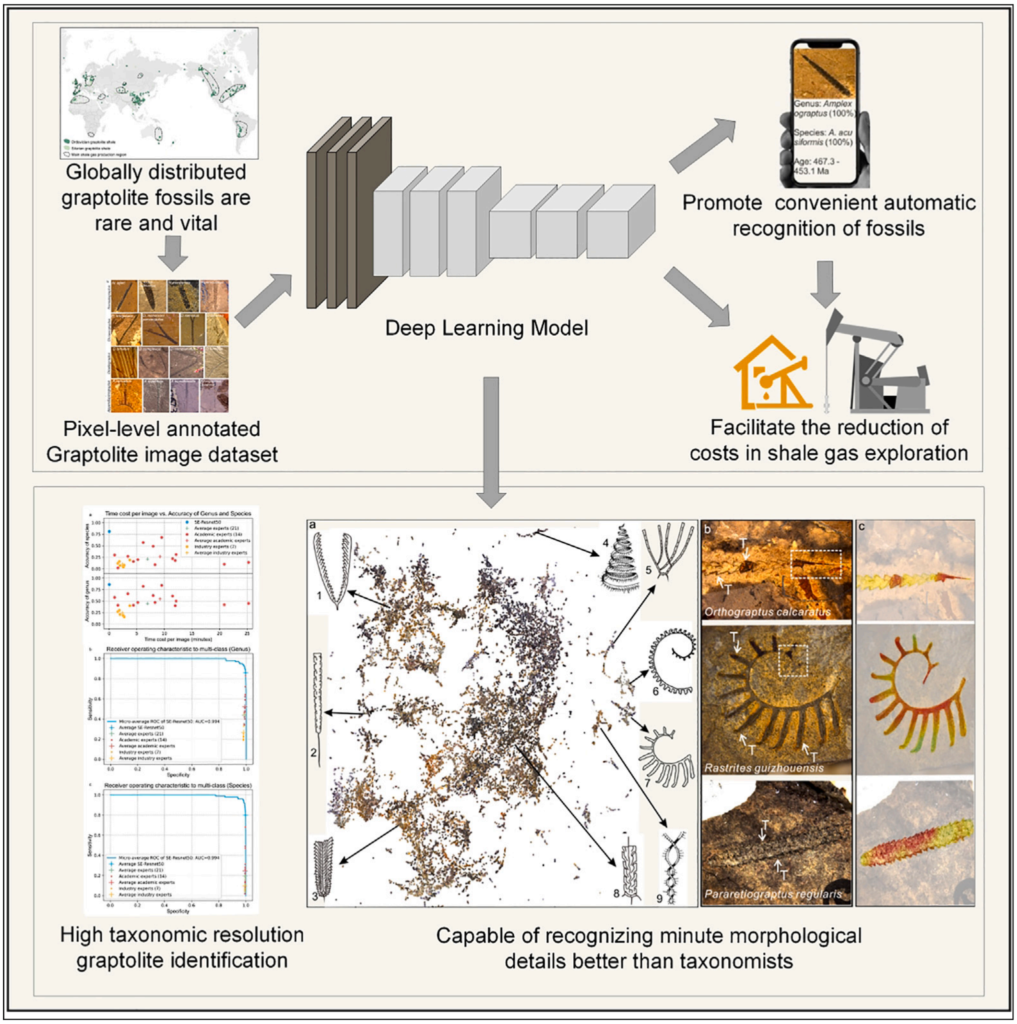
iScience

CellPress
OPEN ACCESS

Article

Automated graptolite identification at high taxonomic resolution using residual networks


Globally distributed graptolite fossils are rare and vital
Pixel-level annotated Graptolite image dataset
Deep Learning Model
Genus: Amplexograptus (100%)
Species: A. acusiformis (100%)
Age: 467.3 - 453.1 Ma
Promote convenient automatic recognition of fossils
Facilitate the reduction of costs in shale gas exploration
Time cost per image vs. Accuracy of Genus and Species
Receiver operating characteristic to multi-class (Genus)
Receiver operating characteristic to multi-class (Species)
Sensitivity
Specificity
High taxonomic resolution graptolite identification
a
b
c
1
2
3
4
5
6
7
8
9
Orthograptus calcaratus
Rastrites guizhouensis
Pararetiograptus regularis
Capable of recognizing minute morphological details better than taxonomists

Zhi-Bin Niu, Si-Yuan Jia, Hong-He Xu

zniu@tju.edu.cn (Z.-B.N.)
hhxu@nigpas.ac.cn (H.-H.X.)

Highlights

Develop a high taxonomic resolution model for the identification of graptolite fossils

Trained using the largest professional single fossil image dataset to date

It outperforms taxonomists in identifying complex graptolite morphological details

Niu et al., iScience *27*, 108549
January 19, 2024 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.108549

iScience

CellPress OPEN ACCESS

Article

Automated graptolite identification at high taxonomic resolution using residual networks

Zhi-Bin Niu,[1,2,3,*] Si-Yuan Jia,[1] and Hong-He Xu[2,*]

SUMMARY

Graptolites, fossils significant for evolutionary studies and shale gas exploration, are traditionally identified visually by taxonomists due to their intricate morphologies and preservation challenges. Artificial intelligence (AI) holds great promise for transforming such meticulous tasks. In this paper, we demonstrate that graptolites can be identified with taxonomist accuracy using a deep learning model. We construct the most sophisticated and largest professional single organisms image dataset to date, which is composed of >34,000 images of 113 graptolite species annotated at pixel-level resolution to train the model, develop, and evaluate deep learning networks to classify graptolites. The model's performance surpassed taxonomists in accuracy, time, and generalization, achieving 86% and 81% accuracy in identifying graptolite genus and species, respectively. This AI-based method, capable of recognizing minute morphological details better than taxonomists, can be integrated into web and mobile apps, extending graptolite identification beyond research institutes and enhancing shale gas exploration efficiency.

Accurate and efficient species identification of graptolite, an extinct group of globally distributed and rapidly evolved macrofossil,[1–5] is indispensable in research on evolution and biostratigraphy and assisting global shale gas exploration.[4] There were 102 Ordovician and Silurian graptolite species selected as biozones for determining global rock age and regional correlation, and contributing to understanding the evolutionary pattern of ancient life,[2] and 16 graptolite species as "gold calipers" to locate shale gas favorable exploration beds (FEBs) in China[4](Figure S2). Graptolite identification has to date been an exclusively human task in paleontology. We developed a computational approach that decreases paleontologist workloads and enables shale gas specialists to accurately identify graptolite species and find shale gas FEBs in seconds. By creating a unique dataset of authoritative, taxonomical, and pixel-level annotated graptolite images, we are able to train the world's first deep neural-network-based taxonomist-level, macrofossil identification model.

Although, computer-aided fossil identification (CAFI) software was introduced as early as 1980s.[6] To our knowledge, no CAFI assistance tool has been used for practical usage due to the lack of generalization capabilities, primarily caused by insufficient labeled fossil images followed by limited performance of the shallow machine learning algorithms. During this formative stage, a plethora of methodologies emerged that amalgamated manual feature extraction with shallow learning approaches such as neural networks and support vector machine algorithms, among others, gaining significant attention.[7–11] Zhang et al.[7] utilized image analysis of texture and shape characteristics for the classification of five taxa of New Zealand pollen types. Despite these developments, fossil identification systems necessitated manual intervention for the precise selection of sample parts for analysis. Ranaweera et al.[9] engineered a computer-aided system for foraminifera identification, tested on a dataset of 244 specimens from five genera. The images of these specimens were normalized for position, rotation, and scale and categorized based on their similarity in a canonical space. Rodriguez-Damian et al.[10] designed an automatic classification system for three prevalent species from the Urticaceae family. The system utilized Hough transform for pollen position detection, applied "snakes" for pollen contour extraction, computed the shape and texture features of the pollen, and subsequently classified them. France et al.[11] amassed image sets from three types of pollen slides, and with the aid of artificial neural networks, they devised a system for the classification and identification of pollen fossils. At this stage, the limited generalizability of machine learning within a narrow category range further impeded the practical implementation in paleontology.

