Reoptimizing neural networks for pollen classification

P. Matavulj[1], S. Jelic[2], S. Brdar[1], M. Radovanovic[3], T. Danijela[3], and B. Sikoparija[1]

[1] BioSense Institute, Serbia
[2] Faculty of Civil Engineering, University of Belgrade, Serbia
[3] Faculty of Sciences, University of Novi Sad, Serbia

Abstract. This study proposes the concept of retraining classification models by utilizing additional data sources to improve the accuracy of automatic pollen classification. The retraining process involves redefining the pre-trained convolutional neural network (CNN) to output pollen concentrations and then comparing automatic and manual measurements. The results indicate a significant improvement in correlation, with an average increase of 0.15 and a maximum increase of 0.25. Additionally, the results demonstrate that multiclass models outperform binary classification models, although each class requires its model. Overall, this study demonstrates that retraining classification models with manual measurements can significantly enhance automatic pollen classification, making it a promising approach for further research in this field.

1 Introduction

Deep learning models have been widely studied across various fields, but their applicability in real-world settings is still being explored. Usually, they are trained and evaluated on a specific dataset split into training and test datasets. The metrics on the test dataset are then compared to determine which model performed better. However, many of these models have yet to be tested in real-world settings, and even if they already have, they often don't use data from these settings to improve the models.

In this research, we investigate the potential of retraining neural networks with additional data different from the original training data. We focus on pollen classification, using data collected by the Rapid-E particle monitor, which automatically records airborne particles. We implement a neural network from previous research [1] on data collected by inserting different pollen species into the device and use this network to classify the yearly records collected by the Rapid-E to obtain pollen concentrations [2]. To evaluate the performance of this automatic method, we compare the automatic pollen concentrations with manually obtained concentrations. The comparison yielded promising results in previous studies, with Pearson's correlation coefficients exceeding 0.7 for 11 pollen types [2, 3]. In this study, we will utilize this manually collected dataset to retrain our classification models, transforming them into regression models that predict pollen concentrations, and in that way, we aim to improve pollen classification.

2 Materials and methodology

2.1 Data

In this study, the performance of automatic measurements of airborne particles is validated through manual measurements. Manual measurements are obtained with a Hirst-type device, while automatic measurements are obtained with the Rapid-E real-time airborne particle monitor (PLAIR SE). The two devices are placed close in close proximity on the rooftop of the Faculty of Sciences at the University of Novi Sad in Serbia. The study extends from 14 Apr 2018 to the end of 2020.

Manual measurements Continuously sampling ambient air, airborne particles fall on a moving adhesive tape. A sample tape for a given hour is then manually analyzed under a light microscope, where a trained domain expert classifies particles by their morphological attributes. This results in a dataset of hourly concentrations of certain pollen species.

Automatic measurements The Rapid-E device automatically samples airborne particles using a deep UV and infrared laser, resulting in three data types: scattered light image, fluorescence spectrum, and fluorescence lifetime [1]. To obtain pollen concentrations from Rapid-E measurements, a multi-modal CNN is trained on the dataset obtained by injecting known pollen in the devices' airflow [2]. This dataset has 26 classes, 24 representing different types of pollen and two representing non-pollen particles. Further details can be found in reference [2]. The multi-modal CNN extracts features from each input separately, concatenates them, and classifies the input particle [1]. Pollen concentrations are obtained by classifying yearly records collected by the Rapid-E.

2.2 Retraining with manual measurements

To improve the accuracy of our automatic pollen classification, we compare the results obtained from manual measurements to those obtained from automatic measurements by calculating Pearson's correlation coefficient. We want to maximize the model's correlation with the manual measurements to improve the automatic pollen classification. To do this, we are modifying the pre-trained model to output hourly concentrations, which we will then compare to the manual measurements using the Mean Squared Error (MSE) formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where n is the number of samples, Y_i is the manual measurement and $\hat{Y}_i$ is the automatic measurement for a given pollen class. After that, the model will be updated by backpropagation to minimize the error, i.e., to minimize the difference between the manual and automatic measurements.

To ensure comparability between the two sets of measurements obtained from different devices, we are normalizing both Y_i and $\hat{Y}_i$ to the 0-1 range. Initially, we tried normalizing the signals to have zero mean and standard deviation of one, but for similar sets of measurements, this normalization resulted in numerically very different vectors and a significant loss when calculating MSE.

To retrain our classification models, we had to change the forward function to take an hour as a sample. For the given hour, m particles are to be classified, where m is different for different hours. To obtain the desired output, we modified the forward function to output $\sum_{i=1}^{m} softmax(y_i)$, where y_i is the i-th sample in that hour for the class we are trying to predict.

First, we trained 23 binary classification models for each pollen type, where one class is the class we want to predict, and the other class represents other pollen and non-pollen classes. Then we retrained each of the 23 models. Furthermore, we retrained the multiclass classifier obtained from [2]. The multiclass loss was now changed to sum over all pollen classes. Furthermore, in the output formula from above, y_i in the multiclass settings represents a vector, where each position of that vector represents one pollen class.

