Marine Micropaleontology 147 (2019) 16–24

Contents lists available at ScienceDirect

Marine Micropaleontology

journal homepage: www.elsevier.com/locate/marmicro

# Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance

R. Mitra[a,1], T.M. Marchitto[a,b,*], Q. Ge[c], B. Zhong[c], B. Kanakiya[c], M.S. Cook[d], J.S. Fehrenbacher[e], J.D. Ortiz[f], A. Tripati[g], E. Lobaton[c]

[a] Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO 80309, USA
[b] Department of Geological Sciences, University of Colorado, Boulder, CO 80309, USA
[c] Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA
[d] Geosciences Department, Williams College, Williamstown, MA 01267, USA
[e] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, USA
[f] Department of Geology, Kent State University, Kent, OH, 44242, USA
[g] Department of Earth, Planetary, and Space Sciences, Department of Atmospheric and Oceanic Sciences, Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095, USA

ARTICLE INFO

*Keywords:*
Foraminifera
Identification
Automation
Artificial intelligence
Neural network

ABSTRACT

Picking foraminifera from sediment samples is an essential, but repetitive and low-reward task that is well-suited for automation. The first step toward building a picking robot is the development of an automated identification system. We use machine learning techniques to train convolutional neural networks (CNNs) to identify six species of extant planktic foraminifera that are widely used by paleoceanographers, and to distinguish the six species from other taxa. We employ CNNs that were previously built and trained for image classification. Foraminiferal training and identification use reflected light microscope digital images taken at 16 different illumination angles using a light-emitting diode (LED) ring. Overall machine accuracy, as a combination of precision and recall, is better than 80% even with limited training. We compare machine performance to that of human pickers (six experts and five novices) by tasking each with the identification of 540 specimens based on images. Experts achieved comparable precision but poorer recall relative to the machine, with an average accuracy of 63%. Novices scored lower than experts on both precision and recall, for an overall accuracy of 53%. The machine achieved fairly uniform performance across the six species, while participants' scores were strongly species-dependent, commensurate with their past experience and expertise. The machine was also less sensitive to specimen orientation (umbilical versus spiral views) than the humans. These results demonstrate that our approach can provide a versatile 'brain' for an eventual automated robotic picking system.

## 1. Introduction

Much of paleoceanographic and biostratigraphic research depends critically on foraminifera. The total number of described foraminifera species, extinct and extant, is > 50,000 (Hayward et al., 2017), with living species accounting for approximately 10,000 (Vickerman, 1992), the vast majority of which are benthic. Typically measuring up to a few hundred microns in size, foraminiferal tests are common in many modern and ancient marine environments, and as such have become invaluable tools for both academic and industrial purposes, such as paleoenvironmental reconstruction and biostratigraphic age control in paleoceanography and petroleum exploration. The relative abundances of different species serve as paleo-environmental indicators, and the isotopic and trace element compositions of their calcium carbonate tests (e.g., O, C, and B isotopes; Mg/Ca, Cd/Ca, and Ba/Ca ratios) are used to infer paleoceanographic parameters such as global ice volume, temperature, salinity, pH, and nutrient content. There is hardly any laboratory in the world that conducts such research and has not employed personnel (ranging from high school students to senior scientists) to sift through ocean sediments to extract relevant species. However, this process is time consuming and can be costly. Training requires many hours of supervision of novice workers by experts, and it benefits from specialized equipment such as dual head microscopes or digital video display systems, so that multiple researchers can

* Corresponding author at: 450 UCB, University of Colorado, Boulder, CO 80309, USA.
*E-mail address:* tom.marchitto@colorado.edu (T.M. Marchitto).
[1] Present address: Interdisciplinary Programme in Educational Technology, Indian Institute of Technology Bombay, Mumbai, Maharashtra 400076, India.

https://doi.org/10.1016/j.marmicro.2019.01.005
Received 21 September 2018; Received in revised form 11 January 2019; Accepted 12 January 2019
Available online 25 January 2019
0377-8398/ 

simultaneously inspect the same sample. Micropaleontological faunal analysis methods require total counts on the order of 500–1000 individuals per sample to obtain statistically meaningful results. A worker might be required to pick hundreds of specimens of a single taxon to obtain 5–10 mg of calcium carbonate, which is the typical mass used to generate a single radiocarbon measurement by accelerator mass spectrometry. Even the extraction of only a few rare specimens may require searching through thousands of foraminifera, and probably every picker has had the experience of spending hours on a sediment sample, only to end up with too few tests for a measurement. After a steep learning curve, picking therefore becomes a repetitive and low-reward task, making it well-suited for automation.

Attempts to reduce such a workload are not new. Automation efforts, directed entirely toward identification and not actual separation of foraminifera from sediments, started with early computer languages such as Prolog and Lisp. The user would still need to input the characteristics of a specimen, but a computer algorithm helped in their classification based on a predetermined set of taxonomic rules (Swaby, 1992; Athersuch et al., 1994). Automation of a rule-based identification system for diatoms was attempted in the Automatic Diatom Identification and Classification (ADIAC) project, which used image processing techniques to isolate valves on microscopic slide images and classify them based on morphometric parameters (such as valve outline and ornamentation) processed through decision trees (du Buf and Bayer, 2002). The need for expert-defined taxonomic rules can be eliminated through the use of artificial neural networks (ANNs), which learn to identify taxa after being trained with labeled images. SYRACO2 (Dollfus and Beaufort, 1999; Beaufort and Dollfus, 2004) automatically identified coccoliths using a convolutional neural network (CNN) with about 800,000 parameters, wherein transmitted-light microscope images were pre-processed for rotational and translational invariance. Another CNN example is COGNIS, developed at ETH Zürich for the identification of marine calcareous nannofossils using either scanning electron microscope or transmitted-light microscope images (Bollmann et al., 2004). CNNs have also been applied to diatom identification with great success (Pedraza et al., 2017). For foraminifera, a potential problem with ANN classification is the demand on resources: a large number of images are required to train against the vast association of microfossils on the seabed, especially if benthic and extinct taxa are included. A compromise between rule-based classification systems and supervised ANN systems was proposed by Ranaweera et al. (2009), in which the authors identified clusters of foraminiferal associations and assigned a template to such clusters that could be identified by experts, thereby reducing human effort but stopping short of fully-automatic recognition.

Ideally, the recent progress that has been made in terms of automated, semi-automated, or hybrid image classification systems can be fine-tuned to provide the necessary backbone for a fully automated system that can not only identify species, but can also pick foraminifera from sediments. This study is intended to be an initial step toward that holistic goal, focusing only on the image classification aspect. Specifically, we describe the hardware and software for such a visual imaging system, and critically evaluate the image identification capabilities of such a system when compared to humans. The imaging hardware and algorithm are expected to be used in building a completely automated FORAminifera roBOT (FORABOT) that will both identify and pick species of foraminifera or similar sized microfossils.

With this study, the aim is to provide a 'proof of concept' for the identification stage of FORABOT. We focus on few taxa of widely used extant planktic foraminifera. There are only ~50 species of extant planktic foraminifera, and ~15 of these are neglected by most paleoceanographers because of their small size (typically < 150 μm in their adult form) (Schiebel and Hemleben, 2017). Hence a full characterization of the > 150 μm planktic foraminiferal assemblage may eventually be attainable using image recognition. However, we chose to initially focus on only a few taxa that are particularly important for geochemical proxy measurements and certain census counts. Instead of *en masse* classification of foraminiferal species, this method is expected to reduce and simplify computation and, if successful, can then be replicated to include other species.

