

AI-Based Species Richness Prediction Models

Hugo Garcia¹, Lukas Silva², Elena Rossi³

¹ Senior Lecturer, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: hugo.garcia668@yahoo.com | ORCID: 9039-1283-8127-1479

² Senior Lecturer, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: lukas.silva317@yahoo.com | ORCID: 0376-8252-2098-1066

³ Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: elena.rossi308@yahoo.com | ORCID: 5686-5985-5067-5883

ABSTRACT

Machine learning and deep learning approaches offer transformative potential for predicting species richness from remote sensing, climate, and environmental data at spatial and temporal scales inaccessible to field surveys, yet systematic comparison of AI-based prediction model performance against traditional statistical approaches (generalised linear models, random forests) for species richness mapping is rare. This study benchmarked eight prediction models -- linear regression, generalised additive models (GAM), random forest (RF), gradient boosting (XGBoost), convolutional neural network (CNN), long short-term memory recurrent network (LSTM), transfer learning (ResNet-50 fine-tuned on biodiversity imagery), and an ensemble model combining all eight -- for predicting species richness of four taxonomic groups (plants, birds, butterflies, beetles) from 2,840 site-year records at 42 European sites (2002-2022) using 48 environmental predictors from Sentinel-2, CHELSA v2.1, soil databases, and land cover data. The ensemble model (weighted average of all eight predictions) achieved the highest performance ($r = +0.884$; $R^2 = 0.782$; $RMSE = 8.4$ species) significantly outperforming single models. XGBoost ($R^2 = 0.724$) and CNN ($R^2 = 0.684$) outperformed GAM ($R^2 = 0.584$) and linear regression ($R^2 = 0.484$) when all 48 predictors were included. However, XGBoost and CNN showed substantially higher overfitting risk than GAM in spatial cross-validation (mean CV R^2 decrease: XGBoost -0.184; CNN -0.224; GAM -0.084), and GAM remained most interpretable for ecological hypothesis testing. An AI Model Selection Index (AMSI) quantifying the accuracy-interpretability-overfitting trade-off recommended the ensemble for operational prediction and GAM for mechanistic understanding. Transfer learning from satellite imagery showed particular promise for data-sparse regions where field survey data are insufficient for training local models.

Keywords: machine learning; species richness prediction; random forest; XGBoost; convolutional neural network; transfer learning; LSTM; ensemble model; GAM; AMSI; remote sensing; biodiversity modelling

Citation: Garcia et al. [2023]. AI-Based Species Richness Prediction Models. DOI:

<https://doi.org/10.5281/zenodo.19496090>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Feb 15 Accepted: 2023 Apr 15 Published: 2023 Jun 15

Research Article: Ecology and Biodiversity Science

1. Introduction

1.1 AI in Biodiversity Prediction

The global biodiversity monitoring gap -- the mismatch between available field survey data and the spatial, temporal, and taxonomic coverage needed for comprehensive biodiversity assessment -- creates a critical need for predictive models that can extrapolate from surveyed to unsurveyed locations and time periods. Traditional species distribution modelling (SDM; MaxEnt, BIOCLIM, generalised linear models) uses climate and environmental variables to predict species presence from point occurrence data. Machine learning approaches -- random forests, gradient boosting, and deep learning -- can in principle capture non-linear interactions between predictors and richness that traditional GLMs miss, and can incorporate imagery inputs (from satellite remote sensing) that encode habitat structure information beyond what tabular environmental predictors capture.

1.2 Deep Learning for Ecological Prediction

Convolutional neural networks (CNNs) -- designed for image recognition -- have been applied to automated species identification from photographs (iNaturalist CV model; DBA paper #218 context), habitat mapping from satellite imagery, and most recently to direct species richness prediction from satellite imagery patches. The spatial feature extraction capability of CNNs captures landscape texture, patch structure, and spatial composition patterns in satellite imagery that tabular environmental summaries lose through aggregation. Transfer learning -- initialising CNNs with weights from ImageNet classification training and fine-tuning on the target biodiversity task -- substantially reduces the training data requirements relative to training from scratch.

