From Risk Labels to Timelines: Predicting When Species Will Go Extinct

Group 5

Alyssa Ray, Solomon Flax, Timothy Padden, Jacob Wolfson, Parth Desai, Christopher Cardwell

Introduction

Biodiversity loss is accelerating at an unprecedented rate (18), with over 48,600 species currently threatened with extinction according to the International Union for Conservation of Nature (IUCN). Because ecosystems rely on complex interdependencies among species, such losses jeopardize the stability of natural systems that sustain life. The consequences extend beyond ecology: over half of the global economy, estimated at $44 trillion (25), is directly dependent on nature and its ecosystem services. Moreover, resource degradation has been linked to political instability and secondary effects such as mass migration, compounding regional pressures. In this context, accurately predicting extinction risk and identifying its key drivers is critical. Understanding which variables most influence species survival and which species are at the most immediate risk, enables targeted conservation strategies, helping mitigate ecological collapse and safeguard long-term economic and social resilience.

Problem Definition

This study introduces a multi-stage analytical framework to assess extinction risk, identify main drivers of extinction, reclassify species status, and estimate extinction timing. Modeling will extrapolate survival probabilities, risk scores, and risk ratings for all species in 2025, identifying those most vulnerable. This output will then be used in combination with intrinsic traits (e.g., body mass, reproductive rate) and extrinsic environmental factors (e.g., conservation actions, habitat loss) to reassess existing IUCN Red List extinction classifications. These updated classifications will then be used to estimate the expected year of extinction for each species. This approach enables a more comprehensive assessment of extinction labeling, highlighting the relative influence of biological and environmental drivers. Results will be visualized through line charts showing how extinction probability evolves over time, alongside an interactive tool that allows users to adjust key variables and observe their impact on species longevity. This framework offers a data-driven foundation for prioritizing conservation efforts and evaluating the mechanisms driving extinction risk.

Literature Survey

Current approaches often rely on reviews (3), which highlight important threats but are largely descriptive, correlation-based, and do not capture the full range of factors such as habitat loss or human activity; Other approaches use Machine Learning methods, which can reveal patterns (1), but they are based on historic data and are not directly transferable to modern species. Overall, today's practices provide valuable insights but lack predictive power and integration across multiple drivers of extinction risk.

For predicting time until extinction, one study used linear regression to estimate mammals' extinction time based on population size and body mass, offering useful insight into expected time until extinction. However, this approach relies on access to fossil or historical extinction data and cannot account for censored species that are still prevalent, making it less suitable for extinction forecasting (4).

Another approach, Optimal Linear Estimation (OLE), uses the most recent sighting records (assuming a Weibull extreme-value distribution) to generate both a point estimate and confidence interval for a species' extinction date. However, this method depends heavily on consistent search effort over time. If sightings are irregular or efforts decline, the model tends to underestimate the extinction date. To compensate, researchers often apply a 5-10 year buffer before officially declaring a species extinct (21). A related method, the sighting-trend index, builds on this idea by analyzing the average time between sightings and whether that interval is increasing or decreasing. This allows researchers to estimate the probability that a species is extinct and calculate an upper 95% confidence bound for the likely extinction year (22).

For predicting extinction risk, one approach utilized a random forest model to predict extinction risk across mammals, incorporating both intrinsic and environmental factors and comparing predictions to International Union for Conservation of Nature (IUCN) Red List labels. Random Survival Forests effectively capture nonlinear relationships and high-dimensional data in extinction modeling, but their limited interpretability constrain insights into variable importance (5). Inversely, some studies that prioritized interpretability over complexity via greedy, stepwise attribute selection, ended with an oversimplified model (11). Reports also compared extinction predictors between birds and mammals and accounted for variable collinearity, which is an important methodological step, but its short five-year dataset limits its usefulness for long-term extinction modeling (7). Finally, other logistic regression models were evaluated; these approaches assessed the role of both intrinsic and environmental factors together to understand key environmental factors to include, however the model did not test the types of factors in isolation so it was unable to explain their individual contributions to extinction risk (8).

For this study, research will utilize data from the IUCN Red List, which offers detailed records of extinction risk classifications and their changes over time. As highlighted in the above studies, a central challenge is the scarcity of complete data for extinct species, particularly accurate extinction timing (4). Most available data pertain to species that are still prevalent, resulting in a highly censored dataset. One way to address this, as noted in other literature, is through a Bayesian sighting-time test. This test uses the time since the last sighting to calculate a Bayes factor, which is the likelihood of a species being extant versus extinct, and converts it into a posterior probability of persistence (20). Like other methods (OLE) it assumes consistent pre-extinction abundance and search effort, meaning declines in either can bias results toward extinction. This study will integrate survival models that can perform well with censored datasets, with species that remain extant treated as right-censored in the analysis. Assessment will be done using Scikit-Survival Random Survival Forests to predict survival probabilities and the risk of extinction using intrinsic and environmental covariates. When substantial multicollinearity exists among predictors, Random Forest

modeling will be employed due to its robustness in the presence of correlated variables and its ability to mitigate bias in feature importance estimation.

