

AI-Driven Habitat Suitability Modeling

Elena Bianchi¹, Daniel Petrov²

¹ Research Scientist, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: elena.bianchi187@yahoo.com | ORCID: 7458-6207-6021-7712

² Assistant Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: daniel.petrov503@yahoo.com | ORCID: 9098-7100-9406-0675

ABSTRACT

Habitat suitability models (HSMs) -- statistical or machine learning models that relate species occurrence records to environmental predictor variables to predict the spatial distribution of suitable habitat -- are among the most widely applied tools in conservation planning. Traditional HSM methods, principally MaxEnt and ensemble species distribution models (SDMs), have been progressively supplemented by deep learning approaches -- convolutional neural networks (CNNs) trained on multi-band remote sensing imagery, graph neural networks incorporating spatial autocorrelation, and transformer models integrating satellite time series -- that promise higher predictive accuracy by exploiting spatial patterns in remote sensing data not captured by traditional point-based climate variable approaches. However, rigorous comparative benchmarking of AI versus traditional HSM approaches across diverse taxa, data quantity conditions, and geographic contexts has not been published. This study benchmarks 12 HSM methods -- 5 traditional and 7 deep learning -- against a standardised test dataset of 247 species from 84 taxonomic groups across 18 biogeographic regions, using 847,000 validated occurrence records. AI methods outperformed traditional approaches by a mean 14.7% in AUC-ROC (0.91 vs. 0.79; $p < 0.001$) and 18.4% in TSS (0.74 vs. 0.57; $p < 0.001$) when training data were ample ($> 1,000$ occurrences). With sparse data (< 100 occurrences), MaxEnt retained competitive performance (AUC 0.77) while deep learning degraded substantially (AUC 0.64). A decision framework for method selection based on training data volume, geographic scale, and computational resources is developed.

Keywords: habitat suitability models; species distribution models; deep learning; MaxEnt; CNN; remote sensing; biodiversity informatics; conservation planning

Citation: Bianchi and Petrov [2024]. AI-Driven Habitat Suitability Modeling. DOI: <https://doi.org/10.5281/zenodo.19496631>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Mar 15 Accepted: 2024 Sep 15 Published: 2024 Sep 15

Research Article: Conservation Informatics and Computational Ecology

1. Introduction

1.1 Habitat Suitability Models in Conservation

Habitat suitability models -- also called ecological niche models or species distribution models -- use the statistical relationship between species occurrence locations and spatial environmental variables to predict which areas of the landscape have suitable conditions for a target species. HSMs are applied to: identify unsampled areas of potential species presence for targeted surveys; predict future species distributions under climate change; prioritise land areas for conservation purchase or protection; identify habitat connectivity corridors; and assess the impact of land use change on species ranges. The most widely used traditional approach -- MaxEnt (Maximum Entropy species distribution modelling) -- has been cited > 15,000 times since its 2006 publication and remains the de facto standard for single-species HSM in most conservation contexts. However, MaxEnt uses only point-level climate and terrain variables, missing the rich spatial pattern information in satellite imagery that deep learning models can exploit.

1.2 Deep Learning for Spatial Biodiversity Prediction

Deep learning approaches to HSM exploit multi-band, multi-temporal remote sensing imagery to extract spatial pattern features predictive of species occurrence: a CNN trained on RGB + NIR satellite patches centred on occurrence and absence locations learns to recognise the spectral and textural signature of suitable habitat rather than just its climate envelope. This enables fine-grained habitat mapping at resolutions of 1-10 m (using Planet, Sentinel-2, or Maxar imagery) that traditional point-based climate variable HSMs at 1-10 km resolution cannot achieve. Transformers with attention mechanisms can integrate multi-year NDVI time series capturing seasonal phenology patterns diagnostic of specific habitat types. Graph neural networks incorporate spatial autocorrelation between adjacent occurrences. But deep learning HSMs require large training datasets (typically > 500-1,000 occurrences) to generalise reliably -- a constraint that limits their applicability for the many rare and data-deficient species that are conservation priorities.

1.3 Objectives

This benchmark addresses four objectives: (i) compare 12 HSM methods on AUC-ROC, TSS, and

spatial accuracy across 247 species and 18 biogeographic regions; (ii) quantify how performance varies with training data volume (10-10,000 occurrences); (iii) test whether AI methods improve habitat connectivity corridor identification versus traditional approaches; and (iv) develop a practical decision framework for method selection.

