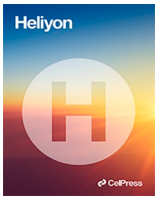
Heliyon 11 (2025) e41656

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Automating airborne pollen classification: Identifying and interpreting hard samples for classifiers

Manuel Milling [a,b,c], Simon D.N. Rampp [c], Andreas Triantafyllopoulos [a,b,c], Maria P. Plaza [d,e], Jens O. Brunner [f,g,h], Claudia Traidl-Hoffmann [d,e,i], Björn W. Schuller [a,b,c,j,k], Athanasios Damialis [d,l,*]

[a] CHI – Chair of Health Informatics, MRI, Technical University of Munich, Munich, Germany
[b] MCML–Munich Center for Machine Learning, Germany
[c] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany
[d] Institute of Environmental Medicine and Integrative Health, Faculty of Medicine, University Clinic of Augsburg & University of Augsburg, Augsburg, Germany
[e] Institute of Environmental Medicine, Helmholtz Center Munich, German Research Center for Environmental Health, Germany
[f] Faculty of Business and Economics, and Faculty of Medicine, University of Augsburg, Augsburg, Germany
[g] Department of Technology, Management, and Economics, Technical University of Denmark, Denmark
[h] Next Generation Technology, Region Zealand, Denmark
[i] Christine Kühne Center for Allergy Research and Education, Davos, Switzerland
[j] MDSI–Munich Data Science Institute, Germany
[k] GLAM–the Group on Language, Audio, & Music, Imperial College London, London, UK
[l] Terrestrial Ecology and Climate Change, Department of Ecology, School of Biology, Faculty of Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

ARTICLE INFO

Keywords:
Pollen recognition
Sample difficulty analysis
Deep learning

ABSTRACT

Deep-learning-based classification of pollen grains has been a major driver towards automatic monitoring of airborne pollen. Yet, despite an abundance of available datasets, little effort has been spent to investigate which aspects pose the biggest challenges to the (often black-box-resembling) pollen classification approaches. To shed some light on this issue, we conducted a sample-level difficulty analysis based on the likelihood for one of the largest automatically-generated datasets of pollen grains on microscopy images and investigated the reason for which certain airborne samples and specific pollen taxa pose particular problems to deep learning algorithms. It is here concluded that the main challenges lie in A) the (partly) co-occurring of multiple pollen grains in a single image, B) the occlusion of specific markers through the 2D capturing of microscopy images, and C) for some taxa, a general lack of salient, unique features. Our code is publicly available under https://github.com/millinma/SDPollen

1. Introduction

Automatic monitoring of airborne pollen has brought up a plethora of challenges at the intersection of biological, medical and

* Corresponding author. Terrestrial Ecology and Climate Change, Department of Ecology, School of Biology, Faculty of Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece.
E-mail address: dthanos@bio.auth.gr (A. Damialis).

https://doi.org/10.1016/j.heliyon.2025.e41656
Received 12 July 2024; Received in revised form 16 December 2024; Accepted 2 January 2025
Available online 3 January 2025

environmental sciences, engineering, and computer science. The idea of accurate systems detecting and classifying airborne pollen in real-time holds exciting visions for the mitigation of allergies in the short- and mid-term, and the tracking of climate change effects, in the long-term [4]. Moreover, such approaches and associated information can comprise an asset for agriculture, phytopathology, and forestry. For instance, relationships with flower, seed and fruit production may be contributing towards timely forecasts of microbial contaminations in economically valuable crops and plantations [4].

Attempts towards such systems involve devices - amongst others - that capture pollen from the atmosphere and create microscopy images from them. A prominent example for this is the commercial device BAA500 (Bio Aerosol Analyzer 500, Hund GmbH, Wetzlar) [25], on the data of which this contribution is based. The detailed processes of the creation and initial processing of the microscopy images are further described in Refs. [11,24].

In a second step, the images are processed with computer vision algorithms, these days powered by deep learning, to detect and classify individual pollen grains [32]. The development of such deep learning models relies on expert-labelled pollen datasets, which has been a major effort of the research community in the past years. Beyond that, attempts towards reliable pollen detectors and classifiers are building upon related computer vision approaches, with particular characteristics of microscopy images [19,23], and are limited by common machine learning challenges, which reach far beyond computer vision, such as class imbalance [15], fairness, e.g., with respect to the applied location [36], and the sparsity of noise-free annotations [33], amongst others.

The most widespread analysis of model performance happens on a dataset level, which, however, does not paint a full picture of how the model learns and makes decisions. In an effort to gain more fine-grained insights into the training dynamics of deep learning algorithms, recent literature has put a heavy focus on sample-level analyses. This particularly manifests in the question of how difficult individual samples are for machine learning algorithms and to understand the underlying reasons [2,21,22]. The original motivation behind sample-level difficulty estimation was curriculum-based learning, i.e., the assumption that performance can be increased, if the model first learns to recognise easy samples before steadily including more and more difficult samples into the training [38]. Nevertheless, this knowledge –even though fulfilling its initial intention by identifying corner cases or even mislabelled data points– still shows room for improvement. In return, it helps in the explainability and confidence of decision processes and might empower architecture design, e.g., by the use of priors. From the biological and environmental perspective, such information can enhance our understanding of technology and particular limitations to the specific field of research and data. Beyond, a better understanding of how different factors contribute to the perceived difficulty from a technological point of view can inspire us to consider different forms of data representations, thus allowing a more straightforward and more robust application of technology.