To effectively implement these models, careful selection of intrinsic and extrinsic covariates is required to ensure biological relevance and predictive accuracy. Prior meta-analyses and comparative studies have synthesized findings across multiple modeling approaches, such as generalized linear models and decision trees, to identify key intrinsic traits most strongly associated with extinction risk, including body mass and geographic range size (6). These insights helped in the selection of intrinsic predictors in this study. However, existing research has often underrepresented environmental and anthropogenic factors, despite consistent evidence linking human-driven pressures such as hunting, deforestation, and climate change to elevated extinction risk (8; 9; 16). To address this gap, the present study integrates both intrinsic and environmental variables into the survival modeling framework. Using these variables within the survival models, metrics will be derived for each species, including the predicted probability of extinction for 2025 and relative risk ratings.

Survival-derived features will be used as inputs to a logistic classification model designed to predict the binary outcome of whether a species is likely to go extinct within a defined timeframe. Although classification models have been applied independently in prior studies (9; 10), few have incorporated predicted time-to-extinction as a feature, limiting their ability to capture temporal risk dynamics. To evaluate the relative contribution of different variable types, this study will construct three classifier models: one using intrinsic traits alone, one using environmental factors, and one combining both. Variable importance will be assessed across models to identify key predictors of extinction risk. Model performance will be evaluated using confusion matrices comparing predicted outcomes to IUCN Red List labels, providing insight into classification accuracy and mislabelling patterns.

Lastly, the classification model's output will be used to estimate the year of extinction for each species. Predictions are to be derived by combining the model-generated labels with each species' last assessment date and mapping these onto a Weibull probability distribution. The Weibull distribution is particularly well-suited for extinction forecasting because it accommodates censored data, captures time-varying extinction risk, and produces interpretable survival probabilities that can inform interactive visualizations and guide conservation planning.

This study addresses a gap in existing research by integrating both intrinsic traits (e.g., body mass, geographic range size) and environmental or anthropogenic factors (e.g., hunting, deforestation) into a survival modeling framework to predict extinction risk (6; 8; 9; 16). The rationale for this integrated approach is rooted in strategic conservation needs: the dominant driver of risk dictates the most effective response, whether it is protecting vulnerable species groups (14) or pursuing ecosystem restoration (15). Evidence from various taxa shows how human pressures accelerate decline (12; 13), underscoring the necessity of including these factors. Therefore, by combining these drivers to generate time-to-extinction predictions and relative risk ratings, this study provides a strengthened basis for targeted conservation action (11; 12).

Proposed Method

The method uses data from the IUCN Red List database and was pulled using the *iucnredlist* library for R. Within the data extraction and transformation, 1.12GB of information was processed which included species with current red list categories in one of the following: Extinct, Critically Endangered, Endangered, Vulnerable and Nearly Threatened. Intrinsic data for species was downloaded from ESA's Ecological Archives (23; 24), Wolfram Research (26), AVONET (27) and AnAge (28) as .txt files, and converted to .csv format in Microsoft Excel with a combined size of 43.7MB. For data storage, this approach utilizes publicly available data and open-source machine learning libraries, which greatly reduce the cost and effort of large-scale extinction risk assessments (2). For storage, an S3 bucket was used in AWS which allows for minimal cloud storage costs; for data access, computing power, and collaboration a Google Colab Pro account was purchased for $10 USD. Only potential license requirements are for certain visualization tools.

These datasets were then further preprocessed and merged based on species' scientific names. Transformations applied include extracting the most recent IUCN Red List assessment, pivoting extrinsic variable columns, and standardizing intrinsic traits. One challenge encountered was the high sparsity of the intrinsic data after joining it with the extrinsic IUCN dataset based on scientific name. Intrinsic columns with 90% or greater missing values were excluded to reduce sparsity. Remaining intrinsic fields were evaluated hierarchically, beginning at the highest taxonomic level (e.g., Kingdom). Categories in which columns contained majority null or unpopulated values for species were excluded from analysis. This filtering process was repeated through four successive taxonomic levels. This approach enabled progressive elimination of sparsely populated categories while preserving finer taxonomic levels where intrinsic data density supported meaningful analysis. Missing trait values will be imputed using either the BHPMF-package or Rphylopars in R; imputing trait data is a known concern in phylogenetic studies and these tools have been shown to be effective (29; 30; 31). Once data have been imputed, logistic regression models will then be developed to quantify variable importance and compare the predictive influence of intrinsic (biological) versus extrinsic (environmental) factors in generating updated extinction labels. The total modeling dataset used for analysis is 2.3MB in size.