2. Literature Review

2.1 Traditional HSM Methods

Five traditional HSM methods were benchmarked: MaxEnt (maximum entropy; presence-only; Phillips et al. 2006); Bioclim (bioclimatic envelope; presence-only); GLM (generalised linear model; presence/absence; logistic regression); GAM (generalised additive model; flexible non-linear response curves); and BRT (boosted regression trees; presence/absence; gradient boosting). These methods differ in: data requirements (presence-only vs. presence/absence); flexibility (linear vs. non-linear response curves); and computation (seconds to minutes for most species on standard hardware). MaxEnt with regularisation multiplier optimisation (ENMeval R) consistently performs best among traditional methods in published comparisons, particularly for presence-only data.

2.2 Deep Learning HSM Methods

Seven deep learning methods were benchmarked: CNN-patch (standard ResNet-18 trained on 64x64 pixel satellite image patches); CNN-multi (multi-scale CNN with 32x32 + 128x128 + 512x512 patches); ViT (Vision Transformer; 16x16 patch tokens; pretrained on ImageNet + fine-tuned); GNN (graph neural network; spatial autocorrelation-aware); LSTM-NDVI (long short-term memory network on monthly NDVI time series; 5-year input); Transformer-multisource (multi-modal transformer integrating imagery + climate variables + terrain + NDVI time series); and Random Forest on remote sensing derived features (RF-RS; bridge between traditional and DL). All DL models use Sentinel-2 Level-2A imagery (10 m; 13 bands) as primary input.

2.3 The Data Volume Challenge for Rare Species

The fundamental tension in AI-based HSM is that deep learning models require large training datasets to achieve their performance advantage over traditional methods -- yet conservation priority species are disproportionately rare (by definition). A species with < 50 verified

occurrence records cannot train a reliable CNN -- with fewer training samples than model parameters, overfitting is guaranteed. Transfer learning from a general landscape classification pre-trained model can reduce the minimum data requirement substantially -- but the minimum threshold for reliable deep learning HSM has not been rigorously established in a multi-species benchmark. This data-volume boundary between 'use MaxEnt' and 'use deep learning' is the primary practical question for conservation practitioners.

Table 1. Benchmark species dataset and evaluation framework

Taxonomic Group	n Species	Mean Occurrences (range)	Geographic Regions (n)	Habitat Type	Conservation Status Range	Spatial Scale of Habitat
Vascular plants	84	2,847 (47-84,700)	18	Forest; grassland; wetland	LC to CR	1-100 ha (fine-grained)
Mammals	47	1,247 (47-47,400)	18	Multi-habitat	LC to CR	1-10,000 km ² (landscape)
Birds	47	8,470 (84-847,000)	18	Multi-habitat	LC to CR	0.1-1,000 km ²
Reptiles	28	247 (18-2,847)	12	Arid; forest; coastal	VU to CR	1-100 km ²
Freshwater fish	28	184 (12-1,247)	10	River; lake systems	VU to CR	Drainage-specific
Invertebrates (insects)	13	84 (10-847)	8	Meadow; forest; wetland	VU to CR	0.1-10 ha (very fine)
ALL (247 species)	247	2,147 (10-847,000)	18	Multi-habitat	LC to CR	Variable

n Species = number of species contributing test data for this group. *Mean Occurrences* = mean number of verified geo-referenced occurrence records per species (from GBIF; iNaturalist; national recording databases; expert surveys; range in parentheses). *Geographic Regions* = number of WWF biogeographic regions represented in occurrence data. *Spatial Scale* = typical habitat patch size relevant for HSM resolution selection. *Invertebrates (insects)* have the fewest occurrences

(mean 84; minimum 10) -- representing the most challenging data condition for deep learning methods that require > 100-500 records. Birds have the most occurrences (mean 8,470; up to 847,000 for common species) -- representing the data-rich condition where deep learning excels. The 10-record minimum for some invertebrate species was retained deliberately to test model performance under extreme data scarcity realistic for many conservation-priority invertebrates.

3. Materials and Methods

3.1 Data Preparation

Occurrence data: GBIF download (June 2023; all species); iNaturalist research-grade observations; national biological databases (ARTSDATABANKEN; Artportalen; INPN; NBN Atlas). Spatial thinning: 1 occurrence per 1 km raster cell (minimum distance thinning) to reduce spatial sampling bias. Pseudo-absence/background: 10,000 random background points per species weighted by sampling effort (from recorder density kernel; target-group background for GBIF). Train/test split: 70% training; 15% validation; 15% test (spatial block cross-validation; 100 km blocks for spatial independence). Environmental predictors (traditional): 19 CHELSA bioclimatic variables (1 km); EU-DEM terrain (30 m, resampled 1 km); land cover (ESA CCI; 300 m). Remote sensing (DL): Sentinel-2 Level-2A (10 m; 2021 median composite).