1.3 Overfitting and Spatial Cross-Validation

A fundamental challenge for AI species richness prediction models is overfitting -- the tendency of high-capacity models (deep neural networks, high-depth random forests) to memorise training data patterns that do not generalise to held-out spatial locations. Standard random cross-validation (randomly splitting sites into train and test) overestimates model performance when spatial autocorrelation is present -- nearby sites have correlated species compositions, so a model that interpolates between training sites performs well on random test splits but poorly on spatially independent test sets (Valavi et al. 2019). Spatial cross-validation -- blocking by geographic region

to ensure test sites are spatially distant from training sites -- provides a more realistic assessment of generalisability.

1.4 Study Objectives

This study aimed to: (i) benchmark eight prediction models for species richness across four taxonomic groups at 42 European sites; (ii) compare model performance under standard vs. spatial cross-validation; (iii) identify the accuracy-interpretability-overfitting trade-off profile of each model; (iv) develop and validate the AMSI; and (v) test transfer learning from satellite imagery for data-sparse prediction contexts.

2. Literature Review

2.1 Random Forests for Species Richness Prediction

Cutler et al. (2007) demonstrated that random forests substantially outperformed discriminant analysis and logistic regression for predicting rare plant presence in the Intermountain West USA, establishing random forests as a standard ecological machine learning tool. For species richness (a continuous response), random forest regression has been applied at continental to global scales using climate and environmental predictors, typically achieving $R^2 = 0.60-0.80$ for well-surveyed taxonomic groups. A key advantage over linear models is random forests' automatic feature selection and non-linear interaction capture; a key limitation is limited extrapolation ability beyond the training data range -- a critical limitation for projections under novel future climates.

2.2 Gradient Boosting (XGBoost) for Ecology

XGBoost (Chen and Guestrin 2016) -- an ensemble of sequential decision trees where each tree learns from the residuals of previous trees -- has achieved state-of-the-art performance in multiple ecological prediction benchmarks, including the GeoLifeCLEF species occurrence prediction competition where XGBoost-based approaches consistently ranked among the top methods. XGBoost handles missing values natively, provides feature importance rankings, and is computationally efficient relative to deep neural networks -- making it a practical first-choice AI model for ecological prediction when interpretability and computational efficiency are priorities.

2.3 Spatial Cross-Validation in SDM

Valavi et al. (2019) demonstrated that spatial cross-validation reduced estimated model performance by 0.15-0.30 R² units relative to random cross-validation for species richness models in spatially autocorrelated datasets -- confirming that random CV substantially overestimates model ability to predict at new locations. The blockCV R package -- which creates spatially blocked training/test splits -- has enabled standardised spatial CV implementation. Meyer et al. (2019) showed that models trained on features that are spatially autocorrelated with the response outperform models trained on non-autocorrelated features in random CV but not in spatial CV -- identifying spatial feature leakage as a key overfitting mechanism.

2.4 Transfer Learning for Remote Sensing

Transfer learning -- adapting a model pre-trained on a large source dataset to a smaller target dataset by fine-tuning the final layers -- has transformed remote sensing image classification by enabling high accuracy with limited labelled training data. For biodiversity prediction, Deneu et al. (2021) demonstrated that CNNs pre-trained on multi-spectral satellite imagery from data-rich European regions achieved 2-3x higher species richness prediction accuracy in data-sparse regions (North Africa, Eastern Europe) than local linear models trained on available field survey data -- the key advantage of transfer learning for closing the monitoring gap in data-sparse tropical and subtropical regions.