It was until recent years, the deep learning methodology reformed the prediction framework, showed promising results in medical diagnoses[12–15] and many other fields,[6,10,14,16–25] and illuminated a possibility of industrial-level CAFI. Initially, identification of microfossils, such as foraminifer,[16] pollen grains,[17–19] and micro-objects from thin sections,[6,20,21] saw a sprout of interdisciplinary trend of paleontology and artificial intelligence.[26,27] Among them, Hsiang et al.[16] designed a supervised species-level classifier for a 34-species planktonic foraminifer identification task using the VGG16 base model. Kong et al.[18] employed the VGG19 model as a foundational network to propose a framework for

[1] College of Intelligence and Computing, Tianjin University, Tianjin 300354, China
[2] State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology and Centre for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Nanjing 210008, China
[3] Lead contact
*Correspondence: zniu@tju.edu.cn (Z.-B.N.), hhxu@nigpas.ac.cn (H.-H.X.)
https://doi.org/10.1016/j.isci.2023.108549

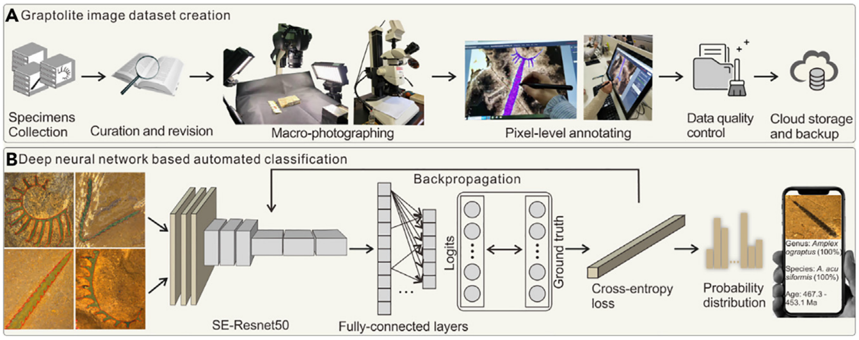


Figure 1. The deep learning approach workflow

(A) The process of creating the dataset. To generate the dataset, we meticulously curated and reviewed graptolite specimens, specifically selecting species with biostratigraphy and application significance. Each specimen was macro-photographed using a single-lens reflex camera and microscope and then pixel-level annotated based on the graptolite body's contours. Following professional revision and cleaning of the annotated images, the entire dataset was then uploaded and securely stored in our cloud server for future reference.

(B) The process of the supervised deep-neural-network-based automated classification. After thorough evaluation of 12 state-of-the-art deep neural network models for our dataset (see Figure S4; Table 1), we selected the squeeze-and-excitation network built on Resnet-50,[24] and it delivered superior performance. Our offline trained model can be conveniently deployed on the cloud server, enabling remote classification via a web interface or mobile device. However, it is important to note that for more precise classification, end-users must annotate the contours of the fossil body.

fossil pollen grain identification. This framework, which incorporated a spatially aware exemplar-based coding approach, was used for species classification, specifically distinguishing between 3 types of fossil spruce pollen. Liu et al.[21] evaluated and compared the performance of four deep convolutional neural network architectures on a microfacies image dataset comprised of 22 groups of fossil and abiotic grains. Wang et al.[28] developed a temporal convolutional neural network for a dataset of five distinct species, emphasizing the importance of high accuracy in brachiopod fossil identification. Pires et al.[29] introduced a transfer learning technique for the classification of a dataset comprising eight distinct fusulinid genera. Liu et al.[30] evaluated the results of three typical deep CNN (convolutional neural network) architectures on a large-scale fossil dataset of more than 50 clades including invertebrates, vertebrates, plants, microfossils, and fossil traces from five hyper-clades. Hou et al.[31] proposed a multi-perspective framework that uses original, gray, and skeleton images of each fossil for training. The final decision is made via soft voting. This was tested on a dataset of 2,400 fusulinids from 16 genera within 6 families for genus recognition. The multi-view ensemble framework enhanced performance across various base models.

The existing body of these work is either mainly conducted on small-scale dataset or focus on microfossils,[7,16,17,19,32] thus not generalizing well to large-scale macrofossils, the more challenging direct evidence of evolution. Microfossils, before photographing, are primarily treated by dissolving the impurities and surrounding rocks, so they are relatively intact, isolated, uniform textured, and rarely overlay, and can be massively photographed cost-effectively using automated slide scanners, providing ample data to feed deep learning models. However, the automated identification of general macrofossils, e.g., animal bone and plant leaf fossils, is much more challenging. This is because they are often with incomplete preservation, clutter and occlusion of multi-fossil remains, a lack of contracts between the fossil remains and surrounding rocks, color confusion caused by mineralization, and varying developmental states. Besides, their specimens and images are more difficult to obtain. Thus, compared with microfossils, obtaining large quantities of high-quality macrofossil images is more difficult and challenging.

This issue is particularly apparent in the study and identification of graptolite species, which has historically relied on human visual inspections of specimens, rather than more modern techniques involving chemistry, spectrum, molecular biology, genomics, or histology. Graptolite organisms have lost their soft tissues during fossilization and most are found flattened and carbonized. Paleontology taxonomists identify graptolite based on examinations using a hand lens in the fieldwork; they also determine accurate species after removing the coverings and applying diagnostic measurements aided with microscope in the laboratory. The identification requires lengthy specific training based on mass observations of limited specimens that are mainly housed in academic institutes, and consequently, the number of qualified graptolite taxonomists cannot meet the massive and urgent demand required for geologic survey and shale gas exploration.[3,4] Shale gas companies have had to deliver the specimens to research institutes for accurate identification, which is a time-consuming and costly procedure.

We fill the gap by creating a unique large dataset of annotated macrofossil graptolite images and training a state-of-the-art deep learning model (Figure 1). The limited automated macrofossil identification has usually required formatted traits[6] with landmarks and even manual measurements; by contrast, our system requires no hand-crafted features, only pixels. We demonstrate generalizable classification ability with the first macrofossil dataset of 34,620 pixel-level annotated graptolite images. We highlight the development of the deep-learning-based approach and its ability to outperform graptolite taxonomists in image-based identification in terms of accuracy, time cost, and generalization capability.

Creating a taxonomist-annotated, multi-modal dataset for graptolite fossils

Our dataset comprises a meticulous collection of 1,565 specimen pieces, obtained from 154 representative geological sections distributed throughout China (Figure 2).[4,33] These specimens were carefully curated and belong taxonomically to 113 graptolite species or subspecies of

