

AI, citizen science data and expert opinions: towards an improved fossil identification process

Isaak Eijkelboom^{1,2}, Laurens E. Hogeweg^{2,3}, Django Brunink², Menno Jansen¹, Hansjorg Ahrens², Charlie Schouwenburg⁴, Floris ter Haar², Natasja den Ouden², Bram W. Langeveld⁵, Luc Amkreutz^{6,7}, Anne S. Schulp^{1,2} and Frank P. Wesselingh^{1,2,8}

¹ Department of Earth Sciences, Faculty of Geosciences, Utrecht University, Utrecht, Netherlands

² Naturalis Biodiversity Center, Leiden, Netherlands

³ Intel Benelux, Eindhoven, Netherlands

⁴ Oertijdmuseum, Boxtel, Netherlands

⁵ Natural History Museum Rotterdam, Rotterdam, Netherlands

⁶ National Museum of Antiquities, Leiden, Netherlands

⁷ Faculty of Archaeology, Leiden University, Leiden, Netherlands

⁸ Faculty of Science and Engineering, Maastricht University, Maastricht, Netherlands

ABSTRACT

Accurate identification of fossils is instrumental to palaeontological research but requires expert knowledge, is time consuming and subject to human biases. Through citizen science platforms and apps, artificial intelligence (AI)-assisted identifications can mitigate those challenges, as shown in successful biodiversity research applications. Digitised fossil data are relatively scarce, but a large and growing body of fossil datasets is becoming available in open data repositories through collection digitisation efforts. Furthermore, data are collected and validated by fossil enthusiasts on citizen science platforms. These datasets can be used as training data for deep learning classification models to provide both experts and citizen scientists with accurate, quick and easy to use tools to collect, validate and analyse palaeontological data. However, AI model performance may be limited by the size and quality of the training dataset. To explore the use of AI in fossil identifications, we present and compare a set of convolutional neural networks (CNNs) that are trained and tested on standardised images from museum and private collections (>46,000 images) and images from the online citizen science platform Oervondstchecker.nl (>74,000 images). Both datasets consist of Quaternary fossils and artefacts from the Netherlands and the southern North Sea Basin. Moreover, we compare model performance with identifications by ten domain experts and expert citizen scientists to gain a measure of data quality. This comparison shows that the AI models can outperform experts when trained and tested on standardised data, but not consistently across both datasets. Furthermore, expert identifications are far from unanimous, affecting the training data quality. Based on these insights we make recommendations on how to account for variable data quality to further optimise AI model training and performance. The synergy between AI model predictions and domain expert identifications can potentially rapidly increase the number of high-quality identifications of fossils and flag potential rare finds. To further increase data acquisition and public engagement, the models have been made publicly available online for use by professional experts, citizen scientists and the general public alike.

Submitted 28 October 2025

Accepted 9 June 2026

Published 4 August 2026

Corresponding author

Isaak Eijkelboom,
isaak.eijkelboom@naturalis.nl

Academic editor

Xiangjie Kong

Additional Information and
Declarations can be found on
page 20

DOI 10.7717/peerj-cs.4049

© Copyright

2026 Eijkelboom et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Neural Networks

Keywords Palaeontology, Machine learning, Convolutional neural network (CNN), Citizen science, Classification, Collection data, Validators, Data quality, Web application

INTRODUCTION

Correct fossil identifications form the basis of datasets for research on the evolution of biodiversity and ecosystems, but assembling such datasets and ensuring the fossils are correctly identified is no trivial task. Fossils are typically rare and often fragmented or otherwise altered, which may hinder high-quality identifications (Fig. 1). Many hours of training are required to gain expert knowledge and even then identifications can be time consuming. Furthermore, (human) biases such as repetition bias, learning bias, ambiguity bias, and recognisability bias may lead to different identifications by different experts (McKay & Kvammen, 2020; Koch et al., 2023). Depending on personal experience, preference, confidence, and time spent on the identification, experts might come to varying conclusions. Beyond this subjectivity, some fossils may be too fragmented or altered for a specific identification with confidence by human experts.

Machine learning (ML) models, such as convolutional neural networks (CNNs), are increasingly used in biodiversity research to assist in the assembly of large, high-quality datasets and thereby mitigate the challenges of the identification process (Shivaprakash et al., 2022; Pichler & Hartig, 2023; Miller et al., 2025). In palaeontology, ML for classification tasks is still much less widespread although the existing body of work has been quickly expanding in recent years (Liu et al., 2023, 2025; Yu et al., 2024; Eijkelboom et al., 2025). ML models have the capacity to quickly classify large quantities of (image) data at or near the accuracy of domain experts if provided with sufficient, high-quality data. In some cases, artificial intelligence (AI) can even outperform domain experts (Li et al., 2022; Feighelestein et al., 2025). In other cases, human-AI collaborations yield the best outcome (Chen et al., 2025). Moreover, CNNs are a tool for citizen scientists and enthusiasts to generate high-quality data, create new insights from large datasets and elevate public engagement (McClure et al., 2020). This citizen science-machine learning (CS-ML) approach (Eijkelboom et al., 2025) is explored and implemented here in the context of AI and expert identifications of Quaternary fossils and artefacts.

The accuracy of AI models largely depends on the size and quality of the training dataset, two factors which can be at odds with one another. Large and growing fossil datasets have become available through collection digitisation efforts, open data practices and professional and citizen scientists collecting, sharing and validating data on online platforms such as the Paleobiology Database (Peters & McClennen, 2015), DiSSCo (Hardisty et al., 2020), GeoCASE (CETAF Earth Sciences Group, 2020), the Fossil Forum (www.thefossilforum.com) and Oervondstchecker.nl¹ (Eijkelboom, Ahrens & Schouwenburg, 2024). However, these different data often vary in quality metrics such as fossil quality, image quality and identification quality (Table 1). With standardised digitisation protocols, image quality can be improved. However, photographing many thousands of fossils is a laborious task and relative data sparsity in palaeontology poses

¹ Translated as 'Prehistoric-find checker'.



Figure 1 Typical fragmented fossil bones from sand-nourished beaches in the Netherlands. Identification of individual bone fragments requires a high level of expertise. In many cases too few diagnostic characters are present to specifically identify a fossil. Scale bar is 180 mm. Photograph by I. Eijkelboom.

Full-size  DOI: 10.7717/peerj-cs.4049/fig-1

Table 1 Data quality metrics for a training dataset of a fossil-ID model.	
Quality metric	Examples
Fossil quality	Fragmentation, abrasion, weathering
Image quality	Lighting, sharpness, number of and relevancy of angles photographed, contrast with background, scale present
Identification quality	Human biases, amount of validators and cross-validation, identifying physical objects or based on pictures

additional challenges for computer vision compared to other fields such as ecology and medicine. CS data may contribute substantially, but can also introduce noise (*e.g.*, non-fossils such as rocks or driftwood) and variety in quality (*e.g.*, blurry or underexposed images, or images taken from non-standardised angles) into the dataset. Regardless of image quality, all datasets have varying degrees of fossil quality (*e.g.*, completeness, varying taphonomic characteristics) and identification quality as a result of *e.g.*, the previously mentioned biases which also carry over to AI model predictions (*Koch et al., 2023*). Fossil data labels, as well as labels from other natural history collections, should therefore not be regarded as an absolute ground truth, but rather a probabilistic ground truth. Setting up data validation protocols can improve labelling quality.

Over the past 150 years, millions of Quaternary fossil bones, teeth, shells and artefacts have been collected from the North Sea and sand-nourished beaches in the Netherlands by fishermen, citizen scientists and enthusiastic collectors alike (*Mol et al., 2006; Moerdijk et al., 2010; Kuitens et al., 2015; Mol, 2016; Amkreutz & Van der Vaart-Verschoof, 2021; Mol & Bakker, 2022*). These finds are housed in both museum and private collections and are instrumental for our understanding of Quaternary landscape and ecosystem evolution of north-western Europe. Since 2014, thousands of citizen scientists and fossil enthusiasts

have amassed over 23,000 fossil observations with images, coordinates, descriptions and identifications by volunteer experts on the online platform Oervondstchecker.nl (Eijkelboom, Ahrens & Schouwenburg, 2024). Digitised collection photographs from various museum and private collections and photographs from the citizen science platform Oervondstchecker.nl are used here as training data for ML models. The two datasets differ in data quantity and quality (see below) and are labelled in different ways which allows for a comparison of these variables.

In this article we aim to improve and accelerate fossil identifications from Dutch beach nourishments and elsewhere through a CS-ML approach, using both a standardised dataset and a CS dataset. We investigate how dataset size and different quality metrics (Table 1) influence AI model performance. Besides standard performance metrics such as accuracy, precision and recall, we compare model performance with identifications by ten domain experts and active citizen scientists. Based on these insights we make recommendations on how to account for variable data quality to optimise AI model performance and rapidly increase the amount of high-quality fossil identifications. Finally, we made the AI models publicly available (<https://museum.identify.biodiversityanalysis.nl/>) as a pilot tool for enthusiasts to help identify their finds.

