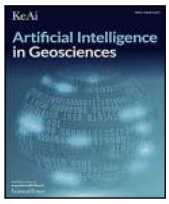
Artificial Intelligence in Geosciences 6 (2025) 100145

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences



Original research articles

Quantifying uncertainty in foraminifera classification: How deep learning methods compare to human experts

Iver Martinsen [a,*], Steffen Aagaard Sørensen [a], Samuel Ortega [a,b], Fred Godtliebsen [a], Miguel Tejedor [c], Eirik Myrvoll-Nilsen [a]

[a] UiT - The Arctic University of Norway, Tromsø, Norway
[b] NOFIMA, Tromsø, Norway
[c] NSE, Tromsø, Norway

ARTICLE INFO

Keywords:
Foraminifera
Uncertainty
Deep learning
Microfossils

ABSTRACT

Foraminifera are shell-bearing microorganisms that are commonly found in marine deposits on the seabed. They are important indicators in many analyses, are used in climate change research, monitoring marine environments, evolutionary studies, and are also frequently used in the oil and gas industry. Although some research has focused on automating the classification of foraminifera images, few have addressed the uncertainty in these classifications. Although foraminifera classification is not a safety-critical task, estimating uncertainty is crucial to avoid misclassifications that could overlook rare and ecologically significant species that are informative indicators of the environment in which they lived. Uncertainty estimation in deep learning has gained significant attention and many methods have been developed. However, evaluating the performance of these methods in practical settings remains a challenge. To create a benchmark for uncertainty estimation in the classification of foraminifera, we administered a multiple choice questionnaire containing classification tasks to four senior geologists. By analyzing their responses, we generated human-derived uncertainty estimates for a test set of 260 images of foraminifera and sediment grains. These uncertainty estimates served as a baseline for comparison when training neural networks in classification. We then trained multiple deep neural networks using a range of uncertainty quantification methods to classify and state the uncertainty about the classifications. The results of the deep learning uncertainty quantification methods were then analyzed and compared with the human benchmark, to see how the methods performed individually and how the methods aligned with humans. Our results show that human-level performance can be achieved with deep learning and that test-time data augmentation and ensembling can help improve both uncertainty estimation and classification performance. Our results also show that human uncertainty estimates are helpful indicators for detecting classification errors and that deep learning-based uncertainty estimates can improve calibration and classification accuracy.

1. Introduction

1.1. Foraminifera

Foraminifera, commonly abbreviated as forams, are microscopic (usually <1 mm) single-celled shell bearing marine organisms that live in the water column (planktic) or at/in the seabed (benthic). Of approximately 50,000 recorded species, around 9000 exist today (Hayward et al., 2024). The foraminiferal shells, formed from various materials, fossilize in sediments where they remain abundant and represent an archive of past climate and oceanic conditions. The faunal composition of foraminifera in sediments, along with measurements of chemical shell properties (such as stable isotopes and trace element ratios) are utilized to reconstruct past oceanic and climate conditions, as well as to evaluate modern marine environments and their ecological state (Emiliani, 1955; Hald et al., 2007; Katz et al., 2010; Kathal et al., 2017). The faunal composition can be analyzed both at the species level and at the group level. The division of foraminiferal fauna into basic groups can be used when examining sediment at lower resolution to plan for more detailed studies within a given sediment core, and the initial grouping of foraminiferal fauna and the ratios between

* Corresponding author.
E-mail addresses: iver.martinsen@uit.no (I. Martinsen), steffen.sorensen@uit.no (S.A. Sørensen), samuel.ortega@nofima.no (S. Ortega), fred.godtliebsen@uit.no (F. Godtliebsen), miguel.tejedor@ehealthresearch.no (M. Tejedor), eirik.myrvoll-nilsen@uit.no (E. Myrvoll-Nilsen).

https://doi.org/10.1016/j.aiig.2025.100145
Received 7 November 2024; Received in revised form 21 June 2025; Accepted 4 July 2025
Available online 16 July 2025

them have been applied as indicators of carbonate dissolution in the Barents and Kara seas, where higher rations of benthic / planktic and in particular agglutinated/calcareous indicate dissolution (Steinsund and Hald, 1994).

1.2. Developments in automatic analysis

After a sediment core is retrieved from the seabed, it undergoes a variety of laboratory procedures before the foraminiferal shells contained there can be studied, counted, and extracted under the microscope from the 100–1000 μm sediment size fraction. For a geoscientist tasked with identifying different foraminiferal groups and/or species, distinguishing features like shell shape, chamber arrangements, ornamentation, aperture shape/size, and shell material are relevant variables. To establish a statistically robust representation of the foraminiferal fauna, 300–500 specimens are identified and extracted per sample. In a typical study, about 100–200 samples are studied, and depending on the complexity (i.e., concentration of foraminifera, number of species, preservation state of species, etc.) of the sample, may take ca. 2–8 working hours per sample.

Due to the time-consuming nature of manual foraminifera classification, efforts have been made to automate the process using deep learning (Marchant et al., 2020; Mitra et al., 2019; Johansen et al., 2021; Johansen and Sørensen, 2020; Zhong et al., 2017). In particular, Johansen and Sørensen (2020) and Johansen et al. (2021) presented promising results for the detection and classification of foraminifera and sediments using images that contained four object classes; benthic calcareous forams, benthic agglutinated forams, planktic forams and sediment grains. There has also been research on chamber segmentation (Ge et al., 2017), the generation of synthetic foraminifera images (Ferreira-Chacua and Koeshidayatullah, 2023), and self-supervised learning utilizing large volumes of unlabeled microfossil images (Martinsen et al., 2024).

1.3. Uncertainty related to foraminifera analysis

Although much work has been done in automatic classification of foraminifera, little work has involved the uncertainty of deep learning predictions and how these predictions relate to human performance. The uncertainty study by Fenton et al. (2018) analyzed the effect of different factors on the identification of manual planktic foraminifera using a random effects model and showed evidence that the confidence, experience, and specimen size reported are important predictors of classification accuracy. However, this study involved uncertainty related to the analysis of specimens through the microscope, not automatic analysis using deep learning. Al-Sabouni et al. (2018) studied the effect of different taxonomic schools on classification inconsistency between different raters, and Mitra et al. (2019) presented a comparison of human and artificial intelligence (AI) performance for foraminifera. This work analyzed pre-trained networks in a setting with an elaborate preprocessing pipeline, but did not involve deep learning uncertainty estimates. Uncertainty estimation of deep learning predictions is potentially an important task because identifying images where AI is uncertain can help uncover rare but important indicator species that may otherwise be overlooked during automatic analysis. For some applications, this is crucial, as unknowingly misclassifying indicator species will lead to wrongful conclusions. In addition, uncertainty estimation increases overall trustworthiness and can improve the reliability of automated systems in environmental and ecological research.

1.4. Uncertainty estimation in deep learning

The perspective of reliable AI models makes uncertainty quantification (UQ) an important task and a fundamental research topic in deep learning. In this context, uncertainty can arise from two main sources: *epistemic uncertainty*, which refers to uncertainty in the model's parameters due to limited knowledge or data, and *aleatory uncertainty*, which is inherent to the data and cannot be reduced by gathering more data (see, e.g., Gawlikowski et al., 2022 or Kiureghian and Ditlevsen, 2009 for a discussion on the topic). An ideal model would accurately capture the aleatory uncertainty while minimizing the epistemic uncertainty. However, due to the complexity of AI models, quantifying uncertainty in a practical and accurate way is often difficult to achieve. In addition to uncertainty about the model parameter values, specifications about model design and training hyperparameters add to the uncertainty. The assumptions of the model, the number of parameters and the learning rate are examples of this. The lack of good uncertainty estimates presents a major challenge when applying AI to tasks that require not only good performance at the population level (i.e. average accuracy). For critical applications, for example diagnosis, we need reliable confidence statements at the individual sample level to obtain an accurate risk assessment (Amodei et al., 2016). Bridging classical statistics with deep learning to address these challenges has led to the development of a range of methods that aim to quantify the uncertainty of deep learning predictions (see, e.g., Gawlikowski et al., 2022 for a review of UQ methods in deep learning). Ensembling (Lakshminarayanan et al., 2017), Monte Carlo dropout (Gal and Ghahramani, 2016), augmentations in testing (Ayhan and Berens, 2018), stochastic weight averaging (Maddox et al., 2019), and stochastic gradient Langevin dynamics (Welling and Teh, 2011) are examples. Investigating the quality of the uncertainty estimates on an instance level is difficult, and we often lack good baselines to assess how reasonable the individual uncertainty estimates are. Accuracy, negative log-likelihood (NLL), Brier score, or confidence calibration error (Guo et al., 2017) are summary statistics that are often reported. Although these metrics are useful for evaluating the overall performance of a model at the population level, they are insufficient to capture the quality of individual uncertainty estimates in a practical setting. To help in this analysis, we propose to evaluate the uncertainty methods from a new perspective.

1.5. Contribution

In this study, we test and compare human participants and deep learning methods. We test the participants from the point of view of the deep learning model and, as such, obtain classification scores and uncertainty estimates that form a realistic benchmark against which other methods can be compared. By testing human participants in classification, we can obtain uncertainty estimates that are representative of the classification difficulty from a human perspective. This allows us to (1) assess the performance of the UQ methods in terms of accuracy and calibration against a realistic benchmark, and (2) assess how deep learning aligns with humans in terms of uncertainty estimation when given the same classification task. This paper adds to the present knowledge in the following ways:

- **Expert uncertainty estimates:** We test geoscientists in a deep-learning-like foraminifera classification task, and produce classification scores and weighted uncertainty estimates that serve as a benchmark to which other methods can compare.
- **Foraminifera classification:** We extend the work done by Johansen and Sørensen (2020), Johansen et al. (2021), and develop an easy-to-use system for the classification of foraminifera at the group level. Our results are on par with or better than the results from geoscientists specialized in foraminifera.
- **Comparison of UQ methods:** We investigate several UQ methods and compare them with the expert benchmark, gaining important insight into how well these methods perform and the quality of their estimates. We also address challenges related to the evaluation and comparison of uncertainty estimates from different sources.

Our research leads to important findings:

- We gain more knowledge about how experts can perform in this field when challenged with the same task as a deep learning model, how to quantify uncertainty from human responses, and the quality of acquired uncertainty estimates.
- We get an understanding about the performance of deep learning methods for this field and how the quality of uncertainty estimates compares to each other, and to the expert uncertainty benchmark.
- Finally, we see how experts and deep learning methods compare both in terms of predictive performance and uncertainty estimation.

2. Material and methods

2.1. Data

The data used in this work consist of images of foraminifera and sediment. Each image shows an object that belongs to one of four possible classes: benthic calcareous, planktic, benthic agglutinated, and sediment grains. The data were produced as follows.

All photographed specimens were handpicked from the 100–1000 μm sediment size fraction under the microscope by a foraminiferal expert. For each class, the specimens were handpicked from a sediment sample, and each specimen was individually manipulated using a picking needle allowing a very high degree of certainty when assigning specimens to the above-mentioned classes and thus the ground-truth has an estimated error rate close to or at zero. In the end, the specimens were dispersed on a new tray that only contained specimens of one class. The specimens were randomly dispersed on the tray, and we did not control the orientation prior the photographing. The photos were produced by capturing images of type-specific samples of the above-mentioned classes using a 51 megapixel Canon EOS 5DS R mounted on a Leica M420 stereo microscope. A Leica cls 150x twin gooseneck combination light was used for lighting at fixed lumen.

All photographed materials (foraminifera and sediment) were extracted from sediment cores retrieved at water depth ranging from ~100–700 m from both fjord and shelf settings in the Arctic Barents Sea region. These areas are influenced by the Atlantic, Arctic, polar, and coastal waters, thus ensuring a broad representation of different ecological environments.