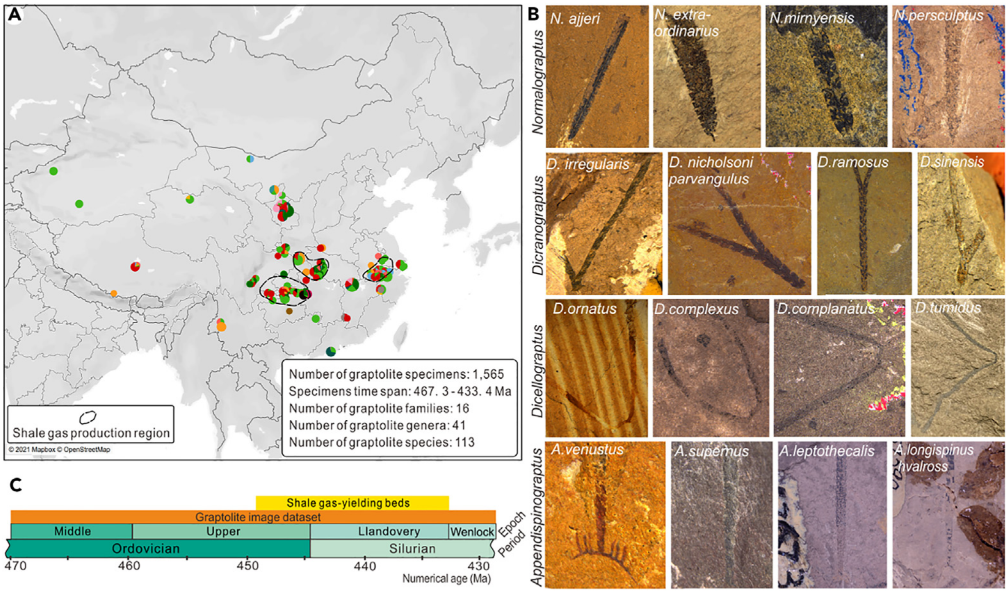


Figure 2. Geological significance and example graptolite images of the dataset

(A) Our dataset includes information regarding the geographic distribution and statistical results of the graptolite specimens. Each locality is denoted by a pie chart where each color represents a graptolite family of the order Graptoloidea. The size of the pie sector reflects the number of specimens for each family, with the radius of the pie chart directly proportional to the total number of specimens from the same locality. Notably, main areas of shale gas production are circled by dashed lines for ease of reference.

(B and C) We present a set of example graptolite images, with each row displaying graptolites of the same genus but different species. Our images illustrate the high intra-class and low inter-class variance characteristics of graptolites. (C) It is worth noting that these graptolite species span from the Middle Ordovician to the Wenlock of the Silurian (black line section), representing all shale-gas-yielding beds from southern China (yellow bar).

41 genera and 16 families of the order Graptoloidea. This comprehensive dataset includes 22 graptolite biozones from the Dapingian stage of the Ordovician to the Homerian stage of the Silurian and 16 ″gold calipers" of shale gas FEBs in cases of 20 to 80 m thick graptolite shale in China, which are frequently used for geological age determination and shale gas FEB indication (Figures 2 and S2). Ground-truth labels for the supervised training were generated by incorporating revision suggestions from distinguished paleontologists, adding to the taxonomical authority of the dataset.

A total of 40,597 images were captured for every specimen at different focusing angles and scales. Following professional revision, unclear and questionable taxonomical status images were removed. Every image was annotated at the pixel level according to the contours of the graptolite's body (Figure 1A). Ultimately, our deposit contains 37,588 pixel-level annotated images, segregated for training, validation, and testing (Figure S3). We employ the COCO Annotator tool[34] for pixel-level annotation to produce a binary mask for each image. This mask, in conjunction with the original image, is subsequently input into the model. The application of pixel-level annotations serves to augment image quality and temper background noise to a certain degree, thereby facilitating a more robust feature extraction process.

Deep-neural-network-based automated taxonomical classification

We chose to use deep neural network as the supervised learning classifier for the deep-learning-based automated taxonomical classification. We conducted experiments and compared the performance of 10 state-of-the-art deep neural network models including CNN-based ones,[24,35–44] the popular ViT,[45] and Swin Transformer[46] on the graptolite image dataset (Table 1; Figure 3) and selected to use the squeeze-and-excitation networks[35,47] built on Resnet-50 (SE-Resnet50) as the base model for its optimal performance on our dataset. The SE-Resnet50[24] introduces excitation and squeezing mechanisms. The squeeze and excitation block, a computational unit in this framework, condenses global spatial information into the channel descriptor via a squeezing operation. Conversely, the excitation operation aims to accurately capture channel dependencies. SE-Resnet utilizes a gating mechanism with a sigmoid activation function, encouraging non-linear channel interactions and enabling the emphasis of multiple channels.

The architecture also uses a transfer learning manner to train the model (see Figure 1B). The using of pre-trained weights leveraging the natural image features learned by ImageNet[48,49] accelerated convergence, especially in the early stages of training. We then fine-tuned the parameters across all layers of our dataset. During the training, we used the stochastic gradient descent optimizer, with a momentum of 0.9 and weight decay of 5×10^{-4}. The initial learning rate was set to 0.001 and decreased following a cosine annealing schedule. It slowly reduced to approximately 2.7×10^{-8} in the last epoch. We used the cross-entropy loss function to calculate the model error. As in Figure 1, the gradients were computed using backpropagation of the error algorithm and the parameters updated using the stochastic gradient descent

Table 1. Performance evaluation results

	Cross-validation results				
Experiments	Valid 1	Valid 2	Valid 3	Average	Gold-standard test
VGG16_bn[39]	63.684%	67.690%	68.281%	66.552%	70%
Resnet50[35]	67.781%	71.872%	71.896%	70.516%	78%
Inceptionv3[40]	67.748%	71.602%	71.965%	70.438%	77%
SE-Resnet50[24]	67.847%	72.040%	71.726%	70.538%	81%
DFL (vgg16_bn)[38]	67.519%	69.511%	71.180%	69.403%	75%
NTS (resnet50)[41]	66.077%	70.152%	69.679%	68.969%	75%
PC (resnet50)[44]	67.060%	71.096%	70.020%	69.392%	78%
MC-loss (resnet50)[42]	65.045%	68.916%	68.405%	67.455%	81%
B-CNN (vgg16)[36]	61.422%	62.226%	61.903%	61.850%	71%
ViT-Base[45]	65.475%	69.723%	70.985%	68.728%	75%
EfficientNet-b5[43]	67.650%	71.747%	71.518%	70.305%	80%
Swin Transformer[46]	60.441%	61.245%	63.179%	61.625%	68%

Three-fold cross-validation and test set accuracy using 12 methods. In the 3-fold cross-validation, we have 34,620 images, with the 3-fold training, and validation image numbers are (31569, 3051), (31655, 2965), and (31688, 2932). We observe the SE-Resnet50 gives the best performance. We also give golden standard test on the "testset-99" in the table.

optimization algorithm. After training each epoch, we fed all test images into the model, performed forward propagation to calculate accuracy, and saved the model parameters with the highest accuracy as our final training model.

Cross-validation for evaluating generalizability in genus- and species-level graptolite identification

We first conducted a cross-validation to evaluate the hyperparameters and generalizability. As the number of specimens of each species ranged from 4 to 20, we chose to use 3-fold cross-validation to evaluate the generalizability performance, so that the cross-validation could cover all species. The validation image set was carefully chosen so that every specimen-related image was completely removed from the training set, and thus there was no intersection between the training and validation test sets. The validation set included 15 to 80 randomly chosen images of each species. In total, there were 34,620 images for training and validation; each iteration had a different number in the validation set (Table 1). We conducted two cross-validations on different taxonomical levels. On the genus and species levels, the deep-learning-based approach achieved 81.8% and 70.538% overall accuracy, respectively (Table 2). These results indicate that the deep learning approach is a feasible means of learning the ultra-fine-grained features of graptolites.

Comparison of the deep learning approach and graptolite taxonomists

The cross-validation results were promising but still inconclusive, as the results depended on a particular random choice for the pairing of training and validation sample sets. To test the generalization ability of the model and conclusively evaluate our deep learning approach, we further compared the performances of our model with graptolite taxonomists on a carefully chosen "golden standard" test set.

We chose 100 images of 35 graptolite species to conduct the golden standard test. These species are directly related to dating sediments and locating shale gas FEBs during the mining. They consist of all 22 graptolite biozone species from the Middle Ordovician to the Wenlock of Silurian and 16 indicator species widely used in shale gas exploration in China (Figure S2). Thirty-five species of the golden standard test set are from 99 pieces of specimens, which are dubbed as "collection-99" specimen dataset. There are totally 2,968 photographs taken from the collection-99; 100 images selected into the golden standard test set are well focused and friendly, showing morphological characters of graptolite for identification. The rest of the images taken from the collection-99, 2,868 images, are removed from the training set. Finally, we use 31,652 images to train the deep-learning-based model.

