Accuracy of a neural net classification of closely-related species of microfossils from a sparse dataset of unedited images

Johan Renaudie Corresp., [1] , **Ryan Gray** [2] , **David B Lazarus** [1]

[1] Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany
[2] Unaffiliated, Reston, Virginia, USA

Corresponding Author: Johan Renaudie
Email address: johan.renaudie@mfn-berlin.de

Identification of biologic objects in images is a major source of biodiversity data. Currently this is done by scarce taxonomic experts and data is thus limited in scope and reproducibility. Automated identification in fields such as plankton research or micropaleontology, where enormous numbers of objects are available, would significantly improve data quantity and quality, particularly in applied studies of environmental and climate change. We describe a machine learning workflow based on the MobileNet convolutional network. The software can identify closely related species of radiolarians, a morphologically challenging group of microfossils, and from complete species populations (not only ideal specimens) as they are normally identified in standard transmitted light microscope preparations. Multiple, partial focus, depth of field limited images were obtained for each fossil specimen from multiple radiolarian microslides. Images were normalized and in one test also cropped to remove most systematic slide-linked image biases (e. g. type of background particles) that could be used by a classifier as non-taxonomic clues to species assignment. An average of 60 specimens per species for 16 species in two distinct clusters of closely related forms (9 species in the *Antarctissa* group and 7 species in the genus *Cycladophora*) were used to train and test the system. An overall average classification accuracy of ca 73% was achieved, and for some species >85%. Using a cutoff for specimens with classifier-calculated low certainty values boosts overall accuracy close to 90%, but at the cost of ca 1/3 reduction in identifiable specimens. This latter accuracy is close to the reproducibility of human experts, albeit with more unidentifiable specimens. The most important constraint to broader use is the time and effort needed by taxonomic experts to collect and label images to be used in training, as many species in these diverse biotas are rare, and the numbers of taxonomic experts available are very limited.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27328v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018

Accuracy of a neural net classification of closely-related species of microfossils from a sparse dataset of unedited images

Johan Renaudie[1], **Ryan Gray**[2], **and David Lazarus**[1]

[1]**Museum für Naturkunde, Invalidenstraße 43, 10115 Berlin, Germany**
[2]**12003 Walnut Branch Road, Reston, VA 20194 USA**

Corresponding author:
Johan Renaudie[1]

Email address: johan.renaudie@mfn.berlin

ABSTRACT

Identification of biologic objects in images is a major source of biodiversity data. Currently this is done by scarce taxonomic experts and data is thus limited in scope and reproducibility. Automated identification in fields such as plankton research or micropaleontology, where enormous numbers of objects are available, would significantly improve data quantity and quality, particularly in applied studies of environmental and climate change. We describe a machine learning workflow based on the MobileNet convolutional network. The software can identify closely related species of radiolarians, a morphologically challenging group of microfossils, and from complete species populations (not only ideal specimens) as they are normally identified in standard transmitted light microscope preparations. Multiple, partial focus, depth of field limited images were obtained for each fossil specimen from multiple radiolarian microslides. Images were normalized and in one test also cropped to remove most systematic slide-linked image biases (e. g. type of background particles) that could be used by a classifier as non-taxonomic clues to species assignment. An average of 60 specimens per species for 16 species in two distinct clusters of closely related forms (9 species in the *Antarctissa* group and 7 species in the genus *Cycladophora*) were used to train and test the system. An overall average classification accuracy of ca 73% was achieved, and for some species $>$85%. Using a cutoff for specimens with classifier-calculated low certainty values boosts overall accuracy close to 90%, but at the cost of ca $1/3$ reduction in identifiable specimens. This latter accuracy is close to the reproducibility of human experts, albeit with more unidentifiable specimens. The most important constraint to broader use is the time and effort needed by taxonomic experts to collect and label images to be used in training, as many species in these diverse biotas are rare, and the numbers of taxonomic experts available are very limited.

INTRODUCTION

Paleontologic and neontologic observations of organism occurrences are central to studies of modern and past biodiversity. While for some types of studies occurrence data for genera or higher taxa (e.g. ‘functional groups’ in ecology) may be acceptable (Richardson, 2006; Barton et al., 2016), for many types of research it is necessary to classify specimens to species level, e.g. for better understanding of modern ecology and ecosystem function, biostratigraphic determination of the geologic age of sediments, reconstruction of past environments, biodiversity dynamics, and many other areas of research (CLIMAP project members, 1976; Bolli et al., 1985; Prance, 1994; Wiese et al., 2016; Tréguer et al., 2018). This data is still overwhelmingly collected by human observation and manual recording of data. This is labor intensive and prone to subjective differences between workers that degrade quality in syntheses of published data. Particularly in combination with a global shortage of expert taxonomists, this style of data collection is a major barrier to the amount and quality of primary data that can be generated for (paleo)biodiversity and other research. Although in many areas of paleontologic work fossils are very rare (e.g. vertebrates, and in particular hominids), in other areas (some fossil invertebrate groups, microfossils), and in biologic research, the amount of material potentially available for observation is extremely large, and faster, more objective methods of species occurrence data generation would

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27328v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018

allow substantial improvement in the scope and quality of research done. In micropaleontology for example, where the current recovered deep-sea sediment archives already contain an estimated 10^{15} fossil specimens, it would open a vast repository of evolution, ecology and climate change data to study (Lazarus, 2011).

The potential of automated species identification of specimens for such materials has long been recognized, and many attempts have been made to develop computerized automatic identification of biologic and paleontologic objects. Early work concentrated on image preprocessing and extraction of exterior shell outlines, and algorithms to extract taxonomically useful data for identification from these (Lohmann, 1983; Hills, 1988). These systems were used either for broad category identification (Benfield et al., 2007), object-background separation (Knappertsbusch et al., 2009), or for extracting general, non-species specific morphologic metrics for ecologic or evolutionary studies (Granlund, 1986; Schmidt et al., 2004). Later work using more advanced programs have made use of whole image data, a variety of general image metrics (both for outlines and internal structures such as texture) and attempt, sometimes via custom, taxon specific algorithms, to classify the imagery (Wu et al., 2015; Apostol et al., 2016; Keçeli et al., 2017). Lastly, recent work has begun to explore using advanced neural network systems to automatically classify objects, e.g. plankton (Zheng et al., 2017) and microfossils (Beaufort and Dollfus, 2004; Keçeli et al., 2017). So far however, these newer studies have mostly made use of a restrictive set of images for training and testing. For example, in both Apostol et al. (2016) and Keçeli et al. (2017) the test images are pre-selected for completeness, orientation, separation from background and other image optimizations, and each image or image set comes from very distinctive taxonomic groups (genus, family, or even ordinal level distinctions). Such work is valuable as it determines the general applicability of these newer software systems to the broad range of morphologies encountered in e. g. microfossil identification. Real world identification of species however deals with a very different set of images, and a different set of classification challenges. Microfossils for example when observed in normal preparations are in mixed assemblages of many species, often include non target objects (other groups of microfossils, inorganic particles), include taxonomically closely related and thus usually morphologically very similar ‘sister’ species, the specimens are presented against cluttered backgrounds, sometimes overlapping with other objects, and in a wide variety of orientations and image quality. To our knowledge the ability of even modern neural network systems to usefully classify images to species level has not yet been adequately tested under such conditions. Only the SYRACO system (Dollfus and Beaufort, 1999; Beaufort and Dollfus, 2004) has been demonstrated to work under such conditions, but with a very limited number (11) of mostly highly distinct, relatively simple image types (bright coccolith images in darkfield illumination).

