RC2: Anonymous Referee #2, 02 Feb 2026

Overview:

In this study, Tardiff et al. sought to distinguish tree, grass & weed pollen in an urban area (Montréal) using flow cytometry data & a machine learning algorithm (specifically random forest). They identified granularity & fluorescence parameters from violet (Violet610_A) & blue (PB450_A) lasers to be the most distinctive features for discerning a pollen species, though species accuracy (F1 = 0.76) in the random forest-model was lower than genus accuracy (F1 = 0.90). I generally found the manuscript to be well written, with few spelling, grammatical issues. This study presents a strong step toward scalable pollen classification using non-imaging cytometry. However, some concerns and requests for clarification are pointed out below:

We thank the reviewer for the positive assessment of our manuscript and for the constructive comments and suggestions.

Major comments:

Line 94-95: For this study, I understand that the purity of the pollen can be critical in analyzing the data obtained from the flow cytometer because the proportion of non-pollen particles that can be similar in size or fluorescence intensity can influence the classification. However, there is no data to validate the pollen's purity using other methods. For example, would it be possible to compare the pollen purity by an image analysis of the extracted pollen subsample under a light/fluorescence microscope?

We thank the reviewer for asking this question. We note also that Referee #1 had a similar concern, which further highlights that we need to be clearer on this issue. To ensure pollen purity, we performed prior to model training a data cleaning step (Figure A2 and Lines 127–131) to separate pollen grains from "debris" classified as non-pollen (other) in our training dataset. To distinguish these two groups, we first used excitation violet laser fluorescence parameters (detectors PB450 and Violet610), while verifying that the distinction held true for the same groups across the other variables. This excitation/emission range is characteristic of sporopollenin which contains fluorophores that in turn are specific to pollen grains [1]. The model is therefore able to identify all particles sharing the fluorescence signature of pollen; everything else is categorised as "other". Even if some non-pollen aerosols (e.g. fungal spores, dust, starch) may emit autofluorescence in similar wavelengths, the probability that any such particle would simultaneously reproduce the sporopollenin-specific fluorescence signature and all multi-parametric characteristics used by the model (size, granularity, multi-channel fluorescence) remains very low. The robustness of our approach relies precisely on this combined multi-parametric classification.

[1] Pöhlker, C., Huffman, J. A., Förster, J.-D., and Pöschl, U.: Autofluorescence of atmospheric bioaerosols: spectral fingerprints and taxonomic trends of pollen, Atmos. Meas. Tech., 6, 3369–3392, https://doi.org/10.5194/amt-6-3369-2013, 2013.

In response to the Reviewer's comment, we made the following changes to the manuscript:

L146-147 : We added : « This excitation/emission range is characteristic of sporopollenin which contains the fluorophores specific to pollen grains (Pöhlker et al., 2013a). »

L136 we also clarify what we consider as debris : « […] debris, that is non-pollen particles, […]»

Line 143-144 = Since the data is unbalanced, the authors chose to use synthetic minority over-sampling to normalise the data. However, I was under the impression that oversampling can cause model overfitting & an inaccurate representation of the smallest minority classes (~300 – 35000 is very unbalanced). Were steps taken to avoid/ensure that didn't happen? It may be at least worth a mention in the discussion.

We thank the reviewer for this comment. Only four taxa (Acer saccharum, Gramineae spp, Juglans cinerea, Picea abies) were oversampled. We chose a threshold of 1,000 pollen grains per species to avoid excessive oversampling of species for which we had insufficient grains. We have added a new table in appendices with this information (Table A2).

We made the following changes to the manuscript to address this issue:

L. 157: We replaced « (min=306; max=35307) » by « (Table A2) »

L161-162: We added « Only four taxa were oversampled (Acer saccharum, Gramineae spp, Juglans cinerea, and Picea abies). »

Line 145-146 = Reads as synthetic minority over-sampling is done on the samples, then the dataset is split into 70% training & 30% validation. If so, would this not lead to overfitting & inflated precision metrics, since the training data is used in the validation set? Perhaps I've misinterpreted what's written, or this doesn't matter. If so, clarification may be needed in the order of steps taken.

We thank the reviewer for this important comment. SMOTE balancing function was indeed applied before splitting the dataset into training and validation sets; however, for our objective, this does not introduce overfitting or inflated accuracy metrics. The validation set, although balanced, was never used in any way during model training. The Random Forest classifier was trained exclusively on the 70% training portion. The 30% validation set, whether balanced or not, was used exclusively to determine whether the model was capable of correctly identifying species it had never encountered during training. It is important to note that the purpose of SMOTE was to provide the classifier with a balanced training set in order to prevent it from being biased toward the majority class. Whether the validation set is balanced or imbalanced has no bearing on this objective, as long as those samples, as was the case here, were not seen during training.

We acknowledge that this ordering of steps may appear unconventional, and we have clarified this in the revised manuscript:

L161-163 : We added « Only four taxa were oversampled (Acer saccharum, Gramineae spp, Juglans cinerea, and Picea abies). The purpose of balancing data was to provide the classifier with a balanced training set to prevent it from being biased toward the majority class. »

L164-165 : We added « The validation set was not used for model training. The Random Forest classifier was trained exclusively on the 70% training portion. »

Line 160 = While F_1-scores are useful for measuring model performance in a single metric, the likely strong class imbalance and use of oversampling would suggest that including metrics such as PR AUC in addition to F_1-scores would provide a more complete summary of the model's performance (Saito & Rehmsmeier, 2015 https://doi.org/10.1371/journal.pone.0118432).

Given that our dataset is balanced and that oversampling was performed only for four species, we believe that the F1 score remains relevant.

Minor comments:

Lines 103-106: The authors state that the pollen's fluorescence depends on the fluorescent proteins on its surface (my understanding is that these are not proteins). If so, please provide a reference.

Thank you for your comment; that is indeed an error. It is true that sporopollenin contains fluorescent phenolic compounds, but it is not a protein; we have changed this sentence in the revised manuscript:

L109-110 We changed « ...which excites the fluorescent proteins on the surface of the pollen grain's outer wall. » by « ... which excites fluorescent phenolic compounds present in the sporopollenin of the pollen grain's outer wall. »

Line 148 = At the beginning of 2.4. It is stated that four supervised classification algorithms were tested, and the random forest performed best. There are many types of random forest classifiers, such as random forest by randomisation, which deals well with unbalanced/noisy data. Which random forest classifiers (Breiman?) were tested? It may be worth mentioning why the particular random forest was chosen over other random forest classifiers.

We used indeed Breiman's original Random Forest model, and since our training dataset is balanced, there was no need to use a different Random Forest model - as a reminder, very few species were oversampled, and the SMOTE function was primarily used for undersampling.

Accordingly, on L154, we have added the reference for random forest model: (Breiman, 2001) Breiman, L.: Random forests. Machine learning, 2001.