Domain adaptation with unlabeled data for model transferability between airborne particle identifiers

Predrag Matavulj[1], Sanja Brdar[1], Miloš Racković[2], Branko Šikoparija[1], and Ioannis N. Athanasiadis[3]

[1]BioSense Institute, University of Novi Sad, Serbia
[2]Faculty of Sciences, University of Novi Sad, Serbia
[3]Geo-Information and Remote Sensing Group, Wageningen University and Research, The Netherlands

Abstract. As the most common causes of seasonal allergies, pollen affects approximately 30% of the world population. The proper information on the number of airborne allergens can significantly reduce its negative health and economic impact. For this reason, there is a growing network of automatic airborne particle monitors deployed. However, the calibration of such devices is a tedious task. Developing a deep learning classifier may allow model transferability between the devices. To investigate this approach, we employed data from two Rapid-E particle identifier devices, in a multi-class pollen identification task. We aim to improve the performance of models trained with data from one device and tested on another device. To our knowledge, this is the first attempt to apply any domain adaptation technique with unlabeled data between automatic airborne particle identifiers. Convolutional Neural Networks were constructed with two outputs to simultaneously perform pollen identification and domain adaptation. A simple gradient reversal layer between the domain classifier and the feature extractor promotes the emergence of not just discriminative features related to the classification task but also features invariant to the domain shifts in data. The development of a method for model transferability has a huge practical value for pollen monitoring since it reduces the costs of collecting labeled data.

Keywords: domain adaptation · pollen classification · neural networks · automatic pollen monitoring

1 Introduction

Each spring, summer and fall, plants release tiny pollen grains to fertilize other plants of the same species. In that period, quality of life significantly decreases for 30% of the world population sensitive to pollen, since pollen is one of the most common triggers of seasonal allergies [1]. Increased pollen production of wind-pollinated plants, the introduction of new, invasive allergenic plant species and human impact are some of the reasons why the sensations related to pollen

are intensified over time with more severe effects on health, including death [2, 3]. The information on the type and number of airborne allergens is a prerequisite for the prevention of allergy symptoms, which can significantly reduce the negative health and economic impact of this widespread non-communicable disease [4].

Pollen classification is mostly performed manually under a microscope. This is a long and costly procedure and the results become available with a delay of a few days. In 2018 in Europe, only 8 out of 525 counting devices were automatic [5]. The best current technologies for automatic pollen detection use laser-induced fluorescence and automatic multi-stack image recording [6]. Rapid-E is an airborne particle identifier that delivers laser-induced fluorescence information about airborne particles in real-time. It records scattered light and laser-induced fluorescence patterns, representing morphological and chemical fingerprints of airborne particles [7]. A convolutional neural network has been developed to identify 24 types of airborne particles with 65% accuracy [8]. This is the current state-of-the-art in terms of accuracy in classifying a high number of pollen classes.

To develop a classification algorithm, one should go through the difficult and costly procedure of obtaining labeled data. This requires collecting pollen samples and exposing them to the device in a controlled environment [7]. We have access to data from two devices, one in Novi Sad, Serbia and the other in Osijek, Croatia. The two devices produce data of the same type, but of different distributions – probably due to the sensibility of lasers and detectors. Therefore, in this work, we investigate how to transfer the model developed for one device to a second device, preferably with a few labeled data from the second device. This would allow us to train a model on collected labeled data from one device and to deploy that model on other devices, without collecting labeled data on those devices.

This study aims to adapt the model learned from one device to the other device, by using unlabeled data from both devices. To put it another way, we want to train a network to learn features that are invariant to domain changes. First, we adapted the domains on pre-trained models without looking at the class labels. We did so by using a gradient reversal layer, a method introduced in [9]. This way the fully connected layers are learning to recognize the domains, while the convolutional layers attempt to produce domain invariant features. This method does not take into account the class labels during training and therefore we will add an additional classifier to simultaneously learn discriminative and domain-invariant features, as proposed in [9].

The rest of the paper is organized as follows: Section 2 is divided into five subsections. Subsection 2.1 describes the data, preprocessing and the final dataset. In Subsection 2.2 a convolutional neural network used for classification purposes is described. Subsection 2.3 explains the expanded version of the network from Subsection 2.2 for dealing with the domains. Subsection 2.4 describes the experimental setup for this paper in detail, i.e. methods for testing the domain adaptation technique from Subsection 2.3 and evaluation with the evaluation metric described in Subsection 2.5. The results of the experiments are presented

in Section 3, along with the discussion. Finally, Section 4 concludes this paper and highlights areas for future research.