For visual mapping, and to mimic geographic range size, the number of locations per species as provided in the IUCN Red List dataset was included as an additional variable. All features were converted to numeric form using categorical encoding or ordinal scaling for use in survival modeling. "Last year seen" was manually pulled from the IUCN website for all extinct species, where available, and set as the Extinction year. For non-extinct species latest assessment date was set as the censored Extinction year.

Analysis began with a Cox Proportional Hazards model using CoxPHFitter from the lifelines Python package, chosen for its ability to handle censored data for species that are still extant. The model estimates the influence of intrinsic and extrinsic traits on survival probability

through 2025 by optimizing beta coefficients via a partial likelihood approach. Each species receives a relative risk score, which scales the average survival curve to produce species-specific survival probabilities over time.

The high dimensionality of the dataset, driven by one-hot encoding of categorical features, posed a challenge for the Cox Proportional Hazards (CPH) model, which requires numeric inputs. To address potential multicollinearity and overfitting, a penalized Cox regression was employed, shrinking some feature coefficients to zero and producing a sparser, more stable model. Despite this, only 3% of variables showed a statistically meaningful effect, likely due to residual multicollinearity confirmed by variance inflation factor (VIF) analysis. Model performance, assessed using the concordance index (C-index), was 0.5, indicating that the model ranked species by extinction order no better than random. These results suggest that, in this context, high feature correlation limits the utility of the CPH model for predicting extinction risk.

Given the limitations of the Cox Proportional Hazards model, the study next employed a Random Survival Forest (RSF) approach, selected for its ability to manage multicollinearity and capture nonlinear relationships. The RSF, implemented using the *RandomSurvivalForest* function from the *scikit-survival* package, constructs an ensemble of decision trees that randomly sample subsets of features at each split, thereby reducing dependence among correlated predictors. Unlike standard Random Forests, the RSF partitions data using the log-rank statistic to maximize differences between survival curves, with the Kaplan–Meier estimator applied to generate subgroup survival functions. This approach effectively handles censored species data and provides interpretable survival probabilities. The RSF achieved a high out-of-bag (OOB) concordance index of 0.896, confirming its superior predictive performance relative to the Cox model.

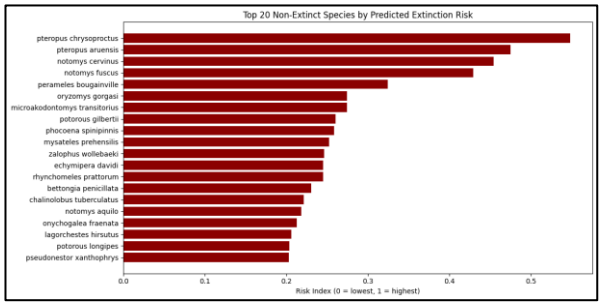

Figure 1. Bar chart displays the top 20 species with the highest predicted extinction evaluation risk, based on model-derived risk scores from a Random Forest Survival. Species such as *Pteropus Chyrsoproctus* and *Notomys Cervinus* show elevated scores, indicating greater vulnerability under current ecological conditions as compared to other non-extinct species.

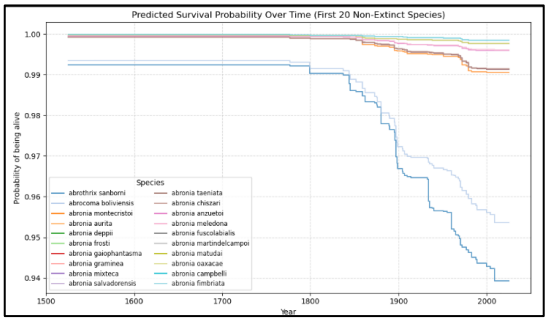

Figure 2. Line graph illustrates the modeled survival probabilities over a 525-year period for 20 species ending in 2025 identified from a Random Forest Survival. Species with steeper declines may face accelerated extinction trajectories without targeted conservation efforts.

The Random Survival Forest models estimated species-level extinction risk rankings, survival probabilities for 2025 and risk scores, the latter of which will then be subsequently integrated into logistic regression models to validate the importance of risk scores on predicting extinction. be applied to the Weibull probability distribution to forecast expected extinction years. Model performance will be assessed using accuracy metrics and concordance indices to ensure robustness and interpretability.

Since none of the previous modeling approaches yielded direct estimates of expected time to extinction, the analysis progressed to explore probability distribution-based methods capable of projecting beyond the most recent assessment year (2025). A Weibull-based Accelerated Failure Time (AFT) model was selected for its flexibility in modeling time-varying hazard rates and handling censored observations. This approach allows intrinsic and extrinsic covariates to accelerate or decelerate the time to extinction, enabling the prediction of median extinction times for species beyond 2025. Extinction labels refined from the logistic regression models were incorporated to improve prediction accuracy. By providing species-specific extinction timelines, this method offers actionable insights for prioritizing conservation interventions. This will allow proper analysis of large ecological shifts (17).