Relevant approaches generally act on the training data and include statistical evaluations on held out subsets of the training partition (*c-score*) [13] or the tracking of sample-based performance during training [35]. In this study, we focus our sample difficulty estimation on the *likelihood* of the correct pollen class as described in Ref. [8], offering a low-complexity solution for difficulty estimation based on the loss of the trained model. While most studies are conducted for benchmark datasets, such as ImageNet, ObjectNet, or CIFAR10, we want to bring insights about a task with more real-world application: The *Augsburg15* dataset [31,32], one of the largest labelled pollen datasets with a particular challenge posed by class imbalance. In this work, we thus aim to answer the following questions.

- Does the application of likelihood of training data points offer a robust and insightful *sample difficulty measure* for the investigated pollen dataset?
- What role does class imbalance play in the estimation of sample difficulty?
- Are there particular challenges posed by distinct pollen classes?
- What other effects are drivers of high difficulty for individual samples?

2. Data

Our analysis is based on the pollen dataset, which we first introduced in Ref. [32] and further analysed in Refs. [14,31]. It consists of automatically captured and pre-processed grayscale pollen images by a BAA500 device in Augsburg, Bavaria, Germany and manual corrections to the automatically generated pollen taxa labels [26]. With more than 50 000 samples, it is still one of the largest pollen datasets of its kind, recently being surpassed by Ref. [20]. We analysed the 15-class version of the dataset, which contains the 15 most common pollen taxa in Bavaria, south Germany. One of the major challenges of the data is the high class imbalance, which corresponds to the abundance of the pollen types in said region and is limited to their occurrence in the gathered data during one pollen season. The number of samples per pollen type ranges from 181 (*Tilia*) to 11 667 (*Corylus*).

3. Methodology

3.1. Neural network architectures

We train the popular architecture *ResNet50* [10] (≈24 Mio. parameters), which was also employed on the *Augsburg15* dataset in Ref. [31], as well as the more recently established EfficientNets [34] in its base version *EfficientNet-B0* (≈4 Mio. parameters) and the scaled version *EfficientNet-B4* (≈18 Mio. parameters); all models have an adjusted classification layer with 15 neurons for the 15-class pollen classification problem. We considere two different types of initialisation for all models: random and pre-trained on the ImageNet corpus [5]. The latter case is generally considered to improve performance for computer vision tasks and has previously shown so for the considered pollen dataset [31].

3.2. Difficulty score

We conduct our analysis of sample difficulty based on the likelihood of the correct class on the training data, which is described in Ref. [8] in the context of curriculum learning, i. e., in a performance investigation of a training paradigm, which iteratively adds more and more difficult data points to the training sets. The likelihood-based difficulty estimation provides us with a difficulty ranking of the data based on the sample-wise values of the cross-entropy loss function.

For this approach, we utilise the model states of the previously described *convolutional neural network*s (CNN) after training them on the *Augsburg15* training dataset. A score for the difficulty of each training data point is obtained by its loss value, i.e., a sample providing a large loss value obtained from the well-trained model state represents a more challenging sample than one with a small loss value. The samples are then ranked based on their difficulty score from the easiest to the most difficult sample. Note that we apply this analysis only to the training data (on which our model states were trained), as was intended by the authors of the approach.

4. Experiments

4.1. Initial model training

The first step of our experiments is the training of model states for the *Augsburg15* dataset with each image being rescaled to 64×64 pixels and normalised. We train the models *ResNet50*, *EfficientNet-B0*, and *EfficientNet-B4* with and without pre-training with the Adam optimiser and a balanced cross-entropy loss (to mitigate the effects of class imbalance) for 100 epochs. We run the same experiments for the learning rates 10^{-3} and 10^{-4} and three random seeds, resulting in overall 36 experiments, and average the performance for each combination of model architecture and initialisation type. We summarise results for *unweighted average recall* (UAR) and unweighted average F1 score in Table 1. The best model performance is slightly lower, but comparable to that reported in Ref. [31]. Differences most likely come down to our simpler training setup omitting techniques such as data augmentation or weight vector normalisation, which would add additional complexity to our discussion. As expected, a performance gap can be reported between pre-trained models and those with random initialisation, while the performance difference amongst pre-trained models is particularly small with the pre-trained *EfficientNet-B4* reaching the highest UAR and F1 score.

4.2. Difficulty score estimation

In the next step, the best model state of each training (wrt. validation UAR) is used to generate difficulty scores according to Section 3.2. However, we average the raw scores for each combination of model architecture and initialisation over learning rates and random seeds before calculating the ranking. Note that at this point, we consider the cross-entropy without balancing weights, such that each sample's difficulty (cfg. likelihood) could be estimated without explicit consideration of its class.

4.3. Inter-model comparison

Fig. 1 shows the pairwise Spearman correlation (i.e., rank correlation) coefficient of sample difficulty for the different model-initialisation combinations. It is apparent that there is a reasonably high correlation of more than 50 % across most pairs, indicating that all models capture a common underlying interpretation of difficulty to a certain degree. Interestingly, correlation between pairs with the same type of initialisation is particularly high, exceeding 55 % in all cases and 63 % for all pairs of models without pre-training. This indicates that a similar knowledge basis leads to a similar sense of difficulty. It seems likely that through the pre-training, models are able to overcome certain challenges – presumably a certain consistent type of features – that poses problems to the randomly initialised models, thus being a reason for lower performance of the latter. It is important to note that a model pre-trained on ImageNet, mostly including coloured, natural images, may not offer a particular sensitivity to the morphological features displayed in the microscopy-images of pollen grains. Pre-training on a more curated dataset including similar microscopy images may result in once again a different sense of difficulty and might be beneficial for pollen recognition performance. Nevertheless, the utility of ImageNet pre-training has been proven across different computer vision tasks [17] including pollen grain recognition [7] and even in other modalities such as audioscalograms [28] or audio-spectrograms [1], the reasons for which however are still not fully understood [9, 12].

