Medium Black Carbon Mixtures, High Black Carbon Mixtures, Coarse Non-Absorbing, and Mixture Non-Absorbing. The aerosol classification based on composition types uses the SSA, FMF, and AE properties. After the data is classified into eight different types, scatter plots AOD versus AE are plotted for each class in Fig. 5(a). For example, the AOD and AE values corresponding to the Black Carbon Medium class are shown using the 'x' marker in blue in Fig. 5(a). Fig. 5(b) shows the scatter plots of 5 classes for source based aerosol classification. The parameters AOD and AE used to do the source/origin type classification scheme are plotted in Fig. 5(b) to separate different composition based aerosol classes for comparison. Similar points are compared from these figures. It is clear from Fig. 5(b) that Desert (source based class) has AE and AOD values from 0 to 1.2 and 0.1–2.2, respectively. The composition based class Dust has similar AOD values, but its AE values are 0–0.6. Coarse Non-Absorbing has similar AE values, but AOD values range from 0.1 to 1.4. Mixture Absorbing and Mixture Non-Absorbing have medium AE values. All these four different composition based classes, Dust, Coarse Non-Absorbing, Mixture Absorbing, and Mixture Non-Absorbing, are from Desert (source based) class. Detailed comparisons between the two classifications are provided in Fig. 6, Fig. 7, and Fig. 8 and their explanations. Fig. 5(c) and Fig. 5(d) represent the output class distribution of aerosols. The Number of occurrences of each aerosol type is plotted. In Fig. 5(c) and Fig. 5(d), the Number in the X-axis indicates the Number of occurrences of each kind of aerosol. The most occurring aerosol type for composition-based and source-based aerosol type is Mixture absorbing and desert, respectively.

Table 3
Threshold values for determining source based aerosol type in Kanpur.

AOD(500 nm)	AE(440–870 nm)	Aerosol type
0.2–0.4	> 1	Urban
< 0.3	0.5–1.7	Maritime
> 0.4	< 1	Desert
> 0.7	> 1	Biomass
> 0.45	> 1.2	Arid

Classifying aerosols based on their composition helps identify a region's toxicity levels and air quality. Also, classification based on aerosol source helps determine the primary source producing the aerosol. A proper comparison between the classification schemes helps backtracking (matching composition to source) and forecasting (formulating laws and regulations) of aerosols. The composition of aerosol emitted by a particular source gives an idea about the aerosols and prepares necessary measures to control a specific type of aerosol.

Fig. 6 and Fig. 7 depict the scatter and 3D plots between the two classification schemes. It is found that mixture absorption occurs from Biomass, Maritime, Urban, and Desert. Black carbon (high, low, and medium) occurs from Maritime, Urban, and Arid BG. Dust occurs in the Maritime and Desert. Coarse non-absorbing also occurs in the desert. Fine non-absorbing occurs from Maritime, Urban, and Arid BG. Mixture non-absorbing occurs from Biomass, Maritime, Urban, and Desert.

3.3. Linking aerosol type and source

Knowledge about the source of a particular type of aerosol is necessary for figuring out a region's climatological and geographical changes. A proper comparison between the classification schemes helps backtracking (matching composition to source) and forecasting (formulating laws and regulations) of aerosols. Fig. 8 shows the heatmap of aerosol composition and aerosol source classification scheme. The primary source of each aerosol type is found and mentioned in Table 5.

Table 4
Comparison between classified aerosol types (Source and composition).

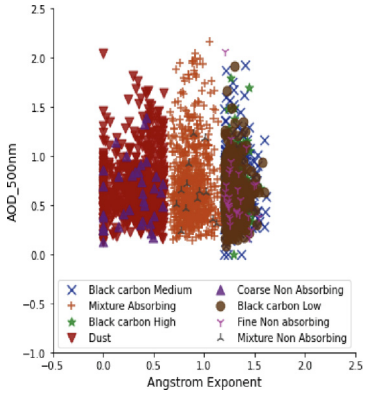
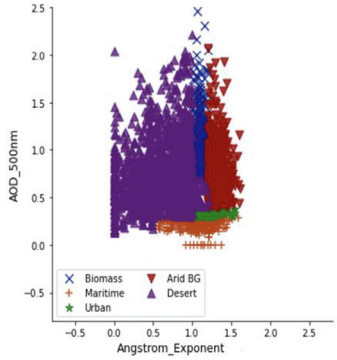
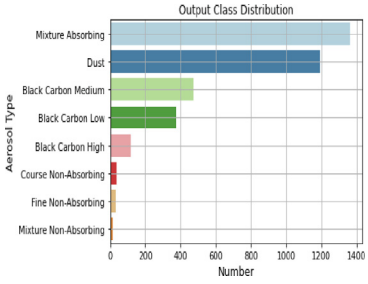
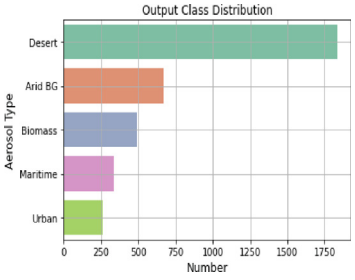
Classification scheme	Composition		Source	
Parameters used	FMF, SSA, AE		AE, AOD	
Classified types	Fine Non-Absorbing Mixtures, Coarse absorbing (Dust), Mixture absorbing, Low Black Carbon Mixtures, Medium Black Carbon Mixtures, High Black Carbon Mixtures, Coarse Non-Absorbing, Mixture Non-Absorbing		Urban, Maritime, Desert, Biomass, Arid BG	
Number of input samples	3604		3604	
Output Distribution	Mixture Absorbing	1365	Desert	1836
	Coarse absorbing (Dust)	1191	Arid BG	671
	Black Carbon Medium	473	Biomass	495
	Black Carbon Low	375	Maritime	340
	Black Carbon High	119	Urban	261
	Coarse Non-Absorbing	36	–	–
	Fine Non-Absorbing	31	–	–
	Mixture Non-Absorbing	14	–	–
Output plots				
	Fig. 5(a) Scatter plot of aerosol types based on composition		Fig. 5(b) Scatter plot of aerosol types based on source	
Output class distribution				
	Fig. 5(c) Output class distribution of aerosol types based on composition		Fig. 5(d) Output class distribution of aerosol types based on source	

Fig. 8 shows the heatmap of aerosol composition and aerosol source classification schemes. Each data class obtained from the composition and source based classification scheme is compared according to the respective date and time values. In the heat map, the Y-axis corresponds to the composition-based classification scheme, and the X-axis corresponds to the source-based composition scheme. A correlation mapping is done concerning the time instance and aerosol properties to show how the classification schemes overlap. As explained before, the four different composition based classes, Dust (64.33 %), Coarse Non-Absorbing (1.96 %), Mixture Absorbing (33.22 %), and Mixture Non-Absorbing (0.49 %), are from Desert (source based) class.

3.4. Seasonal classification of aerosols based on source and composition

The classification results of aerosols based on sources during the Winter, Pre-Monsoon, Monsoon, and Post-Monsoon seasons are shown in Fig. 9(a), 9(b), 9(c), and 9(d), respectively. Table 6 represents the source based aero-

sol type distribution for different seasons. The x-axis in Fig. 9 indicates the Angstrom Exponent measured at 440–870 nm, and the y-axis shows the AOD measured at 500 nm. It is observed from Fig. 9 and Table 6 that there are higher concentration of desert type aerosols (78.60 % and 48.13 %) during the Pre-Monsoon and Monsoon seasons, respectively. The no. of aerosols from the biomass source type is very low in concentration compared to the other classes during the Pre-Monsoon and Monsoon seasons. During the Pre-Monsoon season, the arid background and Maritime aerosols also have significant counts (18.82 % and 15.56 %, respectively). The aerosols coming from the arid background are more, and Maritime aerosols are very low over the Kanpur region during the Winter and Post-Monsoon seasons, respectively. Overall, the desert's aerosols and arid background dominated the Kanpur region during the study period considered.

