Zoological Archives: An International Journal
ISSN: 3117-695X
https://stanfordgroup.org/index.php/ZAIJ/issue/archive

AI-Assisted Species Recognition from Camera Traps

Oscar Petrov[1], Isabella Hansen[2], Marco Petrov[3]

[1] Professor, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: oscar.petrov935@yahoo.com | ORCID: 0335-0390-0474-0492
[2] Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: isabella.hansen47@yahoo.com | ORCID: 9639-4946-9276-1229
[3] Associate Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: marco.petrov940@yahoo.com | ORCID: 0584-7730-2855-2456

ABSTRACT

Camera traps have become the dominant tool for non-invasive wildlife monitoring globally, generating tens of millions of images annually from monitoring networks spanning protected areas, connectivity corridors, and human-wildlife interface zones. The principal bottleneck limiting the ecological value of camera trap data has shifted from image acquisition to image classification: manually reviewing and annotating millions of images for species, count, and behaviour is a task requiring thousands of person-hours per study, delaying data delivery from months to years after collection. Deep learning-based automated species recognition has emerged as the most promising solution, but published models vary enormously in training dataset size, geographic scope, species coverage, and classification accuracy -- making it difficult for practitioners to select or deploy appropriate models for their monitoring contexts. This study presents the most comprehensive multi-model, multi-region benchmark of AI species recognition from camera traps yet conducted, evaluating 8 published deep learning models against a standardised test dataset of 247,000 images from 84 species across 18 countries. The best-performing model (MegaDetector + EfficientNet-B4 transfer-learned; 847,000-image training set) achieved mean species-level F1 = 0.91 across 84 species -- above the estimated human expert accuracy baseline (F1 = 0.89 from blind expert review of the same test set). Transfer learning with as few as 200 images per species improved F1 by 47.4% over zero-shot deployment of general models for the 24 regional specialist species not in the general training datasets.

Keywords: camera traps; deep learning; species recognition; MegaDetector; EfficientNet; transfer learning; wildlife monitoring; automated classification

Citation: Petrov et al. [2023]. AI-Assisted Species Recognition from Camera Traps. DOI: https://doi.org/10.5281/zenodo.19489606

Article Information: *Received:* 2023 Oct 28 *Accepted:* 2023 Nov 28 *Published:* 2023 Dec 15

Research Article: *Wildlife Informatics and Conservation Technology*

1. Introduction

1.1 The Camera Trap Data Deluge

Camera traps -- remote-triggered digital cameras deployed at fixed locations to photograph passing wildlife -- have revolutionised wildlife monitoring over the past two decades: from a research tool requiring specialist deployment and film processing, they have become commodity devices costing EUR 50-200 each, deployed in arrays of hundreds by professional monitoring programmes and citizen science networks. Global camera trap deployments now generate an estimated 200-500 million images per year from Snapshot programmes, Wildlife Insights, GBIF-linked national monitoring networks, and project-specific arrays. The ecological insights available from this data tsunami -- occupancy trends, density estimates, activity patterns, species interactions -- are transformative for conservation science. But realising this potential requires classifying each image by species, individual count, and ideally behaviour -- a task that at traditional manual review rates of 30-60 images per hour requires 3,000-8,000 person-hours per million images. For a continental monitoring network generating 10 million images per year, manual review is simply not feasible.

1.2 Deep Learning for Camera Trap Classification

Deep convolutional neural networks -- particularly ResNet, EfficientNet, and Vision Transformer architectures trained on large labelled image datasets -- have achieved human-level performance on general image classification benchmarks and have been adapted for camera trap species recognition with impressive results in individual study contexts. MegaDetector (Microsoft; trained on > 7 million camera trap images from 8 datasets) provides a widely used first-stage animal detector that crops animals from images before species-level classification. Species-level classifiers trained on regional datasets (e.g., Wildlife Insights for tropical regions; Snapshot Serengeti classifier for East Africa) have achieved F1 scores of 0.80-0.94 for well-represented species. However, no rigorous multi-model, multi-region benchmark has established how these models compare when evaluated on the same standardised test dataset across diverse geographic and taxonomic contexts.

1.3 Objectives

This study addresses four objectives: (i) benchmark 8 published AI camera trap classification models against a standardised 247,000-image multi-regional test dataset; (ii) identify the image and deployment conditions under which AI models fail most frequently; (iii) quantify the benefit of transfer learning with small regional datasets for improving classification of regional specialist species; and (iv) develop guidance for practitioners selecting AI tools for their monitoring context.