Table 1
Performance of different combinations of architecture and initialisation: no suffix corresponds to random initialisation, *-pret* corresponds to pre-training on ImageNet. Performance is averaged across three random seeds and two learning rates. Bold font indicates the best performance.

Model	UAR ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
ResNet50	0.864 ± 0.004	0.865 ± 0.007
EfficientNet-B0	0.824 ± 0.043	0.826 ± 0.038
EfficientNet-B4	0.841 ± 0.014	0.835 ± 0.038
ResNet50-pret	0.895 ± 0.011	0.900 ± 0.008
EfficientNet-B0-pret	0.910 ± 0.004	0.905 ± 0.004
EfficientNet-B4-pret	**0.913** ± 0.004	**0.908** ± 0.006

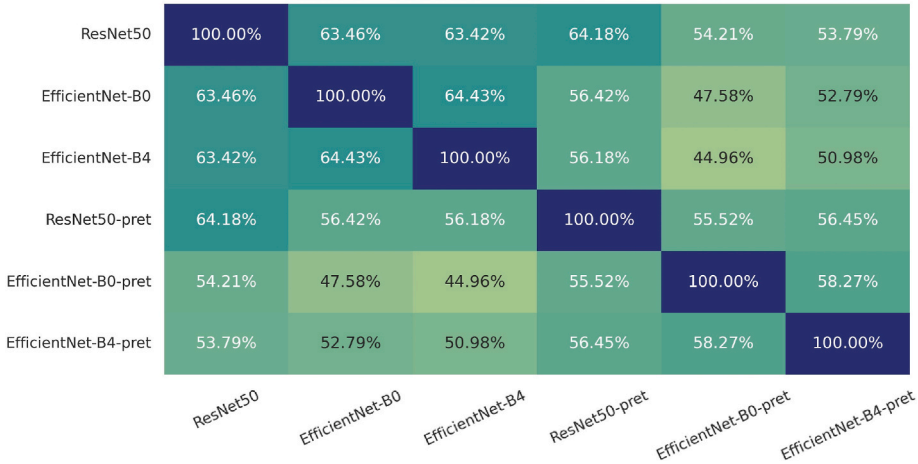


Fig. 1. Pair-wise correlation of difficulty ranks produced with different model architecture. Pre-trained (*pret*) models are considered separately from models with random initialisation.

With the version of *ResNet50* without pre-training showing the highest average pair-wise correlation overall, we further analyse its difficulty ranking in the following.

4.4. Inter-class comparison

In order to analyse the class effects on the difficulty score, we create 20 difficulty score bins and assigned all samples (class-independent) to them with an equal amount of samples per bin. We then look at the difficulty bin distribution per class in a histogram

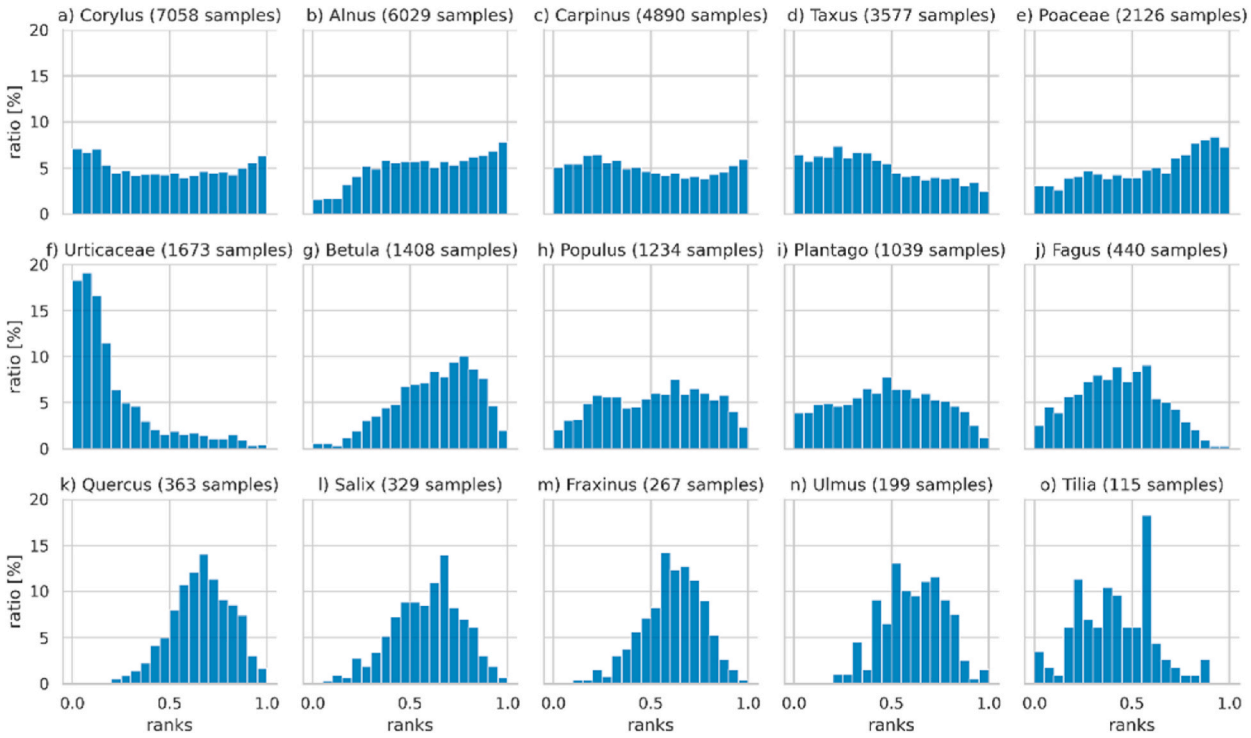


Fig. 2. Relative, normalised distribution of sample difficulty scores for each individual class. Diagrams (a–o) are sorted from top-left to right-bottom by the number of samples per class, from the most abundant (a) to the rarest class (o). Each bar diagram represents the difficulty distribution within each class with respect to the difficulty ordering of the whole dataset. The values of each bar are calculated as the total number of samples for the respective class being found in 5 % bins obtained from the difficulty ordering on the whole dataset. Finally, the bars are normalised with the total number of samples in the respective class to obtain a relative difficulty distribution.