The variations in the composition based aerosol types during the four seasons of India are shown in Fig. 10. Fig. 10(a), 10(b), 10(c), and 10(d) show the seasonal 3D plots of composition based aerosol types during the Winter, Pre-Monsoon, Monsoon, and Post-Monsoon respectively. The parameters such as FMF, SSA, and AE are used in the respective ×, y, and z axes for creating the 3D plots. Table 7 shows the composition based aerosol type distribution for different seasons. It is clear from Fig. 10 and Table 7 that there are higher concentrations of the Dust (Coarse absorbing) type aerosols during the Pre-Monsoon (48.78 %) and Monsoon (35.81 %) seasons. In contrast, Black Carbon low type aerosols dominate (44.98 % and 47.70 %) during the Winter and Post-Monsoon seasons, respectively. In general, the black carbon type aerosols (Black carbon high, back carbon medium, and black carbon low) have significant counts during the Winter and Post-Monsoon seasons in the study region, which is well agreement with (Thamban et al. 2017). It is found from Fig. 10 and Table 7 that the Number of non-absorbing (scattering) type aerosols are least in the study region considered during the study period at all the seasons.

4. Results

In recent days, a machine learning algorithm has been successful in predictive modeling, product recommendation, sentiment analysis, weather prediction, satellite climate modeling, data processing, and language translation. ML algorithms map independent variables (attributes) to a dependent variable (target). In this paper, the aerosol prediction model was implemented by ML algorithms. The AOD, FMF, SSA, and AE are considered attributes, and the aerosol types based on the sources such as Arid background, Biomass, Desert, Maritime, and Urban are the target class. Five ML algorithms, such as KNN, RF, NB, SVM, and DT, are used for aerosol prediction. The data is split into train/test in the ratio of 7:3. The results and performance of ML algorithms are explained in the upcoming sections.

KNN calculates the Euclidean distance between the test point and all the training instances and assigns the majority class label of the first 'k' distances (k-positive integer). The optimum k-value is selected for good prediction accuracy. RF builds several decision trees on bootstrapped datasets and uses majority voting to find the final prediction. The decision trees in the RF take only a subset of attributes, making each decision tree independent of others and reducing overfitting. NB is the probabilistic ML algorithm based on Bayes Theorem. NB assumes conditional independence and calculates the posterior probability of the target (class) from likelihood and prior probabilities. The likelihood and previous chances are determined for each feature (attributes) against the target (type) from the training dataset. SVM finds the optimum hyperplane to classify all the inputs in a high-dimensional space. In this paper, Radial Basis Function (RBF) kernel function transforms the input feature space into high dimensional feature space to find the linear decision boundary that separates the classes. DT ML algorithm builds the tree-like structure by applying an attribute value test on the source set and subsets in the root node and internal nodes. This recursive partitioning is carried out until the leaf node has the same value as the target class, or splitting does not improve the prediction accuracy (Lal et al., 2011).

The parameter configuration is done after analyzing the error and accuracies for various values of parameters for the machine learning algorithms considered. For KNN algorithm, the number of neighbors (k) values is the key parameter. In this paper, the optimum value of k is taken as 40. For random forest algorithm, the number of estimators is 1000. For Naïve Bayes algorithm, the value of variable smoothing is taken as 10. For the SVM algorithm, the C value is taken as 100. For the Decision tree algorithm, the maximum depth of the tree is taken as 20.

4.1. Confusion matrix

The confusion matrix is a summary of the predictions for the classification problem. The Number of incorrect and correct predictions is summarized by count value and categorized by class. After applying the optimization techniques, the best parameter values for all the algorithms are selected. The confusion metrics and accuracy are generated for the five ML algorithms. This is tabulated in Fig. 9.

The confusion matrix of source and composition based aerosol type classifications for each machine learning algorithm are plotted and shown in Fig. 11 and Fig. 12, respectively. This confusion matrix is generated for the test data to determine how accurately the model will classify the aerosols correctly. The accuracy comparison of machine learning algorithms such as KNN, RF, NB, SVM, and DT are shown in Table 8 for both aerosol classification methods (source and composition). It can be observed from Table 8 that the accuracy of the random forest and

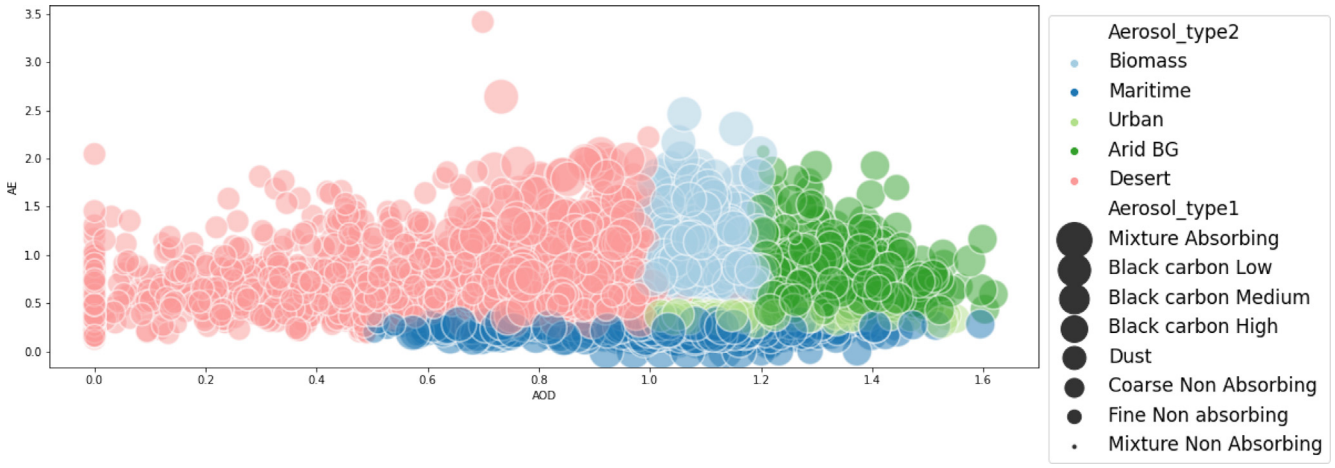


Fig. 6. Scatterplot inclusive of aerosol composition and aerosol source classification schemes.

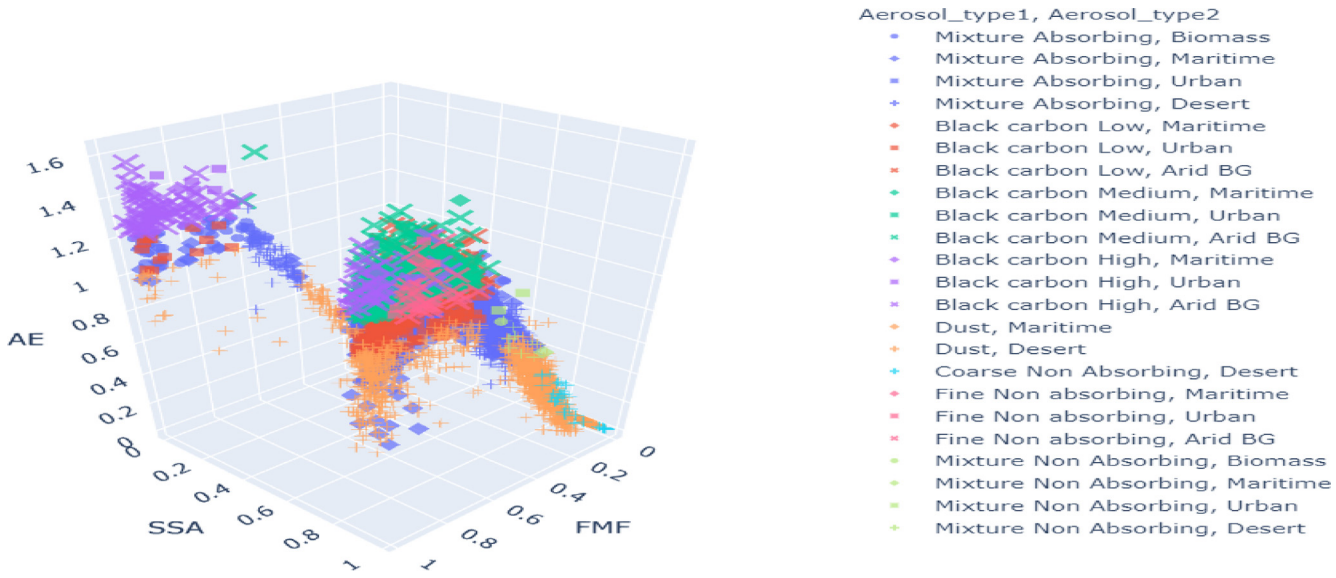


Fig. 7. 3D plot inclusive of aerosol composition and aerosol source classification schemes.

