



HAL
open science

A user-friendly method to get automated pollen analysis from environmental samples

Betty Gimenez, Sébastien Joannin, Jérôme Pasquet, Luc Beaufort, Yves Gally, Thibault de Garidel-Thoron, Nathalie Combourieu-Nebout, Laurent Bouby, Sandrine Canal, Sarah Ivorra, et al.

► To cite this version:

Betty Gimenez, Sébastien Joannin, Jérôme Pasquet, Luc Beaufort, Yves Gally, et al.. A user-friendly method to get automated pollen analysis from environmental samples. *New Phytologist*, 2024, 10.1111/nph.19857 . hal-04593322

HAL Id: hal-04593322

<https://hal.science/hal-04593322>

Submitted on 31 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.












L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Methods

A user-friendly method to get automated pollen analysis from environmental samples

Betty Gimenez¹ , Sébastien Joannin^{1,2} , Jérôme Pasquet^{3,4} , Luc Beaufort⁵ , Yves Gally⁵,
Thibault de Garidel-Thoron⁵ , Nathalie Combourieu-Nebout⁶ , Laurent Bouby¹ , Sandrine Canal¹,
Sarah Ivorra¹ , Bertrand Limier^{1,7}, Jean-Frédéric Terral¹ , Céline Devaux^{1,8*}  and Odile Peyron^{1*} 

¹ISEM, Univ Montpellier, CNRS, IRD, 34090, Montpellier, France; ²School of Earth, Environment & Society, McMaster University, L8S 4K1, Hamilton, ON, Canada; ³AMIS, Univ Paul-Valérie Montpellier 3, 34090, Montpellier, France; ⁴TETIS, INRAE, AgroParisTech, Cirad, CNRS, Univ Montpellier, 34090, Montpellier, France; ⁵CEREGE, Aix Marseille Université, CNRS, IRD, Coll. France, INRAE, 13545, Aix-en-Provence, France; ⁶UMR 7194 CNRS, MNHN, HNHP, Institut de Paléontologie Humaine, 75013, Paris, France; ⁷INRAE, Centre Occitanie-Montpellier, 34000, Montpellier, France; ⁸Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, H1X 2B2, Montreal, QC, Canada

Summary

Author for correspondence:
Betty Gimenez
Email: gimenez.betty@proton.me

Received: 22 February 2024
Accepted: 5 May 2024

New Phytologist (2024)
doi: 10.1111/nph.19857

Key words: artificial intelligence, automated pollen analysis, deep learning, detection errors, environmental real-world samples, guidelines, Mediterranean vegetation monitoring, YOLOv5.

- Automated pollen analysis is not yet efficient on environmental samples containing many pollen taxa and debris, which are typical in most pollen-based studies. Contrary to classification, detection remains overlooked although it is the first step from which errors can propagate. Here, we investigated a simple but efficient method to automate pollen detection for environmental samples, optimizing workload and performance.
- We applied the YOLOv5 algorithm on samples containing debris and c. 40 Mediterranean plant taxa, designed and tested several strategies for annotation, and analyzed variation in detection errors.
- About 5% of pollen grains were left undetected, while 5% of debris were falsely detected as pollen. Undetected pollen was mainly in poor-quality images, or of rare and irregular morphology. Pollen detection remained effective when applied to samples never seen by the algorithm, and was not improved by spending time to provide taxonomic details. Pollen detection of a single model taxon reduced annotation workload, but was only efficient for morphologically differentiated taxa.
- We offer guidelines to plant scientists to analyze automatically any pollen sample, providing sound criteria to apply for detection while using common and user-friendly tools. Our method contributes to enhance the efficiency and replicability of pollen-based studies.

Introduction

Pollen is a major tool for ecological, paleo-environmental, and evolutionary studies, used to monitor plant responses to environmental changes (van der Knaap *et al.*, 2010), inform on plant–pollinator interactions (Morente-López *et al.*, 2018), reconstruct past vegetation and climate (Peyron *et al.*, 2017), anticipate allergology (Anderegg *et al.*, 2021), infer honey origin (Corvucci *et al.*, 2015), or predict harvests (Oteros *et al.*, 2014). Counting and identifying pollen is traditionally performed manually by experts, using a slide under light microscopy. These tasks are time consuming, and limit the size and replicability of pollen studies. Automation of pollen analysis, a long-standing objective in palynology, can help open new research avenues by extending spatial

and temporal resolution of pollen studies, and help obtain standardized data comparable among years, sites, and research teams (Stillman & Flenley, 1996; Holt & Bennett, 2014).

The rapid development of convolutional neural networks (CNNs) for image analysis now makes routine automated pollen analysis achievable, as demonstrated in recent studies (e.g. Khanzhina *et al.*, 2022; Punyasena *et al.*, 2022; Barnes *et al.*, 2023). When applied on slides scanned under light microscopy, automation of pollen analyses relies on (1) the detection of pollen grains, that is finding their position in an image and (2) the classification of the detected objects into predefined classes, for example pollen taxon (Diwan *et al.*, 2022). Classification is usually performed separately from detection, on images containing a single pollen grain (Olsson *et al.*, 2021; Punyasena *et al.*, 2022; Viertel & Koenig, 2022). In contrast to classification, interest, and progress for automated pollen detection remains limited, yet detection is a

*These authors contributed equally to this work.

crucial first step and a challenging task, especially when applied to images containing many pollen and nonpollen objects of many kinds, as in environmental samples. The CNNs-based object-detection algorithms, notably Fast-RCNN (Ren *et al.*, 2015), RetinaNet (Lin *et al.*, 2017), or YOLO (Redmon *et al.*, 2016), and one of its latest versions YOLOv5 (Jocher, 2020), can now perform detection jointly with classification at high speed and accuracy, on almost any object, for example molds in fruits (Jubayer *et al.*, 2021), arctic benthic fauna (Marini *et al.*, 2022), *Plasmodium falciparum* (Zedda *et al.*, 2022), and even on complex images with dense and heterogeneous backgrounds (Diwan *et al.*, 2022; Jiang *et al.*, 2022). In palynology, the implementation of these algorithms is recent (Gallardo-Caballero *et al.*, 2019), and remains mainly limited to allergology, on images containing few pollen taxa and a uniform background, or on airborne samples that contain fresh and well-preserved pollen with few debris (Gallardo-Caballero *et al.*, 2019; Khanzhina *et al.*, 2022; Kubera *et al.*, 2022). Under such conditions, studies report excellent results, for example (1) for 11 pollen types, the combination of Faster R-CNN and RetinaNet correctly detected 98.54% of the annotated pollen grains, and wrongly detected as little as 0.25% of nonpollen objects (Gallardo-Caballero *et al.*, 2019); (2) for 13 allergenic pollen species, a modified RetinaNet ('BayesianRetinaNet network') correctly detected 96.32% of pollen grains, and also correctly classified 97.66% of them (classification F1-score; Khanzhina *et al.*, 2022), and (3) for three Betulaceae pollen taxa, YOLOv5 correctly detected and classified 89.7–98.9% of the pollen grains, and 91.7–97.8% of the predictions were correct (Kubera *et al.*, 2022).

These results however do not apply for pollen-based research typically relying on environmental samples from the 'real-world', such as gravimetric pollen traps, moss pollsters, and sediment records, which can contain many debris and damaged pollen grains, with an uncontrolled and potentially high diversity of pollen taxa. For these environmental samples, the methods for the automation of pollen analyses are still in development. The first two attempts to automate pollen analysis on such samples recently achieved promising results: (1) for gravimetric traps placed in a tropical forest, 83.7% of the pollen grains were correctly detected, and 89.5% of detected grains were correctly classified into 25 selected pollen taxa (Punyasena *et al.*, 2022); (2) for pollen samples from lake sediments, from 87.2% to 99.1% of pollen grains were correctly detected, 84% were correctly classified into 11 selected pollen taxa, and 7% were incorrect classifications (Theuerkauf *et al.*, 2023). These results pave the way for new investigations, especially to understand the impact of detection errors on the accuracy of automated pollen analysis, because any error at this step will inevitably propagate, for example to the classification step.

In this study, we aim to (1) improve the process of pollen detection in environmental samples containing large amounts of debris and pollen taxa, by analyzing the variation in detection errors, (2) find general implementation guidelines applicable to any pollen study, and (3) evaluate the joint detection and classification errors in a full automated analysis. We do so using gravimetric pollen trap samples collected annually in the

Mediterranean area. As we aim to make automated pollen detection accessible to nonexperts of pollen or deep learning, we rely on simple and common tools: mounted slides scanned under light microscopy, and the open-source and user-friendly algorithm YOLOv5. We search for the best annotating strategies that balance workload and performance for studies of a single taxon or an assemblage of taxa, and for studies further extended in time or space, for example as in long-term plant monitoring. We also study in detail the causes of the detection errors, using the information on pollen morphology and the image quality. We finally assess automated and joint detection and classification on five pollen taxa common in our dataset, and also compare automated results to the ones made by an expert palynologist.

Materials and Methods

Pollen samples and image acquisition

We used pollen samples from gravimetric traps collected in 2019, 2020, and 2021 for a project monitoring vegetation in six locations in a Mediterranean massif. The pollen traps consisted of containers with a 5 cm width opening, and a 5 mm mesh containing glycerin and thyme essential oil to retain and preserve pollen grains, and avoid fungus growth. Traps were placed on the ground or attached to a tree in early January, and collected 1 yr later. To calibrate pollen counts, tablets of *Lycopodium* marker spores (*Lycopodium clavatum* L.) were added to the pollen samples (Stockmarr, 1971). The samples were then chemically treated to remove calcium carbonates and silicates, and were acetolyzed for 6–8 min. For image acquisition, one fixed slide per sample was mounted with glycerin jelly, under 16 mm × 16 mm cover slides. Two samples were discarded due to an insufficient amount of pollen. We therefore worked with 16 slides mounted from 16 samples.