representation and for better visibility in the imbalanced dataset, we normalise the absolute count of samples per class. Fig. 2 (a - o) thus shows the relative sample difficulty distribution per class, sorted by the number of samples per class. Interestingly, the difficulty scores seem to be more uniformly distributed amongst classes with many samples and closer to a Gaussian distribution for classes with fewer samples.

One possible explanation for this effect lies in the way the optimisation is performed through the balanced cross-entropy-loss, which assigns stronger weights and, thus, higher importance to samples from under-represented classes. A balanced loss function is a necessary choice for the training to prevent the network from ignoring an otherwise small contribution of the minority classes to the overall training loss. It allows the training to leverage patterns learned in the overall large dataset with more than 50 000 samples to a reasonable recognition rate for classes with less than 200 samples (to some extent even for classes with less than 30 samples), as reported for the dataset at hand in Ref. [31]. As a consequence, though, the model parameter updates are heavily impacted when confronted with a sample of the minority class. This makes it unlikely that any training sample from a minority class will remain with a high loss after training. Consequentially, most samples from minority classes will not be found in the hardest samples of the full dataset. On the other hand, the underrepresented classes might not benefit from the smooth learning process that might be obtained through many parameter updates for the abundant classes. Therefore, the losses of easy samples of the minority classes might not decrease as much as those of the majority classes. This results in fewer samples of the minority class being interpreted as very easy. The observed differences in distribution are, thus, likely to be an artefact of the training paradigm and the difficulty measure rather than being rooted in the pollen types themselves.

A more quantitative perspective on average class difficulty can be taken via Fig. 3, which shows the average difficulty ranks, i.e., the average bar height in Fig. 2, over the number of samples per class on a logarithmic scale. The regression line shows only a minimal trend with negative slope, indicating that effects of class imbalance are largely nullified, presumably through the balanced cross-entropy training. Therefore, we also get a clear picture that the taxa *Quercus*, *Betula*, Poaceae and *Alnus* have particularly difficult samples, while the taxa *Fagus*, *Tilia*, and, above all, Urticaceae have particularly easy samples. This indicates the existence of specific characteristics of the respective pollen taxa impacting their recognition difficulty in the context of the considered setup of 2D microscopy imaging and convolutional neural network classifiers. These results are also backed up by the confusion matrices for *Augsburg15* in Refs. [31,32], in which the "hard" classes show low and the "easy" samples show high recognition rates, even though a trend of better performance for classes with a large number of samples is evident in these studies. It is important to note that the study of Schäfer et al. [31] is not able to nullify the effects of class imbalance on recognition performance despite more sophisticated mitigation strategies, such as focal loss [29] and weight vector normalisation [16]. Nevertheless, a closer investigation of how different techniques for class-imbalanced datasets, such as the ones mentioned above, or other examples like synthetic minority over-sampling technique (SMOTE) [3] or under-sampling, impact the models' interpretations of difficulty.

4.5. Exemplary sample analysis

A closer look at the supposedly easiest and hardest samples from each class in Fig. 4 reveals several interesting insights: *Tilia* pollen shows a very low identification difficulty, despite the lowest occurrence in the dataset, as the morphology of the pollen grain is unique, being larger in size than the commonest ones, colporate and isopolar, cup-shaped and with very characteristic and deep colpi. Likewise, *Fagus* pollen is even larger, with long, narrow colpi and large, round protruding pori. Both *Tilia* and *Fagus* are difficult to misclassify by a well-trained aerobiologist expert. On the other hand, pollen from *Alnus*, *Betula*, Poaceae and *Quercus* are more difficult to classify manually. The first two taxa look alike, with similar size and type and shape of pori, with the main differentiation being that *Alnus* pollen has 5–7 pori in contrast to *Betula* that has only 3. In real-life, not all pori (or any morphological characteristics whatsoever) are visible, which makes the confusion of these two taxa quite high. Interestingly, they are not easily confused with the other

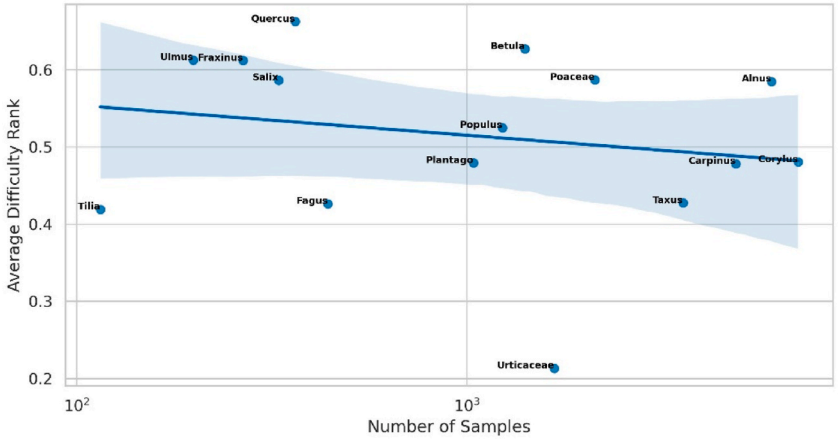


Fig. 3. Average difficulty rank for each class with respect to the number of total samples. The average difficulty rank for each class is computed as the average sum over the position of all corresponding samples in the overall difficulty ranking of all samples.

