

EGUsphere-2025-269: “Aerosol type classification with machine learning techniques applied to multiwavelength lidar data from EARLINET”, by del Águila et al.

This paper presents a novel methodology for automated aerosol classification based on Machine Learning (ML) methods applied to lidar measurements with respect to other aerosol typing procedures. Indeed, aerosol typing, especially with high vertical resolution, is a crucial issue for a better understanding of atmospheric composition and its impact on the climate. Therefore, the outcomes of this work are rather relevant and hence it deserves to be published. However, a Major revision should be performed before it is accepted for publication.

General comments:

- 1) Main findings of this work are critically based on a first aerosol classification manually performed using the EARLINET Granada lidar database (aerosol optical properties and their derived intensive parameters, as shown in Eqs. 1-4). However, no information is introduced in the work about the criteria used for that purpose. This is crucial as the performance of the ML methods applied in aerosol classification is compared against that manual typing. It should be included and clarified (an additional table could help).
- 2) The reference dataset is divided in 80% for training and 20% for testing of the ML methods. Could the results be affected if those percentages are modified? Please, provide an explanation.
- 3) An external validation of the best ML method is performed with an independent dataset just for a dust case. Why it is not validated for another type of aerosols, for instance, less-depolarizing aerosols (e.g. smoke or continental pollution), and thus to be more confident in the ML method performance? This would improve the work.
- 4) What is the difference of using either height-resolved data or conventional average aerosol layer values (i.e. the same value at every height-level of the layer)? This can trigger differences in the results. Please, clarify.

Specific comments:

- 1) Sect. 2.1:
 - a) Once the aerosol layers are identified, the paper indicates that the database includes two representations of the intensive properties (page 4, line 111): height-resolved data and layer-averaged data. The latter averaged value is then assigned to every height-level of each particular identified layer to maintain the lidar resolution. If a single type of aerosol is assigned to each aerosol layer (page 4, line 117), this procedure just serves to increase the same averaged value, which is also associated to a given aerosol type. This should be explained in more detail.
 - b) Also, specify what of the two data representations is used as the reference dataset for comparison with the outputs of each ML method (page 4, lines 115-117).
 - c) The vertical resolution of the lidar profiles used in the reference database for the UGR station is indicated but not their time resolution (page 4, lines 92-94). This should be included.
- 2) Sect. 2.6: Please, provide more information on:
 - a) Table 1: Why did you apply those ranges of values for each hyperparameter?
 - b) Page 8, lines 212-214: Which are the weights to each class applied?
- 3) Sect. 3.1: Figures 3 and 4, and Table 2 (pages 10-11): They can be affected depending on the manual aerosol classification performed to differentiate the ‘Unknown’ type from the rest of aerosols, and between each aerosol class. Please, clarify.
- 4) Sect. 3.2:
 - a) Page 12: Table 3 could be merged with Table 1.

- b) Page 12, lines 281-282: It is not clear if the 416 layers examined in this work correspond to the reference dataset or the testing dataset. Please, clarify.
 - c) Pages 12-13, lines 284-285: In the statement: 'In general, the ML models that incorporate depolarization data demonstrate significantly higher performance compared to those without depolarization', it should be added 'except for the neural network', as shown in Figure 5.
 - d) Page 13, lines 288-289: In the sentence: 'This might be explained due to a stronger influence of other features rather than depolarization on the aerosol classification problem with the NN setup'. Explain what do you mean by 'other features', please.
 - e) Page 16, lines 362-365: In the statement, 'The Continental Polluted and Mixed types show comparable performances near $\sim 70\%$ in both configurations, indicating that these aerosol types are more difficult to be predicted in the classification process. They are also confused with Smoke in a $\sim 30\%$, which makes sense due to nature of the Mixed aerosol type, which can be composed of several types of aerosols, being Smoke one of the major contributors.', why Smoke is one of the major contributors in the Mixed type? Maybe that is related to the way the Mixed class was identified? Please, clarify it.
- 5) Sect. 3.3:
- a) Page 17, line 397: Please, provide more detail for '... particularly when several intensive properties are missing, ...'. Why?
 - b) Pages 18-19, lines 413-418: Please, check this paragraph, there is an inconsistency. If there are not available depolarization information in the EARLINET database during the Saharan dust period, how the depolarization data could be included for the evaluation of the LightGBM performance in this configuration? Please, clarify.
- 6) Discussion Sect.:
- a) Page 19, lines 431-434: Although it was mentioned in this section, the manual aerosol classification is not described in the work, and it can affect the reference dataset used. See General Comment #1.
 - b) Page 19, lines 435-439: See General Comment #4.

Other minor comments:

- 1) Page 5, line 121: What do you mean with 'corresponding height of each layer'? Please, specify.
- 2) Page 5, line 128: What is the meaning of 'row' in this sentence?
- 3) Page 5, line 135: Why do you use the median value for the missing values? Explain.
- 4) Page 5, line 136: If each profile is denoted by j , does it have sense $j=0$? Also, define Δz .
- 5) Page 6, lines 170-172: If k denotes the variables, $k = 0$ should be avoided, shouldn't it? Also, the term 'the predicted class for each layer and height' is correct? Either layer or height?
- 6) Page 7, line 198: Please define 'GroupKFold' (and include reference).
- 7) Page 7, line 198: Please define 'GroupShuffleSplit' (and include reference).
- 8) Page 10, Figure 3's caption: 'ratio' is missing at the end of 'particle linear depolarization'.
- 9) Page 3, Figure 5's caption: Should it be 'in the testing dataset' instead of 'in the reference dataset'?
- 10) Page 14, Figure 6: The title in the colour bar may be 'Feature importance' instead of 'Feature value'?
- 11) Page 17, line 371: Check: '... which was used for neither training nor testing the model, ...'.
- 12) Page 18, Figure 8: Check the x-axis: the title of the first panel, and also the scale in the second panel (it should be multiplied by 10^6).
- 13) Pages 21-28: Check the reference list as wrong and missing details are found.