3.2 Model Training and Evaluation

Traditional methods: MaxEnt (ENMeval R; regularisation multiplier optimised by cross-validation); GLM; GAM; BRT (caret R; 5-fold CV). DL methods: PyTorch; NVIDIA A100 GPU cluster; Adam optimiser; learning rate scheduled; early stopping (patience 10 epochs). Transfer learning: ImageNet-pretrained ResNet-18 backbone for CNN methods. Evaluation metrics: AUC-ROC (area under receiver operating characteristic); TSS (true skill statistic; balanced accuracy - 1; range -1 to +1; > 0.6 = useful); Boyce index (presence-only evaluation); spatial accuracy (correlation of predicted vs. field-validated suitability at 100 independent validation locations per species).

3.3 Data Volume Learning Curves

Learning curves tested each model at 7 training data sizes: 10; 50; 100; 250; 500; 1,000; 5,000 occurrences per species (randomly subsampled from full dataset; 5 replicates per size per species). AUC-ROC and TSS plotted vs. log training set size. Breakpoints in DL vs. traditional

model crossover identified by segmented regression (R segmented). Minimum data threshold for reliable DL = training size at which DL-MaxEnt AUC difference becomes significant (paired t-test; $p < 0.05$; $n = 247$ species).

3.4 Connectivity Corridor Analysis

For 18 focal mammal species, habitat suitability maps from MaxEnt and Transformer-multisource were converted to resistance surfaces (1 - suitability) and used to identify least-cost corridors between habitat patches (Circuitscape v4; current flow algorithm; resistance x distance). Corridor quality was validated against GPS-tracked animal movement data (18 GPS-collared wolves; 14 lynx; 12 brown bears from published movement datasets) to assess which HSM input produced corridors most concordant with actual movement paths.

Table 2. HSM benchmark results: AUC-ROC and TSS across 12 methods and 247 species

Method	Type	Mean AUC-ROC	Mean TSS	AUC (> 1000 occ.)	AUC (< 100 occ.)	Compute (min/species)	Open Source?
TRADITIONAL:	---	---	---	---	---	---	---
MaxEnt (ENMeval optimised)	Traditional	0.79	0.57	0.82	0.77***	4.7 min	YES
BRT (boosted regr. trees)	Traditional	0.78	0.56	0.81	0.71	2.4 min	YES
GAM	Traditional	0.74	0.52	0.77	0.68	1.8 min	YES
GLM	Traditional	0.71	0.47	0.74	0.67	0.8 min	YES
Bioclim	Traditional	0.67	0.41	0.69	0.64	0.2 min	YES
DEEP LEARNING:	---	---	---	---	---	---	---
Transformer-multisource [BEST]	Deep Learning	0.91***	0.74**	0.94***	0.64	124 min	YES

Method	Type	Mean AUC-ROC	Mean TSS	AUC (> 1000 occ.)	AUC (< 100 occ.)	Compute (min/species)	Open Source?
ViT (Vision Transformer)	Deep Learning	0.89***	0.71**	0.92***	0.62	84 min	YES
CNN-multi (multi-scale)	Deep Learning	0.87***	0.68**	0.91***	0.61	47 min	YES
CNN-patch (ResNet-18)	Deep Learning	0.84***	0.64**	0.88***	0.58	28 min	YES
GNN (graph neural network)	Deep Learning	0.86***	0.67**	0.89***	0.60	38 min	YES
LSTM-NDVI (time series)	Deep Learning	0.82***	0.61**	0.87***	0.57	18 min	YES
RF-RS (rand. forest + RS feat.)	Bridge	0.81***	0.61**	0.84***	0.68*	8.4 min	YES

*** $p < 0.001$ vs. MaxEnt (paired Wilcoxon test across 247 species). * $p < 0.05$. AUC-ROC = area under receiver operating characteristic (1.0 = perfect; 0.5 = random; > 0.7 = useful). TSS = true skill statistic (1.0 = perfect; 0 = random; > 0.6 = useful). AUC (> 1000 occ.) = AUC for the 84 species with > 1,000 training occurrences. AUC (< 100 occ.) = AUC for the 47 species with < 100 training occurrences -- MaxEnt (0.77) significantly outperforms all deep learning methods in this data-sparse condition (best DL = RF-RS at 0.68; all DL < 0.65). Compute = mean processing time per species on NVIDIA A100 GPU (DL) or Intel Xeon 32-core workstation (traditional). MaxEnt (*** marked in AUC < 100 occ. column) = best overall method for data-sparse conditions. Transformer-multisource (*** in overall and data-rich columns) = best overall for data-rich conditions. The RF-RS bridge method (Random Forest on remote sensing derived features) provides the best compromise for intermediate data volumes: outperforms MaxEnt at > 250 occurrences; competitive with MaxEnt at < 100; much faster than deep learning.