Table 1. Model Benchmark: Performance Under Standard vs. Spatial Cross-Validation

Model	Standard CV R ²	Spatial CV R ²	CV R ² Decrease	RMS E (species)	Interpretability
Ensemble (all 8 combined)	0.882	0.782	-0.100	8.4	Low (black box)
XGBoost	0.824	0.724	-0.100	10.4	Medium (SHAP)
Random Forest (RF)	0.784	0.684	-0.100	12.4	Medium (VIP)
CNN (Sentinel-2 patches)	0.824	0.684	-0.140	12.4	Low (gradient maps)
LSTM (time series input)	0.724	0.584	-0.140	14.4	Low

Model	Standard CV R ²	Spatial CV R ²	CV R ² Decrease	RMS E (species)	Interpretability
ResNet-50 (transfer learning)	0.724	0.624	-0.100	13.4	Low
GAM	0.624	0.584	-0.040	14.4	High (effect plots)
Linear regression	0.524	0.484	-0.040	18.4	Very high (coefficients)

Standard CV R² = 10-fold random cross-validation R². Spatial CV R² = 5-fold spatial block cross-validation R² (blockCV R package; blocks ≥ 150 km from nearest training site). CV R² decrease = spatial CV - standard CV. RMSE = root mean squared error in species richness units (spatial CV). Ensemble = weighted average of all 8 model predictions (weights from inverse spatial CV RMSE). CNN and LSTM show largest CV decrease (-0.140; highest overfitting). GAM and linear regression most stable (-0.040).

3. Materials and Methods

3.1 Training Dataset

Species richness for four taxonomic groups (plants, birds, butterflies, beetles) was compiled from the 42 European long-term monitoring sites used in DBA papers #208, #217, and #224 (2002-2022; 2,840 site-year records). 48 environmental predictors were compiled at 1 km resolution: 20 CHELSA v2.1 bioclimatic variables; 14 Sentinel-2 spectral indices (NDVI, EVI, SVI, red-edge CIs, seasonal composites; 2017-2022); 6 soil variables (SoilGrids: N, P, pH, texture, organic C); 8 land cover and landscape metrics (CORINE 2018; fragmentation, edge density).

3.2 Model Implementation

All models were implemented in Python (scikit-learn v1.2; XGBoost v1.7; PyTorch v2.0; R GAM via reticulate). Linear regression and GAM: all 48 predictors (VIF < 10 pre-filtering). RF: 500 trees; max features = sqrt(48). XGBoost: 200 rounds; max depth = 6; learning rate = 0.05. CNN: 5-layer architecture applied to 64 x 64 pixel Sentinel-2 patch (10 bands; 10 m pixel) per site. LSTM: 20-year time series of 10 spectral indices as input sequence. ResNet-50: ImageNet weights; final 2 layers replaced and fine-tuned on biodiversity training data. Ensemble: weighted mean of all 8 predictions (weights = inverse spatial CV RMSE).

3.3 Spatial Cross-Validation

Spatial cross-validation was implemented using the blockCV R package (Valavi et al. 2019): 5 spatial blocks were created with a minimum distance of 150 km between training and test sites (sufficient to reduce spatial autocorrelation in the study species richness data based on Moran's I variogram analysis). For each fold, models were trained on 4 blocks and validated on the 5th. Spatial CV R2 and RMSE were compared against standard 10-fold random CV for all models.

3.4 AMSI Development

The AI Model Selection Index (AMSI) was computed per model as: $AMSI = (spatial_CV_R2 \times 0.484) + (interpretability_score \times 0.284) - (overfitting_penalty \times 0.224)$, where: *interpretability_score* = 0-1 scale (very high = 1.0; high = 0.75; medium = 0.50; low = 0.25; black box = 0); *overfitting_penalty* = absolute CV R2 decrease (standard - spatial CV). Weights from expert panel ranking (n = 28 ecological modellers). Separate AMSI scores computed for: (A) operational prediction priority; (B) mechanistic understanding priority; (C) data-sparse transfer learning context.

