Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains

Benjamin Bourel, Ross Marchant, Thibault de Garidel-Thoron, Martin
Tetard, Doris Barboni, Yves Gally, Luc L Beaufort

Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains

Benjamin Bourel*[1], Ross Marchant[1], Thibault de Garidel-Thoron[1], Martin Tetard[1], Doris Barboni[1], Yves Gally[1], Luc Beaufort[1]

[1] CEREGE, Aix Marseille Université, CNRS, IRD, INRAE, Coll. France, Technopole Arbois, 13545 Aix en Provence cedex 4, France

*Corresponding author: bourel@cerege.fr

Link to the code: https://github.com/Beniofh/Multi-CNN

Highlights

- We developed an automated recognition software for pollen data acquisition
- Our software uses multiple convolutional neural network with decision tree (multi-CNNs)
- Augmentation of stacked optical images increases classification accuracy
- Our software successfully recognizes pollen at the genus or species rank
- It achieves robust identification of intact, damaged, modern, and fossil pollen

Abstract

Pollen grains are valuable paleoclimate and paleovegetation proxies which require extensive knowledge of morphotypes and long acquisition time under the microscope. The abundance of damaged, folded, and broken pollen grains in the fossil register and sometimes also in modern soil and sediment samples, has so far prevented automation of pollen identification. Recent improvements in machine learning, however, have allowed reconsidering this approach. Here we present an automated approach which is capable of assisting palynologists with poorly preserved pollen samples. Called multi-CNNs, this approach is based on multiple convolutional neural networks (CNNs) integrated in a decision tree system. To test it, we built a system designed for three botanical families very common in the modern and fossil pollen assemblages of Eastern Africa, namely Amaranthaceae, Poaceae, and Cyperaceae. Our system was tested on stacked optical images of 8 pollen types (6 Amaranthaceae, 1 Poaceae, 1 Cyperaceae) using a training dataset of 1102 intact pollen grains and three validation datasets of intact (276 grains), damaged (223 grains), and fossil pollen (97 grains). We show that our system successfully recognizes intact, damaged, and fossil pollen grains with very low misclassification rates of 0%, 2.8%, and 3.7%, respectively. The use of augmentation on stacked optical images during the training increases classification accuracy. Following a palynologist's approach, our system allows grains without obvious characters to be classified into a class of high taxonomic level or as indeterminable pollen. This is the first software able to process grains with a wide range of taphonomical stages, which makes it the first truly applicable to automated pollen identification of fossil material.

1. Introduction

Since the late 1960s, the palynologist community has been interested in the automation of pollen grain recognition (Flenley, 1968). This desire comes from a real need to reduce the long and tedious process of pollen grain identification and counting, which requires an extended knowledge of palynofloras in order to obtain accurate and statistically valid paleovegetation inferences. Pollen counting and recognition process is an important limiting factor that makes it difficult to build large databases. To date, with the advances of large datasets for paleoenvironmental studies to resolve high spatial and/or temporal resolution, there is a crucial need to reduce the time involved in pollen data acquisition. Palynology is also a discipline at risk today because acquiring sufficient knowledge of palynofloras requires years of experience, particularly in tropical regions where plant diversity is highest. Despite the fact that pollen grains are still the only paleovegetation proxies allowing plant identifications up to the genus and species levels, their study is easily supplemented by other indicators such as $\delta^{13}C$. They require less physical presence of scientists in the acquisition although they do not allow such fine environmental reconstitutions.

In the East African Rift, many paleontological areas have preserved abundant fauna, including hominin remains of Pliocene and Pleistocene age, but few have preserved paleobotanical remains. Yet, pioneering pollen studies in the early 70s have proved to be a powerful tool for inferring paleoenvironments associated with early hominins in Africa (Bonnefille, 1979). Among these, the *Australopithecus afarensis*-bearing Hadar formation in the North Ethiopian Rift was extensively studied, and several levels proved to be rich in pollen grains (Bonnefille *et al.*, 1987, 2004; Bonnefille, 2010). Those levels were analyzed by palynologists of Bonnefille's team, usually over several weeks, in order to obtain adequate pollen counts (generally between 300 and 1000 grains depending on the pollen diversity of each sample) as the depositional and arid climate contexts were not favorable to pollen preservation (R. Bonnefille, pers. com.). Recently, the Hominin Sites and Paleolakes Drilling Project (HSPDP) drilled in this area to obtain paleoenvironmental records with high temporal resolution of the variations of the local hominin environment between 2.9 and 3.6 million years (Ma) ago (Cohen *et al.*, 2016). We have been studying two of those HSPDP cores, collected from near the lacustrine depocenter of Hadar Basin, the Northern Awash Osi-Isi (NAO, 11.31518°N, 40.73689°E) and from a location 3 km east of NAO, the Northern Awash Woranso (NAW, 11.32535°N, 40.76491°E). NAO and NAW were cored respectively to a depth of 187 and 245 meters below surface (Mohan *et al.*, 2016). Analysis of 8 samples out of 175 from these two cores took one of the authors (B.B.) several months to complete, because both the preservation and concentration of pollen grains were low. A high-resolution pollen record at this site would give a fantastic record of paleovegetation changes during the time when *Australopithecus afarensis* thrived and when the first representatives of the *Homo* genus emerged (Maslin *et al.*, 2015; Villmoare *et al.*, 2015), but such a study would require more than one scientist working full time over several months. It is a classic case of "arid" sedimentary archives with very important environmental pollen data that cannot be correctly studied due to human limitations. Testing and exploring the potential of automated recognition to pollen taxa is, therefore, crucial.

In the 1960s, computer performance, image processing, and machine learning techniques were not sufficiently developed to realize Flenley (1968)'s visionary concepts of pollen recognition. During the last decade, however, dramatic improvements opened the way to automated

recognition of pollen, through different approaches such as (1) feature extraction, (2) texture analysis, and (3) the combination of both. (1) Regarding feature extraction, the first breakthrough study by García *et al.* (2012) combined morphological methods based on contour feature extraction of light microscopy (LM) images with a hidden Markov model (HMM) classifier. This approach allowed classifying 17 different pollen types from 11 different botanical families of honey bee's plants with 98.8% accuracy. Kaya *et al.* (2013) used morphological measurements (polar axis length, equatorial axis length, colpus weight, exine thickness, etc.) with a data mining classifier to predict group membership for data instances. This approach allowed 20 very similar species of pollen grains in the genus *Onopordum* to be classified with an accuracy of 90.6%. The work is impressive but requires an initial manual/subjective definition of the recognition characteristic features and is only applicable to intact pollen grains. In addition, the accuracy of Kaya *et al.* (2013)'s model is dependent of the choice of characteristic features (Treloar *et al.*, 2004).

(2) Another set of studies focused on texture analysis, with texture features derived from gray level patterns of the grain surface (Fernandez-Delgado *et al.*, 2003). Using a combination of local linear transformation (LLT) feature vectors and a support vector machine (SVM) classifier, Fernandez-Delgado *et al.* (2003) were able to classify LM images of fresh pollen of five pollen types (in four different families) with an accuracy of 76%. Later, a more complex approach was developed by Marcos *et al.* (2015) using a texture feature vector composed of gray-level co-occurrence matrices (GLCM), log-Gabor filters (LGF), local binary patterns (LBP) and discrete Tchebichef moments (DTM), and Fisher's discriminant analysis (FDA) and k-nearest neighbor (KNN) to reach a final classification at the family level. This study achieved an accuracy of 94.8% for LM images of fresh pollen grains of 15 pollen types in 12 families (Marcos *et al.*,2015), but the texture approach has been shown to be sensitive to grain rotation, lighting, image noise and the quality of texture preservation (Han & Xie, 2018). To circumvent these problems, Han and Xie (2018) proposed to use local decimal patterns (LDP) instead of LBP, an approach robust to the noise and rotation of pollen images but which nevertheless could not tackle the challenge of recognizing damaged pollen.

