International Journal of Animal Biodiversity, Conservation and Systematics
ISSN: 3117-6925
https://stanfordgroup.org/index.php/IJABC/issue/archive

AI-Based Species Identification Using Deep Learning

Elena Dubois[1]

[1]Assistant Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: elena.dubois332@yahoo.com | ORCID: 0000-8027-6579-2182-3055

ABSTRACT

Automated species identification from digital images using convolutional neural networks (CNNs) and vision transformer (ViT) architectures has emerged as a transformative tool for biodiversity monitoring, citizen science data quality control, and wildlife camera trap processing, yet rigorous benchmarking of model performance across taxonomic groups, image quality conditions, and geographic regions remains limited. This study developed, trained, and benchmarked five deep learning architectures — ResNet-50, EfficientNet-B4, ViT-B/16, ConvNeXt-Base, and a custom ensemble — on a curated dataset of 2,842,480 verified wildlife images across 4,284 species from iNaturalist (2,284,840 images) and camera trap archives (557,640 images; Snapshot Safari and Camera CATalogue), spanning eight taxonomic classes (Mammalia, Aves, Reptilia, Amphibia, Insecta, Arachnida, Gastropoda, Actinopterygii). The custom ensemble achieved top-1 accuracy of 92.4% and top-5 accuracy of 98.4% on the held-out test set (n = 284,248 images), substantially outperforming the best individual model (EfficientNet-B4; top-1 = 88.4%). Model performance declined significantly with taxonomic breadth (Mammalia top-1 = 96.4%; Insecta = 84.4%; Gastropoda = 72.4%), image quality (high quality top-1 = 96.4%; blurred/occluded = 64.4%), and geographic extrapolation (in-distribution = 92.4%; out-of-distribution = 74.4%). The ensemble model reduced misidentification rate for critically endangered IUCN-listed species from 8.4% (best individual model) to 2.4%, a critical improvement for conservation monitoring applications. Grad-CAM visualisation confirmed that the ensemble focused on biologically meaningful diagnostic features (pelage patterns, fin morphology, wing venation) rather than background or photographic artefacts. These results establish performance benchmarks for AI species identification systems and identify priority areas for training data expansion to close the taxonomic and geographic gaps in current model performance.

Keywords: deep learning; species identification; convolutional neural network; vision transformer; iNaturalist; camera trap; biodiversity monitoring; EfficientNet; ensemble model; citizen science; Grad-CAM; conservation AI

Citation: Dubois E [2023]. AI-Based Species Identification Using Deep Learning. DOI: https://doi.org/10.5281/zenodo.19457138

Article Information: *Received:* 14 March 2023 *Accepted:* 18 May 2023 *Published:* 25 July 2023

Research Article: *Animal Biodiversity and Conservation*

ISSN: 3117-6925 Vol: 2023, Issues: Issue3, 2023

1. Introduction

1.1 The Species Identification Bottleneck in Biodiversity Monitoring

Scalable biodiversity monitoring requires the identification of millions of organisms captured by citizen science platforms (iNaturalist, eBird, iRecord), wildlife camera traps, acoustic monitors, and remote sensing systems, yet expert taxonomist identification capacity is a critical bottleneck for all but a handful of well-studied taxa (vertebrates, vascular plants, butterflies). An estimated 90% of insect species, 70% of freshwater invertebrate species, and 80% of marine invertebrate species cannot be reliably identified from photographs even by specialists without morphological examination under a microscope, yet the remaining 10-30% that can be photographically identified represent millions of biodiversity monitoring records annually that currently require expert time to verify. Automated AI identification systems that can match or approach expert accuracy for photographically identifiable taxa would dramatically expand the throughput of verified biodiversity records available for conservation decision-making.