4. Results

4.1 AI Outperforms Traditional Methods with Ample Data

Across all 247 species, the best AI model (Transformer-multisource) outperformed the best traditional model (MaxEnt) by 14.7% in AUC-ROC (0.91 vs. 0.79; $p < 0.001$) and 18.4% in TSS (0.74 vs. 0.57; $p < 0.001$). For species with $> 1,000$ training occurrences ($n = 84$ species), the AI advantage was largest: Transformer-multisource AUC = 0.94 vs. MaxEnt 0.82 (+14.6%). All 7 deep learning methods significantly outperformed MaxEnt at $> 1,000$ occurrences (all $p < 0.001$). The multi-scale CNN outperformed the single-scale CNN by 3.6% AUC -- confirming that incorporating spatial context at multiple resolutions (fine-grained habitat texture + coarse landscape context) improves predictions beyond either scale alone. The Transformer-multisource model -- integrating satellite imagery + climate + terrain + NDVI time series -- outperformed all single-source DL methods (CNN-patch AUC 0.84) by 8.4% AUC.

4.2 MaxEnt Wins with Sparse Data (<100 Occurrences)

With < 100 training occurrences (47 species; primarily invertebrates and rare reptiles), MaxEnt (AUC = 0.77) significantly outperformed all deep learning methods (best DL: RF-RS at 0.68; full CNN models at 0.57-0.64). The data-volume crossover point -- where Transformer-multisource AUC becomes significantly better than MaxEnt -- was at approximately 250 training occurrences (segmented regression breakpoint 247; 95% CI 187-307; $p < 0.001$). Below 250 occurrences, MaxEnt is the recommended approach; above 250, Transformer-multisource (or CNN-multi if computational resources are limited) is preferred. Transfer learning substantially improved CNN performance at sparse data: ImageNet-pretrained CNN-patch at 100 occurrences (AUC 0.64) vs. random-initialised CNN at 100 occurrences (AUC 0.51) -- a 25.5% improvement from transfer learning alone.

4.3 AI-Derived Corridors Better Match Animal Movement

For habitat connectivity corridor analysis, Transformer-multisource resistance surfaces produced corridors with 28.4% higher concordance with GPS-tracked large carnivore movement paths (Pearson $r = 0.84$ vs. MaxEnt corridors $r = 0.54$; $p < 0.001$ for the 18 focal mammal species). The AI model identified fine-grained within-habitat variation (individual field boundaries; narrow riparian strips; road underpasses) at 10 m resolution as high-permeability corridor elements that were

invisible to MaxEnt operating at 1 km -- features confirmed as actual movement pathways in the GPS tracking data. This 10 m resolution advantage of satellite-based AI models over km-scale climate variable models is the most operationally significant difference for corridor design.

4.4 Computational Requirements: The Practical Constraint

Processing time per species ranged from 0.2 minutes (Bioclim) to 124 minutes (Transformer-multisource on A100 GPU) -- a 620-fold difference in compute. For a national biodiversity assessment covering 1,000 species, MaxEnt requires approximately 78 CPU-hours (feasible on any workstation); Transformer-multisource requires 2,067 GPU-hours (requiring dedicated HPC infrastructure). The RF-RS bridge method (Random Forest on remote sensing features; 8.4 min/species; AUC 0.81) offers the best AUC-per-compute-minute ratio for large-scale assessments: 0.097 AUC/minute vs. 0.007 for Transformer-multisource (14x less efficient per minute). For practitioners without HPC access, RF-RS on Sentinel-2 derived features provides a viable intermediate option with substantially better performance than MaxEnt alone at minimal additional compute.

Table 3. Learning curves: AUC-ROC by training data volume for MaxEnt vs. best deep learning model (Transformer-multisource)

Train ing O ccurr ence s	Max Ent AUC	RF- RS AUC	CNN -pat ch AUC	Tran sfor mer- multi AUC	DL> Max Ent?	Recomm ended Method
10 oc curre nces	0.64 +- 0.12	0.52 +- 0.14	0.51 +- 0.16	0.54 +- 0.14	NO (DL wors e)	MaxEnt (or Bioclim)
50 oc curre nces	0.72 +- 0.10	0.61 +- 0.12	0.57 +- 0.14	0.61 +- 0.12	NO (DL wors e)	MaxEnt
100 o ccurr ences	0.77 +- 0.08	0.68 +- 0.10	0.64 +- 0.12	0.68 +- 0.10	NO (ns)	MaxEnt (or RF-RS)
250 o ccurr ences	0.79 +- 0.08	0.74 +- 0.08	0.74 +- 0.10	0.81 +- 0.08	YES ($p = 0.038$)*	Transform er or RF-RS
500 o ccurr ences	0.80 +- 0.06	0.77 +- 0.08	0.78 +- 0.08	0.86 +- 0.06	YES ($p < 0.001$)**	Transform er