We have invited 21 graptolite taxonomists to provide a baseline of human performance. All experts are working on graptolite-related education, research, and engineering applications at universities, institutes, or oil/gas companies around the world. They have 5 to 30 years of fossil (especially graptolite) identification experience. During the golden standard test, experts were asked to identify the graptolite and record the genus and species names, time cost, and comments. They could view and refer any literatures, even internet searching, freely.

We use identification accuracy, time cost, and micro-average sensitivity-specificity curves (i.e., *receiver operating characteristics curves, ROC) to evaluate the performances of the deep learning approach and human's classification ability* (Figure 3). The ROC is a widely used method in machine learning to compare the strengths and weaknesses of different models. The x axis is defined as false-positive rate or 1-specificity, and the y axis is defined as true positive rate or sensitivity. The deep learning approach achieves 86% and 81% levels of accuracy for genus and species identification, respectively, in seconds, whereas those of the graptolite taxonomists are only 59.5% (95% confidence

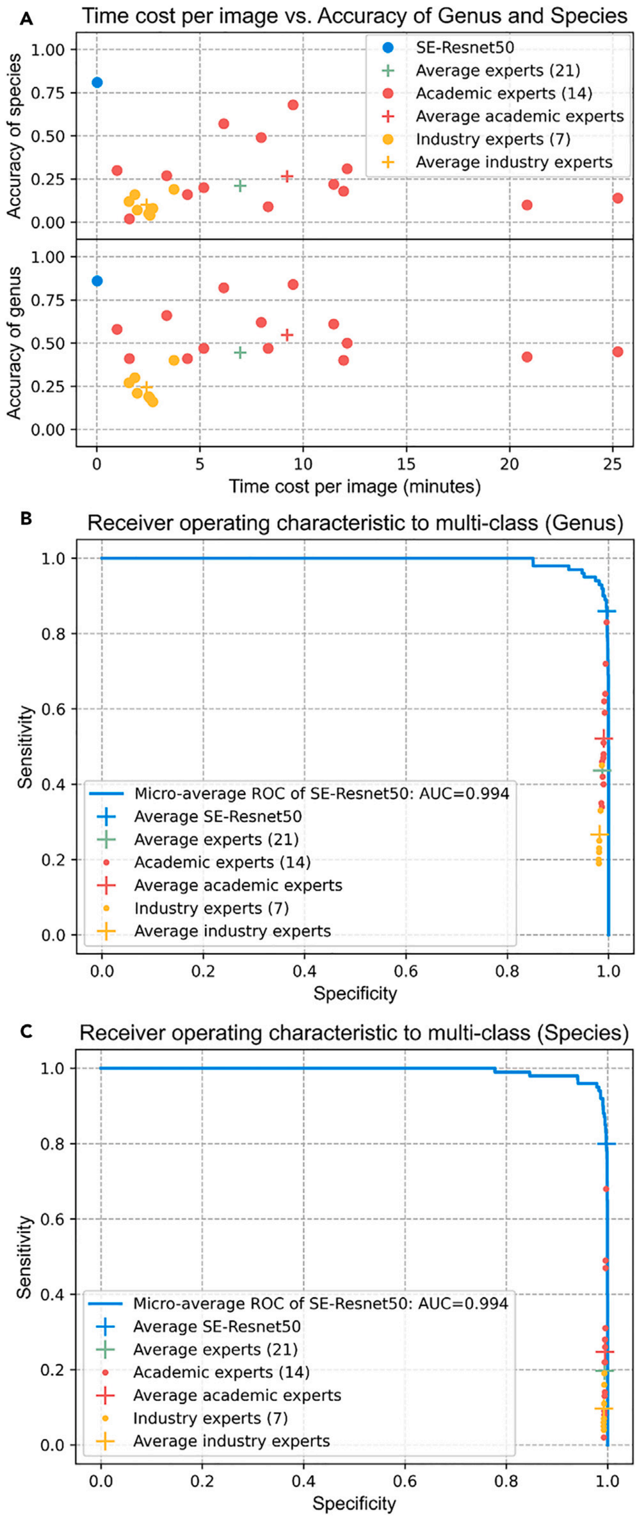


Figure 3. Performances of deep learning approach and experts on gold-standard graptolite classification test

(A) Time cost—genus and species classification accuracy plot. It shows that the deep learning approach (SE-Resnet50) significantly outperforms the experts in genus and species graptolite classification accuracies and the efficiency. The average accuracy of academic experts is higher than that of industry experts at both two levels, indicating that the academic experts can make more reliable predictions, but they usually take longer time.

Figure 3. *Continued*

(B) Comparison of the micro-average ROC curve of SE-Resnet50 with the micro-average experts at genus level. The average point of SE-Resnet50 is located at the upper than that of all experts, indicating that the classification performance of the model is evidently better than that of all experts. The average point of academic experts is located at the upper right of that of industry experts, which means that academic experts have better classification ability.

(C) The comparison of the micro-average ROC curve of SE-Resnet50 with the micro-average experts at species level. The average point of SE-Resnet50 is located at the upper right of all experts, which indicates that the classification performance of the model is obviously better than that of all experts. The average point of academic experts is located at the upper right of that of industry experts, which means that academic experts have better classification ability.

interval [CI] 52.6%, 66.5%) and 45.2% (95%CI 38.9%, 51.5%), taking on average 4.8 min (95%CI 4.2 min, 5.4 min) (Table 3). For further details, in genus identification, the academic experts achieve 68.0% (95%CI 60.0%, 73.5%) and industry experts achieve 45.3% (95CI 33.0%, 57.6%); whereas in species, the numbers are 50.8% (95CI 44.7%, 56.8%) and 35.8% (95CI 21.0%, 50.6%), respectively (Figure 3A).

We also provide micro-average sensitivity-specificity curves to quantitatively measure the classification performances of the deep-learning-based model and experts (Figures 3B and 3C, STAR methods). In genus identification, the experts achieve on average sensitivities of 43.67% (95%CI 35.7%,51.7%) and specificity of 98.8% (95%CI 98.6%, 98.9%) and in species level, 19.7% (95%CI 11.9%, 27.5%) and 99.4% (95%CI 99.3%, 99.5%), respectively. The deep learning model has 86% sensitivity and 99.7% specificity (genus level) and 80% sensitivity and 99.8% specificity (species level). Overall, the deep learning model outperforms the experts in genus 0.89% (95CI 0.68%, 1.1%) specificity and 42.3% (95%CI 34.4%, 50.3%) sensitivity and in species 0.40% (95CI 0.34%, 0.47%) specificity and 60.3% (95%CI 52.5%, 68.1%) sensitivity.

The AUC (area under the ROC curve) illustrates how much the model is capable of distinguishing between classes. The micro-average AUC value is near 1.0 (99.4%), indicating that the deep learning model is generally able to precisely identify the graptolite taxonomy. The human machine contest suggests that our deep learning approach is significantly superior to academic graptolite taxonomists and industry specialists with over 5 years experience; thus, therefore, the relative expensive and time-consuming graptolite identification in the fieldwork would benefit more from the potential CAFI software.

Interpretation through visualization of the deep learning graptolite classification internals

We examined the features learned by the deep neural network through uniform manifold approximation and projection (UMAP)[50] (Figure 4A). Each thumbnail image in the two-dimensional space represents a graptolite projected from the 100,352 dimensional output of the CNN's last hidden layer.[51] The proximity in low-dimensional UMAP space reveals nine major graptolite morphotype clusters[52] (Figure 4A, line drawings showing the major morphotypes of the clusters). In the middle of the map is the dominant cluster, namely the scandent morphotype, including almost all biserial diplograptids. The upper left portion of the map is the V-shaped graptolite morphotype, including all graptolites with typical U- or V-shaped tubarium. We also found that several outliers on the right part of the embedding had distinctive morphology, such as graptolite with spiral or helical shapes (Figures 4A; Table 4).

