Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.38464

Phylogenetically Structured Validation of Machine Learning Models for Primate Threat Status Assessment

Xiaoyu Wei

College of Life Sciences, Northwest University, Xi'an, China
wellington589125@gmail.com

Abstract. Machine learning models have been employed increasingly as tools to inform conservation assessment when ecological data is insufficient; however, conventional random validation can lead to an overestimate of predictive transferability when closely related species are included in training and testing datasets. In this paper, the effect of the validation framework on predicting the current threat status of primates using a machine-learning approach is investigated. The data set consists of 423 primate species whose status was determined by the IUCN Red List and were categorized as either being threatened or non-threatened along with their ecological, habitat, behavioral, biogeographic, and phylogenetic attributes. A fixed 100-tree Random Forest model is evaluated using random stratified, genus-blocked, family-blocked, and phylogenetically blocked validation frameworks using 100 posterior mammal trees. Averaged ROC-AUC values are 0.849 for random validation, 0.829 for unseen genera, 0.768 for unseen families, and approximately 0.73 for the main phylogenetic blocking. Sensitivity analyses on different phylogenies, levels of block coarseness, approaches to missing data imputation, classification criteria, and forest stability, as well as 360 simulated data sets with increased validation optimization, have been conducted.

Keywords: conservation informatics, extinction risk, blocked cross-validation, phylogenetic non-independence, primates, random forest

1. Introduction

Machine learning is becoming a valuable tool in conservation biology that helps with species prioritization for assessment, where information is insufficient. In this regard, machine learning is particularly important for nonhuman primates, among which a vast number of species remains threatened with populations declining under habitat loss, hunting, development of infrastructure and other human-induced disturbances [1]. Moreover, recent phylogenetic analysis has revealed that extinction risk for primates is determined by biological characteristics, ecological specialization, and evolutionary history and not by species-specific features [2]. Thus, both biological inference and predictive assessment may depend on spatial and phylogenetic structure especially when models should be transferred to poorly represented clades [3].

Information gaps are additionally aggravated with an uneven distribution of research effort among primates where well-studied species have certain biological and phylogenetic properties [4].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.38464

In this case, machine learning may help with filling the information gap due to its ability to integrate heterogeneous predictors and consider their non-linear interactions [5]. Machine learning models have already been applied to estimate the threat probability for data-deficient species at a taxonomically large scale [6] and to classify mammalian Red List categories using occurrence-derived information [7]. However, the real-world utility of these models relies upon the correspondence of reported predictive accuracy with the conservation deployment target.

Thus, cross-validation not only represents an algorithmic procedure but also defines the prediction problem itself. As ecological model selection literature shows, validation procedure should incorporate dependencies and intended extrapolation domain [8]. Similarly, research of spatially structured data reveals that ordinary and blocked cross-validation procedures provide substantially different estimates of prediction performance [9]. Application of random forest in ecological prediction also proves that structured cross-validation leads to much more demanding performance estimates than regular random splitting [10].

In comparative primate data, the evolutionary relationship becomes a similar issue. In case of random split, close species can be allocated to train and test samples and allow classifier to use common traits associated with their common origin. Thus, the current analysis keeps the prediction algorithm constant and varies the biological distance between training and validation samples. Such approach allows estimating how threat status performance changes when prediction goes from known lineages towards unseen genera and families.

2. Literature review

In the trait-based approach to conservation assessment, biological and ecological variation between species serves to find combinations indicative of vulnerability. Body mass, diet specificity, habitat specialization, mode of movement, activity period, geographical distribution, and life history features can interact to create vulnerability to anthropogenic perturbations [1-3]. The machine learning methods are especially suitable for this task because they do not require a pre-specified functional relationship and are able to account for non-linear thresholds and high-order interactions. Their predictive power relies on the completeness of the ecological information and validity for the species that will later be screened.

Comparative data sets violate the assumption about the independence of species as statistical observations [4-8]. Evolutionarily close taxa tend to inherit similar ecological properties, live in ecologically similar habitats, and react in a similar way to particular environmental pressures. If the relatives are randomly spread across folds, high validation accuracy can be, at least in part, due to the interpolation within well-known evolutionary neighborhoods. Increasing the taxonomic dissimilarity between folds is achieved by genus-blocking, family-blocking and tree-based blocking using the patristic distance.