Aerosol type	Desert	Arid BG	Biomass	Maritime	Urban
Mixture Absorbing	610	0	494	245	16
Dust	1181	0	0	10	0
Black carbon Medium	0	309	0	34	32
Black carbon Low	0	244	0	39	190
Black carbon High	0	94	0	9	16
Coarse Non Absorbing	36	0	0	0	0
Fine Non Absorbing	0	24	0	2	5
Mixture Non Absorbing	9	0	2	1	2

Fig. 8. Heatmap of aerosol composition and aerosol source classification scheme.

Table 5
Relationship between Aerosol composition and aerosol source classification schemes.

Aerosol type	Major source(s)
Mixture Absorbing	Desert, Biomass, Maritime
Dust	Desert
Black carbon medium	Arid background
Black carbon low	Arid background, Urban
Black carbon high	Arid background
Coarse non-absorbing	Desert
Fine non-absorbing	Arid background
Mixture non-absorbing	Desert

decision tree methods are high, and NB produce less accuracy as compared to the other machine learning models for both aerosol classification type.

Detailed analysis of the performance of these algorithms is possible only by considering the various performance metrics. This will be done in the upcoming section.

Table 6
Source based aerosol type distribution for different seasons.

Aerosol types	Winter	Pre-Monsoon	Monsoon	Post-Monsoon
Desert	87	***694***	***399***	32
Arid BG	***596***	50	156	***486***
Biomass	107	*16*	*59*	56
Maritime	*68*	87	129	*21*
Urban	90	36	86	40
Total no. of all aerosols	948	883	829	635

4.2. Performance metrics

Different performance metrics are used to assess the classification scheme's performance and compare machine learning algorithms. Table 9 shows the performance of var-

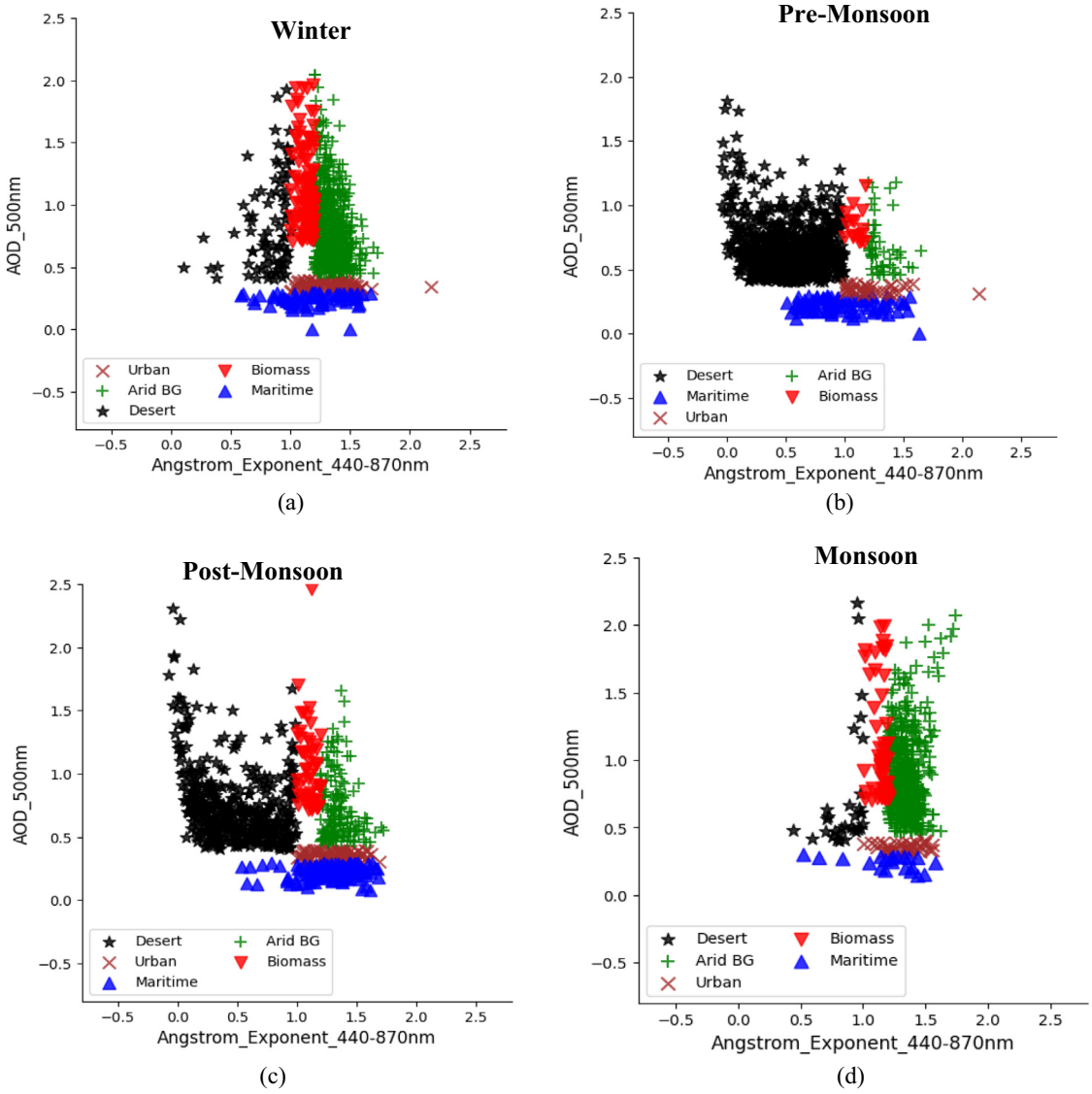


Fig. 9. Seasonal plots of source based aerosol types.

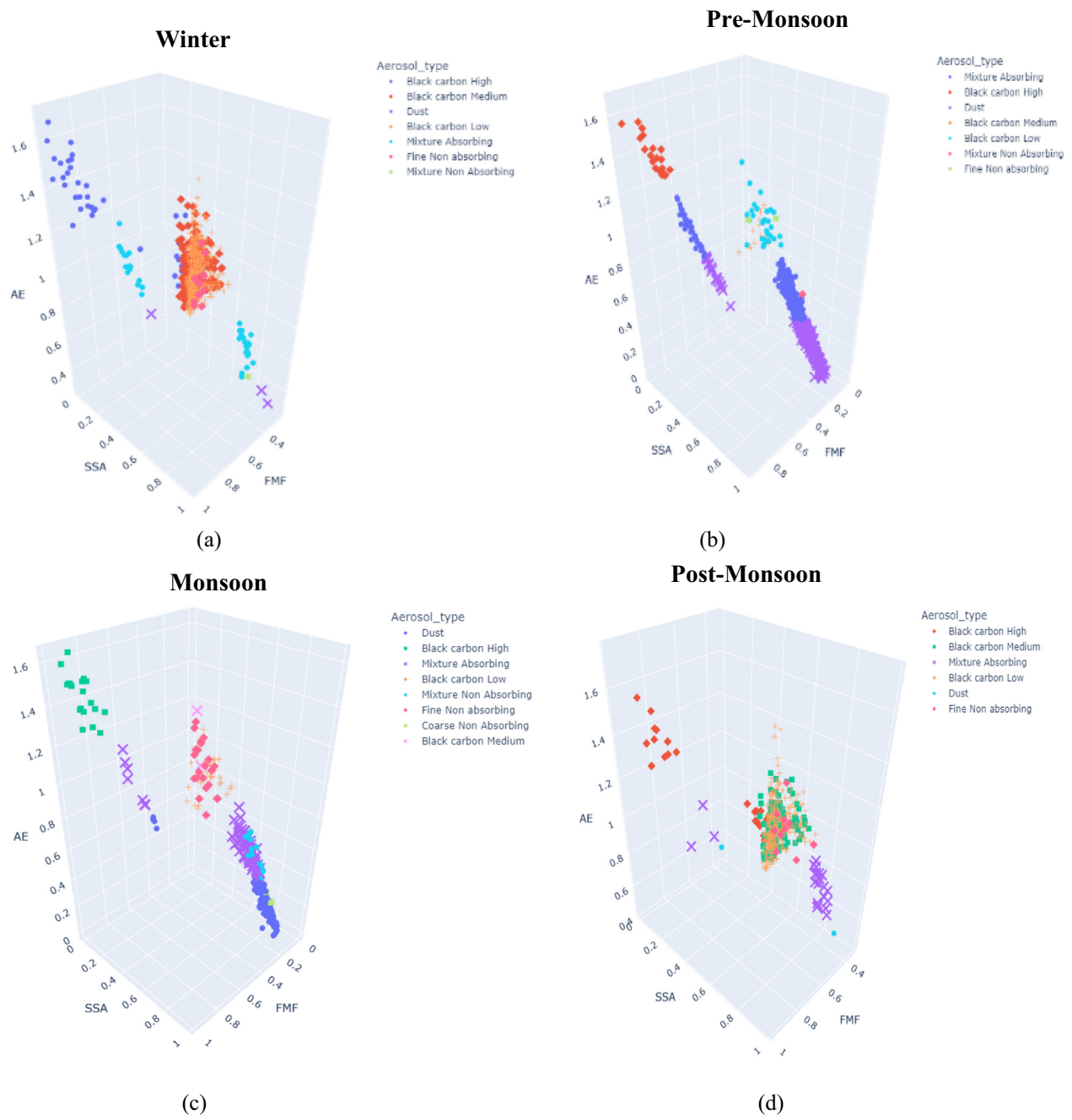


Fig. 10. Seasonal plots of composition based aerosol types.