METHODS

CNN model training and evaluation

The training dataset consists of two major parts: (1) standardised images of fossils and artefacts from Naturalis Biodiversity Center, the Natural History Museum Rotterdam, the Dutch National Museum of Antiquities, and various CS collections (>46,000 images), and (2) images from the online CS platform Oervondstchecker.nl (>74,000 images) (Fig. 2). The standardised image dataset consists of Quaternary fossil bones and teeth (mainly mammalian), and Palaeolithic/Mesolithic artefacts. To represent the known Quaternary mammalian biodiversity of the Netherlands and the southern North Sea, both fossils from very common taxa such as the woolly mammoth, steppe bison and wild horse, and fossils from less common taxa such as various carnivores, marine mammals and humans were photographed. The Oervondstchecker.nl dataset consists of all reported finds (Data S1). It mainly contains mammalian fossil remains (common and rare taxa) and artefacts, but also contains fossil shells, wood and unworked rocks. A complete list of the taxa and osteological parts represented in both datasets can be found in Table S2. The standardised images were taken with a photography setup that can be easily transported to different collections (Fig. S3). Eight photographs were taken from six standardised positions and two random positions by placing fossils at an angle. A standardised naming convention was used for automatic pre-processing of the data. Fossil identifications from Oervondstchecker.nl were done in open text format and the data had to be pre-processed before it was suitable training data for ML models. To achieve standardised image labels, a keyword script was used to categorise the ~23,000 observations (Data S4). The automatically assigned categories were then manually validated or corrected. All images received both an osteological/anatomical label (hereafter referred to as the ‘Anatomy’ label) and a taxonomic label (hereafter referred to as the ‘Taxonomy’ label) (Fig. 2).

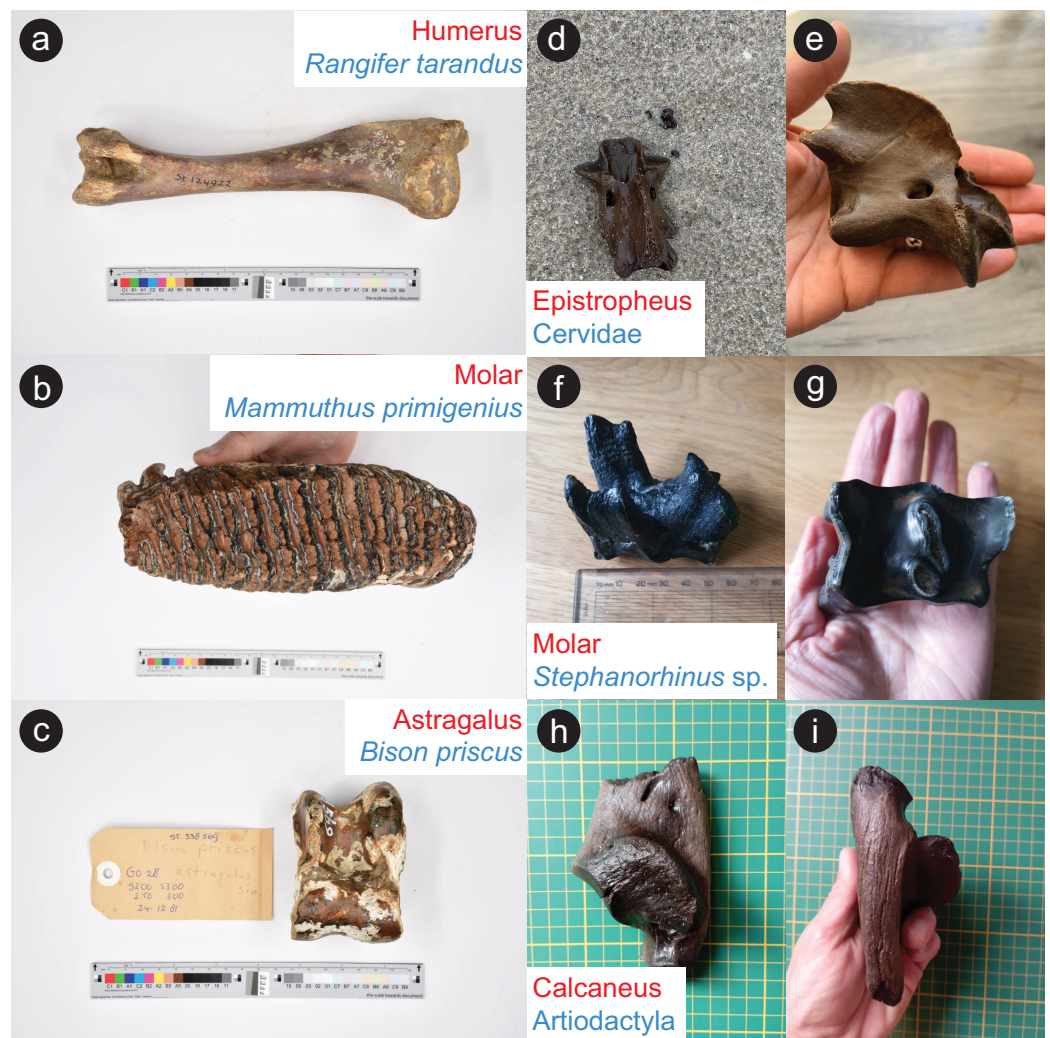


Figure 2 Examples from the standardised collection dataset (A–C) and the Oervondstchecker.nl dataset (D–I). (A) RGM.124922, (B) RGM.400648, (C) RGM.338569, (D–I) uploaded to Oervondstchecker.nl by user Renate Wolthuis (D, E), Anna (F, G), Robin Nietvelt (H, I). In red the ‘Anatomy’ label, in blue the ‘Taxonomy’ label. (A–C) photographed by I. Eijkelboom and F. ter Haar.

Full-size [DOI: 10.7717/peerj-cs.4049/fig-2](https://doi.org/10.7717/peerj-cs.4049/fig-2)

To train, validate and test the ML models, the datasets were split into three parts with each part representing all classes with at least three instances (Data S5). Classes with fewer than three instances were discarded. A 10% test set was kept aside to test the model performance after the training process was completed. The remaining 90% of the data was split into a 90% training set and 10% validation set. Images from eligible classes were split across the training, validation and test sets using a stratified split. This commonly used heuristic dataset split provides sufficient data for model validation and testing, while retaining a large portion of the dataset for training. During each training epoch, the model ‘learns’ from the training set and adjusts its parameters through backpropagation (Lecun, Bengio & Hinton, 2015; Schmidhuber, 2015). After each epoch, the model is tested on the validation set to detect problems like overfitting. Aside from the two separate test sets for

each of the datasets, a third, mixed test set was made (hereafter referred to as the ‘Split data’ test set). It principally consists of images from the Oervondstchecker.nl dataset. Only for those classes that are not represented in the citizen science dataset, standardised images from museum and private collections were included. The ‘Split data’ test set gives the best representation of the type of data we want a publicly available model to perform well on.

Because fossils and fossil image data differ significantly from modern wildlife and biodiversity image data, we developed a data augmentation protocol suited for our dataset (Data S6). Whereas biodiversity image data are typically photographed outdoors with natural lighting and background, our fossil image data are typically captured inside with artificial lighting and a downward-facing camera towards the floor or table. Moreover, in the case of fossils we typically only deal with (fragmented) parts of organisms instead of the complete organisms. Through transportation (*e.g.*, by rivers or construction work) skeletons are disarticulated and bones are broken or eroded. We explored several custom augmentation techniques to reduce generalisation error but found the positive effects to be minimal to insignificant, while these augmentations increased the training time significantly (Fig. S7). We have not further pursued these custom augmentations here.

The CNNs presented here were trained and tested using the Naturalis AI pipeline (Fig. 3; <https://gitlab.com/naturalis/bii/ai/naturalis-ai>). The training pipeline contains a collection of tools and scripts designed to train, evaluate and deploy CNNs based on an ‘image file’ and a ‘taxon file’ (Table S2, Data S8). A set of scripts was used to automatically generate the standardised image and taxon files (Data S9). As the pipeline is already configured for taxonomic classification, we used the recommended, partially pretrained base model, EfficientNetV2-M (Tan & Le, 2021), and most default model settings to train our CNNs. Two sets of models were trained; on the ‘Anatomy’ labels and on the ‘Taxonomy’ labels. The ML model performance was evaluated by calculating commonly used metrics:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

where TP = true positive; TN = true negative; FP = false positive; FN = false negative; F1 = harmonic mean of the precision and recall. Moreover, a set of confusion matrices was generated, all within the Naturalis AI pipeline. The confusion matrices visualise which individual classes perform well, and which classes are often confused for one another (*i.e.*, the false positives and false negatives).

Expert identifications

To contextualise the CNN results and discuss potential applications, we compared the model performance with domain expert identifications of a dedicated set of objects and

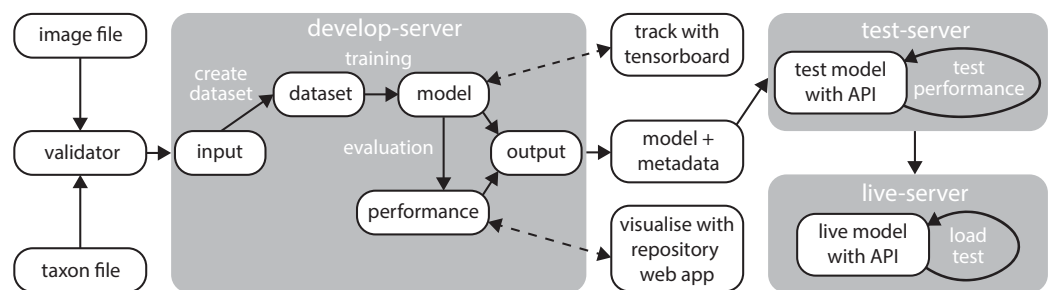


Figure 3 Schematic workflow Naturalis AI pipeline. The image and taxon files are automatically validated (checked for textual errors) and imported onto the develop-server. Here, a dataset is created and parameters are set to train the model. Training progress can be monitored with tensorboard and model performance is evaluated and visualised with a repository web app. The trained model and associated metadata are exported and the model is tested on the test-server. If tests are successful, the model is moved to the live-server where it is available for online use.

