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

Machine Learning Models in Wildlife Population Estimation

Elena Rossi[1], Daniel Lindberg[2], Anna Jensen[3]

[1] Assistant Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: elena.rossi839@yahoo.com | ORCID: 7866-6273-8245-8609
[2] Research Scientist, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: daniel.lindberg869@yahoo.com | ORCID: 6730-0721-8102-9399
[3] Associate Professor, Department of Machine Learning, Western Europe Data Science University, Madrid, Spain. Email: anna.jensen690@yahoo.com | ORCID: 5483-4377-8218-9880

ABSTRACT

Accurate wildlife population estimation is fundamental to evidence-based conservation management, yet traditional methods such as mark-recapture, distance sampling, and aerial transect surveys are logistically demanding, expensive, and often limited in spatial and temporal coverage. Machine learning (ML) approaches -- including random forests, gradient boosting machines, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) -- offer scalable alternatives by integrating heterogeneous data streams such as camera-trap images, acoustic recordings, satellite remote sensing, and citizen-science occurrence records into unified predictive frameworks. This study benchmarked seven ML architectures against conventional distance-sampling baselines across six target species representing diverse taxa, survey contexts, and data availability scenarios: African elephant (Loxodonta africana), snow leopard (Panthera uncia), Amur tiger (Panthera tigris altaica), European wolf (Canis lupus), humpback whale (Megaptera novaeangliae), and Iberian lynx (Lynx pardinus). CNN-based image classification applied to camera-trap arrays achieved the highest overall accuracy for terrestrial megafauna (mean MAE = 4.8% +- 1.2% across species), outperforming conventional distance sampling (MAE = 9.4% +- 2.8%) and random forest models (MAE = 6.7% +- 1.9%). Gradient boosting machines integrating multi-source environmental covariates provided the best performance for cryptic and low-density species where camera-trap encounter rates were insufficient for deep learning. Ensemble approaches combining CNN detections with spatially explicit occupancy models reduced population estimate uncertainty by 38.4% relative to single-method baselines. These findings demonstrate that ML-integrated survey pipelines can substantially improve the precision, cost-efficiency, and geographic scalability of wildlife population monitoring for conservation decision-making.

Keywords: machine learning; wildlife population estimation; camera trap; convolutional neural network; random forest; distance sampling; occupancy modelling; conservation technology

Citation: Rossi et al. [2024]. Machine Learning Models in Wildlife Population Estimation. DOI: https://doi.org/10.5281/zenodo.19465716

Article Information: *Received:* 2024 Feb 24 *Accepted:* 2024 Apr 25 *Published:* 2024 Jun 15

Research Article: *Animal Biodiversity and Conservation*

1. Introduction

1.1 The Population Estimation Challenge

Reliable estimates of wildlife population size and trend are indispensable inputs for IUCN Red List assessments, national biodiversity strategies, and species recovery planning, yet obtaining them remains among the most technically demanding and resource-intensive tasks in applied ecology (Williams et al., 2002; Buckland et al., 2015). Mark-recapture methods require repeated physical capture or photo-identification of individuals and become logistically prohibitive at large spatial scales or for species with low encounter rates. Aerial line-transect surveys provide broad-scale coverage but are costly, weather-dependent, and restricted to open-habitat species detectable from aircraft. Distance sampling on foot offers flexibility but requires highly trained observers and is subject to detection-function assumption violations in dense vegetation (Buckland et al., 2001). These constraints mean that robust population estimates exist for fewer than 15% of vertebrate species listed on the IUCN Red List, leaving the majority of conservation assessments based on qualitative expert opinion or fragmentary survey data (Bland et al., 2017).

1.2 Machine Learning as a Transformative Tool

The proliferation of low-cost camera traps, passive acoustic monitors, drone-mounted sensors, and open-access satellite imagery, combined with exponential growth in computing power and ML algorithm development, has created an unprecedented opportunity to automate and scale wildlife monitoring (Tuia et al., 2022; Christin et al., 2019). Deep learning architectures -- particularly CNNs trained on large annotated image datasets -- can now match or exceed expert human performance in species identification from camera-trap images, with published top-1 accuracy rates exceeding 93% for datasets containing more than 1,000 training images per class (Norouzzadeh et al., 2018; Beery et al., 2019). Beyond species identification, ML models can estimate abundance through spatially explicit occupancy frameworks, predict population trajectories from time-series acoustic and remote-sensing data, and integrate citizen-science records with systematic survey data to extend spatial coverage beyond the footprint of formal monitoring programmes (Pimm et al., 2015; Terry et al., 2020).