Fig. 1 shows examples of the data. Since individual forams were extracted from images that contained multiple specimens of the same class, some images contained additional small-scale forams of the same class (see Fig. 1b). Deep neural networks generally depend on a fixed resolution for all inputs; therefore, all images were cropped and resized to a resolution of $(224, 224)$. The red scale bar in the figure is included to show the scale resolution of the objects and was *not* present in the images used to train and test deep neural networks. This resolution is commonly used in the deep learning literature (Deng et al., 2009; Tan and Le, 2020; He et al., 2015), and was also sufficient to determine the classes in this case. In this study, we also used two disjoint data sets. The training dataset consisted of individual images of 258 agglutinated benthic sediments, 380 calcareous benthic sediments, 332 planktic, and 341 sediments (1311 in total) and was used to tune the hyperparameters and train the neural networks. The test dataset consisted of 65 images from each class (260 in total) originating from the same sediment samples as the training data and was constructed to test both geoscientists and AI models. The test data set was disjoint from the training set and was only used to test the AI after training.

2.2. Testing the participants

To create a benchmark for uncertainty, we challenged four senior geoscientists, highly skilled in foraminifera analysis, with a multiple choice type questionnaire (see Appendix A for a screenshot of the software used). The task of the participants was to classify the images in the test dataset ($n = 260$) assigning each image to one of the four evenly represented classes. In addition to the class response, we also asked the participants to state their confidence level. The confidence statements were restricted to $n_p = 5$ possible values; 0, 25, 50, 75, and 100. The time spent on each test image was also recorded for each participant. Previous studies (Mitra et al., 2019; Fenton et al., 2018) show that human performance is dependent on the level of experience; however, in this work, we consider all participants to be highly experienced. Different taxonomic schools are known to lead to inconsistencies among experts (Al-Sabouni et al., 2018). Because of the well-defined differences between the classes at the group-level, we did not consider this as a potential challenge in this study.

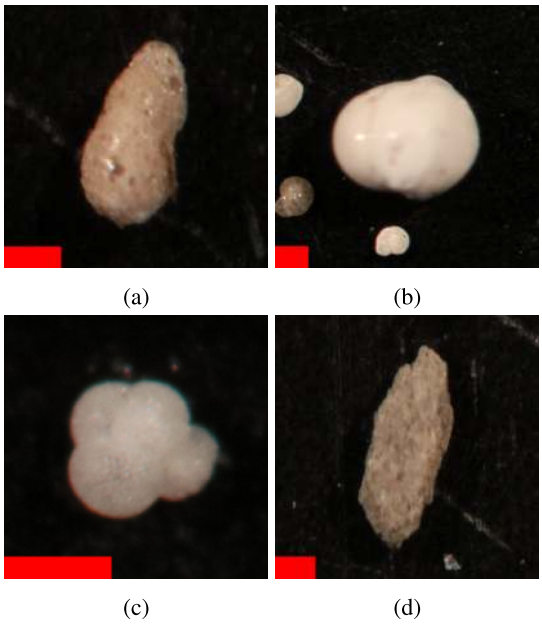


Fig. 1. Foraminifera data The data used in this study consist of images from three groups of foraminifera, in addition to images of sediment grains. The four classes are (a) benthic agglutinated foraminifera, (b) benthic calcareous foraminifera, (c) planktic foraminifera, and (d) sediment grain. The specimens varied in size, and the red line in the bottom left corners represents a length of 240 μm. Fig. (a) and (c) is representative for the size of most of the training and test specimens.

The test specimens were fixed as in Al-Sabouni et al. (2018). This method of testing differs from the way forams are usually analyzed, where each specimen is carefully analyzed using a microscope from different angles to determine the correct species. Our purpose, however, was to obtain a baseline uncertainty estimate for each image in a setting that matched the setting of the deep learning model. Typically in automatic image classification, the model is forced to do prediction of a 2D rendition of the foraminifera, and as such will occasionally encounter specimens that are difficult to classify due to an unfortunate orientation of the specimen. By testing experts on the same examples, we get an estimate of the uncertainty related to classifying the 2D foraminifera image, which indeed is the task of the model.

2.3. Aggregating responses and confidence statements from the participants

Our aim was to aggregate the responses into a single pair (prediction, confidence) for each test image. This is similar to what is commonly done, for example, when looking at an ensemble of deep neural networks. The purpose of aggregating the responses was to take both confidence and inter-rater variability into account to create high-quality predictions and uncertainty estimates that would serve as a prediction and uncertainty benchmark. To construct aggregated responses, we used the following procedure:

- Assume a test image i, and let (y_i^j, p_i^j) represent the class prediction and the confidence statement for participant j.
- Assign a vote to each class c using the formula

$$v_i^c = \sum_j p_i^j \cdot I\{y_i^j = c\} \tag{1}$$

where $I\{y_i^j = c\}$ is an indicator function equal to 1 if $y_i^j = c$, 0 otherwise.
- Let the aggregated prediction be $y_i = \underset{c}{\operatorname{argmax}} \{v_i^c\}_c$

In the presence of ties (two or more classes with equal votes), we used the following rules:

- If two or more classes were given the same score, the majority vote decided the class.
- If two or more classes were still equal, a prediction was drawn at random.[1]

Aggregating multiple confidence statements to obtain good uncertainty estimates is not straightforward. Interpreting the confidence statements as probabilities is in general not justified, and problems will arise when, for example, treating confidence statements of 0 in a four-way classification with a minimum probability of 0.25. To produce aggregated confidence scores (uncertainty estimates) for the predictions, we applied the formula in Eq. (2) where n_0 is the number of nonzero elements of $\{v_i^c\}_{c=1}^4$, and k is the number of participants.

$$p_i = \left(v_i^{y_i} - \sum_{c \neq y_i} v_i^c / (n_0 - 1) \right) / k \tag{2}$$

This formula gives an aggregated confidence statement in the range of 0 to 1 for an image i, where the first term in Eq. (2) is the total confidence for the participants who agreed on the predicted class, and the second term is the total confidence for the participants who proposed a different class. The second term is scaled by a factor $1/(n_0 - 1)$, which reduces the impact of opposing raters in cases where opposing raters disagree among themselves. This way of calculating the uncertainty is slightly more involved than taking, e.g., the mean of the confidence statements or the percentage agreement. The benefit of this approach, however, is that we were able to take both confidence statements and the inter-rater agreement into account. Moreover, this approach provided a better resolution of the uncertainty estimates compared with, e.g., percentage agreement (with at most $k - 1 = 3$ different values) or average confidence (with at most $k \cdot (n_p - 1) + 1 = 17$ possible values).

2.4. Quantifying the uncertainty from deep neural network predictions

We investigated sampling-based deep learning UQ methods in this study, all of which produced a set of M confidence statements $P = \{\hat{y}^1, \ldots, \hat{y}^M\}$ per image. To summarize the ensemble of predictions (P), we computed the mean confidence of all M evaluators

$$\hat{\mathbf{y}}_i = \frac{1}{M} \sum_m \hat{\mathbf{y}}_i^m \tag{3}$$

where $\mathbf{y}_i^m$ is prediction m (a probability vector) for input i. Using this formula, the ensemble of predictions were treated as a mixture of models using Bayesian model averaging with equal weights. The mean confidence is comparable between all uncertainty methods and is also sufficient for computing calibration plots and reliability diagrams (see Section 2.6). A potential weakness in using the mean confidence is that the calculation discards the inter-rater variability. Our experience, however, was that the inter-rater variance ($\mathrm{Var}(P)$) was highly correlated with the mean confidence (Eq. (3)) and added little practical value to the uncertainty estimation (see Appendix E in the appendix for a discussion about this). As $\hat{\mathbf{y}}_i$ is a vector, we defined the uncertainty in terms of the maximum confidence of $\hat{\mathbf{y}}_i$, which also eases the comparison to the human uncertainty estimates. We note however, that other calculations such as the entropy of $\hat{\mathbf{y}}_i$ can also be used to estimate the uncertainty from $\hat{\mathbf{y}}_i$.

We remark that computing uncertainty in a multi-way classification task is not as intuitive as in the case where one has a single numerical output, for instance, in regression. A possible approach is to summarize the uncertainty by the inter-rater variability σ_{Epi} (epistemic uncertainty), or by possibly combining the confidence and variance. A challenge in this, however, is that the confidence statement and the variability between the raters are not generally independent and easily decoupled in classification. There are several possible approaches to estimate the uncertainty in classification tasks that we discuss in the appendix (see Appendix B).

2.5. Uncertainty quantification methods

In this section, we present the uncertainty methods that we applied in this paper. We investigate a handful of methods that are well established and widely used, and we direct the reader to Gawlikowski et al. (2022) for a review of many of the vast number of UQ methods that exist in deep learning.

Ensembling. Uncertainty estimation through ensembling is a well established and conceptually simple method, that has also been adapted for computer vision (Lakshminarayanan et al., 2017). In short, ensembling means training a range of models in parallel with multiple possible states, before aggregating their outputs to obtain a single-class prediction. The output aggregation is usually done with equal weighting for each ensemble member, resembling Bayesian model averaging (BMA) with uniform weighting. The method is appealing for deep learning because it addresses possible challenges related to initial conditions and training trajectories with no extra implementation, and without relying on assumptions about parameter distributions.

Monte Carlo dropout. Monte Carlo Dropout (MC Dropout) is an extension of dropout regularization (Srivastava et al., 2014), which is a popular regularization technique that is commonly used in deep learning. Dropout regularization was originally derived as an approximation of ensembles (Srivastava et al., 2014), with the aim of avoiding overfitting. Dropout is often applied during training, but can also be used for testing to model uncertainty (Gal and Ghahramani, 2016). When used at test time, MC Dropout resembles an ensemble of a combinatorial large number of models. The method is popular in computer vision, as it can help model the uncertainty with little computing overhead in cases where dropout is already implemented for training.

Stochastic gradient langevin dynamics. Stochastic Gradient Langevin Dynamics (SGLD) is a method for parameter sampling that approximates Markov Chain Monte Carlo (MCMC) sampling. MCMC sampling, with Hamiltonian Monte Carlo (Neal, 2012) as the gold standard, is generally not feasible for deep neural networks due to huge computational demands (see, e.g., Izmailov et al.). SGLD was therefore proposed as a seamless transition between Bayesian sampling and optimization (Welling and Teh, 2011). SGLD is analogous to stochastic gradient descent (SGD) with added gradient noise, and is thus computationally fast (see Welling and Teh, 2011). The method has also been adapted for use with the popular RMSprop optimizer (Li et al., 2016).

Stochastic weight averaging — Gaussian. Stochastic weight averaging-Gaussian (SWAG) (Maddox et al., 2019) is an uncertainty method closely related to the Bayesian analysis paradigm and Laplace approximations. SWAG relies on a fundamental assumption that the posterior (after convergence) has a multivariate normal distribution in the close vicinity of the posterior mode. By recording subsequent weight updates after a re-initialization of the training at the mode, one can easily compute the mean and variance estimates required to specify the

[1] The random draw is simply a necessity for handling possible ties.

multivariate normal distribution (see Maddox et al., 2019 for details). The weights (and predictions) are, in turn, sampled after the weight distribution has been specified, and the sampling phase of the method is efficient in terms of computational resources.

Test time data augmentation. Test-time data augmentation (TTA) (Ayhan and Berens, 2018) is a natural extension of the data augmentation paradigm, which is commonly used in computer vision to improve model training. TTA involves applying augmentations during testing, and is simple to implement and can help address the robustness of the model in terms of invariance to input perturbations.