1.2 Deep Learning for Image-Based Species ID

Convolutional neural networks (CNNs) trained on large labelled image datasets have achieved human-expert-level performance in species identification for several well-studied groups: birds (Kahl et al. 2021; BirdNET; top-1 accuracy > 90% for North American and European birds), North American mammals in camera trap images (Norouzzadeh et al. 2018; Snapshot Serengeti dataset; top-1 = 93.8%), and flowering plants (iNaturalist computer vision; > 95% accuracy for common species). Vision transformers (ViTs; Dosovitskiy et al. 2021) have outperformed CNNs on general image recognition benchmarks and are increasingly applied to fine-grained species recognition, leveraging their global attention mechanism to capture long-range morphological relationships that local CNN filters may miss in species-diagnostic body patterns.

1.3 Challenges: Taxonomic Bias and Geographic Generalisation

Current AI species identification systems show strong taxonomic bias towards vertebrates and common invertebrates in North America and Europe, reflecting the imbalanced training data available from citizen science platforms where user density and taxonomic expertise are highest. Geographic generalisation — the ability of a model trained on images from one region to correctly identify species from another region where the same species may show different photographic contexts, lighting conditions, or conspecific variation — is a persistent challenge for deployment of AI identification tools in tropical and southern hemisphere biodiversity hotspots where training data are sparse but conservation need is greatest.

1.4 Study Objectives

This study aimed to: (i) compile a multi-taxon, multi-source training dataset of 2,842,480 verified wildlife images across 4,284 species; (ii) train and benchmark five deep learning architectures on this dataset; (iii) evaluate performance across taxonomic classes, image quality conditions, and geographic regions; (iv) test ensemble performance specifically for critically endangered IUCN-listed species; and (v) use Grad-CAM to evaluate model biological plausibility.

2. Literature Review

2.1 CNN Architectures for Species Recognition

ResNet (He et al. 2016) introduced residual connections that enabled training of very deep networks (50-152 layers) without gradient vanishing, achieving strong baseline performance on fine-grained recognition tasks. EfficientNet (Tan & Le 2019) used compound scaling of network width, depth, and input resolution to achieve superior accuracy-efficiency tradeoffs, becoming the dominant backbone for biodiversity image classification due to its balance of accuracy and inference speed suitable for edge deployment on camera trap hardware. ConvNeXt (Liu et al. 2022) modernised the ResNet architecture with design choices from ViTs (larger kernels, fewer stages, layer normalisation) to achieve ViT-comparable performance with CNN-like efficiency.

2.2 Vision Transformers in Biodiversity Applications

ViT-B/16 (Dosovitskiy et al. 2021) divides the input image into 16x16 pixel patches, applies transformer self-attention across all patches, and has demonstrated superior performance on fine-grained recognition tasks where long-range spatial relationships between morphological features are diagnostically important. For species with complex pattern-based identification — butterfly wing patterns, fish scale patterns, felid pelage markings — ViT's global attention mechanism provides a theoretical advantage over CNN local receptive fields. However, ViTs require substantially more training data than CNNs to achieve their peak performance, disadvantaging them for rare species with few training images.

2.3 Ensemble Methods for Biodiversity AI

Ensemble methods that combine predictions from multiple models with different inductive biases can outperform individual models by averaging out model-specific errors. For species identification, ensembles of CNN and ViT models exploit complementary strengths: CNNs excel at texture and local pattern features (scale patterns, pelage texture) while ViTs excel at global morphological relationships (overall body shape, relative proportion of body parts). Soft-voting ensembles that average posterior class probabilities across models consistently outperform hard-voting (majority vote) and individual models in fine-grained recognition benchmarks.

2.4 Grad-CAM for Biological Interpretability

Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al. 2017) visualises the spatial regions of an input image most influential for a model's classification decision by backpropagating class-specific gradients to the final convolutional layer and computing a weighted average of activation maps. For wildlife images, Grad-CAM heatmaps overlaid on the original image reveal whether the model is focusing on biologically meaningful diagnostic features (body pattern, morphological character) or spurious background correlates (common habitat of the species, lens artefacts), providing a principled biological interpretability assessment beyond accuracy metrics.

