Towards automatic airborne pollen monitoring: From commercial devices to operational by mitigating class-imbalance in a deep learning approach

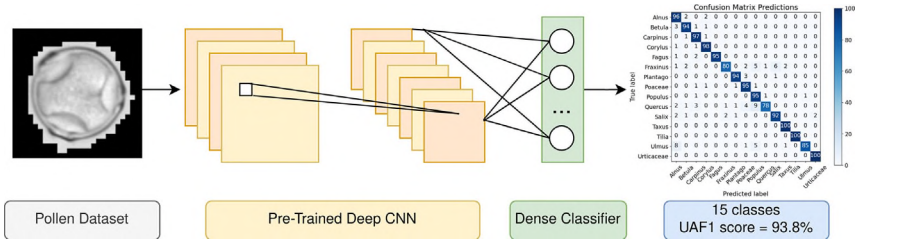
Jakob Schaefer [a], Manuel Milling [a], Björn W. Schuller [a,b], Bernhard Bauer [a], Jens O. Brunner [c], Claudia Traidl-Hoffmann [d,e], Athanasios Damialis [d,*]

[a] Chair of Embedded Intelligence for Health Care & Wellbeing, Faculty of Applied Computer Science, University of Augsburg, Augsburg, Germany
[b] GLAM, "The Group on Language, Audio & Music", Imperial College, London, UK
[c] Chair of Health Care Operations/Health Information Management, Faculty of Business and Economics, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[d] Department of Environmental Medicine, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[e] Institute of Environmental Medicine, Helmholtz Center Munich - German Research Center for Environmental Health, Augsburg, Germany

HIGHLIGHTS

- We used deep learning-based convolutional neural networks to classify pollen.
- Our algorithm was compared against the built-in of an automated device BAA500.
- We achieved an unweighted average F1 score of 93.8% across 15 allergenic taxa.
- The majority of pollen taxa (9 of 15) showed a recall of at least 95%.
- Deep learning algorithms can make automated pollen monitoring devices operational.

GRAPHICAL ABSTRACT



* Corresponding author at: Department of Environmental Medicine, Faculty of Medicine, University of Augsburg, Neusaesser str. 47, 86156 Augsburg, Germany.
E-mail address: thanos.damialis@tum.de (A. Damialis).

1. Introduction

Pollen-induced allergic rhinitis and allergic bronchial asthma pose a substantial burden on the quality of life for a large part of the worldwide population, particularly in the industrialised world (Brożek et al., 2017). In the context of emerging climate change and as the pollen seasons shift significantly earlier, and the pollen peaks have been evidenced as dramatically higher over the last few decades (Ziska et al., 2019; Anderegg et al., 2021), questions are raised regarding the magnitude of these effects on the severity and frequency of allergic diseases. Until today, the first line of defense for pollen allergies is avoidance of the allergen. This can be only achieved by acquiring reliable, accurate and timely information on the airborne pollen concentrations at a fine temporal resolution, so that allergic individuals and their treating practitioners may plan ahead their daily activities and the necessary medication. Muzalyova et al. (2021) have highlighted the existence of consistent diurnal pollen distribution patterns and the importance to take these into account in short-term operational, real-time forecasting models for the optimum allergy management. The importance of integrating hourly-resolution pollen measurements to forecasting models and, even more, using real-time data from novel, automatic monitoring devices has been suggested and discussed by Sofiev (2019), highlighting that such an approach could boost the predictive power of future models. Currently, allergic people are relying on conventional pollen information that exhibits a delay of 1–8 days (or more), as this measurement process involves laborious monitoring methods, requires high taxonomic expertise and achieves a forecasting horizon limited to the daily scale. The above limitations, as well as the necessity for upgrading to automated and short-term health information services, have been highlighted also by Geller-Bernstein and Portnoy (2018).

Hence, during the last decade approximately, automation in airborne pollen monitoring (and less frequently airborne allergenic fungal spores) has been adopted. Until today, continuous and intensive efforts have been made to commercialise such monitoring systems and make them operational so as to substitute the almost 70-year-old conventional method of the Hirst-design (Hirst, 1952). The newly developed techniques, being very promising, showcase already remarkable results and many positive aspects (among which automation and near-real-time temporal resolution); at the same time, with their development being underway, they exhibit also some disadvantages, common with the conventional Hirst-type technique too (among which reliability, comparability and price). While there might still be a long way until they lead the way in atmospheric biomonitoring, their progress is fast-pacing.

At the moment, only few countries stand out developing innovative monitoring sites. Among those, the first to establish such a network were in Japan, even though not able to distinguish among different pollen types (Kawashima et al., 2017); nonetheless, they have recently managed to improve their automation technique (Miki and Kawashima, 2021). Moreover, Germany has also been a pioneer (Oteros et al., 2020), where the Bavarian State has developed a network based on the automatic pollen monitoring devices BAA500 (Bio Aerosol Analyzer 500, Hund GmbH, Wetzlar) (Oteros et al., 2019, 2020). This technique has been described in detail in Oteros et al. (2015, 2020). Furthermore, automatic pollen and spore monitoring devices have been operating also in Lithuania (Šaulienė et al., 2019), Serbia (Šaulienė et al., 2019; Tešendić et al., 2020), and Switzerland (Crouzy et al., 2016; Šaulienė et al., 2019; Sauvageat et al., 2020). There are even more automated devices from additional countries and research teams, like from the U.S.A. (www.pollensense.com), nonetheless, no published information exists yet, to the best of our knowledge.

In Germany, based on previous research using the same automatic device as in the current study, it was stated (Oteros et al., 2015) that they achieved an accuracy score of 93.3% of correct positive classified cases versus the automatically classified cases. More recently, Oteros et al. (2020), in the frame of a State-funded network found that an automatic pollen monitoring network has achieved a 13-class identification average accuracy of 90%, similar to the result in 2015.

Nonetheless, the above findings, being the first of their kind and under lack of cross-validation against other research teams and methods, have been set already under dispute. Recent studies have shown on the one hand the non-biased, unfiltered and much lower performance of the commercial units (including a considerable amount of missing values, particularly during the pollen season peak) of these pollen monitoring devices (Schiele et al., 2019), but at the same time a great potential of automated pollen classification systems, only when trained on large data sets and with sophisticated statistical methods (de Geus et al., 2019; Schiele et al., 2019; Sevíllano et al., 2020).