Full-size DOI: [10.7717/peerj-cs.4049/fig-3](https://doi.org/10.7717/peerj-cs.4049/fig-3)

images. Ten experts (*i.e.*, available domain experts and expert citizen scientists with different fields of expertise and varying levels of expertise. All regularly identify CS finds online) were asked to identify three curated datasets: (1) 20 physical objects from the collections of Naturalis Biodiversity Center (NBC), (2) 20 photographed objects from the NBC collections in a standardised setup and (3) 30 photographed objects from the online CS platform Oervondstchecker.nl (OVC). The 20 physical objects that were classified by the experts, were photographed in the same standardised way for classification by the ML models. These 70 objects were selected to represent a variety of taxa (common/rare; ambiguous/unequivocal) and states of preservation (completeness; diagnostic features present). Moreover, the CS images were also selected based on the image quality (Table 1; lighting, sharpness, number of and relevancy of angles photographed, contrast with background, scale present).

To compare expert identifications, we calculated the inter-annotator agreement, here expressed using Fleiss' Kappa (κ), a standard tool in content analysis (*e.g.*, Drisko & Maschi, 2016). The classifications were translated to numerical ratings after which we calculated κ using the R package *irr* (Gamer et al., 2019) and calculated 95% confidence intervals (CI) based on 1,000 bootstrap iterations using the R package *boot* (Canty, Ripley & Brazzale (2024) after Gooliaff & Hodges (2018). When there is complete agreement among experts $\kappa = 1$; when there is complete disagreement $\kappa = -1$; and when agreement follows what can be expected by chance $\kappa = 0$. Our interpretation of κ was based on Landis & Koch (1977) which is widely used to interpret κ . Note however that Landis & Koch (1977) used a 2-annotator 2-class example. In this study we have ten annotators and a relatively complex classification task with over 150 and 450 plausible classes for the 'Anatomy' and 'Taxonomy' models respectively (Table S2).

Furthermore, comparison of correctness of identifications among experts and between experts and the ML models was expressed using the 'taxon distance' (Hogeweg et al., 2024; Fig. 4), an assessment system involving penalties for incongruent hierarchical identifications. This way, the hierarchical nature of taxonomy is taken into account.

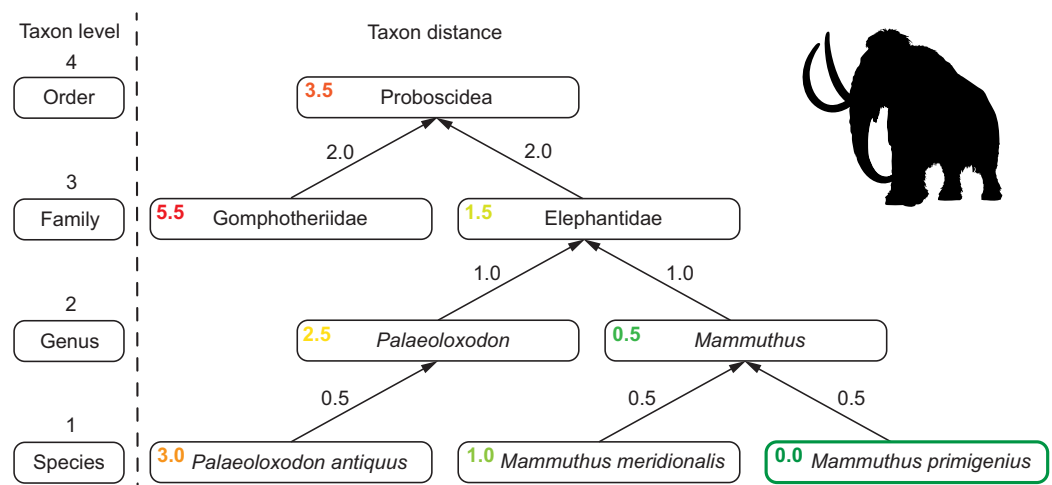


Figure 4 Example of the 'taxon distance' and the 'taxon level'. Here, a correct classification (*Mammuthus primigenius*) yields 0 (penalty) points for the 'taxon distance'. The further a classification is removed from the correct answer in the taxonomic tree, the higher the penalty is. Additionally, the correct classification yields one point for the 'taxon level', as it is the most specific classification (species level).

Full-size DOI: 10.7717/peerj-cs.4049/fig-4

A higher penalty is given as the prediction and the correct label are further removed from each other in the taxonomic tree. Since no absolute ground truth classifications exist, comparisons were made between the classifications by the CNNs and experts based on both the collection labels (that are expert opinions as well, further discussed below) and the consensus vote among the experts. The consensus vote was defined as the most specific taxonomic classification that reached a $\geq 50\%$ vote. In case of a tie, the collection label was used as a tiebreaker. We also defined the 'taxon level' as a value for specificity of taxonomic classifications (Fig. 4). To calculate the significance between expert and AI predictions based on (a subset of) the 70 classified objects, we performed the Wilcoxon Signed-Rank Test (Wilcoxon, 1945). This statistical test is applied instead of a paired t-test when the data distribution deviates significantly from normality, which we calculated with the Shapiro-Wilk Test (Shapiro & Wilk, 1965). Both tests were performed in Python (Data S10).

Finally, we calculated the mean probability that the majority classification of a randomly selected subset of experts matched (1) the consensus classification of the experts or (2) the collection label class in R (Data S11) after Gooliaff & Hodges (2018). Neither consensus classifications nor collection labels are absolute ground truths, so we compare individual expert identifications to both of these probabilistic ground truths. This mean probability (hereafter referred to as the 'majority vote test') expresses the increase in accurate or correct identifications when consulting multiple experts rather than just one expert. The 'majority vote test' provides a sense of the usefulness or necessity of multiple identifications for a single fossil specimen to improve data quality. To see at which taxonomic level the expert opinions start to deviate most, the 'majority vote test' was calculated at class, order, family, genus and species level.

Table 2 Evaluation metrics for the CNN models. Accuracy (for top 1 and top 3 predictions), recall, precision and F1 score (see Methods for definitions) provided in percentages for the ‘Anatomy’ and ‘Taxonomy’ models, and for different Train/Val and Test datasets. OVC data: Oervondstchecker.nl dataset; Coll data: standardised museum and private collections dataset; Split data: 10% test split dataset (see Methods). In bold the highest value for each metric and each test dataset.

Train/Val dataset	Test dataset	Top 1 accuracy (%)	Top 3 accuracy (%)	Top 1 recall (%)	Top 1 precision (%)	F1 score (%)
ANATOMY						
OVC data	OVC data	63.3	84.1	46.8	35.2	40.2
Coll data		25.2	43.9	9.2	7.5	8.3
All data		63.7	84.6	43.4	33.1	37.6
OVC data	Coll data	39.1	57.1	32.7	25.7	28.8
Coll data		85.3	91.6	83.0	77.4	80.1
All data		88.0	94.4	82.0	81.2	81.6
OVC data	Split data	62.9	82.0	39.3	28.3	32.9
Coll data		25.7	43.3	17.9	16.4	17.1
All data		63.7	85.2	46.0	34.8	39.6
TAXONOMY						
OVC data	OVC data	57.3	79.2	31.9	27.2	29.4
Coll data		23.5	45.2	6.0	3.7	4.6
All data		58.0	81.2	33.2	27.3	30.0
OVC data	Coll data	26.5	37.6	8.4	7.5	7.9
Coll data		88.4	94.0	68.7	66.4	67.5
All data		89.8	95.6	70.5	69.2	69.8
OVC data	Split data	56.6	77.2	22.6	18.9	20.6
Coll data		22.5	43.6	20.0	17.6	18.7
All data		57.0	80.5	39.3	32.8	35.8

RESULTS

CNN model performance

A comparison of CNN model performance with different training data illustrates the systematic variation in data contents, quality and quantity (Table 2; Data S12). Moreover, this data ablation approach quantifies the sensitivity of model performance to domain shifts and provides an estimate of the resulting generalisation error. The best performing ‘Anatomy’ and ‘Taxonomy’ models are trained on all data (‘All data’ in Table 2) and tested on the standardised collection dataset (‘Coll data’ in Table 2). With a top-1 accuracy of 88–90%, the models score comparable to experts (discussed below). If provided with standardised data that fall within the dataset distribution, these models can be used in museum collections and by collectors to identify their finds. Also, for the models tested on the Oervondstchecker.nl dataset (‘OVC data’ in Table 2) or the ‘Split data’, we see a general pattern of models trained on all data outperforming the models that are trained on either of the two datasets. Because the ‘Split data’ best represent what input data can be expected on a CS web app (discussed above), the ‘Anatomy’ and ‘Taxonomy’ models trained on all data and tested on the ‘Split data’ were made publicly available (see below) and are hereafter referred to as the ‘public models’. These two models are also further evaluated below and used in the comparison to domain expert classifications.