(3) A combination of texture features and morphological analyses represents the last development. Rodriguez-Damian *et al.* (2006) developed a classification system using both shape and texture analysis of LM images of pollen grains directly extracted from plants. Using a minimum distance classifier (MDC), multilayer perceptron (MLP) and SVM, three morphologically similar pollen types of Urticaceae were classified with an accuracy of 89%. Later, Chudyk *et al.* (2015) used a combination of 15 morphological features and five textures features on a set of LM images including one pollen type of Poaceae, one of Asteroideae and three of Betulaceae. By using SVM and Random Forest classifiers, they reached a success rate of 89% for intact pollen grains. However, this approach still performed poorly on damaged pollen.

Recently, the development of convolutional neural networks (CNNs) has opened a new era in image recognition. CNNs automatically extract the features of interest in an image that are relevant to the classification task (Daood, 2018). For pollen, CNNs with different topology were tested to classify 23 pollen types with large morphological differences (Sevillano & Aznarte, 2018). They achieved an accuracy of 97.2% with a LM images dataset (Gonçalves *et al.*, 2016) of intact pollen grains. Daood *et al.* (2018) recognized z-stacked LM images of intact pollen by combining CNN feature vectors with recurrent neural networks (RNN). They achieved an accuracy of 100% for the classification of 10 pollen types of different families. Comparisons of these methods and their success rates on the same data set of images showed that the approaches based on CNNs are always the most accurate, and point toward the efficiency of CNNs in classifying natural objects such as pollen.

All these approaches proved to be efficient only on fresh or generally intact pollen grains, far from the reality faced by palynologists who must deal with damaged pollen due to taphonomic processes. Holt *et al.* (2011), Holt and Bennett (2014) and recently Khanzhina *et al.* (2018) noticed the absence of a system adapted to "broken, deformed and clumped pollen" and explained it by technological limitations. Han and Xie (2018) pioneered the analysis on damaged pollen grains, unfortunately their approach did not succeed. Punyasena *et al.* (2012), however, successfully classified fossil grains of two *Picea* species (*P. mariana* and *P. glauca*) with an accuracy of 93.8%. They used a KNN with layered learning systems based on kernel density, applied to extracted shape and texture features from 3D fluorescence microscope (FM) images. Using the same set of pollen grains as Punyasena *et al.* (2012), Kong *et al.* (2016) also obtained a high accuracy of 86.1% in classifying fossil grains of three *Picea* species (*P. mariana*, *P. glauca* and *P. critchfieldii*) using dictionary learning sparse coding (DLSC) on shape and texture features extracted from confocal FM images. However, the performances were largely dependent on the selection and quality of sample blocks (Kong *et al.*, 2016; Han & Xie, 2018). These results are nonetheless very promising, but we note that the deterioration of the *Picea* pollen considered in those two studies is very subtle (as is generally observed in peat sediments, e.g. Curry, 2007), and far from resembling the Pliocene pollen records of Eastern Africa.

The aim of this study is to develop a new system of automated identification of pollen that could be applied to down core fossil pollen grains. This system should be able to handle damaged and broken pollen grains, and should be easily applicable to images obtained with light microscopy (LM), as is commonly done in palynological laboratories. It should also allow identifications at higher taxonomical levels when identification at the species level is not reached.

Following recent developments in machine learning we chose to develop our

system using sequential multiple CNNs. As a proof of concept for this study we chose eight pollen types: six among the Amaranthaceae, as they have very similar overall shape (spherical, periporate), one Poaceae (undifferentiated taxa), and one Cyperaceae (undifferentiated taxa), that is two Monocotyledonae with heterogeneous and varying surface patterns (psilate to finely verrucate). We focused on Amaranthaceae, Poaceae, and Cyperaceae because these three families represent 50% to 90% of the total pollen assemblage in both terrestrial and marine archives from Eastern Africa (Bonnefille *et al.*, 1987, Bonnefille, 2010). Pollen-based biome reconstruction for Africa also largely rely on the presence and relative abundance of Amaranthaceae pollen (Sobol & Finkelstein, 2018). Amaranthaceae species, despite exhibiting very similar pollen morphologies are important ecological markers ('African Plant Database', 2019). Being able to identify and count separately different Amaranthaceae could definitely improve paleoenvironment inferences. As a first step, this study makes no attempts to separate types among Poaceae and Cyperaceae, also because such discriminations have only been achieved so far using with Scanning Electron Microscopy (Mander et al., 2014).

2. Material and Methods

2.1. Datasets and image acquisition

We prepared one training and three validation datasets by taking pictures of pollen grains under light microscopy (LM) (Supporting Information, Table **S1**). The training dataset consists of 1102 globally intact pollen grains (hereafter referred to as the training dataset). The three validation datasets consist of 1) 276 grains of intact pollen (hereafter referred to as the intact dataset), 2) 223 deformed, folded, fragmented and/or broken grains (hereafter referred to as the damaged dataset), and 3) 97 pollen grains from Holocene and Pliocene age (hereafter referred to as the fossil dataset) (Fig. **1**). In the fossil dataset, pollen grains also have an altered exine and mineral inclusions (Fig. **1**). Eight pollen types were imaged with specimens taken from the

reference collection of East African pollen species hosted at CEREGE, Aix-en-Provence, France (Supporting Information, Table **S1**) to constitute the training dataset, the intact dataset, and the damaged datasets: *Achyranthes*-type *aspera*, *Aerva*-type *javanica*, *Aerva*-type *lanata*, *Celosia*-type *argentea*, *Cyathula*-type *orthacantha, Digera*-type *muricata*, Poaceae, and Cyperaceae). The fossil pollen dataset includes only five of the eight pollen types available in the reference collection due to the limited availability of fossil material: *Achyranthes*-type *aspera*, *Aerva*-type *javanica*, *Digera-type muricata*, Cyperaceae, and Poaceae (for sampling locations, depth, and ages see Supporting Information, Table **S1**). The damaged and fossil datasets were built using samples which were not considered for the training and intact pollen datasets to avoid biases in the estimation of the effectiveness of our approach.

Each pollen grain was imaged under natural light by taking a stack of pictures at 1 μm steps between the top and bottom surfaces of the grain, using an automated custom Leica DMRBE microscope controlled with a software. This step is semi-automated, whereby the image stack movement (z-axis) is automatic but setting the position (x and y axes) is performed manually. Since the stacking step is fixed to 1μm, the number of fused images is proportional to the size of the pollen. The number of images in the stack ranged from 20 to 40 depending on pollen type. Each stack was split into two halves, and only the half corresponding to the top hemisphere of each pollen grain was kept. The image stack was fused to create a single in-focus image (z-stacked imaging process) with the Helicon Focus 6 software, using the "depth map" method with radius set to 50 pixels and smoothing set to 10 pixels. We chose 1 μm steps empirically, as above this threshold the morphological information was less visible in the fused images and below this value, aberrations appeared into fused images. Each fused image was analyzed using a custom script written for the Fiji image analysis software, to center and crop the pollen into a square image (Schindelin *et al.*, 2012). Finally, all reframed fused images were resized to 128x128 pixels.