## 2. Materials and methods

### 2.1. Core material

Foraminifera were obtained from six marine sediment core samples. Two core sites are in the North Atlantic: DSDP 552A (7–9 cm) from Rockall Plateau (56°03′N, 23°14′W, 2311 m); and HU77029 (902 cm) from Baffin Bay (66°54′N, 58°18′W, 935 m). Four sites are in the Pacific: ODP 807A (22–24 cm) from Ontong Java Plateau (3°36′N, 156°37′E, 2805 m); RC13–114 (319 cm) from the eastern equatorial Pacific (1°39′S, 103°38′W, 3436 m); MV99-PC08 (721.5 cm) from off Baja California (23°28′N, 111°36′W, 705 m); and MV99-PC14 (709–711 cm) also from off Baja California (25°12′N, 112°43′W, 540 m). Samples range in age from Marine Isotope Stage 3 to Holocene. DSDP 552A, HU77029, and ODP 807A provided the bulk of the specimens for this study.

### 2.2. Species selection and identification

Foraminifera were identified and picked from the washed 250–355 μm size fraction, by an expert employing the standard practice of specimen manipulation using a wet brush under a binocular microscope. We selected six species of common, extant planktic foraminifera that are widely used by paleoceanographers (Fig. 1). *Globigerinoides ruber* and *Globigerinoides sacculifer* are commonly used to assess sea surface conditions in low latitudes due to their shallow (mixed layer) habitats (e.g., Spero et al., 2003). In middle latitudes and upwelling regions, emphasis shifts to the more abundant *Globigerina bulloides*, whose abundance has also been used as a proxy for upwelling strength in lower latitudes (e.g., Gupta et al., 2003). At high latitudes, *Neogloboquadrina pachyderma* and *Neogloboquadrina incompta* (formerly known as *N. pachyderma* dextral; Darling et al., 2006) tend to dominate, with the abundance ratio between these two species having historically served as a useful temperature proxy (Ericson, 1959). A number of subsurface taxa may be used to reconstruct properties within the thermocline, but *Neogloboquadrina dutertrei* has probably received the most attention (e.g., Spero and Lea, 2002).

Of the aforementioned species, manual identification of *G. ruber*, *G. sacculifer*, and *G. bulloides* is relatively straightforward. When present, the 'sac-like' final chamber is very diagnostic for *G. sacculifer*. When the sac is absent, the relative size and shape of the final chamber helps in identifying and differentiating between *G. sacculifer* (Fig. 1f) and *G. ruber* (Fig. 1a). Furthermore, the orientation of apertures, test texture, and sometimes color (for the pink form of *G. ruber*) all serve to readily differentiate the two species. *G. bulloides* is equally discernible, with low trochospiral, spherical to subspherical chambers, and a high symmetrical arch-like umbilical aperture (Fig. 1b).

The differentiation of *N. pachyderma*, *N. incompta*, and *N. dutertrei* requires a more nuanced approach. The most significant issue arises in the classification of *N. dutertrei* and *N. pachyderma*, as they form more of a morphological continuum. Indeed, gradational morphotypes between the two, dubbed "P-D intergrades" by Kipp (1976), are quite common. The difference between *N. pachyderma* and *N. incompta* is more straightforward in that they are normally distinguished by coiling direction, left (sinistral) and right (dextral), respectively. However, genetic investigations show that a small percentage of each species (up to ~3%) exhibit aberrant coiling direction (Bauch et al., 2003; Darling et al., 2006). Darling et al. (2006) recommend that in a > 97% sinistrally-coiled population, the small fraction of dextrally-coiled specimens should be identified as dextrally-coiled *N. pachyderma* and not *N. incompta*. The converse would be true for a > 97% dextrally-coiled

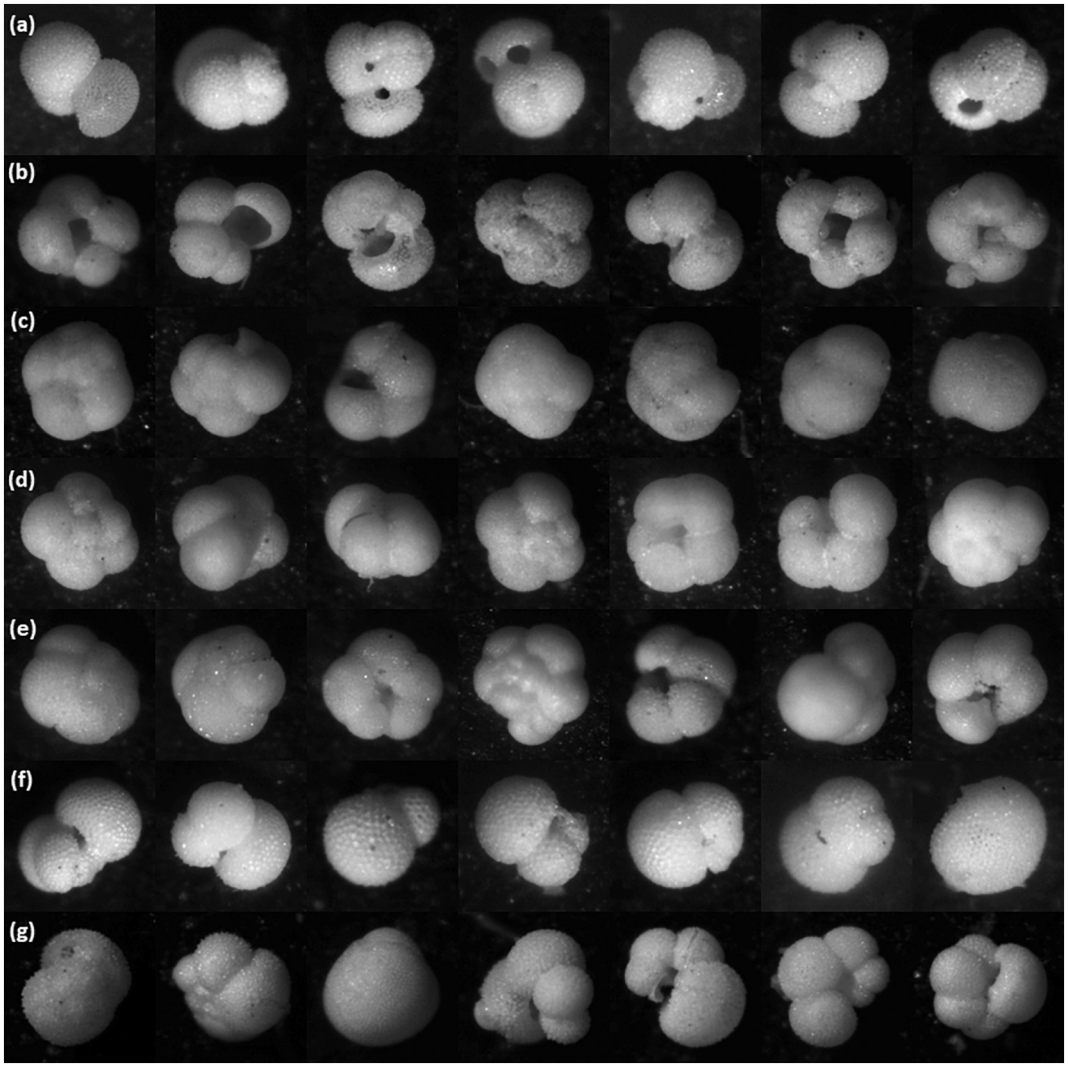


**Fig. 1.** Representative specimens of the species identified for this study. (a) *G. ruber*, (b) *G. bulloides*, (c) *N. pachyderma*, (d) *N. incompta*, (e) *N. dutertrei*, (f) *G. sacculifer*, and (g) other species. Each photograph is one of the 16 frames taken for each specimen by our imaging system.

