Received: 19 October 2020 | Accepted: 28 January 2021

DOI: 10.1111/2041-210X.13575

RESEARCH ARTICLE

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Efficient, automated and robust pollen analysis using deep learning

Ola Olsson[1] | Melanie Karlsson[2] | Anna S. Persson[2] | Henrik G. Smith[1,2] | Vidula Varadarajan[1] | Johanna Yourstone[1] | Martin Stjernman[1]

[1]Department of Biology, Lund University, Lund, Sweden

[2]Centre for Environment and Climate Research, Lund University, Lund, Sweden

Correspondence
Ola Olsson
Email: ola.olsson@biol.lu.se

Funding information
Svenska Forskningsrådet Formas, Grant/Award Number: 215-2014-1603; Region Skånes Miljövårdsfond; Crafoordska Stiftelsen, Grant/Award Number: 20170867

Handling Editor: Robert Freckleton

Abstract

1. Pollen analysis is an important tool in many fields, including pollination ecology, paleoclimatology, paleoecology, honey quality control, and even medicine and forensics. However, labour-intensive manual pollen analysis often constrains the number of samples processed or the number of pollen analysed per sample. Thus, there is a desire to develop reliable, high-throughput, automated systems.
2. We present an automated method for pollen analysis, based on deep learning convolutional neural networks (CNN). We scanned microscope slides with fuchsine stained, fresh pollen and automatically extracted images of all individual pollen grains. CNN models were trained on reference samples (122,000 pollen grains, from 347 flowers of 83 species of 17 families). The models were used to classify images of different pollen grains in a series of experiments. We also propose an adjustment to reduce overestimation of sample diversity in cases where samples are likely to contain few species.
3. Accuracy of a model for 83 species was 0.98 when all samples of each species were first pooled, and then split into a training and a validation set (splitting experiment). However, accuracy was much lower (0.41) when individual reference samples from different flowers were kept separate, and one such sample was used for validation of models trained on remaining samples of the species (leave-one-out experiment). We therefore combined species into 28 pollen types where a new leave-one-out experiment revealed an overall accuracy of 0.68, and recall rates >0.90 in most pollen types. When validating against 63,650 manually identified pollen grains from 370 bumblebee samples, we obtained an accuracy of 0.79, but our adjustment procedure increased this to 0.85.
4. Validation through splitting experiments may overestimate robustness of CNN pollen analysis in new contexts (samples). Nevertheless, our method has the potential to allow large quantities of real pollen data to be analysed with reasonable accuracy. Although compiling pollen reference libraries is time-consuming, this is simplified by our method, and can lead to widely accessible and shareable resources for pollen analysis.

KEYWORDS
agricultural systems, applied ecology, bumblebees, food webs, habitats, palynology, pollination

1 | INTRODUCTION

Pollen analysis is important in a wide range of fields, including medicine (Bastl et al., 2017), honey quality control (von der Ohe et al., 2004), paleoclimatology and paleoecology (Bourel et al., 2020), and pollination ecology (Bertrand et al., 2019). Although significant progress has been made towards automated pollen analysis during recent years (e.g. Bourel et al., 2020; Dunker et al., 2020; Holt & Bennett, 2014; Sevillano et al., 2020), many field studies are still constrained by laborious and costly pollen identification and quantification.

Traditionally, pollen analysis is based on manual microscopy, which requires skilled experts and is labour-intensive (e.g. Beil et al., 2008; Persson et al., 2018; Wood et al., 2017). Methods include both bright field and dark field microscopy, and can be combined with different pollen preparation methods such as staining of fresh pollen or acetolysis (Kearns & Inouye, 1993), with the latter also possible to use on fossil pollen (Bourel et al., 2020). For some applications, manual microscopy might still be preferred, but technological developments are starting to provide alternatives, including molecular methods such as meta-barcoding (e.g. Keller et al., 2015) or genome skimming (Lang et al., 2019), chemotaxonomy (Jardine et al., 2019), and image analysis techniques based on deep neural networks (e.g. Sevillano et al., 2020).

To be useful, pollen analysis methods need to be accurate, quantitative, efficient (Holt & Bennett, 2014; Stillman & Flenley, 1996), and ideally accessible to a wide range of users. Accuracy, in terms of correct identification, is usually assumed high for manual analysis (Holt & Bennett, 2014; although low accuracy of humans is reported by e.g. Gonçalves et al., 2016), and such methods are also quantitative (e.g. Bertrand et al., 2019). However, they are relatively inefficient and there is often a trade-off between number of samples and number of analysed pollen grains per sample, which increases the uncertainty of quantitative estimates (e.g. Bertrand et al., 2019; Persson et al., 2018). It has been claimed that meta-barcoding can be used for quantitative assessments (Keller et al., 2015), but that view has been contested (Bell et al., 2019).

Methods based on image analysis using deep neural networks, or related machine classification, can potentially be accurate, quantitative, efficient and accessible. Recently, Holt and Bennett (2014) reviewed available methods and listed the capacities and requirements automated methods need to fulfil. Since then, several new studies with promising results have been published (e.g. Bourel et al., 2020; Daood et al., 2016; Dunker et al., 2020; Sevillano & Aznarte, 2018; Sevillano et al., 2020). This illustrates the general progress made in the field of machine learning and artificial intelligence, especially using deep neural networks, which are well suited for classifying two-dimensional images (Sevillano & Aznarte, 2018). Convolutional neural networks (CNN) are a class of deep neural networks, which have proven to be very efficient for classifying images without the need for manual feature extraction (e.g. Albawi et al., 2017; Sevillano & Aznarte, 2018).

Recent studies have shown accuracies close to 100% (Bourel et al., 2020; Daood et al., 2016; Dunker et al., 2020; Sevillano & Aznarte, 2018), and even for a severe problem with 46 pollen types, Sevillano et al. (2020) arrived at a correct classification rate of nearly 98%. This is especially impressive as some of the included pollen types are known to be hard to separate even for experienced palynologists. The output from a CNN classification is quantitative, and with modern computers classification is also efficient, with hundred or more objects classified per second (Dunker et al., 2020; Sevillano & Aznarte, 2018). Most of the software are based on open-source code, and therefore open systems can potentially be built.

Some of the most successful previous studies have based their analyses on hybrid methods between CNN and feature extraction (Sevillano & Aznarte, 2018; Sevillano et al., 2020). In this study, we rely on CNN alone, and apply transfer learning. In transfer learning, an existing model, previously developed and trained to separate other types of images, is used to learn a new classification task (reviewed by Lumini & Nanni, 2019; Sevillano & Aznarte, 2018). Currently, over a dozen different CNN models can be used for transfer learning using programming environments such as PyTorch or MATLAB. These models are typically trained to separate 1,000 classes of images from the ImageNet database (cats, dogs, cars, etc. http://www.image-net.org). The models vary in complexity from, for example, AlexNet (Krizhevsky et al., 2012) with eight layers, to DenseNet-201 (Huang et al., 2017) with 201 layers. During transfer learning, the model is changed by replacing the final classification layer, and then training it on a set of new images with the classes to be learnt, in our case the pollen types.

In pollination ecology, it is common to investigate the foraging and choices made by honeybees and wild bees by studying the pollen they collect (Bertrand et al., 2019; Leonhardt & Blüthgen, 2012; Persson et al., 2018). Many species of bumblebees are known to be generalists in their use of flower species, but still focus on just one or a few species during each foraging trip (e.g. Heinrich, 1979). This provides an opportunity for manual pollen analysis, as the person can focus attention to identification of a few species per sample. However, it provides a challenge for an automated pollen analysis system, as misclassification will tend to increase pollen type representation above the actual.