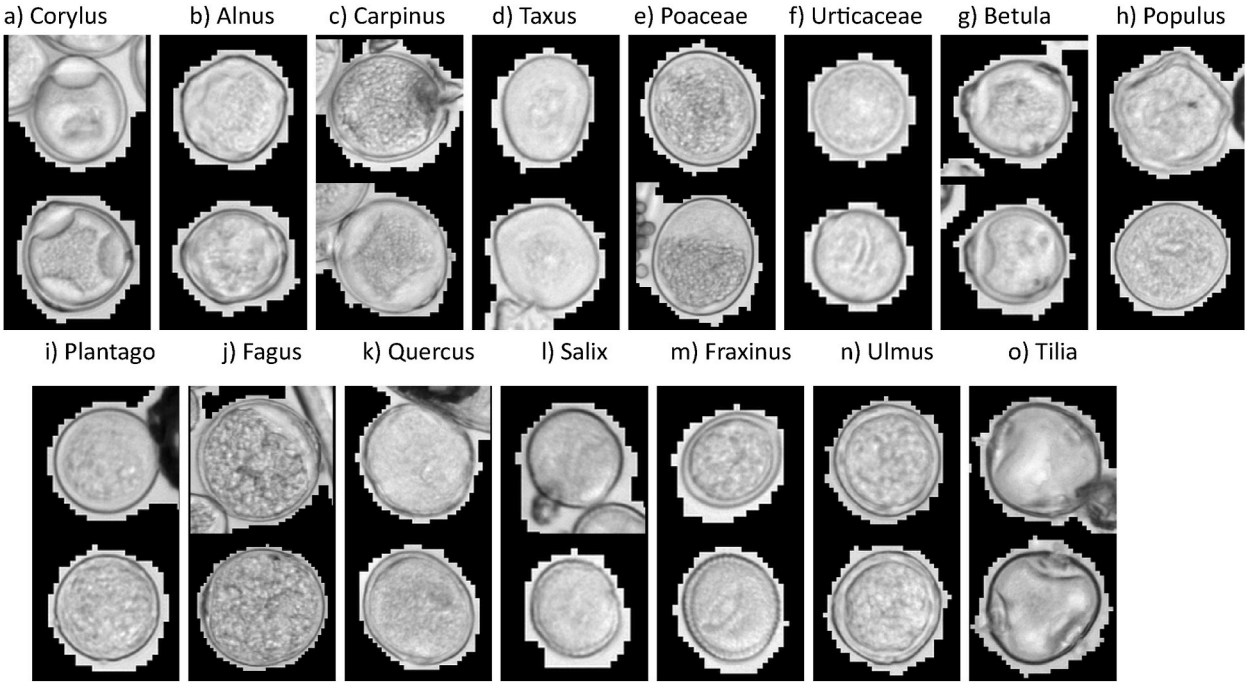


Fig. 4. Overview of each pollen type according to their respective hardest and easiest image. Pollen taxa (a–o) are sorted top-left to bottom-right based on the number of samples per pollen taxa. For each pollen taxa, we display two vertically arranged pollen samples with the top one being the hardest and the bottom one the easiest sample from the dataset as determined with the ResNet50 architecture without pre-training.

representatives of the family, *Carpinus* and *Corylus*, because of different shape (aspidate for the former) and pori morphology (deeper in both taxa). *Fraxinus* and *Quercus* are also harder to classify, with the former literally having no unique features, but 3 colpi, average size, no exine special features, which ranks it as maybe the hardest to recognise pollen type, even for expert aerobiologists.

Fraxinus, is to be confused with other representatives of the family, like *Olea*, *Phillyrea*, *Syringa* and *Ligustrum*, with some distinct features, but those being the most subtle in the case of *Fraxinus*. Poaceae, finally, is the pollen type with the most plant representatives and even though it has one distinct round porus, there a high variability in all other features because of hundreds of different species existing per location.

In conclusion, from a taxon-independent perspective, the largest confounding factor seems to be the coexistence of additional pollen parts, which mostly appear on the edges. An encouraging mitigation strategy for this might be data augmentation through cutmix [39]. However, in contrast to common implementations, application should be limited to the edges while preserving the grain boundaries and without label adjustments. Beyond, it is a crucial problem that some visual features of the 3D pollen grains can be occluded in the 2D images. It seems thus desirable to force the model to be less dependent on these particular features, for instance through the means of priors or other augmentation strategies. For instance, one promising strategy to mitigate these effects is data augmentation that particularly modifies the visibility and occlusion of relevant morphological features by deriving 2D projections from 3D models of pollen grains and their corresponding morphological features. Further, models could be designed to explicitly recognise and model morphological features from the 2D images. The decision process can then be guided by considering the occlusion of some features, further benefiting model interpretability.