population; whereas coiling ratios between 3 and 97% would more likely correspond to a true mixture of the two species. In fact, differences in opinion on species/sub-species level classification are common, even among experts. An entire branch of micropaleontology, morphometric analysis, looks at such subtle variation of forms and their biogeographical significance, and machine learning may be useful there as well (e.g., Beszteri et al., 2018). However, the present scope of our study is much more limited. Therefore, we decided to follow a set of *Neogloboquadrina* rules that are largely consistent with those of Thompson (1976), Gardner and Hays (1976), Hilbrecht (1997), Darling et al. (2006), and Eynaud et al. (2009). We do not claim this scheme to be definitive. Instead, we simply explore the possibilities of such rule-based classification for automated identification of foraminifera.

Within the current context of the three *Neogloboquadrina* species, we have classified any specimen with 4 to 4.5 chambers in the final whorl (with or without an apertural lip) as belonging to either *N. pachyderma* or *N. incompta*, with only the coiling direction distinguishing between the two species (Fig. 1c-d). Most specimens that we successfully identified in this manner had a lip on the final chamber. Specimens with 5 to 6 lobate chambers in the final whorl and with a deep set central umbilicus were classified as *N. dutertrei* (Fig. 1e). Almost half of these specimens had one or more teeth projecting into the umbilicus. Specimens with contradictory and/or borderline evidence, such as those with five or more chambers, but with a lip on the final chamber or without a deep-set central umbilicus, were generally excluded from this study. While such specimens could be considered as P-D intergrades, which Hilbrecht (1997) suggested are morphological variants of *N. pachyderma*, we omit them to minimize the effect of different taxonomic concepts among our human pickers.

### 2.3. Image collection

We aim to develop an affordable system capable of automatically identifying foraminifera using a reflected light microscope. For our experiments, we used an inexpensive AmScope SE305R-PZ binocular microscope at 30× magnification (Fig. 2). We randomly oriented each foraminifer alone near the center of the microscope's field of view, and manually brought the specimen into focus. Since the direction of the light source is an essential factor for highlighting different geometric features in the foraminifera (Ranaweera et al., 2009), we made use of a Light Emitting Diode (LED) ring to get 16 images under various lighting angles (eight such images are shown in Fig. 2). The placement of the LED ring kept the light sources at the same height, and the lighting orientation was changed from 0 to 360° with a uniform spacing. The images were captured using an AmScope MD500, a 5MP USB camera that attaches directly to the microscope and provides an approximate resolution of 450 × 450 pixels. The LED ring was controlled using an Arduino UNO Microcontroller. The camera and Arduino were

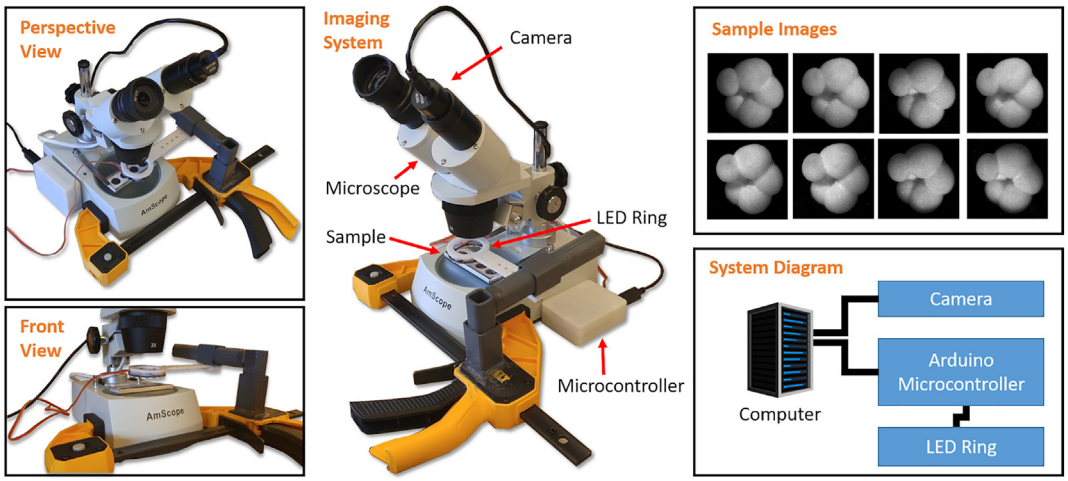


**Fig. 2.** Image acquisition system: Pictures of the system from different perspectives (left) with main components highlighted (middle). Eight images of the same foraminifer specimen using multiple lighting directions (top right). Diagram of the main components of the system (bottom right).

connected to a computer via Universal Serial Bus (USB). An application developed in MATLAB was used for image capture. This design is open source with a list of components, schematics, CAD models, and code available online (https://research.ece.ncsu.edu/aros/foram-identification/).

A data set of 1437 foraminifera was photographed for this study and can be found at https://doi.pangaea.de/10.1594/PANGAEA.897873. This data set includes: *G. bulloides* (178 specimens), *G. ruber* (182), *G. sacculifer* (150), *N. dutertrei* (151), *N. incompta* (174), *N. pachyderma* (152), and examples of other species (450, almost entirely planktic foraminfera). A larger proportion of 'Others' was imaged in order to better capture the larger morphological variability of this group.

### 2.4. Automated identification of foraminifera

Recently, deep CNNs have triggered a revolutionary change in the image classification community. CNNs comprise a special category of ANNs where the hidden layer(s) convolve pixel values of the original image with a filter to extract important features such as edge, color, noise, etc. Although convolutional layers reduce the number of parameters to fit by using parameter sharing techniques, powerful CNNs are usually very deep (have many layers) and thus still require a large amount of training data to fit all the parameters (Goodfellow et al., 2016). However, various pre-trained models have been published and the different levels of features learned from the original data sets have promising potential to be transferred to new data sets (Pan and Yang, 2010). Among these models, Vgg16 (Simonyan and Zisserman, 2015) and ResNet50 (He et al., 2016) are most representative and show competitive performance on various computer vision tasks. In this study, both models are utilized to extract features of the foraminiferal images. We evaluated and compared various methods for recognition, including more standard techniques and features, and determined that these two CNNs gave the best performance (Zhong et al., 2017).

Whereas human experts have to decide which features (e.g., color, texture, contour shape) might be useful for different classification tasks, given sufficient data, the CNNs have the potential to automatically learn which features are important to distinguish different classes. A CNN image classification model is usually made up of several convolutional layers followed by fully-connected (FN) layers (Fig. 3 top). A convolutional layer can be thought of as a function applied at every location in an image to extract local image descriptors that are useful for the recognition task, while the fully-connected layers are the more traditional ANN layers that combine these descriptors in some non-linear ways to come up with a global function that outputs the predicted species. The outputs of the convolutional layers are often referred to as feature maps since they are associated with particular locations in the images. During the training process, the parameters of these functions are optimized to extract discriminative features for the classification task. For example, the low-level convolutional layers may focus on edges or textures, and the middle layers can then combine the edges to be different shapes such as circles or triangles. Finally, the last, high-level convolutional layers transform the shapes to be discriminative components of the objects which are useful for recognition.