Comparing CNN model performance for individual anatomical classes and taxa illustrates strong and weak points within the model. False positives and negatives highlight classes that require additional high-quality data for model predictions to be more accurate. Anatomical and taxonomic groups which are most visually similar are typically also the most confused groups and cause incorrect model predictions. Confusion matrices focused on groups that are anatomically or taxonomically similar or more closely related show this relationship (Fig. 5). For both Cervidae (Fig. 5A) and Bovidae (Fig. 5B), fossils with a ‘Taxonomy’ label at the species level are confused mostly with more general taxonomic groups, rather than with other species from the same family (misclassifications species within Cervidae: 77 with more general taxa, 14 with other species; misclassifications species within Bovidae: 46 with more general taxa, three with other species). The same pattern is observed when evaluating model performance for specific vertebrae (Fig. 5C) and long bones (Fig. 5D) from the ‘Anatomy’ model (misclassifications specific vertebrae: 69 with more general groups, 22 with other specific vertebrae; misclassifications specific long bones: 140 with unidentifiable bones, 59 with other specific long bones).

A taxonomic evaluation focused on more general taxonomic groups also shows a high degree of confusion (Fig. 6). Similarly to the confusion matrices in Fig. 5, mammalian orders (Artiodactyla-Rodentia) are more confused with the general class Mammalia than with other mammalian orders (misclassifications mammalian orders: 488 with class Mammalia, 221 with other mammalian orders). The ‘Split data’ test set mainly contains images from the Oervondstchecker.nl dataset, that are often lower in image and fossil quality (Table 1). As a result, these often-fragmented fossil finds receive a more general ‘Taxonomy’ label and the CNN confuses these fragmentary finds. See Data S13 for a complete confusion matrix for each of the trained models.

Expert agreement and comparison CNNs

The experts scored a fair (species level) to substantial/almost perfect (anatomy and class level) agreement (Table 3). Fossil identifications by experts are far from unanimous. Often different identifications can be related to differences in expert confidence, *e.g.*, one expert identifying a fossil as part of a family and another expert identifying the same fossil as part of a species within that family. Predictably, agreement among experts increases on more general taxonomic levels.

The ‘taxon distance’ results of the classifications by experts show large differences between the three different datasets and between collection labels and consensus votes (Table 4, Table S14). Overall, experts performed better on the photo dataset from the NBC collection compared to the physical NBC collection. Furthermore, the individual experts performed better relative to the consensus vote than relative to the collection labels. When comparing consensus votes, on average the experts performed slightly worse on the photo dataset from the OVC than on the physical NBC collection. For the physical NBC collection dataset, the consensus vote outperformed any individual expert, whereas for the photo dataset from the NBC collection, three individual experts outperformed the consensus vote. Individual experts might perform well relative to the collection labels, while performing poorly relative to the consensus vote and *vice versa*.

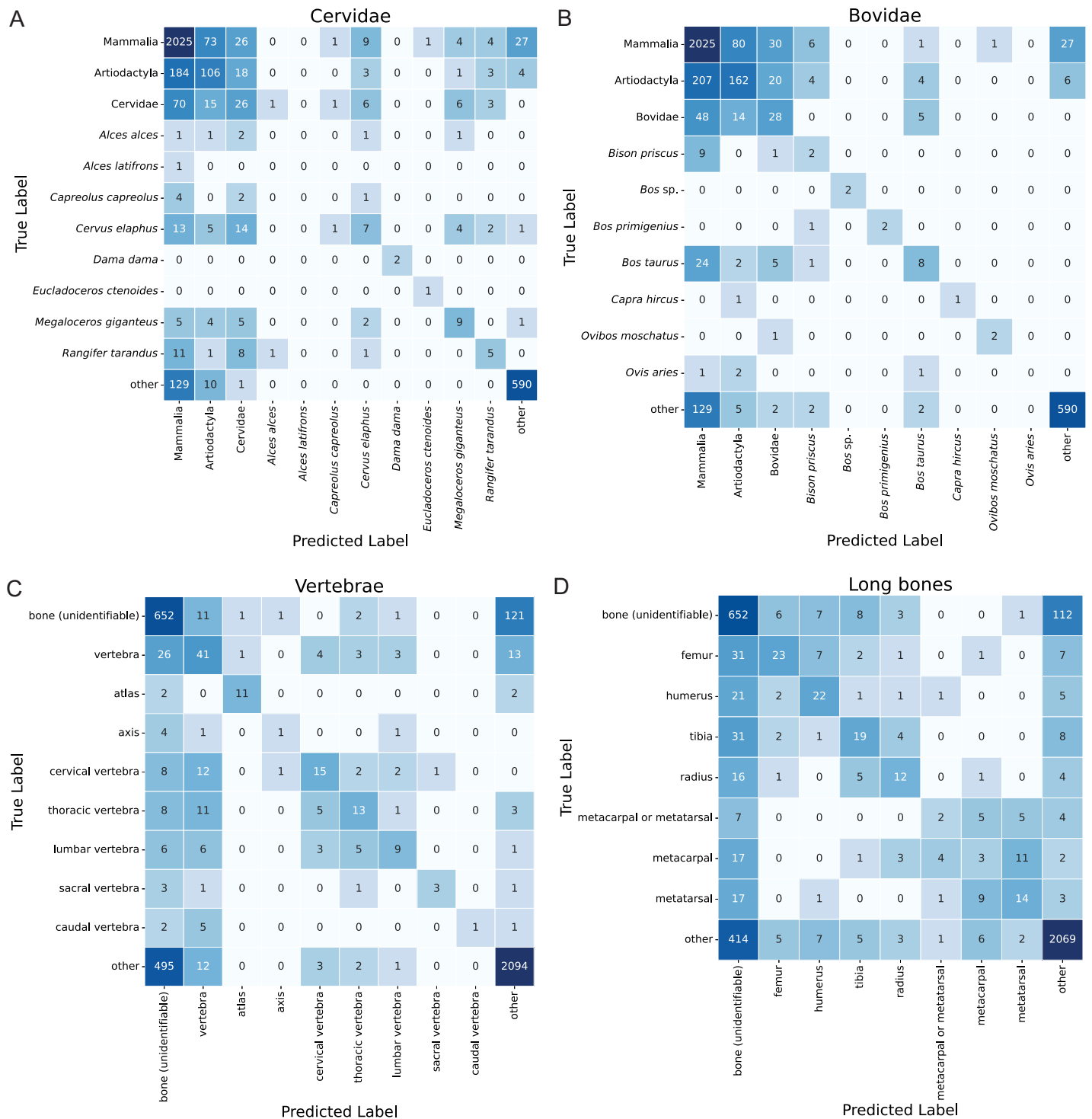


Figure 5 Confusion matrices for the public models focussed on groups that are taxonomically more closely related or morphologically similar. The confusion matrices visualise which individual classes perform well, and which classes are often confused for one another (*i.e.*, the false positives and false negatives). The rows show the true label of an image and the columns show the predicted label of an image. Correct classifications appear on the diagonal line from top left to bottom right. (A) Cervidae; (B) Bovidae; (C) vertebrae; (D) long bones.