We further utilized gradient-weighted class activation mapping (Grad-CAM)[27,53] to visualize and interpret the prediction decisions made by the deep neural network (Figure 4C). The attention map indicates the diagnostic (semantic) image regions by projecting back the weights of the output layer on to the convolutional feature maps, localizing class-discriminative regions that influence the deep-learning-based approach and showing the differences between expert and the deep learning model dealing with graptolite images. The morphology of proximal end (sicula and one or two thecae, the area in the dashed line box in Figure 4B) is critical to identify graptolite genus and superior genus taxa, and measurements are necessary to identify some graptolite species, especially within the genus.[1] Actually, proximal ends are not either readily preserved in every specimen or fully given in our test set images, to which, as a consequence, identification is impossible to experts, except graptolite of specialized type, as shown in the third line of Figure 4B. Additionally, the deep learning model recognizes visual patterns of images as important class-discriminative information better than experts. For example, graptolite with only regular thecae is also identifiable to the deep learning model but not to experts. The attention map demonstrates that the superhuman performance of the deep learning approach is able to recognize morphological nuances of graptolite species without requiring measurements, comprehensively incorporating ultra-fine-grained details, including "traits agnostic" of graptolite.

Table 2. Evaluation results at genus and species levels

SE-resnet50[24]	Valid 1	Valid 2	Valid 3
Training image number	31569	31655	31688
Validation image number	3051	2965	2932
Genus accuracy (41)	79.49%	82.43%	84.60%
Genus average accuracy	81.8%		
Species accuracy (113)	67.847%	72.040%	71.726%
Species average accuracy	70.538%		

Cross-validation for genus and species classification accuracy using SE-resnet50.

Table 3. Expert performance on golden standard test set

	Genus accuracy	Species accuracy	Average time cost (min)
Academic experts (14)	66.5% (95%CI 60.5%, 70.8%)	50.4% (95%CI 44.7%, 56.0%)	5.4 (95%CI 4.9, 5.8)
Industry experts (7)	45.29% (95%CI 33.0%, 57.6%)	34.9% (95%CI 22.7%, 47%)	4.1 (95%CI 3.0, 5.2)

DISCUSSION

Our study effectively demonstrates the powerful application of deep learning in macrofossil graptolite identification. This deep-learning-based approach opens up new possibilities for species-level graptolite identification, exhibiting a performance level that surpasses that of taxonomists, sufficient generalization, and interpretable semantics. Graptolite shale forms over 9% of the hydrocarbon rocks globally (Figure S1) and provides over 61.4% of natural gas for China. This deep learning approach holds the potential to revolutionize macrofossil identification and, when combined with smart mobile devices, can significantly improve the efficiencies of geological surveys and shale gas exploration. Our research paves the way for a transdisciplinary future integrating paleontology, artificial intelligence, and the oil/gas industry. Although our method is currently constrained by limited data sources in China, we are exploring avenues to overcome this limitation.

At this juncture, certain challenges remain. A primary difficulty lies in the scarcity of certain graptolite species that are especially challenging to identify, a scenario that extends to many other fossil types within the field of paleontology. However, we believe that these hurdles can be overcome with strategic improvements in our data collection and model training methods. One promising avenue to explore is the introduction of multimodal data into our deep learning model training. Information such as geographic location and geological age, which often show strong correlations with graptolite species differentiation, could significantly enhance the model's identification accuracy. In addition, we are researching ways to expand our data sources beyond China, potentially incorporating a global range of fossil data. This could greatly enhance the diversity of our dataset, thus improving our model's ability to generalize and accurately identify a broader spectrum of graptolite species. Furthermore, future research endeavors may focus on integrating the expertise and knowledge of paleontologists into our deep-learning-based approach. This melding of human expertise with artificial intelligence could lead to an even more robust and reliable identification system. We are actively working on these improvements and look forward to presenting our progress in the near future.

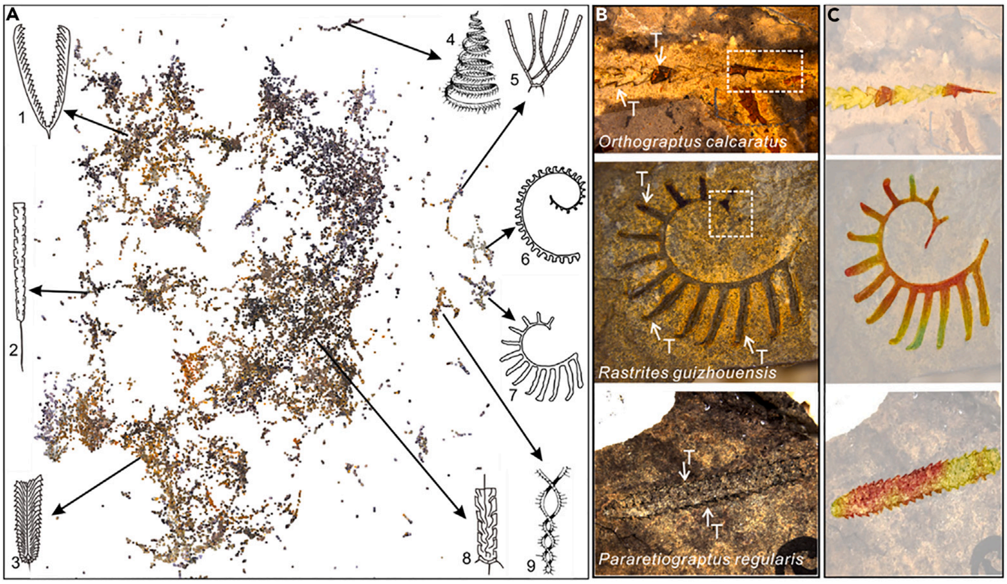


Figure 4. Visualization of our graptolite image dataset structure and attention map of example graptolite images

(A) UMAP embeddings of the last hidden layer representations of our dataset. Every image is projected to a thumbnail on the two-dimensional scatterplot, visually revealing at least nine graptolite morphotype clusters (Table 4).

(B) Three example images from the gold-standard test set, clearly showing graptolite thecae (short or long prominence structures, arrowed, T) and proximal ends (dashed box) of three species, providing good and sufficient morphological information for expert's classification. No scale bar is given in every image, and it is impossible to measure fossil remains, such might affect the species classification of some graptolite. The proximal end is now visible in the lowermost graptolite specimen image, making classification impossible to experts but feasible to the deep learning model.

(C) Three images with highlighted discriminative visual regions generated through CAM. We use a rainbow color map to reweight and render the saliency regions used for image classification. Red regions correspond to high score for class; the next are yellow and green. The high-score regions of the deep learning model coincide with critical areas of expert's classification, e.g., the proximal end, whereas to the graptolite without showing the proximal end, the deep learning model recognizes regular thecae as visual pattern and better identify the graptolite than experts.

Table 4. Nine graptolite morphotypes, clustered from our dataset using UMAP

ID	Graptolite morphotypes	Description
1	V-shaped	Graptolite tubarium with typical V or U structure, including species of *Didymograptus* and *Dicranograptus*, some species of *Dicellograptus* and *Jiangxigraptus*, and most species of *Appendispinograptus*.
2	Thin-shaped	Graptolite tubarium with extremely large ratio of length/width, such as *Normalograptus angustus* and *Coronograptus cyphus*.
3	Flat-shaped	Graptolite tubarium with much smaller ratio of length/width (relative to the type 2), such as *Cardiograptus amplus*, *Phyllograptus anna*, and *Peseudotrigonograptus ensiformis*.
4	Spiral-shaped	Spiral morphotype, as typical as *Spirograptus turriculatus*.
5	Multi-ramose-shaped	Multiramous graptolites, such as species of *Tangyagraptus* and *Pterograptus*.
6	Dorsal-curvature-shaped	Graptolites with proximally accentuated dorsal curvature and isolated thecae, including only *Demirastrites*.
7	Open-spiral-shaped	Graptolites with an open spiral torsion form and isolated thecae, including only *Rastrites*.
8	Scandent-shaped	Scandent morphotype, including almost all biserial diplograptids.
9	Double-helix-shaped	Graptolites with spars double helix shape (as ”8”), such as *Jiangxigraptus spirabilis*.