In recent years, studies for automated classification of pollen grains have gained momentum. Several research groups have collected various types of image-based pollen data for this purpose. Deep learning-driven progress in computer vision has led to high recognition rates. Marcos et al. (2015) prepared pollen, which were collected by bees, in laboratory conditions, before acquiring magnified pollen images under a microscope. On a total dataset of 1800 images from 15 classes they achieved an accuracy of 95% using texture feature extraction and a k-nearest neighbour classification. Daood et al. (2016) utilised a two-stage classification approach based on feature extraction and support vector machines to identify pollen from 30 classes. The total of 10,063 images was provided by Florida Tech's Palaeoecology Laboratory. A 134-class pollen dataset, claimed to be the largest pollen dataset, was introduced by de Geus et al. (2019). The 3640 coloured pollen images were captured under a microscope after preparing the pollen with different reagents. Besides several approaches based on pre-designed feature, de Geus et al. (2019) applied pre-trained convolutional neural networks (CNNs) and achieved an accuracy of up to 96.24%. Further interest in automatic pollen classification has been sparked by the 2020 Pollen Challenge (https://iplab.dmi.unict.it/pollenclassificationchallenge/). The rise of Machine Learning and in particular Deep Learning has led to promising results towards health monitoring systems (Dong et al., 2020; Qian et al., 2021).

Even though part of the research has not been, yet, tested in 'real-life' monitoring conditions, it has been still shown in several cases how sophisticated analytical tools (convolutional neural networks among others) can make a big difference in the accuracy of the classification algorithms in the automated pollen monitoring systems, as highlighted by Gallardo-Caballero et al. (2019), Schiele et al. (2019), Daunys et al. (2021), etc.

The aim of this work was to go beyond the state-of-the-art in automatic pollen monitoring and the commercial pollen classification algorithms and optimise them to the best possible operational level. To achieve this, our approach was based on pre-trained convolutional neural networks (CNNs). We utilised a manually classified database of airborne pollen images, as derived from the automatic device BAA 500 (Hund GmbH, Wetzlar, Germany; as described in detail by Oteros et al., 2015) established in Augsburg, Germany. Finally, this database refers to the whole spectrum of pollen taxa (approximately 40 in total) detected throughout a whole pollen season (year 2016), completely avoiding filters, thresholds and any convenience samples that could bias our results.

Even though there is obviously still a long way to go to be able to discuss about fully operational networks that may provide real-time allergy risk alerts, our work here attempts to unveil the actual status of the research progress on the specific topic, but also the great potential for improvement.

2. Materials

As specified by Schiele et al. (2019), the pollen grains used in our research were gathered between November 2015 and October 2016 by an automated BAA 500 device located at ground-level, in Augsburg, Bavaria, Germany. Airborne pollen is trapped through an orifice on this device, by an intermittent high-throughput inflow of ambient air, thus collecting airborne particles on a sticky surface. A built-in light microscope equipped with a camera then captures images of each air sample and analyses them to extract crops of individual pollen grains. The latter are stored as an image library and then compared to an already existing, pre-classified image library, based on which the commercial device classifies airborne pollen per taxon.

Given the spherical shape of most pollen types, as soon as they are in contact to the sticky substance on the collection surface, most of the cropped images approximate a square shape with image sizes ranging from 56×56 to 179 pixels to 367×411 pixels. Even though the weakness of the BAA500 cropping algorithm has been pointed out before (Schiele et al., 2019), this pollen identification study relies on a high quality subset of the cropped images, which was carved out by a manual choice of suitable crops, and a parallel manual assignment of categorical (true/false) pollen labels.

In addition to the 15-class dataset presented in Schiele et al. (2019), here we introduce an extended and stricter-approached dataset containing samples of 31 different classes. Table 1 lists all considered pollen taxa in this study, as well as their abundance in our datasets. Pollen taxa marked with 'a' are part of the smaller data set Dataset-15, while all listed classes are part of the larger data set Dataset-31.

Fig. 1 shows an example image per each object class for all 31 classes. Table 2 summarises important characteristics of both data records. Since class imbalance was a major issue for both data sets, we introduced an imbalance indicator ρ, which is the ratio of the maximum and minimum number of samples in any class of the dataset.

2.1. Dataset-15

Our first, and smaller, dataset is composed of 15 classes, following the choice of Schiele et al. (2019): these 15 classes represent both the most common pollen types in the atmosphere of the majority of temperate-climate urban environments, as well as some of the most important allergenic pollen taxa worldwide. The overall 51,277 samples were unequally distributed among the classes, with the most frequent class *Corylus* having 11,667 observations, whereas the least frequent taxon *Tilia* accounted for 181 samples. Therefore, the imbalance ratio ρ is equal to 64.46.

Table 1
Pollen taxa (in alphabetical order) and their frequency in Dataset-31.

Latin name	Number of samples	Latin name	Number of samples
Alnus[a]	10063	*Picea*	23
Apiaceae	12	Pinaceae	450
Artemisia	76	*Plantago*[a]	1721
Betula[a]	2399	*Platanus*	63
Cannabaceae	12	Poaceae[a]	3600
Carpinus[a]	8010	*Populus*[a]	2066
Castanea	56	*Quercus*[a]	611
Chenopodiaceae	39	*Rumex*	80
Corylus[a]	11667	*Salix*[a]	526
Cyperaceae	30	*Taxus*[ab]	6944
Fagus[a]	728	*Tilia*[a]	181
Fraxinus[a]	460	*Ulmus*[a]	339
Juglans	110	Urticaceae[a]	2829
Larix	42	Fungal spores[c]	86
Papaveraceae	25	No pollen[c]	578

[a] The most abundant taxa, which are also included in the Dataset-15.
[b] *Taxus* refers to the total of objects deriving from both families Cupressaceae and Taxaceae.
[c] Classes that do not refer to pollen, Fungal Spores and No Pollen (air particles other than Pollen or Spores).