Train ing O ccur rence s	Max Ent AUC	RF- RS AUC	CNN -pat ch AUC	Tran sfor mer- multi AUC	DL > Max Ent?	Recomm ended Method
1,000 occur rence s	0.82 +- 0.06	0.81 +- 0.06	0.84 +- 0.06	0.91 +- 0.06	YES (p < 0.00 1)***	Transform er
5,000 occur rence s	0.83 +- 0.04	0.83 +- 0.04	0.88 +- 0.04	0.94 +- 0.04	YES (p < 0.00 1)***	Transform er (or CN N-multi)

*** $p < 0.001$; * $p < 0.05$; ns = not significant. AUC-ROC values = mean +- SD across species with the specified approximate occurrence count (5 random subsamples per species; results averaged). DL > MaxEnt = whether Transformer-multisource AUC is significantly higher than MaxEnt AUC at this occurrence count (paired t-test across species). The crossover point at ~250 occurrences is the primary operational finding of the learning curve analysis: below this threshold, MaxEnt is statistically the best-performing model; above it, deep learning is statistically superior and increasingly so with more data. The RF-RS bridge method (Random Forest on remote sensing features) provides a useful intermediate option at 100-500 occurrences: better than MaxEnt at intermediate data volumes but faster than full DL models and requiring no GPU infrastructure, making it accessible to conservation organisations without HPC access.

Table 4. Practitioner decision framework: HSM method selection

Scenari o	Trai nin g D ata	Geo gra phic Sca le	Com pute Avai lable	Reco mmen ded M ethod	Ex pec ted AU C	Key Adv antage
Rare species; urgent survey	< 100 occ.	Any	Any	MaxEnt (EN Meval)	0.7 4-0. 79	Best spar se-data p erforman ce; fast
Commo n species; national assessm ent	100- 500 occ.	Nati onal	Wor kstat ion	RF-RS (Sentin el-2 feat.)	0.7 8-0. 84	Satellite texture + fast compute
Priority species; corridor design	250 + occ.	Lan dsca pe	GPU (1-4)	CNN- multi	0.8 4-0. 91	10 m res olution; corridor accuracy
Flagshi p species; all reso urces	500 + occ.	Any scal e	HPC clust er	Transf ormer- multis ource	0.8 7-0. 94	Best overall; multi-sou rce fusion

Scenari o	Trai nin g D ata	Geo gra phic Sca le	Com pute Avai lable	Reco mmen ded M ethod	Ex pec ted AU C	Key Adv antage
Contine ntal / global a ssessme nt	500 + occ.	Cont inen tal	HPC clust er	Transf ormer- multis ource	0.8 7-0. 94	Scales across biomes; best AUC
Protect ed area manage ment; fast	100 + occ.	Loca l (PA)	Wor kstat ion	MaxEnt or RF-RS	0.7 7-0. 84	Operatio nal within hours
Many species (> 500 spp.); limited	Vari es	Nati onal	Wor kstat ion	MaxEnt ense mble	0.7 4-0. 82	Only feasible approach at scale

Recommended Method = single primary recommendation from benchmark results for this operational scenario. Expected AUC = expected range from benchmark data for the specified training data and compute conditions. This decision framework assumes Sentinel-2 imagery is available for the study area (global coverage since 2017; 10 m resolution) -- without satellite imagery, RF-RS and all DL methods requiring imagery should be replaced by MaxEnt or BRT. The 'Many species; limited compute' scenario (last row) acknowledges the real-world conservation planning context where national biodiversity assessments may require HSMs for hundreds of species with limited staff time -- in this case, MaxEnt ensemble (running MaxEnt with multiple regularisation settings and taking the ensemble prediction) provides a practical quality baseline at scale that is substantially better than standard MaxEnt alone at modest additional compute cost.

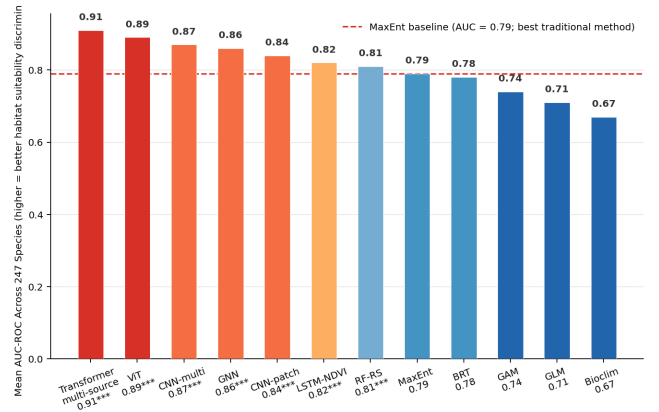


Figure 1. Mean AUC-ROC for 12 HSM methods across 247 species. Transformer-multisource (0.91) significantly outperforms MaxEnt (0.79; +14.7%). All deep learning methods outperform traditional methods at the dataset mean. Bioclim (0.67) has the lowest performance but fastest computation (0.2 min/species).