The final output of this study will be an interactive dashboard developed in Tableau, designed to integrate and visualize results from three modeling approaches: Random Survival Forest (RSF), logistic regression, and Weibull Accelerated Failure Time (AFT) survival models. The dashboard will enable users to manipulate intrinsic and extrinsic variables and observe their influence on predicted extinction timing, facilitating exploration of model dynamics and scenario-based conservation planning. Users will also be able to interact with RSF-derived extinction probabilities for the year 2025, aggregated across taxonomic hierarchies, to identify high-risk groups and prioritize intervention strategies based on class. In addition to model outputs, the dashboard will incorporate species' geographic ranges from the IUCN Red List, visualized as spatially mapped bubbles colored by extinction status. Interactive filters will allow users to explore extinction risk patterns across taxonomic levels, ecological traits, and threat categories. Hover functionality will reveal species-specific metadata—including extinction status, common and scientific names, geographic location, and RSF-estimated survival probability—supporting intuitive, real-time exploration of extinction vulnerability across species.

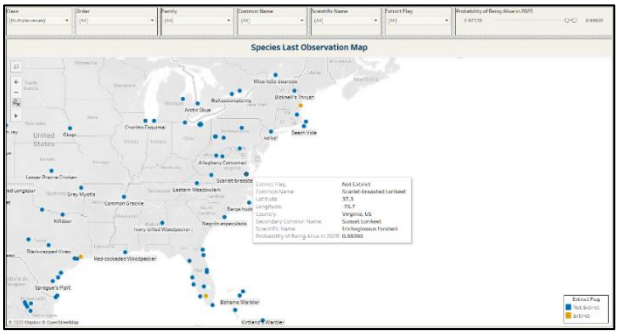

Figure 3. Interactive map of last seen appearances of each species. Each point represents the most recent recorded location of each species with tooltips. Filters (Kingdom, Class, Species, Scientific Name, Common Name, and Probability of being alive in 2025) allow users to refine the view, while the color-blind safe legend colors indicate if a species is extinct. This provides geographic and regional context to species' risk of extinction (19).

Evaluation

To evaluate extinction risk across species, a structured dataset was created by combining intrinsic biological traits, extrinsic environmental pressures, and IUCN Red List classifications. This dataset supported a series of experiments aimed at: (i) estimating species-specific survival probabilities for the year 2025 using a Random Survival Forest (RSF); (ii) generating updated extinction labels through logistic regression; and (iii) modeling predicted extinction years with a Weibull Accelerated Failure Time model. The combined approach enabled comparison between model-derived outputs and official IUCN classifications, while also allowing a sensitivity analysis to examine how changes in key input variables influence survival probability, extinction status, and projected extinction year via an interactive Tableau dashboard.

For all predictive models the following metrics were used to measure classification:

- Random Survival Forest: Concordance index (C-index) measures the probability that the model correctly orders pairs of species by extinction time; a value of 0.5 corresponds to random ranking, and values closer to 1 indicate better discrimination.
- Logistic Regression: Area under the ROC curve (AUC) where 0.5 denotes random classification and 1.0 perfect classification. Accuracy, recall, precision, F1-score and the number of false positives, computed from confusion matrices using a fixed probability threshold.
- Weibull Accelerated Failure Time: Measured through its concordance index (c-index) and Akaike Information Criterion (AIC). A c-index of 1.0 is considered perfectly modelled pair ranking and a value of 0.5 is equivalent to random chance. A lower AIC score generally indicates a better model.

Modeling and evaluation used a transformed and filtered version of the full dataset. The first model developed was a Random Survival Forest with hyperparameters tuned by GridSearchCV using 5-fold cross-validation. The RSF achieved an OOB C-index of 0.8899, indicating very strong discriminatory power and substantially outperforming the Cox Proportional Hazards model used earlier in the pipeline (C-index = 0.5) and surpassed similar tree-based ensemble methods from other studies (AUC score: 0.75-0.84) that used trait-based selection to quantify extinction risk (1). Using the RSF risk scores, a threshold corresponding to the proportion of extinct species in the dataset (1.7%) to identify the most at-risk species was applied. This approach flagged 37 species not currently listed as extinct, including the pteropus chrysoproctus (Ambon flying fox), pteropus aruensis (Aru flying fox), oryctolagus cuniculus (Coney), and pteropus cognatus (Makira Flying Fox) demonstrating the model's potential to inform proactive conservation efforts before species reach official at-risk status (Figure 1).