This research signifies an important step forward, but there is still much to be done as we continue to push the boundaries of what is possible in combining deep learning with paleontological research.

Limitations of the study

The identification process may be impeded by the constraints of limited data sources and incomplete or fragmented samples. Besides, the scarcity of certain graptolite species presents the problem of category imbalance, which could potentially influence the outcomes. We are proactively addressing these issues to ensure the robustness and reliability of our methodologies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABITILY
  - ○ Lead contact
  - ○ Material availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Dataset
  - ○ Data preparation
  - ○ Training procedure and settings
  - ○ Sensitivity-specificity curve
  - ○ Confusion matrix
  - ○ Uniform manifold approximation and projection
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108549.

ACKNOWLEDGMENTS

We thank Prof. Renbin Zhan, Prof. Yuandong Zhang, Prof. Xunlai Yuan, Prof. Maoyan Zhu, Prof. Yi Wang, Prof. Bing Huang, Prof. Bo Wang, Dr. Qing Chen, Dr. Lixia Li, and Dr. Xuan Ma, Nanjing Institute of Geology and Palaeontology (NIGP), Chinese Academy of Sciences (CAS); Dr. Lucy Muir, National museum of Wales (UK); Dr. Petr Kraft, Charles University (Czech); Dr. Petr Štorch, Institute of Geology of the Czech Academy of Sciences (Czech); David Loydell, University of Portsmouth (UK); Dr. Wenhui Wang, Central South University (China); Dr. Yanyan Song, University of Chinese Academy of Sciences; Dr. Ming Li, Chinese Academy of Geosciences; Drs Xin Wang and Jian Wang, Xi'an Geological Survey Center, China Geological Su1rvey; Hongyan Wang, Feng Liang, Qun Zhao, and Shasha Sun, National Energy Shale Gas R & D

(Experiment) Center, Beijing, China; Mr. Jie Tian, China Petroleum Pipeline Bureau, Langfang, China; Prof. Min Zhu and Dr. Zhaohui Pan, Institute of Vertebrate Paleontology and Paleoanthropology, CAS; and a number of anonymous experts, for comments and suggestions; Prof. Wenwei Yuan and Daojun Yuan, NIGP, CAS, for the courtesy in graptolite specimens examination; Mr. Zhiqiu Zhang, Tianjin University, for testing the software; Xiaojing Tong, Yutong Sun, Xiaoyan Dong, Shuangshuang Song, Danni Fan, Yanyan Zhu, and Qing Xia, NIGP, CAS, for preparations of specimens image data; Yansen Chen, NIGPAS, CAS, for technician supports.

This research was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (grant number: XDA19050101, XDB26000000) and the National Natural Science Foundation of China (grant number: 61802278).

AUTHOR CONTRIBUTIONS

Z.N. and H.X. conceived of the whole project. H.X., S.J., and Z.N. wrote and discussed the manuscript. S.J. and Z.N. realized the code and software. H.X. and Z.N. conducted the experimental analyses.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 11, 2023
Revised: August 23, 2023
Accepted: November 20, 2023
Published: November 23, 2023

REFERENCES

1. Maletz, J. (2017). Treatise Online no. 88: Part V, Second Revision, Chapter 13: The History of Graptolite Classification. Treatise Online.
2. Ogg, J. (2020). Geomagnetic polarity time scale. In Geologic Time Scale 2020 (Elsevier), pp. 159–192.
3. Podhalańska, T. (2013). Graptolites–stratigraphic tool in the exploration of zones prospective for the occurrence of unconventional hydrocarbon deposits. Przeglad Geol. *61*, 621–629.
4. Caineng, Z., Jianming, G., Hongyan, W., and Zhensheng, S. (2019). Importance of graptolite evolution and biostratigraphic calibration on shale gas exploration. China Petroleum Exploration *24*, 1–6.
5. Zou, C., Dong, D., Wang, Y., Li, X., HUANG, J., Wang, S., Guan, Q., ZHANG, C., Wang, H., Liu, H., et al. (2015). Shale gas in China: Characteristics, challenges and prospects (I). Petrol. Explor. Dev. *42*, 753–767.
6. Swaby, P.A. (1990). Integrating Artificial Intelligence and Graphics in a Tool for Microfossil Identification for Use in the Petroleum Industry (Citeseer), pp. 203–218.
7. Zhang, Y., Fountain, D.W., Hodgson, R.M., Flenley, J.R., and Gunetileke, S. (2004). Towards automation of palynology 3: pollen pattern recognition using Gabor transforms and digital moments. J. Quat. Sci. *19*, 763–768.
8. MacLeod, N., O'Neill, M., and Walsh, S.A. (2016). A comparison between morphometric and artificial neural network approaches to the automated species recognition problem in systematics. In Biodiversity databases (CRC Press), pp. 37–62.
9. Ranaweera, K., Harrison, A.P., Bains, S., and Joseph, D. (2009). Feasibility of computer-aided identification of foraminiferal tests. Mar. Micropaleontol. *72*, 66–75.
10. Rodriguez-Damian, M., Cernadas, E., Formella, A., Fernandez-Delgado, M., and De Sa-Otero, P. (2006). Automatic detection and classification of grains of pollen based on shape and texture. IEEE Trans. Syst. Man Cybern. C *36*, 531–542.
11. France, I., Duller, A., Duller, G., and Lamb, H. (2000). A new approach to automated pollen analysis. Quat. Sci. Rev. *19*, 537–546.
12. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.
13. Richens, J.G., Lee, C.M., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. Nat. Commun. *11*, 3923–3929.
14. Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., and Potter, H. (2018). Deep neural network improves fracture detection by clinicians. Proc. Natl. Acad. Sci. USA *115*, 11591–11596.
15. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature *577*, 89–94.
16. Hsiang, A.Y., Brombacher, A., Rillo, M.C., Mleneck-Vautravers, M.J., Conn, S., Lordsmith, S., Jentzen, A., Henehan, M.J., Metcalfe, B., Fenton, I.S., et al. (2019). Endless Forams:> 34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. Paleoceanogr. Paleoclimatol. *34*, 1157–1177.
17. Mander, L., Li, M., Mio, W., Fowlkes, C.C., and Punyasena, S.W. (2013). Classification of grass pollen through the quantitative analysis of surface ornamentation and texture. Proc. Biol. Sci. *280*, 20131905.
18. Kong, S., Punyasena, S., and Fowlkes, C. (2016). Spatially Aware Dictionary Learning and Coding for Fossil Pollen Identification, pp. 1–10.
19. Romero, I.C., Kong, S., Fowlkes, C.C., Jaramillo, C., Urban, M.A., Oboh-Ikuenobe, F., D'Apolito, C., and Punyasena, S.W. (2020). Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. Proc. Natl. Acad. Sci. USA *117*, 28496–28505.
20. Kopperud, B.T., Lidgard, S., and Liow, L.H. (2019). Text-mined fossil biodiversity dynamics using machine learning. Proc. Biol. Sci. *286*, 20190022.
21. Liu, X., and Song, H. (2020). Automatic identification of fossils and abiotic grains during carbonate microfacies analysis using deep convolutional neural networks. Sediment. Geol. *410*, 105790.
22. Wu, S., Chang, C.-M., Mai, G.-S., Rubenstein, D.R., Yang, C.-M., Huang, Y.-T., Lin, H.-H., Shih, L.-C., Chen, S.-W., and Shen, S.-F. (2019). Artificial intelligence reveals environmental constraints on colour diversity in insects. Nat. Commun. *10*, 4554.
23. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, f. (2019). Deep learning and process understanding for data-driven Earth system science. Nature *566*, 195–204.
24. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation Networks, pp. 7132–7141.
25. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.
26. van der Valk, T., Pečnerová, P., Díez-Del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J.A., Dehasque, M., Sağlıcan, E., et al. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. Nature *591*, 265–269.
27. Cuthill, J.F.H., Guttenberg, N., and Budd, G.E. (2020). Impacts of speciation and extinction measured by an evolutionary decay clock. Nature *588*, 636–641.
28. Wang, H., Li, C., Zhang, Z., Kershaw, S., Holmer, L.E., Zhang, Y., Wei, K., and Liu, P. (2022). Fossil brachiopod identification using a new deep convolutional neural network. Gondwana Res. *105*, 290–298.
29. Pires de Lima, R., Welch, K.F., Barrick, J.E., Marfurt, K.J., Burkhalter, R., Cassel, M., and Soreghan, G.S. (2020). Convolutional neural networks as an aid to biostratigraphy and