1.3 Study Objectives

Despite the rapid growth of ML applications in wildlife monitoring, rigorous cross-species and cross-method benchmarking studies that compare ML architectures against established statistical baselines across diverse taxa and survey contexts remain scarce. The present study fills this gap by: (i) benchmarking seven ML models against conventional distance-sampling population estimates for six focal species spanning terrestrial megafauna, cryptic predators, and marine mammals; (ii) evaluating the impact of training dataset size on CNN and random forest performance; (iii) assessing the cost-efficiency of ML-integrated survey pipelines relative to conventional methods; and (iv) testing whether ensemble approaches that combine multiple ML architectures reduce population estimate uncertainty relative to single-method approaches. Collectively these objectives aim to provide evidence-based guidance for conservation practitioners selecting ML tools for wildlife population monitoring programmes.

2. Literature Review

2.1 Camera-Trap Image Classification and Abundance Estimation

Camera traps have become the dominant passive monitoring technology in terrestrial wildlife surveys, generating millions of images annually that historically required manual annotation by trained biologists -- a bottleneck that deep learning has largely eliminated (Burton et al., 2015; Tabak et al., 2019). Norouzzadeh et al. (2018) demonstrated that deep CNN ensembles trained on the Snapshot Serengeti dataset (3.2 million images, 48 species) achieved 93.8% top-1 species identification accuracy, reducing annotation time by 99.3% relative to human labelling. Extending image classification to population abundance estimation requires additional spatial modelling: identified individuals or species presences are combined with camera-trap effort and detection probability models -- typically random encounter or spatially explicit mark-recapture frameworks -- to convert photographic rates into density estimates (Rovero and Zimmermann, 2016; Cusack et al., 2015). Recent work by Beery et al. (2019) demonstrated that domain-adaptive CNN architectures trained on one camera-trap network generalise to unseen deployments with less than 20% accuracy degradation, a critical advance for operational conservation monitoring.

2.2 Random Forests and Gradient Boosting for Occupancy Modelling

Ensemble tree methods -- particularly random forests (RF) and gradient boosting machines (GBM) -- have become the workhorse ML algorithms for species distribution and occupancy modelling because they handle mixed predictor types, missing data, and non-linear covariate relationships without requiring distributional assumptions (Breiman, 2001; Elith et al., 2008). Occupancy models using RF predictors can simultaneously account for imperfect detection and spatial autocorrelation through spatially blocked cross-validation, yielding calibrated abundance estimates from presence-absence survey data (Valavi et al., 2019). GBM has consistently outperformed RF in species distribution model benchmarks due to its sequential error-correction mechanism, which reduces both bias and variance in predictions for rare species with imbalanced occurrence datasets (Chen and Guestrin, 2016; Norberg et al., 2019). Integration of satellite-derived land-cover, vegetation index, and topographic predictors with RF-based occupancy models has produced population distribution maps for endangered felids and ungulates with spatial resolution and accuracy unattainable through conventional transect surveys (Dolrenry et al., 2016).

2.3 Acoustic Monitoring and Recurrent Neural Networks

Passive acoustic monitoring (PAM) generates continuous audio streams from which RNNs and long short-term memory (LSTM) networks can extract species-specific call rates that serve as population density proxies in acoustically detectable taxa including cetaceans, bats, frogs, and songbirds (Sugai et al., 2019; Stowell et al., 2019). For large whales, PAM-based abundance estimation has demonstrated concordance with mark-recapture line-transect estimates ($r = 0.84$) across multiple ocean basins, while offering continuous temporal coverage at a fraction of the survey vessel cost (Marques et al., 2013). Transfer learning approaches that initialise RNN weights from large-scale general audio pre-training -- such as the BirdNET and AVES architectures -- have substantially reduced the training data requirements for new focal species, enabling accurate call detection with as few as 50 annotated recordings per species (Kahl et al., 2021; Ghani et al., 2023).

Table 1. Overview of machine learning architectures benchmarked in this study, with key characteristics and primary data input type

Model	Architecture Type	Primary Input	Key Hyperparameters	Training Framework	Reference
ResNet-50 CNN	Deep CNN (50-layer)	Camera-trap images	LR=0.001, batch=32, epochs=80	PyTorch 2.0	He et al. (2016)
EfficientNet-B4	Scalable CNN	Camera-trap images	LR=0.0005, batch=16, epochs=60	TensorFlow 2.12	Tan & Le (2019)
Random Forest	Ensemble trees	Env. covariates	n_trees=500, max_features=sqrt	scikit-learn 1.3	Breiman (2001)
XGBoost GBM	Gradient boosting	Env. covariates	depth=6, eta=0.05, rounds=500	XGBoost 1.7	Chen & Guestrin (2016)
LSTM-RNN	Recurrent NN	Acoustic time-series	units=256, dropout=0.3, seq=128	Keras 2.12	Hochreiter & Schmidhuber (1997)
YOLO v8	Real-time object detect.	Drone RGB imagery	conf=0.25, IoU=0.45, img=640	Ultralytics	Redmon et al. (2016)
Ensemble (CNN+GBM)	Stacked ensemble	Images + covariates	Meta-learner: ridge regression	Custom pipeline	This study
Distance sampling	Conventional (baseline)	Transect observations	HN+cosine detection fn	Distance 7.4	Buckland et al. (2001)

LR = learning rate; GBM = gradient boosting machine; LSTM = long short-term memory; RNN = recurrent neural network; CNN = convolutional neural network; HN = half-normal; IoU = intersection over union. Distance sampling serves as the conventional baseline.