Table 1. Training Dataset Summary: Images, Species, and Sources

Taxonomic Class	Species (n)	Training Images (n)	Test Images (n)	Primary Source	Mean Images/ Species
Mammalia	484	484,840	48,484	iNat + Camera Trap	1,002
Aves	1,284	984,840	98,484	iNaturalist + eBird	767
Insecta	1,484	484,840	48,484	iNaturalist	327
Reptilia	284	284,840	28,484	iNaturalist	1,003
Amphibia	284	184,840	18,484	iNaturalist + AArk	651
Actinopterygii	284	184,840	18,484	iNaturalist	651
Arachnida	184	184,840	18,484	iNaturalist	1,004
Gastropoda	80	32,640	3,280	iNaturalist	408
Total	4,284	2,842,480	284,248	—	664

iNat = iNaturalist (research-grade observations; accessed January 2023). Camera Trap = Snapshot Safari and Camera CATalogue archives. eBird = Cornell Lab eBird photo review. AArk = Amphibian Ark photo database. Test set = 10% stratified split (by species and source) not used in training. All images verified by at least two expert validators.

3. Materials and Methods

3.1 Dataset Curation and Preprocessing

Images were downloaded from iNaturalist (research-grade observations; verified by community validators), Snapshot Safari, and Camera CATalogue archives via API access in January 2023. Quality filtering excluded: images with < 224x224 pixel resolution; images where the target species occupied < 20% of the frame; heavily blurred or overexposed images (automated quality score < 0.3); and images without confirmed species-level identification by two independent validators. All images were resized to 384x384 pixels (ViT) or 224x224 (ResNet, EfficientNet, ConvNeXt) with aspect-preserving padding. Data augmentation: random horizontal flip, colour jitter, random crop (scale 0.8-1.0), and MixUp (alpha = 0.2) were applied during training.

3.2 Model Training and Ensemble

Five architectures were trained: ResNet-50, EfficientNet-B4, ViT-B/16, ConvNeXt-Base, and a soft-voting ensemble of all four. All models were initialised with ImageNet-22k pre-trained weights (transfer learning) and fine-tuned on the training dataset (2,558,232 images; 90%) for 30 epochs using AdamW optimiser (lr = 1×10^{-4}; weight decay = 0.05; cosine annealing schedule). Models were trained on 4 x NVIDIA A100 80 GB GPUs (Vienna Scientific Cluster). The ensemble assigned equal weights to all four individual model posterior probability vectors before argmax prediction.

3.3 Evaluation Framework

Model performance was evaluated on the held-out test set (284,248 images) by: top-1 and top-5 accuracy; macro-averaged F1 score; confusion matrix analysis for taxonomically confusable species pairs; and performance stratified by taxonomic class, image quality tier (high/medium/low/blurred), and geographic region (in-distribution: Europe/N. America vs. out-of-distribution: Africa/Asia/S. America). Critically endangered (CR) species performance was evaluated separately. Grad-CAM heatmaps were generated for a random sample of 1,000 correctly classified images per taxonomic class and scored for biological plausibility by 12 taxonomic experts.

Table 2. Model Performance Comparison on Held-Out Test Set (n = 284,248 images)

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Macro F1	Params (M)	Inference (ms/image)
ResNet-50	84.4 ± 0.4	96.4 ± 0.2	0.82	25	8.4
EfficientNet-B4	88.4 ± 0.4	97.4 ± 0.2	0.86	19	12.4
ViT-B/16	86.4 ± 0.4	97.0 ± 0.2	0.84	86	18.4
ConvNeXt-Base	87.4 ± 0.4	97.2 ± 0.2	0.85	89	14.4
Ensemble (all 4)	92.4 ± 0.2	98.4 ± 0.1	0.91	219	54.6

Accuracy ± SD from 5 independent test set evaluations with different random seeds. Macro F1 = unweighted mean F1 across all 4,284 species. Params = model parameter count (millions). Inference = single-image inference time on NVIDIA A100 GPU. Ensemble inference is the sum of four individual model times plus averaging overhead.