2 Methodology

2.1 Data collection

The Rapid-E devices are recording three types of data for each particle, namely scattered light image, fluorescence spectrum and fluorescence lifetime. After pre-processing, we obtained images of size 1x20x120, 1x4x32 and 1x4x24, respectively (Figure 1). Data preprocessing included normalizing the fluorescence signals into 0–1 range and converting them into images, centering the scattered light image and smoothing the signals with the Savitzky–Golay filter [10] to reduce the noise. The pre-processing steps are further detailed in [7, 8].

The scattering image represents morphological attributes of particles like shape and size. It is obtained by illuminating the particle with an infra-red laser multiple times, depending on the size and shape of the particle, while 24 detectors collect dispersed photons at different angles in the range of 45 to 135 degrees relative to the laser light beam.

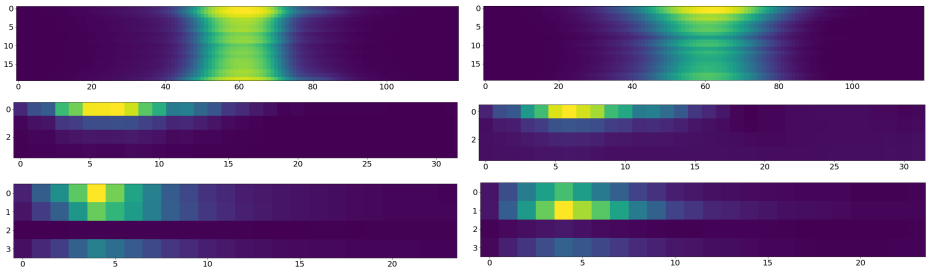
Additional descriptiveness is obtained with the fluorescence spectrum and lifetime, which represent the chemical properties of particles. Measured values of the fluorescence spectrum are derived from the excitation by deep-UV laser light at 337 nm. The spectral values include 32 measured values in the range of 350 to 800 nm, and these measurements were repeated eight times with an interval of 500 ns from the moment the laser excited the particle. Similarly, the fluorescence lifetime is recorded. Particle fluorescence duration is measured for four spectral ranges: 350-400 nm, 420-460 nm, 511-572 nm and 672-800 nm.

We collected data of two types: labeled and unlabeled data. Both were collected on both devices; one located in Novi Sad, Serbia (Station 1) and the other one in Osijek, Croatia (Station 2). Data have been recorded in 24 hours on the two devices.

Datasets From both stations, we had access to both labeled and unlabeled data. Labeled data were available for the same eight pollen labels in both stations. Table 1 summarizes the number of samples for each class and device.

As the devices work in real-time, we also had access to unlabeled data, i.e. data measured by the devices without knowing the substances therein. This data corresponds to any type of airborne particles, not just pollen. The unlabeled datasets comprise 14275 samples from station 1 and 16437 from station 2.

Low dimensional representation Initial exploration of data through low dimensional representation provided us valuable insight into differences between signals coming from different devices. Autoencoders [11] were utilized to learn a low dimensional representation of pollen data and results of spectrum dimensionality reduction presented in Figure 2 illustrate intrinsic differences caused



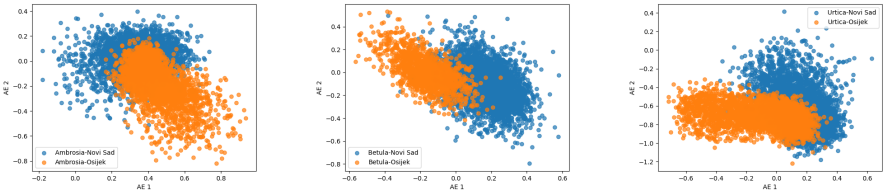
(a) Novi Sad (b) Osijek

Fig. 1: From top to bottom: Sample scattering image, fluorescence spectrum and fluorescence lifetime for the same particle. Station 1 in Novi Sad appears on the left, and Station 2 in Osijek on the right

Table 1: Number of samples for each class and each data source

Source	Station 1	Station 2
Alnus	1225	1075
Ambrosia	3126	2611
Artemisia	4958	2380
Betula	1614	877
Cedrus	522	528
Corylus	497	693
Quercus	563	360
Urtica	3876	3464
SUM	16381	11988

by devices. This raised concern that a classifier learned on data from one device could lose on its performance when used to classify particles on another device and served as the motivation to examine the transferability of classifiers between devices.