In this study we address this issue of selective vs ‘real world’ imagery for species-level classification. We use images as they appear in the microscope, not isolated from the complex background of other microfossils, and with typical image limitations such as only partial sharp focus due to depth of field limitations. Discrimination between closely related taxa is the most difficult task performed by human specialists, and is much more challenging than discriminating between more distantly related, morphologically distinct forms. We thus also explicitly choose species that are morphologically similar to each other, and indeed, are challenging even for human experts to properly separate.

MATERIALS

We have chosen to test automated identification systems on fossil Cenozoic radiolarians. Radiolarians (in this study, we mean only the group Polycystinea, which form fossils) are one of the major groups of organisms used in micropaleontologic research: Cenozoic forms in particular are extensively employed in studies of ocean and climate change, to provide geologic age estimates for sediments and rocks, and in studies of biologic evolution (Lazarus, 2005). Radiolarians have a high global living diversity of ca 400 species and an unusually large range of shell architectures, but also due to their high total diversity, frequently with >100 species in individual samples, many species are found in a single sample that belong to the same genus, and thus have very similar morphologies. Radiolarian shells are constructed as 3-dimensional forms out of an open, pore-dominated lattice-work of transparent opaline silica, and although a great variety of shell forms exist (spheres, discs, and many others, with various spines or other ornamentation), the taxa used in our study are approximately conical in shape. They are normally studied using transmitted light microscopy in mixed assemblages of sieved material, including often not only radiolarians but other types of microfossils such as diatoms, and non-biogenic particles, extracted from sediment or rock samples. Because of their small size (ca $100\mu m$) only a part of the individual

Sample	Imaging by	Age, Ma	*A denticulata*	*A. ballista*	*A. cylindrica*	*A. strelkovi*	*A. deflandrei*	*A. robusta*	*L. setosa*	*H. praevema*	*H. vema*	*C. davisiana*	*C. pliocenica*	*C. conica*	*C. cosma*	*C. spongothorax*	*C. humerus*	*C. golli*	Slide total
751A-1H-1,98 cm	dbl	0.10												32					32
1138A-2R-4 27/31 cm	jr	0.70	16			19			7			13		2					57
751A-1H-2 7/7 cm	jr	1.65	5		5	1			14			52		12					89
693A-6R-5 48/55 cm	dbl	3.50	7		15	4							24						50
751A-3-4 85/87 cm	jr	4.10		1	26				6		29		3						65
689B-2H-5 55 cm	dbl	4.45											50			1			51
693A-18R-4 101/107 cm	jr	6.80	47	19	29	28				6			7						136
689B-3H-3 116/118cm	jr	7.47	4	25	7	10				25			3						74
689B-4H-4 116/118cm	dbl	9.89											3	8	2	18			31
1138A-17-2 105/107cm	dbl	10.30					83								6	12			101
751A-10H-1 98 cm	dbl	11.20												1	38	59	53		151
845A-19-3,107 cm	dbl	12.35												11	8				19
278-20-1 77/78 cm	dbl	16.00					1	35							8			24	68
751A-17-CC	dbl	18.40																39	39
Total specimens by species			79	45	82	62	84	35	27	31	29	65	90	66	62	90	53	63	963

Table 1. List of samples with numbers of specimens of species. Geological ages in millions of years (Ma) from deep-sea drill section age model library at www.nsb-mfn-berlin.de.

shells are in sharp focus (within the microscope's depth of field) at any one time. As the preparation method (Moore Jr, 1973) simply deposits particles randomly over the surface of the microscope slide, specimens are found in all possible 3-dimensional orientations, even if due to their shape, individual species tend to orient in only a few preferred positions on the slide surface. Lastly, the random distribution of particles results in numerous specimens touching, or partially overlapping with other particles on the slide. Radiolarian shells typically have both external and internal structures, both used for classification. Internal structures can usually be imaged, albeit with loss of clarity, due to the transparent nature of the shell material of the outer shell wall, and the ability to image, via control over the depth of field, only a narrow plane within the shell. These aspects – very high variety of objects, highly porous lattice structure, focus limitations, 3D orientation variation, and overlapping objects – all add to the challenge of object identification, beyond those present in most biologic materials such as broken specimens, within-species variation in morphology, and the taxonomic differences between different species' shell morphologies.

The specific materials chosen are limited by two additional practical considerations. Neural network systems work by first being trained on a large number of reference images of the categories that they should learn. There is no fixed number of reference images required, but at least several dozen, and ideally several hundred are normally needed. This requirement intersects with the reality that (as is true of most biotas), within radiolarian assemblages only a few species are common, while most are rather rare ($< 1\%$ abundance). As we wish to present the neural system with images as they are encountered in the microscope in daily work, we have chosen not to use only pre-selected, optimized images from the literature, but new images from specimens taken from actual samples. Thus, to collect reasonable numbers of specimen images (here we chose a minimum of ca 30 specimens per taxon) from sets of closely related taxa, we are compelled to look for species which are mostly rare, requiring substantial effort in scanning assemblages to locate individuals. This in turn limits the number of taxa that can be managed in our study. We have therefore chosen a total of 16 species, distributed in two different clusters of similar morphologies. These provide us with two independent tests of identification performance on closely related taxa, and the opportunity to compare performance as well to identification between the morphologically more dissimilar clusters. We also chose taxa groups and samples (Antarctic, where radiolarian diversity is substantially lower than in the tropics) where many, if not most of our target species were, if not common, also not extremely rare.

The first morphologic group consists of antarctissids - 9 species in the radiolarian genus *Antarctissa* or closely related forms that are currently assigned to the genera *Lithomelissa* and *Helotholus*. Generic assignments in Cenozoic radiolarians are currently rather arbitrary and many closely related forms are still assigned distinct generic names that probably should be synonymized. This situation is due to the highly artifical generic classification created by early workers such as E. Haeckel, which has not yet been formally revised due to limited manpower by current workers (Lazarus et al., 2015). The species chosen are *Antarctissa denticulata* (Ehrenberg 1844), *A. strelkovi* Petrushevskaya 1967, *A. cylindrica* Petrushevskaya 1967, *A. ballista* Renaudie and Lazarus 2012, *A. robusta* Petrushevskaya 1975, *A. deflandrei* (Petrushevskaya 1975), *Helotholus*? *praevema* Weaver 1983, *H*?. *vema* Hays 1965 and *Lithomelissa setosa* Jörgensen 1900 (Figure 1). Full citations to taxon authors for all species used in this study but *A. ballista* can be found in Lazarus et al. (2015). Antarctissid species are topologically segmented-conical, and are constructed out of a large, spherical to thumb-like apical 'segment' (the cephalis) and a larger, more or less cylindrical lower part to the shell (the thorax), plus several bars and struts inside the shell. Their formal taxonomy is not fully resolved, and there are differences of opinion, even among the authors of this paper, on the best set of morphologic characters to use for species and genus assignment. Due to these differences in choice of characters, different specialists sometimes assign identical specimens to different species names. However, a single specialist normally can consistently assign species using his or her preferred criteria, with only a small number ($< 5\%$) of individuals being ambiguous (not including specimens that cannot be confidently assigned due to orientation, preservation or other problems described above). For this study all antarctissids were identified by a single person (JR) to ensure consistency in the training set. Species assignment was based both on the external shape and lattice wall characteristics of the shell, and as well the number and arrangement of the internal bars and struts.