2.2. Image classification

Our system of automatic recognition is developed in the Python language (Pérez et al.,2011) and principally uses the Keras and TensorFlow open-source software libraries (Abadi *et al.*, 2016; Chollet, 2015). It uses convolutional neural networks (CNNs) with a linear architecture similar to VGGNet (Simonyan & Zisserman, 2014). The CNN training was done with the training pollen dataset and the validation dataset with intact pollen. CNNs training was carried out on batches of 101 pictures (out of a grand total of 1102 pictures) over 500 epochs (the learning algorithm is an iterative process; an epoch is one training iteration on all pictures). To optimize the convergent behavior in loss function, which ensures a more precise adjustment of the weights of our network, we used regressive learning rate as a function of the epoch such as epoch [0; 200]=0.001, epoch [201; 300]=0.0006, epoch [301; 400]=0.00036, and epoch [401; 500]=0.000216.

We added an augmentation procedure to most of our trainings to generate new images at each epoch. This procedure generates new images by modifying the originals through random rotations (0°to 90°), random horizontal and vertical flips, and random width and height shifts (0% to 25%). But learning often did not progress when using augmentation because the generation of new images at each epoch made it difficult for the CNN to extract features. To solve this problem, we pre-trained the CNN without augmentation to extract features in its initial layers, then continued training using augmented images. By using this pretrained network for the training with augmentation, the CNN learns to combine these features at the end of the initial layers. This process is commonly called transfer learning (Daood, 2018) (Supporting Information, Fig. **S1**).

The CNNs trained with augmentation are integrated in a decision tree to form a multiple convolutional neural network (multi-CNNs) (Fig. **2**). The multi-CNNs is composed of five main steps. Step 1 aims at isolating Amaranthaceae from

Monocotyledoneae. Step 2 is made to evaluate if pollen identified as Monocotyledonae in step 1 could be altered Amaranthaceae. Indeed, altered pollen of Dicotyledoneae which have lost their ornamentations and part or all their outer layer (ectexine) may show similarities with Monocotyledonae that only have one layer of exine. To do this in step 2, we used a finer mapping with a separation of Poaceae and Cyperaceae which exhibit different morphologies when intact or slightly altered. Altered Monocotyledonae such as Poaceae and Cyperaceae may easily lose parts of their exine and be deformed or broken, because they are made of a single thin layer. To optimize the chances of recognition of altered pollen, two networks with the same mapping were used but with different augmentation, one focused on shape (CNN 2.1), and the other on texture (CNN 2.2). In step 3A, several CNNs were used; each one being specialized in the recognition of one pollen type of Amaranthaceae to maintain CNN high accuracy. The results of all these specialized CNNs are then compiled, the pollen is classified into one Amaranthaceae pollen type or as Amaranthaceae uncertain. In step 3B, the same logic is used as for step 2: two networks with the same mapping are used with different augmentation to separate Poaceae and Cyperaceae. If the results of the two CNNs are incongruous, the pollen is classified as an uncertain Monocotyledoneae. Finally, in step 4, we use the results of the CNNs 2.1 and 2.2 to classify Monocotyledoneae uncertain and Amaranthaceae uncertain in Monocotyledoneae indeterminable, Amaranthaceae indeterminable, or indeterminable pollen.

To evaluate the efficiency of augmentation and multi-CNNs, we tested three other approaches: (i) one using multi-CNNs without augmentation in step 1 and 3A, (ii) one simple-CNN without augmentation, and (iii) another simple-CNN with full augmentation of random rotation of 0-90°, random horizontal and vertical flip, and random width and height shift of 0-25%. The number of palynomorphs analyzed for each pollen type in the intact, damaged, and fossil sets is unbalanced. The number of

fossil grains for each type, in particular, is low. As such, we used the average per-class accuracy (APC%) as it is more appropriate for the evaluation of the algorithm performance in this case (Zheng, 2015).

3. Results

3.1. Global comparison of used/unused augmentation and single-CNNs/multi-CNNs

The four different classification setups yield misclassification rates of 0% to 5.9% for intact pollen, 2.8% to 22.5% for damaged pollen, and 3.7% to 45.3% for the fossil pollen. The multi-CNNs including augmentation always gives the best results (Fig. **3**). Compared with the three other setups, the one using augmentation and multi-CNNs increases the success rate for classification in the lowest taxonomic levels (*Achyranthes*-type *aspera*, *Aerva*-type *javanica*, *Aerva*-type *lanata*, *Celosia*-type *argentea*, *Cyathula*-type *orthacantha, Digera*-type *muricata,* Poaceae, and Cyperaceae) and keeps classification rates at 99.3% for intact pollen, 85.7% for damaged pollen, and 80.8% for fossils pollen (Fig. **3**).

In detail, augmentation increases the success rate for classification in lowest levels for simple-CNN and multi-CNNs by +3.1% and +2.7% for intact pollen, +3.1% and +16.3% for damaged pollen, and +10.5% and +43.8% for fossil pollen, respectively. Image augmentation also reduces misclassification rates for simple-CNN and multi-CNNs by -3.1% and -0.8% for intact pollen, -3.1% and -7.3% for damaged, and -10.5% and -12.6% for fossil pollen, respectively (Fig. **3**). Similarly, the use of multi-CNNs reduces misclassification rates, for the scripts without augmentation (-5.1% for intact pollen, -12.4% for damaged, and -29% for fossil pollen), and for those with augmentation (-2.8% for intact pollen, -16.6% for damaged, and -31.1% for fossil pollen) (Fig. **3**).

Multi-CNNs with augmentation also increase the detection of Amaranthaceae and Monocotyledonae misclassification in step 1 and 2. For the damaged dataset, five misclassified pollen grains of

Monocotyledonae were directed into step 3A normally reserved to Amaranthaceae (Supporting Information, Fig. **S2a**). All of these five misclassified pollen grains of Monocotyledonae were identified as Amaranthaceae uncertain in step 3A (Supporting Information, Fig. **S2a**) then classified as pollen indeterminate in final step 4 (Fig. **2**). This is also observed for fossil data in Step 3A and Step 3B (Supporting Information, Fig. **S2c,d**).

3.2 Details of efficiency of Multi-CNNs with augmentation

Our Multi-CNNs with augmentation classified the intact pollen grains with misclassification rates of 0%. All intact pollen are correctly classified (Fig. **3**). We show that within the Amaranthaceae family, six types with similar morphologies can be distinguished. Over 208 images of intact Amaranthaceae pollen, 99.3% were successfully classified within different Amaranthaceae types and only 0.7% (two *Cyathula*-type *orthacantha*) could not be clearly identified but could still be defined as Amaranthaceae indeterminable (Fig. **3** and **4a**). For Poaceae and Cyperaceae, all were correctly classified as Poaceae and Cyperaceae despite the heterogeneity of textures and features, and the complex shape of Cyperaceae grains.

The damaged pollen grains, however, show a higher misclassification rate than intact pollen with 2.8% misclassification (Fig. **3**). Like for intact pollen, the misclassifications are evenly distributed to different classes (Fig. **4b**). 85.7% of the damaged pollen have been successfully classified to the lowest taxonomic levels (*Achyranthes*-type *aspera*, *Aerva*-type *javanica*, *Aerva*-type *lanata*, *Celosia*-type *argentea*, *Cyathula*-type *orthacantha* and *Digera*-type *muricata*, and Poaceae and Cyperaceae) and 6.6% have been successfully classified to higher taxonomic ranks (5.6% in Amaranthaceae indeterminate, and 1% in Monocotyledonae indeterminate), that is -13.6% and +5.9% than intact pollen, respectively. Of the 167 pollen classified as Amaranthaceae, only 6.6% were classified as Amaranthaceae indeterminate, 6.6% were classified as

pollen indeterminate, and none were misclassified as Monocotyledonae indeterminate (Fig. **3** and **4b**).