2.6. Metrics for comparing and evaluating predictions and uncertainty estimates

Although accuracy is commonly used to evaluate model performance, additional metrics are needed for a more in-depth analysis of the quality of uncertainty estimates. We therefore assess the precision of the confidence statements, the agreement among raters (or models), and the correlation between uncertainty estimates. In this section, we provide a brief overview of some of the metrics used in our analysis. As in Ovadia et al. (2019) and Lakshminarayanan et al. (2017) we also report the Brier score and the negative log-likelihood (NLL) as measures of the precision and quality of the confidence statements.

Calibration error. The perhaps most established metric for evaluating model confidence statements is the so-called calibration error. The calibration error is a measure that aims to assess how well the confidence of the model aligns with the actual probability of error. Following the definition in Guo et al. (2017), the expected calibration error (ECE) is defined by

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{4}$$

where

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} I(\hat{y}_i = y_i), \tag{5}$$

and

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{6}$$

gives the observed and predicted precision for bin m, where M is the number of bins and $|B_m|$ is the number of test examples that have a confidence score that falls under bin m.

Our experience is that the calibration error is highly dependent on the binning strategy used (see, e.g., Nixon et al., 2020). We therefore computed the error by using two different strategies; an equal spacing and an equal sample strategy. In this work, we define the calibration error as the error computed using equal spacing for each bin, as in Guo et al. (2017). We define reliability error as the error computed using an equal sample size per bin, as in Maddox et al. (2019). As suggested in Guo et al. (2017), we employ the so-called one-vs-all strategy to extend the calculations to the multiclass setting. Using this strategy, the calibration error is computed once for each class, and the confidence scores for the remaining classes are summed to resemble a binary classification task. We computed the ECE for each class (four in total) before summing the ECE over all classes.

Cohen's kappa score. Cohen's kappa score (κ) is a coefficient for measuring the agreement (inter-rater reliability) between two raters (e.g., two experts or two neural networks) who are assigning n items into C different classes or categories. The coefficient has previously been used in foraminifera classification (see, e.g. Fenton et al., 2018), and is defined by

$$\kappa \equiv 1 - \frac{1 - p_0}{1 - p_e}, \tag{7}$$

where p_e is the probability of agreement by chance, and p_0 is the observed agreement between the raters. The kappa score has an upper bound of 1 (perfect agreement). See McHugh (2012) for a discussion.

Spearman's rank correlation. The *Spearman's* ρ is a correlation coefficient that determines the rank correlation between two sets of observations $\{x_i\}$ and $\{y_i\}$. Unlike computing the correlation on the variables directly, the rank correlation is computed using the variables $\{R_x(i)\}$, where $R_x(i)$ and $R_y(i)$ are the ranks of $\{x_i\}$ and $\{y_i\}$. The Spearman's rank correlation takes values in the range $[-1, 1]$ and is given by

$$\rho = \frac{\mathrm{Cov}(R_x, R_y)}{\sigma_{R_x} \sigma_{R_y}}. \tag{8}$$

The rank correlation provides an approach for comparing uncertainty estimates between two sources with different scales, for instance between deep learning uncertainty estimates and uncertainty estimates provided by domain experts.

Related metrics. In addition to the metrics mentioned above, there exist other ways of evaluating uncertainty estimates. Hendrycks and Gimpel (2018) proposed to compute the area under the ROC curve and the area under the precision–recall curve for an auxiliary task of detecting misclassified samples. Both are summary statistics that are closely related to our analysis of detecting errors (see Section 4). Another commonly used task related to uncertainty estimation is in detecting out-of-distribution (OOD) data (see, e.g., Ovadia et al., 2019; Hendrycks and Gimpel, 2018; Lakshminarayanan et al., 2017). This application is beyond the scope of this paper, and we leave the evaluation of OOD detection for future work.

3. Calculation

3.1. Model and training configuration

As our data set is relatively small, we investigated small convolutional neural networks (CNNs) in our model training. EfficientNet (Tan and Le, 2020) is a series of CNNs of different sizes (named B0 to B7) that represent different scales of a base architecture in terms of depth, width, and resolution (see Tan and Le for details). Our models were trained with the base architecture (EfficientNet-B0), which has roughly 4.1M parameters when using four classes. This proved to give sufficiently good results with decent hyperparameters, and preliminary tests with B1, B2, and B3 did not reveal a substantial gain in performance. We also used a data augmentation strategy to make the model robust to small changes in input, as is commonly done when training deep neural networks for image classification. Similarly to Johansen and Sørensen (2020), our augmentation strategy consisted of the following operations:

- Random flip (horizontal and vertical)
- Random rotation with a factor of $0.1 \cdot S$ ($\pm 0.2\pi$ radians)
- Random translation with a height/width factor of $0.1 \cdot S$
- Random inwards zoom with a factor of $0.1 \cdot S$
- Random contrast with a factor of $0.2 \cdot S$

We scaled the augmentation factors by a strength parameter S, where our default setting was $S = 2$ during training. We trained all models for 300 epochs using the RMSprop optimizer with a decaying learning rate. We also used L2 regularization (weight decay) of 0.001, which is equivalent to an isotropic Gaussian prior to the weights. More details are provided in Appendix C in the appendix. The source code is available at:
https://github.com/IverMartinsen/ict-plus-uncertainty.

3.2. Configuration for uncertainty methods

We used the following configurations in our implementation of the UQ methods described in Section 2.5.

Ensembles. We trained 10 models using 10 different seeds. This ensured different initial values for the model parameters and different training trajectories.

Monte Carlo dropout. We used the original implementation of EfficientNet-B0 (Tan and Le, 2020), where dropout is applied in all blocks except the first one. There are 9 dropout layers in total in EfficientNet-B0, and the default dropout rate is 0.2. We used this configuration for both training and testing. For testing, we used 30 dropout variations per batch, resulting in 30 predictions per image. We used a batch size of 32 during testing.

Stochastic gradient Langevin dynamics. We trained the SGLD models for 270 epochs, before adding gradient noise (the Langevin dynamics phase) in the last 30 epochs. The weights at the end of the last 30 epochs were stored and used for prediction (that is, 30 predictions per image). We trained our SGLD with the default learning rate from hyperparameter tuning ($5 \cdot 10^{-6}$), in addition to two slightly higher learning rates ($5 \cdot 10^{-5}$ and $1 \cdot 10^{-4}$) to investigate the effect of a larger sampling variance.

Stochastic weight averaging—Gaussian. We implemented the SWAG-Diagonal method as described in Maddox et al. (2019), using an isotropic Gaussian assumption for the posterior. This is a strong but necessary assumption, as it is not feasible to compute the complete multivariate normal covariance matrix with a model that has 4M+ parameters. We trained the models for 30 epochs after re-initialization, using the parameters at the end of each epoch when recording the mean and variance statistics. After training, we used the recorded statistics to sample 30 different weight configurations, resulting in 30 predictions per image.

We note that the estimated variance in SWAG is closely related to the learning rate. Preliminary testing on the training set showed indications that SWAG benefitted from a higher learning rate than that used in training the other models. We did not adjust the SWAG learning rate in our experiments, but instead trained SWAG with four different learning rates ($5 \cdot 10^{-4}$, $1 \cdot 10^{-4}$, $5 \cdot 10^{-5}$, and $5 \cdot 10^{-6}$) and reported the results for each.

Test time data augmentation. We applied the same data augmentation that we used for training in testing. We investigated two different augmentation strengths ($S = 1$ and $S = 2$). We augmented and tested each image 30 times, and used the median probability to aggregate the outputs as in Ayhan and Berens (2018). The median was computed for each class by

$$\hat{\mathbf{y}}_i = \left(\underset{m}{\text{median}}(y_{i1}^m), \ldots, \underset{m}{\text{median}}(y_{i4}^m) \right) / k_i, \tag{9}$$

where k_i is a normalizing factor to ensure that $\mathbf{y}_i$ summed to one. This did not affect the accuracy but was done to correct for the large variance in confidence that would negatively affect the calibration when using the mean.

4. Results and discussion

4.1. Results from participants

As shown in Table 1, the participants performed reasonably well in the test. The accuracy for the participants varied in the range of 89–97%, with a 93% accuracy on average. Although the accuracy varied between participants, our prior assumption was that experts were approximately equally skilled and therefore we treated their responses with equal weighting. The weighted accuracy (see Eq. (1)) was 96% with 249 out of 260 correctly classified images. The experts had an inter-rater reliability (κ) of 0.88, indicating a moderately high agreement among the raters. The participants did not always agree in their confidence statements, reflected in a rank correlation (ρ) of 0.33. This indicates a lack of consistency between experts when stating their confidence, which is also supported by the scatter plots of the confidence statements (see Fig. 2 for an example).

The uncertainty distributions of the participants are shown in Fig. 3. The confidence distribution (Fig. 3a) reveals that the participants on average were less confident when they were wrong (orange bars) compared to when they were correct (blue bars). Analysis showed that when the experts were all close to certain, they were also correct 177 out of 178 times. Although experts were generally likely to be correct when stating high confidence, the significant overlap of the two distributions (correct and incorrect) indicates that confidence statements alone are not sufficient to determine which predictions are likely to be wrong. A more striking plot is perhaps the percent agreement distribution shown in Fig. 3b which shows that when the experts agreed, they were all correct (217 out of 217). To combine confidence with agreement, we use the weighting rule as stated in Eq. (2), and Fig. 3c shows the weighted confidence distribution for correct (blue) and incorrect (orange) responses. The weighted confidence provides finer resolution uncertainty estimates that are also more efficient in separating correct and incorrect predictions.

Table 1
Results from participants. The participants obtained an accuracy of 93% on average, with a weighted accuracy of 96%.

Person 1	Person 2	Person 3	Person 4
0.92	0.97	0.89	0.93
Av. Acc.	**Weight. Acc.**	**κ**	**ρ**
0.93	0.96	0.88	0.33

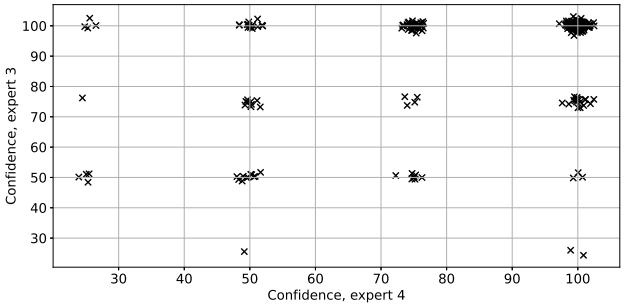
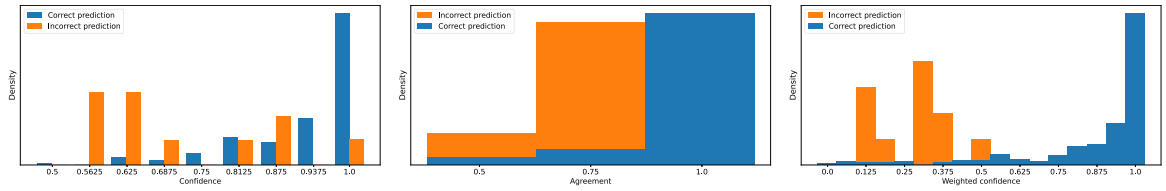


Fig. 2. Confidence scatter plot The confidence statements from the participants' did not generally correlate as shown here for participant 3 and 4. A similar lack of correlation was typical for all the scatter plots.

For comparison purposes, we used the weighted confidence scores to subdivide the test set into *easy* and *hard* images. The images with a weighted confidence of 1.0 were categorized as easy, justified by the fact that all participants gave the correct responses with confidence 100%. Selecting which images are objectively hard to classify is more difficult. In our analysis, we chose a threshold of 0.5625, where test images with uncertainty below the threshold were labeled as hard. Our threshold value was chosen to be the smallest value that included all misclassifications and, as such, probably excluded images that may also be categorized as objectively difficult to classify. However, our aim was to define two groups (easy and hard) that were used in a comparison analysis between humans and AI. Thresholding at this value resulted in a group of 32 images that included all misclassifications, in addition to a handful of correct classifications that the experts were uncertain about. All test images below this threshold are shown in Fig. 4. The hard group consisted of 14 benthic agglutinated forams, 14 sediment grains, and 4 planktic forams. The class distribution in the hard group aligned with our prior expectation that sediment and agglutinated forams could be difficult to distinguish. This is because agglutinated forams largely form their shell from surrounding sediment grains, and if distinguishing features like shell shape, chamber arrangements, aperture shape etc. are not pronounced in the photo they may appear sediment-like.