(a) (b) (c)

Fig. 2: Examples of low dimensional representation learned by autoencoder for the two stations: (a) ambrosia, (b) betula and (c) urtica

2.2 Multi-input CNN for pollen classification

We adopted the architecture of [7, 8] for pollen classification from Rapid-E inputs. The high complexity of data makes CNNs a suitable choice since they can operate automatic feature extraction. An additional benefit is the easy combination of multiple inputs. The three-input architecture, shown in Figure 3, assigns each input channel to the respective image type from the data. Each input channel is followed by two convolutional blocks (three for the fluorescence lifetime image) and one fully connected layer for each branch. Branches are then concatenated and one more fully connected layer is added with n neurons, where n is the number of classes and the softmax activation function is performed (Figure 3). In this way, the network learns to update the weights based on all inputs at the same time, since the gradient flows through the whole network [7, 8].

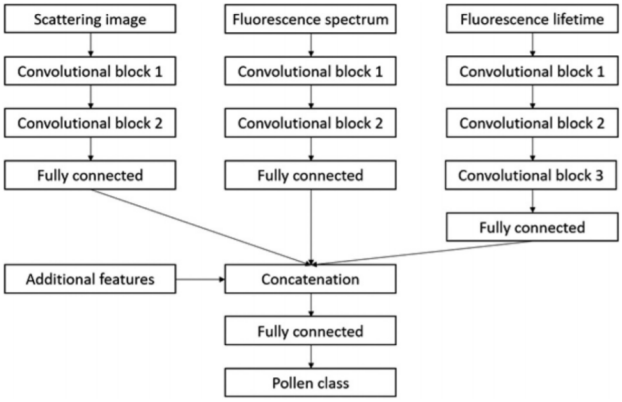


Fig. 3: Neural network architecture

2.3 Domain Adaptation

In this work, we assume that data comes from two different distributions, one from each device, and we want to bring those distributions as close as possible. Based on previous work on domain adaptation [9], we modified the above three-input architecture for pollen classification, by adding a second output that corresponds to a domain label (Figure 4).

The class label classifier is a regular fully-connected neural network. The parameters of both feature extractor and label classifier are updated in such a way to minimize the label prediction loss, which ensures that the features are discriminative and that the pollen classification yields good performance. Conversely, we want to find parameters of the feature extractor which maximize the domain loss to obtain similar feature distributions for two domains, while the domain classifier should be updated to minimize the domain loss. Our loss

is now defined as:

$$Loss = L_{label} + \lambda L_{domain}$$

The domain classifier comprises two fully-connected layers, with the gradient reversal layer (GRL) between the feature extractor and the domain classifier, resulting in the architecture presented in Figure 4. The role of the GRL is to make the features as bad as possible in classifying domain labels. In a forward pass, it acts as an identity. But in a backward pass, it multiplies the gradients with constant $-\lambda$. Now we move in the direction of the gradient, which leads to a local maximum of the loss function. λ controls the trade-off between the two objectives that shape the features.