The fossil grains were classified with 3.7% misclassification, a +0.9% increase compared to damaged pollen (Fig. **3**). In this case, misclassifications are also evenly distributed into different classes (Fig. **4c**). 80.8% of total pollen are successfully classified to the lowest taxonomic levels, compared to -4.9% for damaged pollen. We have one increase of pollen classified to higher taxonomic ranks (+4.8%) compared with damaged pollen. We also show a slight decline of pollen classified as pollen indeterminate with a rate of 4% (Fig. **3**). The presence of mineral inclusions does not seem to interfere with identification (Fig. **1**). Finally, out of the 30 pollen classified as Amaranthaceae, 17% were classified as Amaranthaceae indeterminate.

4. Discussion

Our model has been tested on several datasets with success. The overall computing time required to setup automatic recognition (time required to do all trainings) was about 11 hours using the Nvidia Titan X GPU for the computing. After the training, each test (inference) takes less than a minute on the same machine, and reproducibility tests show slight variations of <2 APC% (not shown). Our study uses image datasets for the "training dataset", "intact dataset", and "damaged dataset" which sizes are in the range of the automatic recognition studies mentioned above, but some may argue that our "fossil dataset" is relatively small. It is, however, a good test for our model because the proportions of Amaranthaceae, Cyperaceae, and Poaceae pollen in the fossil image dataset are close to proportions that are usually observed in modern and fossil pollen assemblages of about 300-400 grains from arid areas of northeastern Africa. The results on the fossil data give a good overview of what can be obtained on real fossil samples. One caveat, not addressed in our study, is that usually, not all pollen types in the fossil/modern sediment or aerial trap are known. As such, the future automatic recognition systems will have to include in

the training dataset pollen types that are not related to any of the analyzed/tested types. We are confident, however, that our system should perform well because in our experiment on the fossil pollen dataset, which had none *Aerva*-type *lanata*, *Celosia*-type *argentea,* and *Cyathula*-type *orthacantha*, we reached significant levels, suggesting that including unknown pollen types should not be a major problem in future applications.

Our study shows that a single CNN is not able to manage the recognition of 8 pollen types when the dataset includes damaged or fossil pollen. With a single CNN, misclassification considerably increases with increased grain deterioration. The origin of this problem is linked to the attempt to classify into low taxonomic ranks highly degraded pollen which have lost most of their surface features and ornamentation. This loss of information complicates the recognition by the CNN when the CNN has been trained to recognize eight pollen types. The new approach we designed and which combines the use of augmentation and multi-CNNs, increases the accuracy of classification for the damaged and fossil pollen. This is principally related to the reclassification of highly degraded pollen into higher taxonomic ranks, and the use of CNNs with augmentation and specialized in the recognition of only one pollen type at a time for the lower taxonomic ranks. The multi-CNNs system is also useful for identifying and self-correcting a significant part of these misclassifications at the high taxonomic level in steps 1 and 2, with the support for the systems set up in steps 3A and 3B. The need for this self-correction is negligible for intact pollen due to the absence of misclassification at the high taxonomic level, but is important for the damaged and fossil pollen which do have misclassification at the high level due to their deteriorated features. Consequently, it is very relevant to use augmentation and a multi-CNNs system for the recognition of fossil or damaged pollen grains.

Our approach based on multi-CNNs with augmentation classifies each class of intact pollen with 0 APC% error. This is as good as in recent studies on intact pollen with a similar number of pollen types (Khanzhina *et al.*, 2018; Sevillano & Aznarte, 2018; Daood, 2018; Daood *et al.*, 2018). For the damaged pollen, we achieved misclassification rate of 2.7 ACP%, which is slightly higher than that of intact pollen (Fig. **1**). To our knowledge, this is the first time that such good accuracy on highly damaged pollen grains has been reached (Holt & Bennett, 2014; Han & Xie, 2018; Khanzhina *et al.*, 2018). Based on all human-made pollen counts of actual soil and mud samples in the African Pollen Database available for Ethiopia (115 samples), we find that indeterminable (i.e. damaged) pollen represent on average about 8% of total pollen count, and that indeterminable Amaranthaceae are most common: they represent about 35% out of total Amaranthaceae ('African Pollen Database', 2019). Our percent of indeterminable pollen out of the total pollen (2.7 APC%) and of indeterminable Amaranthaceae out of the total Amaranthaceae (5.6 APC%) are thus better than the human counts, for modern damaged pollen in the Ethiopian dataset (Bonnefille, 1972; Buchet, 1982; 'African Pollen Database', 2019).

As mentioned before, Punyasena *et al.* (2012) and Kong *et al.* (2016) obtained good results for fossil pollen of *Picea*, but the grains in their study came from a peat; altered pollen in peat cores are far from resembling altered fossil pollen grains from lacustrine Pliocene samples. Our approach is the first to present successful results (only 4.1% misclassification) for fossil pollen with important deterioration, typical for arid areas. In the human-made Pliocene pollen dataset of about 25 samples obtained for the Hadar formation by Bonnefille's team (Bonnefille *et al.*, 1987) indeterminable pollen may account for up to about 30% of total pollen counts (10% in average). Of these 25 samples, 19 have Amaranthaceae of which about 80% on average were classified as "Amaranthaceae indeterminable" (Bonnefille *et al.*, 1987). The percentage of fossil pollen indeterminate that we obtained with our automated system is of 4%, i.e. well within the average of Bonnefille *et al.* (1987), and the taxonomic precision is better as only 17% of Amaranthaceae were classified as indeterminable. Our method, therefore, has

the potential to be applied to both damaged, fossil, and modern pollen grains from arid areas, and may even improve pollen taxa identifications within the Amaranthaceae.

Using our multi-CNNs with augmentation, we were able to classify both morphologically similar pollen (pollen types of Amaranthaceae) at the species or genus level, and morphologically different pollen (Amaranthaceae versus Poaceae versus Cyperaceae), without having significant misclassification for intact, damaged, and fossil pollen (Fig. **4**). For the Monocotyledonae, we can observe an increase in fossil pollen classified as indeterminable Monocotyledonae (+6.8%) in comparison to the damaged pollen, but we do not have significant increases for Amaranthaceae (Fig. **3**). The fossil grains of Monocotyledonae have more mineral inclusions than Amaranthaceae, but we discard the hypothesis that mineral inclusions in pollen is the origin of these increases because fossil pollen of Monocotyledonae bearing mineral inclusions are correctly identified (Fig. **1**). We see on the images that fossil pollen of Monocotyledonae is often "stuck" in residual sediments (Fig. **1**). This is very similar to images of "contaminated grains" of Han and Xie (2018), who suggested that this hinders identification. Our study shows that our system overcomes the sediment contamination and is able to correctly identify fossil pollen of Amaranthaceae even when sediment is attached to the grains. The deformation of pollen grains is another hindrance to identification (Holt *et al.*, 2011; Han & Xie, 2018; Khanzhina *et al.*, 2018) but our fossil grains of Monocotyledonae are no more deformed than damaged grains of Monocotyledonae or fossil grains of Amaranthaceae (Fig. **1**). We posit that, within the biodiversity range we investigated, this issue stems from the structural weakness of Monocotyledonae (in the pollen studied) as this family shows a thinner and weaker exine than Amaranthaceae. As for human eyes, automated pollen recognition requires that the exine texture must be relatively well preserved. This agrees with the idea of Li *et al.* (2004) that the "surface texture is frequently characteristic, even when grains are damaged or fragmented" for the pollen.