The second morphologic group examined are species within the genus *Cycladophora*. Species in this genus have a small, subspherical apical segment, a long, variously shaped conical 'thorax', and depending on species, a short, variously shaped third segment at the base of the shell (Figure 2). The taxonomy of this genus has been revised by prior work (Lombari and Lazarus, 1988) and the criteria for species discrimination (relative size of main segments, conical angle, outline of shell, pore size and shape, etc) are well defined. The differences between some pairs of species are however subtle, and published usage by other radiolarian specialists has not always been consistent with the primary definitions given in Lombari and Lazarus (1988). Seven species were used: *Cycladophora davisiana* Ehrenberg 1873, *C. pliocenica* (Hays 1965), *C. conica* Lombari and Lazarus 1988, *C. cosma* Lombari and Lazarus 1988, *C. spongothorax* (Chen, 1975), *C. humerus* (Petrushevskaya, 1975) and *C. golli* (Chen, 1975). Again, all assignments of specimens were done by a single person (DL) to ensure consistency.

The specimens were imaged from standard radiolarian microscope slides made from ocean sediment samples from various deep-sea drilling sections from the Southern Ocean (DSDP Site 278; and ODP Sites 689B, 693A, 747A, 751A and 1138A) ranging in age from the middle Miocene to the Pleistocene. The use of multiple radiolarian slides from different samples was necessary as different species have distinct, partially non-overlapping geologic age ranges, and not all time intervals are equally well represented or preserved in each deep-sea geologic section. All slides are currently stored in the micropaleontological collections of the Museum für Naturkunde, Berlin. Due to images being collected by different workers, and in part in different locations, three different microscropes were used for image collection.

A total of 963 radiolarian specimens were imaged for this study, with an average of 60 individuals per species (range 27–90). Details of numbers of specimens by species, samples and slides used are given in Table 1.

METHODS

Depending on the species and the specimen's orientation, several focal planes were used to illustrate most specimens, so that all or most determining characters could be observed for each specimen, even if not all in a single image. Target specimens were approximately centered in the image. The images were otherwise not edited or masked, i. e. additional radiolarians and other microfossils were normally present in the image background. All the pictures were either taken as JPEG or converted afterwards to JPEG. The colorspace was normalized to greyscale. Finally, to make all the pictures comparable to one another, they were all resized to a common resolution, i. e. the same pixel to micrometer ratio (8.9 pixel per μm).

For each species, specimens were randomly split into 10 sets: eight of them constitute the training dataset, one the validation dataset and the final one the testing dataset. The training and validation datasets

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27328v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018

were used by the neural network in an iterative process of trying image analysis algorithms on the training images to create classifications of the images, checking with the validation set that the provisionally found algorithm is useful (the classifications were at least partially correct), and repeating many times (i. e. 4000 training steps) to improve the algorithm, i.e. by checking each time the correctness against the validation set. The result of this was then compared to the testing dataset to see how well the algorithm found by the network performed. This entire procedure was itself repeated, each time using a different assignment of the initial 10 sets to the training, validation or testing datasets, to ensure that the specific assignment of images to the test dataset was not biasing the estimate of how well the system worked (10-fold cross-validation: Mosteller and Tukey, 1968; Stone, 1974).

The neural network algorithm selected in this study is the MobileNet convolutional neural network (Howard et al., 2017). The concept of a convolutional neural network is for the algorithm to learn what series of image filters to apply on the picture to optimize the discrimination between the various classes (here species). The image filters (also called kernels) are similar in principle to those found in many image manipulation programs, where small windows (e. g. 3×3 pixels) containing rules for calculating outputs are slid across the source image. Filters of this type are used e.g. for edge detection or other image modifications. The MobileNet convolutional neural network uses a specific type of filters: depth-wise separable filters. As the name implies, MobileNets were conceived to be computable in theory on machines with no more computing power than mobile phones, meaning they are not as computationally intensive as more complex models. Two parameters can be modified to simplify the model: the width multiplier and the resolution multiplier. Here we chose to keep large multipliers (i. e. a width multiplier of 1.0 and an input resolution of 224), to maintain a high level of accuracy.

The optimization metric that the learning process uses is simply the fraction of images correctly classified (ideally 1, when all images are identified correctly). Technically, the software does this by trying to minimize the top-1 classification rate. Top–N classification rate is defined as the sum of pictures for which none of the first N classes attributed by the algorithm are correct, divided by N. As here $N = 1$, it corresponds simply to the number of cases in which the class (here the species) was incorrectly guessed by the algorithm. Classification is based on the confidence value computed by the network that a given image belongs to a certain class (species), with the value ranging from 1 to 0.

Tests suggest that the algorithm performs better if provided with a training dataset including specimens pictured in multiple focal planes. In our study the image in each such set of images for a specimen that produced the most highly ranked assignment confidence was used to identify the specimen to a species. This, at least partially, simulated the way human workers identify species, by looking through the different focal plane views to identify key morphologic characters that distinguish species. Human workers however combine information from different images in making identification decisions, an important ability which our method currently does not include.

As a result of the training phase, a tailored classification network graph is produced, i. e. a set of instructions containing specific filters to apply to the pictures to determine which class (= species) it belongs to. Classification using this corrected graph is then applied to the testing dataset: in real case scenarios, of course, the testing dataset is the dataset composed of pictures of radiolarians that needs to be classified. For each specimen, the classifier outputs a list of top guesses alongside their associated certainty.

All code was run using Python 2.7 (Python Software Foundation, 2010) along with the python module TensorFlow (Abadi et al., 2016) which implements graph-based neural networks.

Three different versions of the images were analyzed. The first version 'raw' consisted of the raw images with only normalization of size. In the second version 'leveled' image grey-levels were adjusted to the same mean values. This was done to ensure that consistent differences in grey levels that existed between microscopes and/or individual microscope slide image sets could not be used by the neural classifier as an indirect source of information for species identification. This would be possible as different species were imaged primarily by different workers using different microscopes. A last version of the images were 'cropped' versions of the leveled images. Here most of the background except that immediately surrounding the specimen was removed. This was done for reasons similar to leveling: some microscope slides that were the source of the majority of specimens for one or another species have characteristically different types of background such as many or few diatoms (a different group of microfossils preserved in the same samples), or other non-species related but nonetheless potentially usable information that might bias our results. The cropping process was automatic, taking a fixed area at

Species	#	raw	levels	crop	average over 3 treatments
Antarctissa strelkovi	62	56.45% (35)	54.84% (34)	51.61% (32)	54.30%
Antarctissa ballista	45	71.11% (32)	71.11% (32)	75.56% (34)	72.59%
Antarctissa cylindrica	82	57.32% (47)	54.88% (45)	59.76% (49)	57.32%
Antarctissa deflandrei	84	85.71% (72)	88.10% (74)	85.71% (72)	86.51%
Antarctissa denticulata	79	62.03% (49)	67.09% (53)	68.35% (54)	65.82%
Antarctissa robusta	35	85.71% (30)	94.29% (33)	88.57% (31)	89.52%
Helotholus praevema	31	67.74% (21)	64.52% (20)	70.97% (22)	67.74%
Helotholus vema	29	86.21% (25)	86.21% (25)	93.10% (27)	88.51%
Lithomelissa setosa	27	59.26% (16)	44.44% (12)	81.48% (22)	61.73%
Cycladophora conica	67	61.19% (41)	62.69% (42)	65.67% (44)	63.18%
Cycladophora cosma	62	66.13% (41)	61.29% (38)	67.74% (42)	65.05%
Cycladophora davisiana	65	81.54% (53)	76.92% (50)	76.92% (50)	78.46%
Cycladophora golli	63	80.95% (51)	80.95% (51)	76.19% (48)	79.37%
Cycladophora humerus	53	73.58% (39)	71.70% (38)	86.79% (46)	77.36%
Cycladophora pliocenica	90	86.67% (78)	90.00% (81)	80.00% (72)	85.56%
Cycladophora spongothorax	90	65.56% (59)	70.00% (63)	82.22% (74)	72.59%
all species		71.47%	71.68%	74.59%	72.58%
Antarctissids		68.99%	69.20%	72.36%	70.18%
Cycladophora		73.88%	74.08%	76.76%	74.90%

Table 2. Classification accuracy of specimens by species, and by image treatment: raw (only size normalization), levels (also greyscale normalization), cropped (most other objects trimmed away).

the center of the images. In about 5% of the images the crop also removed some part of the target species image (usually only a small fraction on one edge), but we do not think this had a significant effect on the results, or if any, was conservative in reducing our reported success rate.