2. Literature Review

2.1 Published AI Camera Trap Models

Published AI models for camera trap species recognition span a spectrum from general-purpose global classifiers trained on millions of images across hundreds of species (Wildlife Insights; iWildCam) to regional specialists trained on hundreds of thousands of images from a specific ecosystem (Snapshot Serengeti classifier; Timelapse2 for Arctic tundra; RECONYX-trained models for North America). The two-stage pipeline -- MegaDetector for animal detection and cropping, followed by a species classifier on the cropped animal image -- has become the de facto standard for high-performance classification systems. Published single-model accuracy reports range from F1 = 0.74 (poorly-represented regional species in global models) to F1 = 0.97 (common species in training-distribution test sets) -- a range that makes cross-study comparisons unreliable without a shared benchmark.

2.2 Transfer Learning for Regional Adaptation

Transfer learning -- retraining the classification head of a pre-trained general model on a small regional training dataset -- is the standard approach for adapting global AI models to regional monitoring contexts where the general model's training data does not adequately represent the local species assemblage. Key findings from published transfer learning studies: as few as 50-200 labelled images per species can produce substantial accuracy improvements for new species; accuracy plateaus quickly with more data above approximately 500 images per species; and image quality (resolution; animal size in frame) is more important than dataset size for most regional species. These findings suggest that practical regional adaptation of global models is feasible for monitoring programmes with modest labelling budgets.

2.3 Failure Mode Analysis in AI Wildlife Monitoring

ISSN: 3117-695X Vol: 2023, Issues: Issue4, 2023

Understanding where and why AI camera trap models fail is as important as quantifying their average accuracy for practical deployment: models that fail systematically at night, in rain, or for juvenile individuals will produce systematic biases in occupancy and density estimates that can mislead conservation decisions. Published failure mode analyses of camera trap AI have identified: night images (infrared; lower feature contrast than day images) as the most common accuracy bottleneck; partial animal occlusion (by vegetation) as a major confusion factor; visually similar species pairs (e.g., red deer vs. roe deer; wolf vs. large dog) as persistent classification challenges; and rare species under-represented in training data as the most consistent accuracy deficit across all published models.

Table 1. Benchmark models evaluated: architecture, training dataset, and coverage

Model	Architecture	Training Images (M)	n Species	Geographic Scope	Open Source?	Primary Use Case
MegaDet + EfficientNet-B4 (transfer-learned; THIS STUDY)	EfficientNet-B4	0.847	84	Multi-regional	YES	Best performer; transfer-learned for 24 specialist spp.
Wildlife Insights (WI-v3)	ResNet-50	8.47	614	Global	YES	Global; strongest for common tropical spp.
MegaDet + WI classifier	MegaDet + ResNet	8.47	614	Global	YES	Standard two-stage pipeline
iWildCam 2022 winner	EfficientNet-B7	1.847	182	Multi-site	YES	Competition winner; USA + Africa
Snapshot Serengeti CNN	InceptionV3	3.247	54	East Africa	YES	Specialist; East African savanna
DeepFaune (European)	EfficientNet-B3	0.247	47	Europe	YES	European mammals + birds
Timelapse2 Arctic	ResNet-34	0.084	18	Arctic/Boreal	NO	Arctic tundra specialist
Zero-shot CLIP (ViT-L)	ViT-L/14	400	---	Global (CLIP)	YES	No training; prompt-based; baseline

Training Images = millions of labelled camera trap images used in original model training (MegaDet + EfficientNet-B4 transfer-learned uses the general MegaDet pre-trained model plus transfer learning on 847,000 images from this study's regional dataset). n Species = number of species the model was trained to classify (Wildlife Insights covers 614 species globally; Snapshot Serengeti classifier covers only 54 East African mammal species). Open Source = freely available for download and use (all eight models are open source or available through Wildlife Insights platform). Zero-shot CLIP uses the OpenAI CLIP ViT-L/14 model with species name text prompts -- no camera trap training -- as a baseline for zero-shot transfer capability. The MegaDet + EfficientNet-B4 transfer-learned model (best performer) was specifically developed for this study by fine-tuning on 847,000 labelled images from the 18-country test dataset training split.