4.2. Results from deep learning

The deep learning models were tested on the same images as the experts, with results summarized in Table 2. We first trained a



(a) (b) (c)

Fig. 3. Uncertainty distributions for the participants The figures show the uncertainty distributions for the responses where the participants were correct (blue) and incorrect (orange). The histograms are displayed on a density scale. (a) Distribution for the mean of the experts' confidence statements. (b) Percent agreement of the experts predictions. (c) Weighted confidence distribution. Plot (a] shows the two histograms interleaved, while the distributions for the incorrect predictions are plotted in the background in plot (b) and (c) for clarity. No orange bars are hidden behind the blue bars.

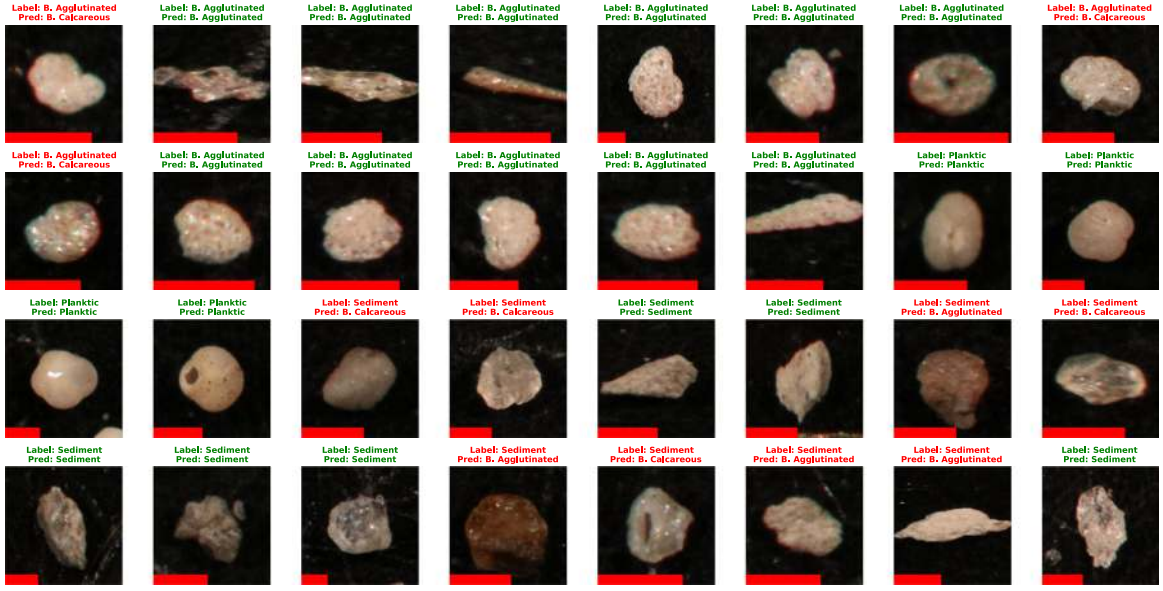


Fig. 4. Hard test images The figure shows the 32 images with the highest expert uncertainty. All misclassifications (red colored text) were assigned to the hard group. The red line in the bottom left corners is a scale representing a length of 240 μm.

single EfficientNetB0 as a baseline (second to top row), for which MC Dropout, SWAG, and TTA were implemented as add-ons for uncertainty estimation. The results for the first baseline, MC Dropout, SWAG and TTA are therefore directly comparable to each other. The second baseline (ensemble baseline) shows the average result when each ensemble member was treated as a single model. This was computed as the mean across all models in the ensemble, and is therefore directly comparable to the ensemble results. The SGLD models were trained independently due to the implementation of noisy gradients.

Accuracy. The first column of Table 2 shows the accuracy after aggregation, along with the plain average accuracy for the individual predictions in parentheses. The accuracy range shows the variability of each set of predictions (i.e., for each expert, model, or sample). Except for TTA (s = 2), deep learning methods achieved an accuracy in the range of 0.91–0.94, which was less than the aggregated expert scores, but on par with what participants achieved on average. For SWAG, TTA, ensemble and SGLD we observed a positive ensembling effect in that aggregation across multiple predictions resulted in increased accuracy. This effect was greatest for ensembles and TTA (s = 1) where the increase was 3 and 2 percentage points, respectively. TTA resulted in a substantial decrease in accuracy at the individual prediction level. However, this was corrected for when aggregating the predictions, and when using a moderate augmentation strength (s = 1) we were able to achieve a substantial increase in accuracy compared to the baseline. Although beyond the scope of this work, we believe that additional improvements can be made by ablating and carefully configuring the augmentations.

Kappa. The purpose of the kappa score in this analysis was to measure to what extent the uncertainty methods caused changes to the model predictions. Table 2 shows that the scores varied between methods, with very low values for MC Dropout, SWAG (5e6, 5e5 and 1e4) and SGLD (5e6 and 5e5), and more moderate scores for SWAG (5e4), Ensembles and SGLD (1e4). TTA severely influenced the predictions, resulting in very low kappa scores. It is difficult to determine what constitutes a good kappa score, but we note that the ensembling obtained a score on par with the kappa between the participants. This was also the case for SWAG and SGLD when increasing the learning rate to 5e4 and 1e4, respectively, but the increased learning rate came at the expense of reduced accuracy.

Easy and detected mistakes. Performing a similar exercise as with the test results from the participants, we used the uncertainty estimates given by the different deep learning methods to differentiate between easy and hard images. For comparison purposes, the *easy* group was predetermined to consist of the 133 most certain images, and the *hard* group to consist of the 32 most uncertain. This was done to create groups of the same size as the experts, and the exercise was repeated for each method. The uncertainty was defined in terms of the mean confidence as described in Section 2.4, except for TTA where the median confidence was used. The mistakes in each group were used to compute the *easy mistakes* and *detected mistakes* statistics shown in Table 2, where we define easy mistakes to be the number of misclassifications among all images with low uncertainty (the easy group), and detected mistakes to be the number of misclassified images that were assigned to the hard group. We note that it is difficult to divide the test data into

Table 2
Deep learning test results. The deep learning methods performed slightly worse than what the humans did (after aggregation), but on par with the average results from the participants. Ensembles and TTA (s = 1) gave the largest improvement in accuracy compared to the baseline, and the methods also resulted in a lower kappa score. The *easy mistakes* column shows the number of misclassifications in the easy (low uncertainty) group, while the *Detected mistakes* column shows the share of misclassifications that were categorized as hard images.

Experts	Acc. (average)	Acc. range	NLL	Brier Score	κ	Easy mistakes	Detected mistakes	
	0.96 (0.93)	0.89–0.97	–	–	0.88	0	11/11	(100%)
Baseline	0.92	–	0.34	0.14	–	2	10/21	(48%)
Baseline + Dropout	0.91 (0.91)	0.89–0.92	0.31	0.14	0.96	2	**14/23**	**(61%)**
Baseline + SWAG (5e6)	0.92 (0.92)	0.92–0.92	0.35	0.14	1.00	2	10/21	(48%)
Baseline + SWAG (5e5)	0.93 (0.92)	0.91–0.93	0.35	0.14	0.99	3	8/19	(42%)
Baseline + SWAG (1e4)	0.92 (0.92)	0.90–0.93	0.38	0.14	0.98	3	9/21	(43%)
Baseline + SWAG (5e4)	0.87 (0.84)	0.72–0.89	0.75	0.23	0.87	6	11/34	(32%)
Baseline + TTA (s = 1)	**0.94 (0.83)**	0.78–0.86	**0.28**	**0.11**	0.72	**0**	7/16	(44%)
Baseline + TTA (s = 2)	0.88 (0.68)	0.63–0.73	0.42	0.19	0.46	1	15/31	(48%)
Baseline	0.90	–	0.44	0.16	–	2.8	**12.5/25.8**	**(48%)**
Ensemble	**0.93**	0.87–0.92	**0.26**	**0.12**	0.89	**2**	8/18	(44%)
SGLD (5e6)	0.92 (0.91)	0.89–0.92	0.34	0.15	0.97	1	7/21	(33%)
SGLD (5e5)	0.92 (0.91)	0.90–0.93	0.30	0.13	0.95	2	11/22	(50%)
SGLD (1e4)	0.88 (0.88)	0.78–0.92	0.34	0.15	0.88	4	18/30	(60%)

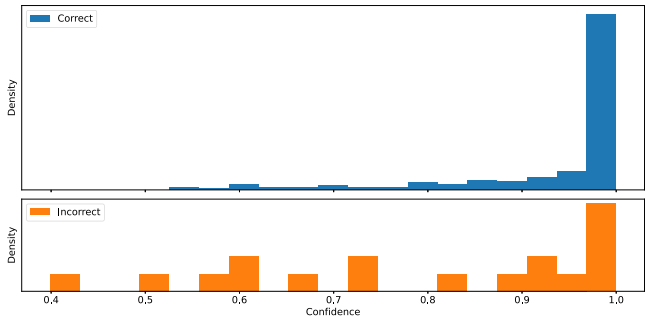


Fig. 5. Ensemble confidence distribution The confidence distribution for incorrect predictions (orange) is very left-skewed, but still has a substantial overlap with the more uniform distribution for correct predictions (blue).

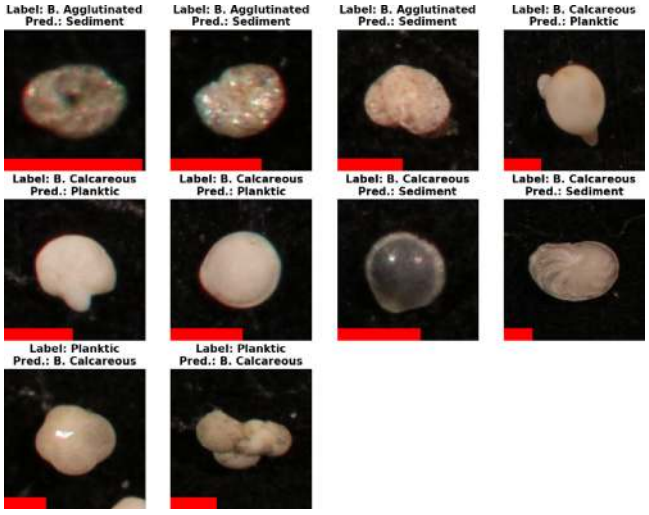


Fig. 6. Deep learning misclassifications The figure shows the 10 forams that both ensembles and TTA (s = 1) misclassified.

easy and hard samples in a way that is fair and comparable across all methods. The purpose of this is analysis is nevertheless to highlight the practical aspect of uncertainty estimation, and to gain insight into the usefulness of the uncertainty estimates in a practical setting.

Among the UQ methods, both TTA and ensembles resulted in a reduction in the number of easy mistakes compared to the baseline. TTA (s = 1) performed the best, with no errors among the easy group. When we look at the number of detected mistakes, the deep learning UQ methods were inefficient at detecting their own mistakes compared to the human uncertainty estimates, and only MC Dropout resulted in an increase among the methods. The challenge of detecting errors in this application is illustrated in Fig. 5, which shows the confidence distribution for ensembles for correct and incorrect predictions. Although the correct and incorrect distributions differ, the degree of overlap is substantial compared to the expert weighted confidence distributions (Fig. 3c). This tendency was also typical for the other methods, which makes it difficult to identify misclassifications from the confidence scores.

