Languages at Risk: A Machine Learning Approach to Predicting Language Endangerment

Helin Yilmaz (helinyilmaz@berkeley.edu) | Courtney Chen (courtney_chen@berkeley.edu)
Brian Woods (brianosaurus@berkeley.edu) | Jordan Andersen (jordanandersen@berkeley.edu)

Abstract:
This machine learning approach to predicting language endangerment attempts to answer the question: What makes a language endangered, and can we predict and anticipate these patterns before it's too late? Our baseline model, which predicts only the majority class, achieves a test accuracy of 43%. After experimenting with gradient boosting, ensemble, extra trees and neural network models with hyperparameter tuning, results show that the gradient boosting and ensemble models achieve the highest performance with test accuracies of 88.3%. Our research finds that further exploration of additional socioeconomic, geographic and historical features may further improve the accuracy, precision and recall of the models.

Introduction:
What makes a language endangered, and can we predict this before it's too late? This question is compelling because language endangerment reflects broader cultural and social challenges. More than 40% of the world's 7,000 languages are at risk of disappearing, according to the Endangered Languages Project. As languages vanish, so do unique histories, identities, and worldviews. Many of these languages are spoken by communities with limited access to education, media, and political support. By identifying the key factors that lead to endangerment, we aim to support earlier and more focused efforts to preserve linguistic and cultural diversity.

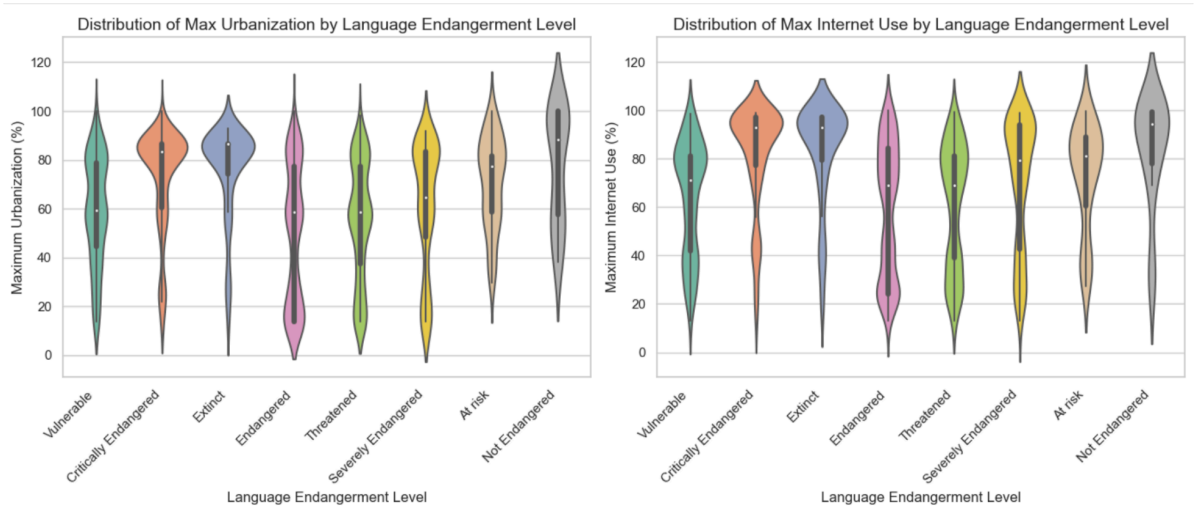
Our goal is to predict a language's level of endangerment, ranging from extinct to not endangered, based on several features. The models are trained on features such as speaker counts, the number of countries where the language is spoken, political status, and the maximum urbanization and internet usage rates in those countries. We apply a range of predictive models to determine the output variable: the classification of the language's endangerment category.

Related work:
Bromham et al.[1] analyzes 51 predictor variables to understand the drivers of language endangerment globally, focusing on language "maintenance" at a global scale. Using a four-level classification with the categories Threatened, Endangered, Critically Endangered, and Sleeping, the authors identify five key predictors of language loss that hold at both global and regional levels: number of L1 speakers, bordering language richness, road density, years of schooling, and nearby endangered languages. Their best-fit linear regression model explains 34% of the variation in language endangerment, with the number of L1 speakers emerging as the strongest predictor, similar to our own findings. The study critiques the assumption that political recognition or minority education policies inherently protect language diversity, noting that such policies often aim at assimilation rather than maintenance. Additionally, the study acknowledges the key limitation of potentially overlooking historical context when relying on current predictors. Ultimately, the work emphasizes that identifying the broader patterns, as well as regional influences, in language vitality can help better target efforts to preserve linguistic diversity.