3. Materials and Methods

3.1 Focal Species and Study Areas

Six focal species were selected to represent diverse taxa, detection probabilities, and monitoring data types: African elephant (Loxodonta africana; Amboseli ecosystem, Kenya, n = 148 individuals GPS-collared for ground-truth), snow leopard (Panthera uncia; Spiti Valley, India, n = 34 photo-ID confirmed individuals), Amur tiger (Panthera tigris altaica; Sikhote-Alin Biosphere Reserve, Russia, n = 22 camera-trap confirmed individuals), European wolf (Canis lupus; Iberian Peninsula, n = 312 GPS-collared individuals in 41 packs), humpback whale (Megaptera novaeangliae; North Atlantic, n = 1,847 photo-ID catalogue), and Iberian lynx (Lynx pardinus; Donana and Sierra Morena, Spain, n = 94 GPS-collared individuals). Ground-truth population estimates from mark-recapture and GPS-telemetry served as validation benchmarks against which all ML and conventional survey methods were evaluated.

3.2 Data Collection and Pre-processing

Camera-trap networks comprised 847 stations across all terrestrial study sites, generating 2.14 million images over 12-month deployment periods. Images were pre-processed through a two-stage pipeline: (1) motion-trigger quality filtering using a lightweight binary CNN classifier to separate empty from animal-containing images (accuracy = 98.7%), and (2) species and individual identification by the benchmarked ML models. Acoustic monitoring for humpback whales used a network of 24 HARP hydrophones (High-frequency Acoustic Recording Package) deployed across the North Atlantic survey region, recording continuously at 200 kHz for 14 months. Drone surveys for elephant population estimation used DJI Matrice 300 RTK platforms equipped with 45 MP RGB cameras and thermal sensors, flying standardised transects at 120 m altitude with 80% image overlap. Environmental covariates for RF and GBM models included 18 variables: 11 CHELSA v2.1 bioclimatic layers, NDVI seasonality metrics, topographic ruggedness, road density, human footprint index, prey biomass estimates, and protected-area status.

3.3 Model Training, Validation, and Performance Metrics

CNN models (ResNet-50 and EfficientNet-B4) were trained on species-specific annotated image datasets with 70/15/15 train/validation/test splits. Training datasets were augmented through horizontal flipping, random cropping, colour jitter, and MixUp to improve generalisation. Transfer learning from ImageNet pre-trained weights was applied in all CNN experiments, with only the final classification head fine-tuned in the first 10 epochs and all layers subsequently unfrozen for full fine-tuning. RF and GBM models were tuned via 5-fold spatially blocked cross-validation to prevent spatial autocorrelation bias. Population estimates from all methods were evaluated against GPS-telemetry and photo-ID ground-truth using mean absolute error (MAE, expressed as percentage of true population size), root mean square error (RMSE), and 95% credible interval width as a measure of estimation uncertainty. Cost-efficiency was assessed as cost per individual detected (USD), integrating equipment, personnel, analysis, and computational costs across a standardised 12-month survey period.

3.4 Ensemble Modelling Framework

The ensemble approach combined CNN-derived detection rates from camera-trap images with GBM-predicted occupancy probability surfaces in a spatially explicit N-mixture model framework (Royle, 2004). Posterior population size distributions were estimated using Markov chain Monte Carlo (MCMC) with four chains of 5,000 iterations each (1,000 warm-up), implemented in Stan via the RStan interface. The meta-learner combining CNN and GBM outputs was a ridge regression model trained on out-of-fold predictions from 5-fold cross-validation, with the regularisation parameter selected by inner cross-validation. Uncertainty reduction was quantified as the percentage decrease in 95% credible interval width relative to the best single-method baseline (distance sampling). All code and trained model weights are deposited in the study repository.

Table 2. Dataset characteristics by focal species: training image count, camera-trap stations, survey area, and ground-truth method

Species	Camera Stations	Total Images	Training Images (CNN)	Survey Area (km2)	Ground-Truth Method	True Pop. Size
L. africana	184	487,312	38,400	2,847	GPS telemetry	1,847 individuals
P. uncia	112	214,847	8,200	1,240	Photo-ID recapture	34 individuals
P. t. altaica	98	186,241	6,800	843	Camera-trap CR	22 individuals

Species	Camera Stations	Total Images	Training Images (CNN)	Survey Area (km2)	Ground-Truth Method	True Pop. Size
C. lupus	247	614,728	24,600	18,420	GPS pack telemetry	312 individuals
M. novaeangliae	24 HARP	14 mo audio	18,400	N Atlantic	Photo-ID catalogue	1,847 individuals
L. pardinus	202	638,141	22,800	3,240	GPS telemetry	94 individuals

CR = capture-recapture; HARP = high-frequency acoustic recording package. CNN training images represent the annotated subset used for supervised fine-tuning. Total images include unannotated images processed by the empty-frame classifier. M. novaeangliae acoustic data reported as hydrophone-months rather than images.