Logistic regression models were then developed to evaluate how intrinsic traits, extrinsic pressures, and the RSF-derived risk scores contribute to the probability of extinction. Three base models were constructed: one using only extrinsic variables, one using only intrinsic variables, and one combining both sets of predictors. Each model was estimated twice, first without the RSF risk scores and then with them. Hyperparameters were selected using LogisticRegressionCV with 5-fold cross-validation. Performance for all models was compared against the majority-class baseline using the metrics above. Among the models without RSF scores, the combined intrinsic-extrinsic model showed the best overall trade-off: it achieved the highest AUC (0.9234) and accuracy (0.7839), maintained a high recall of 0.92 on extinct species, and produced roughly 35% fewer false positives (315 vs. ~490–500) than the intrinsic-only and extrinsic-only models. These results indicate that extinction probability is more accurately captured when both intrinsic and extrinsic information are included. Previous studies using logistic regression for extinction risk usually relied on single, combined models that use both intrinsic and extrinsic predictors without evaluating their individual contributions (8; 9). For instance, Ripple et al. (8) used a combined model to identify body size as a risk factor, but did not separate the predictive power of biological traits from contemporary human threats.

In the combined logistic model, the absence of land or water management significantly increased extinction risk by 47.3% (odds ratio = 1.473), while its presence reduced risk by 32.1% (odds ratio = 0.679). Longer inter-birth intervals also lowered extinction probability by 30% (odds ratio = 0.697), emphasizing the role of conservation interventions and reproductive timing in species survival. Prior research has shown that intrinsic traits are dominant predictors of extinction (5). In one random forest classification, geographic range size was the most important

variable (importance = 0.058), followed by taxonomic order (0.027) and body mass (0.019) (5). Compared to results in this approach, range size may reflect the extent of habitat access enabled by conservation efforts. Other linear models have also identified reproductive rate as a key driver of species loss, aligning with the findings of this study (4). As a validation step, species' risk scores from the random forest survival model were included as a predictive variable in each of the logistic regression models. As expected, the risk scores dominated the models. The fact that the random forest-derived risk scores are such a powerful predictor in the logistic models primarily serves as validation of the random forest's performance, rather than providing new, independent findings from the logistic regression itself.

The Weibull accelerated failure time distribution was chosen because the covariates have a direct impact on the time-to-event, as opposed to a direct impact on the hazard rate. This allows this approach to estimate a time-to-extinction for each species. The model with a small penalty, which is recommended for right censored data, and a lasso regression returned a 0.883 concordance index. Suggesting that this Weibull distribution is a good fit to the data. Using a Weibull accelerated-failure-time distribution is better suited than the Weibull distributions used in other approaches as it allows the model to account for the covariate effects for each species, instead of the entire population. Allowing for time-to-event predictions for each individual species. In other studies, Clements et al. (2013) evaluated a Weibull-based optimal linear estimator to infer extinction time from species sighting. While this approach had higher r-squared performance for ideal conditions, it relies heavily on consistent long-term search effort; data that was not as prevalent in this study (21).

To further explore and communicate key findings, an interactive Tableau dashboard was developed with a focus on user engagement and species conservation awareness. The dashboard features a global map view that visualizes species distributions based on longitudinal coordinates, enabling users to distinguish between currently extant and extinct species. This spatial representation highlights geographic regions with concentrated assessment data, reveals biodiversity hotspots and extinction-prone climates, and facilitates exploration of taxonomic patterns in relation to location (Figure 3). To retrieve said spatial data, a python function was used to pull the last seen location of each species from Global Biodiversity Information Facility and iNaturalist (34; 35). In addition to spatial insights, the dashboard integrates species-level risk scores generated by the random forest model, organized by taxonomic order and common name to provide a consolidated view of extinction vulnerability (Figure 4). Common names were extracted by selecting the first available IUCN common name field for each species. However, because the analysis spans a global dataset, many of the common names appeared in non-English languages such as Chinese and Portuguese. To ensure consistency, a Google Cloud Translation API workflow was implemented to automatically translate all first common names into English. Outputs from the Random Survival Forest, logistic regression, and Weibull AFT models are also linked to interactive visualizations, allowing users to conduct scenario-based analyses in near real time (Figure 5). The dashboard variables were selected based on their relative importance in predicting extinction risk, as identified by the logistic regression model. By prioritizing the most influential intrinsic and extrinsic traits, the dashboard enables users to explore how changes in these key parameters affect species vulnerability, facilitating targeted scenario analysis and conservation planning. Feedback from a small user study described the dashboard as "fun," "engaging," "educational," and "in-depth," highlighting its effectiveness in fostering a deeper understanding of both the focal species and the key factors influencing extinction risk.

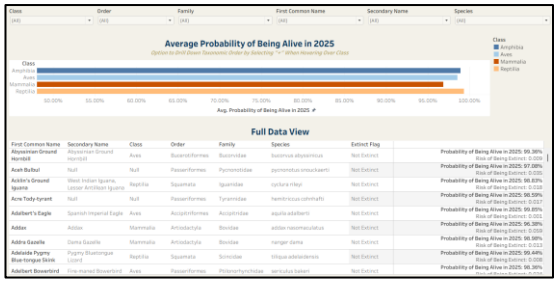

Figure 4. Interactive view of data output from the Python model. Image shows each species Probability of Being Alive in 2025, as well as a scaled Risk of Being Extinct Score. Details are also provided on taxonomic ranks for each species, as well as if the species is extinct as of today. A top-level view across animal Class is also provided, showing how each class ranks on its average probability of being alive in 2025.

