We thank both reviewers for their useful comments and for their time. We did our best to update the last version of our manuscript following their remarks and suggestions and we hope they will be satisfied. Please note that in our answers below, line numbers refer to the author's track-change file.

Reviewer 1

Some suggestions for improvement.

In lines 120–121, "a triple-branch convolutional neural network architecture designed to process holographic and fluorescence input," readers may wonder why three branches are required for only two input modalities. It would therefore be useful to specify that the architecture processes two holographic images together with the fluorescence input.

> We thank the reviewer for this comment. We clarified what the three inputs are:
>
> L119-121:
>
> The 2025-Beta-1, 2025-Gamma-1, 2025-Omega-1, 2025-Beta-2, 2025-Gamma-2, 2025-Omega-2 and 2025-Gamma-3 algorithms follow a triple-branch convolutional neural network architecture designed to process two holographic images and one fluorescence input. The classifier works on events that have both kinds of data, or holography only by setting up the fluorescence input to zero.

Regarding the sentence in lines 300–301, "Ideally, a threshold of 0 would be used, but in practice the confidence thresholds need to be set as low as possible to maximise the number of events kept and optimise the three metrics across classes and the five evaluation sites", it is questionable whether a threshold of 0 is truly optimal. The device also captures particles belonging to classes that are not represented in the trained classifier. To filter out such particles, the threshold should be set sufficiently above zero, while not being so high that it excessively reduces the number of pollen events belonging to the target classes.

> We thank the reviewer for this remark and agree. We clarified this point:
>
> L 310-314
>
> In practice the optimal confidence thresholds need to be set as low as possible to maximise the number of events classified within a class. However, it needs to be high enough to exclude particles that are not represented by any of the model's classes, and that are forced into existing classes; such particles are likely to display low confidence thresholds. Finally, the optimal confidence threshold should be selected to optimise the three metrics across classes and the five evaluation sites.

Line 4 contains a repeated word: "five stations stations."

> Thank you for noticing, we corrected this typo, as well as a few others we identified.

Reviewer 2

Some suggestions for improvement.

The authors should explain in the manuscript to what extent the metrics for evaluating classification model is underestimated/overestimated because the data used to evaluate the eight classification algorithms were not available for entire season when airborne pollen and other aerosols could be detected in the air as seen in Table 2.

In my opinion the metrics does not have same robustness for test on data from 16.01.2024 to 14.09.2024 in Buchs as from 04.05.2024 to 31.08.2024 in Payerne. Especially for pollen classes which season is not fully represented.

> We thank the reviewer for their remark and agree that we should put more emphasis on these aspects. We clarified:
>
> L 249-257:
>
> Additionally, the three metrics also depend on the quality of the evaluation datasets, and one should therefore be careful when comparing metrics across classes of same or different models. The scaling factor in particular is used in operational setup, and the representativeness of the seasons it is calculated from should be assured to achieve a good robustness. For example, the metrics for the class Poaceae in our evaluation datasets is likely to be less robust than most of the other classes, as only two stations encompass Poaceae season (Table 2) compared to four for others. Similarly, very few *Carpinus* pollen counts were detected in the five stations, limiting the robustness of this class. In particular, scaling factors can be inflated if only low-count in-season periods are available. Finally, when evaluation windows are incomplete, as is the case for Poaceae in Luzern, Kendall's Tau can be underestimated if key peaks are missing, or occasionally overestimated if only a small, well-matched subperiod is sampled.
>
> L 438-439
>
> Overall, the evaluation framework would benefit from richer and more diverse evaluation datasets, while keeping in mind that this analysis can be computationally intensive and that data quality should be prioritised over quantity.