RESULTS

Our results are summarized in Table 2, and the detailed, picture by picture, results are given in the Supplementary Material. Over the range of all different image treatments and all specimens the average accuracy in species identification was 72.6%, with only a moderate degree of variation due to image treatment (71–75%). There was only a slightly greater range of accuracy between the two major groups of radiolarians studied, or by treatment: a minimum of 68.99% for antarctissids in the raw image set, vs 76.76% for cropped images of *Cycladophora*. There was however noticeably more variation in performance at the level of individual species. For species, a minimum accuracy was obtained for *A. strelkovi* with ca 52% specimens being on average correctly identified (here we exclude *L. setosa*'s 44.4% in the leveled image set: due to the low number of specimens for this species, its results vary in the different image sets widely from 44 to 81% and are therefore statistically unreliable). A maximum accuracy of 90% was obtained for *Cycladophora pliocenica* (if we again exclude *H. vema*'s 93.1% and *A. robusta*'s 94.3% for the same reason of low specimen numbers).

A striking feature of the results are the differences in the aggregate degree of confidence between the population of specimens that are correctly classified vs those that are incorrectly classified (Figure 3). The distribution of confidence values for correctly identified pictures is very strongly asymmetric and centered near 1.0: for cropped images its median is 0.98, and its skewness -0.51. By contrast, incorrect pictures confidence values are scattered over a wider range of values (between 0.3 and 1) with a flatter distribution: for cropped images its median is 0.77, and its skewness -0.14.

Inspection of specimens incorrectly classified shows that the misindentification occurs mostly between closely-related species (See Figure 4): antarctissids are almost always misindentified for other species of antarctissids and rarely for species of *Cycladophora* and vice versa.

DISCUSSION

The accuracy of species identifications in our study depends on several factors, but the one that makes the largest difference is simply the ‘rejection rate’, i.e. the percent of specimens that are left as ‘unidentifiable’ vs those that are classified into one of the training categories. If we are willing to accept fairly high levels of ‘unidentifiable’ specimens (a bit over 40% of the specimens in the imaged dataset), we can achieve nearly 90% accuracy in those specimens that are classified, vs only 71% accuracy if we try to classify all specimens in the full dataset. The idea of using an identification system that skips over ‘uncertain’ specimens in scientific data collection may seem at first to be of questionable validity. In reality however human workers also routinely skip specimens that they cannot confidently identify. This is true, from our own experience, in all areas of micropaleontologic research, and is presumably true as well for all workers that identify biologic specimens in real-world materials. For example, in this study, we included every specimen encountered on the slides that we, as experts, felt at least moderately confident in correctly identifying, but skipped a significant number of specimens where we were unsure of the correct identification. The reasons for uncertainty were varied, including incomplete preservation of characters, difficult orientation, obscuring of the view by overlapping other specimens, as well as specimens with uncharacteristic/mixed characters. The numbers skipped varied according to observer, slide and taxon. Given the uncertain boundary to specimens not only unidentifiable to species level, but also to the genus-level category (antarctissids or *Cycladophora*), we did not attempt to count the percentages skipped, but subjectively around 30% of potential specimens were skipped by us and not imaged for the study. Thus, an automatic classifier that skips an additional percent of specimens is not doing anything fundamentally different than a human worker does, even if the absolute values of skipped specimens are much higher. Given the huge numbers of specimens available for study, an automated system would simply trade identification completeness for larger numbers of specimens examined. The only problem would be if the ‘unidentified’ category is systematically biased, or even worse, inconsistently biased towards different species.

In Fenton et al. (2018), the authors test the consistency of species identification of planktonic foraminifera among specialists and trained but inexperienced students by providing them with a series of preselected specimens representative of a few species’ morphological ranges and are asked to identify them. The authors find a consistency of 78.5% among specialists and 57% among trained students. With an overall accuracy of 72.6%, the convolutional neural network trained on a small set of radiolarian pictures presented here thus performs slightly worse than a specialist (which was to be expected) but significantly better than a trained student. Similarly, still in Fenton et al. (2018), the authors split this consistency in three categories according to the confidence of the identification (‘confident’, ‘maybe’, ‘not confident’) and find that the consistency is higher in the ‘confident’ category (overall: 77%; specialists: 93.1%; students: 75%). Here, the algorithm confidence is given numerically, so should we defined the ‘confident’ category as values over 0.95 confidence (which concerns 997 of the 1987 pictures; or 639 out of 963 specimens), the algorithm accuracy on the cropped dataset increases to 88.3% of the specimens or 90.5% of the pictures (see Table 3 and Figure 3), again slightly lesser than the consistency among specialists but higher than among trained students. In comparing our results to those of Fenton et al. (2018) it should be noted that the initial ‘universe’ of specimens included in the training set is not fully identical. Whereas we explicitly have not included specimens in our training set that we found difficult to uniquely identify, in Fenton et al. (2018) it is not clear to what extent such specimens were excluded from the initial specimen selection. An additional difference is that in Fenton et al. (2018) all specimens that were selected were subsequently, via additional effort, classified to the same level of fully confident, even if initially there were significant differences in opinion on the identification of the specimen. In our study the filtering of such difficult specimens mostly occurred at the initial step of selection for imaging, rather than subsequently attempting to resolve ambiguous classifications.

Though the difference between treatments is minimal, the cropped dataset is sensibly more accurately identified, which is most probably due to the gain in resolution allowing the algorithm to make use of more morphological details to separate the species, and the elimination of background ‘noise’, such as other type of particles (diatoms, sponge spicules fragments, rock fragments, etc.) that the algorithm could theoretically pick as non-taxonomic, and thus unreliable, clues.