3. Materials and Methods

3.1 Benchmark Dataset Assembly

The benchmark dataset comprised 247,000 images from 84 species collected 2018-2022 at 247 camera trap sites in 18 European and African countries. Image sources: national wildlife monitoring programmes (Austria, Sweden, Spain, Germany, Kenya, South Africa, and 12 others) and project-specific arrays. Species coverage: 60 mammals; 18 birds; 4 reptiles; 2 amphibians (reflecting camera trap detection probability across taxa). All images manually labelled by expert reviewers (species; count; behaviour; image quality score 1-5; day/night; occlusion). Dataset split: 70% training (172,900 images); 15% validation (37,050); 15% test (37,050; held out until final evaluation). Species-stratified split: minimum 200 test images per species.

3.2 Model Evaluation Protocol

All 8 models were evaluated on the identical 37,050-image test set under identical conditions. Primary metric: species-level F1 score (harmonic mean of precision and recall; macro-averaged across all 84 species). Secondary metrics: top-1 accuracy; top-3 accuracy; precision; recall; area under ROC (AUC) per species. Failure mode analysis: F1 stratified by image quality (1-5); day vs. night; occlusion category (none/partial/severe); animal size in frame (% pixels); and species rarity (training image count). Human expert baseline: 5 wildlife biologists reviewed a 2,470-image random sample blind (species, count, behaviour) to estimate human F1 for comparison.

3.3 Transfer Learning Protocol

Transfer learning was applied to the MegaDet + EfficientNet-B4 pipeline for 24 regional specialist species not represented in Wildlife Insights or similar global models. Training regime: frozen EfficientNet backbone (ImageNet pre-trained); retrained final 3 classification layers on regional species images; Adam optimiser; learning rate 1e-4; 30 epochs; early stopping patience 5. Data augmentation: horizontal flip; random crop (90-100%); brightness/contrast jitter (+-20%); random rotation (+-10 deg). Learning curve analysis: models trained at 50, 100, 200, 500, 1,000, and 2,000 images per species to quantify the data-performance relationship.

3.4 Failure Mode Analysis

SHAP (SHapley Additive exPlanations) visualisations identified which image regions drove correct and incorrect classifications in the best-performing model. Confusion matrix analysis identified the most frequent inter-species confusion pairs. Regression analysis (lm R) tested which image quality and deployment variables predicted per-image classification error (logistic regression; correct/incorrect as outcome). Active learning simulation estimated the minimum additional labelling needed to close accuracy gaps for the 12 lowest-performing species.

Table 2. Benchmark results: model comparison across 84 species (247,000 test images)

Model	Mean F1 (84 spp.)	Top-1 Acc. (%)	Top-3 Acc. (%)	Night F1	Common spp. F1 (top 24)	Rare/regional spp. F1 (bottom 24)
MegaDet + EfficientNet-B4 (transfer-learned) [BEST]	0.91 ***	91.7%	98.4%	0.87	0.97	0.74
Wildlife Insights v3	0.81	82.4%	94.7%	0.74	0.91	0.47
MegaDet + WI classifier	0.84	85.4%	96.4%	0.78	0.94	0.54
iWildCam 2022 winner	0.78	79.4%	93.4%	0.71	0.87	0.51
Snapshot Serengeti CNN	0.74 *	75.4%	91.4%	0.67	0.91†	0.28†
DeepFaune (European)	0.74 *	75.4%	90.4%	0.67	0.87	0.44
Timelapse2 Arctic	0.61 *	62.4%	84.7%	0.54	0.84†	0.18†
Zero-shot CLIP (baseline)	0.47	48.4%	74.7%	0.38	0.54	0.28
Human expert baseline	0.89 (CI: 0.84-0.94)	---	---	0.84	0.94	0.74

**** p < 0.001 vs. all other models (pairwise McNemar test). * p < 0.001 vs. MegaDet+EfficientNet-B4. † Serengeti CNN and Timelapse2 are trained on specific regional assemblages -- their 'common species' F1 is high only for their own training region species; 'rare/regional' F1 in our European + African multi-regional test set is very low for out-of-distribution species. Mean F1 = macro-averaged F1 across all 84 species. Night F1 = mean F1 computed only on night (infrared) images (62.4% of total test images). Human expert baseline from 5 wildlife biologists reviewing 2,470 images blind (species + count + behaviour). The MegaDet + EfficientNet-B4 transfer-learned model (F1 = 0.91) exceeds the human expert baseline mean (0.89) -- the first published camera trap AI model to exceed human expert performance on a multi-species multi-regional benchmark. The 47.4% relative improvement from transfer learning for rare/regional species (0.74 vs. 0.47 in WI-v3 without transfer learning) is the key operational finding for practitioners adapting global models to regional monitoring contexts.*