Due to the limited size of our foraminifer data set, we did not train the deep neural networks from scratch; instead, we adopted two CNNs that were already trained using the ILSVRC data set (Russakovsky et al., 2015; Deng et al., 2009), which is a large scale image classification data set with 1000 categories and 1.2 million images. The pre-trained CNNs are well-trained to extract different levels of features for common images (e.g., dogs, airplanes). Although our data set is different from common images, the pre-trained models have the potential to compute informative feature maps that can be generalized to a different data set. The usage of pre-trained CNNs reduces the number of images for training from hundreds of thousands to a few thousands.

The pre-trained models expect color images as input. Our pipeline (Fig. 3 bottom) therefore begins with the creation of a single color image that encodes information from all 16 grayscale images of a given specimen. A Python code takes the 16 individual grayscale values for each pixel, calculates the 10th percentile, median, and 90th percentile, and maps those three values into the red, green, and blue channels, to generate a single color composite image. This image is then fed to the two pre-trained CNNs' convolutional layers, which produce the feature maps. Although the two CNN models are pre-trained with the same data set (ILSVRC), different networks learn to classify in different ways, leading to features that can be complementary to each other. Thus, by concatenating the output features of Vgg16 and ResNet50, the combined model achieves better performance.

The resulting feature maps (i.e., the outputs from the pre-trained convolutional layers) are then passed on to three new fully-connected layers, with dropout layers in between, for classification. That is, the fully-connected layers compose a new function from the features computed by the convolutional layers, in order to recognize the different species of foraminifera. The dropout layers are used to prevent overfitting. They work by randomly removing connections in the network during training to ensure that the ANN can do recognition even if certain features are missing. This has the effect of making the model more robust to variations in the input; hence, performing better when testing new datasets. Detailed descriptions of the functionality of different layers can be found in Krizhevsky et al. (2012). For the three new fully-connected layers, the first two layers have 512 units, and the last

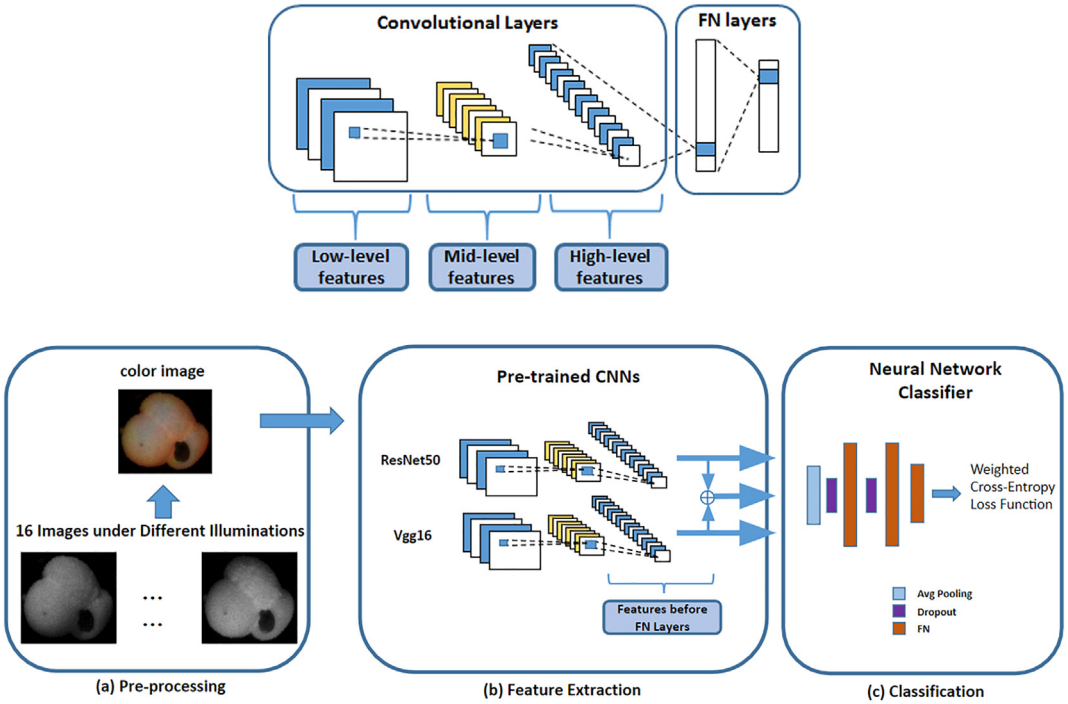


**Fig. 3.** Top: General CNN pipeline. Bottom: Classification pipeline for foraminifera.

layer has 7 units since we have 7 classes to classify. The dropout rates are 0.05 and 0.15 for the first and second dropout layers, respectively (the layers' order is shown in Fig. 3). With comprehensive experiments, Yosinski et al. (2014) demonstrated the effectiveness of transfer learning with pre-trained CNNs as the feature extraction algorithms. Since the identification of foraminifera using deep learning techniques is new, our detailed design (e.g., the number of units of the fully-connected layers and the dropout rates) is based on our experience with natural images, and the manual tuning of hyper-parameters. Alternate classifiers were also explored, namely Random Forest, Support Vector Machine, and K-Nearest Neighbors, but the CNN gives the best performance (Zhong et al., 2017).

The neural networks are implemented in Python using the Keras framework (Chollet, 2015) with TensorFlow (Abadi et al., 2016) as the backend engine. The machine used for experimentation has an Intel Core i7 CPU, 64GB of RAM, and a NVIDIA TITAN GPU. To speed up the training process, we save the pre-trained features for all the samples first, and then train the new layers with those features as input. The entire training process takes around 15 min with 800 epochs (training cycles) and 1000 samples. In testing, it takes around 250 ms for the prediction of one image with ResNet50 + Vgg16.

### 2.5. Human identification of foraminifera

In order to compare the performance of the autonomous foraminifer recognition algorithm against human identification, we collected data from six 'experts' and five 'novices.' The experts included five college professor-level paleoceanographers who have extensive experience (> 15 years each) working with planktic foraminifera, plus the lead author who oversaw the picking of the specimens in this study. The five novices were undergraduate and graduate students with some experience picking foraminifera in the labs of the experts. Novice experience ranged from less than six months to ~2 years. All participants were asked to characterize their familiarity with the six target species as 'High,' 'Medium,' or 'Low.'

For the comparison exercise, the humans and the machine were provided with the same 540 specimens to identify, to ensure fair comparison across the two groups. Each human participant was shown an electronic page with all 16 differently-illuminated images of a given foraminiferal specimen, and was asked to identify the species from a drop-down spreadsheet menu. The choices included the six species considered here, 'Other,' and 'Not Identifiable.' Participants were instructed to choose 'Other' when they thought the specimen was a species other than the six listed, and 'Not Identifiable' when they judged the images to be insufficient to make any decision. Instructions also indicated that participants were allowed to use any references they wished, and an internet link was provided to the scanning electron microscope images from Bé (1977), but that our intent was for them to make identifications primarily from memory, to simulate typical picking practices. Participants labeled the foraminifera in 10 groups of 54 specimens each, for a total of 540 identifications. The photographs covered 60 specimens of each species, plus 180 Others, in random order. This exercise was meant to compare the performance of the automated system to humans when given the same imperfect, single-orientation images.