Full-size DOI: 10.7717/peerj-cs.4049/fig-5

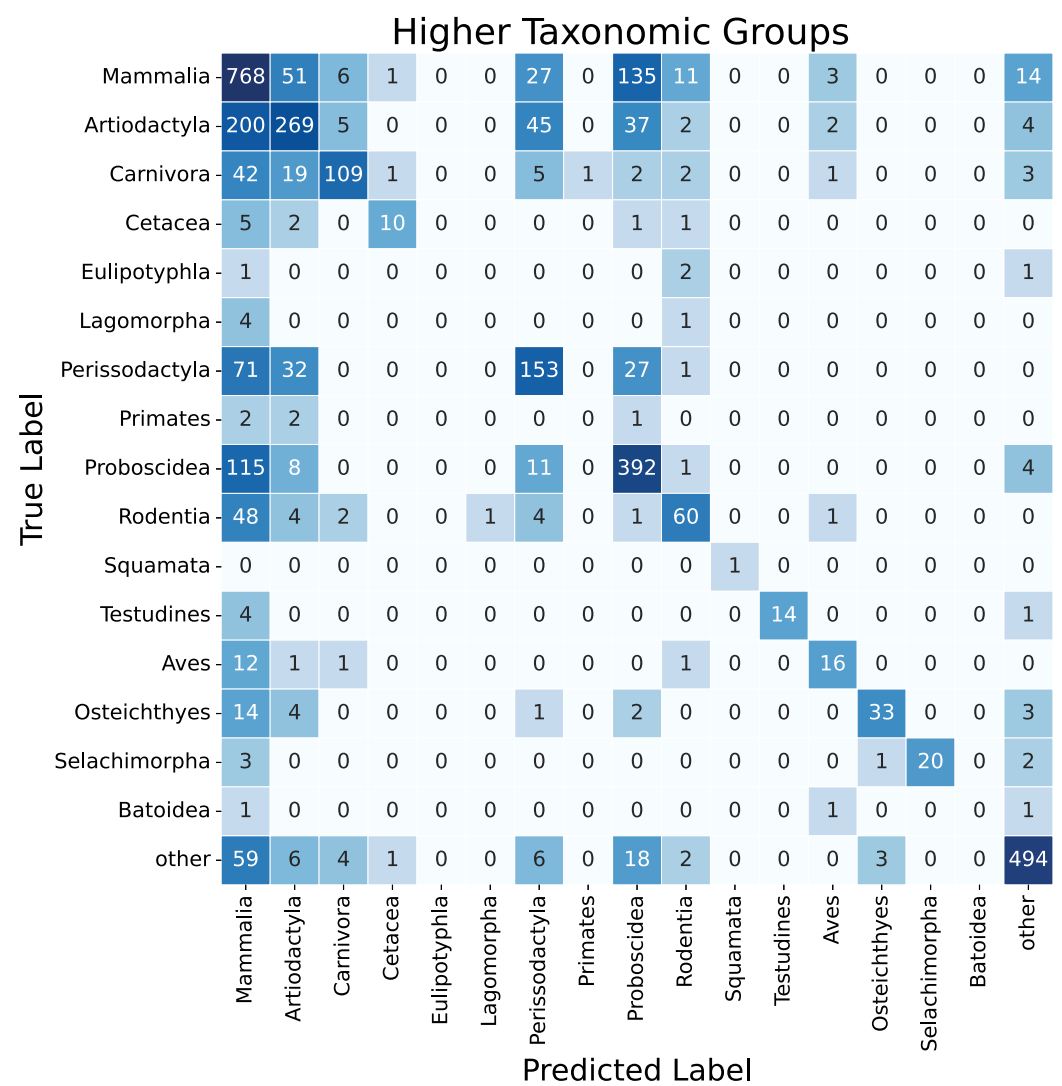


Figure 6 Confusion matrix for the public taxonomy model focussed on higher taxonomical groups (within Chordata). All images with a ‘Taxonomy’ label on a family, genus or species level are grouped in the order of said taxon. The rows show the true label of an image and the columns show the predicted label of an image. Correct classifications appear on the diagonal line from top left to bottom right.

Full-size DOI: 10.7717/peerj-cs.4049/fig-6

Table 3 The inter-annotator agreement for the ten experts. Fleiss’ Kappa (κ) and 95% confidence intervals (CI) for the anatomy and taxonomy at different taxonomic levels. *See Methods for context of interpretation κ .

	Fleiss’ Kappa (95% CI)	Interpretation*
Anatomy	0.76 [0.70–0.83]	Substantial-almost perfect agreement
Species	0.33 [0.28–0.40]	Fair agreement
Genus	0.40 [0.34–0.48]	Fair-moderate agreement
Family	0.53 [0.46–0.62]	Moderate-substantial agreement
Order	0.68 [0.61–0.77]	Substantial agreement
Class	0.79 [0.64–1.00]	Substantial-almost perfect agreement

Table 4 Total ‘taxon distance’ per category, relative to the collection label and the expert consensus vote, for each of the ten experts and for the AI models trained on either or both datasets. Low ‘taxon distance’ indicates better performance. The ‘Photos OVC’ column was normalised, because this dataset contained more objects than the other two (see Methods). AI model ‘taxon distance’ in *italic*: higher than all 10 experts; underlined: within the range of the 10 experts; in **bold**: lower than all 10 experts.

	Physical NBC collection		Photos NBC collection		Photos OVC	
	Label	Consensus	Label	Consensus	Label	Consensus
Expert 1	67.0	39.5	43.0	22.5	x	39.6
Expert 2	40.5	33.0	54.0	16.0	x	60.0
Expert 3	59.0	34.5	50.5	21.5	x	37.0
Expert 4	92.0	66.5	69.5	31.5	x	40.0
Expert 5	40.5	44.0	37.5	31.0	x	47.8
Expert 6	50.5	27.5	38.5	43.5	x	38.5
Expert 7	62.5	39.0	49.0	32.0	x	35.2
Expert 8	54.5	32.5	43.5	27.5	x	30.4
Expert 9	55.0	42.5	39.0	28.5	x	74.8
Expert 10	79.0	73.5	120.5	80.0	x	43.3
AVERAGE	60.1	43.3	54.5	33.4	x	44.7
Consensus	29.5	x	41.5	x	x	x
AI (OVC data)	204.0	178.5	201.5	161.0	x	107.8
AI (Coll data)	166.5	144.0	30.0	<u>50.0</u>	x	188.1
AI (All data)	165.0	141.5	<u>45.0</u>	<u>65.0</u>	x	11.1

All experts outperformed the CNNs when identifying physical objects from the NBC collection or photos from the CS dataset (Table 4). Overall, the experts outperformed the CNNs with significant Wilcoxon Tests ($p < 0.05$; Table S15). Only for the ‘photos NBC collection’ dataset, the CNNs trained on the standardised collection data outperformed some or even all experts (Table 4), resulting in two non-significant ($p > 0.05$) and one significant Wilcoxon Test in favour of the CNN (Table S15). Finally, both the expert identifications and AI model predictions for the ‘photo NBC collection’ have a (much) lower ‘taxon distance’ than the ‘NBC collection’, suggesting the latter contained fossils more difficult to identify.

The specificity of an identification, here expressed as the ‘taxon level’ (Fig. 4), gives an indication of data quality (Table 1) because, in general, more complete and well-photographed fossils can be identified more specifically. On average, the experts identified the physical NBC collection most specifically (lowest total ‘taxon level’) and the OVC photos most generally (highest total ‘taxon level’) (Table 5, Table S14). Similarly, the AI models identified standardised images more specifically than the OVC images. These total ‘taxon levels’ correlate well to the variety in data quality between the different datasets. Moreover, the three best performing experts (lowest ‘taxon distance’, Table 4) provide more specific identifications (lower ‘taxon level’, Table 5) than the three poorest performing experts when compared to the collection labels. On the contrary, when compared to the consensus vote, the best performing experts provide similar or more

Table 5 Total ‘taxon level’ per category for each of the ten experts and the AI models trained on either or both datasets. A higher value indicates more general identifications. The ‘Photos OVC’ column was normalised, because this dataset contained more objects than the other two (see Methods). AI model ‘taxon level’ in *italic*: within the range of the 10 experts; in **bold**: lower than all 10 experts.

	Physical NBC collection	Photos NBC collection	Photos OVC
Expert 1	36	43	58
Expert 2	51	54	76
Expert 3	30	41	56
Expert 4	65	68	83
Expert 5	22	29	55
Expert 6	38	39	59
Expert 7	38	41	74
Expert 8	37	53	56
Expert 9	31	36	43
Expert 10	43	40	65
AVG exp	39	44	63
Consensus	47	55	72
AI (OVC data)	45	40	58
AI (Coll data)	24	21	30
AI (All data)	22	21	67
AVG AI	30	27	52

general identifications (same or higher ‘taxon level’, Table 5). Finally, the consensus vote has a higher ‘taxon level’ than the average of the individual experts for all three datasets. As a higher ‘taxon level’ generally corresponds to poorer data quality, the consensus vote might not necessarily be the best identification method (discussed below).

Relative to the experts, the AI models give more specific (but not necessarily more correct) identifications overall. Similar to the experts, the AI models are more general in the identification of fossils from CS data, which can be explained by the more general labels in this training dataset. Therefore, the AI model that was only trained on CS data also gives more general predictions for the standardised data from the test set. The opposite is true as well: the AI model trained only on standardised data gives more specific predictions for the CS data from the test set. Note however that the correlation between the ‘taxon distance’ and ‘taxon level’ discussed above is not consistent for the AI models, because they are overfitted in many cases and therefore may provide very confident or specific, yet wrong, predictions (discussed below). In a few exceptional cases, this pattern of overfitting is also observed for some experts (e.g., expert 9 for the ‘photos OVC’ dataset). Methods to mitigate overfitting are discussed below.

The ‘majority vote test’ for the consensus vote shows a positive effect of a majority vote (Fig. 7A). However, this is to be expected: as more expert ‘votes’ are considered, the majority vote converges towards the consensus vote of all experts. The ‘majority vote test’ for the collection labels shows no positive effect of a majority vote (Fig. 7B). The mean probability does not increase with more experts identifying a fossil find. This is true for