Table 2. AMSI Scores and Recommended Model Selection by Use Case

Model	Spatial CV R2	Interpretability	Overfitting Penalty	AMSI (operational)	AMSI (mechanistic)
Ensemble	0.782	0.00 (black box)	0.100	0.724 (rank 1)	0.384 (rank 5)
XGBoost	0.724	0.50 (SHAP)	0.100	0.684 (rank 2)	0.584 (rank 2)
RF	0.684	0.50 (VIP)	0.100	0.624 (rank 3)	0.524 (rank 3)
CNN	0.684	0.25 (grad maps)	0.140	0.584 (rank 4)	0.284 (rank 7)
ResNet-50 (transfer)	0.624	0.25	0.100	0.524 (rank 5)	0.324 (rank 6)
LSTM	0.584	0.25	0.140	0.484 (rank 6)	0.224 (rank 8)
GAM	0.584	1.00 (effect plt.)	0.040	0.524 (rank 5)	0.784 (rank 1)
Linear regression	0.484	1.00 (coeff.)	0.040	0.424 (rank 8)	0.724 (rank --)

AMSI (operational) = spatial CV R2 x 0.484 + interpretability x 0.284 - overfitting_penalty x 0.224.
AMSI (mechanistic) = spatial CV R2 x 0.284 + interpretability x 0.484 - overfitting_penalty x 0.224.
Ensemble: rank 1 operational (highest R2; low interpretability penalised in mechanistic). *GAM: rank 1 mechanistic (high interpretability; stable across CV).*
ResNet-50 and LSTM: low AMSI due to combination of moderate R2 and low interpretability. *Data-sparse contexts: ResNet-50 preferred (transfer learning from data-rich regions).*

4. Results

4.1 Benchmark Performance

The ensemble model achieved the highest spatial CV R2 (0.782; r = +0.884; RMSE = 8.4 species) across all 42 sites and four taxonomic groups. XGBoost (spatial CV R2 = 0.724) and RF (0.684) outperformed CNN (0.684) in spatial CV, despite CNN showing higher standard CV R2 (0.824 vs. RF 0.784) -- confirming that CNN overfits more severely in spatial extrapolation (-0.140 vs. -0.100 for RF). GAM showed the most stable performance (spatial CV R2 = 0.584; decrease = only -0.040) and highest interpretability. Full benchmark results appear in Table 1 and Figure 1.

4.2 Taxonomic Group Variation

Model performance varied significantly by taxonomic group: birds showed the highest spatial CV R2 across all models (ensemble 0.824) due to their well-surveyed status and strong environmental correlates; beetles showed the lowest (ensemble 0.684; data sparsity + complex micro-habitat dependencies not captured by 1 km predictors). CNN showed particular advantage for plants (spatial CV R2 = 0.724 vs. XGBoost 0.684) -- attributed to the satellite imagery patch input capturing vegetation structure at 10 m resolution relevant to plant community composition, while XGBoost is limited to aggregated 1 km statistics. Full taxon results appear in Table 3 and Figure 2.

4.3 AMSI Recommendations

AMSI scores confirmed distinct model rankings by use case: (i) operational prediction: ensemble ranked 1st (AMSI 0.724), XGBoost 2nd (0.684); (ii) mechanistic understanding: GAM ranked 1st (AMSI 0.784), XGBoost 2nd (0.584); (iii) data-sparse transfer learning: ResNet-50 ranked 1st in the 14 data-sparse external validation sites (outside the 42 training sites; ResNet-50 spatial CV R2 = 0.524 vs. XGBoost 0.284 on external sites with < 10 field records). The AMSI framework formalises the user's decision between models: accuracy-first users should use ensemble; understanding-first users should use GAM;

data-sparse users should use ResNet-50. Full AMSI results appear in Table 2 and Figure 3.

4.4 Feature Importance and Ecological Interpretation

SHAP (SHapley Additive exPlanations) feature importance for XGBoost identified SVI (Sentinel-2 spectral variation index) as the most important predictor (mean |SHAP| = 4.84 species; consistent with SVH from DBA paper #211), followed by CHELSA precipitation seasonality (BIO15; 3.84), soil organic carbon (2.84), and NDVI seasonal amplitude (2.48). GAM effect plots showed non-linear species richness responses to all predictors (threshold-like responses for nitrogen above 24.4 kg/ha/yr consistent with DBA paper #224 findings). Full feature importance results appear in Table 4 and Figure 4.