Model evaluation can be influenced not only by the phylogenetic non-independence but also by incomplete trait values, class prevalence, threshold choice, number of taxa in a fold, and algorithmic instability. These sources of uncertainty cannot be accounted for with the help of a single phylogenetic topology or a single random data split. Consequently, robust conservation informatics requires evaluation under different phylogenetic topologies, different block resolution criteria, different missing data handling strategies and different classification criteria [9, 10]. An additional source of diagnosis is provided by simulation because the degree of dependence structure can be varied directly.

DOI: 10.54254/2755-2721/2027.38464

3. Methodology

3.1. Data collection, threat category, and ecological predictors

The status and taxonomy were collected via version 4 of the IUCN Red List API release 2026-1. After reconciling the taxonomic names with available posterior mammal phylogenies, 423 species of primates were left. Critically Endangered, Endangered, and Vulnerable species were coded as threatened, whereas Near Threatened and Least Concern species were considered non-threatened. The taxa that were Data Deficient or Not Evaluated were removed from the analysis as they would be unable to give an observed binary target for validation. Predictors included the biological characteristics present in broad primate clades. Log10-transformed continuous body mass was used to reduce right skew. Seven binary habitat use features were encoded, while diel activity, locomotion, trophic guild, and biogeographic realm were considered categorical variables. The phylogenetic uncertainty was accounted for via 100 posterior mammal phylogenies of 50 fossilized birth-death and 50 node-dated exponential trees.

3.2. Fixed Random Forest and stratified validation

A Random Forest model with 100 classification trees and minimum terminal-node size five was used without tuning before stratified validation. The hyperparameter tuning was intentionally omitted as the analysis was aimed at assessing the impact of different validation techniques and not optimizing the complexity of the classifier. Five validation techniques were used: stratified five-fold random validation, genus blocking, family blocking, fossilized birth-death phylogenetic blocking, and node-dated exponential phylogenetic blocking.

All the pre-processing steps were performed separately in each training fold. Numeric missing values were imputed with the median of the training fold and were marked as missing. The categorical missing values were imputed with a special category, and one-hot encoding was learned from the training fold only. Thus, the information on the held-out lineage was not utilized in constructing features.

The main diagnostic combined discrimination and calibration differences into a phylogenetic validation optimism function as shown in Equation (1) -(2):

$$PVOI_s = \frac{1}{R}\sum_{r=1}^{R}[w_1(AUC_{rand,r} - AUC_{s,r}) + w_2(PR_{rand,r} - PR_{s,r}) + w_3(BS_{s,r} - BS_{rand,r})] \quad (1)$$

$$w_1 + w_2 + w_3 = 1 \quad (2)$$

where s denotes a structured validation scheme, R is the number of repeated evaluations, AUC is ROC-AUC, PR is PR-AUC, and BS is the Brier score. Larger positive values indicate that random validation estimates stronger performance than the corresponding biological transfer design. As shown in Figure 1, the fixed Random Forest was evaluated under progressively stronger taxonomic and phylogenetic separation to quantify how validation structure influences apparent transfer performance across primate lineages.