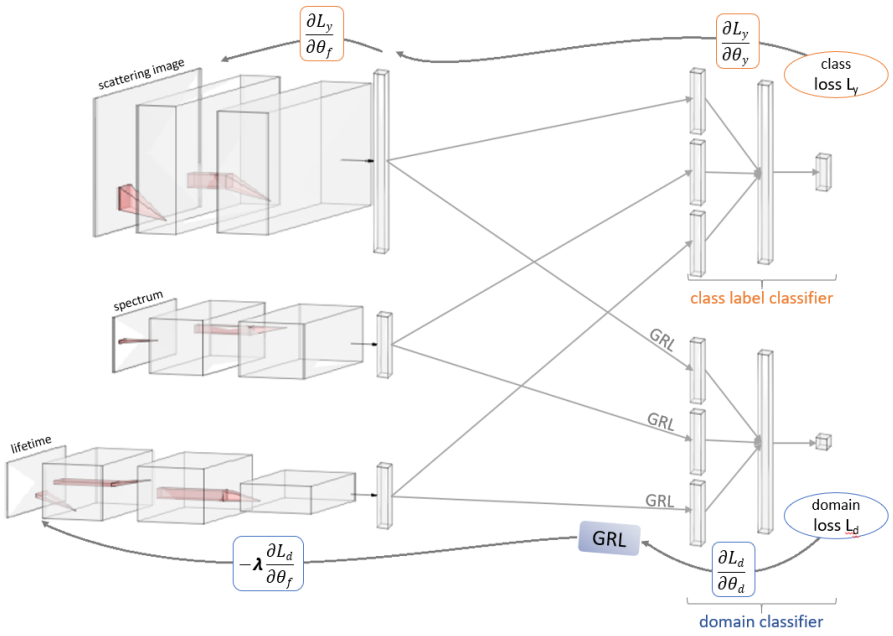


Fig. 4: The proposed architecture for domain adaptation, where θ_f represents the parameters of the feature extractors, and the parameters of the class label classifier and the domain classifier are denoted by θ_y and θ_d, respectively.

2.4 Experimental setup

We used the described pollen dataset for our experiments. Data were split in three, before training. The validation set consists of 10% of random samples from each class as well as the test set, and the rest 80% of the data is used for training. The model was fitted using the training set and evaluated against the validation set. In the end, the best models on the validation set were tested on the test set. Batches were created manually to represent all classes equally,

having 100 samples per class for the multi-class pollen classification and 500 samples per class for the domain adaptation.

We first trained a model on data from only one domain so that we can compare the test performance on both domains. This will give us a baseline which we will try to outperform.

Second, we adapted the domain of the baseline models, by adding a new fully-connected set of layers on top of the pre-trained feature extractor that will classify the domains with the GRL to obtain domain-invariant features. Since we train a whole network with a new classifier, a big gradient is expected, which will lead to a big change in the feature extractor. Therefore we try different learning rates to adjust the feature extractor rather than to change it completely. At the test phase, we will use the retrained feature extracting layers with the fully-connected classifier from the baseline.

Third, we create a new model that will employ the pollen class classifier simultaneously with the domain classifier to learn domain-invariant features relevant to the classification problem. We will explore different λ values representing different learning rates to obtain the best result. In the experiments below, we will adapt the domains with the unlabeled data as well as with the unlabeled class data as follows: each sample will be assigned a new label, depending on which device does the sample belongs to.

2.5 Measures

The measure for validating the models is the weighted F1 score (the Sorensen-Dice index, Dice similarity coefficient), which is more appropriate than accuracy since we have imbalanced data. To define the F1 score, first, we will define values of precision and recall. Consider we have a binary classification problem, and the obtained confusion matrix is:

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

The precision is defined as $\frac{TP}{TP+FP}$. It represents the ability of the classifier not to label a negative sample as positive. The recall is defined as $\frac{TP}{TP+FN}$ and it represents the ability of the classifier to find all the positive samples. Now, the F1 score is defined as the harmonic mean of the two values:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

It is in the range [0, 1], where 1 indicates perfect precision and recall and 0 indicates that either precision or recall (or both) are 0. Since we have a multi-class classification task, we will use the weighted F1 score which calculates the

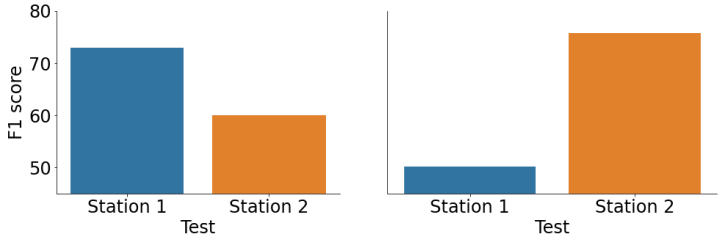
metric for each label and finds their average weighted by the number of instances for each label.

3 Results and discussion

3.1 Baseline

We have trained two models with the architecture presented in subsection 2.3 on labeled data from stations 1 and 2 and tested on both. For each station, we observe a performance drop between testing on the original station and the new station (Figure 5). For station 1 the performance drops by 13%, while for station 2 we observe a drop of 25.6%, almost twice as much as for station 1. We think that station 1 data is more generalized than data from station 2 and therefore the model trained on station 1 fits better on data from station two than the other way around.