Microscopic images of each slide were acquired with an automated bright light microscope Leica DM6 B TL BF (Wetzlar, Germany) (×63 magnification under oil immersion). The imaging was done by a Hamamatsu ORCA FLASH camera (Hamamatsu Photonics K.K., Hamamatsu, Japan) with a 2048 × 2048 pixel camera sensor. We used the image acquisition pipeline developed by Tetard *et al.* (2020), including a LabVIEW interface. We scanned only 17% of each slide to decrease the acquisition time of images, but covered the entire and potentially heterogeneous distributions of pollen within each slide by acquiring images in 16 squared areas, arranged in a 4 × 4 grid; each scanned area consisted of 64 (8 × 8) fields of views (FOV) of 214 × 214 μm with an overlap of 10 μm (smaller than the smallest pollen taxon here). For each FOV, 11 images were taken along the z-axis, spaced 8 μm apart, to produce a stack capturing the vertical details of the pollen grains (Fig. 1 step 1). The depth resolution was a compromise between acquisition time and the need for details of pollen grains, the sizes of which range from c. 12 to c. 150 μm. For each of the 16 slides, piles of images were automatically taken for 1024 FOV in c. 2 h 30 min. Each pile of images was then stacked using Helicon Focus 7 (Tetard *et al.*, 2020), which selects the sharpest areas and discards the unfocused

areas to create a single final composite 2D image, hereafter called FOV (Fig. 1 step 1). The FOVs from four out of the 16 scanned areas, from the top-left/bottom-right diagonal of each slide, were selected to train and evaluate the YOLOv5 algorithm, totaling 4098 FOVs for the full dataset (16 slides \times 4 scanned areas \times 64 FOVs, with two exceptions). The FOVs from the other 12 scanned areas were never used during training, but used to generate the automated pollen counts compared with those obtained by an expert palynologist, on distinct slides mounted from the same samples.

Optimization of the object-detection algorithm for detecting pollen

We selected the light version of the algorithm YOLOv5 among its several releases (Jocher, 2020) as it performs as well as the heavy version on pollen (Kubera *et al.*, 2022). More details on YOLOv5 are available in the Supporting Information Methods S1. The algorithm YOLOv5 performs detection and classification in a single step, by (1) predicting bounding boxes (defined by their width, length, and location within an image) around the detected objects, and jointly (2) predicting a label, among predefined classes, for the objects within the bounding boxes. Confidence scores are also provided for all predictions. The image dataset used to train the algorithm had thus to be manually annotated by (1) tagging the targeted pollen and *Lycopodium* grains, that is placing bounding boxes that frame them, and (2) labeling these grains, for example as a pollen taxon. We used the LABELIMG software (Lin, 2015), and annotated a total of 12 531 pollen and *Lycopodium* grains in the 4098 FOVs of the dataset (Fig. 1 step 2). We split this annotated dataset into 60% for training, 20% for validation, and 20% for testing, unless otherwise mentioned, with an equal contribution of all 16 slides in each subset. We performed a fivefold cross-validation, by interchanging the FOVs from the training, validation, and test datasets for each analysis (Fig. 1 step 3). Models were trained for 150 epochs, that is training iterations, and on images resized to 640 pixels, which took < 3 h per training using Jean-Zay Nvidia V100 GPU (IDRIS, CNRS). The number of epochs was chosen empirically to enhance model performance without inducing overfitting. To evaluate the performance, we applied the model saved after the last training epoch on the test datasets made of annotated FOVs never seen before. We eliminated the predicted bounding boxes that overlapped using an Intersection over Union (IoU) threshold of 0.7 (Methods S2), and that had a confidence score below 0.45 (Methods S3). To evaluate the performance of the sole detection, we then compared the manually annotated and the predicted bounding boxes, without taking into account the labels predicted by the algorithm through its joint classification (Methods S4). We used an IoU of 0.5 between annotated and predicted bounding boxes to determine (1) a true positive (TP) for a bounding box both predicted and manually annotated, for example a pollen grain correctly detected, regardless of its classification (2) a false negative (FN) for a bounding box manually annotated but not predicted, for example a pollen grain not

detected, and (3) a false positive (FP) for a bounding box predicted but not manually annotated, for example a debris falsely detected as pollen. Combinations of these statistics produced (1) the recall, percentage of correctly detected grains among all true grains, (2) the precision, percentage of correctly detected grains among all detected objects, (3) F1-score, the harmonic mean of recall and precision, and (4) the receiver operating characteristics (ROC) curves, which all inform on the power of the models to discriminate pollen grains from the background; see details in Methods S5.

To evaluate the adequacy of the size of our dataset (4098 FOVs including 12 531 annotations), we trained the models on an increasing number of FOVs. We split each annotated dataset into 80% for training, and 20% for validation, with five cross-validations, but we systematically tested the trained models on the same annotated dataset (Fig. 1 step 3b). For these tests, we used the models from both the last training epoch, and from the epoch providing the best performance on validation. F1-scores showed that the best and last epoch models provided similar performances; we therefore chose to use only the last epoch models for all below analyses (Fig. S1c; Notes S1).

Optimization of the annotation strategies tailored for distinct pollen studies

We evaluated the performance of the three following annotation strategies, *a priori* common in any pollen-based studies, to provide useful and efficient guidelines on how to balance performance and workload (Fig. 1 step 2; Methods S6):

- all-0taxon considered the simplest annotation strategy consisting of only three labels associated with the tagged bounding boxes, with no information on pollen taxon: *Lycopodium*, pollen of any taxon, and grains of either pollen or *Lycopodium* cut on the FOVs edge with less than a quarter of their surface visible (Fig. 1 step 2a);
- all-5taxa used the same above dataset but divided the aforementioned pollen category into six labels: Pinaceae (Pinaceae sp.), *Buxus* (*Buxus sempervirens* L.), Poaceae (Poaceae sp.), *Quercus* (*Quercus* sp.), Oleaceae (Oleaceae sp.), and the label pollen for other taxa, could they be determined or not. We kept the labels *Lycopodium* and grains cut on the FOVs edge, thus increasing the number of labels from 3 to 8 to test for the effect of including information on taxonomy. These five taxa selected were the most frequent ones in the samples, together representing 43% of the tagged grains (Fig. 1 step 2b);
- 1taxon-1taxon restricted pollen detection and thus annotation to a single model taxon, thus labeling bounding boxes only for *Lycopodium* and alternatively one of the five taxa mentioned above (Fig. 1 step 2c).

Lycopodium, used for calibration of pollen counts, was systematically tagged and labeled in all annotation strategies, while debris were never annotated. We compared the detection performance and analyzed the detection errors for each of the tagged pollen taxon under the three annotation strategies. At most 12 531 grains were tagged in the 4098 FOVs, among which 919 were labeled as *Lycopodium*, 5349 as one of the five taxa,