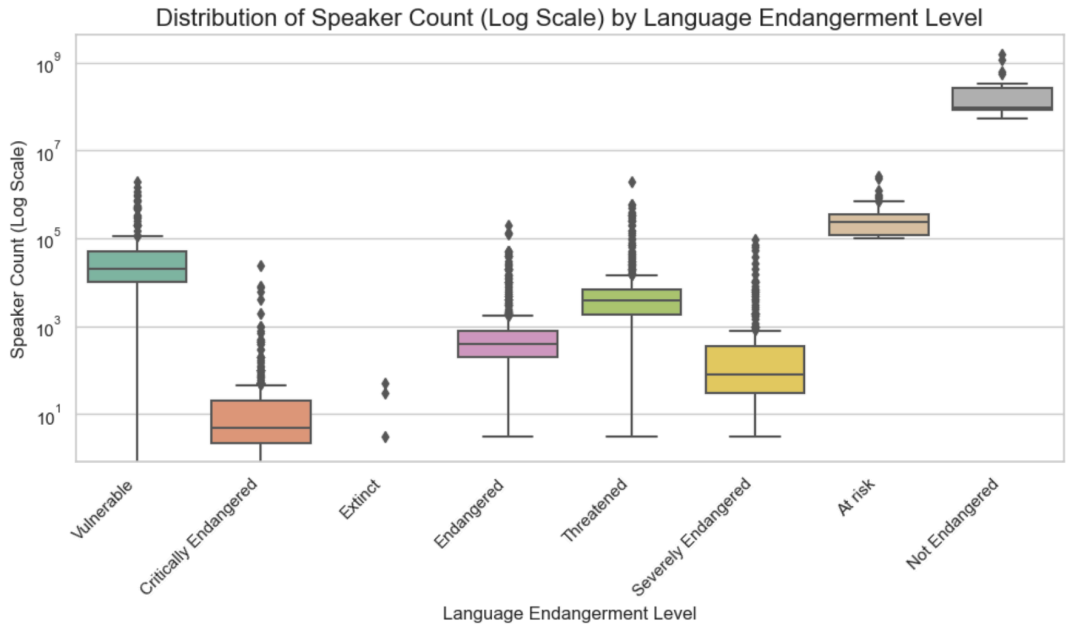
Dwivedi[2] focuses on endangered languages in India and uses linear regression to predict when a language may go extinct. Extinction is estimated by tracking the decline in the proportion of speakers relative to the total population over time. By modeling these trends, the authors aim to forecast the point at which a language's speaker base becomes unsustainably small. While the approach is data-driven, it relies on a simplified definition of extinction and is limited by the availability and granularity of historical language data. Nonetheless, it offers a valuable methodological contribution to forecasting language loss at a national scale.

Dataset:
Our data comes from a cross-sectional dataset compiled with indicators and languages from The Endangered Languages Project, the World Bank, and Ethnologue. We addressed missing values by dropping rows where data was incomplete in consideration of how this would affect the distribution of the data. If the language was “Extinct” and did not have a speaker count, we imputed the speaker count with 0. After pre-processing, the original 3,466 rows were reduced down to 2,870. Since outliers, such as widely spoken languages, were considered informative to our model, we applied log scaling to the speaker count to reduce skew while retaining the full view of language vitality. We also label-encoded the outcome feature, *endangerment level*, for modeling. To improve class balance and account for the unpredictability of revitalization efforts, we merged the "Awakening" and "Dormant" categories into a single "Extinct" group. For feature engineering, we created numeric columns (e.g., “official”, “regional”, etc.) that count the number of countries in which a language holds each legal status. Since multiple countries could correspond to a language, we took the maximum urbanization and internet use rates to reflect a language’s highest potential exposure to modernization. We selected features based on data integrity and completeness, as well as their relevance to language endangerment.

EDA Insights:
Distributions: ‘Not Endangered’ languages tend to have high speaker counts, urbanization, internet use, and broader geographic spread (number of countries). In contrast, other endangerment levels are more concentrated and exhibit lower values across these features.



Correlations: Speaker count and country count show strong positive associations with language vitality. While internet use and urbanization are less predictive overall, ‘Not Endangered’ languages display a slight positive trend with internet access, suggesting some supportive role.