A qualitative analysis of wrong predictions. Fig. 6 shows the forams that both the ensembles and TTA (s = 1) misclassified (10 in total). Among these, 3 were also misclassified by the expert (Number 1, 2, and 9 from the top left). We see that forams 1 and 2 are blurry and, therefore, the chamber structures in the forams were possibly difficult to detect for both the participants and the models. Foram 9 has an outline characteristic of a planktic foram, however, the shine is also typical for a benthic calcareous foram which may explain the mistakes. Foram 3 is a very natural mistake, as it is very difficult to see any foram shell characteristic from the angle the specimen was captured in the photo. Foram 4, 6 and 7 were correctly classified by the participants with very high certainty. These are, however, rare benthic species, and a possible explanation is the lack of specie examples in the training data, leading to the wrongful classification by the methods. Foram 5 has an unusual shape, but was confidently predicted correctly by humans, likely due to subtle but observable chambers and sutures reminiscent of calcareous benthic forams. Foram 8 was confidently predicted correctly by the participants, probably due to an observable planispiral chamber arrangement combined with pronounced sutures, and it is difficult to explain why deep learning models struggled. Foram 10 is partially broken and was also labeled as hard by experts.

We note that, even though the models were trained on a balanced training dataset with respect to foraminifera groups, the balance in terms of species may be vastly different for the training and the test set. Thus, many of the misclassifications in the benthic calcareous group would possibly be avoided if more examples of the benthic calcareous species found in the test set were also present in the training data. This is a challenge that the deep learning model does not share with the participants, as a geoscientist will have a 'training dataset' covering a wide spectrum of benthic calcareous species encountered during their career.

Table 3
Comparing deep learning and experts. The table shows to which extent deep learning uncertainty methods aligns with humans. We report the kappa score on the predictions of the methods compared to humans, and how many images that both the methods and the humans misclassified (*Shared mistakes*). The rank correlation is computed between the uncertainty estimates for the methods and the humans, while *Easy mistakes* shows the number of misclassifications that were categorized as easy by the experts. The *Agree on easy images* and *agree on hard images* columns shows how many test images that were categorized as easy/hard by both the experts and the deep learning methods.

	Kappa	Shared mistakes	Rank corr.	Easy mistakes	Agree on easy images	Agree on hard images
Baseline	0.86	3	0.11	9	73	5
Baseline + Dropout	0.85	3	0.11	10	77	5
Baseline + SWAG (5e6)	0.86	3	0.11	9	73	5
Baseline + SWAG (5e5)	**0.87**	3	0.11	**6**	74	4
Baseline + SWAG (1e4)	0.86	3	0.15	8	76	4
Baseline + SWAG (5e4)	0.79	3	−0.04	14	65	6
Baseline + TTA (s = 1)	**0.87**	2	**0.26**	**6**	**80**	**8**
Baseline + TTA (s = 2)	0.81	3	0.12	9	71	7
Baseline	0.83	2.1	0.20	9.9	**78.2**	5.3
Ensemble	**0.87**	2	**0.21**	**6**	76	**7**
SGLD (5e6)	0.85	2	0.13	7	76	3
SGLD (5e5)	0.86	2	0.11	10	76	3
SGLD (1e4)	0.82	3	0.08	12	73	1

4.3. Comparing deep learning with the participants

In the following analysis, we aim to uncover to what extent the deep learning methods align with the human participants in terms of both the predictions, and the uncertainty estimates. For this purpose, we calculated the Cohen's kappa coefficient and Spearman's rank correlation between human responses and deep learning. Cohen's kappa was computed to measure the agreement in predictions, while Spearman's rank correlation was computed to measure the correlation between the uncertainty estimates. Moreover, we also calculated the intersection between the hard and easy groups of humans and deep learning. The statistics are reported in Table 3, and we discuss the results in the following paragraphs.

Agreement in predictions. Except for SWAG (5e4), TTA (s = 2) and SGLD (1e4), the kappa scores were in the mid-80s range for all methods.[2] Ensembling had the biggest effect on the kappa score, and resulted in a notable increase from 0.83 to 0.87 against the baseline. We also note that mistakes made by the participants and mistakes made by the deep learning methods forms two almost disjoint sets with an intersection of 2 or 3 examples, showing that the participants and the AI methods in general did not make the same mistakes. Although the experts exclusively misclassified benthic agglutinated and sediments, the methods had some additional difficulties with planktic and benthic calcareous (see Appendix G in the appendix for an analysis of the differences).

Uncertainty correlation. Table 3 indicates that there is in general a low correlation between human uncertainty estimates and deep learning. This is also supported by the intersection between the easy group of participants and the easy group of deep learning, which is just above the chance level[3] for most methods. TTA (s = 1) stands out with a higher rank correlation, a higher easy group intersection, a higher hard group intersection, and fewer easy errors compared to the other methods. Ensembling did not increase the easy group intersection, but it scored better on the other metrics compared to the baseline. Both ensemble and TTA(s = 1) made predictions and confidence statements that aligned more with the responses of the participants.

4.4. Confidence calibration

The calibration and reliability errors were computed as defined in Section 2.6, and are reported along with the mean confidence in Table 4. Since the errors are dependent upon the chosen number of bins, we computed the error as an average over 5 different values; 5, 10, 15, 20, and 25 bins. Our results show that ensembling reduced the mean confidence and improved the calibration. When comparing MC Dropout, SWAG and TTA against the baseline model, the results show no changes in mean confidence, but the methods improved the calibration, with TTA (s = 1) obtaining the best results. The table indicates that most of the methods (with the proper configuration) have the potential to improve calibration.

Table 4
Confidence calibration. The table shows the average confidence, along with the calibration and reliability error. The results show that TTA (s = 1) and ensembles has a positive effect on the calibration error. Ensembles also reduced the mean confidence, resulting in more conservative outputs.

	Av. Confidence	Mean Cal. Error	Mean Rel. Error
Baseline	96 ± 09	0.143	0.0494
Baseline + Dropout	95 ± 11	0.137	0.0459
Baseline + SWAG (5e6)	96 ± 10	0.142	0.0457
Baseline + SWAG (5e5)	96 ± 10	0.141	0.0471
Baseline + SWAG (1e4)	96 ± 10	0.141	0.0513
Baseline + SWAG (5e4)	92 ± 14	0.245	0.1015
Baseline + TTA (s = 1)	96 ± 11	**0.134**	**0.0453**
Baseline + TTA (s = 2)	95 ± 10	0.217	0.0775
Baseline	96 ± 10	0.170	0.0680
Ensemble	93 ± 12	**0.123**	**0.0454**
SGLD (5e6)	96 ± 11	0.168	0.0636
SGLD (5e5)	96 ± 10	0.124	0.0435
SGLD (1e4)	92 ± 14	0.135	0.0558

Fig. 7 shows how ensembling affects calibration. The calibration diagrams (a and b) show model confidence (dashed line) against the actual observed frequency of correct predictions (green bar), while the reliability diagrams (c and d) show the confidence subtracted by the observed frequency of correct predictions. It is evident that the deep learning models before ensembling underestimated the uncertainty for high-confidence predictions (Fig. 7a and c). The problem of overconfident predictions was to some extent corrected for by ensembling the predictions (Fig. 7b and d), and the ensemble confidence statements were actually overestimating the uncertainty for predictions with confidence statements in the range 0.4 to 0.9 (Fig. 7b). For completeness, we include calibration and reliability diagrams for the other methods in Fig. 8. TTA(s = 1) in particular has a similar effect on confidence statements, with a clear increase in accuracy for high confident predictions (Fig. 8d and Fig. 8h).

4.5. Combining ensembles and TTA

Our results from the previous sections show that the ensembles and TTA (s = 1) worked best in terms of improving accuracy, reducing the

[2] Given that the methods obtains a slightly lower accuracy than the experts, the theoretical maximum kappa is around 0.96.

[3] The expectation for the hypergeometric distribution is 68 in this case, and the probability of getting at least 74 successes by chance is about 0.95.

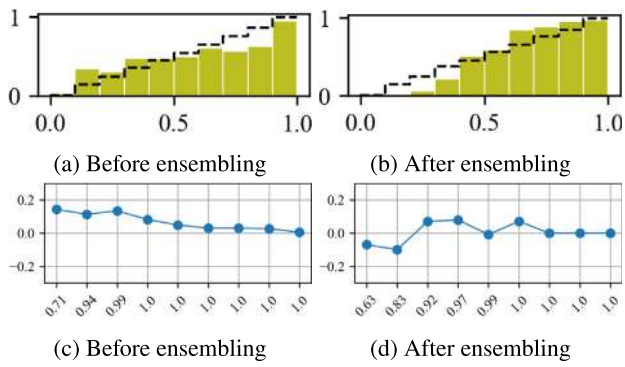


Fig. 7. Ensemble calibration and reliability plots The calibration diagrams (a and b) show model confidence (dashed line) together with model accuracy (green bars). The *x*-axis represents predicted confidence, while the *y*-axis represents observed accuracy. The reliability diagrams (c and d) show calibration with an equal size binning strategy. Here, the *y*-axis represents confidence minus accuracy. The individual deep learning models were in general overconfident in their predictions (a and c). By ensembling the predictions, we obtained confidence scores that were more conservative than what the individual models gave on average (b and d).

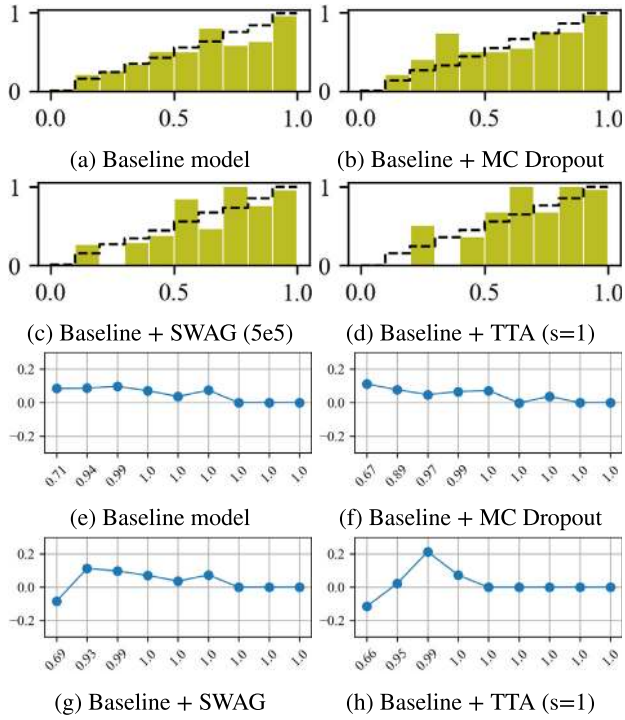


Fig. 8. Calibration and reliability plots The calibration diagrams (a and b) show model confidence (dashed line) together with model accuracy (green bars). The *x*-axis represents predicted confidence, while the *y*-axis represents observed accuracy. The reliability diagrams (c and d) show calibration for bins of equal size. Here, the *y*-axis represents confidence minus accuracy. The calibration diagrams that MC Dropout, SWAG and TTA constructs more conservative confidences compared to a baseline model.

calibration error and increasing the alignment of the uncertainty estimates compared to those of the participants. We therefore investigated the effect of combining the two approaches. For this analysis, we tested 10 different input variations (augmentations) with strength 1 on all 10 members of the ensemble. This resulted in 100 predictions per image, where we aggregated the outputs using the median as when using TTA before. The results of the combined approach are summarized in Table 5. The combined approach resulted in a clear improvement in accuracy, confidence calibration, and correlation of the uncertainty estimates compared to the ensemble and the baseline.

4.6. Discussion

We conclude this section with a summary of our key findings, a discussion of the practical benefits of uncertainty estimation, and a discussion of the limitations of this study.