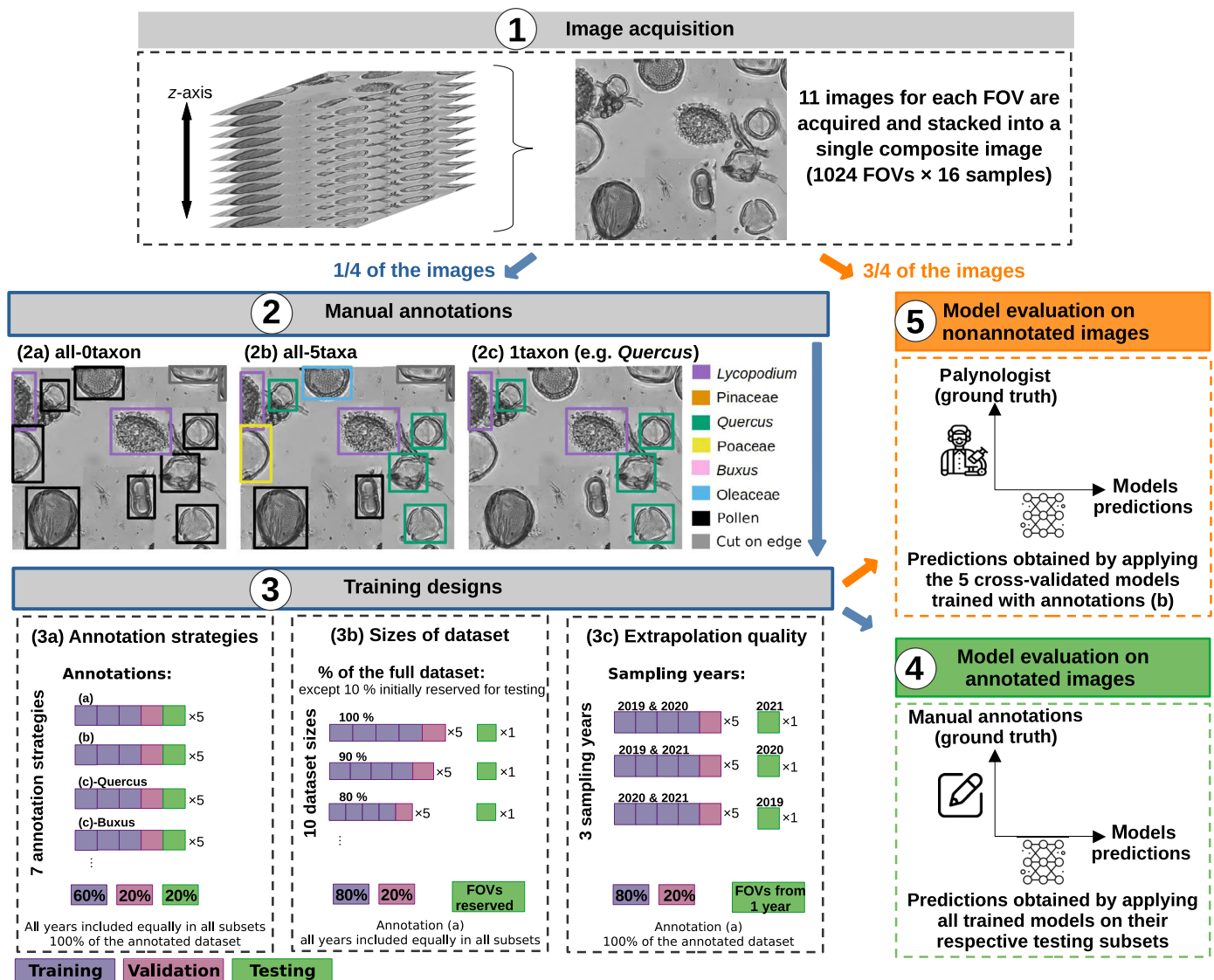


Fig. 1 Experimental design. (1) A pile of 11 images at distinct focal depths is acquired for each field of view (FOV) and then stacked into a single 2D composite image. (2) One-fourth of all images (1024 FOVs × 16 samples × ¼ = 4096) are annotated following three distinct annotation strategies (2a–2c). The annotation strategies are illustrated on a composite image: (a, all-0taxon) tagging 12 bounding boxes with three labels (*Lycopodium*, pollen, and grain cut on edge); (b, all-5taxa) tagging 12 bounding boxes with the same three previous labels and five extra labels (Pinaceae, *Buxus*, Poaceae, *Quercus*, and Oleaceae); (c, 1taxon-1taxon, e.g. *Quercus*) tagging six bounding boxes with two labels (*Lycopodium* and a model taxon); see details in Supporting Information Methods S6. (3) The annotated datasets are split into subsets to train, validate, and test the models with a fivefold cross-validation, following distinct designs (3a–3c) tailored to distinct questions. (4) The performances are evaluated on the testing subsets by comparing manual annotations with predictions obtained with the respective trained models. (5) The models trained with the most detailed annotation strategy 2b (five models from five cross-validations) are applied on nonannotated images (1024 FOVs × 16 samples × ¾ = 12288), and predictions are compared with manual counts made by a palynologist on distinct slides mounted from the same samples.

4615 as pollen with no associated taxon, and 1648 as pollen cut on the FOVs edge. Pollen grains were labeled with a taxon only when they could be confidently determined; otherwise, they were categorized as pollen, meaning that pollen from the five previously mentioned taxa could have been left in this pollen category.

Finally, we assessed the performance of the trained models when applied to new samples never seen by the algorithm. We trained the algorithm on FOVs from all but one sample year

(split randomly with a ratio 80 : 20 for training and validation) and evaluated its performance on 820 random FOVs from the sampling year left apart. We used the simplest all-0taxa annotation strategy, and cross-validated results five times for each year (5 × 3 in total), by interchanging the training and validation datasets, and while keeping the test sets unchanged for each of the three sample years (Fig. 1 step 3c). The sizes of the training datasets (2186 ± 99 FOVs) were larger than the dataset size at which detection performance plateaus (*c.* 1770 FOVs; Fig. S1).

Analyses of the detection errors

We analyzed the variation in the percentage of pollen grains left undetected (FN), using the five following categorical variables, which were included during the annotation process but never used for training: (1) identification, with the two levels determined or not for *c.* 35 taxa, (2) taxon, with the five levels *Buxus*, Pinaceae, Poaceae, Oleaceae, or *Quercus* (the five most common taxa), (3) visible section, with the two levels fully visible within the FOVs or not, (4) image quality, with the two levels good, or poor for grains that are unfocused, covered by a debris or damaged in the image, and (5) deterioration type for poor-quality images only, with the three levels covered, unfocused, or mixed deterioration (Methods S7, S8). False positives predicted as pollen grains were not studied as they all corresponded to debris.

Analyses of the joint detection and classification errors in a full automated analysis

We evaluated the performance of the joint detection and classification produced with the all-5taxa annotation strategy (Fig. 1 step 2b; Methods S6), and compared it to the sole detection error to evaluate whether they can or not compensate. First, we evaluated the performance of the full automation on the same annotated test sets as before (Fig. 1 step 4). We used the confusion matrix (Methods S4a) to compare for each of the five taxa, the counts from automated predictions and from manual annotations, obtained before and after calibration with *Lycopodium* counts. Second, we applied the five cross-validated trained models to the FOVs from the 12 scanned areas of the slides that were not annotated or used before, totaling 768 FOVs per slide. We used an IoU threshold of 0.7 to remove overlapping bounding boxes generated by applying successively the five cross-validated models on the same FOVs. The automated counts were calibrated with their respective *Lycopodium* counts, and finally compared with those obtained manually from nine common gravimetric traps but from different slides (Fig. 1 step 5). Calibrated counts corresponded to the number of pollen grains for 100 *Lycopodium* spores.

Results

Detection performances tended to plateau at *c.* 60% of the full dataset, that is *c.* 1770 training FOVs (Fig. S1; Notes S2), suggesting that the detection performances presented below and obtained with 2452 ± 6 training FOVs are not constrained by the size of the dataset.

Detection of pollen grains regardless of their taxon

We tested whether the models could accurately and precisely detect pollen in the simplest configuration, that is with no distinction of their taxon (all-0taxon; Methods S4). Based on 2506 ± 26 (mean \pm 2SE) manually annotated bounding boxes, the average performance achieved was good (Table S1). The

percentage of grains detected (recall) was $94.8 \pm 0.33\%$, the percentage of correct predictions (precision) was $94.7 \pm 0.38\%$, leaving $5.2 \pm 0.33\%$ of tagged *Lycopodium* or pollen grains left undetected (FN), and falsely detecting $5.3 \pm 0.38\%$ of debris. *Lycopodium* and pollen grains were left undetected with the same frequency ($5.0 \pm 0.25\%$ and $4.4 \pm 0.42\%$, respectively) while grains cut on the FOVs edge were missed twice as frequently ($10.2 \pm 0.98\%$; Table S1). The bounding boxes correctly predicted were well positioned, as they overlapped by $94 \pm 3\%$ (IoU) with manually tagged ones.

We analyzed the variation in FN, using descriptive variables for the annotated grains. Image quality contributed greatly to the detection performance. Pollen and *Lycopodium* grains in images of good quality were left undetected 10 times less frequently than grains with poor visual quality ($1.1 \pm 0.37\%$ vs $10.9 \pm 1.09\%$), especially in unfocused images ($15.6 \pm 2.85\%$) or when covered by a debris ($12.1 \pm 2.21\%$; Fig. 2a). Similarly, pollen grains for which their taxon could not be determined because of the poor image quality, which represent 72% of the impossible identifications, were also left undetected about four times more often ($9.4 \pm 0.50\%$) than grains that could be identified ($2.4 \pm 0.20\%$; Fig. 2b). Missed detections were also higher for pollen and *Lycopodium* grains not fully visible within the FOVs ($8.3 \pm 0.91\%$) compared with those fully visible within the FOVs ($4.1 \pm 0.26\%$; Fig. 2c). Disentangling the effects of taxonomy from the image quality or the visible section of grains was constrained by the uneven distributions of these variables in the dataset. For example, Pinaceae could be identified during annotation, granted to its typical morphology, in any image of bad or good quality, and even for grains not fully visible within the FOVs. By contrast, a good-quality image with a full view of the grain was required to identify *Quercus*, because of its morphological similarity with other taxa (only $5.4 \pm 0.48\%$ of all grains identified as *Quercus* presented nonoptimal conditions compared with $67.9 \pm 3.7\%$ for Pinaceae; Fig. S2). To try to get the independent effect of taxonomy on detection errors, we analyzed only optimal images, that is focused, with grains not covered by debris or deteriorated, and with grains fully visible within the FOVs, totaling 1183 ± 13 annotated grains. Under these conditions, we found grains of *Buxus*, *Quercus*, Poaceae, and Oleaceae were rarely missed (FN below $0.2 \pm 0.37\%$) compared with *Lycopodium* ($1.3 \pm 0.54\%$) and Pinaceae ($3.0 \pm 3.56\%$; Fig. 2d). These differences in detection errors among taxa were not related to the abundance of a given taxon. Similar detection was achieved for *Buxus*, *Quercus*, Poaceae, and Oleaceae despite representing on average $5.3 \pm 0.6\%$ to $27.7 \pm 1.4\%$ of the grains in those optimal images (Fig. 2d). These pollen grains share a common morphology, a circular-shaped monad, which thus is frequent in the dataset. By contrast, Pinaceae, a saccate monad, and *Lycopodium*, a triangular-shaped spore, have specific morphologies, which are thus less abundant in the dataset, accounting, respectively, for $3.5 \pm 0.5\%$ and $10.6 \pm 0.6\%$ grains of optimal images (Fig. 2d). Their shapes are also more irregular, making them more likely to be mistaken for debris. From these results, we conclude that the abundance of a given morphology, not of a given taxon, and its resemblance with debris explained the