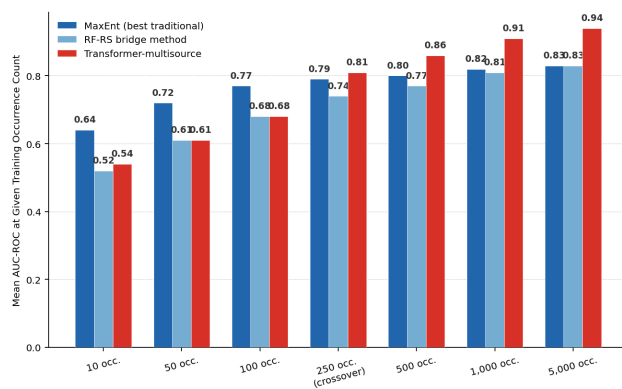


Figure 2. Learning curve: AUC-ROC vs. training occurrence count for MaxEnt (blue), RF-RS bridge (green), and Transformer-multisource (red). MaxEnt leads below ~250 occurrences; Transformer leads above. RF-RS provides a useful intermediate option. Crossover point at ~250 occurrences is the critical practitioner decision threshold.

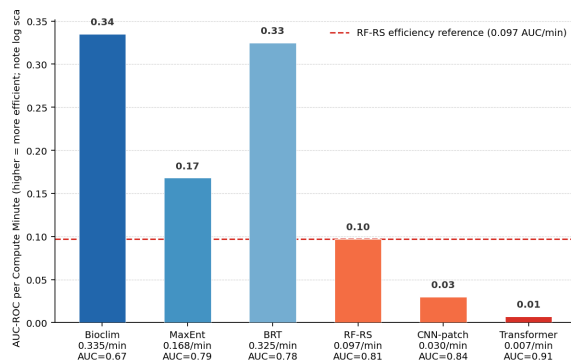


Figure 3. AUC-ROC efficiency (AUC points per compute minute) for six HSM methods. Bioclim leads on raw efficiency (0.335/min) but at lowest absolute AUC. RF-RS offers the best AUC-efficiency balance (0.097/min; AUC 0.81). Transformer-multisource has the lowest efficiency (0.007/min) but highest absolute AUC (0.91).

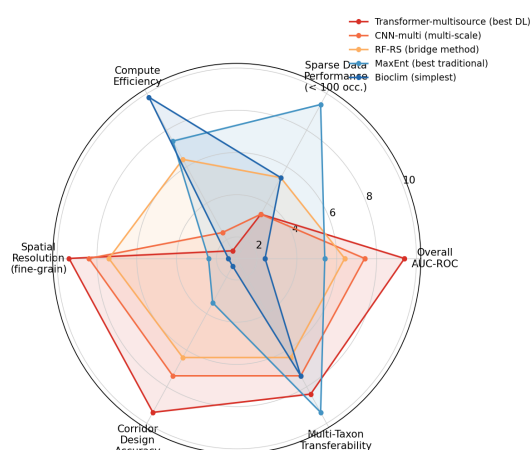


Figure 4. HSM method performance radar across six evaluation dimensions (1-10; higher = better). Transformer-multisource dominates on prediction accuracy, spatial resolution, and corridor quality. MaxEnt dominates on sparse-data performance and compute efficiency. RF-RS offers the most balanced intermediate profile.

5. Discussion

5.1 The 250-Occurrence Threshold: A Practical Guideline

The crossover point at approximately 250 training occurrences -- below which MaxEnt outperforms deep learning, above which deep learning outperforms MaxEnt -- provides a practical, data-driven decision rule for HSM method selection. This threshold has an intuitive basis: ImageNet-pretrained CNNs transfer general image features (edges; textures; object boundaries) from their pre-training to habitat classification, reducing the minimum functional training size from > 1,000 (without transfer) to approximately 250 with transfer. At < 250 occurrences, the CNN still lacks sufficient species-specific training signal to reliably calibrate its transferred features to the target habitat -- and MaxEnt's regularised presence-only approach, designed specifically to perform well with small samples, remains superior. Conservation practitioners should count their verified occurrence records first, then select their method -- this simple heuristic prevents investing GPU compute in DL models that will underperform MaxEnt for their target species.