Table 7
Composition based aerosol type distribution for different seasons.

Aerosol Type	Winter	Pre-Monsoon	Monsoon	Post-Monsoon
Mixture Absorbing	56	387	147	29
Coarse absorbing (Dust)	11	***461***	***260***	3
Black Carbon Medium	132	20	*2*	103
Black Carbon Low	***350***	30	32	***269***
Black Carbon High	214	44	247	150
Coarse Non-Absorbing	*0*	*0*	3	*0*
Fine Non-Absorbing	14	2	23	10
Mixture Non-Absorbing	1	1	12	0
Total no. of all aerosols	**778**	**945**	**726**	**564**

ious machine learning algorithms for the Kanpur data of AERONET.

The bold italics entry denotes the highest value in every row, and the bold value indicates the second-highest value in every row. It can be observed that Random Forest and Decision tree algorithms have the best set of performance metrics, whereas State vector machine and Naïve Bayes have the worst set of performance metrics. The KNN shows the second highest value in every row.

Also, each machine learning algorithm's training and testing stage performance metrics are evaluated and shown in Fig. 13. The value of precision, recall, and accuracy (test, train) is high for both RF and DT. Therefore, it can be concluded that the Decision Tree and Random Forest are the best-suited algorithms for aerosol type classification.

5. Discussion

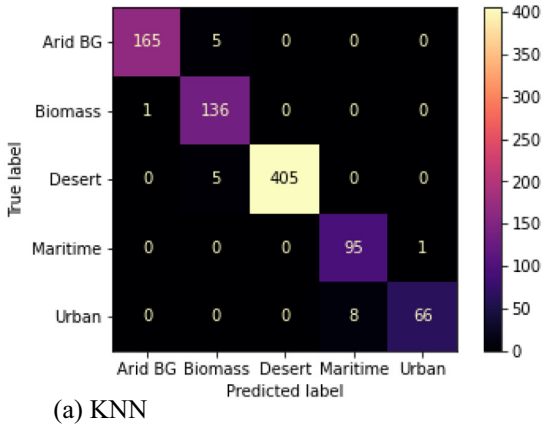
Table 10 provides a comparison of recent existing methods with the proposed method. Christopoulos et al. (2018) use machine learning algorithms to differentiate particles based on their chemistry and size using Single-Particle Mass Spectrometry (SPMS) data. Random forests and feature selection were used to reduce dimensionality and capture minor compositional differences between aerosol mass spectra. The SPMS is located at Karlsruhe Institute of Technology (KIT), Germany, and the classified aerosol types are fertile soils, mineral/metallic particles, biological particles, and all other aerosols. It achieves a classification accuracy of approximately 93 % for broad categorization and 87 % for specific type classification. The trained model was also able to classify a "blind" mixture of aerosols, demonstrating its potential for understanding and predicting the behavior of aerosols.

A machine-learning technique for categorizing different forms of aerosols based on satellite measurements is suggested by (Choi et al. 2021). Using a random forest (RF) model, which was trained using input variables made up of satellite data (MODIS, TROPOMI), the method quantified the satellite input variables to the RF-based model to identify the best input variables. The variables such as PLDR, SSA, AI, CO, NO_2, AOD, AE, TOA reflectance at 412, 470, and 660 nm, Land cover type, Solar zenith angle, and Percent of the urban area obtained from Jan 2018 to July 2020 are used.

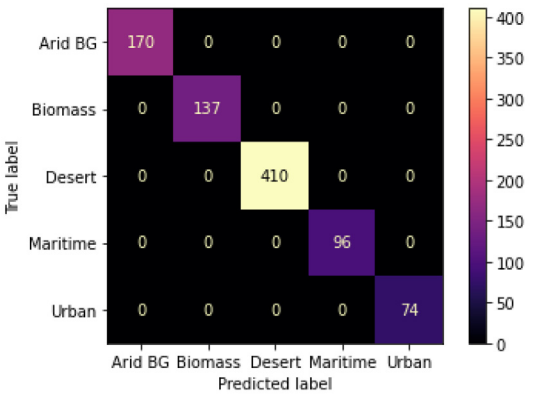
(Hamill et al. 2016) a Mahalanobis distance-based aerosol classification technique was developed using data from 190 AERONET (AErosol RObotic NETwork) locations between 1993 and 2012. For instance, it uses the EAE, AAE, Complex Refractive Index, and SSA Aerosol properties for classification. Using a sun-photometer, these characteristics are measured from the visible portion of the electromagnetic spectrum to categorize aerosols in the atmosphere. Biomass Burning, Industrial, Mixed Aerosol, Urban, Maritime, and Dust are the different categories of aerosols. (Siomos et al. 2020) Developed a novel aerosol categorization method for Thessaloniki, Greece, based on measurements made with a double monochromator Brewer spectrophotometer between 1998 and 2017. This system employs the Mahalanobis distance as a measure, and the decision tree clustering algorithm is used to classify aerosols. UV Single Absorbing Mixtures (FNA) makeup 64.7 % of the output aerosols, followed by Black Carbon Mixtures (BC) at 17.4 %, Mixed at 9.8 %, and Dust Mixtures (DUST) at 8.1 %. Using CIMEL sunphotometer measurement as the training dataset for Brewer spectrophotometer estimation improves the determination. The aerosol properties used from the photometers are AOD, FMF, SSA, and EAE. Compared to manually classified types, the Mahalanobis method clustering potential exhibits a high typing score.

The new method permits the classification of seven aerosol categories, including pure dust, mixed aerosols dominated by dust, polluted aerosols, and pollution aerosols (strongly, moderately, weakly, and non-absorbing). AERONET data that wasn't included in the model training dataset was used to evaluate the model's performance statistically. Model accuracy for identifying the seven categories of aerosols was 59 %, but it increased to 72 % for four types (pure dust, dust-dominant mixed, strongly absorbing, and non-absorbing). Using AERONET data from 2010 to 2013, (Gharibzadeh et al. 2018) examined the seasonal classification of aerosol types in the atmosphere of Zanjan, Iran. It identified distinct aerosol kinds by comparing various aerosol optical properties, including AOD versus AE, EAE versus SSA, EAE versus AAE, FMF AOD versus EAE, and SSA versus FMF AOD. The classified aerosol types are dust, urban industrial, biomass burning, and mixed aerosols. The findings indicated the presence of dust and contaminated dust in the spring, summer, and fall, urban/industrial aerosols throughout the year, particularly in the fall and Winter, and mixed aerosols throughout the year over the research location, but no biomass burning aerosols were discovered.