In this paper, we describe a fully automated CNN-based method for the analysis of fresh, stained pollen samples containing several hundred to thousands pollen grains each, scanned using bright field microscopy. We base the analyses on two

datasets. The first is our 347 pollen reference samples collected from 83 plant species of 17 families, consisting of a total of *c.* 122,000 tagged pollen grains. The second is a dataset with 370 pollen samples, collected from bumblebees foraging in the wild, consisting of 63,500 pollen grains. The latter has been manually identified by us into 29 pollen types, which are either comprised of single species or groups of similar species. We run a series of experiments where we vary the kind of validation data on which the CNN's ability to classify pollen is tested. Thus, our study differs from those previously published by the volume of the reference samples, and the level of challenge posed by the validation, explicitly assessing the robustness of the model on samples other than the ones it was trained on.

2 | MATERIALS AND METHODS

2.1 | Pollen sample preparation

2.1.1 | Reference pollen samples

We took reference pollen samples (Figure 1) in the field from flowers, identified to species, in the provinces of Scania and Småland, southern Sweden, between 2012 and 2020. We focussed on species occurring in bumblebee pollen samples (see below), which included wild species (native and non-native), crops and garden plants.

We collected at least two samples per species (Supporting Information), each from a different flower and preferably from

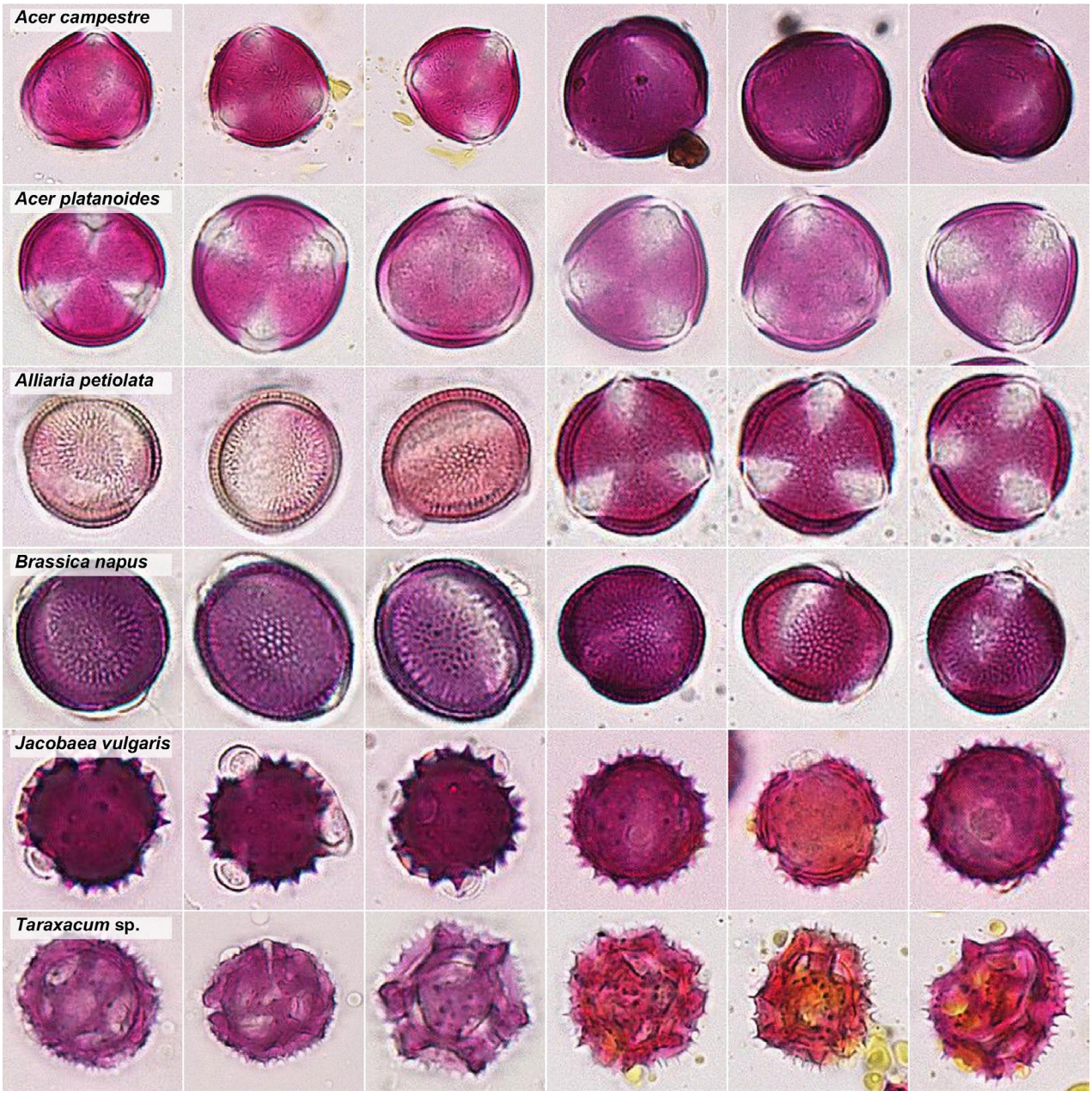

FIGURE 1 Pollen grains of six species, to illustrate within and between species variation in pollen images. For each species, the three grains on the left come from one sample, and the three on the right come from another

different days and localities, to cover different sources of variation. Out of our 83 species, 17 had only two samples.

A small cube (*c*. 1.5 × 1.5 × 1.5 mm) of basic fuchsine gel (Kearns & Inouye, 1993), held by forceps, was rubbed against the anthers of the flowers. We put the fuchsine cubes individually in Eppendorf tubes, and stored them in room temperature for a few days or, if the time before preparation in the laboratory was longer, in a freezer (−18°C). The concentration of fuchsine in the gel affects the intensity of the staining, but such variation in our data appears to have little effect on the accuracy of CNN identification (Supporting Information).

We placed each fuchsine cube on a microscope slide and heated until melting (~50°C). We then covered it with a cover glass, pressing gently to make the sample thin without crushing the pollen grains, and sealed the edges with transparent nail polish to prevent drying. Samples were stored in room temperature over-night before scanning (see below).

2.1.2 | Bumblebee pollen samples

We took 370 pollen samples from bumblebees during May and June 2017 in 36 areas in southern Scania. We caught bumblebees returning to their nest, temporarily placed them in a queen-marking cage and removed the pollen loads from corbiculae on both legs using forceps. Pollen clumps were stored in individual Eppendorf tubes in a freezer (−18°C).

One clump from each bumblebee was dissolved in 70% ethanol, aiming for a 1:7 ratio of pollen to ethanol by weight. We homogenized and separated pollen aggregations in each sample with a vortex mixer, and transferred 2 µl of the suspension with a pipette onto a microscope slide with a fuchsine gel cube. The remaining procedure was the same as for the reference pollen samples.

2.2 | Scanning and image preparation

We scanned pollen samples using a Leica Aperio CS2 slide scanner at 0.25-µm resolution (40× magnification). On each sample, we scanned approximately 6 × 6 mm (*c*. 24,000 × 24,000 pixels), at five different focus (*z*) layers, 3 µm apart. The scanner can load and scan five slides at a time, and saves images as svs-files, that is, 8-bit RGB, multi-layer JPEG files with metadata.

To obtain a single-layered well-focussed image, we stacked the focus layers of the images, using the function fstack (Pertuz et al., 2013) in MATLAB. The resulting images were saved as JPEG at 75% quality, maintaining practically all visible details of lossless TIFF-images, but using only *c*. 10% of disk space (*c*. 100 MB instead of 1 GB per image).