Main findings.

1. **Better uncertainty through weightings** The experts were in high agreement in their prediction. As in Fenton et al. (2018), a high confidence was associated with a low probability of error. However, the rank correlation between the confidence scores suggested a low consistency between the confidence statements of the experts. Using a weighted uncertainty metric, we were able to take this information into account, and our claim is that this metric better captures the true uncertainty associated with each specimen.
2. **Deep learning can achieve human-level performance** With an ensemble of deep learning models, we were able to obtain performance comparable to that of an expert. Due to factors such as specimen orientation and lighting, it is unreasonable to expect perfect performance when classifying 2D foraminifera, which is evident from the suboptimal accuracy obtained by the experts. Our results are consistent with the conclusion in Mitra et al. (2019) where the machine obtained comparable or better results than humans.
3. **Data augmentation and ensembling improves performance** Our results showed that TTA and ensembling stand out among the investigated methods. Both of these methods gave a substantial improvement in accuracy compared to the conventional single-model baseline, were better calibrated, and gave subsequent uncertainty estimates that aligned better with the human-based uncertainty estimates.
4. **Deep learning uncertainty estimates falls short to humans** The participants outperformed the deep learning models to estimate uncertainty, and deep learning uncertainty estimates did not align well with human estimates. However, our impression is that techniques such as ensembling and test-time augmentation (TTA) show promise in improving uncertainty estimation. By incorporating additional elements, it may be possible to bridge the gap between uncertainty estimates and their practical applications.

Practical considerations in uncertainty estimation.

Ensembling and TTA address uncertainty in an intuitive and natural way by elucidating the sensitivity to initial conditions and data sampling (ensembling), and test image characteristics such as orientation and color distribution (TTA). Our hypothesis is that the efficacy of these methods will depend on the robustness of the model, which in turn relies on, for instance, the amount of data available for training. It is reasonable to believe that the improvements will be lower for truly robust models trained on large and diverse datasets, but since these methods do not interfere with the learning process itself, we believe that ensembling provides a net effect that is at least nonnegative. Whether these methods are appropriate for implementation or not will depend on the problem and the data at hand. Although ensembling is commonly applied to ensure robustness in predictive models, input perturbations, as with TTA, may not always be appropriate for a model based on physical measurements.[4]

However, a drawback of implementing these methods is that they scale linearly in computational complexity, and also in memory consumption when using ensembles. This may or may not be a decisive drawback for implementation and will depend on how speed is

[4] See, e.g., Andersson et al. (2021) who uses ensembling, but avoid input perturbations for this reason.

Table 5
Combining ensembling and TTA. By combining ensembling with TTA, we obtain improved predictive performance (accuracy and easy mistakes), as well as more calibrated confidence statements (calibration error). The combined approach provides confidence statements that also aligns better with humans with an increased rank correlation, and more overlap between the easy and hard groups.

	Acc.	NLL	Brier Score	Easy mistakes	Detected mistakes	Cal. Error	Rank corr.	Agree on easy images	Agree on hard images
Baseline	0.90	0.44	0.16	2.8	**12.5/25.8 (48%)**	0.170	0.20	78.2	5.3
Ensemble	0.93	0.26	0.12	2	8/18 (44%)	0.123	0.21	76	7
Ensemble + TTA (s = 1)	**0.94**	**0.25**	**0.09**	**0**	7/15 (47%)	**0.111**	**0.29**	**84**	**10**

weighted against performance. As model training is only done once, increasing the training time from 30 min to 5 h may be perfectly acceptable, but spending 1 min compared to 0.1 s on inference for a single sample may not. A possible solution in that scenario could be a simple screening process for high-speed predictions and a more elaborate uncertainty estimation pipeline for borderline cases.

Uncertainty estimates can be beneficial both from a practical and a research perspective. Identifying instances where humans are certain but the deep learning models are not could provide valuable information in the pursuit of optimizing models until they reach (at least) human performance. On the other hand, in complex problems where uncertainty is high for humans but low for algorithms, post hoc techniques can be used to interpret which features of the images the models are utilizing to perform the classification. This analysis could potentially reveal morphological features that are not even evident to experts.

Uncertainty methods have their limitations, and it is unrealistic to believe that UQ methods alone will correct all models of overconfidence. We note, for example, that nine out of the ten images displayed in Fig. 6 were still misclassified after combining the ensembling and TTA (all but the second). Except for foram no. 3 from the top left, the forams were also misclassified with high confidence. Investigating the reasons for this is beyond the scope of this paper, but we hypothesize that a data set imbalance would potentially be a contributing factor, for example, because background particles and partial forams were unevenly distributed in the training set. An uneven distribution of species across the training and test set could also have a negative impact, even if the training set was balanced on a class level. It is obvious that UQ methods should be an addition to and not a replacement for other methods to increase the robustness of model training. We do believe that, for example, ensembles or augmentations will have an additional benefit in this respect, as they can help uncover systematic weaknesses in the model when mistakes reoccur across multiple sets of weights or augmentations.

Limitations of this study.

- Our study shows how experts and deep learning compare in a still image classification task. We note that experts usually conduct a careful analysis under the microscope studying the foraminifera from different orientations and that the classification results reported here are not representative of how a human expert is expected to perform. In fact, we assume that all experts involved here would obtain a perfect score in a normal analysis setting.
- This work focused on planktic and benthic forams from arctic and subarctic samples, which are extensively studied in the Arctic, and the samples were divided into four groups. Thus, our data set is limited in size and is, for instance, much smaller than Endless Forams (Hsiang et al., 2019).
- Using humans to generate a baseline for comparison of uncertainty estimates relies on the assumption that the human and the deep learning model will rely on the same information when doing the classification. In general, this is, of course, difficult to assess and something that we cannot claim or prove. Some studies (see, e.g., Peterson et al., 2018) suggest a strong correspondence between similarities between latent representations and evaluated similarities from humans. Since the morphological characteristics are what define the groups in foraminifera analysis, it is not unnatural to assume that the deep learning model will learn to classify based on at least partly the same features and, in turn, will have trouble in classifying some of the same images as humans.

5. Conclusions

In this paper, we present test results from humans and AI and create baseline uncertainty estimates from human participants, which deep learning methods can compare. The human baseline provides insight into how human experts perform when given the same classification task as a deep learning model, and we have demonstrated how useful uncertainty estimates can be constructed from the participants. The human uncertainty estimates proved to be efficient in identifying difficult samples and therefore provide a relevant baseline for comparison. Moreover, we have highlighted the qualities of deep learning UQ methods with analysis from a new perspective. In addition to accuracy and confidence calibration, we used Cohen's kappa score, Spearman's rank correlation, and group intersection to measure to what extent deep learning-based uncertainty estimates aligned with the human baseline. Our results show that TTA and the ensembles improved calibration and accuracy and made predictions and confidence statements that were more in line with the human participants. We also saw that the combination of the two approaches provided further benefits in the aforementioned metrics. Combining several approaches that address several dimensions of uncertainty in model training in a natural and intuitive way is a promising avenue for further research.

CRediT authorship contribution statement

Iver Martinsen: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Steffen Aagaard Sørensen:** Writing – original draft, Data curation, Conceptualization. **Samuel Ortega:** Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Fred Godtliebsen:** Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Miguel Tejedor:** Writing – original draft, Methodology. **Eirik Myrvoll-Nilsen:** Writing – original draft, Methodology.

Funding

This document is the result of a research project funded by the Norwegian Research Council (IKTPLUSS-IKT og digital innovasjon, project no. 332901). This document is also the result of a research project funded by SFI Visual Intelligence. Visual Intelligence projects are financially supported by the Research Council of Norway, through its Centre for Research-based Innovation funding scheme (grant no. 309439), and Consortium Partners.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Fig. A.1. Software for testing The image shows a screenshot of the software that was used for foraminifera testing. The participants were tasked with choosing a class from four possible choices, in addition to a confidence statement (certainty level).

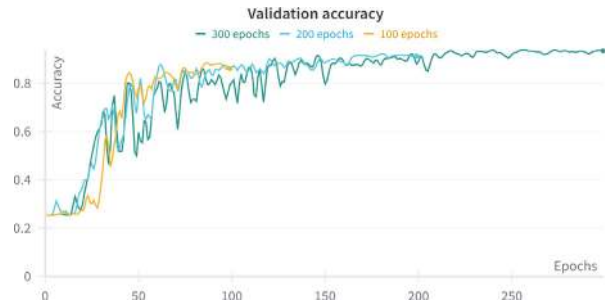


Fig. C.2. Effect of epochs We obtained 86% accuracy after 100 epochs, 91% after 200 and 93% after 300 epochs on the validation data, showing that more epochs gave better training. The model showed no signs of overfitting.

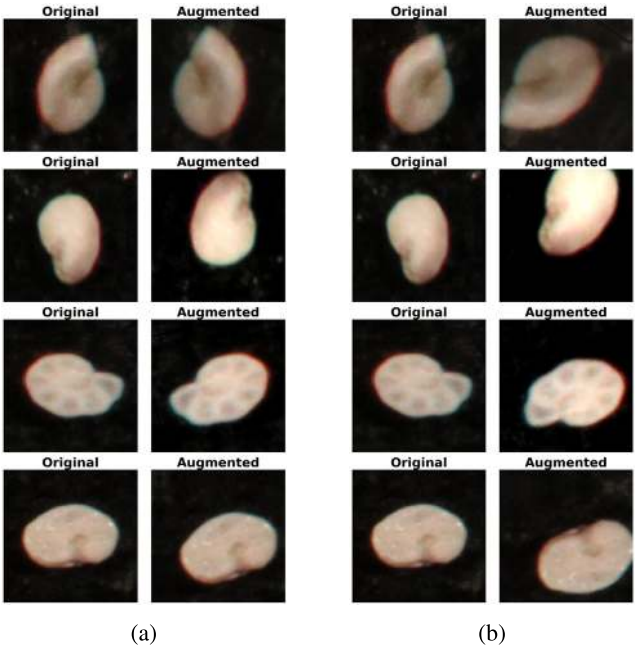


Fig. C.3. Data augmentation The figure displays the effect of data augmentation with strength 1 (a) and 2 (b). The perturbation of the object position and color distribution increases with increased augmentation strength, but the forams still remain recognizable.

Acknowledgments

The authors wish to thank Katarzyna Zamelczyk, Marit-Solveig Seidenkrantz, Morten Hald, and Tine L. Rasmussen for participating in the foraminifera questionnaire. We also thank Kristoffer Wickstrøm and Benjamin Ricaud at *UiT — The Arctic University of Norway* for providing valuable feedback in the writing process.

Appendix A. Software

The GUI for the foraminifera classification software that was used to test the participants is illustrated in Fig. A.1.

Appendix B. Uncertainty quantification background theory

As a supplement to the model confidence, which we use as a measure of uncertainty in this paper, we discuss three other approaches for uncertainty estimation in this section.

Epistemic uncertainty. Epistemic uncertainty in this context is understood as the variability between predictions due to changes in the model (for instance by sampling model weights). By looking at the output variability instead of confidence, we can highlight the epistemic uncertainty in cases where the aleatory term dominates the epistemic term (Ayhan and Berens, 2018). Computing the epistemic uncertainty is straightforwardly done by calculating the percent agreement, Interquartile Range (IQR), output variance, or any other measure of dispersion. None of these measurements will directly address confidence scores, but can be highly correlated with confidence.

Aleatory uncertainty. As proposed by Kendall and Gal (2017) and discussed further in Valdenegro-Toro and Mori (2022), one approach for uncertainty estimation is through the modeling of the aleatory uncertainty (σ_{Ale}) using the *logit* of the model. For context, in the example of binary classification, the *logit* is understood to be the value z_i in Eq. (B.1),

$$p_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}} \tag{B.1}$$

where p_i is the output confidence for example i, and $\sigma(\cdot)$ is the sigmoid function which is commonly used at the output. Assuming heteroscedasticity, this is solved in practice by adding another output to the network. The added output represents the estimated variance of the prediction, which makes the network assess the variance at the logit level.