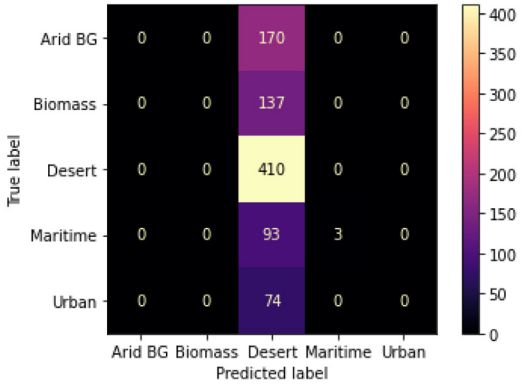
(Li et al., 2022) used the K-means method to categorize the seven basic aerosol features simulated by the EMAC-MADE3 global aerosol model from 2000 to 2013 into the various global aerosol regimes. Among the characteristics are black carbon mass concentration, mineral dust, sea salt, particulate organic matter, the sulfate/nitrate/ammonium system, and the Aitken and accumulation modes' aerosol number concentrations. The method discovered many aerosol clusters distinguished by anthropogenic pollution, mineral dust, emissions from biomass burning, and other sources. Using a satellite-based random forest (RF) Aerosol classification model between 2018 and 2020, with or without AERONET observations, a study was carried out to categorize and examine the most prevalent aerosol types in Asian capital cities (Choi, Lee, and Park et al. 2021). The properties such as FMF, SSA, depolarization ratio, effective radius, and volume size distribution, AI, CO, NO_2, Land cover type, Solar zenith angle, Percent of urban area, AOD, AE obtained from AERONET, TROPOMI Sentinel 5P, MODIS were used. According to the



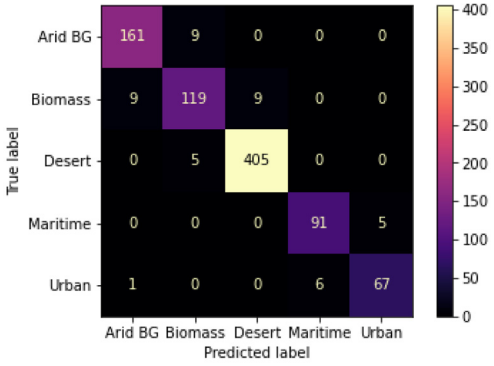
(a) KNN



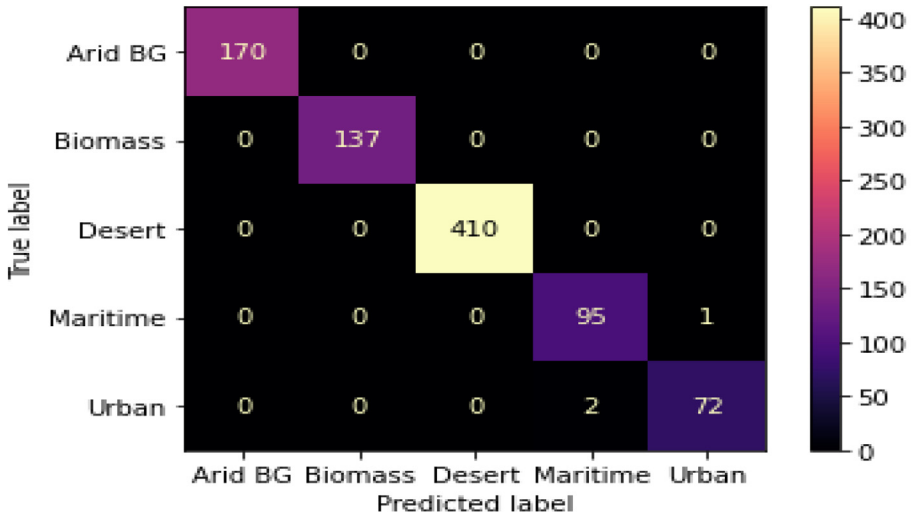
(b) Random Forest



(c) Naïve Bayes



(d) Support Vector Machine

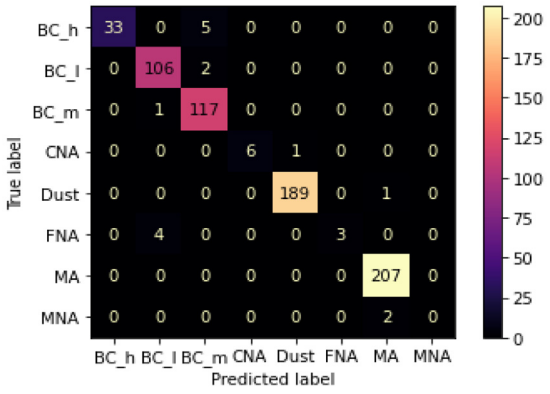


(e) Decision Tree

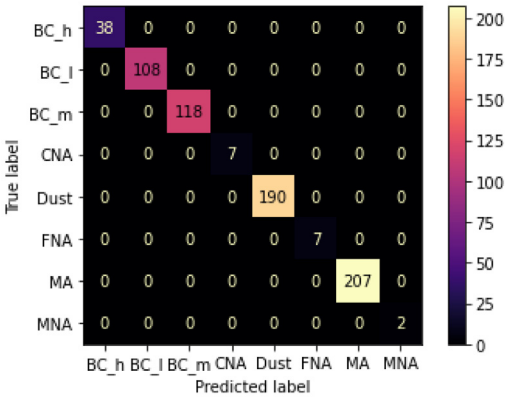
Fig. 11. Confusion matrix of machine learning algorithms for classifying aerosols based on source.

study, there are four types of aerosols: non-absorbing, heavily absorbing, dust-dominated, and pure dust aerosols. It was discovered that non-absorbing aerosols, mainly, were mostly produced by pollution in Asian capital cities. Nonetheless, seasonal observations of natural dust particles were also made in East Asia (March to May) and South Asia (March-August).

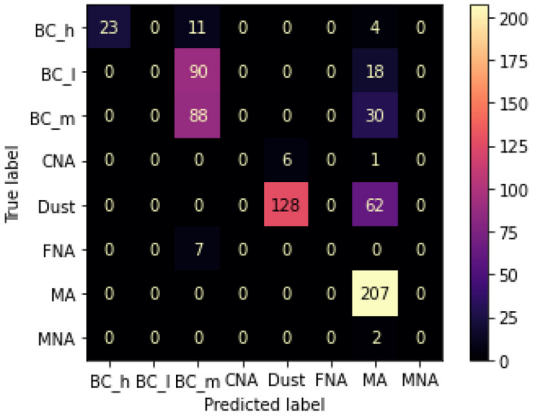
Based on the SKYNET data products, changes in column-integrated aerosol optical characteristics over Beijing, including AOD, AE (400–870 nm), SSA, ASY, and



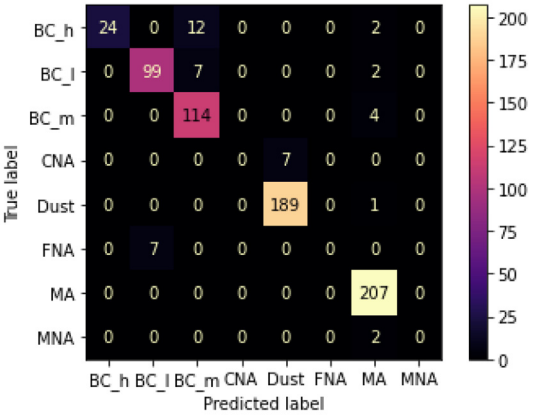
(a) KNN



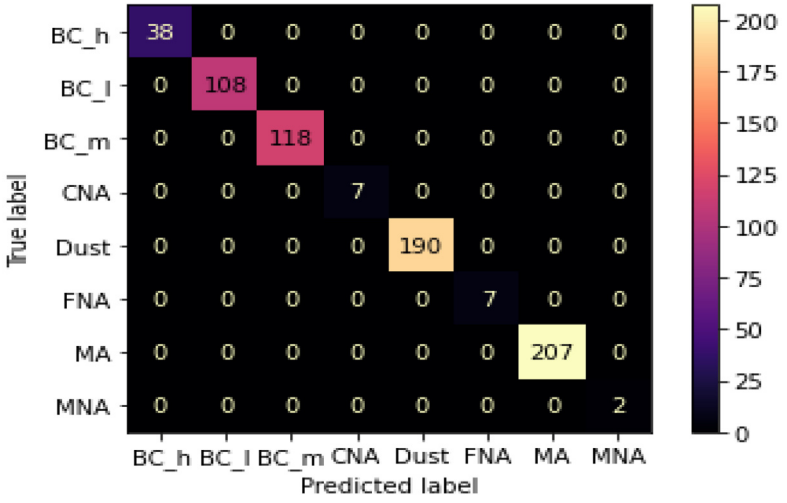
(b) Random Forest



(c) Naïve Bayes



(d) Support Vector Machine



(e) Decision Tree

Fig. 12. Confusion matrix of machine learning algorithms for classifying aerosols based on composition.

refractive index, were examined from March 2016 to December 2019 (Dong Xiaofei et al. 2023). Higher AOD and AE readings during the summer suggest that it is crucial to pay closer attention to fine particle pollution. They discovered that monthly changes in aerosol optical depth calculated from SKYNET observation were mainly compatible with the AERONET data by fitting the AOD data of SKYNET with AERONET at the same waveband. The spring aerosols displayed a good scattering ability and limited absorption capacity, per the SSA and ASY index.

Table 8
Accuracy comparison of machine learning algorithms for classifying aerosols based on source and composition.

