AI Species Identification Using Image and Sound Recognition for Citizen Science, Collection Management and Biomonitoring: From Training Pipeline to Large-Scale Models

Laurens Hogeweg[‡], Ni Yan[‡], Django Brunink[‡], Khadija Ezzaki-Chokri[‡], Wilfred Gerritsen[‡], Rita Pucci[‡], Burooj Ghani[‡], Dan Stowell[‡], Vincent J. Kalkman[‡]

[‡] Naturalis Biodiversity Center, Leiden, Netherlands

Corresponding author: Ni Yan (ni.yan@naturalis.nl)

Abstract

Biodiversity data are currently being generated at an unprecedented rate from deployed field monitoring sensors (e.g., wildlife and insect cameras, sound recorders, radars), citizen science observations, digitised museum collections, and biodiversity- and environmental-generated research. Deep neural networks have made it possible to automatically identify species on multimedia (e.g., image, sound, radar, DNA) with increasing accuracy and efficiency, a task that would otherwise be impossible for taxonomic experts to perform at the rate and scale at which these data are being generated. Artificial intelligence (AI) models can help understand biodiversity data and automate tasks.

At Naturalis Biodiversity Center, we developed several AI species identification models using image or sound recognition for citizen science, collection management and biomonitoring purposes. We present here a pipeline for training large-scale AI species identification models combining multiple sources of image training data that cover the most commonly encountered macro-organisms in Europe.

The training pipeline is shown in Fig. 1. First, 45.4 million images from a total of 133,367 taxonomic names from six different data sources were pooled and mapped into 35.5 million images of 41,014 unique taxa using a custom-developed software tool TaxonMap. From the pooled data, shared models for eight different plants, animals and fungi species groups were trained using imbalance mitigating techniques to increase data efficiency. Subsequently, the shared models were finetuned using data from each source to adapt to its species frequency distribution and local taxonomies. For the three insect species groups, specialised models that predict the life stage of the organism were also trained.

As shown in Fig. 2, the resulting species identification algorithm consists of 39 specialised models, which includes one main model and eight species group models, each customised for four European organisations, plus three life-stage models for the three insect groups.

Measured on the same test data, which have not been used for training the models, the 2023 large-scale multi-source model (MSM), fine-tuned and customised for Observation.org, showed significant performance improvement compared to the 2021 model trained with only their own data. As shown in Fig. 3, t he 2023 model not only includes more taxa, but also identifies species with greater accuracy, especially for the rarer taxa, as measured by average recall.

Fig. 4 presents the effect of class imbalance on model performance by showing the relationship between the number of training samples (right vertical axis) and the accuracy and average recall including the number of most common taxa. Analysis is performed in the mollusks species group for Observation.org. The right vertical axis shows the strong class imbalance in data out of the around 800 taxa in this species group, with the rarest taxon having only ten training images and the most common taxon having about a thousand training images. Measured on the 2023 test data, the average accuracy for all taxa (right-most point in the figure) in this species group was 86%, with the average recall being 64%. By including rarer taxa, average recall drops as expected, while accuracy drops less, as accuracy is mostly influenced by common taxa.

Fig. 5 shows how the analysis of Fig. 4 can be used to compare different models, in this case the multi-source 2023 arthropod model customised for Norway vs. the 2022 arthropod model trained on only the Norwegian data. The 2023 model showed an improved accuracy of 5% for all taxa included, and an even larger improvement on the identification of rarer taxa of about 11%.

The large-scale species identification model, with its 39 specialised models, has been deployed as an auto-scaling web service used by seven (in 2024) biodiversity portals in Europe, and has performed about 65 million identifications in the past 12 months (Aug 2023–Aug 2024), allowing citizen scientists and interested public to identify European flora and fauna using web interface and/or interactive mobile apps, increasing the speed of collecting citizen science data.

Continuous developments of advanced features for this large-scale species identification model are taking place. In the 2023 model, we have implemented explicit probability calibration of AI identifications, allowing automatic validation. Auto-validation is a feature that suggests those AI identifications of the data with low risk, without the need for expert review. Advanced features to be implemented in the 2024 model include providing prediction probabilities at all taxonomic levels (only species level in the 2023 model) and developing life-stage models for other species groups. Planned advanced features for 2025 include context-aware identification (using location, time and neighbouring species to improve identification), rejecting invalid and unusable input such as selfies, poor

quality and unknown taxa (Hogeweg 2024), and image search (returning images similar to the input image).