### 2.6. Performance metrics

To evaluate the performance of both the automated system and humans, we used standard classification metrics, namely: precision, recall, and F1 score (Sokolova and Lapalme, 2009). In any classification task, true positives (TP) are the instances in which an item was correctly classified under a target category, false positives (FP) are those in which an item was incorrectly classified, true negatives (TN) are those in which an item was correctly excluded from the category, and false negatives (FN) are those in which an item was incorrectly excluded. Therefore, within each taxonomic category, precision is the ratio TP/(TP + FP), and recall is the ratio TP/(TP + FN). In other words, for each taxonomic group, precision is the fraction of identifications that is correct, and recall is the fraction of specimens that are recovered. F1 score is the harmonic mean of precision and recall, which is often used as an overall indicator of classification performance. There is a tradeoff between precision and recall. If the identifier is being 'overzealous' he/she will have good recall but poor precision, because Other or Not Identifiable (ambiguous) species will be mislabeled as target species. On the other hand, one can be 'overcautious,' resulting in good

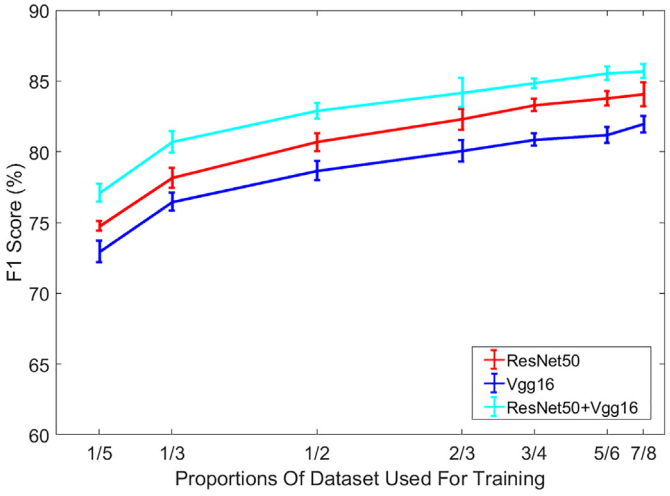


**Fig. 4.** Foraminiferal classification performance (F1 scores) of the two CNNs used for this study, plotted against the fraction of the 1437 specimens that was used for training. Scores are averaged across the six target species, and error bars indicate the standard deviation of repeated iterations (see text for details). The best performance corresponds to the model that uses the concatenation of features from both CNNs, labeled as ResNet50 + Vgg16.

precision but poor recall because target species are mislabeled as Others or deemed Not Identifiable. When reporting these metrics averaged across the six target species, we weighted them according to the species' proportions in the data set.

## 3. Results and discussion

### 3.1. Machine performance

Given the tradeoff between precision and recall, we use a loss function to find similarly good performance for each, which maximizes the F1 score (Zhong et al., 2017). The classification performance of each computer vision algorithm is closely related to the amount of data used for training, with larger data sets leading to better performance (Fig. 4). To obtain training and validation sets of different sizes, the 1437 total specimens were first split into k groups (k = 2, 3, 4, 5, 6, or 8), and then either 1/k or (k-1)/k specimens were used to train the classification of the remainder. The process was repeated until every group had been used for training once, and the whole procedure was further replicated 10 times, with the mean and standard deviation plotted. With the current amount of data, the ResNet50 + Vgg16 overall F1 score is ~86% when ⅞ of the images (1257 specimens) are used for training. Having been thus trained on ~130–160 specimens of each of the six target species (and ~400 Others), the machine should be able to achieve ~86% accuracy when applied to any new sediment sample, as long as the specimens in this study have adequately captured the morphological variability that is likely to be encountered. There exists an increasing trend at the end of Fig. 4, which means that with more training data, the proposed algorithm has the potential to have even higher classification accuracy.

We are not aware of any published reports of automated foraminiferal identification performance to compare our results to. For 11 species of coccoliths, SYRACO2 was able to achieve recall of better than 90% using a similar level of training (150 images per species), but precision was poorer, with ~40% of non-coccolith particles being misclassified as coccoliths (Beaufort and Dollfus, 2004). Various diatom classifiers, mostly rule-based but including a CNN, report accuracies between ~80 and 99.5% (summarized by Pedraza et al., 2017). The best performance for diatoms (99.5% across 80 species) was achieved using a CNN trained on 11,000 specimens augmented with rotation and flipping to produce a training set of 160,000 images (Pedraza et al., 2017). However, the diatom results are not directly comparable to ours due to differences in experimental design, such as a lack of non-target species in the exercises.

To compare machine performance to human performance, we needed a training set that is substantial yet leaves enough data for the actual classification task. After removing the 540 images to be used for the comparison exercise, the remaining 897 images (62%) were used for training the CNN. With this smaller training set (~90–120 specimens of each species), the machine F1 score was reduced to 81% (Fig. 5). Reported machine statistics are average values across 10 iterations of the exercise, using the same training and unknown data sets each time.

### 3.2. Human performance and comparison to machine

Averaged across the six target species, expert precision ranged from 59 to 83% (mean 74%), and recall ranged from 32 to 83% (mean 60%), resulting in F1 scores between 39 and 83% (mean 63%) (Fig. 5). As expected, novice performance was lower on average, with precision of 49–65% (mean 56%), recall of 47–64% (mean 53%), and F1 scores between 47 and 63% (mean 53%). In comparison, the machine had precision (80%) comparable to experts and better than novices; and recall (82%) and F1 scores (81%) better than 10 out of 11 participants.

It is important to note that the participants were given an option that the machine did not have, namely to label specimens as Not Identifiable. Theoretically, this choice could improve precision by not forcing participants to guess, but it could hurt recall for the same reason. In particular, one expert was very cautious and used this option

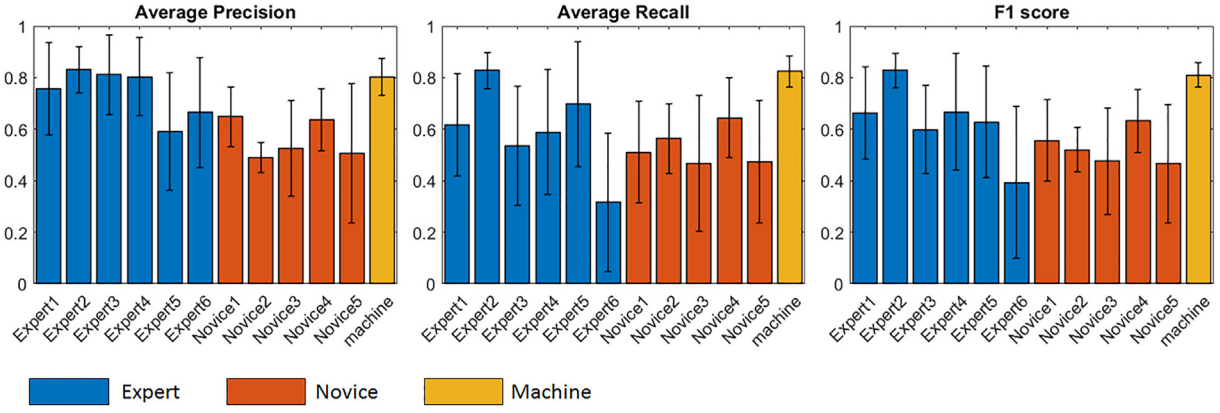


**Fig. 5.** Precision, recall, and F1 score for each expert, each novice, and the machine, averaged across the six target species, with ± 1σ error bars to indicate the spread across taxa. The machine scores are additionally averaged across 10 iterations.

often, resulting in poor recall and therefore the lowest F1 score. In this sense, our exercise may overestimate the precision and underestimate the recall abilities of the participants, relative to a scenario in which they are forced to decide. If we reanalyze the results excluding the Not Identifiable selections, overall recall improves to 67–88% (mean 79%) for experts and 52–73% (mean 62%) for novices. However, this would limit the analysis to those specimens that each participant felt reasonably confident about, resulting in overestimation of their true recall abilities.