Both models were initialized with the same predefined weights before training to have the same starting point. Two learning rates were explored: 0.001 and 0.0001. In both setups the model converges, but in different time frames. Furthermore, the error and the F1 score of a model trained with the higher learning rate oscillates more than with the lower learning rate. Therefore our final setup was to start training with a high learning rate and to gradually decrease the learning rate to obtain a steady error.



(a) Train: station 1 (b) Train: station 2

Fig. 5: Models trained on labeled data from a) station 1 and b) station 2, tested on both stations

3.2 Domain adaptation - training only with the domain classifier

The naive approach to obtain domain invariant features is to add new fully-connected layers on top of the pre-trained feature extractor from the baseline which will now classify the domains. Employing GRL and training the new network the features should become domain invariant over time. Different learning rates were explored: 0.01, 0.001, 0.0001, 0.00001, 0.000001 and 0.0000001 and the

models were trained for 100 epochs. We have tested our models on labeled data from both devices by replacing the domain classifier with the already trained class label classifier from the baseline. For each learning rate, we have experienced the same result as depicted in Figure 6. After some time, depending on the learning rate, the performances started to decrease on both stations, which was expected since we have adapted the domains without looking at the class labels. Therefore the feature extractor "forgot" how to be discriminative since its only concern was to produce domain-invariant features. The same results were obtained on both stations, with the unlabeled data and the unlabeled class data, as described in subsection 2.4.

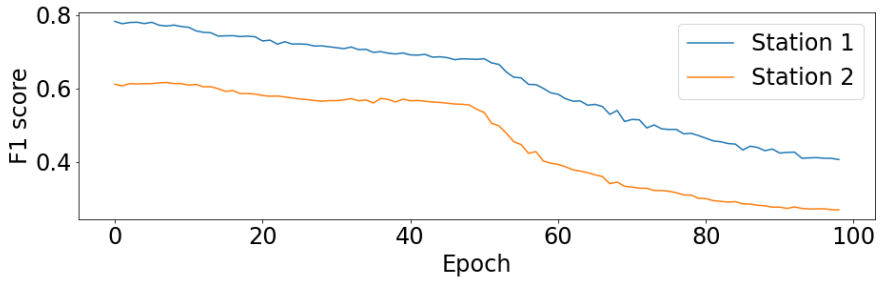


Fig. 6: Adapting domains on model trained on labeled data from station 1

3.3 Domain adaptation - simultaneous training of the domain and the class label classifier

Finally, we have trained a model to simultaneously classify class labels and the domains with architecture depicted in Figure 4. This model takes as input labeled class data from only one station to classify pollen and unlabeled data from both stations to classify the domains. Since we have two types of unlabeled data, we will have two experiments. The first experiment will use unlabeled class data and the second one will use random data when the devices were working in real-time.

The feature extractor and the class label classifier were initialized with the predefined weights from the baseline before training, while the domain classifier was initialized with a new set of random weights. We have used the same setup with the learning rate as for the baseline, starting with a higher one and gradually reducing it to obtain a steady error when the model converges. Along with the learning rate, we have evaluated different λ rates: 0.1, 1, 5, 10 and 20, defining how big of an impact the domain loss has on the overall loss. The best λ on the validation data was $\lambda = 5$, even though bigger lambdas such as 10 and 20 were on par with the best results.

The F1 score for the models trained on station 1 and tested on the same data stays approximately the same as for the baseline (in the range of 0.6%), while tested on station 2 the F1 score improves by 11.8% for experiment 1 with

the unlabeled class data, and 11% for the experiment 2 with the random data. The situation is similar to the models trained and tested on station 2. Tested on station 1, however, the model shows an improvement of 11.5% for experiment 1 and only 1.9% improvement for experiment 2 with the random data (Table 2). More examination needs to be done to explore this phenomenon. One of the reasons can be the choice of random data. Another may be the order in which we provided the data to the model during training. Similar results were obtained on the validation set, with negligibly small differences (Table 3).