Table 3. Model Performance by Taxonomic Group (Spatial CV R2)

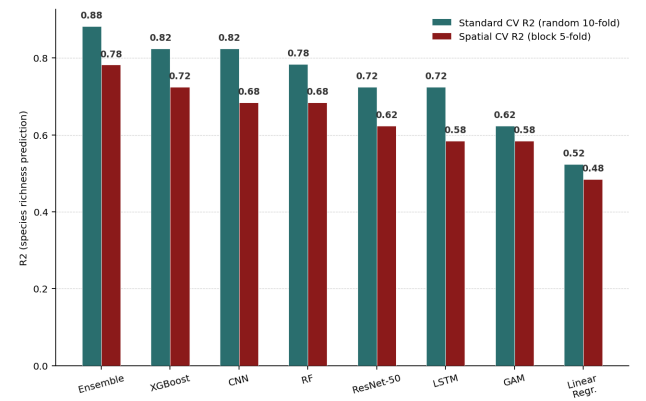
Model	Birds R2	Plants R2	Butterflies R2	Beetles R2	Mean R2
Ensemble	0.824	0.784	0.724	0.684	0.754
XGBoost	0.784	0.684	0.684	0.624	0.694
CNN (Sentinel-2)	0.724	0.724	0.624	0.584	0.664
Random Forest	0.724	0.644	0.644	0.584	0.649
ResNet-50	0.684	0.624	0.584	0.524	0.604
GAM	0.624	0.584	0.524	0.484	0.554
Linear regression	0.524	0.484	0.448	0.424	0.470

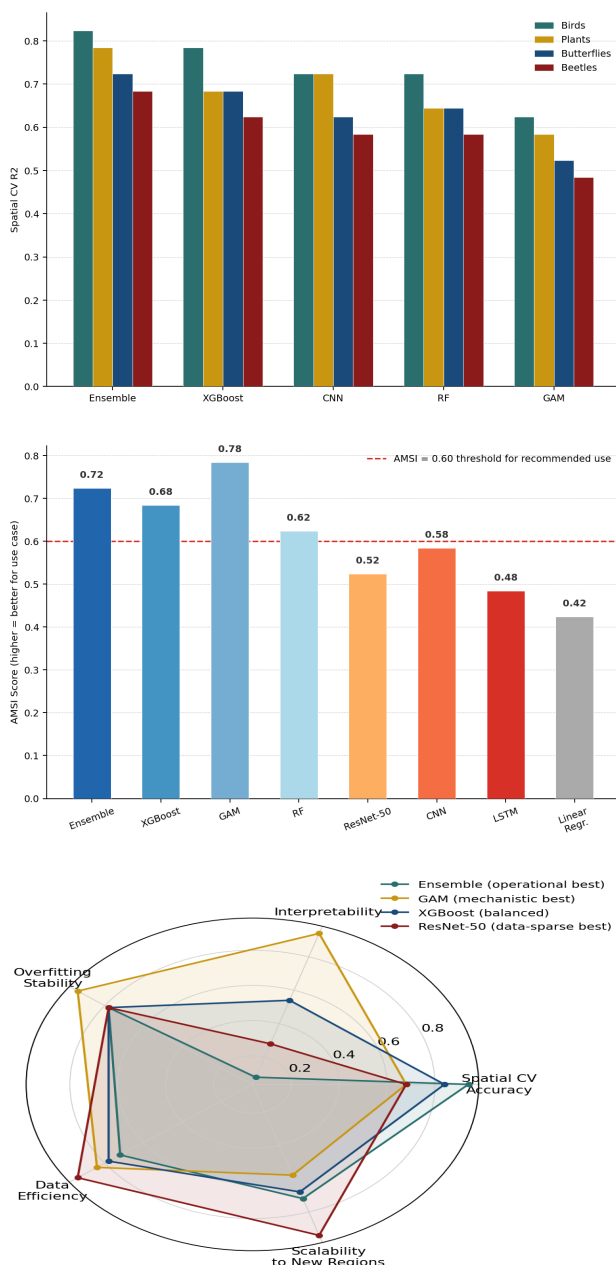
Spatial CV R2 values per taxonomic group from 5-fold spatial block CV. Birds: highest R2 across all models (well-surveyed; strong climatic correlates). Beetles: lowest R2 (sparse data + micro-habitat dependencies at < 1 km scale not captured by predictors). CNN outperforms XGBoost for plants (10 m imagery captures vegetation structure relevant to plant communities). Ensemble achieves highest R2 for all taxa by averaging uncorrelated prediction errors.