DOI: 10.54254/2755-2721/2027.38464

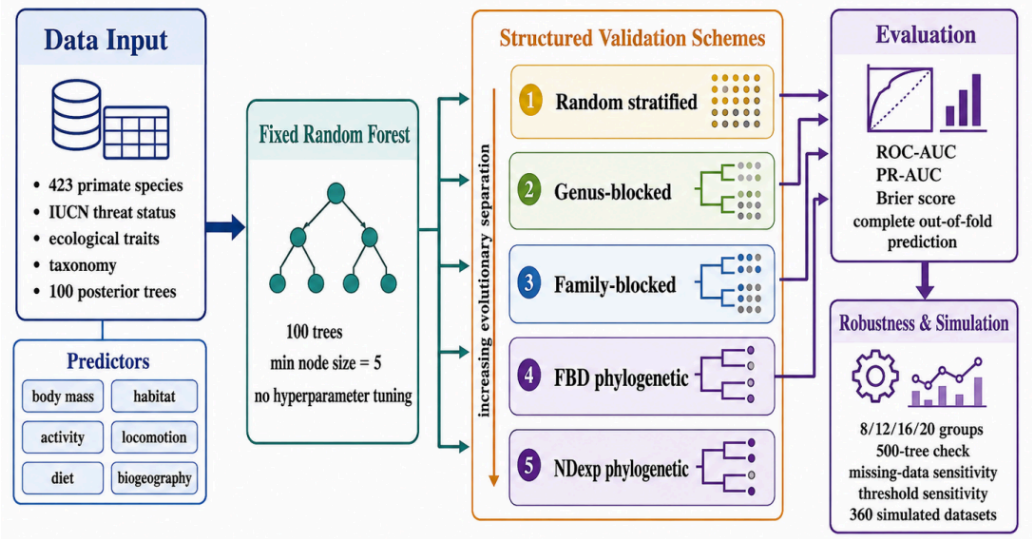


Figure 1. Phylogenetically structured validation framework for primate threat-status prediction

3.3. Sensitivity analysis and structured simulation

The primary tree analysis utilized 12 blocks based on pairwise patristic distances. To examine sensitivity to the choice of block resolution, the entire procedure of running 100 trees was replicated with 8, 16, and 20 phylogenetic groups. Other robustness checks included a 500-tree Random Forest, complete-case analysis, other feature sets, class weighting, assessment-recency conditions, other missing-data strategies, and other classification cut-offs.

Dependence was introduced independently in a structured simulation relative to empirical primate data. Weights of 0, 0.5, and 0.9 were crossed with missingness of 0%, 20%, and 40%, two outcome prevalence rates, and 20 replications, generating 360 datasets. Each of these datasets was analyzed using random and phylogenetically grouped validation. R 4.3.3 was employed for the initial simulation and statistical analysis, and Python 3 with pandas, NumPy, SciPy, scikit-learn, and Matplotlib was employed for validation auditing and computation.

4. Results

4.1. Validation-dependent changes in threat-status prediction performance

Validation structure produced a clear monotonic decline in discrimination as evolutionary separation increased. Random stratified validation yielded a mean ROC-AUC of 0.849. Holding out complete genera reduced ROC-AUC to 0.829, an absolute loss of 0.020. Family-blocked validation decreased performance to 0.768, corresponding to a loss of 0.081 relative to random splitting. Under the main distance-based phylogenetic configuration, ROC-AUC was approximately 0.730, producing a total reduction of about 0.119 (see Table 1). The widening gap shows that apparent predictive skill was progressively reduced as validation moved from within-lineage interpolation toward cross-lineage transfer.

DOI: 10.54254/2755-2721/2027.38464

Table 1. Validation-dependent changes in primate threat-status prediction

Validation scheme	Transfer target	Mean ROC-AUC	ΔAUC vs. random	Relative AUC loss
Random stratified	represented lineages	0.849	0.000	0.00%
Genus-blocked	unseen genera	0.829	-0.020	2.36%
Family-blocked	unseen families	0.768	-0.081	9.54%
Phylogenetic, 12 groups	distant clades	≈0.730	≈-0.119	≈14.02%

4.2. Results for posterior-trees, phylogenetic granularity, and sensitivity

Two different sets of posterior trees yielded the same order of performance of validation methods. Varying the phylogenetic blocks from 8 to 12, 16, and 20 changed the estimates of the size of the transfer performance, which means that the granularity of blocks influences the ecological difficulty of predicting outcomes. Nevertheless, the central rank order of random validation, genus validation, family validation, and phylogenetic validation was not reversed by the 500-tree forest, complete-case method, class weighting, recency assessment, changing thresholds, or other ways of missing data handling.