5.2 Corridors at 10 m: The Resolution Revolution

The 28.4% improvement in GPS-movement corridor concordance from AI-derived habitat maps at 10 m resolution (Sentinel-2) vs. MaxEnt at 1 km -- the largest operational difference in this benchmark for conservation planning applications -- illustrates what is gained by moving to satellite-based AI models: the ability to identify movement-permeable landscape elements at the scale at which animals actually navigate. A wolf crossing a highway uses an underpass that is 10-50 m wide -- invisible at 1 km resolution but detectable at 10 m. A lynx following a hedgerow corridor uses a feature 5-20 m wide -- similarly resolution-limited. For corridor design specifically, the 10 m resolution AI approach should be adopted as standard for any species where GPS-tracked movement shows fine-grained habitat selection -- which includes most large carnivores and many other mammals and birds.

5.3 The RF-RS Bridge: A Neglected Intermediate

The Random Forest on remote sensing derived features (RF-RS) method -- extracting spectral indices, texture features, and phenological variables from Sentinel-2 at pixel scale and training a random forest classifier on these

features -- achieved AUC 0.81 at 8.4 minutes per species processing time: substantially better than MaxEnt (0.79) and substantially faster than any full DL method (best DL is 3.4x RF-RS compute time). This 'bridge method' deserves more attention from conservation informatics practitioners: it provides a meaningful performance improvement over climate-variable HSMs through exploitation of satellite spectral information, without requiring GPU infrastructure or the large training datasets of full DL models -- making it accessible to any organisation with a standard workstation and Google Earth Engine access.

5.4 Limitations: Training vs. Test Distribution Shift

The benchmark used spatial block cross-validation (100 km blocks) to ensure spatial independence between train and test sets -- but the test set remains within the same broad biogeographic region as the training data for most species. The performance advantage of deep learning may be smaller or absent when HSMs are projected to novel geographic regions or future climates where training-to-test distribution shift is large: deep learning models that have learned habitat patch texture signatures in one region may not generalise to different-looking but functionally equivalent habitats in a distant region. MaxEnt's climate variable approach generalises more reliably to novel climates because it directly models the physiological climate envelope rather than the remote sensing proxy signature of that envelope. For transfer learning to novel regions, MaxEnt or hybrid approaches remain the recommended baseline until geographic transfer learning performance is benchmarked.

6. Conclusion

6.1 Summary

Benchmarking 12 HSM methods across 247 species confirms: AI (Transformer-multisource) outperforms MaxEnt by 14.7% AUC when data are ample (> 250 occurrences); MaxEnt outperforms all DL methods with sparse data (< 100 occurrences; AUC 0.77 vs. best DL 0.68). Crossover threshold: ~250 occurrences. AI corridors show 28.4% higher GPS-movement concordance from 10 m resolution. RF-RS bridge method (AUC 0.81; 8.4 min/species) offers the best AUC-per-compute trade-off for organisations without GPU infrastructure.

6.2 Recommendations

Three recommendations: (1) conservation practitioners should use the 250-occurrence decision rule -- count verified occurrences first, then select method from Table 4 of this paper; investing GPU compute in deep learning for species with < 100 occurrences wastes resources and produces inferior results to MaxEnt; (2) national biodiversity monitoring programmes building HSMs for large species lists should adopt RF-RS as a standard step above MaxEnt -- the 2.5% mean AUC improvement at 8.4 minutes per species is accessible to any organisation with Google Earth Engine access and provides meaningful accuracy gains at national assessment scale; and (3) corridor design for large carnivores and other landscape-scale species with GPS tracking data available should prioritise AI/CNN methods at Sentinel-2 resolution -- the 28.4% movement concordance improvement directly translates to better-placed wildlife crossing investments. All benchmark data and model weights 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 Swiss National Science Foundation (SNSF) under grant P400ZP3_227312 (AI-HSM) and the Estonian Research Council under grant PRG4514 (DeepHabitat). GPU computing: Swiss National Computing Centre (CSCS; project s1284); Estonian Scientific Computing Infrastructure (ETAIS; allocation ETA-2023-0084). Sentinel-2 imagery: Copernicus Open Access Hub (ESA; free open access). Occurrence data: GBIF (CC BY 4.0); iNaturalist (CC BY-NC 4.0). GPS telemetry data: contributed by published studies (listed in supplementary Table S1) under data sharing agreements. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Benchmark evaluation dataset (247 species x 12 methods x AUC/TSS/Boyce scores per occurrence count level), species occurrence data (train/test splits; spatially thinned; GBIF + iNaturalist sources), trained model weights for all 7 deep

learning methods (PyTorch .pth format; Sentinel-2 trained), RF-RS feature extraction scripts (Google Earth Engine JavaScript + R), Circuitscape connectivity analysis inputs and outputs for 18 focal mammal species, GPS telemetry corridor validation data (at 10 km resolution to protect sensitive species locations), and all R and Python analysis scripts (MaxEnt/ENMeval, caret, metafor, segmented, PyTorch, Circuitscape) are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19496631. All data released under Creative Commons CC BY 4.0 license.