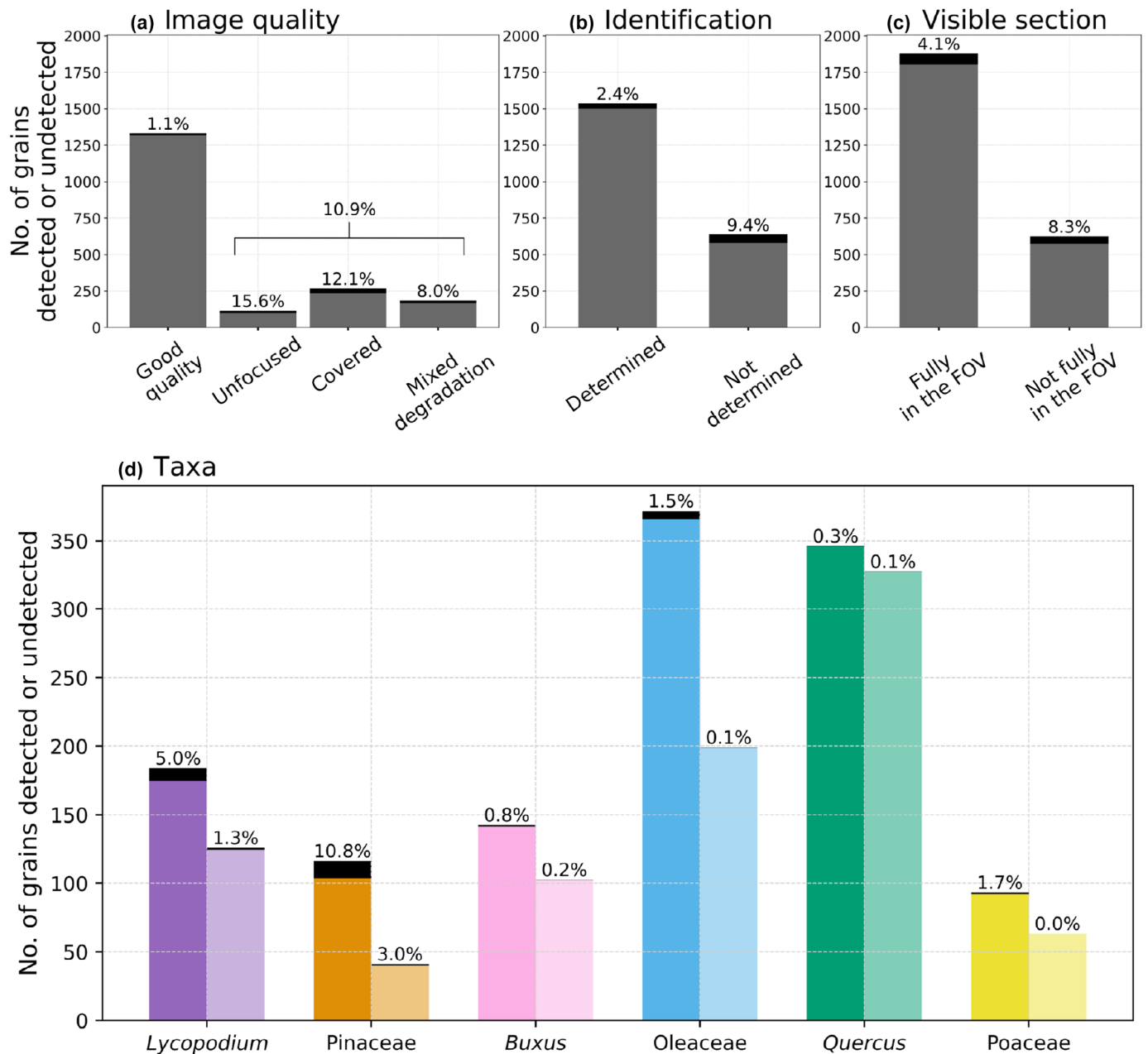


Fig. 2 Analysis of detection errors. Number of pollen grains correctly detected (true positives (TP), gray, and colored bars) or left undetected (false negatives (FN), black bars, and numbers above bars) according to: (a) image quality, (b) identification, discarding the grains not fully visible within the fields of views (FOVs, a and b), (c) visible section, (d) taxon for five taxa and accounting for all grains (darker bars, left) and for grains in optimal conditions (lighter bars, right), that is with good visual quality and fully visible within the FOVs. Values represent means over the five cross-validation tests, and bars' height corresponds to the number of annotated bounding boxes.

performance of detection. Using grains in optimal conditions only also reduced the percentage of grains left undetected by a factor of 10.

Effects of the annotation strategy

Application to new pollen samples Detection was good when models were applied to the sampling years 2019 and 2021, not

included in model training. F1-scores remained above $93.5 \pm 0.45\%$, corresponding to 1% decrease compared with that of the reference models (Fig. 3; Table S2; ROC curves in Fig. S3), and the average percentage of FN ($6.5 \pm 0.45\%$) and FP ($6.5 \pm 0.13\%$) increased at most by 1.2% compared with the reference models (Fig. 3b). Surprisingly, detection performance evaluated on the 2020 sampling year achieved a higher F1-score of $98.1 \pm 0.12\%$. Detection errors decreased by at least 3.4%

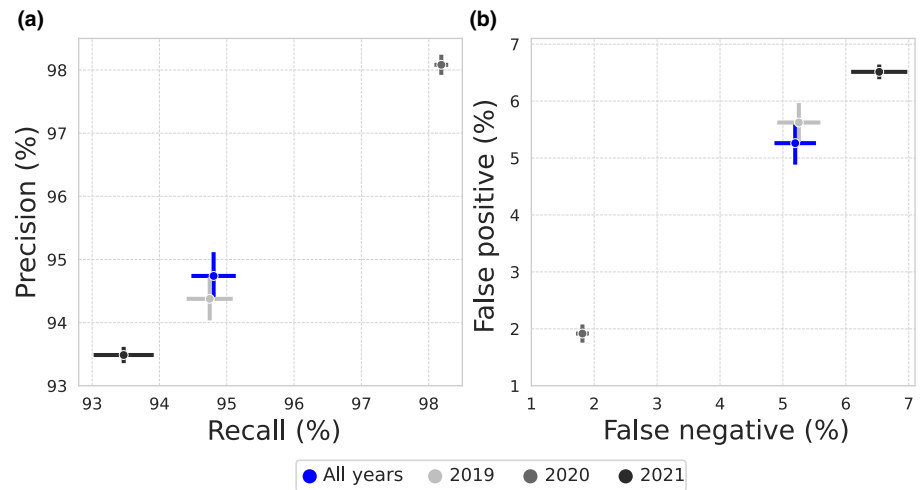


Fig. 3 Performance of the detection method measured with (a) recall and precision, (b) the percentages of false positives and false negatives, when trained on all (blue), or all but one sampling years: 2019 (lighter gray), 2020 (intermediate gray), or 2021 (darker gray). Points and bars represent means and 95% confidence intervals over the five cross-validation tests.

Table 1 Performance metrics of the detection for all annotation strategies (mean \pm 2SE for the five cross-validated test datasets).

Annotation strategy	No. of annotated bounding boxes	Percentage of false negative	Percentage of false positive	Precision	Recall	F1-score
all-0taxon	2506 \pm 26	5.20 \pm 0.33	5.26 \pm 0.38	94.74 \pm 0.38	94.80 \pm 0.33	94.77 \pm 0.08
all-5taxa	2506 \pm 26	6.70 \pm 0.40	5.16 \pm 0.25	94.84 \pm 0.25	93.31 \pm 0.39	94.07 \pm 0.11
1taxon-Pinaceae	300 \pm 9	10.32 \pm 1.09	15.64 \pm 0.85	84.36 \pm 0.85	89.68 \pm 1.09	86.93 \pm 0.43
1taxon-Buxus	326 \pm 12	9.41 \pm 0.68	22.96 \pm 2.12	77.04 \pm 2.12	90.59 \pm 0.68	83.24 \pm 1.15
1taxon-Poaceae	277 \pm 10	10.38 \pm 0.58	24.32 \pm 2.05	75.68 \pm 2.05	89.62 \pm 0.58	82.04 \pm 1.10
1taxon-Oleaceae	556 \pm 19	9.38 \pm 0.59	15.86 \pm 0.78	84.14 \pm 0.78	90.62 \pm 0.59	87.26 \pm 0.45
1taxon-Quercus	530 \pm 13	12.63 \pm 1.74	30.26 \pm 2.15	69.74 \pm 2.15	87.37 \pm 1.73	77.54 \pm 1.72

for both FN ($1.81 \pm 0.10\%$) and FP ($1.92 \pm 0.16\%$). This pattern could result from the greater quality of the pollen images for that year. There were 1865 annotated bounding boxes in the test set in 2020 compared with 2787 in 2019 and 3050 in 2021, and 58.6% of the grains were in images of good quality in 2020, compared with 46.3% in 2019 and 55.1% in 2021 (Fig. S4).

Increasing the label details Including taxonomic details when manually annotating the dataset, with labels for the five most frequent Mediterranean pollen taxa in the dataset, had no effect on the percentage of FP, while the percentage of FN increased on average by 1.5% (Table 1; Fig. 4; ROC curves in Fig. S5a). This slight increase was mostly due to grains cut on the FOVs edge being more frequently missed (67 ± 10 compared with 34 ± 4 not detected, out of 330 ± 19 grains; Fig. 4b).