Although both the machine and humans were given the same images to identify, it is possible that the humans were more negatively impacted by image quality (cf. Fig. 1) than the machine. The machine was trained on images of similar quality, while humans are trained with the ability to change light intensity, magnification, focus, and specimen orientation (orientation is discussed in Section 3.4). We note, however, that the machine performed poorly on some under-illuminated specimens that otherwise should have been easily identified.

In identifying foraminifera, the relative importance of precision and recall depends on the task at hand. If picking specimens for geochemical analysis, the picker needs to achieve excellent precision, but they can afford to be selective (low recall) as long as the species is sufficiently abundant. In contrast, assemblage work requires high recall because all specimens need to be identified. Our automated system was able to optimize both precision and recall at ≥80% in this exercise, which exceeded the overall performance of all but one participant. Three other experts who had similar precision to the machine achieved that at the expense of lower recall.

### 3.3. Comparison by species

The automated system achieved comparable performance across the different taxa, including Others, with F1 scores ranging from 76 to 89% (Fig. 6). Only one species, *N. dutertrei*, had machine precision lower than 70%, and all taxa had recall scores above that value. Since we purposely excluded ambiguous P-D intergrades (Section 2.2), it is not yet clear how machine performance on *N. pachyderma* and *N. dutertrei* would be affected by including them. A true morphological continuum would require the human experts who label the training data set to assign an arbitrary division, and close to that division both machine and humans would likely have difficulty making a firm decision. In such a case, the concept of distinct morphospecies may be of limited value.

In contrast to the machine, the performance of the human participants varied considerably between taxa. Highest average F1 scores were for *G. ruber* and *G. sacculifer* among both experts (85 and 82%, respectively) and novices (71 and 63%). *Neogloboquadrina. incompta* and *Neogloboquadrina. pachyderma* had the lowest F1 scores among both experts (38 and 44%) and novices (36 and 33%). For *Globigerina. bulloides*, *Globigerinoides. ruber*, *Globigerinoides. sacculifer*, and Others, average expert precision was better than machine precision, but expert recall and F1 scores were worse than the machine for all taxa except *Globigerinoides. sacculifer*. This illustrates that the overall superior performance of the machine (Fig. 5) is partly a result of its taxon-independent abilities, compared to the taxon-specific pitfalls that exist for most of the human participants.

Insight into the reasons for poor performance on certain species may be gleaned from a confusion matrix (Fig. 7). This diagram compares assigned identities to true identities, thereby illuminating which species are mistaken for others. Among the experts, *Neogloboquadrina. incompta* and *Neogloboquadrina. pachyderma* were sometimes mislabeled as *Neogloboquadrina. dutertrei*. *Neogloboquadrina. incompta* and *Neogloboquadrina. pachyderma* were less often confused with each other, indicating that coiling direction was not easily mistaken. The novices more frequently confused *G. ruber* and *G. sacculifer* than the experts did. Like the experts, the novices mislabeled some *N. incompta* and *N. pachyderma* as *N. dutertrei*, and they additionally confused *N. incompta* and *N. pachyderma*, suggesting some lack of clarity about coiling direction. In both groups of participants, the Not Identifiable category steals some identifications and contributes to low recall, most notably for *N. incompta*, *N. pachyderma*, and Others. Interestingly, the machine occasionally mislabeled *G. sacculifer* as *N. dutertrei*, a rather egregious mistake that the humans never made.

There are at least three factors that can potentially explain the taxon-dependent performance of the human participants. First, most pickers are not equally expert in all planktic foraminiferal species. This uneven expertise may arise from working predominantly in one part of the ocean (e.g., the tropics or the Arctic) where not all taxa will be routinely encountered; or, in the case of novices, being trained to pick only a few target species relevant to the laboratory's current work. Averaged across the experts and across the novices, species-level performance correlates with self-reported familiarity (Fig. 8). All six experts reported high familiarity with *G. ruber* and *G. sacculifer*, commensurate with their high F1 scores on those two species (85 and 82%, respectively). The species with the lowest scores, *N. incompta* and *N. pachyderma*, were the least familiar to the average expert and, especially, novice. Second, as described in Section 2.2, there can be differences in taxonomic concepts among different pickers. That is, even a self-identified expert in the Neogloboquadrinids might not fully agree with our ground-truth labels. We tried to minimize this effect in our study by avoiding P-D intergrades, but it could still play a role. Third, the inability to flip specimens over may have been a handicap, as discussed in the next section.

### 3.4. Effect of orientation

Identification of foraminifera by human pickers often involves multiple views. When unsure, a picker may turn the specimen over for confirmation, as was frequently done when the specimens in this study were first selected. Admittedly, that affordance was not available in this

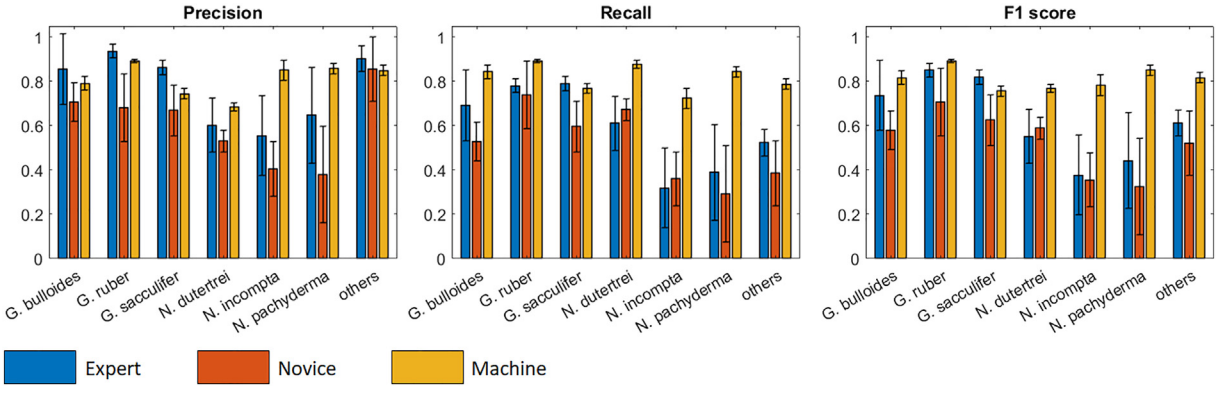


**Fig. 6.** Precision, recall, and F1 score for each taxon, averaged across the experts, the novices, and the machine iterations, with ± 1σ error bars.