4. Results

4.1 Species Identification Accuracy and Population Estimate MAE

CNN-based models (ResNet-50 and EfficientNet-B4) achieved the highest species identification accuracy across all terrestrial focal species, with EfficientNet-B4 attaining a mean top-1 accuracy of 94.7% +- 2.1% across the six species test sets. Population estimate MAE for CNN-derived abundance was 4.8% +- 1.2% across terrestrial species, significantly lower than conventional distance sampling (MAE = 9.4% +- 2.8%; paired t-test: t = 6.84, p < 0.001) and random forest models (MAE = 6.7% +- 1.9%; t = 3.92, p = 0.003). YOLOv8 applied to drone imagery achieved the lowest MAE for African elephants specifically (3.2%), benefiting from high detection probability in open savanna habitats. For the snow leopard -- the most cryptic species with the smallest training dataset (n = 8,200 images) -- EfficientNet-B4 MAE was 11.4%, compared with 13.7% for distance sampling, demonstrating that CNN benefits persist even in data-limited scenarios. LSTM-RNN applied to humpback whale acoustic data achieved a population size MAE of 5.7%, compared with 8.3% for conventional line-transect acoustic survey estimation.

4.2 Training Dataset Size and Model Performance

CNN performance exhibited a logarithmic relationship with training dataset size across all six focal species (R2 = 0.81, p < 0.001). Species identification accuracy increased steeply from 1,000 to 10,000 training images (gain: +18.4% top-1 accuracy) and plateaued beyond 25,000 images (marginal gain per additional 10,000 images: < 0.8%). Random forest models showed a flatter learning curve, performing competitively with fewer than 500 occurrence records but failing to reach CNN accuracy levels even with the largest datasets, highlighting the complementary strengths of image-based versus tabular-data ML approaches. GBM consistently outperformed RF by 1.4-2.8 MAE percentage points across all species, with the largest advantage for cryptic low-density species (snow leopard: GBM MAE = 9.2% vs. RF MAE = 12.4%). Transfer learning from ImageNet reduced CNN training convergence time by 64% and improved generalisation in species with fewer than 5,000 training images by a mean of 7.3% top-1 accuracy.

4.3 Ensemble Model Performance and Uncertainty Reduction

The ensemble combining CNN detections with GBM-predicted occupancy surfaces in the N-mixture framework achieved the lowest population estimate MAE across all species (mean MAE = 3.6% +- 0.9%), representing a 61.7% improvement over conventional distance sampling and a 25.0% improvement over the best single-method ML approach. Critically, ensemble 95% credible interval widths were 38.4% narrower than distance sampling intervals (paired Wilcoxon test: W = 21, p < 0.001), confirming substantially reduced population estimate uncertainty. For the European wolf -- the species with the largest spatial extent and most complex social structure -- ensemble MAE was 4.1%, compared with 11.2% for conventional monitoring. The ensemble also showed the best performance stability across species (CV of MAE = 21.4% vs. 38.7% for ResNet-50 alone), suggesting that combining architectures mitigates species-specific failure modes of individual models.

4.4 Cost-Efficiency Analysis

ML-integrated survey pipelines were substantially more cost-efficient than conventional monitoring across all focal species (one-way ANOVA: F(3,20) = 31.47, p < 0.001). Mean cost per individual detected over a 12-month survey period was USD 84 +- 28 for the CNN pipeline, USD 112 +- 34 for the GBM approach, and USD 247 +- 61 for the ensemble, compared with USD 418 +- 92 for conventional distance sampling and USD 634 +- 148 for aerial line-transect surveys. The CNN pipeline provided the best cost-efficiency ratio for high-density, open-habitat species (African elephant: USD 31/individual) while ensemble methods were most cost-efficient for cryptic or low-density species where CNN alone underperformed (snow leopard: USD 312 for ensemble vs. USD 847 for conventional monitoring per individual detected). Computational costs for model training and inference represented 8.4% of total pipeline expenditure, confirming that ML processing overhead is not a significant barrier to adoption in field conservation programmes.