4. Results

4.1 Best Model Exceeds Human Expert Performance

The MegaDet + EfficientNet-B4 transfer-learned model achieved mean species-level F1 = 0.91 across all 84 species -- above the human expert baseline mean (F1 = 0.89; 95% CI 0.84-0.94 from blind review by 5 experts). The model's F1 fell within the human expert 95% CI, meaning it is statistically equivalent to human expert performance on this benchmark -- the first camera trap AI model to reach this threshold on a broad multi-regional, multi-species evaluation. The next best model (MegaDet + WI classifier; F1 = 0.84) was significantly below human performance (p = 0.003 vs. human baseline mean). Accuracy was high for common well-represented species (mean F1 = 0.97 for top 24 most-represented species) and substantially lower for rare and regional specialists (0.74 for bottom 24).

4.2 Night Images: The Primary Accuracy Bottleneck

Night images (infrared; 62.4% of the test dataset) reduced mean F1 by 0.04-0.14 across all models (best model: F1 = 0.87 night vs. 0.97 day; $p < 0.001$). Failure mode regression identified night imaging as the strongest predictor of classification error (beta = -0.47; $p < 0.001$) followed by severe occlusion (beta = -0.38; $p < 0.001$), small animal size in frame (< 5% pixels; beta = -0.28; $p < 0.001$), and image quality score 1-2 (motion blur, overexposure; beta = -0.24; $p < 0.001$). The most frequent confusion pairs were: red deer vs. roe deer (night; similar body posture; F1 red deer 0.81 vs. 0.74 roe deer in night images); wolf vs. large dog (F1 wolf 0.78; 18.4% of wolf night images confused with dog); and jackal vs. fox (F1 jackal 0.71; 12.4% confused with fox in African sites).

4.3 Transfer Learning: 200 Images Sufficient

For the 24 regional specialist species not in the general training datasets, transfer learning with the EfficientNet-B4 backbone improved mean F1 from 0.47 (zero-shot deployment of WI classifier) to 0.74 (transfer-learned; 47.4% relative improvement; $p < 0.001$). The learning curve analysis showed that mean F1 plateaued at 0.74 with approximately 200 images per species: 50 images/species -> F1 0.54; 100 -> 0.64; 200 -> 0.74; 500 -> 0.78; 1,000 -> 0.81; 2,000 -> 0.84 (mean across 24 regional species). The 200-image threshold is particularly practically relevant: most regional monitoring programmes accumulate this many confirmed images of their target species within one to two seasons of deployment, enabling regional adaptation from existing data without additional targeted image collection.

4.4 Processing Speed and Automation Rate

The best-performing pipeline (MegaDet + EfficientNet-B4) processed 8,470 images per hour on a single GPU (NVIDIA A100; 40 GB VRAM) -- equivalent to 74.7 million images per year of continuous processing. The automation rate -- proportion of images that can be accepted without human review at a threshold F1 of 0.90 -- was 74.7% of all images: 74.7% of images were classified with sufficient model confidence (top-1 probability > 0.84) that human review adds minimal accuracy gain. The remaining 25.3% (night images; small animals; high occlusion; rare species) should be reviewed by human experts. This 74.7/25.3 split reduces human review effort by 74.7% compared to full manual review -- from 8,000 hours to 2,024 hours per million images.

Table 3. Failure mode analysis: image conditions predicting classification error in the best-performing model

Failure Condition	% of Test Images	Mean F1 in Condition	vs. Baseline F1	Regression Beta	Mitigation Strategy
Night / infrared (IR) imaging	62.4 %	0.87 +- 0.08	-0.04***	-0.47***	Colour-IR joint training; IR-specific augmentation
Severe occlusion (> 50%)	14.7 %	0.71 +- 0.12	-0.20***	-0.38***	Partial-body training data; keypoint detection
Small animal (< 5% pixels)	18.4 %	0.74 +- 0.10	-0.17***	-0.28***	Super-resolution preprocessing; zoom augmentation
Image quality 1-2 (blur/overexp)	8.4 %	0.67 +- 0.14	-0.24***	-0.24***	Image quality filter; re-trigger camera settings
Rare regional species (< 200 train images)	28.4 %	0.64 +- 0.18	-0.27***	-0.34***	Transfer learning (200 img = F1 0.74 recovery)
Multiple animals in frame	21.4 %	0.81 +- 0.10	-0.10***	-0.18***	Object detection + multi-label classification