4.6. Discussion and limitations

Even though the trends appearing in Fig. 1 are similar across the different models, details are dependent on the chosen model states and further carry some bias through the choice of likelihood-based difficulty measure. Approach-dependent notions of difficulty will be hard to overcome as a certain level of subjectiveness of difficulty has to be expected due to the different ways in which different machine learning algorithms, and for reference, humans, approach a given problem. This aspect is further emphasised through the abundance of difficulty estimations available in literature [18,37], including methods based on ensembles or on the learning history of a model. Nevetherless, across different automated approaches for difficulty estimation, a reasonable overlap, at least across deep

learning architectures of a similar type, can be observed, for instance, in the case of fully connected versus convolutional neural networks in Ref. [38]. Similarly, some forms of human annotation-based difficulty estimates have shown a reasonable level of agreement with some of the automated difficulty estimates [21]. Still, an in-depth investigation into the similarity across human-based and machine-based difficulty estimation is still missing in literature. Beyond, we analyse only a specific type of data from one device being collected essentially as 2D microscopy images in context of a specific capturing mechanism. The ability to generalise our discovered insights to other devices thus needs to be further analysed. Nevertheless, we are optimistic that some findings are transferable to other problems as in particular the usage of optical capturing devices in combination with deep learning models and more specifically convolutional neural networks have been established as a common practice in recent years [27]. In this setting, some encountered challenges should be common, and they intuitively pose challenges to the deep learning algorithms, like, for instance, the partial co-occurrence of other pollen grains. Beyond, other studies show similarity in certain pollen taxa being more difficult to recognise for machine learning systems under different data collection methods. For instance, in Ref. [30] data is captured based on multi-angle scattering images and fluorescence spectra of pollen grains. Despite the different recording approach, *Quercus* and *Betula* show particularly low recognition rates, while *Fagus* and *Taxus* show high recognition rates, all of which is in-line with our reported difficulty analysis. Furthermore, in Ref. [6] pollen grain information is captured via scattered light from a laser source, which is in the following used for pollen classification with the more traditional machine learning methods support vector machine (SVM), k-nearest-neighbour (KNN) and multi-layer-perceptron (MLP). Within the quite different choice of pollen taxa, *Quercus*, *Fraxinus*, *Betula*, and *Alnus* show sub-par recognition rates, all of which we also conclude to have an above-average difficulty. Interestingly however, *Corylus* appears with the lowest recognition rate in the study, which provides a rather average difficulty according to our results. Despite the differences in technologies, there seem to be some trends in the difficulty of pollen taxa for automated machine learning-based pollen recognition that generalise beyond our study. Finally, we have to mention limitations in the representativeness of the investigated samples in Fig. 4(a–o). Even though they were selected based on the highest and lowest difficulty score per class, there is only some insight to be obtained from the analysis of a single sample per class.

5. Conclusion

This work explored sample-level difficulty for an imbalanced microscopy pollen classification dataset. Our results indicate that the difficulty score is consistent across different architectures with higher discrepancies between different initialisations and that effects of class imbalance on sample difficulty are not apparent with a balanced cross-entropy training. However, consistent with previous work, we find that some classes are particularly easy or hard to recognise for deep learning model; taxa from the Betulaceae and Poaceae families, along with *Quercus* species seem to be the hardest to identify overall. In future work, we hope to leverage insights from the sample difficulty into practice for more robust and justified training strategies and, in general, to make a broader case for dataset analyses via sample difficulty in machine learning for bio-imaging. Beyond, an investigation into biology-inspired difficulty estimators, particularly suited to the recognition of pollen grains, offers a promising venue for future research.

CRediT authorship contribution statement

Manuel Milling: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Simon D.N. Rampp:** Writing – review & editing, Methodology, Investigation. **Andreas Triantafyllopoulos:** Writing – review & editing, Methodology, Investigation. **Maria P. Plaza:** Writing – review & editing, Investigation, Data curation. **Jens O. Brunner:** Writing – review & editing, Investigation. **Claudia Traidl-Hoffmann:** Writing – review & editing, Project administration, Investigation. **Björn W. Schuller:** Writing – review & editing, Investigation. **Athanasios Damialis:** Writing – review & editing, Project administration, Investigation, Data curation.

Data and code availability

Data will be made available upon reasonable request. Our code is publicly available under https://github.com/millinma/SDPollen.

Ethics declaration

This study did not involve any human participants, animals, or sensitive data, and it does not raise any ethical concerns.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Corresponding author, Athanasios Damialis, serves as Section Editor in Heliyon Agriculture. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, B. Schuller, Snore Sound classification using image-based deep spectrum features, Proc. Interspeech 2017 (2017) 3512–3516, https://doi.org/10.21437/Interspeech.2017-434.

[2] R. Baldock, H. Maennel, B. Neyshabur, Deep learning through the lens of example difficulty, Adv. Neural Inf. Process. Syst. 34 (2021) 10876–10889.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[4] A. Damialis, C. Traidl-Hoffmann, R. Treudler, Climate change and pollen allergies, Biodiver. Health Face Climate Change (2019) 47–66.

[5] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet:Alarge-scalehierarchicalimagedatabase, in: 2009IEEEconference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. Ieee.

[6] H. El Azari, J.B. Renard, J. Lauthier, T. Dudok de Wit, A laboratory evaluation of the new automated pollen sensor beenose: pollen discrimination using machine learning techniques, Sensors 23 (2023) 2964.

[7] A.R.d. Geus, C.A. Barcelos, M.A. Batista, S.F.d. Silva, Large-scale pollen recognition with deep learning, in: 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5, https://doi.org/10.23919/EUSIPCO.2019.8902735.

[8] G. Hacohen, D. Weinshall, On the power of curriculum learning in training deep networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 2535–2544.

[9] K. He, R. Girshick, P. Dollár, Rethinkingimagenetpre-training, in: ProceedingsoftheIEEE/CVFinternationalconferenceoncomputer Vision, 2019, pp. 4918–4927.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[11] U. Heimann, J. Haus, D. Zuehlke, Op3-fully automated pollen analysis and counting: the pollen monitor baa500, in: Proceedings OPTO, 2009, pp. 125–128, 2009 & IRS2 2009.

[12] M. Huh, P. Agrawal, A.A. Efros, What Makes Imagenet Good for Transfer Learning? arXiv Preprint arXiv:1608.08614, 2016.

[13] Z. Jiang, C. Zhang, K. Talwar, M.C. Mozer, Characterizing structural regularities of labeled data in overparameterized models, in: International Conference on Machine Learning, PMLR, 2021, pp. 5034–5044.