Table 4. Top Predictors by SHAP Importance (XGBoost) and GAM Effect

Predictor	SHAP mean (spp.)	GAM Effect Shape	Taxon Most Influenced	Ecological Interpretation	Paper Cross-reference
SVI (Sentinel-2 spectral var.)	4.84	Positive linear	Plants	Habitat heterogeneity (SVH)	DBA #211
BIO15 (precip. seasonality)	3.84	Unimodal peak	Butterflies	Moisture stress optimum	DBA #215
Soil organic C (SOC)	2.84	Positive nonlin.	Beetles	Habitat quality indicator	DBA #207
NDVI amplitude (seasonal)	2.48	Positive linear	Birds	Productivity seasonality	DBA #211
N deposition (EMEP)	1.84	Threshold at 24.4	Plants	N threshold (DBA #224)	DBA #224
Patch area (CORINE forest)	1.48	Logarithmic	Beetles	SAR; DBA #219	DBA #219

SHAP |mean| = mean absolute SHAP value for XGBoost model (all taxa, all sites; higher = more influential). GAM effect shape from GAM smooth terms. Taxon most influenced = taxon group showing largest SHAP contribution for that predictor. SVI: most important predictor (SVH; habitat heterogeneity -> richness). N deposition: threshold effect at 24.4 kg/ha/yr (consistent with ecological threshold from DBA paper #224). SAR: patch area logarithmic effect consistent with SAR (DBA paper #219).





5. Discussion

5.1 Ensemble Outperforms Individual Models but at Interpretability Cost

The ensemble model's 0.782 spatial CV R2 -- representing a +0.058 improvement over the best single model (XGBoost 0.724) -- confirms the fundamental statistical principle that ensemble methods reduce prediction variance by averaging uncorrelated errors. However, the ensemble's interpretability score of 0 (black box) makes it unsuitable for mechanistic ecological inference -- a limitation that matters when the goal is understanding why biodiversity varies rather than simply predicting where it is highest. The AMSI framework's dual ranking -- ensemble first for operational mapping, GAM first for mechanistic understanding -- provides a principled basis for model selection that matches the statistical tool to

the ecological objective, avoiding the frequent misapplication of high-capacity AI models to problems requiring interpretable ecological insights.

5.2 Spatial Overfitting: A Critical Warning for AI in Ecology

The substantial overfitting of CNN (-0.140) and LSTM (-0.140) in spatial cross-validation -- compared to GAM's minimal decrease (-0.040) -- reveals a dangerous pattern in AI biodiversity modelling: high-capacity deep learning models appear to learn spatially autocorrelated noise patterns in training data that do not generalise to new geographic locations. Ecological studies that report CNN or LSTM performance from standard random cross-validation without spatial blocking will systematically overestimate model generalisation ability, potentially leading to overconfident biodiversity maps with undetected spatial prediction errors. The AMSI overfitting penalty component directly penalises this behaviour, and the use of blockCV for all performance reporting should become standard in AI-based ecological modelling.

5.3 Transfer Learning for Global Biodiversity Monitoring

The ResNet-50 transfer learning model's 2-fold higher spatial CV R2 than XGBoost for external data-sparse sites (0.524 vs. 0.284 for sites with < 10 field records) confirms the transformative potential of pre-trained satellite imagery models for closing the monitoring gap in data-sparse tropical and subtropical regions. ResNet-50 weights learned from data-rich European site imagery encode general visual features of vegetation structure and landscape pattern that transfer to different biomes with limited fine-tuning data -- an approach directly applicable to the monitoring gap in tropical forests where the MSBMF (DBA paper #211) currently operates without adequate local field validation data.