Our multi-CNNs with augmentation is operational. At CEREGE, it is used occasionally to assist with identification of fossil Amaranthaceae when identification to type is not straightforward. The main bottleneck to an extended use of our recognition system is the acquisition time, which is manual and therefore, time-consuming, preventing an easy addition of new pollen types. Adding an automated acquisition step would therefore provide a complete automatic pollen counting system, saving a lot of time for palynologists of arid areas and possibly elsewhere. With such a system, it would probably be easy to acquire images of large pollen reference collections within in a few months. The prospect of a fully automated pollen counting system for paleoclimatic and paleoenvironmental reconstructions, similar to the operating systems developed for other microfossils like coccoliths or foraminifera (Beaufort & Dollfus, 2004; Marchant et al., in preparation), can be foreseen in the very near future. These automated systems will pave the way to high-throughput palynology studies, including morphometric studies.

Conclusions

In this study, we proposed a robust system for classification of very damaged and badly preserved fossil grains. We have developed a classification topology that has flexibility in the choice of taxonomic levels used for classification. It can, like a palynologist, assign higher taxonomic ranks if the characteristics of the image are not sufficient for robust identification at a lower taxonomic level. Moreover, if images cannot be properly classified, it can declare them as indeterminable. Our system has the advantage of using simple LM images and does not require the implementation of complex and expensive systems such as confocal fluorescence microscopes. With the damaged pollen, we have obtained results similar to or better than those obtained from human pollen counting performed by CEREGE palynologists on modern soils and

Pliocene sediments of the Awash Valley in Ethiopia.

The system developed here does not yet allow a complete count of fossil pollen. It remains limited to a few families and to a very limited number of taxa in comparison to overall pollen diversity. Evaluating and improving how the system reacts to pollen types or other grains not included in the training set is also important because usually not all pollen types in the sediment core are known. To handle this, one solution may be the addition of an "unknown" category. Our results, based on a limited dataset but representing a wide range of biases affecting pollen grains from their production to their fossilization, allow the possible development of automatic pollen recognition for routine palynological studies of samples coming from multiple depositional environments and with different degrees of pollen alteration. It would also be interesting to see if it is possible to increase the accuracy of our system with the use of a much larger number of training images.

It is also important to specify that the system presented here cannot be directly extended to other pollen families as it relies more on the morphology of the pollen types than on the taxonomic relationship, although the two are related. It is also important to have a good knowledge of how the morphology of each type of pollen changes with degradation or fossilization in order to build relevant mappings and logical steps in the classification system to cover all stages of pollen grain degradation. The expertise of paleo-palynology is therefore critical to the extension of this system.

The next step of this work will be to extend the system to automatically take pictures of pollen grains from microscope slides. The automatic acquisition will be inspired by the processing software developed for coccoliths at CEREGE (Beaufort & Dollfus, 2004) and modified by Martin Tetard and Yves Gally for Radiolaria tests (Martin Tetard, personal communication). It can be summarized in 6 steps: (1) color pollen grains with Safranin for easier segmentation, (2) take grayscale images of microscope fields of view (FOVs) by incrementally stepping through the slide (1 µm step), (3) segment each pollen grains in each stacked FOV (saving their x and y coordinates on the original unstacked FOV), (4) generate sub-stacks for each previously recorded pollen grain from each original unstacked FOV, (5) remove the unfocused pollen images from each pollen sub-stack using a grey level threshold, and (6) apply the method in the second paragraph of part 2.1 for finalizing image processing prior to automated recognition. With this automatic acquisition, it will be possible to fully scan reference pollen collections and thus quickly build a large database of pollen images to develop an automatic pollen counting system (acquisition and recognition) for modern and fossil pollen from eastern Africa, for example.

Computer code availability

All files and folders presented here are downloadable at https://github.com/Beniofh/Multi-CNN. We used the script Multi-CNN_test_V3.5.2.py (40.9 Ko) to recognize pollen types using multi-CNNs with augmentation. Multi-CNNs with augmentation is compared to single-CNN with augmentation (script CNN_test_v2.1.0.py, 16.5 Ko), single-CNN without augmentation (CNN_test_v2.1.0

(wihout_augm).py, 16.5 Ko), and multi-CNNs without augmentation (Multi-CNN_test_V3.5.2(wihout_augm).py, 40.9 Ko). In the CNN_solo folder, we have for each CNN model parameters, a folder with file of CNN model parameter used in .ckpt, histograms of accuracy, loss and learning rate during the training in .bmp and table of accuracy, loss and learning rate during the training in .csv. The folder "intact", "damaged" and "fossil" contain respectively the images for "training dataset"/"intact pollen dataset", "damaged dataset", and "fossil dataset". The User_manual.pdf explains how to use these scripts to reproduce our results.

These scripts have been developed by Benjamin BOUREL (chebenjamin@laposte.net ; +33 6 63 49 85 24) with the help of Ross MARCHANT. These scripts were first available the 13th October 2019. They were created in the open-source Anaconda 3 v.5.2.0 (https://www.anaconda.com) with Spyder v.3.3.6 in python v.3.6 (Pérez et al.,2011). We also used NumPy (Oliphant, 2006), Scikit-image (van der Walt et al., 2014), Scikit-learn (Pedregosa et al. 2011), and Matplotlib (Hunter, 2007) packages. The details of the python libraries used (name and version) are indicated in User_manual.pdf. All software and python libraries used here are open sources. For optimal use, please use the indicated versions.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, *et al.* 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:* 1603.04467.

[Database] African Plant Database. 2019. *African Plant Database (version 3.4.0).* htpp://ville-ge.ch/musinfo/bd/cjb/africa/. (accessed 25 March 2019).

[Database] African Pollen Database. 2019. *African Pollen Database.* http://apd.sedoo.fr/accueil.htm. (accessed 06 May 2019).

Beaufort L, Dollfus D. 2004. Automatic recognition of coccoliths by dynamical neural networks. *Marine Micropaleontology* 51: 57–73.

Bonnefille R. 1972*. Associations polliniques actuelles et quaternaires en Ethiopie (vallées de l'Awash et de l'Omo).* PhD Dissertation, Paris VI University, Paris, France.

Bonnefille R. 1979. Methode palynologique et reconstitutions paleoclimatiques au Cenozoique dans le Rift est-africain. *Bulletin de la Société Géologique de France* S7-XXI: 331–342.

Bonnefille R. 2010. Cenozoic vegetation, climate changes and hominid evolution in tropical Africa. *Global and Planetary Change* 72: 390–411.

Bonnefille R, Potts R, Chalié F, Jolly D, Peyron O. 2004. High-resolution vegetation and climate change associated with Pliocene Australopithecus afarensis. *Proceedings of the National Academy of Sciences* 101: 12125–12129.

Bonnefille R, Vincens A, Buchet G. 1987. Palynology, stratigraphy and palaeoenvironment of a pliocene hominid site (2.9-3.3 M.Y.) at Hadar, Ethiopia. *Palaeogeography, Palaeoclimatology, Palaeoecology* 60: 249–281.

Buchet G. 1982. *Transport des pollens dans les fleuves Omo et Awash (Ethiopie): étude des vases actuelles.* PhD Dissertation, École pratique des hautes études, Montpellier, France.

Chollet F. 2015. *Keras.* https://github.com/fchollet/keras. (accessed 20 May 2019).