2.3 | Image extraction

The CNN algorithms need small images of fixed size, with a single object (pollen grain) in each. To extract images, we developed an algorithm that performs segmentation and extraction based on a sequence of operations using functions from MATLAB's Image Processing Toolbox (Figure 2; see Supporting Information for details). Resulting bounding boxes, based on centroids and major axis lengths (shown in Figure 2f) , are used to extract and save individual images of each object after rescaling to the appropriate size (in our case 224 × 224 or 299 × 299 pixels depending on CNN model; Figure 1). As rescaling equalizes size among pollen, size information is not part of our analyses. Thus, the whole process rapidly locates, separates, measures and extracts all objects in large images on current hardware (Intel i9 with 64 GB RAM; NVIDIA GeForce RTX 2060 Super with 8 GB GRAM). The image extraction allows for complete quantification of all pollen on a

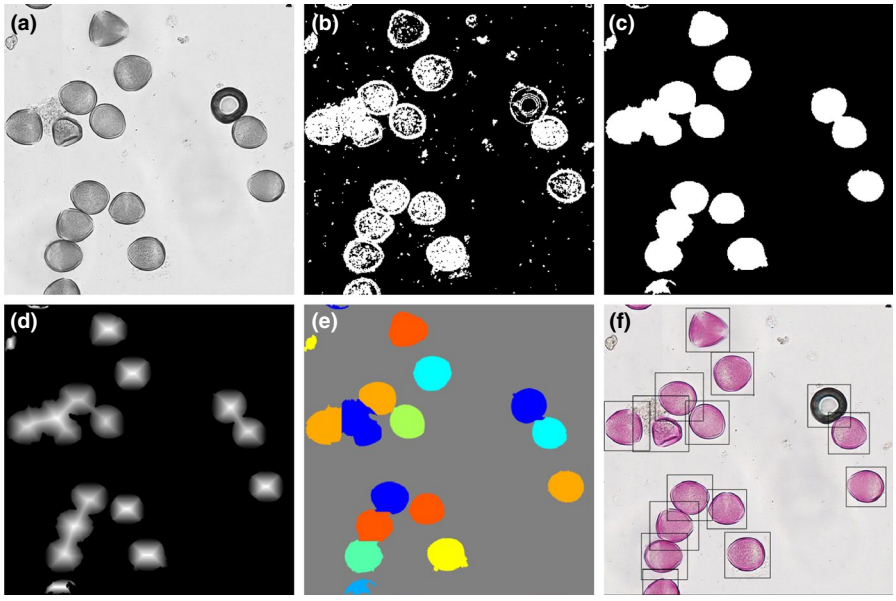

FIGURE 2 Steps involved in detecting and locating objects in images: (a) conversion to greyscale, (b) conversion to gradient mask and dilation using linear structuring elements, (c) filling, erosion and extraction of connected components, (d) calculation of distance to edge, (e) identification of individual grains by watershed analysis. (f) Resulting bounding boxes in original image

slide, but here we primarily focus on estimating relative proportions of species.

2.4 | Reference library preparation and CNN training

We sorted the extracted images from the reference dataset by manually selecting pollen grain images ascertained to be correctly determined and well representative of the species in that sample. Aiming for at least 1,000 pollen grains per species, we selected 200–500 images per sample per species (in total *c*. 122,000 images) from different samples accounting for variation in, for example, staining, debris, density and pollen development. Within each sample, we made sure to cover variability in shape, colouring and orientation among grains.

In our bumblebee data (see below), we only classified pollen items and not debris, bubbles or other non-pollen objects. Therefore, we only trained the models on pollen. However, to illustrate the possibility to separate pollen from non-pollen objects, we provide an example of that in the Supporting Information.

We applied transfer learning to train pre-trained deep learning CNNs in MATLAB 9.8 (R2020a). We used ResNet-18 (He et al., 2016) for all basic analyses but also trained one model with GoogLeNet (Szegedy et al., 2015), and one with Xception (Chollet, 2017), for comparison. Initial learning rate was set to 0.01 with a drop factor of 0.9 (i.e. 10% reduction in learning rate) in every epoch (complete run of the training data), and mini-batch size was 200 when running ResNet-18, 100 when running GoogLeNet and 24 for Xception (which requires more gpu memory). We trained the models for six epochs, at which point mini-batch accuracy was 100% (or very close to) in all cases.

2.5 | Cross-validation through splitting experiment

As a first evaluation of the CNN classification, we treated all pollen grains of a certain species as a single sample, regardless of which reference sample it originally came from. From this pooled set of pollen grains, a total of 1,000 per species were randomly picked, and then split in two groups, using 90% for training, and 10% for validation (Figure 3).

2.6 | Cross-validation through leave-one-out experiments

Convolutional neural networks models may become over-fitted to the training data (Xu et al., 2019), and validation based on the splitting experiment may therefore overestimate the accuracy achievable under new conditions. To create a more challenging validation situation, we kept the reference samples separate and, for each species, randomly picked one sample as the validation sample and trained the CNN model on the remaining samples (Figure 3). We iterated this procedure five times, making sure not to use the same sample for testing more than once. The procedure was possible to perform for species where we had five or more samples (36 out of 83). For the remaining species, we only had two to four samples, and had to reuse some of them during the iterations. Then, iterations do not represent entirely new situations, but we ensured that available samples were, as much as possible, equally represented as test samples. In

FIGURE 3 Schematic illustration of the splitting and leave-one-out experiments. On the left, three samples each of two species are shown. In the splitting experiment (upper panel), the samples within a species are pooled, and a fraction of all the pollen grains are used for CNN training, and the remainder for validation. In the leave-one-out experiment (lower panel), all grains in two of the samples are used for training, and the grains in the third sample are used for validation

each session, 500 pollen grains were selected from available samples of each species.

We repeated this procedure using 28 pollen types, formed by groups of related species with pollen hard to separate (e.g. Beug, 2004). The pollen types are shown in the right and lower margins of Figure 5 and in Figure S2. As we had more samples per type than per species, we did not have to reuse test samples between iterations.

2.7 | Cross-validation using bumblebee samples

Using Aperio ImageScope software (https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope/), we manually identified pollen from 370 bumblebee samples. We visually compared the grains in the samples to grains in our reference samples, and consulted published pollen floras (Beug, 2004; Sawyer, 2006) and picture libraries online (e.g. https://pollen.tstebler.ch/MediaWiki/). Thus, we classified the pollen into 28 identifiable types, plus one type (i.e. 29 in total) consisting of collapsed or damaged grains that could not be identified to the other types (hereafter termed damaged).

To quantify proportions of each pollen type in each bumblebee sample, we randomly placed five non-overlapping circles over the scanned area using the tools in ImageScope, and kept adding new circles until covering at least 150 identifiable pollen grains (not counting damaged grains) in each image. Selected grains were labelled with information about their type and location in the image. Unidentified pollen grains were labelled as unknown (0–5 per sample).

We used the label information (saved by ImageScope as xml-files) to extract individual images of all selected pollen grains. The total sample size of identified pollen used for cross-validation was 63,650.

Here, we first used a model trained on the 29 pollen types as described above, using 3,000 pollen grains per type. We then compared this with one alternative model trained on a dataset including six additional pollen types, for which we had sufficient reference samples (Supporting Information), but that were not present in the bumblebee data. The purpose of this was to test to what extent these 'empty types' attract false positives. We also created one alternative model where we trained the CNN on all 84 separate species (including the damaged pollen type), and then grouped the resulting frequencies into the 29 pollen types during analysis.

2.8 | Evaluation and optimization of CNN/model output

The CNN model outputs a vector $\mathbf{p}_i$ of classification scores for each pollen grain i in the test set, with one element for each class (i.e. species, or pollen type) that the model is trained for. The score of pollen grain i, for a particular species j represents the probability, $p_{i,j}$ that the grain belongs to that species, according to the model.