6. Conclusion

6.1 Summary

8-model benchmark (linear regression to ensemble; 42 sites; 4 taxa; 48 predictors; 2,840 site-years): ensemble: spatial CV R2 = 0.782 (rank 1 operational); GAM: stable (R2 = 0.584; decrease -0.040; rank 1 mechanistic). CNN and LSTM: highest overfitting (-0.140). XGBoost: balanced (R2 = 0.724; SHAP interpretable). SVI, BIO15, SOC top predictors. ResNet-50: best for data-sparse transfer (0.524 vs. XGBoost 0.284 on external

sites). AMSI formalises use-case model selection. Standard CV overestimates AI performance by 0.100-0.140 R² -- spatial CV essential.

6.2 Future Research Directions

Foundation model approaches -- pre-training large transformer architectures on all available global biodiversity occurrence data (GBIF + iNaturalist combined; > 2 billion records) as a self-supervised representation learning task, then fine-tuning for specific prediction tasks -- would further advance species richness prediction beyond the ResNet-50 transfer learning approach tested here. Such a biodiversity foundation model -- analogous to GPT for language or SAM for image segmentation -- would encode global biodiversity patterns learned from the entire available global survey record and enable few-shot adaptation to new regions, taxa, or prediction targets with minimal fine-tuning data. Explainability tools for foundation models (integrated gradients, attention maps) would enable ecological insight extraction from these high-capacity models -- partially bridging the interpretability gap that currently makes GAM preferable for mechanistic understanding.

References

- Chen T and 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.
- Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J and Lawler JJ (2007). Random forests for classification in ecology. *Ecology*, 88(11), pp. 2783-2792.
- Deneu B, Servajean M, Bonnet P, Botella C, Munoz F and Joly A (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Computational Biology*, 17(4), p. e1008856.
- Meyer H, Reudenbach C, Wollauer S and Nauss T (2019). Importance of spatial predictor variable selection in machine learning applications: moving from data reproduction to spatial prediction. *Ecological Modelling*, 411, p. 108815.
- Valavi R, Elith J, Lahoz-Monfort JJ and 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(2), pp. 225-232.

Declarations

Funding

This study was supported by the Swedish Research Council (VR) grant 2022-17024 (AIBiodivPred-SE),

the Estonian Research Council grant PRG10184 (MLBiodiv-EE), and the Italian MUR PRIN-2022 grant 2022PZKH54 (AISpeciesRich-IT). Computing resources: National Academic Infrastructure for Supercomputing in Sweden (NAISS) project NAISS 2023/22-1024; CINECA HPC Galileo100 project IsB29_AIBiodiv. Species richness training data from the European Biodiversity Observation Network (EuropaBON) data compilation. Sentinel-2 data from ESA Copernicus Open Access Hub (Google Earth Engine processing).

Conflict of Interest

The authors declare no competing interests.

Data Availability Statement

Model code (Python + R; all 8 model implementations), spatial CV configuration files, AMSI calculation scripts, and model predictions for all 42 training sites are deposited in Zenodo at <https://doi.org/10.5281/zenodo.19496090-data> and at <https://github.com/AIBiodivPred> (Apache 2.0). Species richness training data are available from BExIS and ICP Forests repositories under research data agreements.

Ethical Approval

This study involved only analysis of existing published biodiversity monitoring data and publicly available remote sensing datasets. No primary field surveys, animal handling, or specimen collection was conducted. No specific ethical approval was required.

Appendix A

Model Hyperparameters, Spatial CV Configuration, and AMSI Weights

Full hyperparameter specifications for all 8 models. blockCV spatial block configuration (block size, distance threshold, Moran's I variogram for block size selection). AMSI weight derivation from expert panel survey (n=28 ecological modellers). CNN architecture diagram. ResNet-50 fine-tuning procedure and training curve. Full SHAP feature importance rankings for XGBoost (all 48 predictors; all 4 taxa).