Features: Speaker count and country count are the most reliable indicators of lower endangerment. Urbanization and internet use do not directly prevent language loss but may aid in preservation or documentation, especially for 'Critically Endangered' or 'Extinct' languages.

PCA: Principal Component Analysis reveals a broad separation between 'Not Endangered' languages and the rest, with endangered levels forming a dense, overlapping cluster based on the selected features. This indicates that while these features can identify general endangerment, they are less effective for finer distinctions within the endangered spectrum. See Appendix D.

Methods:

Baseline model:

The baseline model uses majority class prediction, always assigning the most frequent endangered language category found in the training data. This simple approach is appropriate as a reference point to evaluate more advanced models' improvements, especially given the class imbalance and multiple endangered categories. It works by ignoring input features entirely and predicting the dominant class, resulting in low accuracy on minority classes but providing a clear performance floor to compare against. Its simplicity helps demonstrate the difficulty of the classification task and justifies the need for models that can learn patterns from linguistic and sociocultural features.

Gradient Boosting model:

Gradient Boosting is an ensemble algorithm that builds sequential decision trees, with each tree correcting the errors of the previous one. It is appropriate for predicting language endangerment because it captures non-linear relationships well, such as the diminishing impact of gaining additional speakers as a language becomes more widespread. Combined with SMOTE to address class imbalance, it offers a strong predictive performance with flexible hyperparameter tuning, making it a substantial improvement over simpler baselines.

Ensemble Model:

Ensemble models incorporate multiple models such as random forests, gradient boosting and extra trees, then use methods such as voting, bagging or boosting which aim to generate the best possible estimator for the task.This approach is well-suited to predicting language endangerment because it reduces overfitting, handles noisy and imbalanced data better, and is more resilient to outliers such as the few extinct languages in the dataset. By

aggregating the strengths of different models, ensemble methods improve generalization from training to unseen data.

Neural Network Model:

A neural network is a multi-layered model that processes inputs through successive transformations involving weighted sums and non-linear activation functions. This architecture allows it to learn complex, non-linear patterns within the data. In this context, the model uses engineered features derived from linguistic prevalence, digital presence, and socioeconomic metrics to classify language endangerment levels. Its flexibility makes it appropriate for capturing interactions between variables in a multi-class classification setting.

Extra Trees model:

Extra Trees is appropriate for language endangerment classification because it adds more randomization than Random Forest, which can help when dealing with potentially noisy or incomplete data. The additional randomness can better handle the inherent uncertainty in language assessment and may discover unexpected patterns in the relationship between features and endangerment status.

Experiments, Results and Discussion:

The baseline model's performance included a 26.0% test accuracy for granular classification and a 43.0% test accuracy for grouped classification. All of the following models improved from this baseline.

Gradient boosting model (experiments):

For the Gradient Boosting model, we tested 50, 100, and 200 estimators; depths of 1, 3, and 5; learning rates of 0.05 and 0.1; and subsample ratios of 0.6 and 0.8, using validation accuracy as our primary metric. Initial models showed signs of overfitting, so we grouped the endangerment levels, which significantly improved generalization. The best configuration (50 estimators, depth of 3, learning rate of 0.1, and subsample of 0.8) achieved a test accuracy of 88% and a validation accuracy of 90%. The final model's balance of simpler trees, substantial learning improvements, and randomness allowed the model to perform and generalize well (Appendix A). For subgroup evaluation, we define a language as legally recognized if it is classified as official or national in at least one country. The tuned model achieves a higher test accuracy on legally recognized languages (96%) than on unrecognized languages (88%) (Appendix B), despite legally recognized languages representing only ~5% of the test set. This suggests that legally recognized languages are easier to classify, likely due to distinctive patterns associated with them.

Ensemble model (experiments):

We tested both voting and bagging ensemble methods using the engineered features from the dataset. For the voting approach, we implemented both hard and soft voting, evaluating models with and without class imbalance handling via SMOTE. For bagging, we explored various imbalance strategies, including SMOTE, SMOTETomek, and class weighting. We used grid search to tune hyperparameters such as the number of estimators, maximum tree depth, and learning rate. Additionally, we evaluated classification performance using both fine-grained levels and grouped categories, where critically and severely endangered languages were combined, as were at-risk, vulnerable, and threatened languages. Grouping these classes improved overall model accuracy. The best-performing configuration was a hard voting ensemble with grouped classification levels. It combined a gradient boosting model with a learning rate of 0.01, a max depth of 5, and 200 estimators; an extra trees model with a max depth of 15 and 100 estimators; and a random forest with a max depth of 10 and 200 estimators. None of these models used class weighting. This ensemble significantly improved generalization to unseen data, reducing overfitting compared to the untuned model, and achieved a training accuracy of 91.1% and a test accuracy of 88.3%. The confusion matrix results indicating the performance by category can be seen in Appendix E.