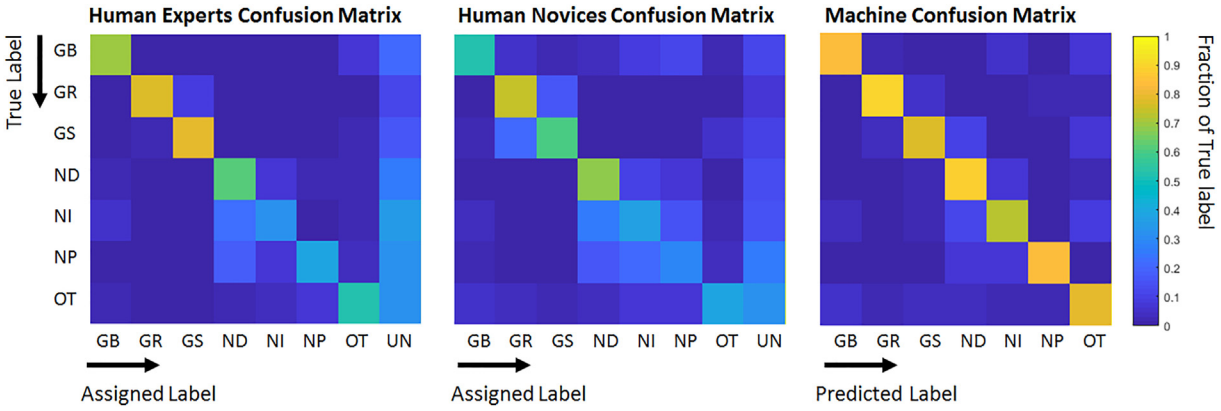


**Fig. 7.** Confusion matrices averaged across the experts, the novices, and the machine iterations. Each row shows how the 'true' images of a given species were labeled during the exercises (called 'assigned' in the case of humans and 'predicted' in the case of the machine). Each row sums to 1, and perfect performance would be indicated by yellow squares (values of 1) across the diagonal. The fraction given on the diagonal is equivalent to the recall for that species. Labels are ordered as in Fig. 6, with the addition of UN for 'unidentifiable.' (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exercise. Instead, the participants were asked to identify a foraminifer based on imperfect photographs taken from a single point of view (but multiple illuminations). Four of the six species considered here have very distinct umbilical versus spiral sides: *G. bulloides* and the three Neogloboquadrinids. The impact of orientation on the participants' skill was most dramatic for *G. bulloides* (Fig. 9). Typically, experts and novices alike converged on positive *G. bulloides* identification only when an umbilical view was provided. Human recall was substantially degraded when a spiral view was provided, especially for novices. It is in such circumstances that most pickers would flip the specimen over to provide a more diagnostic view. For spiral and other unconventional (side) views, we usually found a greater number of experts than novices identifying *G. bulloides* correctly. This validates the notion that experts have had more experience and thereby are more acquainted with unconventional views. Machine performance, in comparison, suffered less when comparing spiral views to umbilical views. However, the machine did poorly with some side views and atypical (e.g., poorly illuminated) spiral views in comparison to humans.

*Neogloboquadrina incompta* and *N. pachyderma* were similar in that human performance was moderately better when based on umbilical views, but the difference was less dramatic than for *G. bulloides*, probably because overall skill was poorer. Machine performance was slightly better when using the spiral side for *N. incompta*, but for *N. pachyderma* the orientation made no notable difference. For *N. dutertrei*, participant skill was higher when provided with a spiral view,

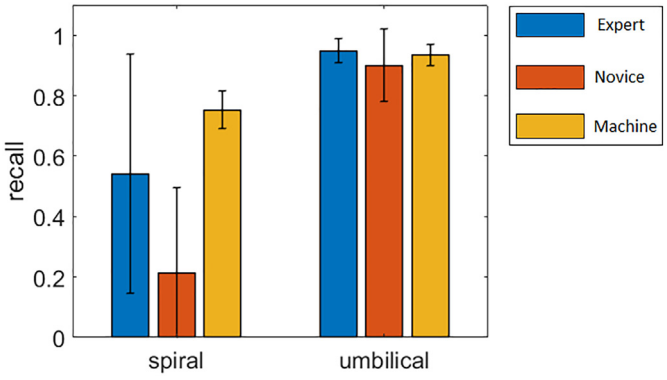


**Fig. 9.** Recall for spiral and umbilical views of *G. bulloides*, averaged across the experts, the novices, and the machine iterations, with ± 1σ error bars.

especially among novices. Machine performance was nearly identical between umbilical and spiral views of *N. dutertrei*.

Overall, human pickers appear to be more sensitive to specimen orientation than the machine. That is not an indictment of humans, because they are not normally constrained to one view and hence have little incentive to learn taxonomy equally well from all angles. Conversely, this observation is critical for the success of an automated system, since specimens need not be oriented in any particular

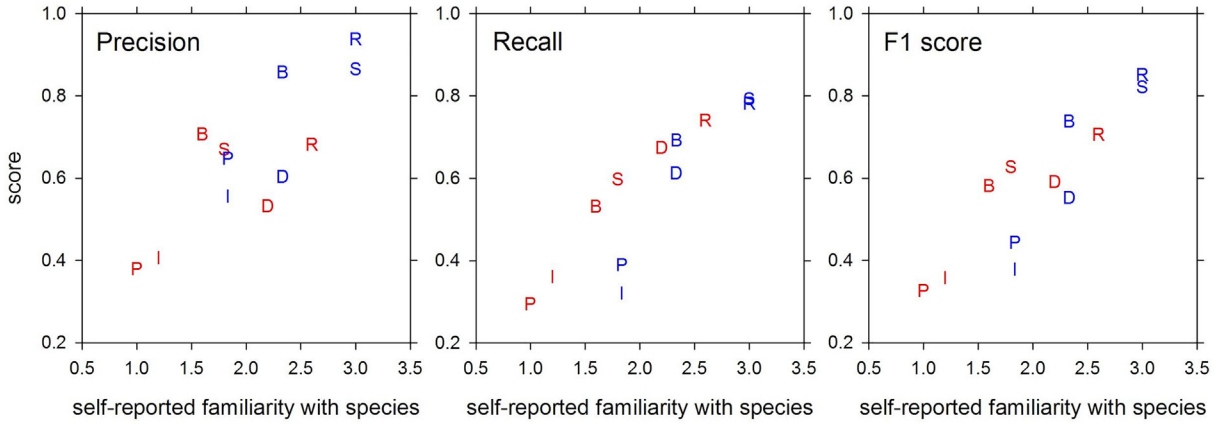


**Fig. 8.** Species-level performance averaged across experts (blue) and novices (red) as a function of self-reported level of familiarity (1 = low, 2 = medium, 3 = high). Symbols refer to the first letter of the species name. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

direction.

## 4. Conclusions and outlook

We successfully trained a CNN-based system to identify six species of extant planktic foraminifera using reflected light microscope digital images. Machine performance was compared to that of six human experts and five novices, wherein all were tasked with identifying the same 540 specimens. For this exercise the machine was trained on ~90–120 images of each species plus 270 Others, and achieved 80% precision, 82% recall, and an F1 score of 81%. Experts displayed comparable precision but poorer recall, resulting in an average F1 score of 63%. Not surprisingly, novices displayed less skill, with an average F1 score of 53%. Human performance was compromised by being species-dependent, which we attribute mainly to training and experience. For some species, most notably *G. bulloides*, the machine was less sensitive to specimen orientation than the human participants.

Overall, the machine's strength in this exercise is grounded in its relatively uniform performance across taxa and orientations. These results suggest that similarly good results could be achieved for the entire extant planktic foraminiferal assemblage above some nominal size, ideally 150 μm. Select benthic foraminifera could also be added. By increasing the training set to several hundred or more specimens per species, performance could theoretically be improved beyond the scores presented here. Performance might also benefit from higher quality microscopes and cameras, especially for smaller specimens. We therefore conclude that our CNN approach can provide the 'brain' for a viable robotic picking system.