2.2. Dataset-31

Dataset-31 extends Dataset-15 by additionally including 2549 samples from 16 more classes, which are less frequently captured. This dataset was constructed by taking each class from the entire data record with at least 10 sample images. Due to the very low number of samples in specific taxa in this group, for example only 12 samples from the taxa of Apiaceae and Cannabaceae, the imbalance indicator dramatically increased to $\rho = 972.25$. An increased number of classes as well as a much higher imbalance indicator compared to Dataset-15 poses additional challenges to the handling of Dataset-31. However, Dataset-31 mitigates the problem of comparability pointed out by Schiele et al. (2019) as both the classifiers for Dataset-31, as well as the classification algorithm of the BAA500 are designed to distinguish between more than 30 classes. Nevertheless, a true comparison is actually not possible, as the BAA500 has been reported to be able to recognise at least 34 classes, based on the commercial, built-in image library, and therefore test sets are not identical and there is no possibility for direct comparisons and independent evaluation.

3. Methodology

Below, the major concepts are introduced that built up our deep learning-based classification approaches, as well as the evaluation metrics, which lead to our obtained results. The best configurations were found by running multiple experiments exploring a predefined hyperparameter space. All our approaches apply transfer learning, i.e., by fine-tuning a CNN that has been pre-trained on the ImageNet dataset. Moreover, data augmentation and weight penalties have been adopted as regularisation techniques to reduce overfitting. Finally, focal loss, class weights and weight vector normalisation were employed to mitigate biases resulting from datasets' imbalances.

3.1. Neural network design

Transfer learning approaches for CNNs are based on the idea that the first convolutional layers of a network learn generic features like edges which are crucial for most image-related tasks. Pre-training these layers on a large data set can often improve results on tasks with small data sets. In order to train the model on the small dataset, we replace the classification layer of the pre-trained model and we continue to optimise the model, which is referred to as fine-tuning. In some cases, it can be beneficial to freeze a certain number of the model's early layers, i.e., said layers are removed from the optimisation procedure. Our general network architecture is illustrated in Fig. 2. Four different architectures trained on the 2012 ImageNet image data set (Deng et al., 2009) are utilised and fine-tuned on the pollen data. We employ the publicly available networks ResNet50 (He et al., 2016a), ResNet101-V2 (He et al., 2016b), InceptionV3 (Szegedy et al., 2016) and DenseNet121 (Huang et al., 2017) as basemodels. The fully-connected layer on top of each basemodel was replaced by a dropout layer followed by a fully connected layer with softmax activation. The number of neurons in the latter layer is equal to the number of classes and consequently produces a normalised probability prediction for each class. Optimisation is based on the cross-entropy loss.

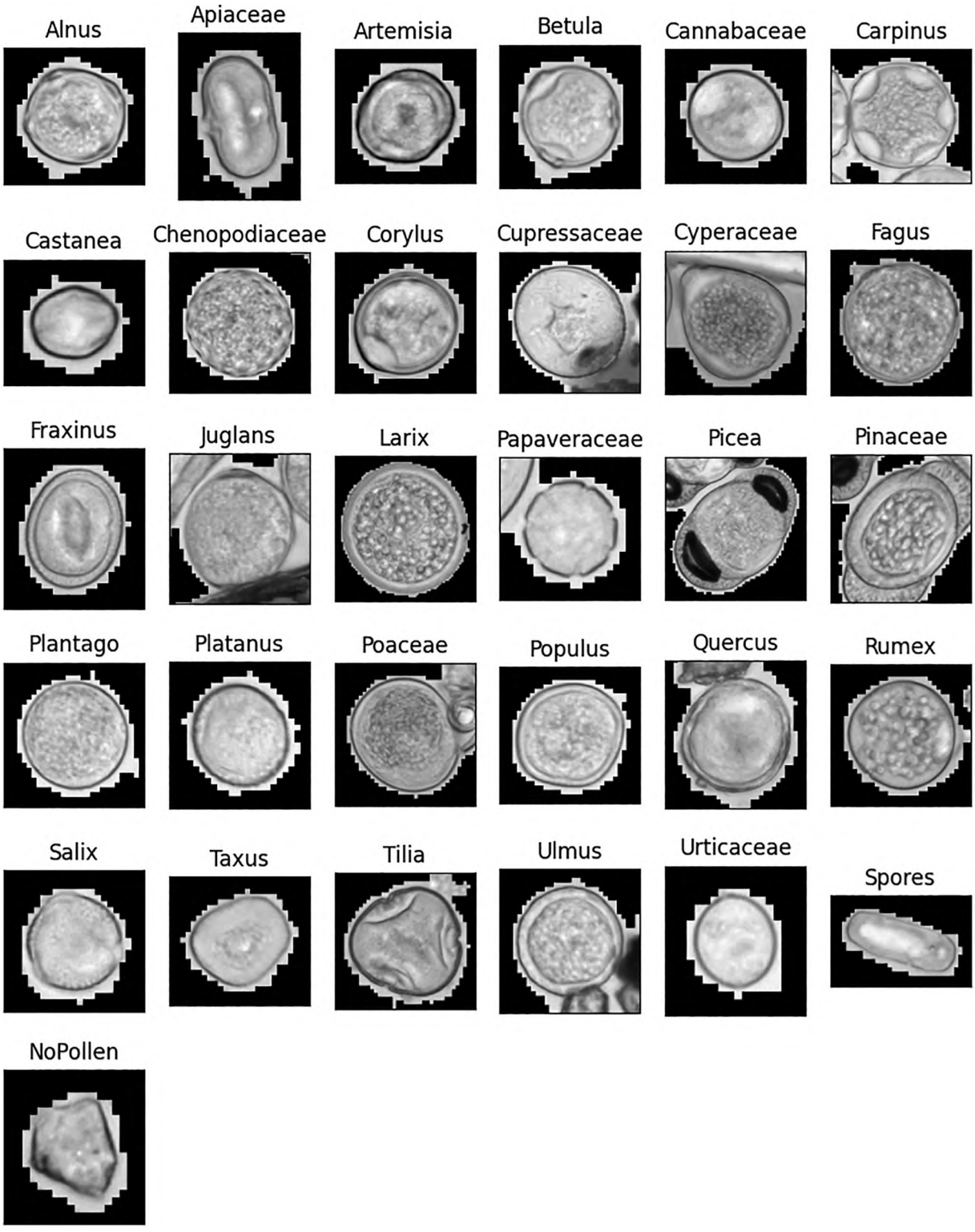


Fig. 1. Example images for each identified object class (29 different pollen types, Spores and NoPollen).
Size scales differ among classes, for better visualisation.