We based our analyses on the full information of classification scores, rather than on predicted classes (the species/type with maximum $p_{i,j}$; cf Harrell, 2015; Parmigiani & Inoue, 2009). This enables utilization of more information when assessing model certainty and fit to validation data, as well as when estimating sample proportions. Thus, we can measure the uncertainty in classification of individual pollen grains by the Shannon entropy, $H_i = -\sum_{j=1}^{C} p_{i,j} \log p_{i,j}$, where there are C species (Parmigiani & Inoue, 2009).

We use the true-positive probability (i.e. $r_i = p_{i,J*}$, where J^* is the true species identity of the pollen), to estimate three measures of performance, true positive rate, accuracy and deviance. Averaging the true-positive probabilities (r_i) across all pollen grains of a species, gives a continuous estimate of that species' recall rate, that is, the fraction of pollen grains of that species that the machine can correctly identify. Overall (or micro-averaged) accuracy is the fraction of correctly classified pollen grains in the whole sample, across all species, which we estimate as average r_i within samples, and is also equivalent to the multi-class 'micro-averaged' F_1-score (Powers, 2019). Class-averaged (macro-averaged) accuracy is average recall rates of all species in a sample, thus weighing each species (rather than each grain) equally, regardless of its frequency.

As r_i is the likelihood of data i (the manual classification of grain i) under the model, it is relevant to calculate deviance for each model, based on the summarized (log) scores for the true species: $D = -2\sum_{k=1}^{K}\sum_{i=1}^{N_k} \log r_{k,i}$, where K is the number of samples in the dataset, and sample k has N_k pollen grains.

We apply our method to pollen samples from individual bumblebees, which mostly contain only a few (1–4) species of pollen in our samples. In such cases, misidentification of pollen grains by the CNN is likely to estimate small proportions of species not actually in the sample, and thus overestimate the number of species.

To reduce the risk of such overestimation, we propose and test a method that adjusts the scores of individual pollen grains within each sample by the total frequencies in that sample. The estimated frequencies of species j in sample k is $f_{k,j} = \frac{1}{N_k}\sum_{i=1}^{N_k} p_{i,j}$, which is a vector $\mathbf{f}_k$. These frequencies can be used to adjust the scores for each pollen grain in the output from the CNN, by multiplying the original vectors of scores with the vector of sample frequencies, $\mathbf{p}'_i = \mathbf{p}_i \mathbf{f}_k^x$. The exponent x (≥0) is the number of times that $\mathbf{f}_k$ is multiplied into the original scores, and can be optimized during cross-validation (Supporting Information). Finally, the adjusted scores, $\mathbf{p}'_i$, must be normalized to sum to 1.

The logic of the adjustment assumes that there are few true species in the sample, and the machine identifies the majority of pollen correctly. Therefore, the high values of $\mathbf{f}_k$ are likely correct and the low more likely incorrect. Thus, proportions of rare species are reduced and those of common species boosted, and x determines the extent of this adjustment. This adjustment is only recommended when samples can safely be assumed to be comprised of a few species, and rare species are likely to be misclassifications.

3 | RESULTS

Our proposed method had high throughput, as the automated image extraction processed an average sample image (*c.* 24,000 × 24,000 pixels, 100 MB) containing *c.* 10,000 pollen

grains, per minute. Training the base CNN model (see below) took *c*. 30 min, after which it classifies at least 600 pollen grains per second. Preparation of slides was the same as for manual microscopy analysis, and once samples were prepared, only a few minutes per sample needed to be spent by a person. It took 5–10 min to prepare a rack of five slides for scanning, and 25–30 min to scan five slides (each with 5 *z*-layers). Setting up the *z*-stacking requires a few minutes manual time for a batch of images, and stacking takes *c*. 15 min machine time per image. Sorting of the reference library required more manual time (10–25 min per sample). A complete specification of processing times, by humans and machines, is given in Supporting Information.

The accuracy of the splitting experiment (Figure 3) was 0.98 (overall and class-averaged values identical as validation dataset is balanced), and recall rate ranged from 0.92 to 1.00 for the 83 species in the model (Figure 4a). Five species (*Crataegus monogyna*, *Prunus spinosa*, *Salix caprea*, *S. euxina* and *Thlaspi arvense*) had a recall rate <95%, and in all these cases the species had primarily been confused with congeneric species (or other members of the Brassicaceae family in the case of *T. arvense*).

Overall accuracy of the leave-one-out experiment (Figure 3) was 0.41 and class-averaged 0.49, but the confusion matrix indicated that confusions mostly occurred among closely related species (Figure 5). Species with more training samples generally had higher recall rates ($F_{4,78} = 7.21$, $p < 0.0005$; Figure 4b), mainly due to a difference between a single versus more training images.

By contrast, the leave-one-out experiment at the pollen type level, that is, the level used for manual classification where species are grouped into types, demonstrated that most groups could be reliably separated (Supporting Information). Accuracy across all pollen types in this case was 0.68, class-averaged 0.76. For 20