Chudyk C, Castaneda H, Léger R, Yahiaoui I, Boochs F. 2015. Development of an Automatic Pollen Classification System Using Shape, Texture and Aperture Features. In: Proceedings of the LWA 2015 Workshops: KDML FGWM IR and FGDB Trier. Trier, Germany, 65–74.

Cohen A, Campisano C, Arrowsmith R, Asrat A, Behrensmeyer AK, Deino A, Feibel C, Hill A, Johnson R, Kingston J, *et al.* 2016. The Hominin Sites and Paleolakes Drilling Project: inferring the environmental context of human evolution from eastern African rift lake deposits. *Scientific Drilling* 21: 1–16.

Curry BB. 2007. The late-glacial and early Holocene geology, paleoecology and paleohydrology of the Brewster Creek Site, a proposed wetland restoration site, Pratt's Wayne Woods Forest Preserve and James 'Pate' Philip State Park, Bartlett, Illinois. *Circular 571. Illinois Department of Natural Resources. Illinois State Geological Survey.*

Daood A. 2018. *Pollen Grains Recognition Based on Computer Vision Methods.* PhD Dissertation, Florida Institute of Technology, Melbourne, Florida, USA.

Daood A, Ribeiro E, Bush M. 2018. Sequential Recognition of Pollen Grain Z-Stacks by Combining CNN and RNN. In: The Thirty-First International Florida Artificial Intelligence Research Society Conference. Melbourne, Florida, USA, 8–13.

Fernandez-Delgado M, Carrion P, Cernadas E, Galvez JF. 2003. Improved Classification of Pollen Texture Images Using Svm and Mlp. In: 3rd IASTED International Conference on Visualization, Imaging and Image Processing. Benalmadena, Spain.

Flenley JR. 1968. *The problem of pollen recognition*. Canberra, Australia: CSIRO.

García NM, Chaves VAE, Briceño JC, Travieso CM. 2012. Pollen Grains Contour Analysis on Verification Approach. In: International Conference on Hybrid Artificial Intelligence Systems. Salamanca, Spain, 521–532.

Gonçalves AB, Souza JS, Silva GG da, Cereda MP, Pott A, Naka MH, Pistori H. 2016. Feature Extraction and Machine Learning for the Classification of Brazilian Savannah Pollen Grains. *PLOS ONE* 11: e0157044.

Han L, Xie Y. 2018. Local Decimal Pattern for Pollen Image Recognition. In: Artificial Neural Networks and Machine Learning – ICANN 2018. Rhodes, Greece, 47–55.

Holt K, Allen G, Hodgson R, Marsland S, Flenley J. 2011. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology* 167: 175–183.

Holt KA, Bennett KD. 2014. Principles and methods for automated palynology. *New Phytologist* 203: 735–742.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, 9, 90-95.

Kaya Y, Pınar SM, Erez ME, Fidan M. 2013. An expert classification system of pollen of Onopordum using a rough set approach. *Review of Palaeobotany and Palynology* 189: 50– 56.

Khanzhina N, Putin E, Filchenkov A, Zamyatina E. 2018. Pollen grain recognition using convolutional neural network. In: Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018). Bruges, Belgium, 409–414.

Kong S, Punyasena S, Fowlkes C. 2016. Spatially Aware Dictionary Learning and Coding for Fossil Pollen Identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, Nevada, USA, 1– 10.

Li P, Treloar WJ, Flenley JR, Empson L. 2004. Towards automation of palynology 2: the use of texture measures and neural network analysis for automated identification of optical images of pollen grains. *Journal of Quaternary Science* 19: 755–762.

Mander L, Baker SJ, Belcher CM, Haselhorst DS, Rodriguez J, Thorn JL, Tiwari S, Urrego DH, Wesseln CJ, Punyasena SW. 2014. Accuracy and consistency of grass pollen identification by human analysts using electron micrographs of surface ornamentation. *Applications in Plant Sciences*, 2(8), 1400031.

Marchant R, Tetard M, de Garidel-Thoron T. In preparation. A system for classification of foraminifera images using deep convolutional neural networks.

Marcos JV, Nava R, Cristóbal G, Redondo R, Escalante-Ramírez B, Bueno G, Déniz Ó, González-Porto A, Pardo C, Chung F, *et al.* 2015. Automated pollen identification using microscopic imaging and texture analysis. *Micron* 68: 36–46.

Maslin MA, Shultz S, Trauth MH. 2015. A synthesis of the theories and concepts of early human evolution. *Philosophical Transactions of the Royal Society B:*

Biological Sciences 370: 20140064–20140064.

Mohan J, Stone JR, Campisano CJ. 2016. Three novel species of Bacillariophyta (Diatoms) belonging to Aulacoseira and Lindavia from the Pliocene Hadar Formation, Afar Depression of Ethiopia. *Phytotaxa* 272: 235–247.

Oliphant TE. 2006. *A guide to NumPy*, USA: Trelgol Publishing.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825-2830.

Pérez F, Granger BE, Hunter JD. 2011. Python: An Ecosystem for Scientific Computing. *Comput. Sci. Eng.* 13, 13–21.

Punyasena SW, Tcheng DK, Wesseln C, Mueller PG. 2012. Classifying black and white spruce pollen using layered machine learning. *New Phytologist* 196: 937–944.

Rodriguez-Damian M, Cernadas E, Formella A, Fernandez-Delgado M, Sa-Otero PD. 2006. Automatic detection and classification of grains of pollen based on shape and texture. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 36: 531–542.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, *et al.* 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9: 676–682.

Sangster, AG, Dale HM. 1964. Pollen grain preservation of underrepresented

species in fossil spectra. *Canadian Journal of Botany*, *42*(4): 437-449.

Sevillano V, Aznarte JL. 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.

Simonyan K, Zisserman A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:* 1409.1556.

Sobol MK, Finkelstein SA. 2018. Predictive pollen-based biome modeling using machine learning. *PLOS ONE* 13: e0202214.

Treloar WJ, Taylor GE, Flenley JR. 2004. Towards automation of palynology 1: analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning electron microscope images. *Journal of Quaternary Science* 19: 745–754.

Villmoare B, Kimbel WH, Seyoum C, Campisano CJ, DiMaggio EN, Rowan J, Braun DR, Arrowsmith JR, Reed KE. 2015. Early Homo at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia. *Science* 347: 1352–1355.

van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, the scikit-image contributors. 2006. Scikit-image: Image processing in Python, *PeerJ 2*: e453.

Zheng A. 2015. *Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls.* Sebastopol, Ukrainian/ Russia: O'Reilly Media, 2015. [online] Available: http://www.oreilly.com/data/free/files/ evaluating-machine-learning-models.pdf.

Captions

Figure 1: Pollen images with automatic z-stack of 128x128 pixels for all pollen types of the intact, damaged, and fossil pollen datasets. The name under each image is the final label assigned by the multiple convolutional neural networks (multi-CNNs) with augmentation.

Figure 2: Schematic of the classification system based on multiple convolutional neural networks (multi-CNNs) with augmentation. aa: *Achyranthes*-type *aspera*; aj: *Aerva*-type *javanica*; al: *Aerva*-type *lanata*; amar: Amaranthaceae; c: Cyperaceae; ca: *Celosia*-type *argentea*; co: *Cyathula*-type *orthacantha*; dm: *Digera*-type *muricata*; mono: Monocotyledonae; p: Poaceae; uncert: uncertain; indet: indeterminable.