Table 2
Characteristics of the data sets.

	Dataset-15	Dataset-31
Number of classes	15	31
Number of samples	51277	53826
Imbalance ratio ϱ	64.46	972.25

3.2. Regularisation

We use several regularisation techniques in order to decrease the generalisation error. Besides the already mentioned dropout layer, this includes data augmentation and weight penalties.

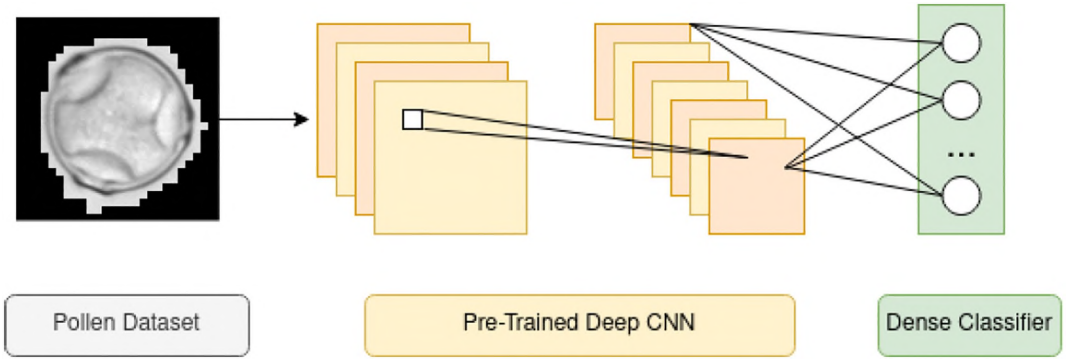


Fig. 2. Schematic network architecture consisting of a pre-trained basemodel as feature extractor and one-layered classifier.

3.2.1. Data augmentation

We aimed to make our model more robust against translational and rotational variances in images by artificially extending the training data with transformed images. We implemented five different data augmentation techniques, each of which was applied to each image that the algorithm 'reads' during training with 25% probability:

- Random crop (RC) to a bounding box, which contains between 70 and 90% of the original image pixels.
- Random rotation (RR) with angles between 45 and −45 degrees.
- Horizontal flip (HF).
- Vertical flip (VF).
- Additive white Gaussian noise (GN) with a mean of 0 and a standard deviation of 1, multiplied by a constant value of 15 is added to the pixel values in the range of [0; 255].

3.2.2. Weight penalties

We add a weight penalty - proportional to the L2 norm $\|\bullet\|_2$ of the network weights W - to the loss function L, in order to encourage small weights and therefore less complex models. The updates for the weights W from step t to step $t+1$ can be expressed as

$$W_{t+1} = W_t - \eta \nabla_{W_t}(L + \lambda ||W_t||_2), \tag{1}$$

with the learning rate η and the weight penalty factor λ (Goodfellow et al., 2016). We set $\lambda = 0.0005$ in our experiments.

3.3. Class imbalance

A major problem of the classification tasks, especially for Dataset-31, is a strong imbalance between the different classes. The bias towards the majority of classes can either be mitigated by altering the training data to decrease imbalance or by modifying the model's underlying learning process to increase sensitivity towards the minority group. In our approach, we focus on different algorithm-level techniques, class weighted loss, focal loss, and weight vector normalisation, to shift the bias towards the minority classes (Johnson and Khoshgoftaar, 2019).

3.3.1. Class weighted loss

To increase the importance of minority classes, the learning process can be adjusted by assigning weights to samples to match a given data distribution. Our approach is to weigh the cross-entropy loss with a factor equal to the ratio between the number of samples in the majority class n_{max} and the considered class n_i according to Schiele et al. (2019):

$$L_\alpha = \frac{n_{max}}{n_i} \log(p_i). \tag{2}$$

3.3.2. Focal loss

The focal loss designed by Lin et al. (2020) addresses the extreme imbalance between foreground and background classes in an object detection task. The loss function weighs down easy examples such that they contribute less to the total loss, even if their number is large. Although the initial purpose of focal loss was to improve object detection, it has since been successfully applied to different classification tasks (Nemoto et al., 2018; Wang et al., 2020). Focal loss is derived from cross entropy, which measures the deviation of the predicted probability from the actual label, and can be calculated as

$$FL(p_i) = -(1-p_i)^\gamma \log(p_i), \tag{3}$$

where p_i is the model's estimated probability for the ground-truth class and γ is a hyperparameter. Samples from minor classes tend to have lower prediction probabilities for the ground truth class. By increasing the value of γ, we increase the contribution of seemingly more difficult samples to the overall loss. In our experiments, we set $\gamma = 2$.

3.3.3. Weight vector normalisation

Previous research by Kim and Kim (2020) suggests that imbalanced data leads to a difference in weight vector norms in the classification layer of a neural network, which causes a decision boundary bias towards classes with lower sample frequency. It is hypothesised that a normalisation of the weight vectors and an adjustment of the decision boundary reduces this bias. During the training, we therefore normalise all weight vectors w_i – connecting the second-to-last layer of the network with the ith neuron of the classification layer – after each gradient descent step according to

$$w_i \leftarrow \frac{w_i}{||w_i||_2}. \tag{4}$$

After the training is finished, we finally adjust the decision boundary by rescaling all weight vectors of the classification layer applying

$$w_i \leftarrow \left(\frac{n_{max}}{n_i}\right)^\beta w_i, \tag{5}$$

where again n_i denotes the number of samples belonging to class i and n_{max} is the maximum number of samples of one class in the data set. The hyperparameter β correlates with the feature space size for infrequent classes, with higher values of β leading to increased feature space sizes. In our experiments we employ $\beta = 0.3$.