Selective annotation of a single model taxon As expected, tagging and labeling a single model taxon decreased annotation time compared with tagging and labeling all pollen (1385–2778 bounding boxes depending on the taxon instead of 12 531). This strategy increased the percentage of FP, and to different extents depending on taxon (Table 1; Figs S5, S6). This increase was mainly caused by the wrong detection of grains cut on the FOVs edge and not determined, and grains from other taxa than the targeted one, but not of debris falsely detected as pollen (Fig. 4a).

For example, when detecting *Buxus*, among the 88 ± 9 falsely detected bounding boxes, only 12 ± 4 were debris while 26 ± 4 were Oleaceae pollen, which has a similar reticulated exine, and 23 ± 4 were pollen cut on the FOVs edge. The amount of undetected grains (FN) for Pinaceae and *Lycopodium* grains was not affected by the annotation strategy (Fig. 4b; Table S3). By contrast, pollen with common morphologies, that is *Quercus*, Poaceae, *Buxus*, and Oleaceae, were left undetected more frequently when tagged solely with *Lycopodium* compared to when tagged along with the other pollen grains (Fig. 4b; Table S3).

Combination of detection and classification errors in a full automated pollen analysis

Here, we analyzed errors combining both detection and classification to evaluate how the method predicts pollen counts, assessing the drivers of potential biases.

Compensation between undetected grains and falsely detected debris In this study, we chose a confidence score threshold to balance on average the percentage of undetected grains and of falsely detected debris (Methods S3). We found that FN were unevenly distributed among taxa; *Lycopodium* and Pinaceae grains were left undetected more frequently than others because of their specific morphologies. Using the classification

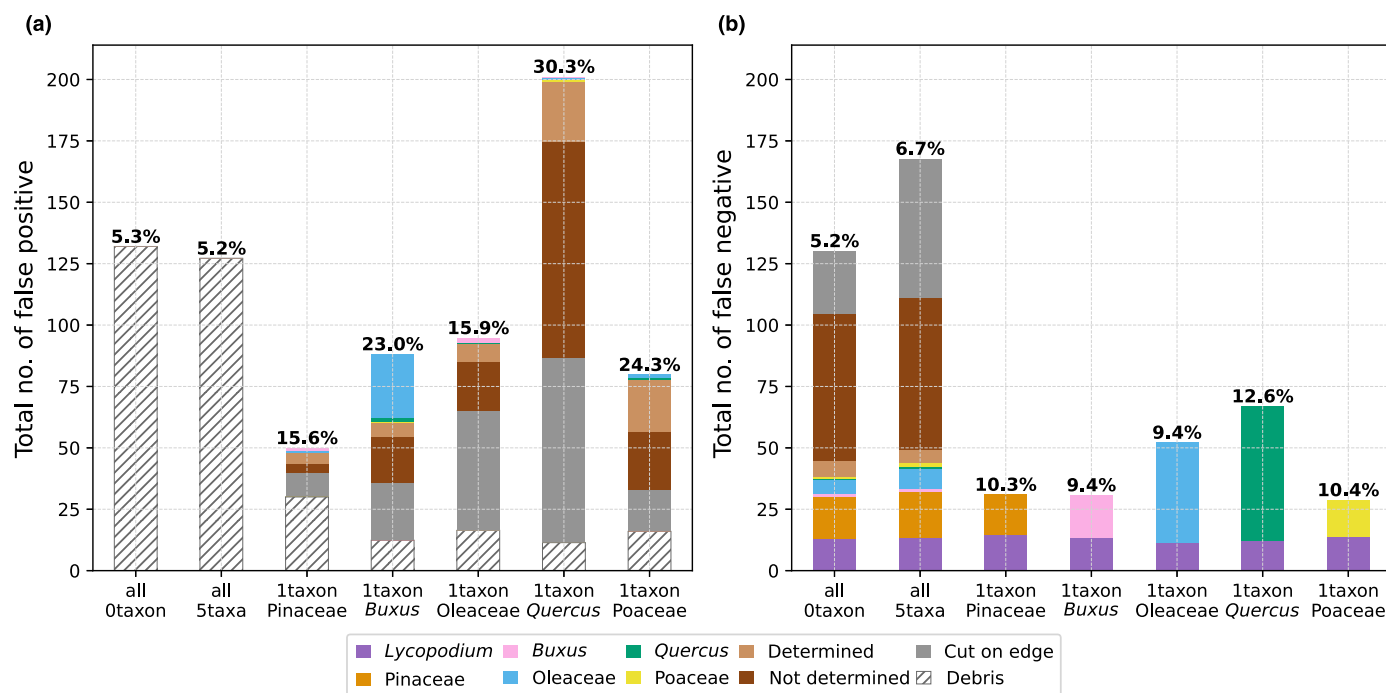


Fig. 4 Effect of annotation strategy on detection errors. Number (bars height) and percentage (values above bars) of (a) false positives, that is detected but not annotated debris and pollen grains, and (b) false negatives, that is annotated but not detected pollen grains, obtained with each annotation strategy, and averaged over the five cross-validation tests.

information for falsely detected debris, we found that their distribution was also uneven among taxa, and followed the same above pattern for detection: Pinaceae and *Lycopodium* grains were also associated with a higher percentage of falsely detected debris (4.2% and 7.6%) compared with the other four taxa defined by common and regular morphologies (below 2.3%; black in Fig. 5b; Table S4a). Therefore, errors during detection from both undetected grains and detected debris partially compensated with each other.

Accuracy of combined detection and classification Once detected, the five target taxa were almost never misclassified within the five categories, with the exception of *c.* 2% of grains tagged as Oleaceae but misclassified to *Buxus*, and 0.7% of grains tagged as *Buxus* but misclassified to Oleaceae, which both have a reticulated exine (light blue; Fig. 5; Table S4b). Most classification errors for the target taxa, with the exception of *Lycopodium* and Pinaceae with specific morphologies, were made to the class of pollen from other taxa simply labeled as pollen (dark blue; Fig. 5), and pollen cut on the FOVs edge (gray, Fig. 5). The large amount of grains from the pollen category and misclassified to target taxa was partly compensated by the amount of grains from the target taxa misclassified to the pollen category, thus limiting the overestimation of the predicted counts (dark blue; Fig. 5). A *post hoc* visual examination of the images showed that some misclassifications to other pollen taxa corresponded to pollen with similar morphologies, for example few *Viburnum* and Brassicaceae pollen grains were misclassified as Oleaceae, all of which have a reticulated exine, while most misclassifications were to

pollen that could not be determined during annotation. We cannot exclude that some of these latter misclassifications could actually correspond to the correct pollen taxon, that is a true annotation error or a lack of confidence to label a grain. Similarly, grains labeled as cut on the FOVs edge and misclassified to a pollen taxon could correspond to a correct identification. *Post hoc* visual inspections confirmed this hypothesis for Pinaceae. Once detected, the classification of *Lycopodium* was very efficient (Fig. 5). Less than 0.9% were misclassified, and all grains cut on the FOVs edge and classified to *Lycopodium* were actually correct. This last result confirms that automated detection of pollen using *Lycopodium* for calibration should provide accurate counts.

Finally, predicted counts for all labeled categories were systematically overestimated compared with manual annotations, except for the category of pollen cut on the FOVs edge, which were directly classified to pollen taxa (Fig. 6a). As *Lycopodium* counts were also overestimated, the inferred pollen counts after calibration, matched very well the calibrated pollen counts from the manual annotations (Fig. 6b).

Application of the automated method to routine conditions and comparison with a palynologist Once the method was established and the images were acquired, predictions for each slide produced up to 5700 detected and classified pollen grains in one step and only a few minutes, compared with *c.* 2–3 h for the standard microscopy analysis by the palynologist expert. Calibrated counts for the five most common taxa predicted by the models and obtained by the palynologist, for the same gravimetric traps but not the same slides, were close to each other,

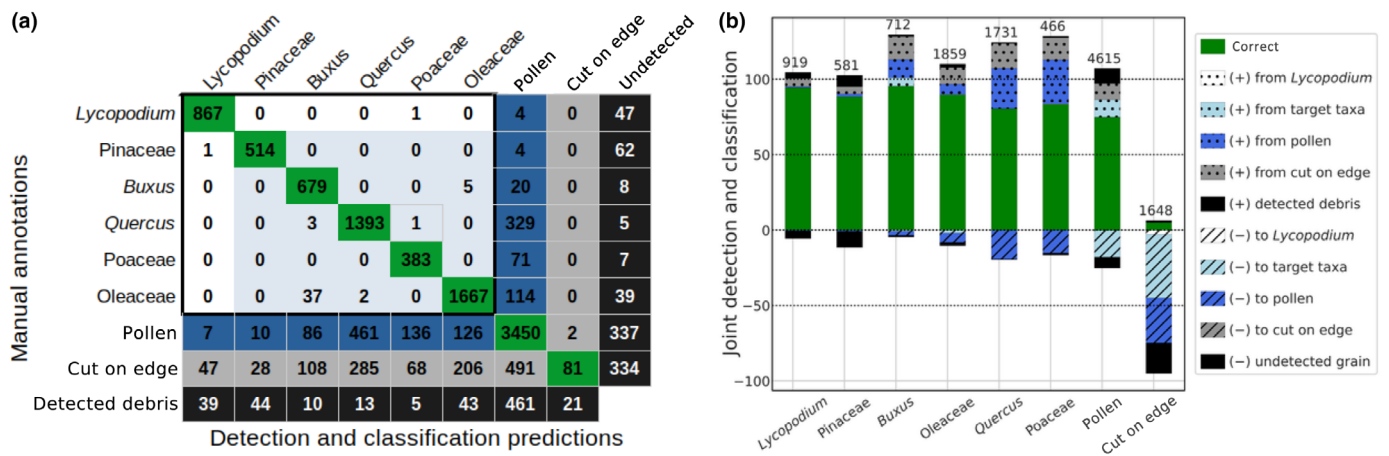


Fig. 5 Performance of the joint detection and classification. Grains correctly or incorrectly detected and/or classified (a) as the confusion matrix, (b) plotted for each label separately. Results are summed over the five cross-validation tests from the all-5taxa strategy, and show correct detection and classification (green), detection errors (black), and classification errors from confusion with: *Lycopodium* (white), the five target taxa (light blue), the category pollen (dark blue), or grains cut on the edge of the fields of views (FOVs, gray). In (b), gridded bars represent errors under-estimating the class (grains undetected or classified into another class), and dotted bars represent errors overestimating the class (detected debris or grains not belonging but classified into the class); values above bars are the number of bounding boxes annotated for each label.