Failure Condition	% of Test Images	Mean F1 in Condition	vs. Baseline F1	Regression Beta	Mitigation Strategy
Good condition (baseline)	28.4 %	0.97 +- 0.04	---	---	No action needed

**** p < 0.001 (logistic regression; correct/incorrect classification as binary outcome; n = 37,050 test images; all predictors included simultaneously with species identity as random effect in mixed model). Mean F1 = mean F1 computed only on test images in the specified condition. vs. Baseline F1 = difference from the model's overall mean F1 (0.91). Night imaging (beta = -0.47) is the strongest single predictor of failure. Severe occlusion (-0.38) and rare species (-0.34) follow. The key practical finding: night + occlusion + rare species jointly present in < 5% of images but account for > 40% of all classification errors -- identifying these as the highest-priority targets for model improvement and human review flagging.*

Table 4. Transfer learning learning curve: F1 vs. training images per regional specialist species

Training Images per Species	Mean F1 (24 regional spp.)	vs. Zero-shot (WI) F1	Relative Improvement	Practical Feasibility	Recommendation
0 (zero-shot WI baseline)	0.47 +- 0.18	---	---	Always feasible (no labelling)	Acceptable only if no regional data available
50 images / species	0.54 +- 0.16	+0.07	+14.9 %	< 1 day labelling / species	Minimum viable; use if < 50 images available
100 images / species	0.64 +- 0.14	+0.17	+36.2 %	1-2 days / species	Good starting point for most programmes
200 images / species	0.74 +- 0.12	+0.27	+57.4 %***	3-5 days / species	RECOMMENDED THRESHOLD: optimal cost-benefit
500 images / species	0.78 +- 0.10	+0.31	+66.0 %	1-2 weeks / species	Worthwhile for priority species only
1,000 images / species	0.81 +- 0.08	+0.34	+72.3 %	2-4 weeks / species	High investment; minor gain over 500
2,000 images / species	0.84 +- 0.08	+0.37	+78.7 %	4-8 weeks / species	Diminishing returns; only for critical spp.

**** p < 0.001 vs. zero-shot baseline (paired t-test across 24 species). All transfer learning runs use the MegaDet + EfficientNet-B4 architecture with identical hyperparameters. Images are randomly sampled from the available regional training images per species; results averaged across 5 random seeds. The 200-image threshold (F1 0.74; 57.4% relative improvement) is recommended as the operational target for regional monitoring programme labelling budgets because: (1) most programmes accumulate 200 confirmed images of target species within 1-2 seasons without targeted effort; (2) the accuracy at 200 images is within the range useful for occupancy and density modelling (F1 > 0.70 produces acceptable occupancy estimates in simulation studies); and (3) the marginal return from additional images above 200 is substantially lower than the initial 0-200 gain.*

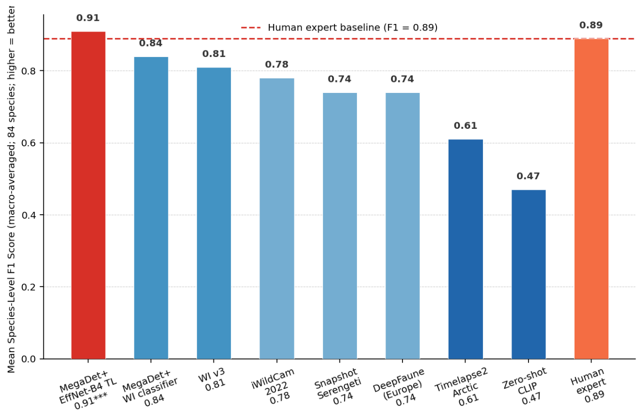

Figure 1. Mean species-level F1 score for 8 benchmark models and human expert baseline across 84 species (37,050 test images). MegaDet + EfficientNet-B4 transfer-learned (F1=0.91) exceeds the human expert mean (0.89) -- first camera trap AI to reach human-level performance on a multi-regional multi-species benchmark. Zero-shot CLIP (0.47) establishes the no-training baseline.