3.4. Evaluation metrics

The performance evaluation of our experiments is based on the common metrics precision $p = TP/(TP + FP)$, recall $r = TP/(TP + FN)$, and F1 measure $f_1 = 2pr/(p + r)$, expressed in terms of true (T) and false (F) positives (P) and negatives (N). Considering the imbalance of the data

set and an equal importance of each pollen taxa, these metrics are calculated for each class before the unweighted average is formed, i.e., we report unweighted average precision (UAP), unweighted average recall (UAR), and unweighted average F1 measure (UAF1) for each experimental setup. To compute the latter quantity, we choose the approach of the unweighted arithmetic mean over harmonic means, which can be expressed as

$$\mathrm{UAF1} = \frac{1}{N}\sum_{x}^{N} \mathrm{F1}(x) = \frac{1}{N}\sum_{x}^{N} \frac{2P(x)R(x)}{P(x)+R(x)}, \tag{6}$$

with N being the number of classes.

3.5. Experiments

Before conducting our experiments, we randomly split the available data set into a training (60% of the data), a validation, and test set (20% of the data, each). The validation set is only used to evaluate the performance of the model and is excluded from gradient-descent optimisation. Every experimental setup is trained five times to detect statistical variance and is evaluated on the test set based on the UAP, UAR, and UAF1 measures. Potential overfitting could be detected by monitoring the performance on the validation set. During training, whenever the validation F1 measure stops improving for more than four epochs, the learning rate is halved. A setup is generally trained for 40 epochs deploying the Adam optimiser (Kingma and Ba, 2014) and a mini-batch size of 64 samples. Based on an adapted strategy of early stopping we choose the model state, which achieves the best UAF1 measure on the validation set, for evaluation on the test set.

For each data set, we investigate 30 experimental setups with different hyperparameter configurations. Hyperparameters are assigned to one of the three categories basemodel, regularisation, and class balancing, the former of which includes the initial learning rate as well as the number of layers to freeze. Hyperparameter optimisation is performed in an iterative manner by subsequently optimising each of the mentioned categories. Even though this procedure might not lead to the best possible hyperparameter configuration, it seems the most reasonable approach under the given resource limitations.

4. Results

The best among the tested configurations for each data set is shown in Table 3.

4.1. Results for Dataset-15

For Dataset-15, the best setup uses the DenseNet121 architecture with an initial learning rate of 10^{-4}, while the first 36 layers in this network are excluded from weight updates during training. We combine that with a dropout rate of 50% and L2-normalised weight penalties. In addition, augmentation in the form of vertical and horizontal flips plus random rotations and crops is employed. Focal loss and weight vector normalisation are applied to reduce the bias of the learning algorithm towards minority classes. Table 4 shows the obtained results by our model, compared with the results obtained also by Schiele et al. (2019) and the BAA 500 internal classification (commercial) algorithm. Note that the considered random data split is different from the evaluation in Schiele et al. (2019).

Table 3
Overview of the network's best configurations.

	Dataset-15	Dataset-31
Basemodel	DenseNet121 Initial LR 10^{-4} First 36 layers frozen	InceptionV3 Initial LR 10^{-4} No layers frozen
Regularisation	Dropout rate 0.5 L2 weight penalties Augmentations: HF, VF, RR, RC	Dropout rate 0.3
Class balancing	Focal loss WVN	Class weighted loss WVN

Table 4
Results on the test data sets (in %) for 15 object classes (pollen types), compared among three different classification algorithms.

Algorithms tested	UAP	UAR	UAF1
Commercial algorithm (Schiele et al., 2019)	59.4	54.5	56.4
Algorithm by Schiele et al. (2019)	83.0	77.1	79.1
Current algorithm	94.3 ± 0.4	93.5 ± 0.1	93.8 ± 0.1

For our algorithm, we report the average results and standard deviations over five runs.

Overall, as can be concluded from Table 4, our model manages to improve the commercial system's rates of UAP, UAR, and UAF1 relatively by 58.8%, 71.6% and 66.3%, respectively. Our proposed model shows superior performance, achieving a UAP of 94.3%, a UAR of 93.5%, and a UAF1 of 93.8%. The normalised confusion matrix depicted in Fig. 3 shows the performance of the model for each individual class. Each row of the confusion matrix considers the samples, which are assigned to one given class, considering the ground truth, and indicates the percentage of these samples, which are assigned to each class by the classification algorithm. The confusion matrix shows very high performance measures for all classes, with the majority (9 out of the 15 classes) showing a recall equal to or greater than 95% and only *Quercus* barely missing a recall of 80%. The pollen types of *Taxus*, *Tilia* and Urticaceae show a remarkable 100% recall rate.

4.2. Results for Dataset-31

Table 5 shows the results obtained by our model on the Dataset-31 data set, next to an evaluation of the BAA 500 classification based on 34 classes done by Schiele et al. (2019). The best setup relies on a completely trainable InceptionV3 backbone with an initial learning rate of 10^{-4}. A dropout rate of 30% is utilised for regularisation, class weighted loss, and weight vector normalisation mitigate the high class imbalance. We are aware that a direct comparison between our approach and the BAA 500 algorithm is not feasible, since the latter has been designed for an even broader variety of classes and not all classes of Dataset-31 are present in the 34 class-evaluation of the BAA 500

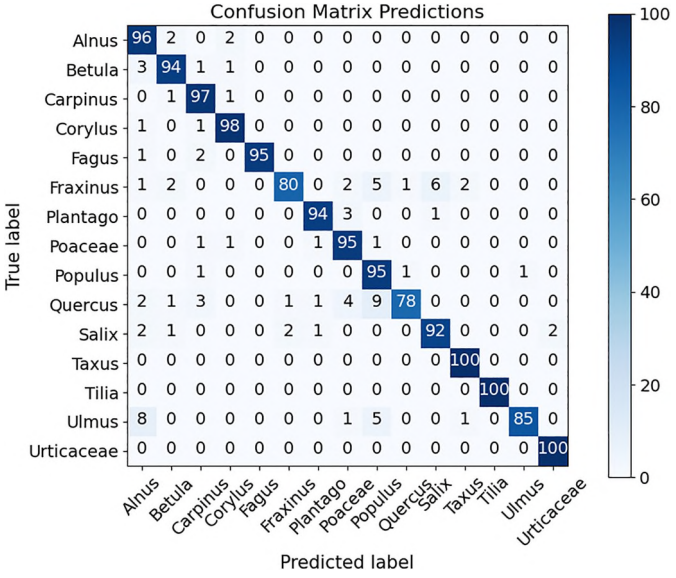


Fig. 3. Normalised confusion matrix for our best Dataset-15 configuration, averaged over five runs.
Results are rounded to integer percentage points.