Figure 3: Results of our classification system (multi-CNNs with augmentation) as a function of pollen dataset and in comparison with other setups. Pollen types of the lowest taxonomic levels: *Achyranthes*-type *aspera*, *Aerva*-type *javanica*, *Aerva*-type *lanata*, *Celosia*-type *argentea*, *Cyathula*-type *orthacantha* and *Digera*-type *muricata*, and Poaceae and Cyperaceae. indet.: indeterminable. The APC% is the percentages based on average per-class.

Figure 4: Confusion matrix for the intact (a), damaged (b) and fossil (c) pollen dataset for final classes. aa: *Achyranthes*-type *aspera*; aj: *Aerva*-type *javanica*; al: *Aerva*-type *lanata*; amar: Amaranthaceae; c: Cyperaceae; ca: *Celosia argentea*; co: *Cyathula*-type *orthacantha*; dm: *Digera*-type *muricata*; mono: Monocotyledonae; p: Poaceae; indet: indeterminable.

Supporting Information captions

Figure S1: Effect of transfer learning (the use of pretrained CNN) on the training with augmentation of CNN. Dataset: training pollen dataset; mapping: *Achyranthes*-type *aspera*, *Aerva*-type *javanica*, *Aerva*-type *lanata*, *Celosia*-type *argentea*, *Cyathula*-type *orthacantha*, *Digera*-type *muricata*, Cyperaceae, and Monocotyledonae.

Figure S2: Confusion matrix for the damaged pollen dataset (classes of step 2A (a) and classes of step 2B (b)) and the fossil pollen dataset (classes of step 2A (c) and classes of step 2B (d)). aa: *Achyranthes*-type *aspera*; aj: *Aerva*-type *javanica*; al: *Aerva*-type *lanata*; amar: Amaranthaceae; c: Cyperaceae; ca: *Celosia*-type *argentea*; co: *Cyathula*-type *orthacantha*; dm: *Digera*-type *muricata*; mono: Monocotyledonae; p: Poaceae; indet: indeterminable.

Table S1: Origin of pollen grains and distribution in the datasets. ka: thousand years.