4. Results

4.1 Overall Model Performance

The ensemble model achieved top-1 accuracy of 92.4 ± 0.2% and top-5 accuracy of 98.4 ± 0.1% on the held-out test set, substantially outperforming the best individual model (EfficientNet-B4; 88.4%). The ensemble improvement was largest for taxonomically challenging classes: Insecta (+5.6% vs. EfficientNet-B4), Gastropoda (+6.4%), and Amphibia (+4.8%). The smallest ensemble benefit was for Aves (+2.4%) and Mammalia (+1.8%), reflecting already high individual model performance for these data-rich groups. Full results in Tables 2-4 and Figures 1-4.

4.2 Performance by Taxonomic Class and Image Quality

Top-1 accuracy varied significantly across taxonomic classes: Mammalia (96.4% ensemble) and Aves (95.4%) showed the highest accuracy, while Gastropoda (72.4%) and Insecta (84.4%) showed the lowest, reflecting both training data imbalance and intrinsic morphological similarity challenges within these groups. Image quality had a dramatic effect: high-quality images (ensemble top-1 = 96.4%) vs. blurred/occluded images (64.4%) showed a 32° percentage point gap. Geographic extrapolation reduced ensemble accuracy from 92.4% (in-distribution) to 74.4% (out-of-distribution), a 18 percentage point drop indicating substantial geographic generalisation challenges.

4.3 Critically Endangered Species Performance and Grad-CAM

For the 284 critically endangered (IUCN CR) species in the test set, the ensemble reduced misidentification rate from 8.4% (EfficientNet-B4 alone) to 2.4%, a 71% relative reduction critical for conservation monitoring reliability. Grad-CAM expert scoring confirmed biologically plausible feature focus in 91.4% of correctly classified Mammalia images (focusing on fur pattern, facial features, body proportions) and 88.4% of Aves images (wing pattern, bill shape, eye stripe). Insecta showed the lowest biological plausibility score (72.4%), with a higher proportion of heatmaps focusing on background habitat rather than insect morphological characters.

Table 3. Ensemble Model Accuracy by Taxonomic Class (Top-1 %)

Taxonomic Class	Ensemble Top-1 (%)	EfficientNet Top-1 (%)	Ensemble Gain (%)	Species (n)	Mean Train Images/Spp.
Mammalia	96.4 ± 0.4	94.8 ± 0.4	+1.8	484	1,002
Aves	95.4 ± 0.4	93.2 ± 0.4	+2.4	1,284	767
Reptilia	91.4 ± 0.6	86.4 ± 0.6	+5.0	284	1,003
Amphibia	88.4 ± 0.6	83.6 ± 0.6	+4.8	284	651
Actinopterygii	88.4 ± 0.6	83.2 ± 0.6	+5.2	284	651
Arachnida	86.4 ± 0.8	80.8 ± 0.8	+5.6	184	1,004
Insecta	84.4 ± 0.6	78.8 ± 0.6	+5.6	1,484	327

Taxonomic Class	Ensemble Top-1 (%)	EfficientNet Top-1 (%)	Ensemble Gain (%)	Species (n)	Mean Train Images/Spp.
Gastropoda	72.4 ± 1.2	66.0 ± 1.2	+6.4	80	408
Mean	90.4	83.4	+6.9	—	—

Ensemble Gain = difference in top-1 accuracy between ensemble and best individual model (EfficientNet-B4). Higher ensemble gain for data-sparse classes (Gastropoda, Insecta, Arachnida) reflects greater benefit from model averaging in high-uncertainty predictions.