micropaleontology: a test on late Paleozoic microfossils. Palaios *35*, 391–402.

30. Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y., and Song, H. (2023). Automatic taxonomic identification based on the Fossil Image Dataset (> 415,000 images) and deep convolutional neural networks. Paleobiology *49*, 1–22.
31. Hou, C., Lin, X., Huang, H., Xu, S., Fan, J., Shi, Y., and Lv, H. (2023). Fossil Image Identification using Deep Learning Ensembles of Data Augmented Multiviewspreprint at arXiv *230*.
32. Punyasena, S.W., Tcheng, D.K., Wesseln, C., and Mueller, P.G. (2012). Classifying black and white spruce pollen using layered machine learning. New Phytol. *196*, 937–944.
33. Xu, H.-H., Niu, Z.-B., and Chen, Y.-S. (2020). A status report on a section-based stratigraphic and palaeontological database–the Geobiodiversity Database. Earth Syst. Sci. Data *12*, 3443–3452.
34. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014). Microsoft Coco: Common Objects in Context (Springer), pp. 740–755.
35. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition, pp. 770–778.
36. Lin, T.-Y., RoyChowdhury, A., and Maji, S. (2015). Bilinear CNN Models for Fine-Grained Visual Recognition, pp. 1449–1457.
37. Chen, Y., Bai, Y., Zhang, W., and Mei, T. (2019). Destruction and Construction Learning for Fine-Grained Image Recognition, pp. 5157–5166.
38. Wang, Y., Morariu, V.I., and Davis, L.S. (2018). Learning a Discriminative Filter Bank within a Cnn for Fine-Grained Recognition, pp. 4148–4157.
39. Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognitionpreprint at arXiv *1409*.
40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision, pp. 2818–2826.
41. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. (2018). Learning to Navigate for Fine-Grained Classification, pp. 420–435.
42. Chang, D., Ding, Y., Xie, J., Bhunia, A.K., Li, X., Ma, Z., Wu, M., Guo, J., and Song, Y.-Z. (2020). The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Trans. Image Process. *29*, 4683–4695.
43. Tan, M., and Le, Q. (2019). Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks (PMLR)), pp. 6105–6114.
44. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., and Naik, N. (2018). Pairwise Confusion for Fine-Grained Visual Classification, pp. 70–86.
45. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scalepreprint at arXiv *2010*.
46. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, pp. 10012–10022.
47. Juan, D.-C., Lu, C.-T., Li, Z., Peng, F., Timofeev, A., Chen, Y.-T., Gao, Y., Duerig, T., Tomkins, A., and Ravi, S. (2020). Ultra Fine-Grained Image Semantic Embedding, pp. 277–285.
48. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A Large-Scale Hierarchical Image Database (IEEE), pp. 248–255.
49. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. *25*.
50. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv *802*.
51. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning Deep Features for Discriminative Localization, pp. 2921–2929.
52. Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R.M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imag. *35*, 1285–1298.
53. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization, pp. 618–626.
54. Science. 60th Anniversary Celebration - Nanjing Institute of Geology and Palaeontology. https://www.sciencemag.org/advertorials/60th-anniversary-celebration-nanjing-institute-geology-and-palaeontology.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Graptolite image dataset	Nanjing Institute of Geology and Palaeontology (NIGP), Chinese Academy of Sciences (CAS)	https://zenodo.org/record/5205216
Software and algorithms		
PyTorch	FAIR	https://www.pytorch.org/
UMAP Algorithm	McInnes et al.[49]	https://github.com/YaleDHLab/pix-plot
Python version 3.7	Python Software Foundation	https://www.python.org
COCO Annotator	Lin et al.[34]	https://github.com/jsbroks/coco-annotator

RESOURCE AVAILABITILY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Zhi-Bin Niu (zniu@tju.edu.cn).

Material availability

This study did not generate new unique materials.

Data and code availability

- The graptolite images dataset is open accessed at http://fossil-explorer.com/ (https://doi.org/10.1101/2022.06.12.495851), and https://zenodo.org/record/5205216 which are publicly available as of the date of publication. For early testing evaluation, visit our model deployed on cloud servers at http://ai.fossil-explorer.com and are publicly available as of the date of publication.
- The COCO annotator code can be downloaded from https://github.com/jsbroks/coco-annotator. This annotation tool comes with object labeling and polygon annotation capabilities and provides the annotation data in COCO format. As for UMAP Visualization, one may access it from https://github.com/YaleDHLab/pix-plot. This tool projects the learned image features of the model onto a 2D space, making it possible to cluster similar images based on their features. The PyTorch deep learning framework can be accessed from https://www.pytorch.org/. It is a python-based machine learning library based on Torch, which facilitates powerful GPU-accelerated tensor calculation and deep neural network with automatic derivation.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The research used the PyTorch(PyTorch, RRID:SCR_018536) deep learning framework.

METHOD DETAILS

Dataset

Our graptolite images came from 1,565 pieces of specimens housed at the Nanjing Institute of Geology and Palaeontology (NIGP),[54] Chinese Academy of Sciences (CAS). The NIGP is the world's largest palaeontological research center and one of the top three specimen collection centers. With approximately 180 palaeontological researchers and laboratory technicians, and collecting over pieces of fossil specimens,[53] all specimens underwent professional curation and photography. Single-lens reflex digital Nikon D800E cameras with Nikkor 60 mm macro-lens and Leica M125 and M205C microscopes equipped with Leica cameras were used for capturing the images.

To generate our dataset, we hired 7 data entry clerks and 3 engineers specializing in specimen photography, pixel-level annotation, and data cleaning. The creation of the dataset took over 2 years, during which we took 40,597 images. This included 20,644 camera photos (each with a resolution of 4,912 × 7,360) and 19,953 microscope photos (each with a resolution of 2,720 × 2,048). After thorough examination of all photos, we removed 5,977 unclear or questionable taxonomical status images, and kept 34,620 valid images for validation, training, and testing.

Our dataset underwent a 3-fold validation process, where each iteration used between 2,932 and 3,051 images for validation while the rest were used for training. We subsequently separated the dataset into a train set comprising 31,652 images and a golden standard test set consisting of 100 images. The golden standard test set, representing specimens photographed using camera and microscope, was selected from 99 pieces of specimens.