[14] B. Jin, M. Milling, M.P. Plaza, J.O. Brunner, C. Traidl-Hoffmann, B.W. Schuller, A. Damialis, Airborne pollen grain detection from partially labelled data utilising semi-supervised learning, Sci. Total Environ. 891 (2023) 164295, https://doi.org/10.1016/j.scitotenv.2023.164295. URL: www.sciencedirect.com/science/article/pii/S0048969723029169.

[15] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (2019) 1–54.

[16] B. Kim, J. Kim, Adjusting decision boundary for class imbalanced learning, IEEE Access 8 (2020) 81674–81685.

[17] H.E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M.E. Maros, T. Ganslandt, Transfer learning for medical image classification: a literature review, BMC Med. Imag. 22 (2022) 69.

[18] F. Liu, T. Zhang, C. Zhang, L. Liu, L. Wang, B. Liu, A review of the evaluation system for curriculum learning, Electronics 12 (2023) 1676, https://doi.org/10.3390/electronics12071676, 10.3390/electronics12071676.

[19] Z. Liu, L. Jin, J. Chen, Q. Fang, S. Ablameyko, Z. Yin, Y. Xu, A survey on applications of deep learning in microscopy image analysis, Comput. Biol. Med. 134 (2021) 104523.

[20] P. Matavulj, M. Panić, B. Šikoparija, D. Tešendić, M. Radovanović, S. Brdar, Advanced cnn architectures for pollen classification: design and comprehensive evaluation, Appl. Artif. Intell. 37 (2023) 2157593, https://doi.org/10.1080/08839514.2022.2157593, 10.1080/08839514.2022.2157593.

[21] D. Mayo, J. Cummings, X. Lin, D. Gutfreund, B. Katz, A. Barbu, How hard are computer vision datasets? calibrating dataset difficulty to viewing time. https://neurips.cc/virtual/2023/poster/73596, 2022.

[22] K. Meding, L.M.S. Buschoff, R. Geirhos, F.A. Wichmann, Trivial or impossible–dichotomous data difficulty masks model differences (on imagenet and beyond), arXiv preprint arXiv:2110.05922 (2021).

[23] M. Milling, M. Lienhart, Y. Oksymets, A. Gebhard, M. Brugger, C. Westerhausen, B.W. Schuller, Neurocellcentredb: exploring a novel dataset for neuron-like cell centre detection with deep neural networks, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2023, pp. 1–4. IEEE.

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

[25] J. Oteros, M. Sofiev, M. Smith, B. Clot, A. Damialis, M. Prank, M. Werchan, R. Wachter, A. Weber, S. Kutzora, S. Heinze, C.E. Herr, A. Menzel, K.C. Bergmann, C. Traidl-Hoffmann, C.B. Schmidt-Weber, J.T. Buters, Building an automatic pollen monitoring network (epin): selection of optimal sites by clustering pollen stations, Sci. Total Environ. 688 (2019) 1263–1274, https://doi.org/10.1016/j.scitotenv.2019.06.131, sciencedirect.com/science/article/pii/S0048969719327020.

[26] M.P. Plaza, F. Kolek, V. Leier-Wirtz, J.O. Brunner, C. Traidl-Hoffmann, A. Damialis, Detecting airborne pollen using an automatic, real-time monitoring system: evidence from two sites, Int. J. Environ. Res. Publ. Health 19 (2022) 2471.

[27] M. Qi, F.K. Du, F. Guo, K. Yin, J. Tang, Species identification through deep learning and geometrical morphology in oaks (quercus spp.): pros and cons, Ecol. Evol. 14 (2024) e11032.

[28] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, B. Schuller, Deep scalogram representations for acoustic scene classification, IEEE/CAA J. Automat. Sinica 5 (2018) 662–669, https://doi.org/10.1109/JAS.2018.7511066.

[29] T.Y. Ross, G. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2980–2988.

[30] I. Šauliene, L. Šukien˙e, G. Daunys, G. Valiulis, L. Vaitkevičius, P. Matavulj, S. Brdar, M. Panic, B. Sikoparija, B. Clot, et al., ˙ Automatic pollen recognition with the rapid-e particle counter: the first-level procedure, experience and next steps, Atmos. Meas. Tech. 12 (2019) 3435–3452.

[31] J. Schaefer, M. Milling, B.W. Schuller, B. Bauer, J.O. Brunner, C. Traidl-Hoffmann, A. Damialis, Towards automatic airborne pollen monitoring: from commercial devices to operational by mitigating class-imbalance in a deep learning approach, Sci. Total Environ. 796 (2021) 148932, https://doi.org/10.1016/j.scitotenv.2021.148932 sciencedirect.com/science/article/pii/S0048969721040043.

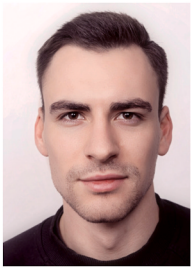
[32] J. Schiele, F. Rabe, M. Schmitt, M. Glaser, F. Häring, J.O. Brunner, B. Bauer, B. Schuller, C. Traidl-Hoffmann, A. Damialis, Automated classification of airborne pollen using neural networks, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC), 2019, pp. 4474–4478, https://doi.org/10.1109/EMBC.2019.8856910.

[33] M.P. Schilling, L. Klinger, U. Schumacher, S. Schmelzer, M.B. López, B. Nestler, M. Reischl, Ai2seg: a method and tool for ai-based annotation inspection of biomedical instance segmentation datasets, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC), 2023, pp. 1–6, https://doi.org/10.1109/EMBC40787.2023.10341074.