We have developed this large-scale multi-source model using citizen science observation data from several European biodiversity portals. This AI training pipeline can be applied to develop other large-scale, multi-source algorithms for biodiversity monitoring with sensor input (e.g., insect cameras), digitised museum collection identification as part of the digitisation and collection management workflow, and sound recognition models for citizen science and biomonitoring.

Keywords

artificial intelligence, machine learning, deep learning, hierarchical model, accuracy, recall, class imbalance

Presenting author

Ni Yan

Presented at

SPNHC-TDWG 2024

Funding program

Some of our AI work is partially funded by the European Commission through the project MAMBO, GUARDEN and TETTRIs. Views and opinions expressed are those of the author(s) only.

Hosting institution

Naturalis Biodiversity Center, Leiden, The Netherlands

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Hogeweg L, et al. (2024) COOD: Combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification.

In: CVPR, et al. (Ed.) IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024.

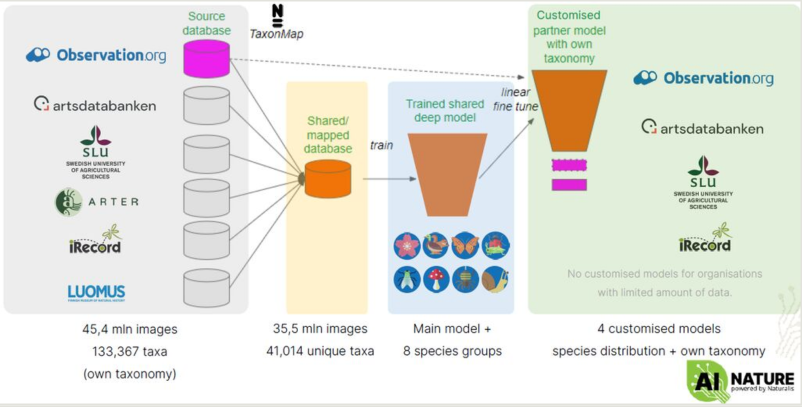
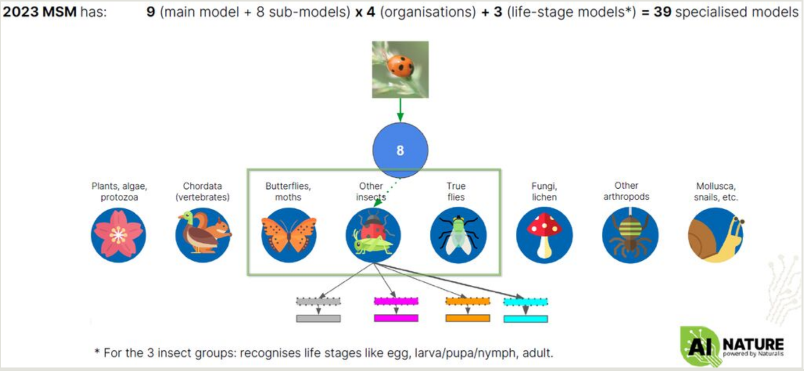


Figure 1.

Training pipeline for a large-scale Multi-Source Model (MSM).

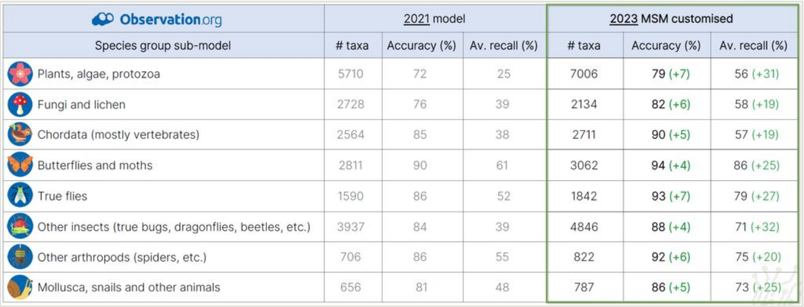


2023 MSM has: **9** (main model + 8 sub-models) **x 4** (organisations) **+ 3** (life-stage models*) **= 39** specialised models

* For the 3 insect groups: recognises life stages like egg, larva/pupa/nymph, adult.