ML algorithms	KNN	RF	NB	SVM	DT
Accuracy (%) - source based	97.75	***100***	46.56	95.04	***99.66***
Accuracy (%) - composition based	97.64	***100***	65.88	93.50	***100***

Table 9
Performance metrics.

Measure	KNN	RF	NB	SVM	DT
Sensitivity/Recall (%)	**98.31**	***99.05***	*46.38*	97.52	***99.05***
Specificity (%)	**99.69**	***99.68***	*89.04*	99.34	***99.68***
Precision (%)	**97.34**	***99.16***	*73.50*	97.89	***99.16***
Negative Predictive Value (%)	**99.66**	***99.69***	*92.16*	99.35	***99.69***
False Positive Rate (%)	**0.31**	***0.32***	*10.96*	0.65	***0.32***
False Negative Rate (%)	**1.69**	***0.95***	*53.62*	2.47	***0.95***
False Discovery Rate (%)	**2.65**	***0.83***	*26.49*	2.10	***0.83***
Accuracy (%)	**99.47**	***99.55***	*85.49*	99.05	***99.55***
F1 Score (%)	**97.76**	***99.10***	*45.18*	97.70	***99.10***
Matthews Correlation Coefficient	**0.9825**	***0.9851***	*0.5217*	0.9684	***0.9851***

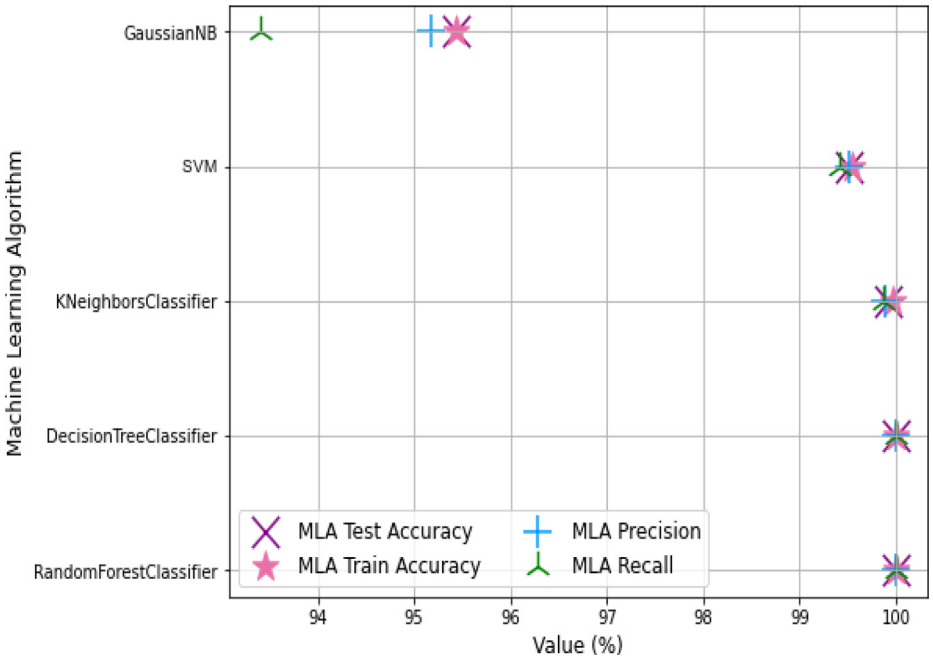


Fig. 13. Performance metrics of machine learning algorithms for training and testing.

Aerosols could be further classified into dust aerosol and polluted aerosol using the CALIPSO lidar and SKYNET data, and the results are compatible with previous categorization techniques. These showed that Beijing's primary composition modes were small and mixed and that aerosol had weak absorption. The results of fitting with MODIS demonstrated that the technique based on deep learning can effectively supplement the missing value. The random forest approach was used to fill in missing data.

Using an Automated Machine Learning (AutoML) technique, which chooses based on the data set, (Zhonghua Zheng et al. 2023) analyzed the information diversity of tropospheric trace gas columns for $PM_{2.5}$ estimates over India within a modeling testbed. For calculating $PM_{2.5}$ over four Indian sub-regions on daily and monthly time scales, they measured the relative information content of tropospheric trace gas columns, AOD, meteorological fields, and emissions. The results imply that including trace gas modeled columns enhance $PM_{2.5}$ predictions, regardless of the specific machine learning model assumptions. They used the ranking scores generated by the AutoML system and Spearman's rank correlation to infer or link the potential relative relevance of primary vs. secondary sources of $PM_{2.5}$. The comparison of selected baseline machine learning models with AutoML-derived models showed that AutoML is at least as effective as

Table 10
Comparison of the proposed method with existing papers.

Ref	Measurement Type	Location	Year	Variables used	ML algorithm	Classified aerosol types	Performance metrics
Christopoulos et al. 2018	single-particle mass spectrometry	Karlsruhe Institute of Technology (KIT) Germany	–	–	Random Forest	Fertile soils, mineral/metallic particles, biological particles, and all other aerosols.	Accuracy = 93 %
Hamill et al. 2016	AERONET	at 190 AERONET sites	1993 to 2012	EAE, AAE, SSA, RRI, IRI	Mahalanobis Distance	Urban-Industrial, Biomass Burning, Mixed Aerosol, Dust, and Maritime.	–
Siomos et al. 2020	double monochromator Brewer spectrophotometer	Thessaloniki, Greece	1998–2017	AOD, FMF, SSA, EAE	Mahalanobis Distance	Fine Non Absorbing Mixtures, Black Carbon Mixtures, Dust Mixtures, and Mixed.	Typing score compared with training dataset FNA: 77.0 %, BC: 63.9 %, and DUST: 80.3 %
	CIMEL Sunphotometer			SSA, EAE, AOD			Typing score compared with manually classified dataset FNA: 100.0 %, BC: 66.7 %, and DUST: 83.3 %
Choi et al. 2021	AERONET, TROPOMI Sentinel 5P, MODIS	At all AERONET sites	Jan 2018 to July 2020,	PLDR, SSA, AI, CO, NO_2, AOD, AE, TOA reflectance at 412, 470, and 660 nm, Land cover type, Solar zenith angle, Percent of urban area	Random Forest	Pure dust, dust-dominant mixed, pollution-dominant mixed aerosols, and pollution aerosols (strongly, moderately, weakly, and non-absorbing).	Accuracy = 59 %
Gharibzadeh et al. 2018	AERONET, TROPOMI Sentinel 5P, MODIS	Zanjan, Iran	2010 to 2013	AOD, AE, EAE, SSA, AAE, FMF	–	dust, urban industrial, biomass burning, and mixed aerosol	–
Xiaofei Donget al. 2023	SKYNET	Beijing	Mar 2016 to Dec 2019	AOD, AE	–	Container clean, marine, biomass burning / urban-industrial, desert dust, mixed	–
				SSA, AE		Strong absorption fine mode, strong absorption mixed mode, strong absorption coarse mode, moderate absorption fine mode, moderate absorption mixed mode, moderate absorption coarse mode, weak absorption fine mode, weak absorption mixed mode, weak absorption coarse mode	
Kuifeng Luan et al. 2023	CALIPSO	China	2007 to 2020	Attenuation backscattering, particle depolarization, surface type, layer top and foundation height	–	clean marine, dust, polluted continental/smoke, clean continental, polluted dust, elevated smoke, dusty marine.	–

(*continued on next page*)

Table 10 (*continued*)