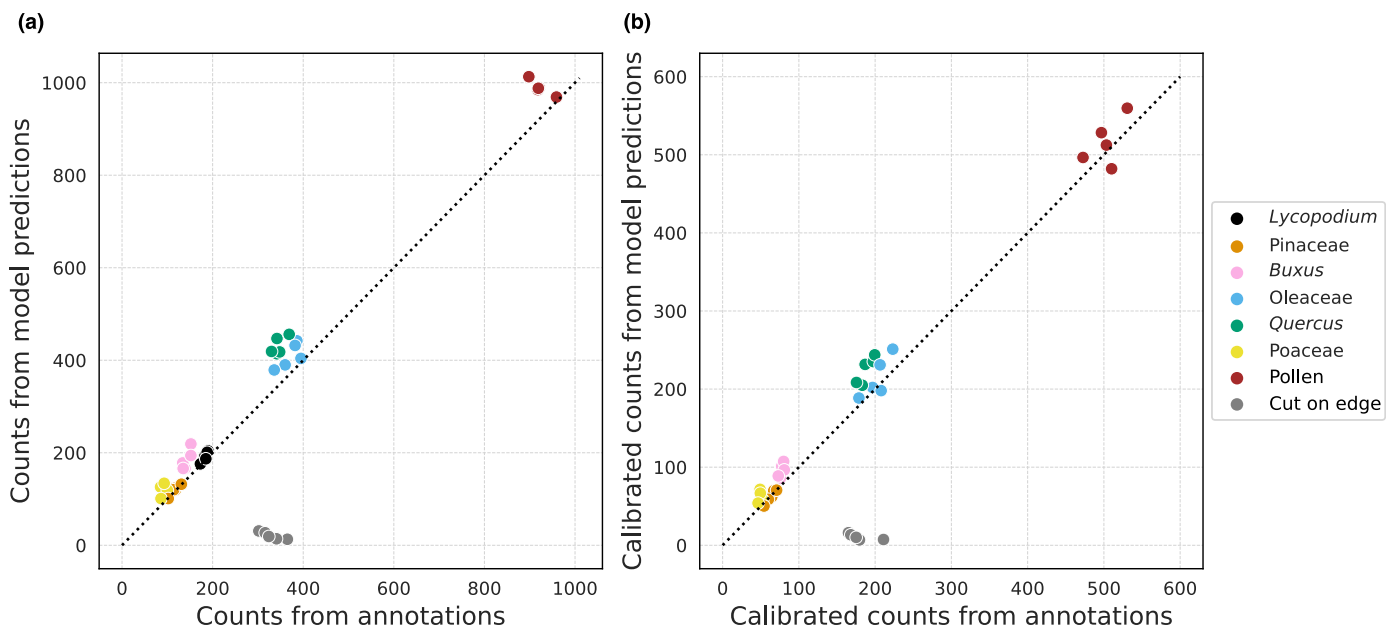


Fig. 6 Comparison of automated predictions and manual annotations. (a) Counts for the five taxa and *Lycopodium*, and (b) counts calibrated with *Lycopodium*; both for the manual annotations and from the model predictions in each of the five cross-validation tests (some points overlap), and obtained with the annotation strategy all-5taxa; the dotted line is the 1 : 1 line.

except for *Buxus*, rare taxa in these samples for which sampling variance is expected to be large (Fig. 7). Predicted counts for *Quercus* were systematically underestimated compared with manual counts, potentially because we labeled *Quercus* simply as pollen when we could not confidently identify it, and which was replicated by the model, while the expert could identify *Quercus* in many more conditions. Predicted counts for Pinaceae tended to be overestimated compared with manual counts, potentially because Pinaceae grains were manually annotated as 1/2 when torn but counted as a full pollen grain by the model.

Discussion

Environmental pollen samples are no longer a barrier to automated detection

Despite challenges inherent to environmental samples, with images not manipulated before the analyses, and thus containing many debris and pollen taxa, the performance of pollen detection here is consistent with two similar and recent studies. From European lake sediment samples, 87.2–99.1% of pollen was retrieved for a Faster

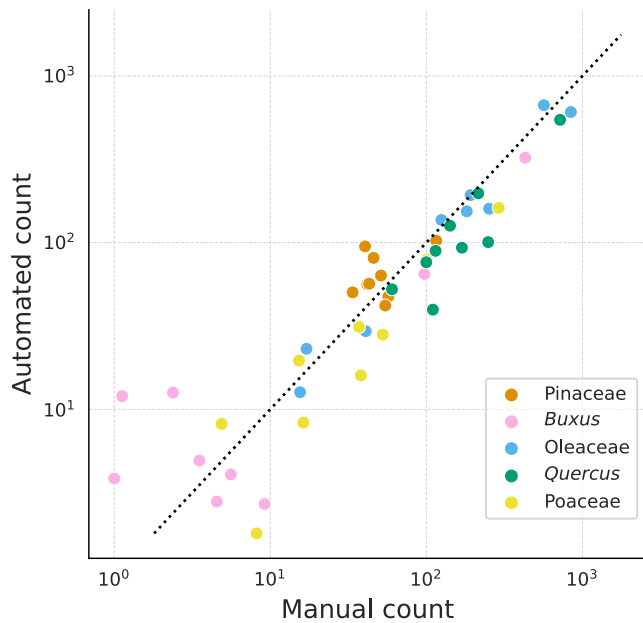


Fig. 7 Comparison of automated predictions and count from palynologist. Manual (expert palynologist) vs automated pollen counts from five key taxa; the models used for automated counts were trained with the strategy all-5taxa and applied to images not used before (704 fields of views for each of nine slides). Results are log-transformed, the dotted line is the 1 : 1 line.

R-CNN detector on *c.* 1450 annotations (Theuerkauf *et al.*, 2023), and from pollen trap samples collected in a tropical forest, 83.7% of pollen was retrieved for a modified ResNet34 architecture on 14 764 annotations (Punyasena *et al.*, 2022), while our method retrieved 95.6% of pollen and 95.0% of *Lycopodium* using as many as 12 531 annotations. Only 5.3% of the total detected objects were debris misidentified as pollen (Table S1), and the predicted bounding boxes were very well positioned (IoU of 0.94), possibly making subsequent classification efficient. Missed detections were mainly caused by grains covered by debris, unfocused grains, or grains not fully visible within the FOVs (Fig. 2). These cases are likely to occur regardless of pollen taxon, and thus not likely to bias the pollen counts in the detected assemblage. These grains with poor visual quality also lack discernible identification criteria, making them irrelevant for classification, be it by an algorithm or an expert palynologist. Excluding these grains with visual poor quality increased pollen detection to 99.4% recall, a performance consistent with that obtained on reference pollen images taken under optimal conditions, that is with one or very few taxa from flower anthers with no debris (Gallardo-Caballero *et al.*, 2019; Khanzhina *et al.*, 2022; Kubera *et al.*, 2022).

How to efficiently increase detection performance

Debris generated FN by masking pollen grains, and also FP, since debris could be misidentified to pollen grains. Thus any strategy that can eliminate debris will undoubtedly increase detection performance, for example through chemical treatment or sieving, or

by diluting the material in the slides. The imaging process, fully automated here, also contributed to detection errors by generating images of unfocused pollen grains, which were left undetected more frequently than pollen in focused images (Fig. 2a). Therefore, improving the focus of pollen grains shall decrease the percentage of undetected pollen, for example by increasing the number of focal planes during image acquisition, or reducing the thickness of the slide preparation.

Our results also show that increasing the taxonomic details, and thus number of labels, required tedious identification effort but did not improve the performance of pollen detection. Slightly more grains were left undetected for similar numbers of debris falsely detected (Fig. 4). If the goal is to detect pollen regardless of taxon, we thus recommend to not spend time identifying and labeling taxa, but instead use a few general categories as done here (*Lycopodium*, pollen and cut on edge). Apart from the high performance this strategy can achieve, it has the advantage that it can be done by nonpalynologists. We also found that the annotation, and thus detection, of a single model taxon was effective only for morphologically distinct taxa such as Pinaceae or *Lycopodium* spores (Fig. 4). This strategy can be used, as it saves time, only if the model taxon has a distinct shape and can be identified with a good confidence in all images. Lastly, although often overlooked, the trade-off between FP and FN can be adjusted to the study objectives, by modifying the confidence score threshold to filter the predicted bounding boxes, for instance by increasing it to reduce the detection of debris (Methods S3).