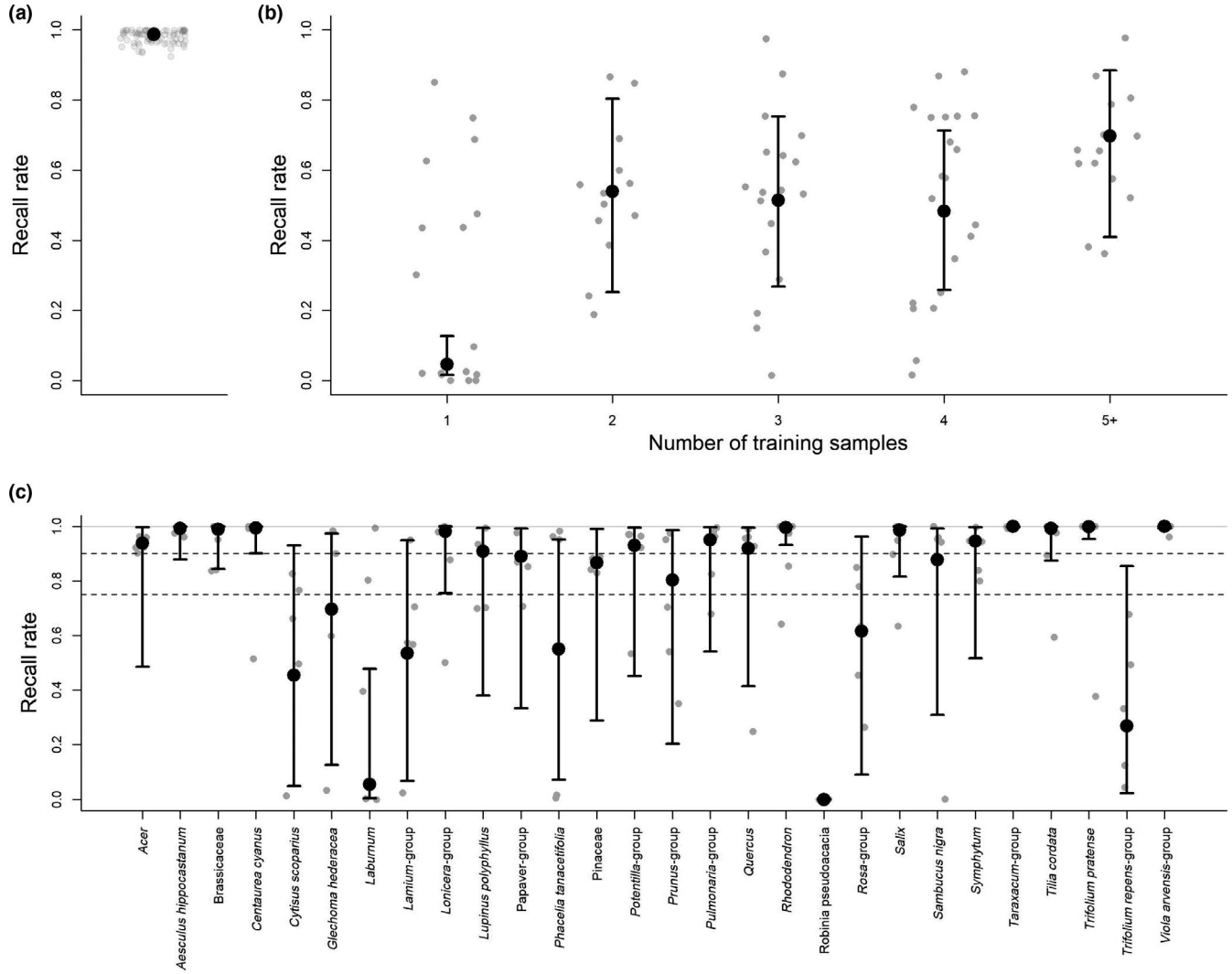

FIGURE 4 (a) Recall rates for 83 species in the splitting experiment. (b) Recall rates of species in the leave-one-out experiment. Grey dots are means of the five iterations for each species. The large black dot with error bars is mean value of the group with confidence interval, calculated at the logit scale. The horizontal axis shows the number of samples that were used for training for the species, where 5+ means five or more. (c) Recall rates of pollen types in the leave-one-out experiment. Grey dots are the individual values from the five iterations. The large black dot with error bars is mean value with confidence interval, calculated at the logit scale

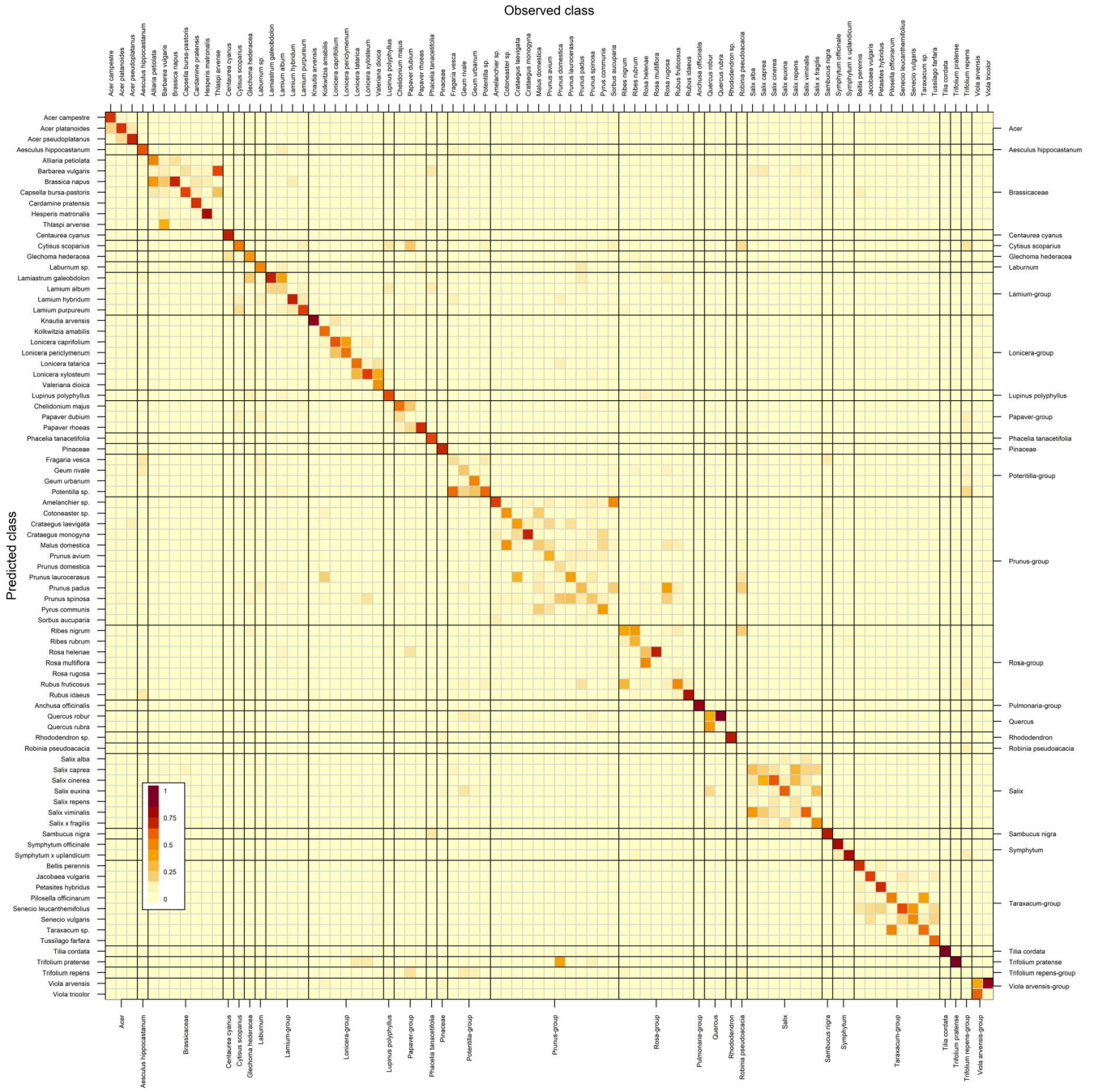

FIGURE 5 Confusion matrix resulting from the leave-one-out experiment at the species level. True species are shown in the columns, and predicted species in the rows. Colours indicate the frequency in each cell, as shown by the inset legend. The bold lines separate species into the pollen types shown in the opposite margins, and used in the subsequent analyses

out of 28 types, recall rate was above 0.75 and in 16 types above 0.90 (Figure 4c). The relationship between the number of training samples and recall rate was not significant at the pollen type level ($F_{1,26} = 1.127$, $p = 0.30$), but the pollen type (species) with the lowest recall (*Robinia pseudoacasia*) only had a single training sample.

When cross-validating our CNN model against the bumblebee data (63,650 pollen grains in 370 samples) overall accuracy was 0.79 across all pollen types and samples, and the class-averaged accuracy was 0.67. Although recall rate varied among pollen types, and was very low for some, it was higher than 0.7 for all pollen types represented by at least 2000 pollen grains in the bumblebee dataset (Figure 6; Supporting Information). Deviance of this base model was 124,063.

We adjusted classification scores by the estimated frequency per sample. Figure 7 shows three representative samples to exemplify the adjustment process. The adjusted scores resulted in higher overall accuracy (0.85), and adjusted recall rates were higher for 22 of 29 types (Paired $t_{28} = 2.58$, $p = 0.008$; Figure 6). Deviance was 17,853 lower compared to the base model, using an optimized adjustment factor of 1.4 (cf. Supporting Information).