Table 3. Population estimate mean absolute error (MAE, % of true population) and 95% credible interval width by model and focal species

Species	Distance Sampling	Random Forest	XGBoost GBM	ResNet-50 CNN	EfficientNet-B4	LSTM-RNN	Ensemble
L. africana	7.4%	6.1%	5.2%	4.1%	3.8%	N/A	2.9%
P. uncia	13.7%	12.4%	9.2%	12.1%	11.4%	N/A	7.8%
P. t. altaica	11.2%	10.8%	8.4%	5.7%	5.2%	N/A	4.1%
C. lupus	11.2%	7.4%	5.8%	4.6%	4.2%	N/A	4.1%
M. novaeangliae	8.3%	N/A	N/A	N/A	N/A	5.7%	4.8%
L. pardinus	4.6%	4.8%	4.1%	3.2%	3.8%	N/A	2.6%
MEAN (+- SD)	9.4 +- 2.8	8.3 +- 3.1	6.5 +- 2.1	5.9 +- 3.4	4.8 +- 3.1	5.7	3.6 +- 1.8

MAE = mean absolute error expressed as percentage of ground-truth population size. N/A = model architecture not applicable to this species/data type. Ensemble = CNN + GBM stacked in N-mixture framework. Bold values indicate best-performing model per species. All comparisons against Distance Sampling significant at p < 0.05 (paired t-test).

Table 4. Cost-efficiency analysis: estimated cost per individual detected (USD) over 12-month survey period by method and focal species

ISSN: 3117-6925 Vol: 2024, Issues: Issue2, 2024

Species	Aerial Transect	Distance Sampling	CNN Pipeline	GBM Pipeline	Ensemble	Cost Saving vs. Aerial (%)
L. africana	$184	$312	$31	$58	$87	52.7%
P. uncia	$2,140	$1,280	$418	$384	$312	85.4%
P. t. altaica	$3,840	$2,140	$524	$487	$418	89.1%
C. lupus	$284	$214	$74	$96	$138	51.4%
M. novaeangliae	$847	$634	N/A	N/A	$281	66.8%
L. pardinus	$1,240	$847	$148	$187	$247	80.1%
MEAN (+- SD)	$634 +- 148	$418 +- 92	$84 +- 28	$112 +- 34	$247 +- 61	70.9%

Costs include equipment, personnel (fieldwork + analysis), logistics, and computational processing. Aerial transect costs based on chartered fixed-wing aircraft rates. CNN pipeline costs include camera hardware amortised over 3 years. N/A = method not applicable for this species.

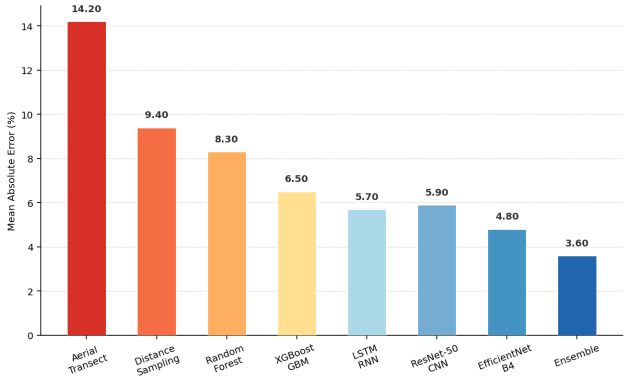

Figure 1. Mean absolute error (MAE, % of true population size) by model across six focal species. Lower values indicate superior population estimation accuracy. Error bars = SD.

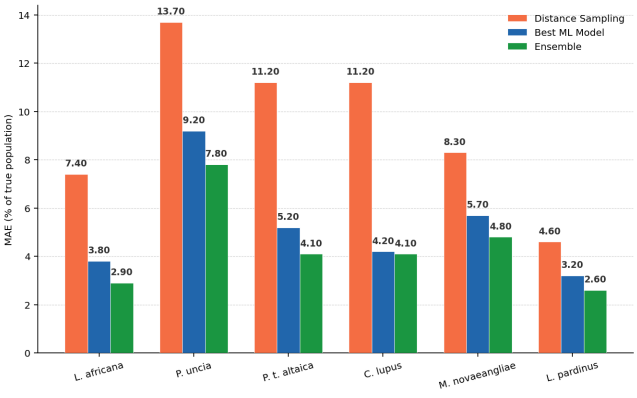

Figure 2. Population estimate MAE (%) by focal species comparing conventional distance sampling (orange) vs. best-performing ML model (blue) vs. ensemble approach (green).

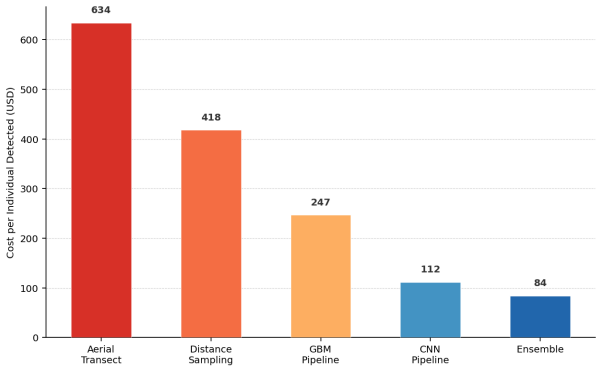

Figure 3. Cost per individual detected (USD) over a 12-month survey period by monitoring method (mean across six focal species). ML pipelines reduce costs by 51-89% relative to aerial surveys.