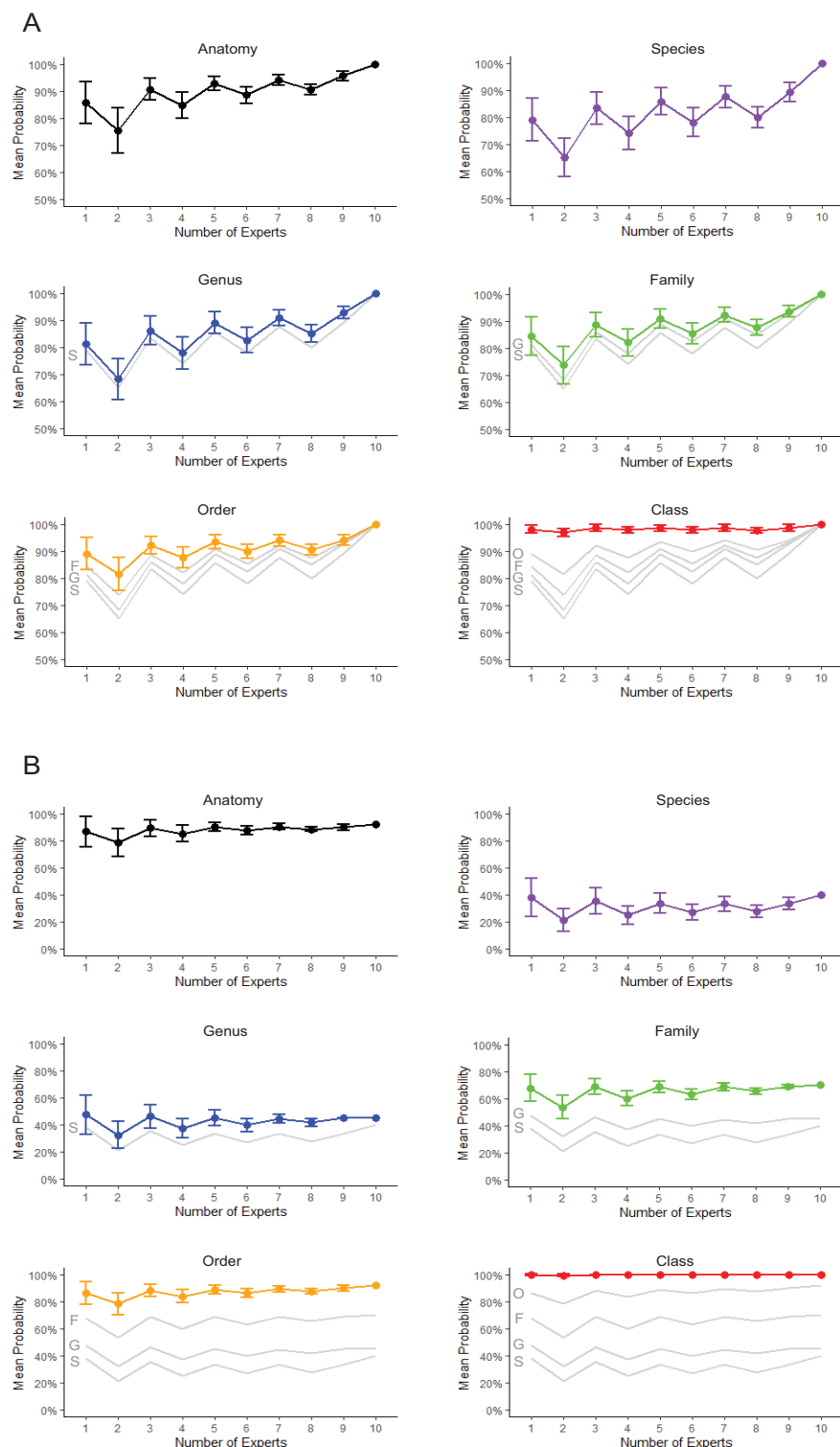


Figure 7 The ‘majority vote test’ or mean probability of a correct identification by a random subset of experts relative to (A) the consensus vote of all the experts and (B) the collection labels. Bars indicate 95% confidence intervals. In light grey the mean probabilities of more specific taxonomic levels are shown. Even numbers of experts score lower for the ‘majority vote test’ because of the likelihood of drawing a split vote, which is not possible for odd numbers of experts (Gooliaff & Hodges, 2018). Note that the graphs in (A) only depict 50–100% on the y-axis.

Full-size DOI: 10.7717/peerj-cs.4049/fig-7

both the ‘Anatomy’ and ‘Taxonomy’ identifications, across taxonomic levels. Like the correlation between ‘taxon distance’ and ‘taxon level’, the label data quality of these fossil datasets does not improve with a majority vote.

DISCUSSION

Data quantity

The performance of ML models such as the ones presented here depends strongly on both the quantity and the quality of the dataset used to train the model. Compared to general computer vision datasets (e.g., ImageNet-21K, >14M images; [Ridnik et al., 2021](#)) and datasets used to train modern species recognition models (e.g., ObsIdentify, >40M images (data L. E. Hogeweg)) the dataset used here is small. We are a long way from reaching such dataset sizes, but continued digitisation of collections and the constant flow of new CS data allow us to iteratively increase model performance. Combining data from different sources does improve model performance ([Table 2](#)), suggesting that the models do learn shared features across datasets. By identifying gaps in the training datasets and weak areas in the model performance, specific groups of fossils can be photographed for further model performance gains. Also, continuous advances in computer vision model architectures allow us to increase model performance with relatively small datasets such as ours (e.g., [Chai et al., 2021](#)).

Data quality

Data quality varies strongly within our training dataset ([Table 6](#), [Table S16](#)). First, the standardised image dataset from museum and private collections has consistently high image quality ([Table 1](#)). On the contrary, the quality of the images in the CS dataset varies hugely. For future data collection, citizen scientists could get additional instructions on how to best photograph fossils specifically for ML model identifications ([Eijkelboom, Ahrens & Schouwenburg, 2024](#)). Second, fossil specimens from the standardised image dataset are generally relatively complete and well preserved (high fossil quality, [Table 1](#)). This enables the CNN model to extract meaningful features for a classification. Whereas fossils in collections are typically selected based on their quality, the fossil quality of the CS dataset is again highly variable. The inclusion of more (artificially or digitally) fragmented material from museum and CS collections in our training dataset could help increase model performance on the fragmented material from Oervondstchecker.nl. Third, the collection material was identified and labelled based on the physical objects themselves, whereas the CS dataset was identified based on the provided images and description (identification quality, [Table 1](#)). Improving the identification protocol with e.g., experts cross-validating each other’s identifications or AI pre-selection of fossils for experts in different subcategories, could further improve model performance.

The model predictions from the same 100 CS data examples from [Table 6](#) were analysed to see whether the misclassifications were due to a prediction that was more general/specific, but within the correct taxonomic hierarchy, or due to a prediction from a different taxonomic branch ([Table S16](#)). Out of the 100 examples 55 were classified correctly and 45 were misclassified. However, 35 misclassifications were within the same taxonomic branch

Table 6 Data quality metrics for 100 CS data examples from the test set of the public taxonomy model. Provides a measure of data quality of our CS dataset and helps guide efforts to increase the data quality. See Table S16 for a complete overview of the 100 test set examples.

Category	Quality metric	Very poor	Poor	Intermediate	Good	Very good
Fossil quality	Fragmentation	0–24%	25–49%	50–74%	75–90%	91–100%
		complete (50)	complete (14)	complete (9)	complete (13)	complete (14)
	Abrasion	76–100%	51–75%	26–50%	10–25%	0–9%
		surface abraded (12)	surface abraded (12)	surface abraded (19)	surface abraded (35)	surface abraded (22)
	Weathering	76–100%	51–75%	26–50%	10–25%	0–9%
		surface weathered (0)	surface weathered (0)	surface weathered (3)	surface weathered (24)	surface weathered (73)
Image quality	Lighting	Very over/under lit, no outline visible (0)	Over/under lit, no details visible (12)	Somewhat, over/under lit, some details visible (51)	Fairly well lit, most details visible (30)	Very well lit, all details visible (7)
	Sharpness	Very blurry, no outline visible (0)	Blurry, no details visible (4)	Somewhat blurry, some details visible (32)	Sharp, most details visible (61)	Very sharp, all details visible (3)
	Number of and relevancy of angles photographed	One image, not relevant (1)	1–2 images, somewhat relevant (22)	3+ images, somewhat relevant (35)	1–2 images, relevant angles (9)	3+ images, relevant angles (33)
	Contrast with background	Same/similar colour (8)		Somewhat-contrasting colour (23)		Well-contrasting colour (69)
	Scale present	No scale (15)	Approximate scale (e.g., hand, background) (16)		Exact scale (e.g., coin) (6)	Exact scale (e.g., ruler, gritted cutting mat) (63)

(22 more general, 13 more specific predictions) and only 10 misclassifications belonged to a different branch. Out of the 22 more general predictions 18 were identified as the general and uninformative ‘Mammalia’ and the remaining four were identified as ‘Cervidae’. Out of the 13 more specific predictions nine were at the species level with six being identified as ‘*Mammuthus primigenius*’. Out of the 10 misclassifications belonging to a different branch seven were at the species level. However, the last common group between the true labels and the predicted labels was at the order, class or phylum level for 9/10 misclassifications, indicating that these misclassifications were largely completely wrong and yield high ‘taxon distance’ penalties. Common taxa labels such as ‘Mammalia’ and ‘*Mammuthus primigenius*’ were overrepresented in the wrong predictions, similarly to what can be observed for the complete ‘Split data’ test set in the confusion matrices (Fig. 5).

Based on (1) the difference between model performance metrics (accuracy, precision, recall) for the training dataset and the test dataset and (2) the discrepancy between model performance metrics and the confidence of the model predictions from the test set, the public models appear to generalise poorly to data outside the training dataset (overfitting). Methods of improving data quantity and quality discussed elsewhere in this article should reduce the overfitting of the models. Moreover, ML methods such as further regularisation and data augmentations, and early stopping could be explored and integrated in the

training of future model iterations (Bejani & Ghatee, 2021; Santos & Papa, 2022). Although our efforts yielded no substantial results, continued exploration of domain-specific data augmentation techniques can also help to mitigate overfitting. Finally, further data ablation strategies (e.g., by data quality metrics (Table 1)) can be used to explore the effect on model performance.