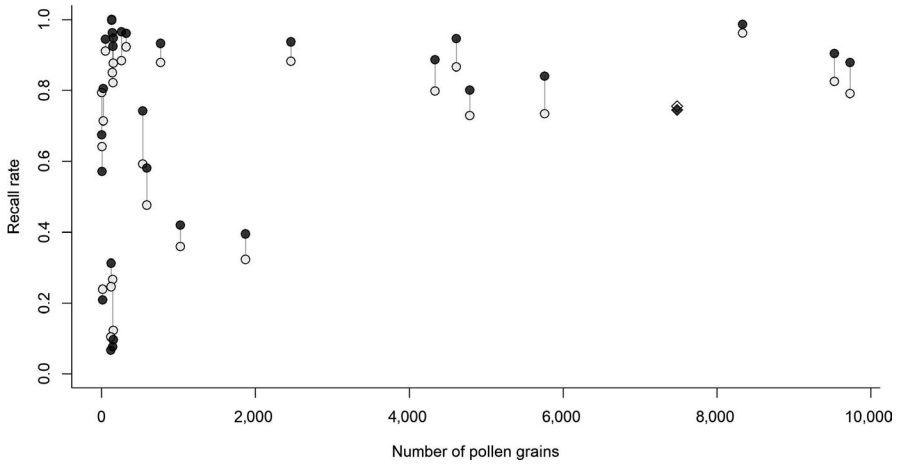

FIGURE 6 Recall rates of 29 pollen types sampled from bumblebees. Open symbols are the original recall rates, and filled are after adjustment. Diamonds are for the damaged pollen type

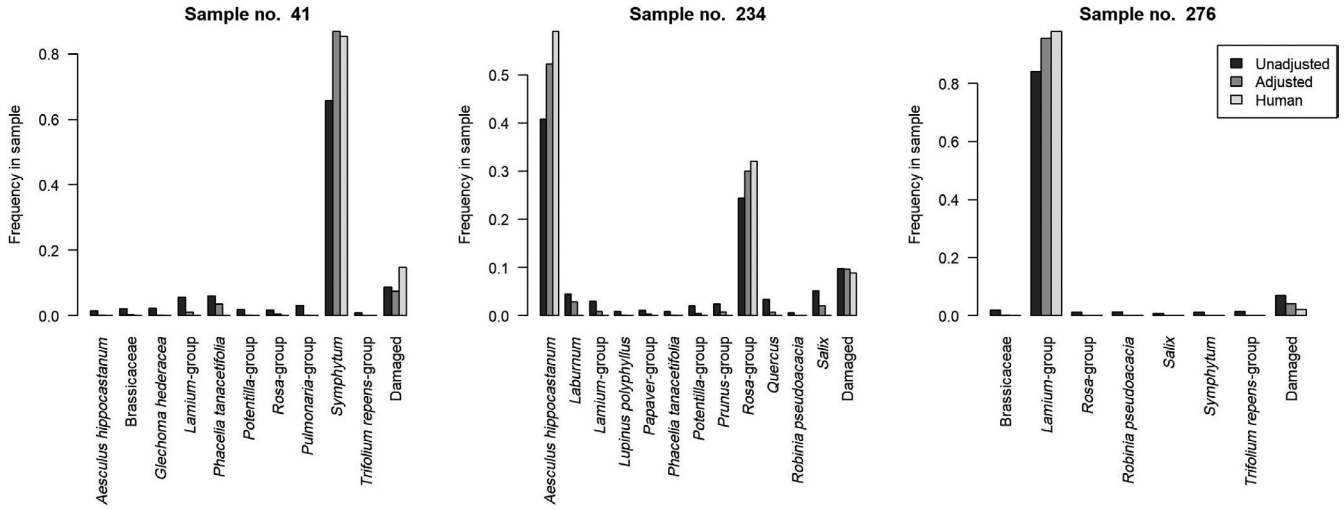

FIGURE 7 Relative frequencies of pollen types in three representative samples. Dark bars are initially estimated frequencies, medium grey bars are the adjusted frequencies and light bars are the frequencies as estimated by the human classification

The mean entropy per pollen grain was 0.23, in the unadjusted data, and 0.11 in the adjusted, that is, uncertainty in pollen type of each grain decreased with the adjustment (Figure 8a). Adjustment resulted in higher accuracies per sample (Paired $t_{369} = 25.1$, $p < 0.0005$; Figure 8b). Furthermore, the number of effective species (Shannon diversity) per sample was closer to the manual classification using the adjusted frequencies (linear regression, $r^2 = 0.30$; Figure 8d) than the original frequencies (linear regression, $r^2 = 0.22$; Figure 8c).

A CNN model trained with six additional pollen types, not present in the bumblebee dataset, produced an overall accuracy similar to the base model (0.77). However, the deviance was 1783 higher than for the base model. In total, there were 450 grains (incorrectly) classified to one of the six additional pollen types, that is, 0.7% of all grains.

To assess the effect of training on pollen types or individual species, we first trained a model on the individual 84 species (i.e. including damaged pollen as a separate species) and then summed the classification scores within the 29 pollen types, and could thus cross-validate this model in the same manner as for the previous ones. This model had an average accuracy of 0.70, and deviance 15,563 higher than for the base model.

A CNN model using GoogLeNet had an accuracy of 0.77 and a deviance 4,926 higher than the ResNet-18 (base) model. As mini-batch size had to be lowered to 100 (in comparison to 200 for ResNet-18) to fit in memory, the model was slightly slower to train (*c*. 40 min). A model based on Xception had an accuracy of 0.82, and deviance of 37,795 lower than the base model. However, this model took more than five times longer to train.

4 | DISCUSSION

We present a high-throughput method to analyse fresh pollen samples on microscope slides. Our results go beyond recent studies (Bourel et al., 2020; Dunker et al., 2020; Sevillano et al., 2020 and references reviewed therein) in several respects. First, using a larger pollen reference library, we were able to classify as many as 83 species accurately as determined by validation in the splitting experiment. Second, when exposing the model to a much more challenging

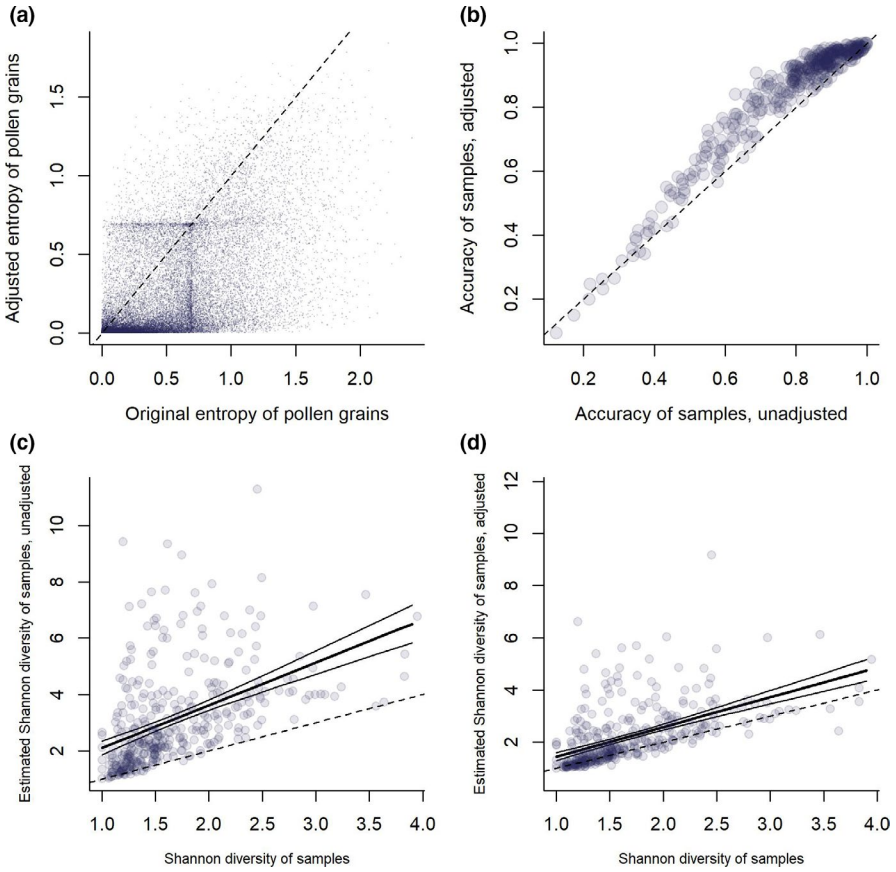

FIGURE 8 Effects of frequency adjustment on end results. (a) shows entropy of the individual pollen grains adjusted and unadjusted. (b) Shows the accuracy in samples, (c) shows the unadjusted machine estimated Shannon pollen type diversity in samples, against the same entity as estimated by human classification, (d) as in (c), but adjusted numbers

test (keeping training and validation samples separate), we show that the robustness is lower than a simple splitting experiment might indicate. Third, we apply our model to a large real dataset, collected from bumblebees and hence independent from the reference samples used in training. Fourth, we propose an adjustment of the classification scores to reduce the risk of overestimating diversity in situations where it can be assumed that samples contain only a few species, such as in the case of pollen collected by bees.