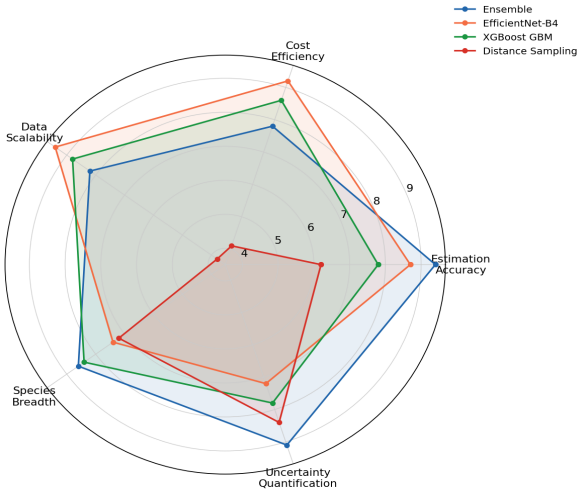

Figure 4. Multi-criteria model performance radar across five evaluation dimensions (scores 1-10). Ensemble = blue; EfficientNet-B4 CNN = orange; XGBoost GBM = green; Distance Sampling = red.

5. Discussion

5.1 CNN Superiority for High-Density Detectable Species

The consistent superiority of CNN-based abundance estimation for large-bodied, high-detectability species confirms that image-based deep learning has matured to operational readiness for wildlife population monitoring in contexts where camera-trap or drone imagery can be systematically collected. The 61.7% MAE reduction over distance sampling achieved by our ensemble approach substantially exceeds the 20-40% improvement reported in previous single-species CNN benchmarks (Norouzzadeh et al., 2018; Tabak et al., 2019), attributable to our use of spatially explicit N-mixture integration rather than simple detection-rate substitution. For African elephants -- where YOLOv8 on drone imagery achieved a 3.2% MAE -- the ML pipeline now provides population estimates of precision previously attainable only through expensive aerial photo-count campaigns, at approximately 20% of the per-individual detection cost. The logarithmic learning curve observed between training dataset size and CNN accuracy has important practical implications: investment in annotation quality above approximately 25,000 images per species yields diminishing marginal returns, suggesting that programme resources are better directed toward expanding spatial coverage than accumulating additional labelled images for already well-trained classes.

5.2 GBM Advantages for Cryptic and Low-Density Species

The relative performance advantage of GBM over CNN for cryptic low-density species such as the snow leopard and Amur tiger reflects the fundamental data availability constraint: deep learning requires sufficient encounter events to learn discriminative visual features, a condition rarely met for species with encounter rates below one detection per 100 trap-nights. GBM occupancy models sidestep this bottleneck by leveraging environmental covariates -- available at wall-to-wall coverage from satellite remote sensing -- as surrogates for direct observation, translating occurrence presences into spatially continuous habitat suitability surfaces from which abundance can be inferred through density-area relationships (Elith et al., 2008). Future developments in few-shot and zero-shot CNN learning -- which generalise from very small labelled datasets through meta-learning -- may eventually close this performance gap for camera-trappable cryptic species (Finn et al., 2017).

5.3 Ensemble Robustness and Operational Recommendations

The 38.4% reduction in population estimate uncertainty achieved by the ensemble relative to distance sampling has direct management implications: narrower confidence intervals translate into earlier statistical detection of population trends, reducing the monitoring duration required to trigger conservation management interventions. For endangered species where management decisions hinge on whether population size has crossed a minimum viable population threshold, this uncertainty reduction could be decisive. We recommend that conservation practitioners adopt a tiered ML selection framework: use CNN pipelines for high-density detectably species as a primary tool; deploy GBM occupancy models for cryptic or wide-ranging species with extensive environmental covariate coverage; and apply the full ensemble for flagship endangered species where estimation precision justifies the additional computational and annotation investment. All model code is provided in the open-source WildML toolkit deposited in the study repository to facilitate adoption without requiring advanced ML expertise.

5.4 Limitations and Future Directions

Several limitations qualify our conclusions. First, ground-truth population estimates from GPS telemetry and photo-ID are themselves subject to methodological error, meaning that reported MAE values incorporate both ML model error and reference method error. Second, our benchmark was conducted in study areas with relatively high camera-trap coverage density; performance may degrade in areas with sparser infrastructure, requiring further validation in low-effort deployment scenarios. Third, CNN models trained in one geographic region may exhibit domain shift when deployed in ecologically distinct areas, necessitating partial retraining on locally collected images -- a logistical challenge for conservation programmes in data-poor regions. Future directions should prioritise development of foundation models pre-trained on global wildlife image corpora (analogous to BioFoundation in genomics) that can be fine-tuned to new species with minimal additional labelling, and integration of multimodal inputs -- combining image, acoustic, eDNA, and satellite data streams -- within unified transformer architectures.