[34] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[35] M. Toneva, A. Sordoni, R.T.d. Combes, A. Trischler, Y. Bengio, G.J. Gordon, An empirical study of example forgetting during deep neural network learning, arXiv preprint arXiv:1812.05159 (2018).
[36] A. Triantafyllopoulos, M. Milling, K. Drossos, B.W. Schuller, Fairness and underspecification in acoustic scene classification: the case for disaggregated evaluations, arXiv preprint arXiv:2110.01506 (2021).
[37] X. Wang, Y. Chen, W. Zhu, A survey on curriculum learning. https://arxiv.org/abs/2010.13166, 2020.
[38] X. Wu, E. Dyer, B. Neyshabur, When Do Curricula Work? arXiv Preprint arXiv:2012.03107, 2020.
[39] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.

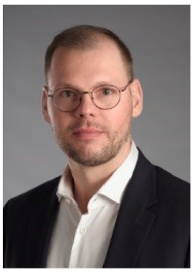
Manuel Milling received his Bachelor of Science in Physics and in Computer Science from the University of Augsburg in 2014 and 2015, respectively and his Master of Science in Physics from the same university in 2018. He is currently a PhD candidate in Computer Science at the chair of Health Informatics, Technical University of Munich. His research interests include machine learning with a particular focus on the core understanding and applications of deep learning methodologies.



Simon D. N. Rampp received his Bachelor of Science in Computer Science from the University of Augsburg in 2022, with a focus on deep learning. He is currently pursuing a Master of Science in Computer Science at the same university, working on his thesis that explores example difficulty and curriculum learning across image and audio modalities.

Andreas Triantafyllopoulos received the diploma in ECE from the University of Patras, Greece, in 2017. He is working toward the doctoral degree with the Chair of Health Informatics, Technical University of Munich. His current focus is on deep learning methods for auditory intelligence and affective computing.

Maria P. Plaza graduated in Forestry Engineering in 2007, obtained a Master's degree in Natural Environment in 2011 and a PhD in Aerobiology in 2017 at the University of Cordoba, Spain. Subsequently, she completed a Master's degree in Bioinformatics and Biostatistics at the University of Catalonia in 2020. She is the group leader of Human Exposure Science Group at the Institute of Environmental Medicine at the Augsburg University Hospital and Helmholtz Center Munich, Germany, since 2022. In particular, her expertise is in Aerobiology, detection and quantification of aeroallergens, influence of different environmental factors, impacts on human health and forecasting of spatio-temporal interactions at all levels with state-of-the-art technology and programming languages, aiming to provide accurate and timely risk alerts.



Jens O. Brunner has been permanently appointed as Professor of Decision Science in Healthcare at the Department of Technology, Management, and Economics at the Technical University of Denmark. He was appointed as Professor of Health Care Operations/Health Information Management at the Faculty of Business and Economics at the University of

Augsburg until March 2023. From 2013 to 2020 he was (co-) director of the University Center for Health Care at Klinikum Augsburg (UNIKA-T). He received a PhD from the TUM School of Management in 2009 and a diploma degree in Business Administration from the University of Mannheim in 2006.

Claudia Traild-Hoffmann is Professor of Environmental Medicine at the University of Augsburg and Director of the Institute of Environmental Medicine at Helmholtz Munich. Since 2014, she has also headed the University Outpatient

Clinic for Environmental Medicine at Augsburg University Hospital as Chief Physician and was Deputy Director of the

ZIEL - Institute of Food and Health in Weihenstephan from 2017-2022. She is a founding member and vice-director of the Center for Climate Resilience at the University of Augsburg since 2020. Since 2013, she has been a member of the board of directors of CK-CARE and spokesperson since 2020. She studied medicine at RWTH Aachen University and has board certification in dermatology, venerology and allergology. Her research focuses on human-environment interactions with special emphasis on allergies and the impacts of climate change on health. For the appointment period 2022-2026 she has been appointed to the Commission "Environmental Public Health" of the RKI and WBGU member since December 2022.

Björn W. Schuller received his diploma in 1999, his doctoral degree in 2006, and his habilitation and was entitled Adjunct Teaching Professor in 2012 all in electrical engineering and information technology from TUM in Munich/Germany. He is Full Professor of Artificial Intelligence and the Head of GLAM at Imperial College London/UK, Chair of CHI – the Chair for Health Informatics, MRI, Technical University of Munich, Munich, Germany, amongst other Professorships and Affiliations. He is a Fellow of the IEEE and Golden Core Awardee of the IEEE Computer Society, Fellow of the ACM, Fellow and President-Emeritus of the AAAC, Fellow of the BCS, Fellow of the ELLIS, Fellow of the ISCA, and Elected Full Member Sigma Xi. He (co-)authored 1,400+ publications (60,000+ citations, h-index=111).

Athanasios Damialis received his BSc and MSc in Forestry and Natural Environment in 2000, his second MSc in Environmental Biology in 2002, and his PhD in 2010. He served as an Adjunct Lecturer in the Aristotle University of Thessaloniki in 2010, postdoctoral researcher - amongst others - in Royal Holloway University of London, UK, in 2012-2013, Group Leader - amongst others - in the Technical University of Munich, Germany, in 2015-2021, and - currently - As. Professor of Terrestrial Ecology and Climate Change in the Aristotle University of Thessaloniki, Greece. He has published a total of more than 200 papers, abstracts, and book chapters. He is the President of the European Aerobiology Society, the Europe Councilor in the International Society of Biometeorology, member of the Working Group of Aerobiology and Pollution of the European Academy of Allergy and Clinical Immunology, and Working Group Leader (Impacts) in the EUMETNET-Autopollen project, while he also served as Reviewer in the latest IPCC Report (WGII AR6).