Our splitting experiment gives the same accuracy as recently published studies (Bourel et al., 2020; Dunker et al., 2020; Sevillano et al., 2020), despite our study having nearly twice the number of species. Our pollen library is also much larger than in previous studies, and we expect that as pollen libraries are extended, the capabilities of this kind of models will increase.

The lower accuracy of the model during the leave-one-out experiment at the species level demonstrates that true robustness might be much lower than indicated by a splitting experiment. In our case, this is probably because different reference samples from the same species vary in ways unrelated to species identity, leading to an over-fitted model (Xu et al., 2019). Other recent studies (Bourel et al., 2020; Dunker et al., 2020; Sevillano et al., 2020), using different preparation procedures, possibly have less between-sample, within-species variation, and therefore less over-fitted and more robust models. However, without a leave-one-out validation procedure similar to ours that is difficult to judge.

Most of our misclassifications occur within pollen types (i.e. groups of closely related species; Figure 5). The leave-one-out experiment based on 29 such pollen types indicates that at this level the model achieves acceptable accuracy, even if lower than for the splitting experiment. However, some misclassifications remain at the pollen type level (Supporting Information). In particular, *R. pseudoacacia* and the *Trifolium repens*-group were rarely correctly identified. The former was represented by two reference samples only (i.e. one for training and one for validation), which may explain the poor performance. The poor validation performance for the *T. repens*-group, consisting of several species with a total of eight reference samples, is more difficult to explain. At this point, we do not know if identification to species level would be possible with more reference samples for each species, but it appears possible since recall rates increase with number of samples.

We found similar model performance on the bumblebee samples compared to the reference samples (leave-one-out experiment using pollen types). This is important because the bumblebee data constitutes real data from a field study, a situation for which the model needs to be useful (e.g. Beil et al., 2008). However, a shortcoming is that we implicitly assume that our manual classification has been done without error, which is unlikely (Gonçalves et al., 2016), and unfortunately we do not have any independent estimate of the accuracy of the manual classification.

The system we propose has very high throughput, with the possibility to process (slide preparation and scanning) 50–100 samples per person per day (Supporting Information). Using a trained CNN model, 30–60 samples, with up to 10,000 pollen grains each, can be processed by the machine every hour. That is, even large projects can

be accommodated by modest laboratory and computing facilities. Building and extending the pollen library is time-consuming, but once in place it will be useful for future projects. Research in pollination ecology and conservation (e.g. Kleijn & Raemakers, 2008), and other fields where pollen analysis is important (Holt & Bennett, 2014), can benefit from our method, facilitating analysis of a higher number of samples than manual visual identification. Moreover, in comparison to meta-barcoding (Bell et al., 2019), our method allows quantitative analyses of pollen content.

Our analyses are based on bright-field microscopy of fuchsine stained pollen (Kearns & Inouye, 1993). This is one of the simplest and fastest methods to process samples, but not necessarily the most precise (Jones, 2014). Some previous studies have been based on acetolysed pollen (e.g. Bourel et al., 2020; Sevillano et al., 2020), and Sevillano et al. (2020) processed their images in ways to further enhance three-dimensionality, bringing out additional visible details, likely improving possibilities for classification. By contrast, our method has the benefit of maintaining pollen shape, and other visual structures lost during acetolysis. For our case, working with fresh samples and aiming at a fast, high-throughput protocol, the method we used has benefits. However, for other applications, where fossil or subfossil pollen is analysed (e.g. Bourel et al., 2020; Stillman & Flenley, 1996), acetolysis is necessary, and our sample preparation is not an option. As yet another option for fresh pollen, Dunker et al. (2020) used flow cytometry in combination with CNN, which provides additional possibilities to obtain detailed structural information.

There will obviously be cases where either acetolysed or fresh samples are preferred. Therefore, parallel libraries should be built. Currently, there are several publicly available libraries that are useful for deep learning applications (e.g. the POLEN23E library Gonçalves et al., 2016; Sevillano & Aznarte, 2018; or the Classifynder system Sevillano et al., 2020). Pollen libraries are investments in time and effort, that can be shared and open for user contributions, providing an infrastructure is built including systems for reviewing samples, etc. We envision a system where users contribute their reference samples, tailor a model for the pollen types they require, and analyse their samples online, in a way similar to molecular bioinformatics.

We have maintained a probabilistic approach throughout analyses. That is, rather than classifying each pollen grain according to the highest score in a binary manner, we kept the classification scores (probabilities of species identity) as is. This allowed evaluations of the classifications in more detail (Harrell, 2015; Parmigiani & Inoue, 2009) as well as adjustment of classification scores using sample frequencies. The benefit of the frequency adjustment should be limited to cases where it can be safely assumed that samples are dominated by a few species. Optimizing the adjustment factor (in our case to 1.4), probably requires a large validation dataset, which is often impractical. However, using an adjustment factor of 1, may still be a preferred alternative to, for example, deciding to ignore classes with frequencies below a certain percentage (cf. Beil et al., 2008; Kleijn & Raemakers, 2008). The adjustment improved recall rates of species by correcting misidentifications, except for a few rare species and improved accuracy of the vast majority of samples. Importantly, it also brought the estimated species diversity of samples closer to the value estimated by the manual analysis. The probabilistic approach additionally allows efficient manual validation of model performance through a rather simple protocol. First, it should be made sure that pollen grains classified with high certainty (low entropy) are indeed correct (a small subset of a few hundred grains should suffice). Second, a subset of pollen grains with high entropy (i.e. classified with low certainty) should be checked, as these indicate problematic cases, for example, pollen types missing during model training.

In conclusion, we have shown that a useful number of pollen types can be well analysed in an automated system, which is reasonably accurate, highly quantitative, very efficient, and could be made widely accessible. However, there is still ample room for improvement. We expect that some of these improvements will be in machine learning algorithms, but that even more improvement will be made by extending pollen libraries with more samples of more species.

ACKNOWLEDGEMENTS

This study was funded by grant 20170867 from the Crafoord Foundation, grant 215-2014-1603 from Formas and a grant from Region Skånes Miljövårdsfond to O.O., H.G.S. and A.S.P. The strategic research area BECC (Biodiversity and Ecosystem Services in a Changing Climate) provided support to infrastructure. We thank Albin Andersson, Anna Berg, Björn Klatt and Sara Hellström, for help in the field and laboratory, and Matina Donaldson-Matasci for initial inspiration for this project.

AUTHORS' CONTRIBUTIONS

O.O. conceived the idea together with H.G.S. and A.S.P.; O.O., A.S.P., V.V., J.Y. and M.K. collected, prepared and compiled the reference library; M.K. performed the manual pollen analysis; O.O. programmed models and functions; O.O. and M.S. analysed the data; O.O. and M.S. wrote the manuscript with substantial input from all other authors.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13575.

DATA AVAILABILITY STATEMENT

Data available from the Dryad Digital Repository https://doi.org/10.5061/dryad.n8pk0p2tw (Olsson et al., 2021).