3 Results

All models were trained over six epochs with a batch size of 24 hours, and every 50 batches, the models were saved. They were trained using 2019 data, which consisted of 5793 samples (hours). However, using the entire dataset to train binary models resulted in poor performance due to the fact that each pollen type's presence in the air typically lasts for usually one to two months. To address this issue, for each binary model representing a specific pollen type, we limited the training data to the season when that pollen type is present in the air. For the multiclass model, updating the weights based on all pollen classes simultaneously also presented a problem, as the model began to output more and more zeros as time passed. Therefore, we compared the model's output only for the pollen present in the air at that specific time.

The models were tested on 2020 data, and the best model for each pollen class was chosen. They were then tested on 2018 and 2019 data, resulting in improvements for almost every pollen class. Additionally, even with a slight improvement in the season, some models significantly improved their correlations when tested over a whole year, reducing false-positive identifications. For example, the correlation for Ambrosia in the season was improved by only 0.03, but when tested on the whole year, there was a 0.22 improvement. Additionally, some pollen types (such as Platanus) were significantly improved in the season (by 0.42) and the year (by 0.45).

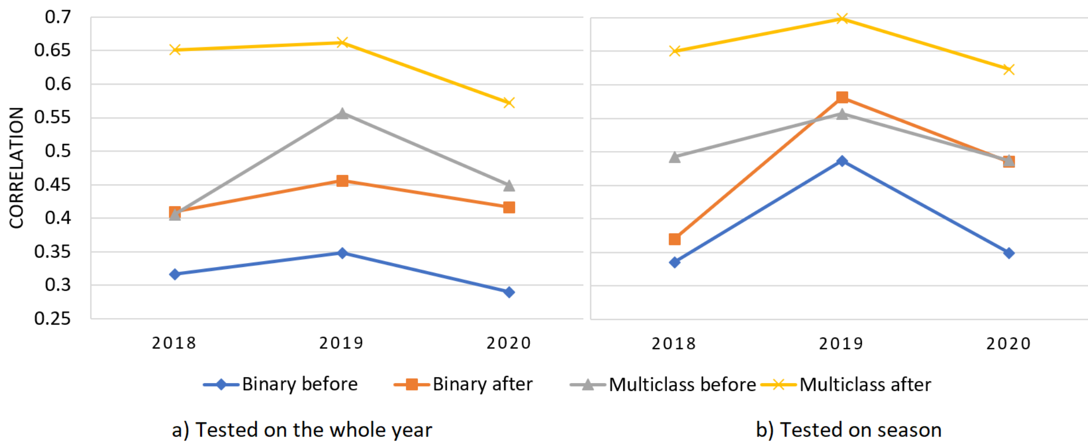

Fig. 1: Average correlation over multiple pollen types calculated a) over the whole year or b) over the pollen season, before and after retraining

We present only the average correlation over all pollen classes for a more straightforward result presentation. In the binary case, the average correlations are significantly improved (Figure 1). However, even the improved binary models are close to but do not exceed the results of the multiclass model before retraining. The multiclass model improves for the first 150 batches and then decreases performance. However, looking at the best result for each pollen class, independent of the model version, we obtain an average correlation improvement of 0.15 in season, while on the whole year, the correlation can be improved up to 0.25 (Figure 1).

4 Conclusion

In this study, we proposed the concept of retraining classification models using additional data sources. Utilizing CNNs, our classification model demonstrated promising results in previous studies compared to ground-truth Hirst data, achieving Pearson's correlation coefficients greater than 0.7 for 11 pollen types [2, 3]. By leveraging Hirst data, we aimed to improve the correlation with manual measurements. The retraining process involved comparing parts of two signals, calculating the MSE loss, and updating the network accordingly. Our results indicate a significant improvement, although further testing is necessary. The average correlation for binary classification models improved by 0.1, but it still falls short of the performance of the multiclass model prior to retraining. The performance of multiclass classification models is even more impressive, with a 0.15 increase in seasonal correlation and up to a 0.25 improvement over the course of a year. However, this approach requires training and utilizing one multiclass model for each pollen class.

Future research should focus on retraining a multiclass model for each of the 23 pollen classes independently, incorporating additional data out of season in a controlled manner. The retrained models should be evaluated using reference data, and the effect of time on training should be examined. Gradient descent should be implemented on the entire dataset, if possible, instead of the stochastic version, and cross-validation should be performed to obtain more robust results.

References

1. Sauliene, I., Sukiene, L., Daunys, G., Valiulis, G., Vaitkevicius, L., Matavulj, P., 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech. 12 (6), 3435–3452. doi: 10.5194/amt-12-1-2019.
2. Tesendic, D., Boberic-Krsticev, D., Matavulj, P., Brdar, S., Panic, M., Minic, V., Sikoparija, B., 2020. RealForAll: real-time system for automatic detection of airborne pollen. Enterp. Inf. Syst. doi: 10.1080/17517575.2020.1793391.
3. Matavulj, P., Cristofori, A., Cristofolini, F., Gottardini, E., Brdar, S., and Sikoparija., B. 2022. Integration of reference data from different rapid-E devices supports automatic pollen detection in more locations. The Science of the Total Environment 158234. doi: 10.1016/j.scitotenv.2022.158234.