Figure 5. Interactive dashboard that displays multiple model outputs for a user to select species. The image shows the Weka bird, scientifically known as *gallirallus australis*, and the probability of surviving in 2025, the predicted extinction label, and predicted year of extinction based on user set parameters. Each model takes in user input on animal lifespan, litter/clutches per year, Land/Water Protection & Management, Interbirth interval, and Forest/Habitat.

Data utilized to generate evaluation scores was sourced from the IUCN Red List and multiple trait databases (ESA, Wolfram, AVONET, AnAge), totaling over 1.16GB. After merging by species name, the dataset was cleaned to remove intrinsic traits with over 90% missing values and filtered hierarchically across taxonomic levels to retain meaningful columns. Missing intrinsic values were imputed using phylogenetic modeling via Rphylopars. This package utilized a phylogenetic tree generated from the Open Tree of Life (32) and matched with the species in the dataset; this eliminated 100 species for which there was no match in the Open Tree of Life tree. Previous studies often handled

missing values by applying mean substitution or generic model-based estimates (7), by using phylogeny to generate maximum-likelihood estimates, the Rphylopars approach produced species-specific estimates that aligned more closely with evolutionary context, minimizing noise from unrelated taxa and enhancing the overall relevance of the imputed data. Because the R package is designed for continuous data, categorical intrinsic traits were excluded from analysis. The final dataset of 2.3MB included 4,890 species across 4 classes, each with up to 44 standardized intrinsic traits and 26 one-hot encoded extrinsic features. Due to the heavy imbalance between non-extinct and extinct species, a stratified sampling method was used in all three models. Extinction timing was derived from IUCN records and transformed into a duration variable for survival modeling. Data were stored in an AWS S3 bucket.

The experiment used a combination of open-source tools and cloud-based platforms. Data preprocessing and merging were performed in Python using Google Colab Pro, with packages like pandas, numpy, duckdb, and boto3. Survival modeling was conducted using Python libraries: scikit-survival and scikit-learn for Random Survival Forests with GridSearchCV tuning, and lifelines and statsmodels for Weibull AFT modeling. Logistic regression was implemented with sklearn and matplotlib (for visualization) Python packages. Visualization was done in Tableau Desktop, integrated with Python models via TabPy, allowing real-time updates based on user input.

Conclusions and Discussion

The purpose of this study was to develop a multi-stage analytical framework to assess extinction risk, reclassify species status, and estimate extinction timing by integrating survival analysis, classification and forecasting to provide a more detailed, forward-looking understanding of species vulnerability.

Across this modeling framework the outcomes were a robust risk ranking with very high predictability (0.8899 OOB c-index) and identifying potentially high-risk species, identifying the combined intrinsic and extrinsic features (land and water management and interbirth interval) that most influence extinction risk through a combined model that outperformed (AUC: 0.9234) single-feature type models and a novel parametric forecast through a Weibull Accelerated Failure Time model.

Despite these advances, several limitations should be noted. Sparse data was a frequent issue for species' intrinsic traits and imputation was needed to mitigate this; however, it also introduces risk of noise and bias for species with limited trait information. Further, to avoid imputing nonsensical figures for species (such as egg length/beak length for mammals), traits were limited to those shared by all classes in the study (Reptilia, Mammalia, Aves, Amphibia). Future research might investigate class-specific predictors of extinction. Additionally, the high dimensionality of the data made collinearity a major concern with VIF values being used to reduce the feature set prior to regression, however appropriate caution should be taken. Finally, the chained modeling approach, where risk scores from one model feed into another, raises the potential for compounding errors or overfitting despite the cross-validation safeguards used. Another limitation in observing the 2025 survival probabilities was the heavy skew toward extant species (98% of the dataset was censored) resulting in most survival probabilities exceeding 95%. Although extinct and extant species were stratified across training and test sets, future iterations would benefit from a more balanced representation to yield a broader and more informative range of survival estimates. However, sourcing well-documented data on extinct species remains a challenge due to limited historical records.

The findings from this study can be used to support more proactive conservation planning. The risk rankings from the Random Survival Forest, the probability estimates from the logistic regression models, and the projected extinction timelines from the Weibull model jointly provide an evidence-based foundation for identifying species that warrant early intervention. The framework also offers a scalable analytical tool for institutions that seek to monitor global extinction risk using large trait and threat datasets. In the long term, this approach may help shift conservation decision-making toward earlier detection and prevention rather than reactive response. The fact that water protection and management were among the strongest predictors is especially important, as ecosystems facing water scarcity are among the most vulnerable under climate change. Recent work shows that declining moisture availability amplifies ecological stress, reduces resilience, and increases the likelihood of system-wide collapse (33).