The output uncertainty is summarized through the posterior predictive variance, which can be written as in Eq. (B.2),

$$\text{Var}\{z(x_i)\} = \text{E}\{\sigma_\theta(x_i)\} + \text{Var}\{z_\theta(x_i)\} \tag{B.2}$$

disentangling the uncertainty into an aleatory and an epistemic term (Kendall and Gal, 2017; Valdenegro-Toro and Mori, 2022). This way of modeling aleatory uncertainty is however less natural than in regression, as we do not have a ground truth for the logits, and since the estimated variance is not directly incorporated into the loss function (see Kendall and Gal, 2017; Valdenegro-Toro and Mori, 2022).

Integrated approach. An approach for uncertainty estimation is to assume that the class label y_i is a Bernoulli.[5] random variable with

[5] Or a categorical variable in the multiclass setting.

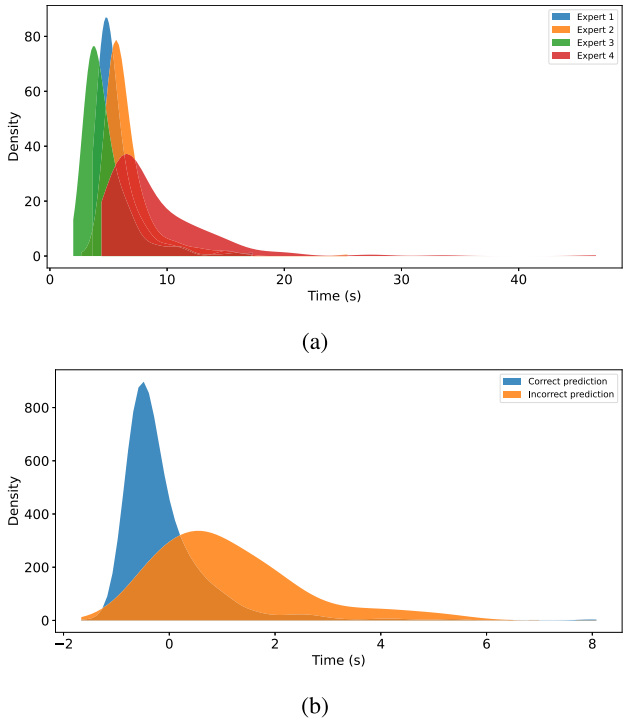


Fig. D.4. Expert results (a) Time distribution for the individual participants. (b) Overall time distribution after standardization. Subtraction of the mean results in a negative x-axis for the standardized distribution.

probability p_i. Since the Bernoulli distribution (in contrast to, e.g., the normal distribution), is fully specified by a single parameter, we can interpret p_i as an implicit modeling of the aleatory uncertainty. The posterior predictive variance in this case is given by Eq. (B.3)

$$\mathrm{Var}\{y_i\} = \mathrm{E}\{\mathrm{Var}_\theta(y_i)\} + \mathrm{Var}\{E_\theta(y_i)\} \tag{B.3}$$

Since the variance is highly correlated with p for the Bernoulli distribution, however, the epistemic term (the second term) will be negligible compared to the aleatory term when the output p_i is large.

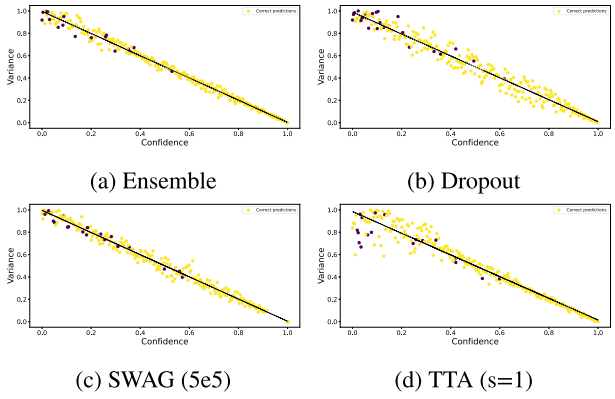


Fig. E.5. Epistemic uncertainty The plot shows mean confidence plotted against sample variance for a handful of the methods. The yellow dots show correct predictions, while the purple dots show incorrect predictions. There is a strong correlation between the confidence statements and the variance, suggesting that there is little to gain by including variance across predictions in the uncertainty calculations.

Appendix C. Hyperparameters

Learning rate. We used the RMSprop optimizer with a polynomial decaying learning (Eq. (C.1)) rate schedule. The learning rate schedule also included a linear warm-up phase for the first 10 epochs. The warm-up phase helps to prevent optimization challenges early in training (Goyal et al., 2018), while the decaying learning helps to ensure a proper exploration of the loss surface early in training and a stable loss towards the end of training. The learning rate α_t for time step t is given by Eq. (C.1)

$$\alpha_t = (\alpha_0 - \alpha_T)(1 - t/T)^\beta + \alpha_T. \tag{C.1}$$

where T equals the total number of training steps, and α_0 and α_T are the learning rates at the start and end of training. Values of $\alpha_0 = 5{\cdot}10^{-4}$, $\alpha_T = 5 \cdot 10^{-6}$ and $\beta = 2$ gave reasonable results during hyperparameter tuning. An important property of the learning rate decay is that it satisfies the quadratic requirement of convergence in SGLD (Welling and Teh, 2011).

Epochs. The networks were trained with a batch size of 64 for 300 epochs. We investigated training with 100, 200, and 300 epochs. We used a fraction of the training data for validation, and we observed

Table F.6
Agreement between models. Cohen's kappa score between the different methods show that there is a high agreement between baseline, dropout, and SWAG. TTA creates predictions that deviate the most from the other methods.

	Baseline	Dropout	SWAG (1e4)	SWAG (5e5)	TTA (s = 1)	TTA (s = 2)
Baseline	1.00	0.98	0.98	0.98	0.92	0.84
Dropout	0.98	1.00	0.96	0.97	0.92	0.84
SWAG (1e4)	0.98	0.96	1.00	0.98	0.92	0.84
SWAG (5e5)	0.98	0.97	0.98	1.00	0.93	0.85
TTA (s = 1)	0.92	0.92	0.92	0.93	1.00	0.91
TTA (s = 2)	0.84	0.84	0.84	0.85	0.91	1.00

Table F.7
Rank correlation between models. The table shows Spearman's rank correlation coefficient between uncertainty estimates from the different methods. TTA affects the uncertainty rankings to a greater extent than the other methods, reflected by the low correlation between the other methods.

	Baseline	Dropout	SWAG (1e4)	SWAG (5e5)	TTA (s = 1)	TTA (s = 2)
Baseline	1.00	0.94	0.97	0.99	0.63	0.42
Dropout	0.94	1.00	0.91	0.92	0.63	0.48
SWAG (1e4)	0.97	0.91	1.00	0.99	0.64	0.40
SWAG (5e5)	0.99	0.92	0.99	1.00	0.62	0.42
TTA (s = 1)	0.63	0.63	0.64	0.62	1.00	0.74
TTA (s = 2)	0.42	0.48	0.40	0.42	0.74	1.00

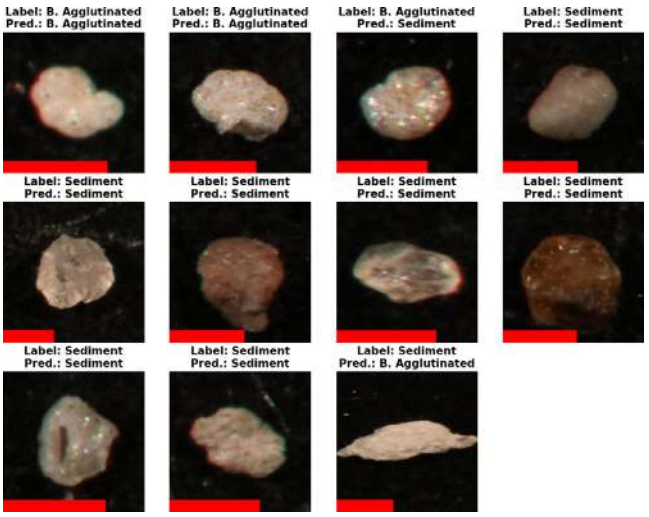


Fig. G.6. Expert mistakes The figure shows the ensemble performance on the forams where the expert were wrong.

increased performance when increasing the number of epochs (Fig. C.2).

Data augmentation. We investigated two different augmentation strengths, the effect of which is illustrated in Fig. C.3.

Appendix D. Expert time usage

The experts time usage distributions are shown in Fig. D.4, and we see from (a) that the mean time usage differed between the participants. We therefore standardized the time usage by subtracting individual means and dividing by individual standard deviations to obtain an overall time distribution for all participants (Fig. D.4b). The standardized plot shows that the participants had a higher mean time usage when they were wrong (orange curve) compared to when they were correct (blue curve), indicating the possibility of using information from time usage in future work.

Appendix E. Epistemic uncertainty results

To investigate the correlation between the mean confidence and the variance between predictions, we plotted the mean confidence against the variance for some of the methods (Fig. E.5). The figure shows a high negative correlation between confidence and variance, indicating that there is little loss of information when excluding the output variance in the uncertainty calculations. Moreover, the plot did not indicate that wrong predictions (purple dots) exhibited a higher epistemic variance than correct predictions (yellow dots).

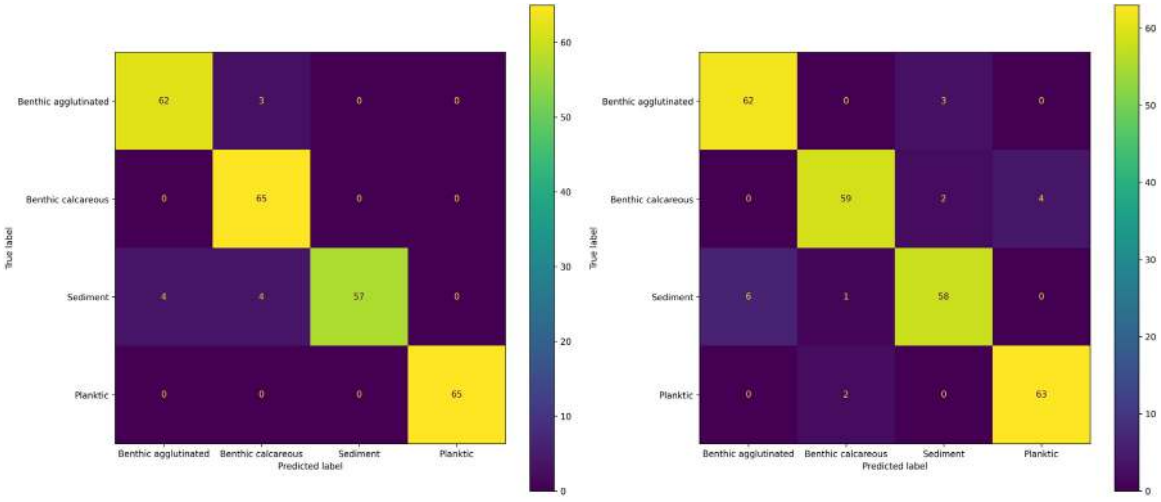
Appendix F. Correlation between methods

We include an analysis to answer whether the different methods lead to the same conclusions in terms of predictions and confidence statements. Table F.6 shows the Cohen's kappa scores between predictions from the different methods, and Table F.7 shows the rank correlation for the confidence scores for the different methods. The tables show that MC Dropout and SWAG agrees with the baseline in both prediction and confidence, while TTA stands out compared to the others. The low rank correlation between TTA and the other methods also shows that TTA has a large effect on the uncertainty rankings.