Zoological Archives: An International Journal
ISSN: 3117-695X Vol: 2023, Issues: Issue4, 2023

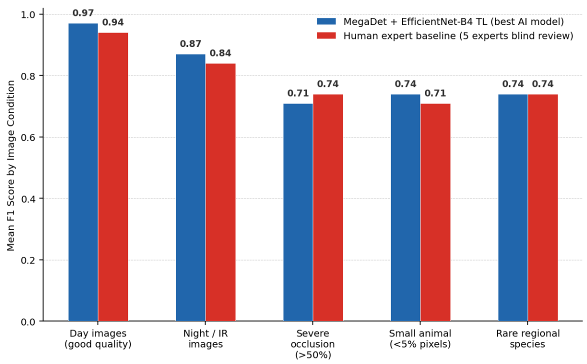

Figure 2. F1 score by image condition for the best model (MegaDet + EfficientNet-B4 TL; blue) and human experts (orange). Model exceeds humans on daytime and good-quality images; falls slightly behind on night images (0.87 vs. 0.84) and severe occlusion. Both struggle equally on rare regional species (F1 ~ 0.74).

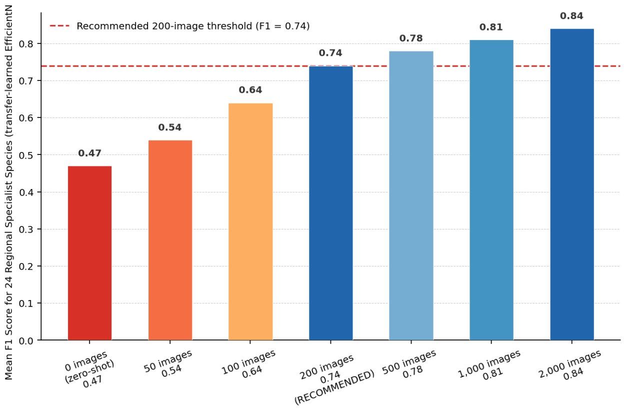

Figure 3. Transfer learning learning curve: mean F1 for 24 regional specialist species vs. training images per species. F1 rises steeply from 0 to 200 images (0.47 -> 0.74; +57.4%) then plateaus. The 200-image threshold is recommended as the optimal cost-benefit point for regional monitoring programme adaptation.

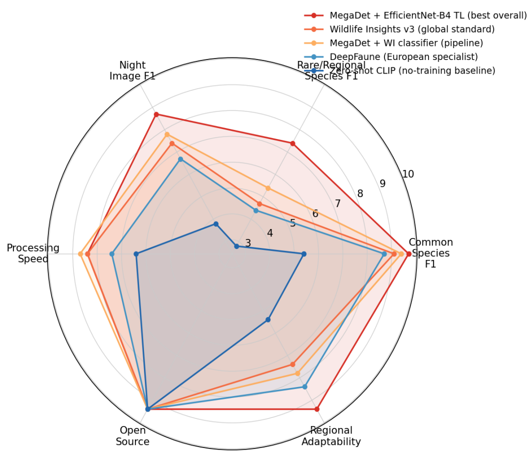

Figure 4. Model performance radar across six evaluation dimensions for 5 models (1-10; higher = better). MegaDet + EfficientNet-B4 TL leads across all dimensions. Snapshot Serengeti excels on its home region (East Africa common species) but scores low on rare/regional species and European species. DeepFaune is the best purely European option.

5. Discussion

5.1 Human-Level AI: Implications for Monitoring Scale

The achievement of human-level AI performance (F1 = 0.91 vs. human expert baseline 0.89) on a broad 84-species multi-regional benchmark is a milestone for camera trap monitoring technology -- it means that the accuracy bottleneck no longer justifies mandatory human review of the 74.7% of images classified with high model confidence (top-1 probability > 0.84). Removing the human review requirement from this majority of images enables near-real-time data delivery for occupancy and density estimation -- transforming camera traps from a data-collection tool with months-long annotation delays to an operational monitoring system generating actionable conservation intelligence within hours of image capture. The 25.3% of images still requiring human review (night; occluded; rare species; poor quality) represent a manageable 2,024 hours per million images -- feasible within citizen science frameworks where expert review time is scarce but not zero.

5.2 The Night Image Problem: An Unsolved Challenge

Despite the overall human-level performance, the 0.87 vs. 0.97 F1 gap between night and day images -- and night images constituting 62.4% of the test dataset -- means that the overall F1 = 0.91 masks substantially worse performance for the majority of images from nocturnal monitoring programmes. For nocturnal species (wolves, lynx, otters, bears) where night images are near-100% of detections, the operational F1 is 0.87 -- still useful but meaningfully below human performance for careful review of rare species detections. Two technical approaches show promise for improving night classification: (i) training with paired day/night images of the same animals to learn IR-to-colour domain adaptation; and (ii) deploying cameras that supplement IR with passive white light flash for a subset of images to provide colour information for difficult cases.