Using a model pretrained on pollen images, even if outside the data to be analyzed, can help limit the number of pollen images to annotate and increase pollen detection. Therefore, we make our best-trained model here on the anemophilous flora of the Mediterranean region available to all, calling also for a large share among research teams of models and data to increase the performance of all future automated pollen analyses.

Robustness of the automated detection method to extrapolation

Our models achieved the same detection performance whether they were applied to samples used or not to train the model (Fig. 3). This result has important consequences for many pollen-based studies, based on long-term monitoring (van der Knaap *et al.*, 2010), or long fossil sequences (Donders *et al.*, 2021) or at large spatial scales. For a given study, good detection performance can be achieved by training the algorithm only once on annotated images from a few samples, and without the need to reiterate the training process on new sampling years and/or new locations. Of course, the inherent variation in data between studies, such as the chemical treatment of samples, microscope settings, or the sample content itself, will affect the detection performances. It should however be noted that using models pretrained on images from any pollen study can help increase detection performances, and limit the annotation workload required to build a new training dataset.

Comprehensive evaluation of the joint detection and classification to achieve full automated pollen counts

The proposed method detected and classified *Lycopodium* spores and five pollen taxa, which is a small fraction of pollen diversity present in the samples. Our goal was to provide guidelines for other pollen studies while focusing on detection errors and on how they propagated to classification, and not to make a full analysis of the samples. The chosen taxa accounted for 43% of all pollen grains in the samples, and also corresponded to a diversity of morphologies, which thus allowed to address our goals. Detection errors were affected by the abundance of the grain morphologies and resemblance with debris, rather than by the taxon abundance. We found fewer FN for taxa that share a standard and thus abundant pollen morph (Oleaceae, Poaceae, *Buxus*, and *Quercus*) compared with taxa that have rare and irregular morphologies (Pinaceae and *Lycopodium* spores, Fig. 2d), which were thus more underestimated in the detected pollen assemblage. Nonetheless, when detection was performed jointly with classification, grains with rarer and irregular morphologies were also assigned a higher number of falsely detected debris, thus partially compensating errors. When classification is conducted separately from detection, debris falsely detected are often processed with dedicated classes or processes (Crouzy *et al.*, 2022; Zhao *et al.*, 2022), and pollen grains left undetected are usually not considered in the final assessment, which may generate biases. The proposed method, based on an algorithm jointly conducting detection and classification, thus integrates detection errors through both stages, which effectively mitigates both types of detection errors.

The detected grains of *Lycopodium* and of the five target taxa were classified with very little confusion between themselves (Figs 5a, S6). Classification errors of the target taxa mainly occurred for grains of other morphologically similar taxa, and for grains that could not be identified with confidence, which we both labeled here as simply pollen (Fig. 5a). The proposed method requires to annotate, and thus identify, pollen grains directly in 2D images containing many different taxa and debris, sometimes unfocused, and in which pollen may be not well oriented or sufficiently visible to use the appropriate discriminant criteria for identification. Such conditions make identification, when annotating the images, challenging and sometimes not possible, which will generate ambiguity when training the models, and contribute to increase detection errors. The presence of grains not identified and annotated in the test dataset also prevents the accurate evaluation of errors, as some predictions are likely to be correct, for example a pollen grain is indeed a *Quercus* one but we did not label it as such. Nonetheless, misclassifications from the target taxa to the category of other grains simply labeled as pollen, and from the category pollen to target taxa partly compensated each other. The distinction of grains cut on the FOVs edge also generated many misclassifications, although some were correct; we thus recommend avoiding this label, and instead, to directly label cut pollen grains with their taxa or as simply pollen. Despite these challenges and potential biases, automated counts matched those manually

done, but also counts obtained by a palynologist expert with standard microscopy analysis.

If the goal is to get pollen counts per taxon for many taxa, improving the confidence of manual identifications will for sure improve classification performances. Identifications used for training could be improved by enhancing the visibility of pollen structures in the images, for example by decreasing pollen density, using colored images, or using images of reference pollen collected in flowers' anthers. This latter suggestion though may be less effective for detecting pollen in environmental samples afterwards, as the model also trains on the background of the images, clear of debris in reference samples. Conducting a separate classification, after the detection and segmentation of pollen of any taxon from their original sample images, would bypass the challenging and error-prone step of manual pollen identification and annotation on images taken from environmental samples. In that case, we recommend that classification will be performed on images of pollen from known species and plant individuals. This approach should especially improve classification performance for taxa of common morphologies that are difficult to identify. It would also allow the identification of rare taxa that *de facto* have too few images in the samples to train the models (e.g. only one pollen grain of *Juglans* was found in our dataset). Conducting a separate classification may finally benefit from the extensive work carried out by many these last few years on pollen. Large image classification datasets such as POLLEN23E (Sevillano & Aznarte, 2018), POLLEN73S (Astolfi *et al.*, 2020), or POLLEN13K (Battiatto *et al.*, 2020) were made open source. Diverse new CNN-based image classification methods have recently been developed to improve the performances of pollen classification. One approach relies on both taxonomical and morphological labels to train the classification models (Barnes *et al.*, 2023), others use multi-CNN architectures with a decision tree (Bourel *et al.*, 2020), or add a preliminary image deblurring process before classification, combined to a multi-scale architecture to also include image sizes in the training (Chen & Ju, 2022).

Our study is based on purpose on simple tools and thus on images from slides scanned under light microscopy. Other types of pollen data have been tested for the automation of pollen analysis, and gave good classification performances, by-passing some limitations encountered in our study: pollen images acquired with flux cytometry (Dunker *et al.*, 2021; Barnes *et al.*, 2023), with scanning electric microscopy (Li *et al.*, 2023), or with confocal microscopy then classified with a 3D-classification algorithm (Wang *et al.*, 2021).

Concluding remarks

Our work represents a comprehensive attempt to assess joint detection and classification of pollen using artificial intelligence, and shows that pollen detection is a critical step for getting accurate pollen counts in pollen assemblage. Overall, our method, which relies on standard equipments, simple tools, and rules, provides excellent performance on environmental samples from the 'real-world' containing many debris and pollen taxa. The

method can generate for a slide, once the images are acquired and labeled, and the models are trained, up to 5700 detected and labeled pollen grains in only a few minutes, compared with *c.* 500 grains in 2–3 h for a palynologist, with no subjectivity or fatigue, and can be confidently extrapolated to new samples not seen by the models.

Acknowledgements

We thank the Doctoral school GAIA, Univ Montpellier (Montpellier, France) (Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau), the CNRS PEPS project DETECT to CD, the Univ Montpellier MUSE to OP and CD, the OSU OREME, Observatoire de REcherche Montpelliérain de l'Environnement (Montpellier, France) SO POLLIMED, and SO POLLUMINE (<https://oreme.org/observation/>), which all funded this study. We also thank three anonymous reviewers for their helpful comments. This work was performed using HPC resources from GENCI-IDRIS (Grant no.: 2023-(AD011013554R1)). This work is a ISEM contribution (ISEM 2024-096). For the purpose of Open Access, a CC-BY public copyright license has been applied by the authors to the present document and will be applied to all subsequent versions up to the author-accepted manuscript arising from this submission.












Competing interests

None declared.

Author contributions

BG, CD, OP and SJ conceived the ideas, designed the methodology, analyzed the data, and led the writing of the manuscript. JP provided ideas on the methodology of YOLO. L Beaufort, TG-T, and YG helped with the image acquisition process. L Bouby, J-FT, SC, SI, BL and BG collected pollen in the field and conducted wet laboratory work. NC-N counted pollen manually on the microscope and helped with image analysis. All authors helped with the writing of the manuscript. CD and OP contributed equally to this work.

ORCID

Luc Beaufort  <https://orcid.org/0000-0001-6055-9373>
 Laurent Bouby  <https://orcid.org/0000-0002-3633-9829>
 Nathalie Combourieu-Nebout  <https://orcid.org/0000-0002-3604-5986>
 Céline Devaux  <https://orcid.org/0000-0002-5192-2828>
 Thibault de Garidel-Thoron  <https://orcid.org/0000-0001-8983-9571>
 Betty Gimenez  <https://orcid.org/0009-0005-1727-0669>
 Sarah Ivorra  <https://orcid.org/0000-0003-0314-8054>
 Sébastien Joannin  <https://orcid.org/0000-0001-8345-9252>
 Jérôme Pasquet  <https://orcid.org/0000-0002-7993-4724>
 Odile Peyron  <https://orcid.org/0000-0002-7028-6302>
 Jean-Frédéric Terral  <https://orcid.org/0000-0003-1921-2161>

Data availability

The data that support this study (images, annotation metadata, and the weights of the YOLOv5 trained models) are openly available in Zenodo (<https://zenodo.org/>) at doi: [10.5281/zenodo.11126431](https://doi.org/10.5281/zenodo.11126431).