Like the consensus vote, the collection labels often have a varying identification quality as a result of human biases involved in the identification process. In some cases, the labels date from decades ago and may be outdated. To ensure the most up-to-date identification for a specimen, it is best to check the online repository if collections were digitised and if there have been updates to the identifications. Also, in some cases the collection piece may have been identified by someone who was not an expert on the specific group of fossils, as museum collections often house a wide variety of natural history objects and do not have a dedicated curator for each group. This may also have been the case in our dataset. For example, one fossil from the expert identifications was labelled as a metatarsal of the moose *Alces alces*, but 8/10 experts identified it as the metatarsal of the giant deer *Megaloceros giganteus* (and none as *Alces alces*). Also, a radius of a seal was labelled at the family level (Phocidae), but a marine mammal expert was confident in identifying the fossil to the species level as a grey seal *Halichoerus grypus*. To improve the data labels on which the models are trained, we need to specifically address the biases or blind spots experts might have by exploring peer verification in the identification process. AI models could provide a second opinion or preselect fossil finds on a more general taxonomic level to send to the expert with the right specialisation. Moreover, in this study the experts were explicitly asked to identify the fossils the same way they do on a daily or weekly basis at home on CS platforms or social media. Some of the experts consulted relevant literature, but most of the identifications were done based on memorised knowledge. The most thorough, tried-and-tested method of identifying a fossil is taking exact measurements, comparing those to measurements from the literature and taking the fossil to a comparative collection at a museum (Von Den Driesch, 1976). When publishing about a specific fossil, this is typically done to ensure the best identification possible. However, this takes time and cannot be done for every fossil identification, especially when basing the identification on images from a CS platform or social media.

The fossil identification process

The results for the ‘majority vote test’ relative to the collection label (Fig. 7B) suggest that a majority vote yields little to no improvement in this particular case. On average, the consensus vote of all ten experts performs roughly as well as the vote of any one expert. Counterintuitively, a single expert can reach the same or even better results than all ten experts through a majority vote. For future work we recommend assessing the effect of peer verification by experts on the quality of identifications and seeing whether the majority vote can outperform the individual expert votes in the ‘majority vote test’. Second, the lower percentages for the ‘majority vote test’ for the collection labels point to a discrepancy between collection labels and expert predictions. This difference partially arises from collection labels generally being at the species level, and many experts

identifying some of the fossils at a higher ‘taxon level’ (Fig. 4). A similar observation can be made for the AI model predictions from the ‘Split data’ test set (discussed above).

Depending on what would be considered (partially) correct, both the expert and the AI model performance could increase.

As there is no ground truth for the classifications of fossils, the methods used to measure both ML model and expert performance (Fleiss’ Kappa and ‘taxon distance’) are inherently subjective. Although Fleiss’ Kappa provides a good measure of inter annotator agreement, the statistical method does not take classes into account that are not voted on which limits its use. However, by combining these different methods we gain a well-supported measure of performance.

The results presented here cannot directly be compared with results from other ML models trained on palaeontological data, as they differ in data quality and quantity, taxonomical group studied, and taxonomical level at which the images were classified. For example, *Liu et al. (2023)* classify a broad range of fossils at a broad clade level, whereas the models presented here can identify fossils across taxonomical levels up to the species level. The models presented by *Hsiang et al. (2019)*, for example, identify planktonic foraminifers to the species level and those models perform similarly as ours on accuracy. However, we do see similar general results and limitations in results. Our best performing models scored comparable to experts, similar to comparisons between AI and human experts in other domains (*Li et al., 2022; Feighelein et al., 2025*). Moreover, studies in neighbouring domains also found that expert identifications are often far from unanimous and a reliable majority consensus cannot always be achieved (*Austen et al., 2016; Al-Sabouni et al., 2018*), and that expert identifications are more consistent when presented with the actual specimens instead of digital specimens (*Al-Sabouni et al., 2018*).

By making the fossil-identification models public, we help accelerate data collection and validation as well as increase public engagement. At Naturalis Biodiversity Center in Leiden, Netherlands, CNN models were trained on digitised collection material from various extant and extinct organism groups, and were made available online, including the ‘public models’ presented here. Users can go to <https://museum.identify.biodiversityanalysis.nl/> and follow the instructions on how to best use the model to identify their own Quaternary vertebrate fossil finds. The current models are regarded as a prototype and need to be iteratively improved before yielding research grade results. We made the model weights available at <https://doi.org/10.5281/zenodo.20054200>. With increased engagement through the web application, and continued data collection and validation, model performance can be improved over time.

CONCLUSION

In this article we presented convolutional neural networks (CNNs) that were trained to identify fossils and artefacts from the Dutch beaches, based on a standardised image dataset from museum and citizen science (CS) collections and the dataset from the CS platform Oervondstchecker.nl. Second, we compared the two datasets and assessed how well-suited this CS dataset is for an AI training dataset. Third, we discussed the possible applications of the CNNs by comparing the performance of the CNNs with ten domain

experts. We discussed the effects varying data quality have on AI model performance and provided pathways for AI model performance improvements that enable generation of high quality palaeontological and archaeological research data.

The best performing models scored comparable to experts, similar to comparisons between AI and human experts in other domains ([Li et al., 2022](#); [Feighelstein et al., 2025](#)). If provided with standardised data that fall within the training dataset distribution, these models can be used in museum collections and by collectors to identify their finds. These models provide a launching pad for more specialised models that can identify certain groups of fossils or even artefacts that are especially difficult to distinguish. The ‘public models’ were trained to identify CS data with more varying data quality. Although these models perform well on common fossils that are relatively well-preserved and well-photographed, many more rare fossil taxa or artefacts and fragmented material cannot yet be accurately identified.

Expert identifications remain instrumental to sound palaeontological and archaeological research. Although, based on this study, majority votes do not provide significant improvements in identifications, we recommend peer verification by experts to improve data quality. Moreover, being aware of potential biases and taking the time to consult the literature and comparative collections further improves data quality. Because this is time consuming, we believe AI can play a role in aiding experts, for example by providing a second opinion or preselecting fossil finds on a more general taxonomic level to send to the expert with the right specialisation. CS data improve model performance, regardless of the test dataset. Moreover, by making the models available online for use by professional experts, citizen scientists and the general public alike, we can further increase data acquisition and public engagement.

ACKNOWLEDGEMENTS

We want to thank collectors and citizen scientists Niels van Steijn, Henk Mulder and Wil Mulder for actively contributing to our standardised image dataset. We thank Menno Sman and Eva Sell for sharing images from their well-documented collection. Fred Sieval provided additional fragmented fossil finds that were added to the standardised dataset. We thank the ten anonymous experts who helped and lent their expertise for this study. Elsa Girard helped with data processing for [Fig. 7](#). Finally, we want to thank the many Dutch palaeontology and archaeology collectors and citizen scientists involved in the LegaSea project, as well as all the enthusiasts who have been uploading their beach finds to Oervondstchecker.nl over the past 11 years.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Dutch Research Council (NWO) Open Competitie ENW-M: OCENW.M20.360. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Dutch Research Council (NWO).

Open Competitie ENW-M: OCENW.M20.360.

Competing Interests

Isaak Eijkelboom, Laurens E. Hogeweg, Django Brunink, Hansjorg Ahrens, Natasja den Ouden, Anne S. Schulp and Frank P. Wesselingh are employed by Naturalis Biodiversity Center. Floris ter Haar is a volunteer at Naturalis Biodiversity Center. Charlie Schouwenburg is a volunteer at Oertijdmuseum. Bram W. Langeveld is employed by Natural History Museum Rotterdam. Luc Amkreutz is employed by National Museum of Antiquities.

Author Contributions

- Isaak Eijkelboom conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Laurens E. Hogeweg conceived and designed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Django Brunink conceived and designed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Menno Jansen conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Hansjorg Ahrens performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Charlie Schouwenburg performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Floris ter Haar performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Natasja den Ouden analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Bram W. Langeveld analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Luc Amkreutz analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Anne S. Schulp conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Frank P. Wesselingh conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available in the [Supplemental Files](#).

The dataset of Naturalis Biodiversity Center is available on the institute's online repository:

- <https://medialib.naturalis.nl/>.

- To access this data, specify in search: Producent = Paleontologie-CRS-Algemeen; Sinds (yyyy-mm-dd) = 2026-01-21; Tot (yyyy-mm-dd) = 2026-01-29).

The dataset of the Natural History Museum Rotterdam is available on GBIF: Slieker F J A, van der Es H, Andeweg R, Langeveld B W, Schnörr S (2025). Natural History Museum Rotterdam—Specimens. Version 1.46. Natural History Museum Rotterdam. Occurrence dataset <https://doi.org/10.15468/kwqaay>.

The dataset of the Dutch National Museum of Antiquities is available on DANS and on the institute's online repository:

- I. Eijkelboom. 2026. "LegaSea Project—Image Dataset Dutch National Museum of Antiquities." DANS Data Station Archaeology. <https://doi.org/10.17026/AR/NRYH0C>.

- <https://www.rmo.nl/collectie/collectiezoeker/>.

- The 271 specific collection reference numbers are available in the [Supplemental Files](#).

The Naturalis AI Pipeline is available at GitLab:

- <https://gitlab.com/naturalis/bii/ai/naturalis-ai>.

The model weights are available at Zenodo: Eijkelboom, I. (2026).

Eijkelboom_et_al_CNN_model_weights_labels_train_log [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.20054200>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj-cs.4049#supplemental-information>.