Differences in accuracy between species are also significant. In particular the algorithm has difficulty differentiating closely-related species of *Antarctissa* (*A. denticulata* from *A. cylindrica* and *A. strelkovi* from *H. praevema* for instance; see Figure 4). These species are known to be difficult to distinguish by

Confidence band	% Pictures in confidence band	% of them correctly classified	% Specimens in confidence band	% of them correctly classified
$(0, 0.5]$	5.08% (101)	32.67% (33)	1.55% (15)	32.67% (6)
$(0.5, 0.95]$	44.74% (889)	52.19% (464)	32.30% (312)	52.19% (151)
$(0.95, 1]$	50.18% (997)	90.47% (902)	66.15% (639)	88.26% (564)

Table 3. Classification accuracy of pictures and specimens (in the cropped dataset) by confidence categories, in percent, followed by the actual number of correct identifications vs total number of identifications in the categorie.

experts, and in fact are often lumped in may studies together as '*Antarctissa* spp.', or in the case of *A. strelkovi–praevema*, are often considered to be one variable species. Separating such species necessitate the combined observations of external and internal characters. One limitation of the MobileNet algorithm is that it automatically resizes the pictures to a lower resolution in order to process them in a more manageable timeframe (as the algorithm complexity increases significantly with the size of the pictures). Problematically, nassellarian radiolarians (and in particular this cluster of *Antarctissa* species) have a complex series of internal spicules that can be diagnostic at the species- or genus-level: these spicules (as can be seen on Figure 1.1B, 3 or 6B) are usually only a few micrometer wide (usually close to 1 μm). This spicular system complexity is probably mostly lost to the downsized image resolution used by MobileNet and thus also the taxonomic information that could have been used to identify species.

Similarly, in the confusion matrix (Figure 4), *Cycladophora spongothorax* and *C. humerus* cluster together (meaning that they are often mistaken for one another by the algorithm): one of the main differences between the two species is the presence in most, but not all, specimens of *C. spongothorax* of a spongy outer layer that obscures the primary morphologic characters of the main shell, and which, if not in sharp focus, creates a blurry grey image similar to typical background image areas - clumps of sediment, out-of-focus centric diatoms, etc.

However without knowing the details of how the network classifiers for these species were constructed we cannot be sure whether or not these specific features were strongly weighted by the algorithm. This points to another aspect of neural networks which remain problematic when employed for species identification: they produce good results, but how they were calculated, and in particular which aspects of the morphology were used to discriminate species are not clear. It is possible to extract a report of the image processing steps (filter series) but as these are extremely long and complicated, and do not directly link to image features their meaning for a taxonomist or other user is quite limited. Efforts to back-map the processing steps to image morphologic characters have been attempted (e. g. Zeiler and Fergus, 2014) but the methods are not accessible to non-technical specialists and the outputs still too vague to be of much use for taxonomic work. Although classical methods e. g. linear measurements, landmarks or outline analysis (Lohmann, 1983) lack the classification power of network methods they provide a better link to classical taxonomic knowledge of species, and thus their results can be better placed in a scientific context.

The above suggests that it would also be of interest to know how well classical morphometrics would perform with the same materials. Unfortunately there has been only a limited amount of such work done on these taxa. Granlund (1986, 1990) used manually measured linear and hand-digitized image outline data to study morphologic variation in *Antarctissa*, but did not distinguish between species. Lombari and Lazarus (1988) gave hand-recorded linear measurements made with an eyepiece micrometer for most species pairs of *Cycladophora* used in the current study and showed that bivariate plots of selected characteristics were usually sufficient to distinguish between most taxa. For neither genus, to our knowledge, has their been any prior attempt to identify and discriminate between species with automated methods. Recent work (Christodoulou et al., 2018) suggests that (automated) use of classic morphometric methods would result in a similar accuracy range as the CNN algorithm used here.

In our study, multiple images of the same specimen taken at different focal planes were used in specimen classification, but each image was treated individually by the network classifier, with synthesis of information limited to choosing the image with the highest confidence value for specimen assignment. A stronger approach would be to combine the images together before submission to the network. This in principle would allow more characters to be in focus in a single image and thus available to the network for building more sophisticated classification rules that better simulate the methods used by human experts.

There are however challenges to doing this. As noted above, for some radiolarian species internal as well as external characters are important in taxonomy, and no single image can therefore fully represent the significant characters needed for accurate classification. Additionally, although several software programs are available that automatically composite microscope generated images of several focal planes into a single sharp, deep-focus image, the technical requirements (automated microscope controls and control systems) substantially raise the costs, and image acquisition times, vs capture of simple multiple images as used in our study.

As mentioned above, the resolution limitation of MobileNet might impede its ability to identify species for which the minute inner spicular system is diagnostic. Using a convolutional neural network that uses full resolution pictures should thus improve the accuracy rate. Furthermore, in a real-case scenario, the accuracy of the algorithm should theoretically increase with a larger number of training specimens used for each species, and with an increase in training steps (which we limited here to 4000) and a lowering of the learning rate (here 0.01).

In addition to the basic issue of identification accuracy, there are many other issues that need examination before pilot studies such as this can be transformed into systems for routine use in micropaleontology or other similar areas of research.

Our study examined only a very small fraction of the known diversity of fossil radiolaria, which is estimated at several thousand valid species descriptions over their entire Phanerozoic geologic range (Lazarus, 2005), with probably an even larger number of forms not yet described. How well our system would scale to diversity of this magnitude has not been tested. However, our initial tests using our own, limited set of taxa (SOM) show that by using transfer learning methods, which make use of the 'general' knowledge about classes learned to speed up training times for related classes (Pan et al., 2010), we obtain a linear behavior of run time to classes (taxa), and ca 5 minutes per taxon, so that even several thousand species should be manageable with current technology.

The time and effort needed however to collect and expertly identify large numbers of images, particularly for the many rare species is more problematic. Despite selecting faunal assemblages of only moderate diversity, and taxa groups where many member species are (relatively) common, we required ca 5–10 minutes per specimen to collect imagery, with the time spent approximately equally between scanning to find specimens, and taking/labelling images. Although these numbers could be improved by optimizations to the work flow (autolabeling of images, etc), the time needed per species (50–100 specimens on average i. e. ca 1 man-day of effort) is substantial. Indeed, although the ultimate use of such systems would be to amplify the rare resource of human expertise by taking over routine counting tasks, for many taxonomic groups the available human resources are already so limited that they would be stretched simply to collect sufficient images (for e. g. just the Cenozoic radiolaria, several man-years of work would be required), and this diversion of effort would be to the serious detriment of other types of research. Better tools to harvest existing published imagery, and/or to enable users to post casual new images in central, accessible online repositories would help, particularly for rare species, and taxonomic groups with few specialists. Machine learning systems that can be trained effectively with fewer images would also be extremely beneficial. A more radical approach that could dramatically reduce the time required (< 1 minute per specimen) would be where imaging was done for all specimens of radiolarians on a slide (i. e. all species at once, not scanning just for a few, mostly rare species within selected groups), with software support for improved workflow, e. g. labeling by experts. Such a large scale, dedicated effort may be the best approach to converting pilot studies into routine use.

CONCLUSIONS

Automatic identification of morphologically complex imagery from radiolarian microfossils, in normal mixed assemblages viewed in transmitted light microscopy, is practical using currently available technology (convolutional neural networks), at least for tasks where false positive classification rates of ca 10% are acceptable, and a substantial fraction (ca. a third) of the potentially identifiable specimens are not included in the collected data.

Although more studies are needed to confirm and generalize our results, it is possible that routine assemblage composition tasks can be automated, such as species counts of common taxa in applied environmental and climate change studies.

Collecting and labeling enough images of each species in order to train the network is a potential bottleneck, given low numbers of the available taxonomic specialists that are needed to do this work.

ACKNOWLEDGMENTS

We thank the creators of the open-source software used in this study and Google for the use of cloud services.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Apostol, L. A., Márquez, E., Gasmen, P., and Solano, G. (2016). Radss: A radiolarian classifier using support vector machines. In *2016 7th International Conference on Information, Intelligence, Systems Applications (IISA)*, pages 1–6.