References

- Anderegg WRL, Abatzoglou JT, Anderegg LDL, Bielory L, Kinney PL, Ziska L. 2021. Anthropogenic climate change is worsening North American pollen seasons. *Proceedings of the National Academy of Sciences, USA* 118: e2013284118.
- Astolfi G, Gonçalves AB, Menezes GV, Borges FSB, Astolfi ACMN, Matsubara ET, Alvarez M, Pistori H. 2020. POLLEN73S: an image dataset for pollen grains classification. *Ecological Informatics* 60: 101165.
- Barnes CM, Power AL, Barber DG, Tennant RK, Jones RT, Lee GR, Hatton J, Elliott A, Zaragoza-Castells J, Haley SM *et al.* 2023. Deductive automated pollen classification in environmental samples via exploratory deep learning and imaging flow cytometry. *New Phytologist* 240: 1305–1326.
- Battiato S, Ortis A, Trenta F, Ascari L, Politi M, Siniscalco C. 2020. Pollen13k: a large scale microscope pollen grain image dataset. *IEEE International Conference on Image Processing (ICIP)* 2456–2460. doi: [10.1109/ICIP40778.2020.9190776](https://doi.org/10.1109/ICIP40778.2020.9190776).
- 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.
- Chen X, Ju F. 2022. Automatic classification of pollen grain microscope images using a multi-scale classifier with SRGAN deblurring. *Applied Sciences* 12: 7126.
- Corvucci F, Nobili L, Melucci D, Grillenzoni F-V. 2015. The discrimination of honey origin using melissopalynology and Raman spectroscopy techniques coupled with multivariate analysis. *Food Chemistry* 169: 297–304.
- Crouzy B, Lieberherr G, Tummon F, Clot B. 2022. False positives: handling them operationally for automatic pollen monitoring. *Aerobiologia* 38: 429–432.
- Diwan T, Anirudh G, Tembhurne JV. 2022. Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications* 82: 9243–9275.
- Donders T, Panagiotopoulos K, Koutsodendris A, Bertini A, Mercuri AM, Masi A, Combourieu-Nebout N, Joannin S, Kouli K, Kousis I *et al.* 2021. 1.36 million years of Mediterranean forest refugium dynamics in response to glacial-interglacial cycle strength. *Proceedings of the National Academy of Sciences, USA* 118: e2026111118.
- Dunker S, Motivans E, Rakosy D, Boho D, Mäder P, Hornick T, Knight TM. 2021. Pollen analysis using multispectral imaging flow cytometry and deep learning. *New Phytologist* 229: 593–606.
- Gallardo-Caballero R, García-Orellana CJ, García-Manso A, González-Velasco HM, Tormo-Molina R, Macías-Macías M. 2019. Precise pollen grain detection in bright field microscopy using deep learning techniques. *Sensors* 19: 3583.
- Holt KA, Bennett KD. 2014. Principles and methods for automated palynology. *New Phytologist* 203: 735–742.
- Jiang P, Ergu D, Liu F, Cai Y, Ma B. 2022. A review of Yolo algorithm developments. *Procedia Computer Science, The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19* 199: 1066–1073.
- Jocher G. 2020. Ultralytics YOLOv5 (7.0). *Zenodo*. doi: [10.5281/zenodo.3908559](https://doi.org/10.5281/zenodo.3908559).
- Jubayer F, Soeb JA, Mojumder AN, Paul MK, Barua P, Kayshar S, Akter SS, Rahman M, Islam A. 2021. Detection of mold on the food surface using YOLOv5. *Current Research in Food Science* 4: 724–728.
- Khanzhina N, Filchenkov A, Minaeva N, Novoselova L, Petukhov M, Kharisova I, Pinaeva J, Zamorin G, Putin E, Zamyatina E *et al.* 2022. Combating data

- incompetence in pollen images detection and classification for pollinosis prevention. *Computers in Biology and Medicine* 140: 105064.
- van der Knaap WO, van Leeuwen JFN, Svitavska-Svobodova H, Pidek IA, Kvavadze E, Chichinadze M, Giesecke T, Kaszewski BM, Oberli F, Kalnina L *et al.* 2010. Annual pollen traps reveal the complexity of climatic control on pollen productivity in Europe and the Caucasus. *Vegetation History and Archaeobotany* 19: 285–307.
- Kubera E, Kubik-Komar A, Kurasinski P, Piotrowska-Weryszko K, Skrzypiec M. 2022. Detection and recognition of pollen grains in multilabel microscopic images. *Sensors* 22: 2690.
- Li J, Xu Q, Cheng W, Zhao L, Liu S, Gao Z, Xu X, Ye C, You H. 2023. Weakly supervised collaborative learning for airborne pollen segmentation and classification from SEM images. *Life* 13: 247.
- Lin T. 2015. *HumanSignal/LABELIMG*. [WWW document] URL <https://github.com/HumanSignal/labelImg> [accessed 21 May 2024].
- Lin T-Y, Goyal P, Girshick R, He K, Dollar P. 2017. *Focal loss for dense object detection*. Presented at the Proceedings of the IEEE International Conference on Computer Vision, 2980–2988.
- Marini S, Bonofiglio F, Corgnati LP, Bordone A, Schiaparelli S, Peirano A. 2022. Long-term automated visual monitoring of Antarctic benthic fauna. *Methods in Ecology and Evolution* 13: 1746–1764.
- Morente-López J, Lara-Romero C, Ornosca C, Iriondo JM. 2018. Phenology drives species interactions and modularity in a plant – flower visitor network. *Scientific Reports* 8: 9386.
- Olsson O, Karlsson M, Persson AS, Smith HG, Varadarajan V, Yourstone J, Stjernman M. 2021. Efficient, automated and robust pollen analysis using deep learning. *Methods in Ecology and Evolution* 12: 850–862.
- Oteros J, Orlandi F, Garcia-Mozo H, Aguilera F, Ben Dhiab A, Bonofiglio T, Abichou M, Ruiz-Valenzuela L, Mar del Trigo M, Diaz de la Guardia C *et al.* 2014. Better prediction of Mediterranean olive production using pollen-based models. *Agronomy for Sustainable Development* 34: 685–694.
- Peyron O, Combourieu-Nebout N, Brayshaw D, Goring S, Andrieu-Ponel V, Desprat S, Fletcher W, Gambin B, Ioakim C, Joannin S *et al.* 2017. Precipitation changes in the Mediterranean basin during the Holocene from terrestrial and marine pollen records: a model–data comparison. *Climate of the Past* 13: 249–265.
- Punyasena SW, Haselhorst DS, Kong S, Fowlkes CC, Moreno JE. 2022. Automated identification of diverse Neotropical pollen samples using convolutional neural networks. *Methods in Ecology and Evolution* 13: 2049–2064.
- Redmon J, Divvala S, Girshick R, Farhadi A. 2016. You only look once: unified, real-time object detection. *arXiv*: 1506.02640. doi: [10.48550/arXiv.1506.02640](https://doi.org/10.48550/arXiv.1506.02640).
- Ren S, He K, Girshick R, Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*. Curran Associates, 28. *arXiv*: 1506.01497. doi: [10.48550/arXiv.1506.01497](https://doi.org/10.48550/arXiv.1506.01497).
- 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.
- Stillman EC, Flenley JR. 1996. The needs and prospects for automation in palynology. *Quaternary Science Reviews* 15: 1–5.
- Stockmarr J. 1971. Tablets with spores used in absolute pollen analysis. *Pollen et Spores* 13: 615–621.
- Tetard M, Marchant R, Cortese G, Gally Y, de Garidel-Thoron T, Beaufort L. 2020. Technical note: a new automated radiolarian image acquisition, stacking, processing, segmentation and identification workflow. *Climate of the Past* 16: 2415–2429.
- Theuerkauf M, Siradze N, Gillert A. 2023. A trainable object finder, selector and identifier for pollen, spores and other things: a step towards automated pollen recognition in lake sediments. *The Holocene* 34: 297–305.
- Viertel P, Koenig M. 2022. Pattern recognition methodologies for pollen grain image classification: a survey. *Machine Vision and Applications* 33: 18.
- Wang Z, Wang Z, Wang L. 2021. Automatic 3D pollen recognition based on convolutional neural network. *Scientific Programming* 2021: 5577307.
- Zedda L, Loddo A, Di Ruberto C. 2022. A deep learning based framework for malaria diagnosis on high variation data set. In: Sclaroff S, Distanto C, Leo M, Farinella GM, Tombari F, eds. *Image analysis and processing, ICIAP 2022, Pt II*. Cham, Switzerland: Springer, 358–370.
- Zhao L-N, Li J-Q, Cheng W-X, Liu S-Q, Gao Z-K, Xu X, Ye C-H, You H-L. 2022. Simulation palynologists for pollinosis prevention: a progressive learning of pollen localization and classification for whole slide images. *Biology* 11: 1841.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Performance of the detection for each size of the dataset from the last epoch model, and from the last vs the best epoch model.

Fig. S2 Distribution of grains in optimal conditions per taxon and category.

Fig. S3 Performance of the detection of models trained on all or all but one sampling year.

Fig. S4 Performance of the detection of models trained on all or all but one sampling year, and considering the image properties.

Fig. S5 Performance of the detection obtained for all annotation strategies.

Fig. S6 Confusion matrix obtained when annotating only one model taxon and *Lycopodium*.

Methods S1 Description of YOLO algorithms.

Methods S2 Illustration of the process that eliminates predicted bounding boxes that overlap.

Methods S3 ROC curve of the detection, showing the trade-off between precision and recall.

Methods S4 Confusion matrix for joint detection and classification and for detection only.

Methods S5 Definition of the detection metrics.

Methods S6 Description of the annotation strategies and the associated datasets.

Methods S7 Variables used to characterize manually annotated grains.

Methods S8 Illustration of the categorical variables used to evaluate the causes of undetected grains.

Notes S1 Analysis of the performance of the detection from the last vs the best epoch models.

Notes S2 Analysis of the performance of the detection for each size of the dataset.

Table S1 Detection metrics obtained using the annotation strategy all-0taxon.

Table S2 Performance of the detection of models trained on all or all but one sampling year.

Table S3 Proportion of grains left undetected measured separately for each annotated labels, and for each annotation strategy.

Table S4 Confusion between categories and compensation of false negatives and false positives confusions.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.