Ref	Measurement Type	Location	Year	Variables used	ML algorithm	Classified aerosol types	Performance metrics
Xinyu Yuet al. 2022	AERONET	Hong Kong	2006 to 2021	AOD, AE SSA, AE, FMF	–	Marine, Dust, Mixed 1, Mixed 2, Urban/Industrial, Biomass Burning Coarse absorbing, Coarse non-absorbing, Mixed absorbing, Mixed non-absorbing, Fine highly absorbing, Fine medium absorbing, Fine slightly absorbing, Fine non-absorbing	–
(Li et al., 2022)	EMAC-MADE3 aerosol model	Global	2000–2013	Mass concentration of black carbon, mineral dust, sea salt, particulate organic matter, the sulphate/nitrate/ammonium system, and the aerosol number concentrations of the Aitken and accumulation modes.	K-means algorithm	biomass burning or biogenic activity, mineral dust, anthropogenic pollution, background conditions, mixture	–
Choi, Lee, and Park et al. 2021	AERONET, TROPOMI Sentinel 5P, MODIS	Asian capital cities	2018 to 2020	FMF, SSA, depolarization ratio, effective radius, and volume size distribution, AI, CO, $NO_{2,}$,Land cover type, Solar zenith angle, Percent of urban area, AOD, AE	Random Forest	Pure dust, dust-dominated aerosols, strongly absorbing aerosols, and non-absorbing aerosols	–
This work	AERONET	Kanpur	2002–2018	AOD, AE SSA, AE, FMF	KNN, RF, NB, SVM, DT	Urban, Maritime, Desert, Biomass, AridBG Mixture Absorbing, Dust, Black Carbon Medium, Black Carbon Low, Black Carbon High, Coarse Non-Absorbing, Fine Non-Absorbing, Mixture Non-Absorbing	Sensitivity/Recall (%) = 99.05 Specificity (%) = 99.68 Precision (%) = 99.16 Negative Predictive Value (%) = 99.69 False Positive Rate (%) = 0.32 False Negative Rate (%) = 0.95 False Discovery Rate (%) = 0.83 Accuracy (%) = 99.55 F1 Score (%) = 99.10 Matthews Correlation Coefficient = 0.9851

user-selected models. The idealized pseudo-observations (chemical-transport model simulations) used in this study set the stage for using satellite retrievals of tropospheric trace gases to estimate acceptable particle concentrations in India.

In (Luan Kuifeng et al. 2023), using statistical analysis of Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) L3 data from 2007 to 2020, the horizontal and vertical distributions of aerosol properties in the Taklimakan Desert (TD), North Central Region of China (NCR), North China Plain (NCP), and Yangtze River Delta (YRD) were examined. TD (0.38), NCP (0.49), and YRD (0.52) have high annual averages for AOD. These rates show a decline after the long-term pollution control measures were put into place; AOD values decreased by 5 %, 13.8 %, 15.5 %, and 23.7 % in TD, NCR, NCP, and YRD, respectively, when comparing 2014–2018 to 2007–2013, and by 7.8 %, 11.5 %, 16 %, and 10.4 % when comparing 2019–2020 to 2014–2018. The aerosol extinction coefficient and a distinct regional pattern tended to decline gradually with increasing height. The variations in AOD and extinction coefficients between TD and NCR and NCP and YRD were caused by dust and polluted dust, respectively. With a change in longitude in TD, dust aerosol first rose and then progressively fell, peaking in the middle. Compared to TD and NCR, the raised smoke aerosols in NCP and YRD were much higher. The relatively weak aerosol extinction coefficients (>0.001 km-1) were primarily distributed between 5 and 8 km, and the high aerosol extinction coefficient values (>0.1 km-1) were primarily distributed below 4 km, suggesting that NCP and YRD are impacted by the high-altitude long-range transport of TD and NCR dust aerosols.

The work by (Yu Xinyu et al. 2022) used AERONET data and satellite-based observations based on the extreme-point symmetric mode decomposition (ESMD) model to investigate seasonal patterns and long-term fluctuations in aerosol optical properties in Hong Kong from 2006 to 2021. Mixed and urban/industrial aerosols with fine-mode sizes and slightly absorbing or non-absorbing qualities are the two most common types of aerosol in Hong Kong. Aerosol optical depth (AOD), Angstrom exponent (AE), and single scattering albedo (SSA) all changed according to the season, with summer showing lower AOD but higher AE and SSA and spring and Winter showing higher AOD but lower AE and SSA. With an upward tendency in AOD and AE before 2012 and a lower trend after 2012, the long-term variations indicated that 2012 marked a turning point in the data. However, SSA showed a growing tendency in pre- and post-2012 eras, with the first period showing a steeper gradient. Aerosol optical characteristics had shorter-term, non-linear variations with alternating ascending and descending trends, according to the ESMD study. Extreme gradient boosting (XGBoost)-based analysis of the correlations between AOD and meteorological parameters revealed that the impacts of weather on AOD are complex and non-monotonic. The reduction of aerosol loads in Hong Kong is helped by decreased relative humidity, greater southwesterly winds, and lower temperatures.

It is evident from Table 10 that the majority of the articles categorize aerosols based on either source or composition. This work describes two aerosol categorization techniques (Source and composition). A relationship is also made between source-based classifications (Urban, Maritime, Desert, Biomass, and Arid) and composition-based classifications (Mixture Absorbing, Dust, Black Carbon Medium, Black Carbon Low, Black Carbon High, Coarse Non-Absorbing, Fine Non-Absorbing, Mixture Non-Absorbing). This can aid in figuring out the origin of a specific aerosol and, in turn, help determine the type of aerosol that particular source is emitting. This paper used four aerosol properties, AOD, AE, SSA, and FMF, acquired from the globally accessible AERONET for Aerosol classification based on machine learning algorithms. Much literature also noted that many aerosol properties obtained from different measurement techniques such as spectrometers (Brewer, CIMEL, Single particle mass), satellites (TROPOMI, MODIS), and models were used for aerosol classification. In most of the literature publications, aerosol categorization was carried out using machine learning techniques. The most appropriate machine learning algorithm for the classification of aerosols is found after validating the derived classification schemes using the five machine learning algorithms Naive Bayes (NB), K Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). This work uses a variety of performance metrics to determine the best machine learning algorithm for the class, including accuracy, F1 score, specificity, precision, negative predictive value, false positive rate, false negative rate, false discovery rate, and sensitivity/recall. Most previous studies only used accuracy and typing score to validate the performance of machine learning algorithms.

6. Conclusion

In this paper, aerosol data obtained from AERONET over a long period from 2002 to 2018 is used for classifying aerosols. Two classification schemes are presented in this work for the Kanpur region. The aerosol properties such as AE, AOD, SSA, and FMF obtained from AERONET are used for this classification. The classification schemes are based on the composition (Mixture Absorbing, Dust, Black Carbon Medium, Black Carbon Low, Black Carbon High, Coarse Non-Absorbing, Fine Non-Absorbing, Mixture Non-Absorbing) and Source (Urban, Maritime, Desert, Biomass, and Arid) of the Aerosol. After classification, the relationship between composition and Source of Aerosol is found. Once this is done, various machine learning algorithms are applied to evaluate classification performance with different performance metrics. After validating the derived classification schemes with

the five machine learning algorithms Naive Bayes (NB), K Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF), the most appropriate machine learning algorithm for the classification of aerosols is discovered. It employs a range of performance criteria, such as accuracy, F1 score, specificity, precision, negative predictive value, false positive rate, false negative rate, false discovery rate, and sensitivity/recall, to identify the best machine learning method. The source based aerosols of the desert and arid background and the composition based aerosols of types, Mixture Absorbing, Coarse absorbing (Dust), and Black Carbon dominated the Kanpur region during the study period. The Number of non-absorbing (scattering) type aerosols are least in the study region considered during the study period at all the seasons. Random forest and decision tree were the best algorithms for classifying aerosol data. The RF and DT metrics values are Sensitivity/Recall (%) = 99.05, Specificity (%) = 99.68, Precision (%) = 99.16, Negative Predictive Value (%) = 99.69, False Positive Rate (%) = 0.32, False Negative Rate (%) = 0.95, False Discovery Rate (%) = 0.83, Accuracy (%) = 99.55, F1 Score (%) = *99.10*, and Matthews Correlation Coefficient = *0.9851* and found to be best as compared to other algorithms. Also, the proposed method is compared with existing methods, and the results are provided in the discussion section.

The present work revolves around aerosol distribution in the Kanpur region, India, from 2002 to 2018. The model can be evaluated for denser datasets of different areas with different geographical conditions, specifically in the Indo-Gangetic Plain, to study how changes in aerosol patterns have affected the region, its air, and its agricultural produce. The classification schemes presented can backtrack the origin of a particular type of aerosol, which can be used to cut off or diminish the aerosol's further emission and implement strict rules to preserve the air quality. The current work can be extended to establish a link between the source-composition-based aerosol types and the air quality index. Also, particulate matter 2.5 can be incorporated into the studies. The data from the MODIS satellite can also be included to get a denser dataset for deep learning applications

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Anitha, M., Kumar, L.S., 2020. Ground Based Remote Sensing of Aerosols Using AERONET in Indian Region. In 2020 international conference on wireless communications signal processing and networking (WiSPNET) IEEE, 72-77.

Anitha, M., Kumar, L.S., 2023. Development of an IoT-Enabled Air Pollution Monitoring and Air Purifier System. Mapan J. Metrol. Soc. India. https://doi.org/10.1007/s12647-023-00660-y.