Future extensions of this work include making the Tableau dashboard publicly accessible and packaging the full solution for streamlined deployment across platforms. Establishing a direct, regularly updated connection to the IUCN Red List would enhance model responsiveness to new assessments. Sharing the tool with IUCN could support a more interactive and informative user experience on their site. Additionally, evaluating the refined dataset across alternative modeling approaches, such as those from literature referenced, would further benchmark performance and assess robustness.

All team members contributed meaningfully to the project, with responsibilities distributed across key phases of the workflow. Timothy Padden focused on data extraction and modeling; Jacob Wolfson contributed to data cleansing; Solomon Flax contributed to data extraction, cleansing, and modeling; Christopher Cardwell supported both data extraction and cleansing; Parth Desai focused on the data visualization efforts; and Alyssa Ray played a multifaceted role across data cleansing, modeling, and visualization. While individual roles varied in scope, the overall distribution of effort was balanced, reflecting collaborative engagement across technical and analytical domains.

References

1. Foster WJ, Ayzel G, Münchmeyer J, et al. Machine learning identifies ecological selectivity patterns across the end-Permian mass extinction. *Paleobiology*. 2022;48(3):357-371. doi:https://doi.org/10.1017/pab.2022.1
2. Zizka A, Andermann T, Silvestro D. IUCNN – Deep learning approaches to approximate species' extinction risk. *Diversity and Distributions*. 2022;28:227-241. doi:https://doi.org/10.1111/ddi.13450
3. Dueñas MA, Hemming DJ, Roberts A, Diaz-Soltero H. The threat of invasive species to IUCN listed critically endangered species: A systematic review. *Global Ecology and Conservation*. 2021;26:e01476. doi:https://doi.org/10.1016/j.gecco.2021.e01476
4. Michael Ellman Soulé, University Of Cambridge. *Viable Populations for Conservation*. Cambridge University Press; 1996:35-57.
5. Davidson AD, Shoemaker KT, Weinstein B, et al. Geography of current and future global mammal extinction risk. Kamilar JM, ed. *PLOS ONE*. 2017;12(11):e0186934. doi:https://doi.org/10.1371/journal.pone.0186934
6. Chichorro F, Juslén A, Cardoso P. A review of the relation between species traits and extinction risk. *Biological Conservation*. 2019;237:220-229. doi:https://doi.org/10.1016/j.biocon.2019.07.001
7. Ma X, Dong R, Hughes A, Corlett RT, Jens-Christian Svenning, Feng G. Population trends are more strongly linked to environmental change and species traits in birds than mammals. *Proceedings of the Royal Society B Biological Sciences*. 2024;291. doi:https://doi.org/10.1098/rspb.2024.1395
8. Ripple WJ, Wolf C, Newsome TM, Hoffmann M, Wirsing AJ, McCauley DJ. Extinction risk is most acute for the world's largest and smallest vertebrates. *Proceedings of the National Academy of Sciences*. 2017;114(40):10678-10683. doi:https://doi.org/10.1073/pnas.1702078114
9. Mashayekhi M, MacPherson B, Gras R. A machine learning approach to investigate the reasons behind species extinction. *Ecological Informatics*. 2014;20:58-66. doi:https://doi.org/10.1016/j.ecoinf.2014.02.001
10. Branco VV, Correia L, Cardoso P. The use of machine learning in species threats and conservation analysis. *Biological conservation*. 2023;283:110091-110091. doi:https://doi.org/10.1016/j.biocon.2023.110091
11. Purvis A, Gittleman JL, Cowlishaw G, Mace GM. Predicting extinction risk in declining species. *Proceedings of the Royal Society of London Series B: Biological Sciences*. 2000;267(1456):1947-1952. doi:https://doi.org/10.1098/rspb.2000.1234
12. Borges CM, Terribile LC, de Oliveira G, Lima-Ribeiro MS, Dobrovolski R. Historical range contractions can predict extinction risk in extant mammals. *PLoS ONE*. 2019;14(9):e0221439-e0221439. doi:https://doi.org/10.1371/journal.pone.0221439
13. Cardoso P, Barton PS, Birkhofer K, et al. Scientists' warning to humanity on insect extinctions. *Biological Conservation*. 2020;242:108426. doi:https://doi.org/10.1016/j.biocon.2020.108426
14. Hardesty-Moore M, Deinet S, Freeman R, et al. Migration in the Anthropocene: how collective navigation, environmental system and taxonomy shape the vulnerability of migratory species. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2018;373(1746):20170017. doi:https://doi.org/10.1098/rstb.2017.0017
15. Prugh LR, Hodges KE, Sinclair ARE, Brashares JS. Effect of habitat area and isolation on fragmented animal populations. *Proceedings of the National Academy of Sciences*. 2008;105(52):20770-20775. doi:https://doi.org/10.1073/pnas.0806080105
16. Lee Jay Hannah. *Saving a Million Species : Extinction Risk from Climate Change*. Island Press; 2012.
17. Botkin DB. *The Moon in the Nautilus Shell : From Climate Change to Species Extinction, How Life Persists in an Ever-Changing World : Discordant Harmonies Reconsidered*. Oxford University Press; 2012.
18. Bernhard Grzimek, Macleod N, J David Archibald, Levin PS. *Grzimek's Animal Life Encyclopedia : Exintction*. Gale/Cengage Learning; 2013.
19. Tek Raj Bhatt, J. Guy Castley, R Sims-Castley, Hem Sagar Baral, Alienor L. M. Chauvenet. Connecting tiger (Panthera tigris) populations in Nepal: Identification of corridors among tiger‐bearing protected areas. *Ecology and Evolution*. 2023;13(5). doi:https://doi.org/10.1002/ece3.10140
20. Solow AR. Inferring Extinction from Sighting Data. *Ecology*. 1993;74(3):962-964. doi:https://doi.org/10.2307/1940821
21. Clements CF, Worsfold NT, Warren PH, et al. Experimentally testing the accuracy of an extinction estimator: Solow's optimal linear estimation model. *Journal of Animal Ecology*. 2012;82(2):345-354. doi:https://doi.org/10.1111/1365-2656.12005
22. Jarić I, Ebenhard T. A method for inferring extinction based on sighting records that change in frequency over time. *Wildlife Biology*. 2010;16(3):267-275. doi:https://doi.org/10.2981/09-044
23. Jones K, Bielby J, Cardillo M, et al. Ecological Archives E090-184. Esapubs.org. Published 2009. Accessed October 30, 2025. https://esapubs.org/archive/ecol/E090/184/#data
24. Wilman H, Belmaker J, Simpson J, de la Rosa C, Rivadeneira MM, Jetz W. Ecological Archives E095-178-metadata. Esapubs.org. Published 2014. Accessed October 31, 2025. https://www.esapubs.org/archive/ecol/E095/178/metadata.php
25. Dasgupta P. *The Economics of Biodiversity: The Dasgupta Review*.; 2021:431. https://assets.publishing.service.gov.uk/media/602e92b2e90e07660f807b47/The_Economics_of_Biodiversity_The_Dasgupta_Review_Full_Report.pdf