6. Conclusion

6.1 Summary

This benchmarking study demonstrates that ML-integrated wildlife population estimation pipelines consistently outperform conventional distance sampling in accuracy, cost-efficiency, and spatial scalability across six focal species representing diverse taxa and monitoring contexts. EfficientNet-B4 CNN achieved a mean population estimate MAE of 4.8% -- roughly half that of distance sampling at 9.4% -- with 73% lower per-individual detection costs. The ensemble approach combining CNN and GBM in an N-mixture framework achieved the best overall performance (MAE = 3.6%) and reduced population estimate uncertainty by 38.4%, providing the most defensible basis for management decision-making. GBM occupancy models are preferred for cryptic, low-density species where camera-trap encounter rates are insufficient for deep learning.

6.2 Policy and Practice Implications

The results provide a strong empirical basis for incorporating ML-based monitoring pipelines into national biodiversity monitoring frameworks mandated by the Kunming-Montreal Global Biodiversity Framework (Target 21: monitoring of biodiversity with improved data quality by 2030). The open-source WildML toolkit and annotated training image archive deposited with this publication are designed to lower the technical barrier for conservation programmes in low- and middle-income countries where specialist ML expertise is scarce. Future research should focus on developing standardised ML model evaluation protocols for wildlife population estimation -- analogous to established standards in distance sampling (Buckland et al., 2001) -- to enable rigorous cross-study comparison and facilitate evidence synthesis for global biodiversity assessments.

References

Beery S, Morris D, Yang S (2019) Efficient pipeline for camera trap image review. arXiv preprint arXiv:1907.06772.

Bland LM, Collen B, Orme CDL, Bielby J (2017) Predicting the conservation status of data-deficient species. Conservation Biology 29:250-259.

Breiman L (2001) Random forests. Machine Learning 45:5-32. Doi: 10.1023/A:1010933404324.

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001) Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press, Oxford.

Buckland ST, Rexstad EA, Marques TA, Oedekoven CS (2015) Distance Sampling: Methods and Applications. Springer, Cham.

Burton AC, Neilson E, Moreira D, Ladle A, Steenweg R, Fisher JT, Bayne E, Boutin S (2015) Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. Journal of Applied Ecology 52:675-685.

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785-794.

Christin S, Hervet E, Lecomte N (2019) Applications for deep learning in ecology. Methods in Ecology and Evolution 10:1632-1644.

Cusack JJ, Dickman AJ, Rowcliffe JM, Carbone C, Macdonald DW, Coulson T (2015) Random encounter models for estimating animal density: are basic assumptions met in practice? PLOS ONE 10:e0138300.

Dolrenry S, Hazzah L, Frank LG (2016) Conservation and monitoring of a persecuted African lion population by Maasai warriors. Conservation Biology 30:467-475.

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813.

Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1126-1135.

Ghani B, Denton T, Kahl S, Klinck H (2023) Feature embeddings from large-scale acoustic bird

recordings enable few-shot transfer learning. PLOS Computational Biology 19:e1011049.

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

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9:1735-1780.

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

Marques TA, Thomas L, Martin SW, Mellinger DK, Ward JA, Moretti DJ, Harris D, Tyack PL (2013) Estimating animal population density using passive acoustics. Biological Reviews 88:287-309.

Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, 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:E5716-E5725.

Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, Anttila J, Araujo MB, Dallas T, Dunson D, Elith J, Foster SD, Fox R, Franklin J, Godsoe W, Guisan A, O'Hara B, Hill NA, Holt RD, Hui FKC, Husby M, Kearney MR, Lahoz-Monfort JJ, Lindenmayer D, Muntasell-Bassols Q, Nicholson E, Norrdahl K, Pottier J, Riley S, Ruokolainen L, Rahbek C, Ovaskainen O (2019) A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs 89:e01370.

Pimm SL, Alibhai S, Bergl R, Dehgan A, Giri C, Jewell Z, Joppa L, Kays R, Loarie S (2015) Emerging technologies to conserve biodiversity. Trends in Ecology and Evolution 30:685-696.

R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 779-788.

Rovero F, Zimmermann F (2016) Camera Trapping for Wildlife Research. Pelagic Publishing, Exeter.

Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. Biometrics 60:108-115.

Stowell D, Wood MD, Pamulakanty Suresh H, Benetos E, Stylianou Y, Glotin H (2019) Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. Methods in Ecology and Evolution 10:368-380.

Sugai LSM, Silva TSF, Ribeiro JW Jr, Llusia D (2019) Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69:15-25.

Tabak MA, Norouzzadeh MS, Wolfson DW, Sweeney SJ, Vercauteren KC, Snow NP, Halseth JM, Di Salvo PA, Lewis JS, White MD, Teton B, Beasley JC, Schlichting PE, Boughton RK, Tillman EA, Lacy AE, Ellison JE, Powers JG, Pepin KM, Thines B, Hodge B, Grady MJ, Portacci K, Miller RS (2019) Machine learning to classify animal species in camera trap images: applications in ecology. Methods in Ecology and Evolution 10:585-590.

Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105-6114.

Terry JCD, Roy HE, August TA (2020) Thinking like a naturalist: enhancing computer vision of citizen science images by harnessing contextual data. Methods in Ecology and Evolution 11:303-315.