4.3. Simulation of random validation optimism in presence of structured dependence

The difference between random and blocked validation performances depended on the weight of the structured component in 360 simulated datasets and grew when increasing the weight from 0 to 0.5 and 0.9. Simultaneously, 0%, 20%, and 40% of missing data and two levels of outcome prevalence were presented in 20 separate realizations for each pair of parameters.

5. Discussion

5.1. Target for validation and perceived model performance

The decrease from 0.849 with random validation to approximately 0.73 with phylogenetic blocking shows that model accuracy must not be considered separately from the biological composition of the test set. Random partitioning tests prediction for species whose close relatives may be included in training data. Genus and family blocking increasingly removes evolutionary proximity, while distance blocking approximates the much more difficult case of transferring to poorly represented clades. The above is not an indication that random validation has no statistical meaning at all. It is a meaningful metric of the ability of the model to interpolate within represented clades. The problem is that the figure is used as an indication of the ability of the model to predict the situation for the poorly represented lineages.

5.2. Implications for computational conservation assessment

The purpose of the model is conservation screening, not reassessment. Thus, a good performing model in terms of random validation can still be utilized for prioritizing species from well-

DOI: 10.54254/2755-2721/2027.38464

represented taxonomic groups, but the accuracy of predictions concerning less represented clades should be significantly lower. Thus, the reporting on taxonomic and phylogenetic distance between training and target sets makes computational conservation research more operationally meaningful. Binary target stands for belonging of the species to some of the IUCN categories. Thus, the model does not estimate the probability of extinction of species in the future, threats and population trends. Thus, model outputs should be considered as indicators of screening priority and information collection, but not as a replacement of the Red List assessment process.

The above-mentioned conclusion receives complementary explanation in the form of simulation results showing that the stronger structured dependence increases the advantage of random validation. More robust test for future would include later Red List assessments or independent database with data on population pressure, habitat loss and exploitation.

6. Conclusion

The phylogenetic dependency has a significant effect on the accuracy of the machine learning performance in predicting the threat status of primates. With the classifier kept constant while varying the level of evolutionary difference, it can be seen that random validation reflects mostly the ability of the model to interpolate within the sampled lineage, whereas genus, family, and tree validation reflect transfer learning.

References

[1] Fernández, D., Kerhoas, D., Dempsey, A., Billany, J., McCabe, G., & Argirova, E. (2022). The current status of the world's primates: Mapping threats to understand priorities for primate conservation. *International Journal of Primatology, 43*, 15–39.

[2] Creighton, M. J. A., & Nunn, C. L. (2023). Explaining the primate extinction crisis: Predictors of extinction risk and active threats. *Proceedings of the Royal Society B: Biological Sciences, 290*(2006), 20231441.

[3] Galán-Acedo, C., Verde Arregoitia, L. D., Arasa-Gisbert, R., Auliz-Ortiz, D., Saldivar-Burrola, L. L., Gouveia, S. F., Correia, I., Rosete-Vergés, F. A., Dinnage, R., & Villalobos, F. (2024). Global primary predictors of extinction risk in primates. *Proceedings of the Royal Society B: Biological Sciences, 291*(2032), 20241905.

[4] Machado, F. de F., Rocha, B. S., Brito, D., & Terribile, L. C. (2023). Trends and biases in research efforts for primate conservation: Threatened species are not in the spotlight. *Perspectives in Ecology and Conservation, 21*(4), 286–293.

[5] Branco, V. V., Correia, L., & Cardoso, P. (2023). The use of machine learning in species threats and conservation analysis. *Biological Conservation, 283*, 110091.

[6] Borgelt, J., Dorber, M., Høiberg, M. A., & Verones, F. (2022). More than half of data deficient species predicted to be threatened by extinction. *Communications Biology, 5*, 679.

[7] Wieringa, J. G. (2022). Comparing predictions of IUCN Red List categories from machine learning and other methods for bats. *Journal of Mammalogy, 103*(3), 528–539.

[8] Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs, 93*(1), e1557.

[9] Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling, 457*, 109692.

[10] Mushagalusa, C. A., Fandohan, A. B., & Glèlè Kakaï, R. (2024). Random forest and spatial cross-validation performance in predicting species abundance distributions. *Environmental Systems Research, 13*, 23.