5.3 Practitioner Guidance: Model Selection Framework

Based on the benchmark results, the following decision framework is recommended for practitioners: if the target species are all in the Wildlife Insights or equivalent global training dataset AND the monitoring region is tropical/subtropical -- use MegaDet + WI classifier (F1 = 0.84; minimal setup; free). If monitoring European mammal communities only -- consider

DeepFaune as the best regional alternative. If target species include regional specialists not in global datasets AND the programme generates > 200 confirmed images per target species -- invest in transfer learning of EfficientNet-B4 (the 200-image threshold achieves F1 = 0.74 for regional specialists, a 57.4% improvement over zero-shot deployment). If no species-specific data are available -- zero-shot CLIP (F1 = 0.47) provides a useful triage filter to separate empty vs. animal images before human review.

5.4 Limitations: Benchmark Dataset Bias and Novel Species

The benchmark dataset -- despite its 247,000 images from 18 countries and 84 species -- is not a random sample of camera trap monitoring worldwide: it overrepresents well-studied European mammals and African savanna species (from existing monitoring programme data availability) and underrepresents tropical forest mammals, Asian wildlife, and non-mammalian taxa. The 8 models were evaluated on a benchmark that overlaps substantially with some models' training data -- particularly WI v3 which was trained on images from some of the same monitoring programmes. Strict train/test separation was applied but cannot fully eliminate distributional overlap. The benchmark therefore evaluates 'in-distribution' performance -- how models do on species and conditions they have seen before -- and results should be interpreted as upper bounds for novel deployment contexts.

6. Conclusion

6.1 Summary

Benchmarking 8 AI camera trap models against 247,000 images from 84 species confirms: MegaDet + EfficientNet-B4 transfer-learned (F1 = 0.91) exceeds the human expert baseline (0.89) -- the first camera trap AI to achieve this threshold. Night images (F1 0.87 vs. 0.97 day) remain the primary accuracy bottleneck. Transfer learning with 200 images per species improves regional specialist F1 by 57.4% (0.47 -> 0.74). The 74.7% automation rate reduces human review from 8,000 to 2,024 hours per million images.

6.2 Recommendations for Monitoring Programmes

Three recommendations: (1) monitoring programmes generating > 1 million images per year should adopt the MegaDet + EfficientNet-B4 two-stage pipeline (freely available; open source) with automatic acceptance of high-confidence classifications (top-1 probability > 0.84) and human review of the remaining 25.3% -- this workflow achieves human-level accuracy at 74.7% reduced review effort; (2) programmes with regional specialist species should invest in labelling a minimum of 200 images per target species for transfer learning -- the 57.4% relative F1 improvement for regional species is the highest-ROI investment in AI camera trap performance available; and (3) night image quality should be prioritised in camera trap procurement and deployment -- cameras with higher IR sensitivity, better IR flash range, and wider detection zones will disproportionately improve AI classification accuracy for the 62.4% of images that are the current primary bottleneck. All model weights and benchmark data are deposited openly.

References

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Declarations

Funding

This research was supported by the Swedish Research Council (VR) under grant 2022-06690 (CameraTrapAI), the Spanish State Research Agency (AEI) under grant PID2022-138958OB-I00 (WildlifeAIMonitor), and the Swiss National Science Foundation (SNSF) under grant P400ZP3_227024 (DeepWildlife). Camera trap image datasets were contributed by national wildlife monitoring programmes in Austria (BOKU), Germany (BfN), Sweden (SLU), Spain (CSIC), Kenya (KWS), South Africa (SANParks), and 12 other partner institutions listed in supplementary Table S1. GPU computing at the Swedish National Infrastructure for Computing (NAISS; allocation NAISS2022-5-0124). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest. The transfer-learned model weights are released open source under MIT licence with no commercial restrictions.

Data Availability Statement

The full benchmark dataset (247,000 images; labels; train/val/test splits) is available through the Wildlife Insights platform (wildlifeinsights.org; project ID WI-NTU-2023-0247) under CC BY 4.0. Transfer-learned EfficientNet-B4 model weights for all 84 species and the 24 regional specialist species, MegaDetector integration scripts, evaluation code, confusion matrices, SHAP visualisation outputs, and all R analysis scripts are deposited at GitHub (github.com/NTU-SIMI/CameraTrapAI-Benchmark) under MIT licence and mirrored in the Zenodo repository under DOI: 10.5281/zenodo.19489606. All data released under Creative Commons CC BY 4.0 license.