Table 5
Results on the test data sets (in %) for 31 object classes, compared among two different classification algorithms.

Algorithms tested	UAP	UAR	UAF1
Algorithm by Schiele et al. (2019)	66.6	62.3	60.1
Current algorithm	74.9 ± 2.7	78.3 ± 1.7	75.9 ± 1.8

For our algorithm, we report the average results and standard deviations over five runs.

evaluation. The main reason for this is that only the classes present in Dataset-31 offer at least 10 labelled samples within the collected data. Nevertheless, the 15.8% difference in UAF1 score leads us to expect that our approach will still be competitive in a fair comparison with the BAA 500 algorithm.

The confusion matrix in Fig. 4 shows that many classes still achieve very high scores and only few classes achieve a rather low score. Overall, as expected, the least performing classes refer to the smallest classes in the database (Apiaceae: 12 samples, recall = 16%; *Platanus*: 63 samples, recall = 29%), whereas the best performances were observed in the most abundant classes, i.e. those included in Dataset-15; in Dataset-31, the recall drops, but never lower than 79%.

5. Discussion

In this paper, we investigated transfer learning-based CNN models for classifying airborne pollen grains. Our models were evaluated based on two large data sets, which were collected by a BAA 500 device. We have used different techniques of regularisation and class balancing to cope with emerging issues of overfitting and bias towards majority classes. Our best models achieve an unweighted F1 measure of 93.8% across 15 classes and an unweighted average F1 measure of 75.9% across 31 classes. In the 15-class model, the majority of classes achieves a recall higher than 95%.

Our findings here point out that there is plenty of room for improvement in the commercial, built-in algorithm of BAA 500, although a direct comparison to the BAA 500 algorithm proves to be difficult. At this stage, we managed to obtain relative improvements of 58.8% - 71.6% depending on the considered evaluation metric.

Our results showed a remarkable prediction of specific pollen types, namely those of *Taxus*, *Tilia*, and Urticaceae, which are among the easiest to microscopically classify. On the other hand, the worst predictions

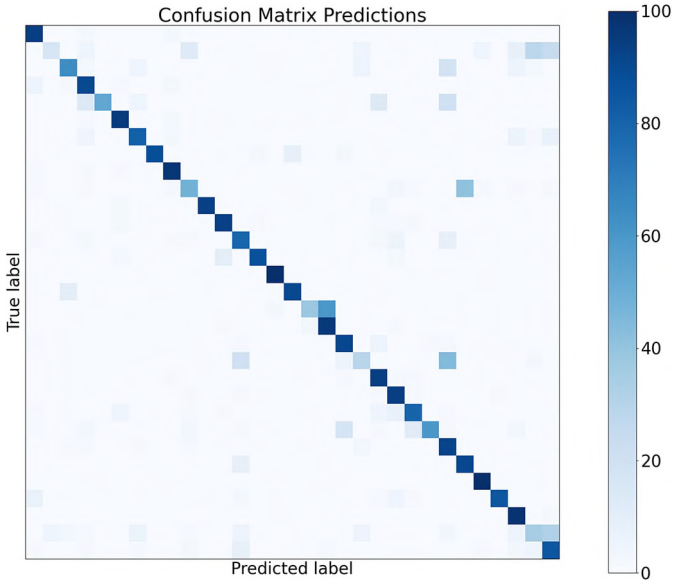


Fig. 4. Normalised confusion matrix for our best Dataset-31 configuration, averaged over five runs.

of our algorithm are for the pollen types of (in ascending order of accuracy): *Quercus*, *Fraxinus*, and *Ulmus*. These taxa are among those with the smallest abundance in the database, but also some could be characterised of comparatively higher difficulty to manually classify their pollen, as it might be in the case of *Quercus*; apparently, this is also dependent on the co-occurrence of other pollen taxa too, as well as the environmental conditions in the atmosphere that may obscure the quality of the pollen samples. Therefore, one can observe that our own-developed algorithm seems to exhibit the same advantages, disadvantages and particularities like the conventional manual classification procedure, even though to a lesser extent, because of the automation.

Since the length of the datasets is obviously of decisive importance, we have been already expanding our dataset by labelling additional pollen samples, which have been continuously collected by two different BAA 500 devices, from two different locations and from two different years. We anticipate that such a dataset would allow not only for cross-calibration and improvement of existing algorithms, but also for testing the comparability of the monitoring devices.

Moreover, for further improving classification algorithms, the hyperparameter space seems to be important to explore. Due to the fast pace of the deep learning research community, state-of-the-art models in computer vision at present tend to become outdated quickly. Experiments with novel pre-trained base models might therefore lead to even better results, for instance, utilising Meta Pseudo Labels, the currently best performing model on ImageNet (February 2021) (Pham et al., 2020). Further, data-level methods for class-balancing have been successfully applied to other related problems, for instance, in the classification of ocean plankton (Lee et al., 2016).

Moreover, as has been highlighted by Schiele et al. (2019), the cropping technique of acquired images via the commercial devices is not the most efficient. It clearly appears to be effective mainly for round objects, thus focusing on the usually circle-shaped pollen grains, which, however, fails to correctly classify non-round objects, either deformed or broken pollen grains, or other airborne particles, like variable-shaped fungal spores. This, exactly, is one of the limitations of our study presented here: our currently developed algorithm still relies on a flawed cropping algorithm of the commercial BAA 500 device. The ultimate goal would be to move forward to an operational system of real-time, automatic pollen monitoring, which would improve and, when needed, bypass the existing commercial algorithms and other units' traits.