REFERENCES

- Al-Sabouni N, Fenton IS, Telford RJ, Kučera M. 2018. Reproducibility of species recognition in modern planktonic foraminifera and its implications for analyses of community structure. *Journal of Micropalaeontology* 37(2):519–534 DOI 10.5194/jm-37-519-2018.
- Amkreutz L, Van der Vaart-Verschoof SE. 2021. *Doggerland. Lost world under the North Sea*. Leiden: Sidestone Press.
- Austen GE, Bindemann M, Griffiths RA, Roberts DL. 2016. Species identification by experts and non-experts: comparing images from field guides. *Scientific Reports* 6(1):3364 DOI 10.1038/srep33634.
- Bejani MM, Ghatte M. 2021. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review* 54(8):6391–6438 DOI 10.1007/s10462-021-09975-1.
- Canty A, Ripley B, Brazzale AR. 2024. R package "boot". DOI 10.32614/CRAN.package.boot.
- CETAF Earth Sciences Group. 2020. Geoscience collections access service (GeoCAsE). Available at <https://geocase.eu>.

- Chai J, Zeng H, Li A, Ngai EWT. 2021. Deep learning in computer vision: a critical review of emerging techniques and application scenarios. *Machine Learning with Applications* 6:100134 DOI 10.24433/CO.0411648.v1.
- Chen C, Liu H, Yang J, Mervak BM, Kalaycioglu B, Lee G, Cakmakli E, Bonatti M, Pudu S, Kahraman O, Pamuk GG, Oto A, Chatterjee A, Tan C. 2025. Can domain experts rely on AI appropriately? A case study on AI-assisted prostate cancer MRI diagnosis. ArXiv DOI 10.48550/arXiv.2502.03482.
- Drisko JW, Maschi T. 2016. *Content analysis*. Oxford, UK: Oxford University Press.
- Eijkelboom I, Ahrens H, Schouwenburg C. 2024. 10 jaar Oervondstchecker. *Cranium* 41(2):44–68.
- Eijkelboom I, Schulp AS, Amkreutz L, Verheul D, der Vaart WVV, van der Vaart-Verschoof S, Hogeweg L, Brunink D, Mol D, Peeters H, Wesselingh F. 2025. Making sense of fossils and artefacts: a review of best practices for the design of a successful workflow for machine learning-assisted citizen science projects. *PeerJ* 13:e18927 DOI 10.7717/peerj.18927.
- Feighelestein M, Luna SP, Silva NO, Trindade PE, Shimshoni I, van der Linden D, Zamansky A. 2025. Comparison between AI and human expert performance in acute pain assessment in sheep. *Scientific Reports* 15(1):626 DOI 10.1038/s41598-024-83950-y.
- Gamer M, Lemon J, Fellows I, Singh P. 2019. R package “irr.”. DOI 10.32614/CRAN.package.irr.
- Gooliaff T, Hodges KE. 2018. Measuring agreement among experts in classifying camera images of similar species. *Ecology and Evolution* 8(22):11009–11021 DOI 10.1002/ece3.4567.
- Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalgo A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L. 2020. Conceptual design blueprint for the DiSSCo digitization infrastructure—DELIVERABLE D8.1. *Research Ideas and Outcomes* 6:e54280 DOI 10.3897/rio.6.e54280.
- Hogeweg LE, Gangireddy R, Brunink D, Kalkman VJ, Cornelissen L, Kamminga JW. 2024. COOD: combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification. ArXiv DOI 10.48550/arXiv.2403.06874.
- Hsiang AY, Brombacher A, Rillo MC, Mleneck-Vautravers MJ, Conn S, Lordsmith S, Jentzen A, Henahan MJ, Metcalfe B, Fenton IS, Wade BS, Fox L, Meilland J, Davis CV, Baranowski U, Groeneveld J, Edgar KM, Movellan A, Aze T, Dowsett HJ, Miller CG, Rios N, Hull PM. 2019. Endless forams: >34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. *Paleoceanography and Paleoclimatology* 34(7):1157–1177 DOI 10.1029/2019PA003612.
- Koch W, Hogeweg L, Nilsen EB, O’hara RB, Finstad AG. 2023. Recognizability bias in citizen science photographs. *Royal Society Open Science* 10(2):221063 DOI 10.1098/rsos.221063.
- Kuitema M, Kolfshoten T, Busschers F, De Loecker D. 2015. *Geoarchaeological and palaeontological research in the Maasvlakte 2 sand extraction zone and on the artificially created Maasvlakte 2 beach: a synthesis*. Rotterdam, Netherlands: Bureau Oudheidkundig Onderzoek Rotterdam.
- Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. *Biometrics* 1:159–174.
- LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–444 DOI 10.1038/nature14539.
- Li MD, Huang ZR, Shan QY, Chen SL, Zhang N, Hu HT, Wang W. 2022. Performance and comparison of artificial intelligence and human experts in the detection and classification of colonic polyps. *BMC Gastroenterology* 22(1):517 DOI 10.1186/s12876-022-02605-2.

- Liu X, Jiang S, Wu R, Shu W, Hou J, Sun Y, Sun J, Chu D, Wu Y, Song H. 2023. Automatic taxonomic identification based on the fossil image dataset (>415,000 images) and deep convolutional neural networks. *Paleobiology* 49(1):1–22 DOI 10.1017/pab.2022.14.
- Liu Y, Torres LN, Wang B, Pei W, Na Y, Song Q, Shi X. 2025. Artificial intelligence in paleobotany and palynology. *Geological Journal* 61(3):748–769 DOI 10.1002/gj.70007.
- McClure EC, Sievers M, Brown CJ, Buelow CA, Ditria EM, Hayes MA, Pearson RM, Tulloch VJD, Unsworth RKF, Connolly RM. 2020. Artificial intelligence meets citizen science to supercharge ecological monitoring. *Patterns* 1(7):100109 DOI 10.1016/j.patter.2020.100109.
- McKay D, Kvammen A. 2020. Auroral classification ergonomics and the implications for machine learning. *Geoscientific Instrumentation, Methods and Data Systems* 9(2):267–273 DOI 10.5194/gi-9-267-2020.
- Miller T, Michoński G, Durlík I, Kozłowska P, Biczak P. 2025. Artificial intelligence in aquatic biodiversity research: a PRISMA-based systematic review. *Biology* 14(5):520 DOI 10.3390/biology14050520.
- Moerdijk PW, Janssen AW, Wesselingh FP, Peeters GA, Pouwer R, van Nieulande FAD, Janse AC, van der Slik L, Meijer T, Rijken R, Cadée GC, Hoeksema D, Doeksen G, Bastemeijer A, Strack H, Vervoenen M, Poorten T. 2010. *De fossiele schelpen van de Nederlandse kust*. Leiden: Nederlands Centrum voor Biodiversiteit Naturalis.
- Mol D. 2016. Mammoth fossils recovered from the seabed between the British Isles and the European continent. *Bulletin Du Musée d'anthropologie Préhistorique de Monaco* 6:129–142.
- Mol D, Bakker R. 2022. Quaternary terrestrial megafaunal remains and their localities in the southern bight of the North Sea between the British Isles and the Netherlands: an overview. *Staringia* 17(1):88–127.
- Mol D, Post K, Reumer JWF, van der Plicht J, de Vos J, van Geel B, van Reenen G, Pals JP, Glimmerveen J. 2006. The Eurogeul—first report of the palaeontological, palynological and archaeological investigations of this part of the North Sea. *Quaternary International* 142–143(3):178–185 DOI 10.1016/j.quaint.2005.03.015.
- Peters SE, McClennen M. 2015. The Paleobiology Database application programming interface. *Paleobiology* 42(1):1–7 DOI 10.1017/pab.2015.39.
- Pichler M, Hartig F. 2023. Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution* 4:994–1016 DOI 10.1111/2041-210X.14061.
- Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L. 2021. ImageNet-21K pretraining for the masses. ArXiv DOI 10.48550/arXiv.2104.10972.
- Santos CFGD, Papa JP. 2022. Avoiding overfitting: a survey on regularization methods for convolutional neural networks. *ACM Computing Surveys* 54(10s):1–25 DOI 10.1145/3510413.
- Schmidhuber J. 2015. Deep learning in neural networks: an overview. *Neural Networks* 61(3):85–117 DOI 10.1016/j.neunet.2014.09.003.
- Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52(3):591–611 DOI 10.1093/biomet/52.3-4.591.
- Shivaprakash KN, Swami N, Mysorekar S, Arora R, Gangadharan A, Vohra K, Jadeyegowda M, Kiesecker JM. 2022. Potential for artificial intelligence (AI) and machine learning (ML) applications in biodiversity conservation, managing forests, and related services in India. *Sustainability* 14(12):7154 DOI 10.3390/su14127154.
- Tan M, Le QV. 2021. EfficientNetV2: smaller models and faster training. In: *International Conference on Machine Learning*. Westminster: PMLR, 10096–10106.

- Von Den Driesch A. 1976.** *A guide to the measurement of animal bones from archaeological sites.* Vol. 1. Cambridge, Massachusetts: Peabody Museum Press.
- Wilcoxon F. 1945.** Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6):80–83 DOI [10.2307/3001968](https://doi.org/10.2307/3001968).
- Yu C, Qin F, Watanabe A, Yao W, Li Y, Qin Z, Liu Y, Wang H, Jiangzuo Q, Hsiang AY, Ma C, Rayfield E, Benton MJ, Xu X. 2024.** Artificial intelligence in paleontology. *Earth-Science Reviews* **252**(28–39):104765 DOI [10.1016/j.earscirev.2024.104765](https://doi.org/10.1016/j.earscirev.2024.104765).