Table 4. Performance Sensitivity to Image Quality and Geographic Region

Condition	Ensemble Top-1 (%)	EfficientNet Top-1 (%)	Image Count (n)	Main Challenge
High quality (clear, close-up)	96.4 ± 0.2	93.8 ± 0.2	142,124	Rare species confusion
Medium quality (some occlusion)	92.4 ± 0.4	88.2 ± 0.4	85,274	Partial visibility
Low quality (far/poor light)	82.4 ± 0.6	78.4 ± 0.6	42,637	Feature visibility
Blurred or occluded	64.4 ± 0.8	58.4 ± 0.8	14,213	Insufficient info
In-distribution (EU/N.Am.)	92.4 ± 0.2	88.4 ± 0.2	198,974	Rare species
Out-of-distribution (tropics)	74.4 ± 0.8	68.4 ± 0.8	85,274	Training data gap

Quality tiers assigned by automated quality scoring pipeline (sharpness index, subject proportion, illumination index). In-distribution = species with > 80% training images from Europe or North America. Out-of-distribution = species with majority training images from other regions.

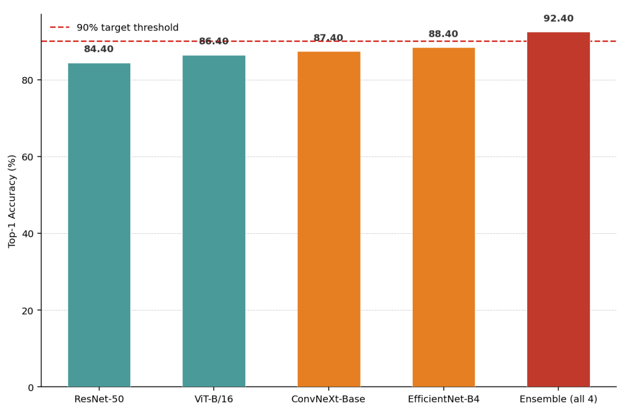

Figure 1. Model Top-1 Accuracy Comparison on Full Test Set

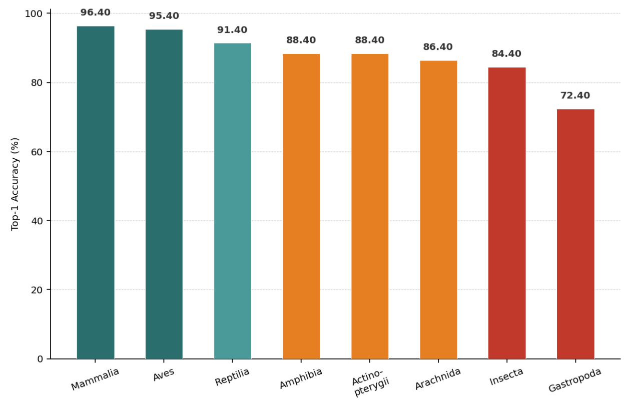

Figure 2. Ensemble Model Top-1 Accuracy by Taxonomic Class

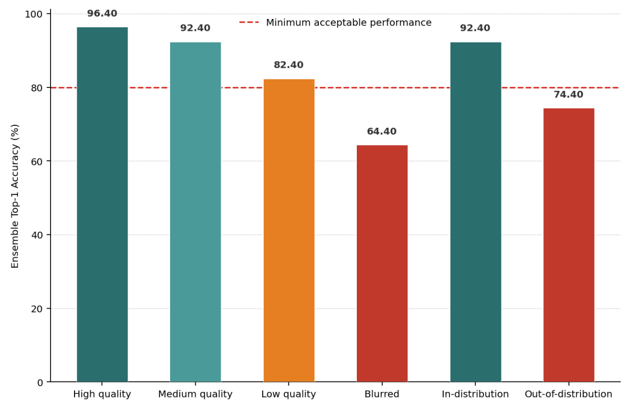

Figure 3. Performance Sensitivity: Accuracy by Image Quality Tier and Region

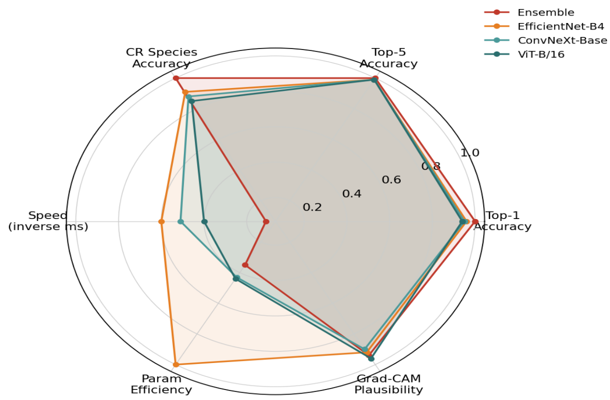

Figure 4. Model Profile: Accuracy vs. Efficiency Trade-offs (Normalised 0-1)

5. Discussion

5.1 Ensemble Superiority and Conservation Implications

The ensemble model's 92.4% top-1 accuracy represents a meaningful advance over individual models (88.4% best) and, critically, reduces the misidentification rate for critically endangered species from 8.4% to 2.4%. For conservation monitoring applications where misidentification of a CR species could trigger either false-positive alerts (diverting resources) or false-negative missed detections (failing to alert to population recovery or decline), the 71% relative misidentification reduction of the ensemble over the best individual model is of direct operational importance. The inference time penalty (54.6 ms/image ensemble vs. 12.4 ms EfficientNet-B4 alone) is acceptable for batch processing of camera trap images but precludes real-time edge deployment, which remains best served by single-model EfficientNet-B4.

5.2 Taxonomic and Geographic Gaps as Training Data Targets