Figure 2.

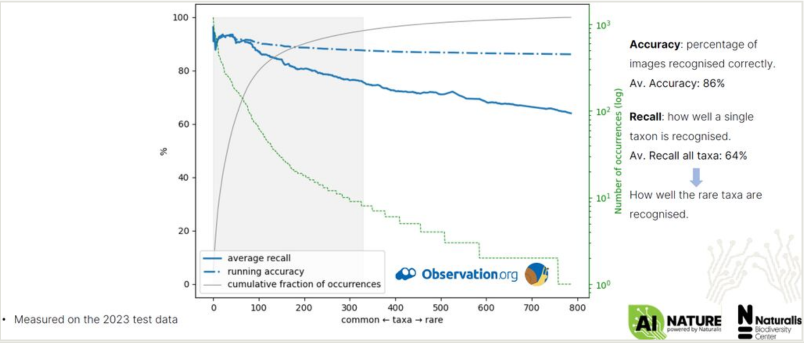
Large-scale species identification hierarchical model.



Observation.org	2021 model			2023 MSM customised		
Species group sub-model	# taxa	Accuracy (%)	Av. recall (%)	# taxa	Accuracy (%)	Av. recall (%)
Plants, algae, protozoa	5710	72	25	7006	79 (+7)	56 (+31)
Fungi and lichen	2728	76	39	2134	82 (+6)	58 (+19)
Chordata (mostly vertebrates)	2564	85	38	2711	90 (+5)	57 (+19)
Butterflies and moths	2811	90	61	3062	94 (+4)	86 (+25)
True flies	1590	86	52	1842	93 (+7)	79 (+27)
Other insects (true bugs, dragonflies, beetles, etc.)	3937	84	39	4846	88 (+4)	71 (+32)
Other arthropods (spiders, etc.)	706	86	55	822	92 (+6)	75 (+20)
Mollusca, snails and other animals	656	81	48	787	86 (+5)	73 (+25)

Figure 3.

Measured performance improvement of MSM (2023) vs Observation model (2021) using the same test data. Accuracy is measured as the percentage of observations in which the first prediction is correct. Average recall is measured as the average recognition rate per taxon. Higher values indicate better recognition of rarer taxa.



Accuracy: percentage of images recognised correctly.
Av. Accuracy: 86%

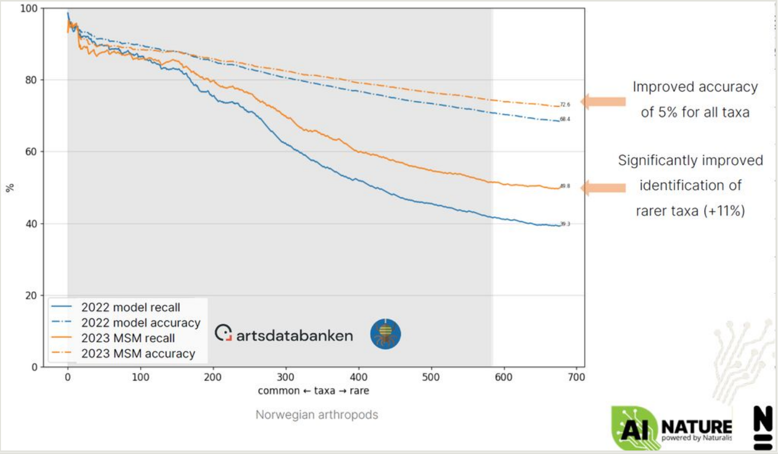
Recall: how well a single taxon is recognised.
Av. Recall all taxa: 64%

How well the rare taxa are recognised.

- Measured on the 2023 test data

Figure 4.

Strong class imbalance in data and its effect on accuracy and average recall in 2023 MSM for Observation.org mollusks (Grey area: 95% of observations).



Improved accuracy of 5% for all taxa

Significantly improved identification of rarer taxa (+11%)

Figure 5.

Measured performance improvement of MSM (2023) vs artsdatabanken model (2022) for Norwegian arthropods (Grey area: 95% of observations).