Babu, S.S., Manoj, M.R., Moorthy, K.K., Gogoi, M.M., Nair, V.S., Kompalli, S.K., Satheesh, S.K., Niranjan, K., Ramagopal, K., Bhuyan, P.K., Singh, D., 2013. Trends in Aerosol optical depth over Indian region: Potential causes and impact indicators. Journal of Geophysical Research-Atmospheres 118 (11), 794–806.

Cazorla, A., Bahadur, R., Suski, K.J., Cahill, J.F., Chand, D., Schmid, B., Ramanathan, V., Prather, K.A., 2013. Relating aerosol absorption due to soot, organic carbon, and dust to emission sources determined from in-situ chemical measurements. Atmospheric Chemistry and Physics 13 (18), 9337–9350.

Choi, W., Kang, H., Shin, D., Lee, H., 2021. Satellite-based aerosol classification for capital cities in Asia using a random forest model. Remote Sens.-Basel 13 (2464), 1–13.

Choi, W., Lee, H., Park, J., 2021. A first approach to aerosol classification using space-borne measurement data: Machine learning-based algorithm and evaluation. Remote Sens.-Basel 13 (609), 1–21.

Choudhry, P., Misra, A., Tripathi, S.N., 2012. Study of MODIS derived AOD at three different locations in the Indo Gangetic Plain: Kanpur, Gandhi College, and Nainital. Annales de Geophysique 30 (10), 1479–1493.

Christopoulos, C.D., Garimella, S., Zawadowicz, M.A., Möhler, O., Cziczo, D.J., 2018. A machine learning approach to aerosol classification for single-particle mass spectrometry. Atmospheric Measurement Techniques 11 (10), 5687–5699.

Dey, S., Tripathi, S.N., Singh, R.P., Holben, B.N., 2005. Seasonal variability of the aerosol parameters over Kanpur, an urban site in Indo-Gangetic basin. Advances in Space Research 36, 778–782.

Emetere, M.E., 2019. Environmental modeling using satellite imaging and dataset re-processing. Springer International Publishing.

Gharibzadeh, M., Alam, K., Abedini, Y., Bidokhti, A.A., Masoumi, A., Bibi, H., 2018. Characterization of Aerosol optical properties using multiple clustering techniques over Zanjan, Iran, during 2010–2013. Applied Optics 57 (11), 2881–2889.

Hamill, P., Giordano, M., Ward, C., Giles, D., Holben, B., 2016. An AERONET- Based aerosol classification using the Mahalanobis distance. Atmospheric Environment 140, 213–236.

HaoChen, ChengTianhai., Gu, Xingfa., Li, Zhengqiang., Wu, Yu., 2016. Characteristics of aerosols over Beijing and Kanpur derived from the AERONET dataset. Atmos Pollut Res 7, 162–169. https://doi.org/10.1016/j.apr.2015.08.008.

Kuifeng, L., Cao, Z., Song, H.u., Qiu, Z., Wang, Z., Shen, W., Hong, Z., 2023. Aerosol characterization of Northern China and Yangtze River Delta based on multi-satellite data: Spatiotemporal variations and policy implications. Sustainability 15 (3), 1–24.

Kumar, S., Kumar, S., Singh, A.K., Singh, R.P., 2012. Seasonal variability of atmospheric aerosol over the North Indian region during 2005–2009. Advances in Space Research 50, 1220–1230.

Laakso, L., Koponen, I.K., Mönkkönen, P., Kulmala, M., Kerminen, V. M., Wehner, B., Wiedensohler, A., Wu, Z., Hu, M., 2006. Aerosol particles in the developing world; a comparison between New Delhi in India and Beijing in China. Water Air Soil Poll. 173, 5–20.

Lal, S.N., Chandra, K.J., Mahavir, S., Manum, S., Raj, G., 2011. Characteristics of aerosol optical depth and Ångström parameters over Mohal in the Kullu Valley of Northwest Himalayan Region, India. Acta Geophysica 59, 334–360. https://doi.org/10.2478/s11600-010-0046-1.

Lee, J., Kim, J., Song, C.H., Kim, S.B., Chun, Y., Sohn, B.J., Holben, B. N., 2010. Characteristics of aerosol types from AERONET sunphotometer measurements. Atmospheric Environment 44, 3110–3117.

Li, J., Hendricks, J., Righi, M., Beer, C.G., 2022. An aerosol classification scheme for global simulations using the K-means machine learning method. Geoscientific Model Development 15 (2), 509–533.

Maciszewska, A., Markowicz, K., Witek, M., 2010. A Multiyear Analysis of Aerosol Optical Thickness over Europe and Central Poland Using NAAPS Model Simulation. Acta Geophysica 58, 1147–1163.

Mahesh, B., Rama, B.V., Spandana, B., Sarma, M.S.S.R.K.N., Niranjan, K., Sreekanth, V., 2019. Evaluation of MERRAero $PM_{2.5}$over Indian cities. Adv. Space Res. 64, 328-334.

Mohan, A.S., Manisekaran, A., Kumar, L.S., 2021. Aerosol classification using machine learning algorithms. Indian J. Radio Space Phys. (IJRSP) 50, 217–223.

Ojha, N., Sharma, A., Kumar, M., Girach, I., Ansari, T.U., Sharma, S.K., Singh, N., Pozzer, A., Gunthe, S.S., 2020. On the widespread enhancement in fine particulate matter across the Indo-Gangetic Plain towards Winter. Sci. Rep.-UK 10 (5862), 1–9.

Payra, S., Gupta, P., Bhatla, R., El Amraoui, L., Verma, S., 2021. Temporal and spatial variability in aerosol optical depth (550 nm) over four major cities of India using data from MODIS onboard the Terra and Aqua satellites. Arabian Journal of Geosciences 13 (1256), 1–11. https://doi.org/10.1007/s12517-021-07455-y.

Raman, R.S., Ramachandran, S., Kedia, S., 2011. A methodology to estimate source-specific aerosol radiative forcing. Journal of Aerosol Science 42 (5), 305–320.

Siomos, N., Fountoulakis, I., Natsis, A., Drosoglou, T., Bais, A., 2020. Automated Aerosol classification from spectral UV measurements using machine learning clustering. Remote Sens-Basel. 12, 965–965.

Szkop, A., Pietruczuk, A., Posyniak, M., 2016. Classification of aerosol over Central Europe by cluster analysis of aerosol columnar optical properties and backward trajectory statistics. Acta Geophysica 64, 2650–2676.

Tariq, S., Ali, M., 2015. Analysis of optical and physical properties of aerosols during crop residue burning event of October 2010 over Lahore, Pakistan. Atmospheric Pollution Research 6, 969–978.

Thamban, N.M., Tripathi, S.N., Moosakutty, S.P., Kuntamukkala, P., Kanawade, V.P., 2017. Internally mixed black carbon in the Indo-Gangetic Plain and its effect on absorption enhancement. Atmospheric Research 197, 211–223.

Xiaofei, D., Chen, B., Yamazaki, A., Shi, G., Tang, N., 2023. Variations in aerosol optical characteristics from SKYNET measurements in Beijing. Atmospheric Environment 302 (119747), 1–10.

Xinyu, Y.u., Nichol, J., Lee, K.H., Li, J., Wong, M.S., 2022. Analysis of long-term aerosol optical properties combining AERONET sunphotometer and satellite-based observations in Hong Kong. Remote Sens.-Basel 14 (20), 1–17.

Zheng, F., Hou, W., Sunm, X., Li, Z., Hong, J., Ma, Y., Li, L., Li, K., Fan, Y., Qiao, Y., 2019. Optimal estimation retrieval of aerosol fine-mode fraction from ground-based sky light measurements. Atmos.-Basel 10 (196), 1–15.

Zheng, C., Zhao, C., Zhu, Y., Wang, Y., Shi, X., Wu, X., Chen, T., Wu, F., Qiu, Y., 2017. Analysis of influential factors for the relationship between PM 2.5 and AOD in Beijing. Atmospheric Chemistry and Physics 17 (21), 13473–13489. https://doi.org/10.5194/acp-17-13473-2017.

Zhonghua, Z., Arlene, M.F., Daniel, M.W., George, P.M., Goldsmith, J., Karambelas, A., Curci, G., et al., 2023. Automated machine learning to evaluate the information content of tropospheric trace gas columns for fine particle estimates over India: A modeling testbed. Journal of Advances in Modeling Earth Systems 15 (3), 1–17.