Data preparation

Our data cleaning was overseen by graptolite palaeontologists from the NIGP at the CAS. During the data cleaning process, we removed graptolite images that were: 1) incorrect or poorly focused with very low contrast, 2) of poorly preserved specimens and substantially deformed, 3) incorrectly labeled and impossible to identify, or 4) showing graptolite bodies with confusing textural information that hindered the ability to see the morphological characteristics. In our study, we maintained a strict segregation between the training, validation, and test sets, ensuring that no image appeared in more than one set. This stringent approach was adopted to guarantee the integrity and reliability of our results. By ensuring non-overlap between the datasets, we eliminate any potential bias that could arise from the model having prior exposure to an image during the training phase, and subsequently encountering the same image in the validation or test set. This approach thus ensures that our model's performance evaluation is based exclusively on novel data, thereby enhancing the validity and reliability of our results. Images that were not in sharp focus or partially corrupted were solely included in the training set. This strategy was employed to improve the model's robustness to variations in image quality, without compromising the accuracy of our validation and testing procedures. Our data entry team annotated the graptolite bodies at a pixel-level using COCO annotator,[34] a web-based image annotation tool. All the images were stored and backed up in our cloud server for future use.

Training procedure and settings

To prepare the training images, we initially processed them through an augmentation and normalization operation. The pixel-level annotated graptolite images were resized to 480 × 480 pixels using interpolation and padded with zero background around each graptolite body. We kept the graptolite body aspect ratio constant as much as possible. Subsequently, we randomly cropped the images to 448 × 448 pixels, as this was an empirical trade-off between learning accuracy and information redundancy in fine-grained image classification.[34,36–38,47] We also augmented the images by flipping, rotating, and color jittering operations to enhance the model's ability to generalize and adapt to various test images. The images were then normalized with the *Z* score algorithm on each image channel, accelerating the optimization convergence speed. We followed a similar procedure for centre-cropping and normalizing test graptolite images.

To conduct the deep learning automated taxonomical classification, we opted to use Convolutional Neural Network (CNN) as the supervised learning classifiers. We experimented with and compared the performance of twelve state-of-the-art deep learning models on the graptolite image dataset (see Table 1; Figure 3). Ultimately, we selected the squeeze-and-excitation networks[35,47] built on Resnet-50 (SE-Resnet50) for its best performance as a representative deep learning approach. Our SE-Resnet50 architecture involved pre-training the deep neural network on approximately 1.3 million images from the ImageNet dataset.[48,49] The pre-trained model was expected to extract sufficient general image information for feature extraction.[14,44] In practice, we copied the parameters of the pre-trained model and fine-tuned them based on our target dataset, obtaining better results than from random initialization. The use of pre-trained weights may accelerate convergence, particularly in the early stages of training. We subsequently fine-tuned the parameters across all layers using our dataset.

We trained our model for 300 epochs with a mini-batch size of 32. In each epoch, we first shuffled all the images in the training set and then iteratively fed the 32 mini-batch images into the model until all the training images were loaded. The convolutional and pooling layers were used to obtain a 32 × 2048 × 1 × 1 feature map, which was then flattened to a 32 × 2048 matrix and mapped to the label space through the fully connected layer. Lastly, we employed a softmax layer to obtain the probability distribution matrix of 32 × 113, which gave the probability of each image belonging to any of the 113 taxonomies.

For training, we used a stochastic gradient descent optimizer with a momentum of 0.9 and weight decay of 5×10^{-4}. The initial learning rate was set at 0.001 and decreased following a cosine annealing schedule. As such, the learning rate initially reduced slowly, then accelerated, and then slowly reduced again to about 2.7×10^{-8} in the last epoch. This decay strategy was well-combined with the stochastic gradient descent algorithm to optimize the objective function. During early stages of training, the learning rate was higher to accelerate the network's convergence while gradually reducing it to ensure that the network better converged to the optimal solution.

In general, learning in neural networks involves an optimization procedure. We utilized the cross-entropy loss function to calculate the model error, which cast the learning problem as a search optimization problem. To compute the gradients, we employed the backpropagation of the error algorithm and updated the parameters using the stochastic gradient descent optimization algorithm. After training each epoch, we fed all test images into the model and performed forward propagation to calculate accuracy. We saved the model with the highest accuracy parameters as our final training model.

Sensitivity-specificity curve

In our task, we utilized the sensitivity-specificity curve, which is a probability curve generated by varying threshold settings. It is commonly referred to as the receiver operating characteristic (ROC) curve. The area under the curve, known as the AUC, represents the degree of separability. Since our task constituted a multiclass classification problem, we compared our results based on macro-average and micro-average sensitivity and specificity.

$$\text{Sensitivity}_i = \frac{\textit{true positive}_i}{\textit{positive}_i}$$

$$\text{Specificity}_i = \frac{\textit{true negative}_i}{\textit{negative}_i}$$

$$\text{Sensitivity}_{macro} = \frac{\text{Sensitivity}_1 + \text{Sensitivity}_2 + \cdots + \text{Sensitivity}_k}{k}$$

$$\text{Specificity}_{macro} = \frac{\text{Spectivity}_1 + \text{Spectivity}_2 + \cdots + \text{Spectivity}_k}{k}$$

$$\text{Sensitivity}_{micro} = \frac{\textit{true positive}_1 + \textit{true positive}_2 + \cdots + \textit{true positive}_k}{\textit{positive}_1 + \textit{positive}_2 + \cdots \textit{positive}_k}$$

$$\text{Specificity}_{micro} = \frac{\textit{true negative}_1 + \textit{true negative}_2 + \cdots + \textit{true negative}_k}{\textit{negative}_1 + \textit{negative}_2 + \cdots \textit{negative}_k}$$

where *true positive*$_i$ is the number of correctly predicted fossils, *positive*$_i$ is the number of the type *i* fossil shown, *true negative*$_i$ is the number of correctly predicted fossils *i*, and *negative* is the number of fossils of type *i* shown. When a test set is fed through the deep network, it outputs a probability, *P*, of fossil types by image. In macro-averaging, all classes are equally weighted when contributing their portion of the value to the total, while in micro-averaging, each observation receives equal weight. This gives greater power to the classes with the most observations. We computed both averaged sensitivity-specificity values by adjusting the threshold of the deep neural network classifier. The sensitivity-specificity curve is a probability curve at various threshold settings and its area under the curve (AUC) represents the degree or measure of separability.

We chose to use micro-instead of macro-averaging because the former computes the metric for each class independently and then averages them, whereas the latter treats all classes equally, and thus therefore, micro-averaging is preferable in multi-class classification tasks with imbalanced classes.

Confusion matrix

In evaluating classification accuracy, we utilized the confusion matrix, which provides a metric for classification evaluation. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in the actual class (or vice versa). As for our deep learning approach using the test set, the confusion matrix is provided in the Figures S5 and S6.

Uniform manifold approximation and projection

To validate the feature extraction ability of our deep learning model and visualize the distribution of all images, we employed the uniform manifold approximation and projection (UMAP) algorithm,[50] a dimension reduction and visualization technique. In doing so, we reshaped the last convolutional block of the deep learning model, generating feature maps (2,048 × 7 × 7) into 1 × 100,352 dimensional features. We then fed these features into the open-source tool "pix-plot" to build the UMAP layout. The "min_distance" parameter of the embedding was set to 0.001, and the trade-off between local and global clusters ("n_neighbors") was set to 6. We employed "correlation" as the "metric" parameter.

QUANTIFICATION AND STATISTICAL ANALYSIS

We only involved statistical analyses in the calculation of recognition accuracy, time cost, and micro-averaged sensitivity-specificity curves (i.e., receiver operating characteristic curves, ROCs) for evaluating deep learning methods. The identification accuracy in Figure 3 and Figures S4–S6 is calculated as ''Identification accuracy = number of test samples recognized accurately/number of all test samples''. For the sensitivity-specificity curves, we give the definition in the main text of the manuscript, and the definition and calculation method of the sensitivity-specificity curves are given in detail in the ''method details'' section.