It is neither desirable nor advisable to eliminate human expertise from this process. Rather, we envision FORABOT as a labor-saving device to execute the bulk of a given picking task, whether that be the characterization of the full assemblage or the removal of a single species for chemical analysis. Once that task is completed with ~80–90% accuracy, a person would then validate and, as needed, correct the identifications. By allowing the human picker to focus on subtle differences such as morphotypes or intergrades, we suggest that they will quickly achieve a deeper level of taxonomic expertise than is currently practical in most laboratories.

## Acknowledgments

We thank Payton Birdsong, Riff Denbow, Ben Elliott, Brigitta Rongstad, and Brenna McBride for participating in this study as novices; Jean Lynch-Stieglitz, Niklas Meinicke, Minda Monteagudo, Valerie Voisin, and Jody Wycech for participating in a test phase of the identification exercise; Anne Jennings for supplying the sample from HU77029; and Rohini Sharma and Charles Hirsch for hardware and software design of the image acquisition system. This work was supported by US National Science Foundation grant OCE-1637023 to Lobaton and Marchitto. Tripati was supported by US Department of Energy grant DE-SC0010288. Images used for this work are available at https://doi.pangaea.de/10.1594/PANGAEA.897873, and code is available at https://research.ece.ncsu.edu/aros/foram-identification/.

## Declarations of interest

None.

## References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. USENIX Association, Savannah, GA, USA, pp. 265–283.

Athersuch, J., Banner, F.T., Higgins, A.C., Howarth, R.J., Swaby, P.A., 1994. The application of expert systems to the identification and use of microfossils in the petroleum industry. Math. Geol. 26 (4), 483–489.

Bauch, D., Darling, K., Simstich, J., Bauch, H.A., Erlenkeuser, H., Kroon, D., 2003. Palaeoceanographic implications of genetic variation in living North Atlantic *Neogloboquadrina pachyderma*. Nature 424 (6946), 299.

Bé, A.W., 1977. An ecological, zoogeographic and taxonomic review of recent planktonic Foraminifera. In: Ramsay, A.T.S. (Ed.), Oceanic Micropaleontology. Acad. Press, London, pp. 1–100.

Beaufort, L., Dollfus, D., 2004. Automatic recognition of coccoliths by dynamical neural networks. Mar. Micropaleontol. 51, 57–73.

Beszteri, B., Allen, C., Almandoz, G.O., Armand, L., Barcena, M.Á., Cantzler, H., Crosta, X., Esper, O., Jordan, R.W., Kauer, G., Klaas, C., Kloster, M., Leventer, A., Pike, J., Rigual Hernández, A.S., 2018. Quantitative comparison of taxa and taxon concepts in the diatom genus *Fragilariopsis*: a case study on using slide scanning, multiexpert image annotation, and image analysis in taxonomy. J. Phycol. 54, 703–719.

Bollmann, J., Quinn, P.S., Vela, M., Brabec, B., Brechner, S., Cortés, M.Y., Hilbrecht, H., Schmidt, D.N., Schiebel, R., Thierstein, H.R., 2004. Automated particle analysis: calcareous microfossils. In: Image Analysis, Sediments and Paleoenvironments. Springer, Dordrecht, pp. 229–252.

Chollet, François (2015). "Keras" programming package, https://github.com/keras-team/keras.

Darling, K.F., Kucera, M., Kroon, D., Wade, C.M., 2006. A resolution for the coiling direction paradox in *Neogloboquadrina pachyderma*. Paleoceanography and Paleoclimatology 21 (2). https://doi.org/10.1029/2005PA001189.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

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

du Buf, H., Bayer, M.M., 2002. Automatic Diatom Identification. Series in Machine Perception and Artificial Intelligence, 51. World Scientific, Singapore, pp. 316.

Ericson, D.B., 1959. Coiling direction of *Globigerina pachyderma* as a climatic index. Science 130 (3369), 219–220.

Eynaud, F., Cronin, T.M., Smith, S.A., Zaragosi, S., Mavel, J., Mary, Y., Mas, V., Pujol, C., 2009. Morphological variability of the planktonic foraminifer *Neogloboquadrina pachyderma* from ACEX cores: Implications for late Pleistocene circulation in the Arctic Ocean. Micropaleontology 55, 101–116.

Gardner, V.J., Hays, J.D., 1976. Responses of sea-surface temperature and circulation to global climatic change during the past 200,000 years in the eastern equatorial Atlantic Ocean. Memoirs of the Geological Society of America 24 (4), 221–246.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning. Vol. 1 MIT Press, Cambridge.

Gupta, A.K., Anderson, D.M., Overpeck, J.T., 2003. Abrupt changes in the Asian southwest monsoon during the Holocene and their links to the North Atlantic Ocean. Nature 421 (6921), 354.

Hayward, P.J., Ryland, J.S. (Eds.), 2017. Handbook of the Marine Fauna of North-West Europe. Oxford University Press.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hilbrecht, H., 1997. Morphologic gradation and ecology in *N. pachyderma* and *N. dutertrei* (planktic foraminifera) from core top sediments. Mar. Micropaleontol. 31 (1–2), 31–43.

Kipp, N.G., 1976. New transfer function for estimating past sea-surface conditions from sea-bed distribution of planktonic foraminiferal assemblages in the North Atlantic. Geological Society of America, Memoir 145, 3–41.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105. https://doi.org/10.1145/3065386.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Pedraza, A., Bueno, G., Deniz, O., Cristóbal, G., Blanco, S., 2017. Automated Diatom Classification (Part B): a Deep Learning Approach. Appl. Sci. 7 (460), 1–25.

Ranaweera, K., Harrison, A.P., Bains, S., Joseph, D., 2009. Feasibility of computer-aided identification of foraminiferal tests. Mar. Micropaleontol. 72 (1–2), 66–75.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Schiebel, R., Hemleben, C. (Eds.), 2017. Planktic Foraminifers in the Modern Ocean. Springer-Verlag, Berlin, Heidelberg, pp. 358.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, International Conference on Machine Learning. 14.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45 (4), 427–437.

Spero, H.J., Lea, D.W., 2002. The cause of carbon isotope minimum events on glacial terminations. Science 296 (5567), 522–525.

Spero, H.J., Mielke, K.M., Kalve, E.M., Lea, D.W., Pak, D.K., 2003. Multispecies approach to reconstructing eastern equatorial Pacific thermocline hydrography during the past 360 kyr. Paleoceanography 18 (1). https://doi.org/10.1029/2002PA000814.

Swaby, P.A., 1992. VIDES: an expert system for visually identifying microfossils. IEEE Expert 7 (2), 36–42.

Thompson, P.R., 1976. Planktonic foraminiferal dissolution and the progress towards a Pleistocene equatorial Pacific transfer function. The Journal of Foraminiferal Research 6 (3), 208–227.

Vickerman, K., 1992. The diversity and ecological significance of Protozoa. Biodivers. Conserv. 1 (4), 334–341.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Proces. Syst. 27, 3320–3328.

Zhong, B., Ge, Q., Kanakiya, B., Marchitto, Mitra, R., & Lobaton, E. (2017). A comparative study of image classification algorithms for Foraminifera identification. 2017 IEEE Symposium Series on Computational Intelligence, 8 pp., DOI: https://doi.org/10.1109/SSCI.2017.8285164