Given that airborne pollen monitoring usually is conducted in the frame of environmental health services to inform and protect allergic individuals on high-risk time intervals, it is timely and important to develop the most accurate monitoring systems. Nevertheless, most pollen and spore monitoring networks are not publicly funded and data are not freely available; towards this direction, an increasing number of various automatic pollen and spore monitoring systems across the world has been established (Buters et al., 2018). The fundamental question immediately raised is whether there is one single pollen monitoring system that can outperform the others and this can assess the 'genuine pollen exposure'.

Albeit an increasing number of countries have already established innovative automated pollen and spore monitoring devices, only few are completely operational or belonging to open real-time information networks. Still, it is obvious that the bioaerosol monitoring methods gradually move to a new era of automated systems. Hence, there is an apparent need for the evaluation of the reliability, performance and accuracy of emerging automatic devices, particularly if one takes into account the current cost of renting or purchasing such an automatic device. Towards this direction, dedicated campaigns have been already set up across Europe (Clot et al., 2020). The establishment of external, independent panels of experts to conclude on this seems essential.

In the long run, the herein presented deep learning models can provide a valuable tool for accurate and real-time pollen classification algorithms in some of the existing automatic monitoring devices. Such

advancements, in combination with hardware improvements, and online automated platforms for data flow, including mobile technologies, will all significantly contribute to the optimum provision of most efficient allergenic pollen information services. In combination with novel pollen apps, a real-time environmental health service can be created for the benefit of those affected by aeroallergens, usable as first-line of defense against high-risk pollen exposure intervals.

Such pioneer methods and automation in detection methods, along with real-time, open-access pollen data, can additionally serve a huge societal purpose: not only they would be necessary as a first-line prophylaxis tool against allergic diseases, but also for emerging health risks as for viral infections. It has been recently reported by Damialis et al. (2021) that airborne pollen concentrations are positively correlated with increased SARS-CoV-2 infection rates (Damialis et al., 2021). Real-time, automated, pollen flight information is more important than ever, especially if climate change effects are also considered: earlier pollen seasons and higher pollen abundances (Ziska et al., 2019; Anderegg et al., 2021) would make the co-occurrence of any winter-spring viruses and high pollen concentrations a common phenomenon with additive health implications in the forthcoming decades.

CRediT authorship contribution statement

JS, MM: Conceptualization, methodology, data analysis, visualization, writing original draft, reviewing and approving final draft; BS, BB: methodology, reviewing and approving final draft; JOB, CTH: provision of aerobiological data, funding acquisition, reviewing and approving final draft; AD: provision of aerobiological data and expertise, data curation, writing original draft, reviewing and approving final draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The study was partly implemented in the frame of the EU-COST Action ADOPT (New approaches in detection of pathogens and aeroallergens), Grant Number CA18226 (EU Framework Program Horizon 2020).

CTH was supported by:

a) the Helmholtz Climate Initiative (HI-CAM), Mitigation and Adaptation.

b) Christine Kühne–Center for Allergy Research and Education (CK-CARE).

References

Anderegg, W.R.L., Abatzoglou, J.T., Anderegg, L.D.L., Bielory, L., Kinney, P.L., Ziska, L., 2021. Anthropogenic climate change is worsening North American pollen seasons. Proc. Natl. Acad. Sci. 118, e2013284118.

Brożek, J.L., Bousquet, J., Agache, I., Agarwal, A., Bachert, C., Bosnic-Anticevich, S., Brignardello-Petersen, R., Canonica, G.W., Casale, T., Chavannes, N.H., Correia de Sousa, J., Cruz, A.A., Cuello-Garcia, C.A., Demoly, P., Dykewicz, M., Etxeandia-Ikobaltzeta, I., Florez, I.D., Fokkens, W., Fonseca, J., Hellings, P.W., Klimek, L., Kowalski, S., Kuna, P., Laisaar, K.-T., Larenas-Linnemann, D.E., Lødrup Carlsen, K.C., Manning, P.J., Meltzer, E., Mullol, J., Muraro, A., O'Hehir, R., Ohta, K., Panzner, P., Papadopoulos, N., Park, H.-S., Passalacqua, G., Pawankar, R., Price, D., Riva, J.J., Roldán, Y., Ryan, D., Sadeghirad, B., Samolinski, B., Schmid-Grendelmeier, P., Sheikh, A., Togias, A., Valero, A., Valiulis, A., Valovirta, E., Ventresca, M., Wallace, D., Waserman, S., Wickman, M., Wiercioch, W., Yepes-Nuñez, J.J., Zhang, L., Zhang, Y., Zidarn, M., Zuberbier, T., Schünemann, H.J., 2017. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines-2016 revision. J. Allergy Clin. Immunol. 140, 950–958.

Buters, J.T.M., Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allergy 8, 9.

Clot, B., Gilge, S., Hajkova, L., Magyar, D., Scheifinger, H., Sofiev, M., Bütler, F., Tummon, F., 2020. The EUMETNET AutoPollen programme: establishing a prototype automatic pollen monitoring network in Europe. Aerobiologia https://doi.org/10.1007/s10453-020-09666-4.

Crouzy, B., Stella, M., Konzelmann, T., Calpini, B., Clot, B., 2016. All-optical automatic pollen identification: towards an operational system. Atmos. Environ. 140, 202–212.

Damialis, A., Gilles, S., Sofiev, M., Sofieva, V., Kolek, F., Bayr, D., Plaza, M.P., Leier-Wirtz, V., Kaschuba, S., Ziska, L.H., Bielory, L., Makra, L., del Mar Trigo, M., COVID-19/POLLEN study group, Traidl-Hoffmann, C., 2021. Higher airborne pollen concentrations correlated with increased SARS-CoV-2 infection rates, as evidenced from 31 countries across the globe. Proc. Natl. Acad. Sci. 118, 2019034118.

Daood, A., Ribeiro, E., Bush, M., 2016. Pollen recognition using a multi-layer hierarchical classifier. 23rd Int. Conf. Pattern Recognit. (ICPR), pp. 3091–3096.

Daunys, G., Šukienė, L., Vaitkevičius, L., Valiulis, G., Sofiev, M., Šaulienė, I., 2021. Clustering approach for the analysis of the fluorescent bioaerosol collected by an automatic detector. PLoS ONE 16, e0247284.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conf. Computer Vision Pattern Recognit, pp. 248–255.