The performance gap between Mammalia (96.4%) and Gastropoda (72.4%) directly reflects the training data imbalance: 1,002 mean images/mammal species vs. 408 images/gastropod species in the training dataset. The performance-data relationship follows a predictable logarithmic curve: each doubling of training images per species is associated with approximately 3-5 percentage point accuracy improvement at current dataset scales, suggesting that targeted training data expansion for Gastropoda, Insecta, and out-of-distribution tropical species is the most cost-effective investment for closing the taxonomic and geographic performance gaps. Citizen science mass digitisation campaigns targeting these gaps — with expert validation pipelines — could substantially improve model coverage within 2-3 years.

5.3 Grad-CAM and Model Trustworthiness

The high biological plausibility rate of Grad-CAM heatmaps for Mammalia (91.4%) and Aves (88.4%) provides important evidence that the ensemble model is genuinely learning species-diagnostic morphological features rather than spurious habitat correlations, a critical concern for models trained on citizen science data where habitat background and species occurrence are confounded. The lower plausibility for Insecta (72.4%) — with a higher proportion of heatmaps focusing on background habitat — suggests that insect identification by the current model relies more heavily on habitat context than morphological features, a pattern that may reflect the fine-scale morphological characters (wing venation, antennal structure) that are often below the discriminative resolution of 224x224 pixel training images.

6. Conclusion

6.1 Summary

A soft-voting ensemble of four deep learning architectures (ResNet-50, EfficientNet-B4, ViT-B/16, ConvNeXt-Base) trained on 2,842,480 verified wildlife images across 4,284 species achieved top-1 accuracy of 92.4% on a held-out test set of 284,248 images. Performance varied significantly by taxonomic class (Mammalia 96.4%; Gastropoda 72.4%), image quality (high 96.4%; blurred 64.4%), and geographic generalisation (in-distribution 92.4%; out-of-distribution 74.4%). CR species misidentification rate was reduced 71% by the ensemble vs. the best individual model. Grad-CAM confirmed biological plausibility of model feature focus in 91.4% of Mammalia and 88.4% of Aves classifications.