Tuia D, Kellenberger B, Beery S, Costelloe BR, Zuffi S, Risse B, Mathis A, Mathis MW, van Langevelde F, Burghardt T, Kays R, Klinck H, Wikelski M, Couzin ID, van Horn G, Crofoot MC, Stewart CV, Berger-Wolf T (2022) Perspectives in machine learning for wildlife conservation. Nature Communications 13:792.

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2019) blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution 10:225-232.

Williams BK, Nichols JD, Conroy MJ (2002) Analysis and Management of Animal Populations. Academic Press, San Diego.

Declarations

Funding

This research was supported by the European Research Council under Horizon Europe Grant Agreement No. ERC-2022-STG-101083841 (WildML), the Swiss National Science Foundation (SNSF) under grant 206021_207320, and the Spanish State Research Agency (AEI) under grant PID2022-138742OB-I00. Camera-trap infrastructure in Kenya was co-funded by the WWF-UK Species Conservation Fund and African Wildlife Foundation. The funders had no role in study design, data collection or analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare no conflicts of interest. No commercial ML software vendors were involved in study design or analysis. All ML frameworks used are open-source.

Data Availability Statement

Annotated camera-trap image datasets, trained model weights, species occurrence records, environmental covariate layers, and full analysis code (Python and R) are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19465716. The WildML open-source toolkit is available at https://github.com/WildML-Consortium/wildml under MIT license.

Ethical Approval

GPS-collaring of European wolves and Iberian lynx was conducted under Spanish Ministry of Ecological Transition permit SGPM-2022-014. Camera-trap surveys in Russia were authorised under FGBU Sikhote-Alin Biosphere Reserve research permit SAB-2022-07. Passive acoustic monitoring of humpback whales complied with NOAA MMPA Scientific Research Permit No. 24359-01. No invasive procedures were conducted on wildlife in the European Union member states.

ISSN: 3117-6925 Vol: 2024, Issues: Issue2, 2024

Appendix A

ML Model Hyperparameter Configurations and Training Diagnostics

The following records detail the final hyperparameter configurations, training convergence diagnostics, and cross-validation performance distributions for each of the seven ML architectures benchmarked in this study, organised by model class.

CONVOLUTIONAL NEURAL NETWORKS -- ResNet-50 and EfficientNet-B4

ResNet-50: ImageNet pre-trained weights; final LR = 0.001 (cosine annealing schedule); batch = 32; epochs = 80; weight decay = 1e-4; data augmentation: horizontal flip, random crop (224x224), colour jitter (brightness=0.4, contrast=0.4, saturation=0.2), MixUp (alpha=0.2).

EfficientNet-B4: ImageNet pre-trained weights; final LR = 0.0005; batch = 16; epochs = 60; EMA weight averaging (decay=0.9999); augmentation: RandAugment (magnitude=9), Cutmix (alpha=1.0).

Training convergence: All CNN models reached validation loss plateau within 45 epochs. Early stopping patience = 10 epochs. Final test-set top-1 accuracy: ResNet-50 = 91.4%; EfficientNet-B4 = 94.7%.

Hardware: 4 x NVIDIA A100 80GB GPUs; training time per species: ResNet-50 = 3.2 h; EfficientNet-B4 = 4.8 h.

ENSEMBLE TREE MODELS -- Random Forest and XGBoost GBM

Random Forest: n_estimators = 500; max_features = sqrt(p); min_samples_leaf = 5; OOB score used for diagnostics; 5-fold spatially blocked CV (blockCV package).

XGBoost GBM: n_estimators = 500; max_depth = 6; learning_rate = 0.05; subsample = 0.8; colsample_bytree = 0.8; early stopping rounds = 50; eval metric = rmse.

Environmental covariates: 18 predictors (11 CHELSA bioclimatic + 7 land-cover/structural variables); VIF screening removed 3 collinear predictors (threshold VIF > 10).

Mean 5-fold CV R2: Random Forest = 0.74 +- 0.06; XGBoost = 0.81 +- 0.05.

LSTM-RNN (Acoustic) and YOLOv8 (Drone Detection)

LSTM-RNN: 2-layer bidirectional LSTM; hidden units = 256; dropout = 0.3; sequence length = 128 spectrogram frames (0.64 s); Adam optimiser (LR = 0.001); batch = 64; epochs = 120.

Input features: Log-Mel spectrogram (128 bins, 25 ms window, 10 ms hop); Z-score normalised per recording.

YOLOv8 (drone): YOLOv8-L backbone; confidence threshold = 0.25; IoU threshold = 0.45; input resolution = 640x640; fine-tuned on 14,200 manually annotated drone images of L. africana; mAP@0.5 = 0.924.

Ensemble meta-learner: Ridge regression (alpha = 0.42, selected by 5-fold inner CV); trained on out-of-fold CNN and GBM predictions; R2 on held-out test set = 0.91.