Dong, F., Qian, K., Ren, Z., Baird, A., Li, X., Dai, Z., Dong, B., Metze, F., Yamamoto, Y., Schuller, B.W., 2020. Machine listening for heart status monitoring: introducing and benchmarking HSS—the heart sounds Shenzhen corpus. IEEE J. Biomed. Health Inform. 24 (7), 2082–2092.

Gallardo-Caballero, R., García-Orellana, C.J., García-Manso, A., González-Velasco, H.M., Tormo-Molina, R., Macías-Macías, M., 2019. Sensors 19, 3583.

Geller-Bernstein, C., Portnoy, J.M., 2018. The clinical utility of pollen counts. Clin. Rev. Allergy Immunol. 57, 340–349.

de Geus, A.R., Barcelos, C.A.Z., Batista, M.A., de Silva, S.F., 2019. Large-scale Pollen Recognition With Deep Learning. 27th Eur. Signal Process. Conf. (EUSIPCO). pp. 1–5.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press http://www.deeplearningbook.org.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. IEEE Conf. Computer Vision Pattern Recognit. (CVPR), pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. Eur. Conf. Computer Vision, pp. 630–645.

Hirst, J.M., 1952. An automatic volumetric spore trap. Ann. Appl. Biol. 39, 257–265.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. IEEE Conf. Computer Vision Pattern Recognit. (CVPR), pp. 2261–2269.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 27.

Kawashima, S., Thibaudon, M., Matsuda, S., Fujita, T., Lemonis, N., Clot, B., Oliver, G., 2017. Automated pollen monitoring system using laser optics for observing seasonal changes in the concentration of total airborne pollen. Aerobiologia 33, 351–362.

Kim, B., Kim, J., 2020. Adjusting decision boundary for class imbalanced learning. IEEE Access. vol. 8, pp. 81674–81685.

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv, 1412.6980.

Lee, H., Park, M., Kim, J., 2016. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. 2016 IEEE Int. Conf. Image Process. (ICIP), pp. 3713–3717.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42, 318–327.

Marcos, J.V., Nava, R., Cristóbal, G., Redondo, R., Escalante-Ramírez, B., Bueno, G., Déniz, Óscar, González-Porto, A., Pardo, C., Chung, F., Rodríguez, T., 2015. Automated pollen identification using microscopic imaging and texture analysis. Micron 68, 36–46.

Miki, K., Kawashima, S., 2021. Estimation of pollen counts from light scattering intensity when sampling multiple pollen taxa – establishment of an automated multi-taxa pollen counting estimation system (AME system). Atmos. Meas. Tech. 14, 685–693.

Muzalyova, A., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2021. Forecasting *Betula* and Poaceae airborne pollen concentrations on a 3-hourly resolution in Augsburg, Germany: toward automatically generated, real-time predictions. Aerobiologia https://doi.org/10.1007/s10453-021-09699-3.

Nemoto, K., Hamaguchi, R., Imaizumi, T., Hikosaka, S., 2018. Classification of rare building change using CNN with multi-class focal loss. IEEE Int. Geosci. Remote Sensing Symp, pp. 4663–4666.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T., 2015. Automatic and online pollen monitoring. Int. Arch. Allergy Immunol. 167, 158–166.

Oteros, J., Sofiev, M., Smith, M., Clot, B., Damialis, A., Prank, M., Werchan, M., Wachter, R., Weber, A., Kutzora, S., Heinze, S., Herr, C.E., Menzel, A., Bergmann, K.-C., Traidl-Hoffmann, C., Schmidt-Weber, C.B., Buters, J.T., 2019. Building an automatic pollen monitoring network (ePIN): selection of optimal sites by clustering pollen stations. Sci. Total Environ. 688, 1263–1274.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition. Environ. Res. 191, 110031.

Pham, H., Xie, Q., Dai, Z., Le, Q.V., 2020. Meta Pseudo Labels. arXiv, 2003.10580.

Qian, K., Zhang, Z., Yamamoto, Y., Schuller, B.W., 2021. AIoT for the elderly: an overview from assisted living to healthcare monitoring. IEEE Signal Process. Mag. 38.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, 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. Atmos. Meas. Tech. 12, 3435–3452.

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., Vasilatou, K., 2020. Real-time pollen monitoring using digital holography. Atmos. Meas. Tech. 13, 1539–1550.

Schiele, J., Rabe, F., Schmitt, M., Glaser, M., Haring, F., Brunner, J.O., Bauer, B., Schuller, B., Traidl-Hoffmann, C., Damialis, A., 2019. Automated classification of airborne pollen using neural networks. 41st Annual Int. Conf. IEEE Engin. Med. Biol. Soc. (EMBC), pp. 4474–4478.

Sevíllano, V., Holt, K., Aznarte, J.L., 2020. Precise automatic classification of 46 different pollen types with convolutional neural networks. PLoS ONE 15, e0229751.

Sofiev, M., 2019. On possibilities of assimilation of near-real-time pollen data by atmospheric composition models. Aerobiologia 35, 523–531.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2818–2826.

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. Enterp. Inf. Syst. https://doi.org/10.1080/17517575.2020.1793391.

Wang, L., Wang, C., Sun, Z., Cheng, S., Guo, L., 2020. Class balanced loss for image classification. IEEE Access 8, 81142–81153.

Ziska, L.H., Makra, L., Harry, S.K., Bruffaerts, N., Hendrickx, M., Coates, F., Saarto, A., Thibaudon, M., Oliver, G., Damialis, A., Charalampopoulos, A., Vokou, D., Heiðmarsson, S., Guðjohnsen, E., Bonini, M., Oh, J.-W., Sullivan, K., Ford, L., Brooks, G.D., Myszkowska, D., Severova, E., Gehrig, R., Ramón, G.D., Beggs, P.J., Knowlton, K., Crimmins, A.R., 2019. Temperature-related changes in airborne allergenic pollen abundance and seasonality across the northern hemisphere: a retrospective data analysis. Lancet Planet. Health 3, e124–e131.