Barton, A. D., Irwin, A. J., Finkel, Z. V., and Stock, C. A. (2016). Anthropogenic climate change drives shift and shuffle in north atlantic phytoplankton communities. *Proceedings of the National Academy of Sciences*, 113(11):2964–2969.

Beaufort, L. and Dollfus, D. (2004). Automatic recognition of coccoliths by dynamical neural networks. *Marine Micropaleontology*, 51(1):57 – 73.

Benfield, M. C., Grosjean, P., Culverhouse, P. F., Irigoien, X., Sieracki, M. E., Lopez-Urrutia, A., Dam, H. G., Hu, Q., Davis, C. S., Hansen, A., et al. (2007). Rapid: research on automated plankton identification. *Oceanography*, 20(2):172–187.

Bolli, H. M., Saunders, J. B., and Perch-Nielsen, K. (1985). *Plankton Stratigraphy*. Cambridge University Press, Cambridge, UK.

Chen, P.-H. (1975). Antarctic radiolaria. *Initial Reports of the Deep Sea Drilling Project*, 28:437–513.

Christodoulou, M. D., Battey, N. H., and Culham, A. (2018). Can you make morphometrics work when you know the right answer? Pick and mix approaches for apple identification. *PLOS ONE*, 13(10):1–17.

CLIMAP project members (1976). The surface of the ice-age earth. *Science*, 191:1131–1137.

Dollfus, D. and Beaufort, L. (1999). Fat neural network for recognition of position-normalised objects. *Neural Networks*, 12(3):553 – 560.

Ehrenberg, C. G. (1844). Einige vorläufige Resultate seiner Untersuchungen der ihm von der Südpolreise des Capitain Ross, so wie von den Herren Schayer und Darwin zugekommenen Materialien über das Verhalten des kleinsten Lebens in den Oceanen und den grössten bisher zugänglichen Tiefen des Weltmeeres. *Monatsberichte der Königlichen Preuss. Akademie der Wissenschaften zu Berlin*, pages 182–207.

Ehrenberg, C. G. (1873). Mikrogeologische Studien uber das kleinste Leben der Meeres-Tiefgrunde aller Zonen und dessen geologischen Einfluss. *Konigliche Akademie Wissenschaften zu Berlin, Abhandlungen, Jahre 1872*, pages 131–399.

Fenton, I. S., Baranowski, U., Boscolo-Galazzo, F., Cheales, H., Fox, L., King, D. J., Larkin, C., Latas, M., Liebrand, D., Miller, C. G., Nilsson-Kerr, K., Piga, E., Pugh, H., Remmelzwaal, S., Roseby, Z. A., Smith, Y. M., Stukins, S., Taylor, B., Woodhouse, A., Worne, S., Pearson, P. N., Poole, C. R., Wade, B. S., and Purvis, A. (2018). Factors affecting consistency and accuracy in identifying modern macroperforate planktonic foraminifera. *Journal of Micropalaeontology*, 37(2):431–443.

Granlund, A. (1986). Size and shape patterns in the Recent radiolarian genus *Antarctissa* from a south Indian Ocean transect. *Marine Micropaleontology*, 11:243–250.

Granlund, A. (1990). Evolutionary trends of *Antarctissa* in the Quaternary using morphometric analysis. *Marine Micropaleontology*, 15(3-4):265–286.

Hays, J. D. (1965). Radiolaria and late tertiary and quaternary history of antarctic seas. In Llano, G. A., editor, *Biology of the Antarctic Seas II*, pages 125–184. American Geophysical Union.

Hills, S. J. (1988). Outline extraction of microfossils in reflected light images. *Computer & Geosciences*, 14:481–488.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jörgensen, E. H. (1900). Protophyten und protozöen in plankton aus der norwegischen westküste. *Bergen Museums Aarbog*, 6:51–112.

Keçeli, A. S., Kaya, A., and Keçeli, S. U. (2017). Classification of radiolarian images with hand-crafted and deep features. *Computers & Geosciences*, 109(Supplement C):67 – 74.

Knappertsbusch, M. W., Binggeli, D., Herzig, A., Schmutz, L., Stapfer, S., Schneider, C., Eisenecker, J., and Widmer, L. (2009). Amor - a new system for automated imaging of microfossils for morphometric analyses. *Palaeontologica Electronica*, 12(2):1–20 (web).

Lazarus, D. (2005). A brief review of radiolarian research. *Paläontologische Zeitschrift*, 79(1):183–200.

Lazarus, D. (2011). The deep-sea microfossil record of macroevolutionary change in plankton and its study. In Smith, A. and McGowan, A., editors, *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies*, pages 141–166. The Geological Society, London.

Lazarus, D., Suzuki, N., Caulet, J.-P., Nigrini, C., Goll, I., Goll, R., Dolven, J. K., Diver, P., and Sanfilippo, A. (2015). An evaluated list of Cenozoic-Recent radiolarian species names (Polycystinea), based on those used in the DSDP, ODP and IODP deep-sea drilling programs. *Zootaxa*, 3999(3):301–333.

Lohmann, G. P. (1983). Eigenshape analysis of microfossils: a general morphometric procedure for describing changes in shape. *Mathematical Geology*, 15:659–672.

Lombari, G. and Lazarus, D. (1988). Neogene cycladophorid radiolarians from North Atlantic, Antarctic, and North Pacific deep-sea sediments. *Micropaleontology*, 34(2):97–135.

Moore Jr, T. C. (1973). Method of randomly distributing grains for microscopic examination. *Journal of Sedimentary Research*, 43(3).

Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology*, 2:80–203.

Pan, S. J., Yang, Q., et al. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Petrushevskaya, M. (1967). Radiolyarii otryadov Spumellaria i Nassellaria antarkticheskoi oblasti. *Issledovaniya Fauny Morei*, 4(12):1955–1958.

Petrushevskaya, M. (1975). Cenozoic radiolarians of the Antarctic, Leg 29, DSDP. *Initial Reports of the Deep Sea Drilling Project*, 29:541–675.

Prance, G. T. (1994). A comparison of the efficacy of higher taxa and species numbers in the assessment of biodiversity in the neotropics. *Philosophical Transactions of the Royal Society*, 345(1311):89–99.

Python Software Foundation (2010). Python language reference, version 2.7. http://www.python.org/.

Renaudie, J. and Lazarus, D. B. (2012). New species of Neogene radiolarians from the Southern Ocean. *Journal of Micropalaeontology*, 31(1):29–52.

Richardson, A. J. (2006). Using continuous plankton recorder data. *Progress in Oceanography*, 68(1):27–74.

Schmidt, D. N., Thierstein, H., Bollmann, J., and Schiebel, R. (2004). Abiotic forcing of plankton evolution in the cenozoic. *Science*, 303:207–210.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147.

Tréguer, P., Bowler, C., Moriceau, B., Dutkiewicz, S., Gehlen, M., Aumont, O., Bittner, L., Dugdale, R., Finkel, Z., Iudicone, D., Jahn, O., Guidi, L., Lasbleiz, M., Leblanc, K., Levy, M., and Pondaven, P. (2018). Influence of diatom diversity on the ocean biological carbon pump. *Nature Geoscience*, 11(1):27–37.

Weaver, F. (1983). Cenozoic radiolarians from the southwest Atlantic, Falkland Plateau region, Deep-Sea Drilling Project Leg 71. *Initial Reports of the Deep Sea Drilling Project*, 71(SEP):667–686.

Wiese, R., Renaudie, J., and Lazarus, D. (2016). Testing the accuracy of genus-level data to predict species diversity in cenozoic marine diatoms. *Geology*, 44(12).