Table S2: Excel file with detailed calculations for percentages of average per-class accuracy (APC%).
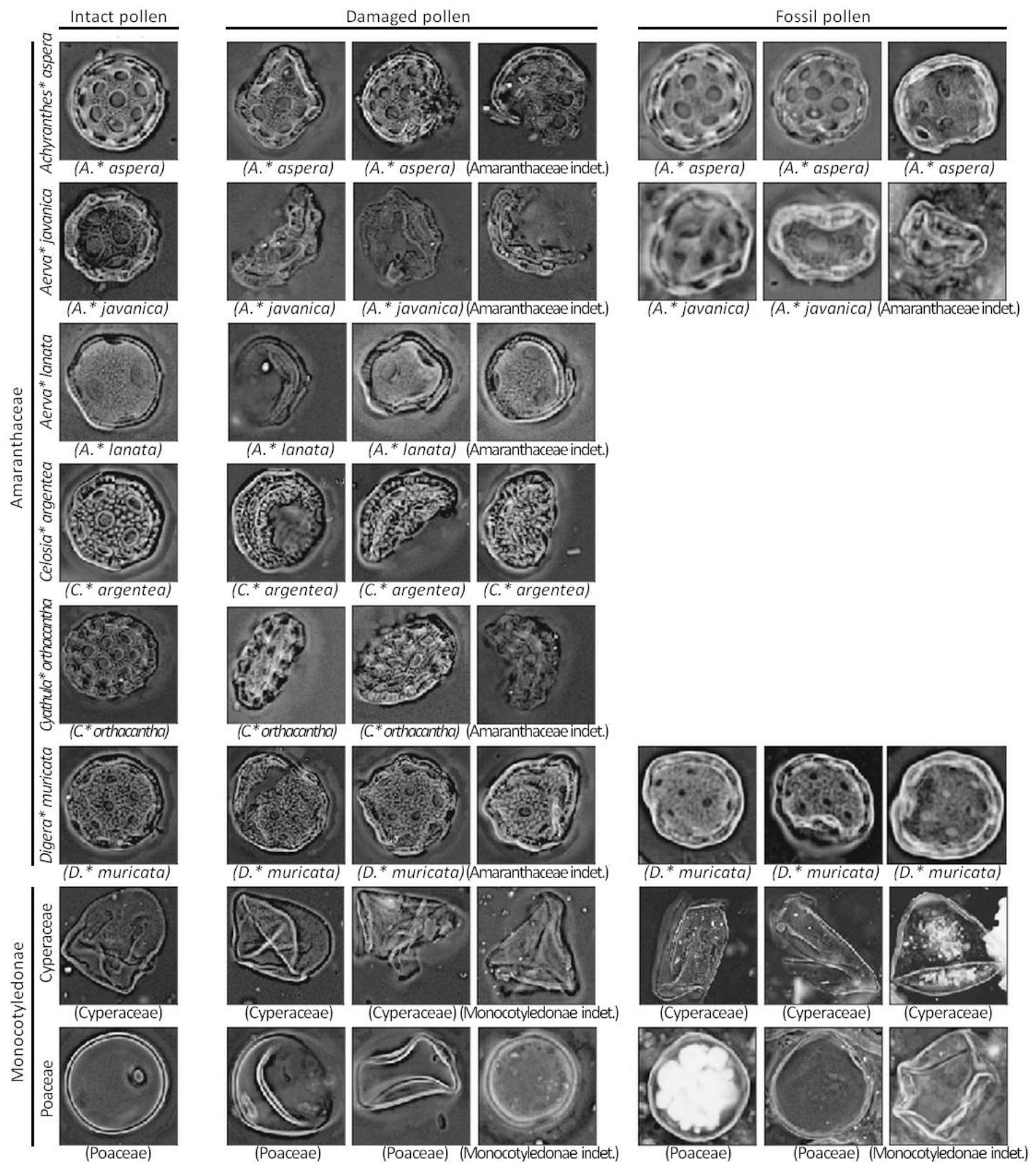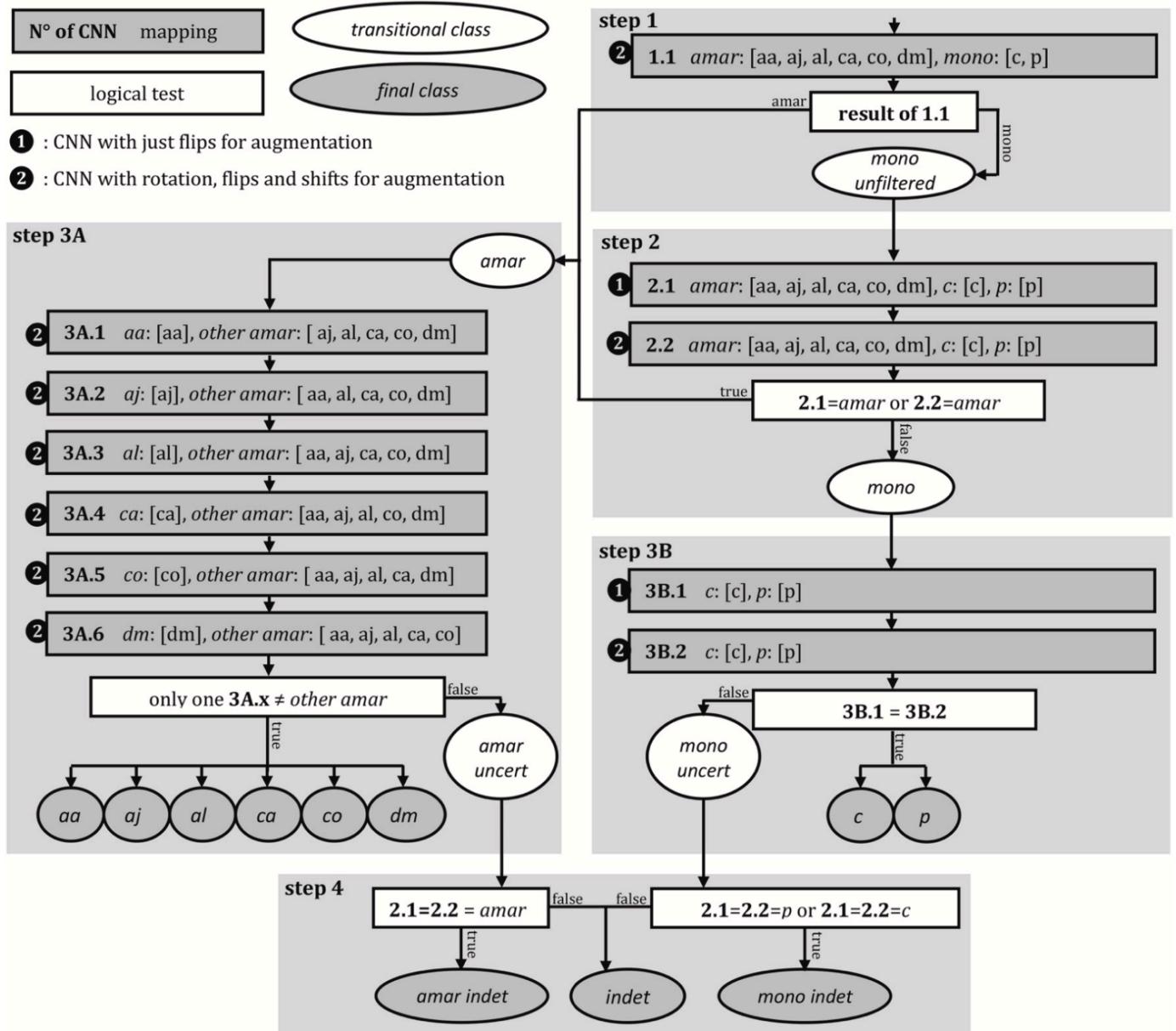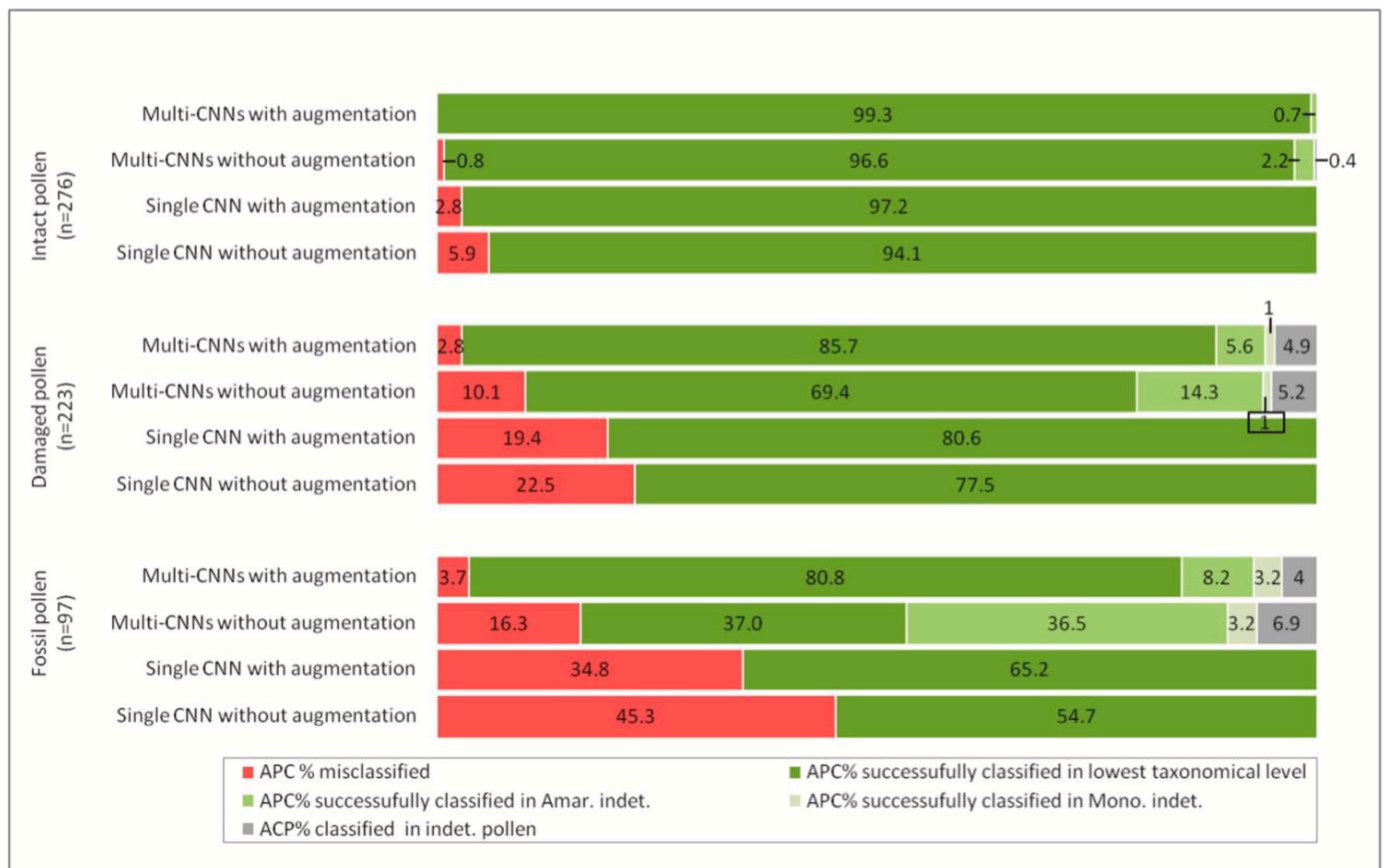
Figure 1

	Intact pollen	Damaged pollen	Fossil pollen

Amaranthaceae

Achyranthes aspera*: (A. * aspera) | (A. * aspera) (A. * aspera) (Amaranthaceae indet.) | (A. * aspera) (A. * aspera) (A. * aspera)

Aerva javanica*: (A. * javanica) | (A. * javanica) (A. * javanica) (Amaranthaceae indet.) | (A. * javanica) (A. * javanica) (Amaranthaceae indet.)

Aerva lanata*: (A. * lanata) | (A. * lanata) (A. * lanata) (Amaranthaceae indet.)

Celosia argentea*: (C. * argentea) | (C. * argentea) (C. * argentea) (C. * argentea)

Cyathula orthacantha*: (C.* orthacantha) | (C.* orthacantha) (C.* orthacantha) (Amaranthaceae indet.)

Digera muricata*: (D. * muricata) | (D. * muricata) (D. * muricata) (Amaranthaceae indet.) | (D. * muricata) (D. * muricata) (D. * muricata)

Monocotyledonae

Cyperaceae: (Cyperaceae) | (Cyperaceae) (Cyperaceae) (Monocotyledonae indet.) | (Cyperaceae) (Cyperaceae) (Cyperaceae)

Poaceae: (Poaceae) | (Poaceae) (Poaceae) (Monocotyledonae indet.) | (Poaceae) (Poaceae) (Monocotyledonae indet.)

Figure 2

Figure 3

Figure 4

a. Intact pollen dataset

b. Damaged pollen dataset

c. Fossil pollen dataset