26. Myhrvold NP, Baldridge E, Chan B, Sivam D, Freeman DL, Morgan SK. Amniote Life History Database | Wolfram Data Repository. Wolframcloud.com. Published January 4, 2016. Accessed October 31, 2025. https://datarepository.wolframcloud.com/resources/Amniote-Life-History-Database/
27. Tobias JA, Sheard C, Pigot AL, et al. AVONET: morphological, ecological and geographical data for all birds. Coulson T, ed. *Ecology Letters*. 2022;25(3):581-597. doi:https://doi.org/10.1111/ele.13898
28. DE MAGALHÃES JP, COSTA J. A database of vertebrate longevity records and their relation to other life-history traits. *Journal of Evolutionary Biology*. 2009;22(8):1770-1774. doi:https://doi.org/10.1111/j.1420-9101.2009.01783.x
29. Goolsby EW, Bruggeman J, Ané C. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. Fitzjohn R, ed. *Methods in Ecology and Evolution*. 2016;8(1):22-27. doi:https://doi.org/10.1111/2041-210x.12612
30. Johnson TF, Isaac NJB, Paviolo A, González-Suárez M. Handling Missing Values in Trait Data. Schrodt F, ed. *Global Ecology and Biogeography*. 2020;30(1):51-62.
31. Moura MR, Ceron K, Guedes M, et al. A phylogeny-informed characterisation of global tetrapod traits addresses data gaps and biases. *PLoS Biology*. 2024;22(7):e3002658-e3002658. doi:https://doi.org/10.1371/journal.pbio.3002658
32. Michonneau F, Brown JW, Winter DJ. rotl: an R package to interact with the Open Tree of Life data. *Methods in Ecology and Evolution*. 2016;7(12):1476-1481. doi:10.1111/2041-210X.12593
33. Clarke H, Nolan RH, De Dios VR et al. Forest fire threatens global carbon sinks and population centres under rising atmospheric water demand. *Nature Communications*. 2022;13(716). doi:https://doi.org/10.1038/s41467-022-34966-3
34. GBIF Home Page. Gbif.org. Published 2025. Accessed November 21, 2025. https://techdocs.gbif.org/en/openapi/
35. iNaturalist. iNaturalist.org. iNaturalist.org. Published 2025. https://www.inaturalist.org/