Wu, H., Wang, L., Zhang, F., and Wen, Z. (2015). Automatic leaf recognition from a big hierarchical image database. *International Journal of Intelligent Systems*, 30(8):871–886.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zheng, H., Wang, R., Yu, Z., Wang, N., Gu, Z., and Zheng, B. (2017). Automatic plankton image classification combining multiple view features via multiple kernel learning. *BMC Bioinformatics*, 18(16):570.

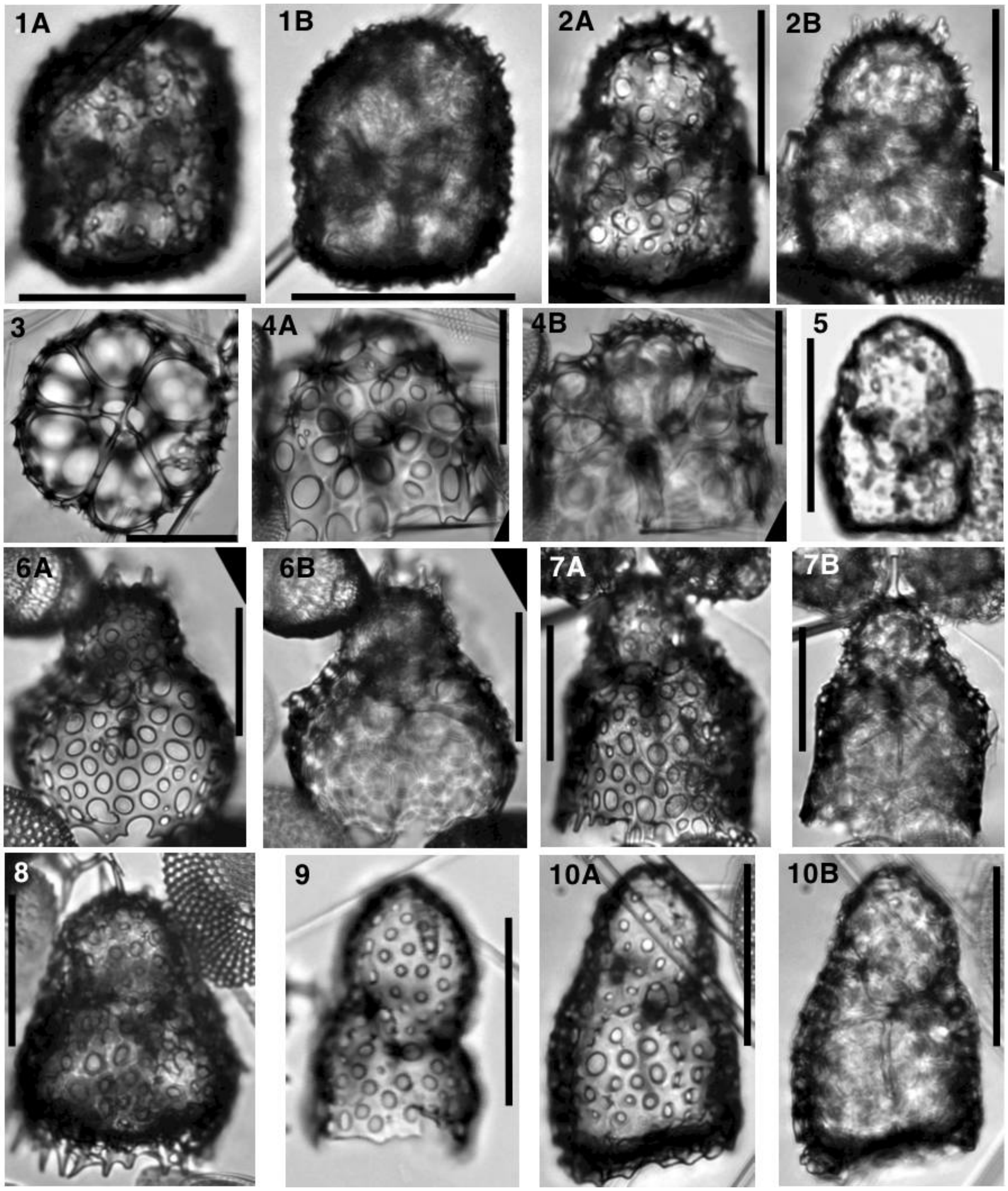

Figure 1. Antarctissid species used in this study. 1. *Antarctissa ballista* with focus on external shell (A) and internal spicules (B). 2. *Antarctissa cylindrica*. 3-4. *Helotholus*? *vema* with a basal view (3), a sagittal view with a focus on the external shell (4A) and on the internal spicule (4B). 5. *Antarctissa robusta*. 6. *Helotholus*? *praevema*. 7. *Antarctissa strelkovi*. 8. *Antarctissa denticulata*. 9. *Antarctissa deflandrei*. 10. *Lithomelissa setosa*. Samples: 278-20-1,77 (5), 689B-3-3,116 (1, 2, 6, 7, 8), 751A-3-4,85 (3, 4, 10), 1138A-17-2,105 (9). Scale bar: 50μm

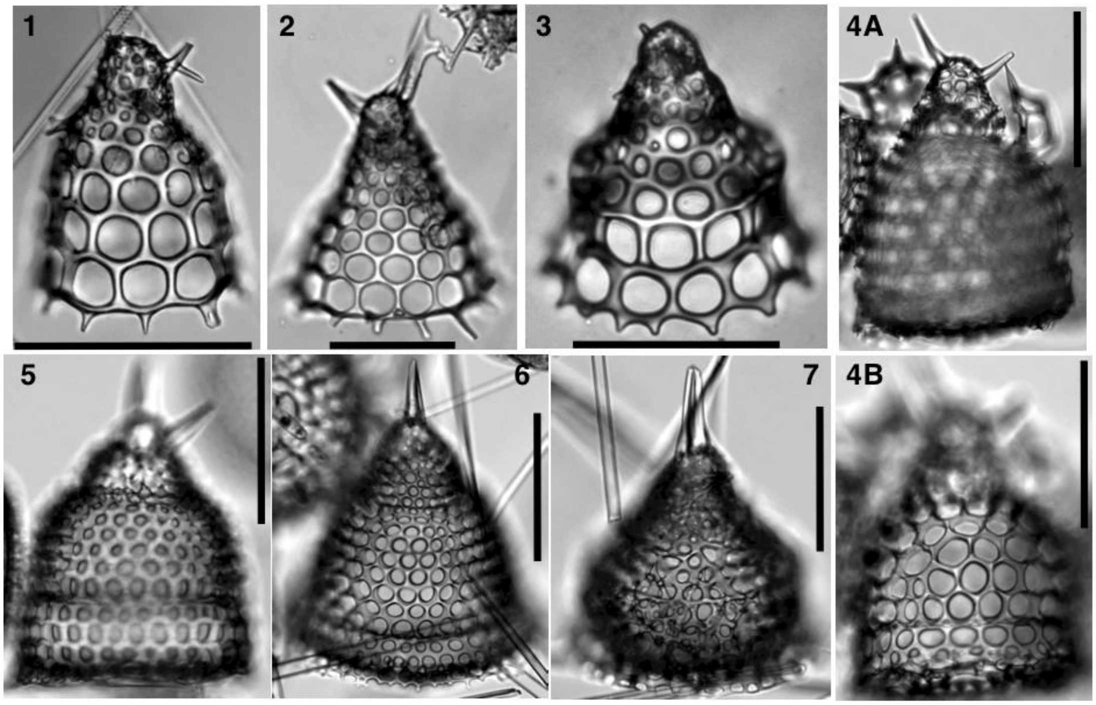

Figure 2. *Cycladophora* species used in this study. 1. *Cycladophora conica*. 2. *C. cosma*. 3. *C. davisiana*. 4. *C. pliocenica*. 5. *C. golli*. 6. *C. humerus*. 7. *C. spongothorax*. Samples: 278-20-1,77 (2), 689B-2-5,55 (4), 751A-1-1,98 (1), 751A-1-2,7 (3), 751A-10-1,98 (6, 7), 751A-17-CC (5). Scale bar: 50μm