Table 2: The F1 scores on test set

Train	Station 1		Station 2	
Test	Station 1	Station 2	Station 1	Station 2
Baseline	73	60	50.2	75.8
DA exp1	73.5	71.8	61.7	76.8
DA exp2	73.6	71	52.1	75.1

Table 3: The F1 score on validation set

Train	Station 1		Station 2	
Test	Station 1	Station 2	Station 1	Station 2
Baseline	78.4	60.7	56.7	77.6
DA exp1	76.9	72.9	65	79.4
DA exp2	77.8	72.4	54.9	79

4 Conclusion and future work

In this work, we have experimented with the domain adaptation method for training classification models in such a way so that the data shift between the domains is as small as possible, where the domains represent different Rapid-E airborne particle monitors. The devices produce data of the same type, but still there exists a shift in data distributions which can be seen in subsection 2.1. We have shown that we can improve the performance by approximately 11% when training on station 1 and testing on station 2 with both types of unlabeled data. The results were similar when training on station 2 and testing on station 1 but only for the experiment with the unlabeled class data. The experiment with the random data, however, did not show any improvement which needs to be investigated further since many stations will not have class data available.

Future work will include new random data and explore if the order in which we provide the data to the model during training affects the results. On the other hand, exploration of the number of classes needed for the domain adaptation with the unlabeled class data is necessary to show how much class data should

be gathered to obtain the best possible results. Future experiments should also include more stations to further validate our results. Finally, more advanced domain adaptation techniques should be considered.

Acknowledgements This research was funded by the BREATHE project from the Science Fund of the Republic of Serbia PROMIS program, under grant agreement no. 6039613 and partial financial support was received from the Ministry of Education, Science and Technological Development of the Republic of Serbia grant agreement no. 451-03-68/2020-14/200358. Measurement campaign in Novi Sad was supported by RealForAll project (2017HR-RS151) co-financed by the Interreg IPA Cross-border Cooperation program Croatia - Serbia 2014-2020 and Provincial secretariat for Science, Autonomous Province Vojvodina, Republic of Serbia (contract no. 102-401-337/2017-02-4-35-8). The results presented here relate to COST Action CA18226 "New approaches in detection of pathogens and aeroallergens" (ADOPT), www.cost.eu/actions/CA18226. This work was partially supported from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810775 (DRAGON).

References

1. Burbach GJ, Heinzerling LM, Edenharter G, Bachert C, Bindslev-Jensen C, Bonini S, et al. GA(2)LEN skin test study II: clinical relevance of inhalant allergen sensitizations in Europe. Allergy. 2009;64(10):1507–15
2. Beggs PJ. Allergen aerosol from pollen-nucleated precipitation: A novel thunderstorm asthma trigger. Atmos Environ. 2017;152:455–7.
3. Marselle, M. R., Korn, H., Irvine, K. N. (2019). Biodiversity and Health in the Face of Climate Change., Springer International Publishing, doi: 10.1007/978-3-030-02318-8
4. Zuberbier T, Lötvall J, Simoens S, Subramanian SV, Church MK. Economic burden of inadequate management of allergic diseases in the European Union: a GA(2) LEN review. Allergy. 2014 Oct;69(10):1275-9. doi: 10.1111/all.12470. Epub 2014 Aug 1. PMID: 24965386.
5. Buters JTM, Antunes C, Galveias A, Bergmann KC, Thibaudon M, Galán C, Schmidt-Weber C, Oteros J. Pollen and spore monitoring in the world. Clin Transl Allergy. 2018 Apr 4;8:9. doi: 10.1186/s13601-018-0197-8. PMID: 29636895; PMCID: PMC5883412.
6. Alex, H. J., A. E. Perring, N. J. Savage, B. Clot, B. Crouzy, F. Tummon, O. Shoshanim, et al. 2019. "Realtime Sensing of Bioaerosols: Review and Current Perspectives." Aerosol Science and Technology. 1–31. doi:10.1080/02786826.2019.1664724.
7. Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panić, M, Šikoparija, B., Clot, B., Crouzy, B., Sofiev, M. 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmospheric Measurement Techniques, 12, 3435-3452, doi: 10.5194/amt-12- 3435-2019
8. Tešendić D., Boberić Krstićev D., Matavulj P., Brdar S., Panić M., Minić V., Šikoparija B. 2020. RealForAll: real-time system for automatic detection of airborne pollen, Enterprise Information Systems, DOI: 10.1080/17517575.2020.1793391

9. Ganin Y., Lempitsky V., Unsupervised Domain Adaptation by Backpropagation., ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, 1180–1189
10. Savitzky, A., and M. J. E. Golay. 1964. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." Analytical Chemistry. 36(8):1627–1639.
11. Wang, Y., Yao, H. and Zhao, S., 2016. Auto-encoder based dimensionality reduction. Neurocomputing, 184, pp.232-242.