ORCID

Ola Olsson https://orcid.org/0000-0003-0789-1064
Melanie Karlsson https://orcid.org/0000-0002-2865-2284
Anna S. Persson https://orcid.org/0000-0002-2711-0344
Henrik G. Smith https://orcid.org/0000-0002-2289-889X
Johanna Yourstone https://orcid.org/0000-0002-5667-1538
Martin Stjernman https://orcid.org/0000-0002-5088-8840

REFERENCES

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1–6). IEEE Xplore. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Bastl, K., Berger, U., & Kmenta, M. (2017). Evaluation of pollen apps forecasts: The need for quality control in an eHealth service. *Journal of Medical Internet Research*, *19*, e152. https://doi.org/10.2196/jmir.7426

Beil, M., Horn, H., & Schwabe, A. (2008). Analysis of pollen loads in a wild bee community (Hymenoptera: Apidae) – A method for elucidating habitat use and foraging distances. *Apidologie*, *39*, 456–467. https://doi.org/10.1051/apido:2008021

Bell, K. L., Burgess, K. S., Botsch, J. C., Dobbs, E. K., Read, T. D., & Brosi, B. J. (2019). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology*, *28*, 431–455. https://doi.org/10.1111/mec.14840

Bertrand, C., Eckerter, P. W., Ammann, L., Entling, M. H., Gobet, E., Herzog, F., Mestre, L., Tinner, W., & Albrecht, M. (2019). Seasonal shifts and complementary use of pollen sources by two bees, a lacewing and a ladybeetle species in European agricultural landscapes. *Journal of Applied Ecology*, *56*, 2431–2442. https://doi.org/10.1111/1365-2664.13483

Beug, H.-J. (2004). *Leitfaden der Pollenbestimmung*. Verlag Dr Friedrich Pfeil.

Bourel, B., Marchant, R., de Garidel-Thoron, T., Tetard, M., Barboni, D., Gally, Y., & Beaufort, L. (2020). Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains. *Computers & Geosciences*, *140*, 104498. https://doi.org/10.1016/j.cageo.2020.104498

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258). https://doi.org/10.1109/CVPR.2017.195

Daood, A., Ribeiro, E., & Bush, M. (2016). Pollen recognition using a multilayer hierarchical classifier. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 3091–3096). https://doi.org/10.1109/ICPR.2016.7900109

Dunker, S., Motivans, E., Rakosy, D., Boho, D., Mäder, P., Hornick, T., & Knight, T. M. (2020). Pollen analysis using multispectral imaging flow cytometry and deep learning. *New Phytologist*, *229*(1), 593–606. https://doi.org/10.1111/nph.16882

Gonçalves, A. B., Souza, J. S., Silva, G. G. D., Cereda, M. P., Pott, A., Naka, M. H., & Pistori, H. (2016). Feature extraction and machine learning for the classification of Brazilian Savannah Pollen Grains. *PLoS ONE*, *11*, e0157044. https://doi.org/10.1371/journal.pone.0157044

Harrell, F. E. J. (2015). *Regression modeling strategies - With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed). Springer International Publishing.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Heinrich, B. (1979). 'Majoring' and 'Minoring' by Foraging Bumblebees, Bombus Vagans: An Experimental Analysis. *Ecology*, *60*, 246–255. https://doi.org/10.2307/1937652

Holt, K. A., & Bennett, K. D. (2014). Principles and methods for automated palynology. *New Phytologist*, *203*, 735–742. https://doi.org/10.1111/nph.12848

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243

Jardine, P. E., Gosling, W. D., Lomax, B. H., Julier, A. C. M., & Fraser, W. T. (2019). Chemotaxonomy of domesticated grasses: A pathway to understanding the origins of agriculture. *Journal of Micropalaeontology*, *38*, 83–95. https://doi.org/10.5194/jm-38-83-2019

Jones, G. D. (2014). Pollen analyses for pollination research, acetolysis. *Journal of Pollination Ecology*, *13*, 203–217. https://doi.org/10.26786/1920-7603%282014%2919

Kearns, C. A., & Inouye, D. W. (1993). *Techniques for pollination biologists*. University Press of Colorado.

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., Rost, S., Härtel, S., & Steffan-Dewenter, I. (2015). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology (Stuttgart, Germany)*, *17*, 558–566. https://doi.org/10.1111/plb.12251

Kleijn, D., & Raemakers, I. (2008). A retrospective analysis of pollen host plant use by stable and declining bumble bee species. *Ecology*, *89*, 1811–1823. https://doi.org/10.1890/07-1275.1

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105). https://doi.org/10.5555/2999134.2999257

Lang, D., Tang, M., Hu, J., & Zhou, X. (2019). Genome-skimming provides accurate quantification for pollen mixtures. *Molecular Ecology Resources*, *19*, 1433–1446. https://doi.org/10.1111/1755-0998.13061

Leonhardt, S. D., & Blüthgen, N. (2012). The same, but different: Pollen foraging in honeybee and bumblebee colonies. *Apidologie*, *43*, 449–464. https://doi.org/10.1007/s13592-011-0112-y

Lumini, A., & Nanni, L. (2019). Deep learning and transfer learning features for plankton classification. *Ecological Informatics*, *51*, 33–43. https://doi.org/10.1016/j.ecoinf.2019.02.007

Olsson, O., Karlsson, M., Persson, A. S., Smith, H. G., Varadarajan, V., Yourstone, J., & Stjernman, M. (2021). Data from: Efficient, automated and robust pollen analysis using deep learning. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.n8pk0p2tw

Parmigiani, G., & Inoue, L. (2009). *Decision theory: Principles and approaches*. John Wiley & Sons.

Persson, A. S., Mazier, F., & Smith, H. G. (2018). When beggars are choosers—How nesting of a solitary bee is affected by temporal dynamics of pollen plants in the landscape. *Ecology and Evolution*, *8*, 5777–5791. https://doi.org/10.1002/ece3.4116

Pertuz, S., Puig, D., Garcia, M. A., & Fusiello, A. (2013). Generation of all-in-focus images by noise-robust selective fusion of limited depth-of-field images. *IEEE Transactions on Image Processing*, *22*, 1242–1251. https://doi.org/10.1109/TIP.2012.2231087

Powers, D. M. W. (2019) What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. *arXiv*, 1503.06410.

Sawyer, R. (2006). *Pollen identification for beekeepers*. Northern Bee Books.

Sevillano, V., & Aznarte, J. L. (2018). Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. *PLoS ONE*, *13*, e0201807. https://doi.org/10.1371/journal.pone.0201807

Sevillano, V., Holt, K., & Aznarte, J. L. (2020). Precise automatic classification of 46 different pollen types with convolutional neural networks. *PLoS ONE*, *15*, e0229751. https://doi.org/10.1371/journal.pone.0229751

Stillman, E. C., & Flenley, J. R. (1996). The needs and prospects for automation in palynology. *Quaternary Science Reviews*, *15*, 1–5. https://doi.org/10.1016/0277-3791(95)00076-3

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR) (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

von der Ohe, W., Persano Oddo, L., Piana, M. L., Morlot, M., & Martin, P. (2004). Harmonized methods of melissopalynology. *Apidologie*, *35*, S18–S25. https://doi.org/10.1051/apido:2004050

Wood, T. J., Holland, J. M., & Goulson, D. (2017). Providing foraging resources for solitary bees on farmland: Current schemes for pollinators benefit a limited suite of species. *Journal of Applied Ecology*, *54*, 323–333. https://doi.org/10.1111/1365-2664.12718

Xu, Q., Zhang, M., Gu, Z., & Pan, G. (2019). Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs. *Neurocomputing*, *328*, 69–74. https://doi.org/10.1016/j.neucom.2018.03.080

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Olsson O, Karlsson M, Persson AS, et al. Efficient, automated and robust pollen analysis using deep learning. *Methods Ecol Evol.* 2021;12:850–862. https://doi.org/10.1111/2041-210X.13575