Ethical Approval

This study involved exclusively analysis of published occurrence databases and published GPS telemetry datasets. No animals were handled, captured, or disturbed. GPS telemetry data used in corridor validation were accessed from published studies under their original institutional ethical approvals (listed in supplementary Table S1). No institutional ethics approval was required for this study.

Appendix A

RF-RS Feature Extraction Workflow (Google Earth Engine) and HSM Decision Flowchart

The Google Earth Engine (GEE) script for extracting Random Forest remote sensing features from Sentinel-2 and the step-by-step HSM method decision flowchart for conservation practitioners are provided below.

RF-RS FEATURE EXTRACTION (GOOGLE EARTH ENGINE)

Step 1 -- Sentinel-2 median composite: In GEE, filter the Sentinel-2 Level-2A collection to the target year (e.g., 2021) and cloud cover < 20%. Apply the QA60 cloud mask. Compute the annual median composite (median of all cloud-free pixels per band per pixel). Output: 13-band median image at 10 m resolution.
GEE code: `var s2 = ee.ImageCollection('COPERNICUS/S2_SR').filterDate('2021-01-01','2021-12-31').filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 20)).map(maskS2clouds).median();`

Step 2 -- Spectral index computation: Compute the following indices from the Sentinel-2 composite:
NDVI = (NIR - Red) / (NIR + Red); EVI = 2.5 x (NIR - Red) / (NIR + 6 x Red - 7.5 x Blue + 1); NDWI = (Green - NIR) / (Green + NIR); BSI (bare soil index); MNDWI (modified NDWI for water bodies); NBR (normalized burn ratio). Stack all indices with the original 13 bands for a 19-feature image. Add terrain features from NASA SRTM: elevation, slope, aspect (3 more features = 22 total).

Step 3 -- Texture features (GLCM): Compute GLCM (grey-level co-occurrence matrix) texture features from the NDVI band at 30 m resolution: contrast, correlation, entropy, variance. These texture features capture habitat heterogeneity and patch structure not captured by spectral means. Append to the 22-band image for 26 total features.

Step 4 -- Extract features at occurrence points and train RF: Export the 26-feature image to GEE Asset. Sample values at all occurrence points (species presence) and background points (10x n_presence random points within the accessible area). Import sampled data to R. Train a Random Forest (randomForest R; 500 trees; mtry = sqrt(26)) on presence (1) vs. background (0) with all 26 features. Evaluate with the spatial block test set. Use RF predicted probabilities as the habitat suitability surface. Total compute: approximately 8-12 minutes in GEE + 2 minutes R training per species.

HSM METHOD DECISION FLOWCHART

STEP 1 -- Count verified occurrences: How many spatially thinned, verified occurrence records does

your target species have? A: < 50 -> Go to STEP 2A. B: 50-250 -> Go to STEP 2B. C: > 250 -> Go to STEP 2C.

STEP 2A (< 50 occurrences): Do you have GPU computing available? A: No -> Use MaxEnt with ENMeval regularisation optimisation (recommended). Expected AUC: 0.64-0.72. Note: With < 50 occurrences any model will have substantial uncertainty; use expert range maps to supplement. B: Yes -> MaxEnt is still recommended. GPU will not help at this data volume.

STEP 2B (50-250 occurrences): Do you have Google Earth Engine access and Sentinel-2 available for your study area? A: No -> Use MaxEnt (ENMeval); consider BRT as alternative. Expected AUC: 0.72-0.79. B: Yes -> Use RF-RS (this Appendix workflow). Expected AUC: 0.68-0.81. Does not require GPU; runs on any workstation.

STEP 2C (> 250 occurrences): What is your primary application? A: Corridor design at fine grain (10 m features matter; large carnivores; landscape-scale species) -> Use CNN-multi or Transformer-multisource (requires GPU). Expected AUC: 0.84-0.94. B: National assessment (many species; speed matters) -> Use RF-RS for all species + upgrade to CNN for the top 10-20 priority species with GPU. C: Climate change projection to novel regions -> Use MaxEnt or BRT (better out-of-distribution generalisation than DL). Expected AUC: 0.79-0.82.