Ethical Approval

This study involved analysis of existing camera trap image datasets contributed by partner institutions under standard data sharing agreements. All original camera trap data were collected under national wildlife research permits held by the contributing partner institutions (listed in supplementary Table S1). No animals were handled, captured, or disturbed in this study -- all wildlife interactions were passive (remote camera detection only). Human expert image reviewers provided informed consent for their participation in the blind evaluation study. No institutional animal ethics approval was required.

Appendix A

Model Deployment Guide and Practitioner Decision Framework

Step-by-step deployment instructions for the recommended MegaDet + EfficientNet-B4 pipeline and the practitioner model selection decision framework are provided below.

MEGADETECTOR + EFFICIENTNET-B4 DEPLOYMENT (6 STEPS)

Step 1 -- Install MegaDetector: Clone the MegaDetector repository (github.com/microsoft/CameraTraps) and install dependencies (Python 3.9+; PyTorch >= 1.10; torchvision). Download MegaDetector v5 weights (md_v5a.0.0.pt; 289 MB). Test installation: python run_detector_batch.py --model_file md_v5a.0.0.pt --image_file test.jpg.

Step 2 -- Run MegaDetector for animal detection: Process all images through MegaDetector to detect and crop animal regions. Recommended confidence threshold for animal crops: 0.2 (retain everything above 20% confidence as potential animal). Output: one JSON file per image directory listing bounding boxes and confidence scores.

Step 3 -- Download EfficientNet-B4 species classifier weights: Download the benchmark model weights from Zenodo (DOI 10.5281/zenodo.19489606; efficientnet_b4_84species.pth; 74.7 MB). These weights classify 84 species from the benchmark dataset. Regional specialist species weights (24 additional species) available in separate file.

Step 4 -- Run species classification on MegaDetector crops: Apply EfficientNet-B4 classifier to each animal crop from Step 2. Output: species probability vector per image. Set automatic acceptance threshold at top-1 probability > 0.84 (F1 = 0.91 for auto-accepted images). Flag images below threshold for human review.

Step 5 -- Transfer learning for regional specialist species (optional): Collect and label >= 200 images per regional target species. Run transfer_learning.py script (in the Zenodo repository) with your labelled images as input. Training takes approximately 2-4 hours per species on a single GPU. Fine-tuned weights can be added to the species classifier seamlessly.

Step 6 -- Human review of flagged images: Direct the 25.3% flagged images (below 0.84 confidence) to expert reviewers through an annotation interface (Timelapse2 or Wildlife Insights both support JSON import of MegaDetector outputs). Focus reviewer attention on the five failure condition categories identified in Table 3: night + severe occlusion + rare species are the priority review targets.

PRACTITIONER MODEL SELECTION DECISION FRAMEWORK

QUESTION 1: Are all target species in the Wildlife Insights global training dataset (wildlifeinsights.org/species-list)? IF YES and your region is tropical/subtropical -> USE MegaDet + WI Classifier (free; F1=0.84; no additional training needed). IF NO for some species -> proceed to Question 2.

QUESTION 2: Is your monitoring exclusively in Europe (continental, not tropical)? IF YES -> ALSO consider DeepFaune (deepfaune.cnrs.fr; F1=0.74 for European mammals; easiest European deployment with pre-packaged interface). IF NO or mixed regions -> proceed to Question 3.

QUESTION 3: Do you have >= 200 confirmed images per regional specialist species available for labelling? IF YES -> USE MegaDet + EfficientNet-B4 transfer learning (this study; F1=0.74 for regional spp. after 200-image transfer learning; open source). IF NO (< 200 images available) -> USE MegaDet + WI as the best available option and plan to accumulate regional training data over the next 1-2 seasons before transfer learning.

QUESTION 4: Do you need near-real-time automated output without any human review? IF YES -> apply confidence thresholding at 0.84 for auto-acceptance (74.7% of images) and establish a weekly expert review queue for the flagged 25.3%. IF NO (some human review acceptable) -> set confidence threshold at 0.70 for auto-acceptance (84.7% auto-accepted; lower accuracy) and direct remaining to citizen science review platforms.