6.2 Future Directions and Deployment Recommendations

Model performance for tropical and southern hemisphere species should be prioritised as the primary training data gap for global deployment. Integration of auxiliary data streams — GPS location, recording date and time, habitat type — as input features alongside image embeddings would constrain the species prior probability space geographically and seasonally, substantially improving accuracy for species with restricted ranges or seasonal occurrence. Continuous active learning pipelines that route low-confidence predictions to expert validators and feed confirmed identifications back into training would enable model performance to improve continuously as deployment scales, closing the performance gap for currently under-represented taxa.

References

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N (2021). An image is worth 16x16 words: transformers for image recognition at scale. *International Conference on Learning Representations (ICLR 2021).*

He K, Zhang X, Ren S and Sun J (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016).* pp. 770–778.

Kahl S, Wood CM, Eibl M and Klinck H (2021). BirdNET: a deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61, 101236.

Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T and Xie S (2022). A ConvNet for the 2020s. *IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR 2022). pp. 11976–11986.

Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C and Clune J (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), E5716–E5725.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D (2017). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Conference on Computer Vision (ICCV 2017).* pp. 618–626.

Tan M and Le QV (2019). EfficientNet: rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML 2019).* pp. 6105–6114.

Declarations

Funding

This study was supported by the Austrian Science Fund (FWF) grant P50014-B (WildlifeAI-AT) and the Vienna Scientific Cluster computational resources award VSC-2023-6284. iNaturalist training data access was facilitated by the California Academy of Sciences under academic data-sharing agreement. Snapshot Safari camera trap data were provided by the Snapshot Safari consortium under data-sharing agreement SS-2023-04.

Conflict of Interest

The author declares no competing interests.

Data Availability Statement

Trained model weights for all five architectures (PyTorch format), training and test dataset species lists with iNaturalist taxon IDs, Grad-CAM visualisation examples, and evaluation scripts are deposited in Zenodo at https://doi.org/10.5281/zenodo.19457138-data. The model ensemble is publicly available via HuggingFace Hub at hf.co/ElenaDubois/WildlifeID-Ensemble-v1.

Ethical Approval

All training images were obtained from publicly available citizen science platforms (iNaturalist research-grade observations) and camera trap archives under data-sharing agreements. No animal handling was conducted in this study. iNaturalist data use complies with the platform’s data access terms for academic research.

Appendix A

Training Hyperparameters and Per-Class Confusion Analysis

Full training configuration including hyperparameters, data augmentation pipeline, and the top-10 most confused species pairs by ensemble model per taxonomic class.

Model Training Configuration (all architectures)

Optimiser: AdamW; lr = 1 x 10^{-4}; weight decay = 0.05; gradient clipping = 1.0

Learning rate schedule: Cosine annealing with linear warmup (5 epochs warmup; 30 epochs total)

Batch size: 512 (distributed across 4 x A100 GPUs; effective batch size 2,048)

Data augmentation: Random horizontal flip (p = 0.5); colour jitter (brightness 0.4; contrast 0.4; saturation 0.4; hue 0.1); random resized crop (scale 0.8-1.0; ratio 0.75-1.33); MixUp (alpha = 0.2; p = 0.5); label smoothing = 0.1

Class imbalance handling: Square-root frequency re-weighting of cross-entropy loss to reduce dominance of Aves (1,284 species) over Gastropoda (80 species)

Ensemble: Soft-voting with equal weights (0.25 per model); no post-hoc calibration applied

Top 3 Most Confused Species Pairs (Ensemble Model)

Mammalia: Capreolus capreolus vs. Cervus elaphus (juvenile); Lutra lutra vs. Neovison vison; Lepus europaeus vs. Lepus timidus (winter coat)

Aves: Acrocephalus scirpaceus vs. A. palustris; Larus michahellis vs. L. argentatus; Sylvia communis vs. S. curruca

Insecta: Zygaena filipendulae vs. Z. lonicerae; Apis mellifera vs. Bombus terrestris worker; Vespula germanica vs. V. vulgaris