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27328v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018

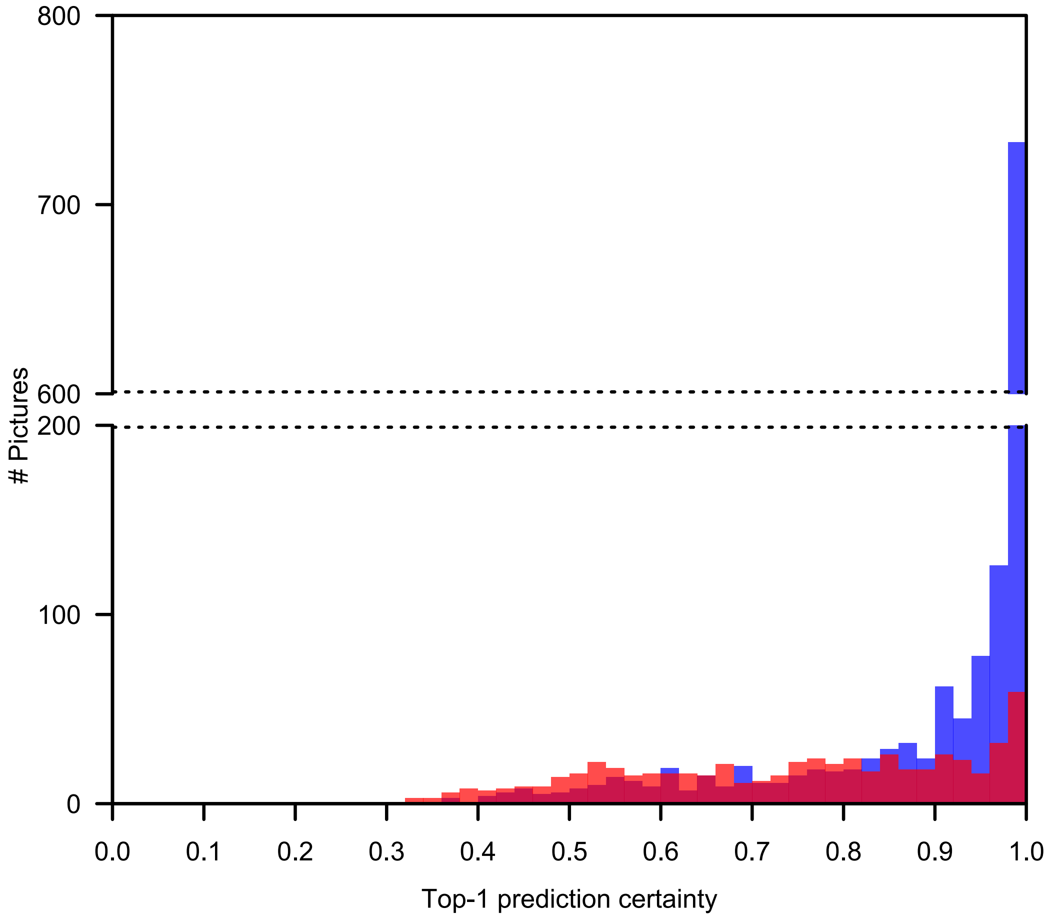

Figure 3. Top-1 certainty scores for correctly (blue) and incorrectly (red) identified pictures (cropped images).

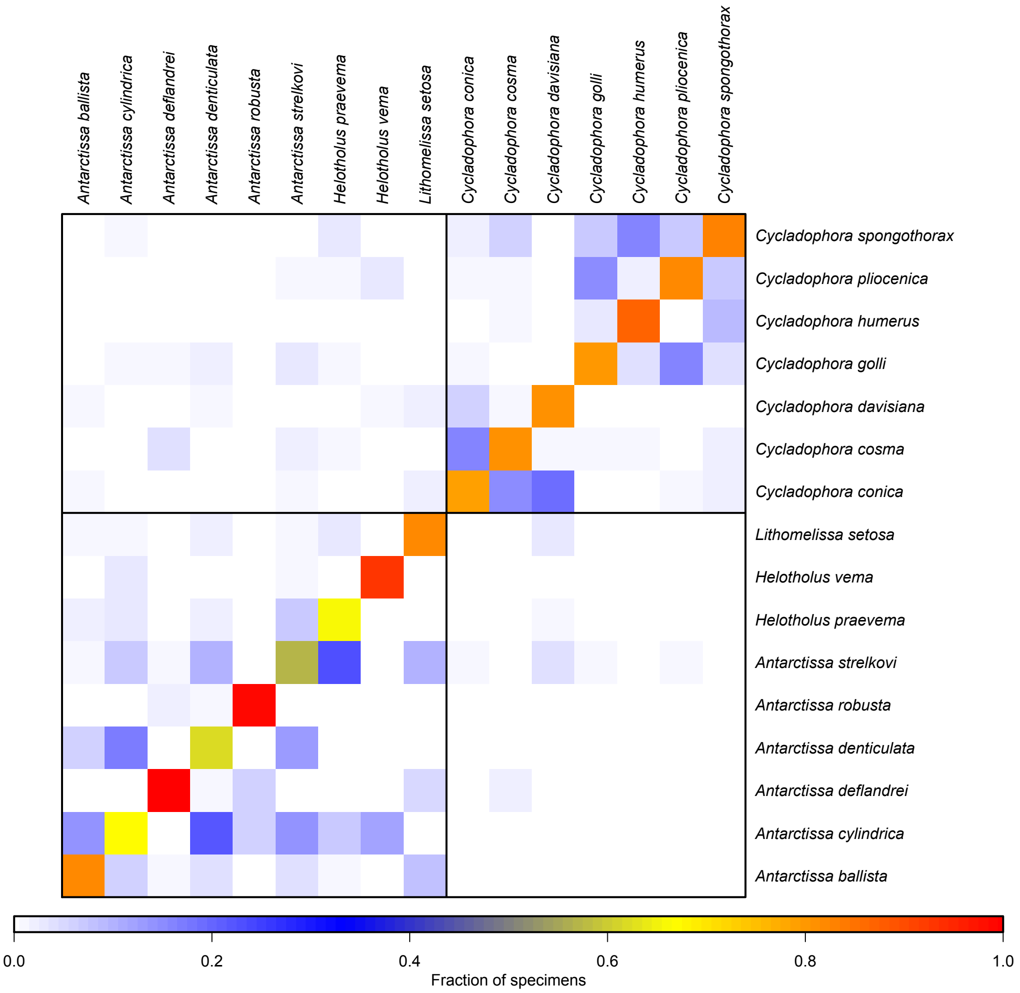

Figure 4. Confusion matrix of the algorithm on the cropped dataset, reported by picture (and not by specimen). Rows correspond to the actual, correct identification of the picture while columns correspond to the algorithm best guess. Cell color corresponds to the relative amount of (mis-)identifications.