Appendix G. Comparison between experts and methods

The mistakes made by the experts and the methods were in general not the same. Fig. G.6 shows the ensemble predictions for the forams where the expert were wrong.

To help analyze for which groups the experts and the methods had trouble, we show the confusion matrices for experts and ensemble in Fig. G.7. Both the experts and the method had trouble with benthic agglutinated and sediment, however the method has some additional difficulties in classifying Benthic calcareous forams. A possible reason for this is the large diversity in the benthic calcareous group, which contains a range of species that may not be adequately covered in the training set when training the ensemble.



(a) (b)

Fig. G.7. Confusion matrices The expert confusion matrix (a) shows that the participants mistakes were restricted to sediment and benthic agglutinated forams only. As shown in (b), the ensemble also had trouble with benthic agglutinated and sediments, and in addition confused benthic calcareous and planktics.

References

Al-Sabouni, N., Fenton, I.S., Telford, R.J., Kučera, M., 2018. Reproducibility of species recognition in modern planktonic foraminifera and its implications for analyses of community structure. J. Micropalaeontol. 37 (2), 519–534. http://dx.doi.org/10.5194/jm-37-519-2018, URL https://jm.copernicus.org/articles/37/519/2018/, Publisher: Copernicus GmbH.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D., 2016. Concrete Problems in AI Safety. http://dx.doi.org/10.48550/arXiv.1606.06565, URL http://arxiv.org/abs/1606.06565, arXiv:1606.06565.

Andersson, T.R., Hosking, J.S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D.C., Wilkinson, J., Phillips, T., Byrne, J., Tietsche, S., Sarojini, B.B., Blanchard-Wrigglesworth, E., Aksenov, Y., Downie, R., Shuckburgh, E., 2021. Seasonal Arctic sea ice forecasting with probabilistic deep learning. Nat. Commun. 12 (1), 5124. http://dx.doi.org/10.1038/s41467-021-25257-4, URL https://www.nature.com/articles/s41467-021-25257-4, Number: 1 Publisher: Nature Publishing Group.

Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: Proceedings of the 2018 Medical Imaging with Deep Learning Conference. URL https://openreview.net/forum?id=rJZz-knjz.

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. IEEE, pp. 248–255.

Emiliani, C., 1955. Pleistocene temperatures. J. Geol. 63 (6), 538–578. http://dx.doi.org/10.1086/626295, URL https://www.journals.uchicago.edu/doi/10.1086/626295, Publisher: The University of Chicago Press.

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., Purvis, A., 2018. Factors affecting consistency and accuracy in identifying modern macroperforate planktonic foraminifera. J. Micropalaeontol. 37 (2), 431–443. http://dx.doi.org/10.5194/jm-37-431-2018, URL https://jm.copernicus.org/articles/37/431/2018/, Publisher: Copernicus GmbH.

Ferreira-Chacua, I., Koeshidayatullah, A., 2023. ForamViT-GAN: Exploring new paradigms in deep learning for micropaleontological image analysis. http://dx.doi.org/10.48550/arXiv.2304.04291, URL http://arxiv.org/abs/2304.04291, arXiv:2304.04291 [cs, eess].

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning. PMLR, pp. 1050–1059, URL https://proceedings.mlr.press/v48/gal16.html, ISSN: 1938-7228.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2022. A Survey of Uncertainty in Deep Neural Networks. http://dx.doi.org/10.48550/arXiv.2107.03342, URL http://arxiv.org/abs/2107.03342, arXiv:2107.03342 [cs, stat].

Ge, Q., Zhong, B., Kanakiya, B., Mitra, R., Marchitto, T., Lobaton, E., 2017. Coarse-to-fine foraminifera image segmentation through 3D and deep features. In: 2017 IEEE Symposium Series on Computational Intelligence. SSCI, pp. 1–8. http://dx.doi.org/10.1109/SSCI.2017.8280982.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2018. Accurate, large minibatch SGD: Training ImageNet in 1 hour. http://dx.doi.org/10.48550/arXiv.1706.02677, URL http://arxiv.org/abs/1706.02677, arXiv:1706.02677 [cs].

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR, pp. 1321–1330, URL https://proceedings.mlr.press/v70/guo17a.html, ISSN: 2640-3498.

Hald, M., Andersson, C., Ebbesen, H., Jansen, E., Klitgaard-Kristensen, D., Risebrobakken, B., Salomonsen, G.R., Sarnthein, M., Sejrup, H.P., Telford, R.J., 2007. Variations in temperature and extent of Atlantic Water in the northern North Atlantic during the Holocene. Quat. Sci. Rev. 26 (25), 3423–3440. http://dx.doi.org/10.1016/j.quascirev.2007.10.005, URL https://www.sciencedirect.com/science/article/pii/S0277379107002697.

Hayward, B., Le Coze, F., Vachard, D., Gross, O., 2024. World foraminifera database. http://dx.doi.org/10.14284/305, URL https://www.marinespecies.org/foraminifera.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. http://dx.doi.org/10.48550/arXiv.1512.03385, URL http://arxiv.org/abs/1512.03385, arXiv:1512.03385 [cs].

Hendrycks, D., Gimpel, K., 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks. http://dx.doi.org/10.48550/arXiv.1610.02136, URL http://arxiv.org/abs/1610.02136, arXiv:1610.02136 [cs].

Hsiang, A.Y., Brombacher, A., Rillo, M.C., Mleneck-Vautravers, M.J., Conn, S., Lordsmith, S., Jentzen, A., Henehan, M.J., Metcalfe, B., Fenton, I.S., Wade, B.S., Fox, L., Meilland, J., Davis, C.V., Baranowski, U., Groeneveld, J., Edgar, K.M., Movellan, A., Aze, T., Dowsett, H.J., Miller, C.G., Rios, N., Hull, P.M., 2019. Endless forams: 34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. Paleoceanogr. Paleoclimatology 34 (7), http://dx.doi.org/10.1029/2019PA003612, URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2019PA003612.

Izmailov, P., Vikram, S., Hoffman, M.D., Wilson, A.G., 2021. What are Bayesian neural network posteriors really like?. http://dx.doi.org/10.48550/arXiv.2104.14421, URL http://arxiv.org/abs/2104.14421, arXiv:2104.14421 [cs, stat].

Johansen, T.H., Sørensen, S.A., 2020. Towards detection and classification of microscopic foraminifera using transfer learning. http://dx.doi.org/10.48550/arXiv.2001.04782, URL http://arxiv.org/abs/2001.04782, arXiv:2001.04782 [cs, stat].

Johansen, T.H., Sørensen, S.A., Møllersen, K., Godtliebsen, F., 2021. Instance segmentation of microscopic foraminifera. Appl. Sci. 11 (14), 6543. http://dx.doi.org/10.3390/app11146543, URL https://www.mdpi.com/2076-3417/11/14/6543, Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

Kathal, P.K., Nigam, R., Talib, A., 2017. Micropaleontology and Its Applications. vfGJDwAAQBAJ, Scientific Publishers, Google-Books-ID.

Katz, M.E., Cramer, B.S., Franzese, A., Hönisch, B., Miller, K.G., Rosenthal, Y., Wright, J.D., 2010. Traditional and emerging geochemical proxies in foraminifera. J. Foraminifer. Res. 40 (2), 165–192. http://dx.doi.org/10.2113/gsjfr.40.2.165.

Kendall, A., Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. http://dx.doi.org/10.48550/arXiv.1703.04977, URL http://arxiv.org/abs/1703.04977, arXiv:1703.04977 [cs].

Kiureghian, A.D., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Risk Acceptance and Risk Communication, Struct. Saf. Risk Acceptance and Risk Communication, 31 (2), 105–112. http://dx.doi.org/10.1016/j.strusafe.2008.06.020.URL https://www.sciencedirect.com/science/article/pii/S0167473008000556,

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Li, C., Chen, C., Carlson, D., Carin, L., 2016. Preconditioned stochastic gradient langevin dynamics for deep neural networks. Proc. the AAAI Conf. Artif. Intell. 30 (1), http://dx.doi.org/10.1609/aaai.v30i1.10200, URL https://ojs.aaai.org/index.php/AAAI/article/view/10200, Number: 1.

Maddox, W.J., Izmailov, P., Garipov, T., Vetrov, D.P., Wilson, A.G., 2019. A simple baseline for Bayesian uncertainty in deep learning. In: Advances in Neural Information Processing Systems. Vol. 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

Marchant, R., Tetard, M., Pratiwi, A., Adebayo, M., de Garidel-Thoron, T., 2020. Automated analysis of foraminifera fossil records by image classification using a convolutional neural network. J. Micropalaeontol. 39 (2), 183–202. http://dx.doi.org/10.5194/jm-39-183-2020, URL https://jm.copernicus.org/articles/39/183/2020/, Publisher: Copernicus GmbH.

Martinsen, I., Wade, D., Ricaud, B., Godtliebsen, F., 2024. The 3-billion fossil question: How to automate classification of microfossils. Artif. Intell. Geosci. 100080. http://dx.doi.org/10.1016/j.aiig.2024.100080, URL https://www.sciencedirect.com/science/article/pii/S2666544124000212.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Medica 22 (3), 276–282, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/.

Mitra, R., Marchitto, T.M., Ge, Q., Zhong, B., Kanakiya, B., Cook, M.S., Fehrenbacher, J.S., Ortiz, J.D., Tripati, A., Lobaton, E., 2019. Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. Mar. Micropaleontol. 147, 16–24. http://dx.doi.org/10.1016/j.marmicro.2019.01.005, URL https://www.sciencedirect.com/science/article/pii/S0377839818301105.

Neal, R.M., 2012. MCMC using Hamiltonian dynamics. http://dx.doi.org/10.1201/b10905, URL https://arxiv.org/abs/1206.1901v1.

Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., Tran, D., 2020. Measuring calibration in deep learning. http://dx.doi.org/10.48550/arXiv.1904.01685, URL http://arxiv.org/abs/1904.01685, arXiv:1904.01685.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. Vol. 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Peterson, J.C., Abbott, J.T., Griffiths, T.L., 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. Cogn. Sci. 42 (8), 2648–2669. http://dx.doi.org/10.1111/cogs.12670, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12670, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12670.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958, Publisher: JMLR. org.

Steinsund, P.I., Hald, M., 1994. Recent calcium carbonate dissolution in the Barents Sea: Paleoceanographic applications. Mar. Geol. 117 (1), 303–316. http://dx.doi.org/10.1016/0025-3227(94)90022-1, URL https://www.sciencedirect.com/science/article/pii/0025322794900221.

Tan, M., Le, Q.V., 2020. EfficientNet: Rethinking model scaling for convolutional neural networks. http://dx.doi.org/10.48550/arXiv.1905.11946, URL http://arxiv.org/abs/1905.11946, arXiv:1905.11946 [cs, stat].

Valdenegro-Toro, M., Mori, D.S., 2022. A deeper look into aleatoric and epistemic uncertainty disentanglement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 1508–1516. http://dx.doi.org/10.1109/CVPRW56347.2022.00157, URL https://ieeexplore.ieee.org/abstract/document/9857056?casa_token=g_ERrfwOwO0AAAAA:Duz9jttG-wBrgCq9oBFS2DDOkc0omtDKnI2HlelVxllg8KuoMARu7mUtnhXq3JXlVyA9KpJffg, ISSN: 2160-7516.

Welling, M., Teh, Y.W., 2011. Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning. ICML-11, pp. 681–688.

Zhong, B., Ge, Q., Kanakiya, B., Marchitto, R.M.T., Lobaton, E., 2017. A comparative study of image classification algorithms for Foraminifera identification. In: 2017 IEEE Symposium Series on Computational Intelligence. SSCI, pp. 1–8. http://dx.doi.org/10.1109/SSCI.2017.8285164.