Neural Network Model (experiments):

We trained and evaluated neural network models to predict language endangerment using two label schemes: granular (8 classes) and grouped (5 classes). The baseline model consisted of two dense layers with 128 and 64

units, using ReLU activations and dropout, trained with the Adam optimizer and early stopping. It achieved test accuracies of 72.8% on granular labels and 87.6% on grouped labels, but showed lower recall for minority classes such as "At Risk" and "Severely Endangered." We then applied hyperparameter tuning using Keras Tuner, optimizing layer sizes (160 and 128 units), dropout rates (0.3 and 0.2), and learning rate (0.001). The best tuned model was retrained with SMOTE to address class imbalance. Although the tuned model showed slightly lower overall accuracy (74.2% granular test accuracy and 85.7% grouped test accuracy), it maintained or improved recall for important minority classes. For example, recall for "At Risk" remained high in the grouped scheme and modestly improved for "Severely Endangered" in the granular scheme. Overfitting was controlled via validation monitoring and dropout, and detailed subgroup performance was assessed through class-wise recall and F1 scores.

Extra Trees model (experiments):
We utilized an exhaustive grid search across four key hyperparameters. The tuning space comprised 192 unique parameter combinations (4 × 4 × 3 × 3), with n_estimators ranging from 100 to 500 trees to balance computational cost with ensemble strength. The max_depth parameter included None to allow fully grown trees, while constrained options (10, 15, 20) served to control model complexity. The best-performing model (100 estimators, depth of 15, 5 minimum number of samples split, and 1 minimum number of samples at a leaf node) achieved a final train accuracy of 91.2%, a validation accuracy of 89.4%, and a test accuracy of 87.8%. This model was optimal because it balanced complexity and generalization, achieving high accuracy across train, validation, and test sets without overfitting.

Conclusions:
Final model selection:
Our models demonstrate strong predictive performance, with the number of speakers consistently identified as the most influential feature across all approaches. Grouping similar endangerment categories improved classification performance and reduced overfitting. The best-performing models were the Gradient Boosting classifier and the hard voting ensemble, which achieved test accuracies of 88.0% and 88.3% respectively. Their success can be attributed to a combination of feature learning, ensemble diversity, and reduced class imbalance from label grouping.

A key limitation of this work is the lack of temporal, historical, and policy-related variables that likely influence language survival, but are not captured in the dataset. Additionally, underrepresented classes posed challenges for recall, particularly in the granular classification scheme. If more time or computational resources were available, future improvements could include incorporating longitudinal speaker data, socio-economic indicators, and geographic features. These additions may help capture deeper patterns behind language endangerment and improve performance beyond the current 88% threshold.

Contributions:
Helin: Created both the Baseline Model notebook to establish a baseline for classification performance using the majority class prediction, and the Neural Network notebook which expanded the baseline model to include neural network classification on endangered languages. Performed hyperparameter tuning of the neural network using Keras Tuner. Evaluated both models and summarized the results in the Methods and Experiments sections. Contributed to the Introduction, EDA, and completed the Conclusion.

Brian: Wrote code for initial models and experimented with number of language-specific domains (TLDs) as an input feature for modeling. Contributed to data preparation by performing SMOTE balancing and creating interaction features. Built and tuned the Extra Trees model and contributed to the final report.

Courtney: Created the Data Cleaning notebook, which involved cleaning five datasets and merging them into a single master dataset. Performed Gradient Boosting hyperparameter tuning to identify optimal parameters and

evaluated model performance and generalization. Wrote the Gradient Boosting sections in the Methods and Experiments portions of the report and contributed to the final presentation.

Jordan: Contributed to the exploratory data analysis. Conducted hyperparameter tuning for the Ensemble Model and evaluated its performance and generalizability. Summarized the results in the Methods and Experiments sections and contributed to the Abstract, Introduction and Related Work sections.

References:

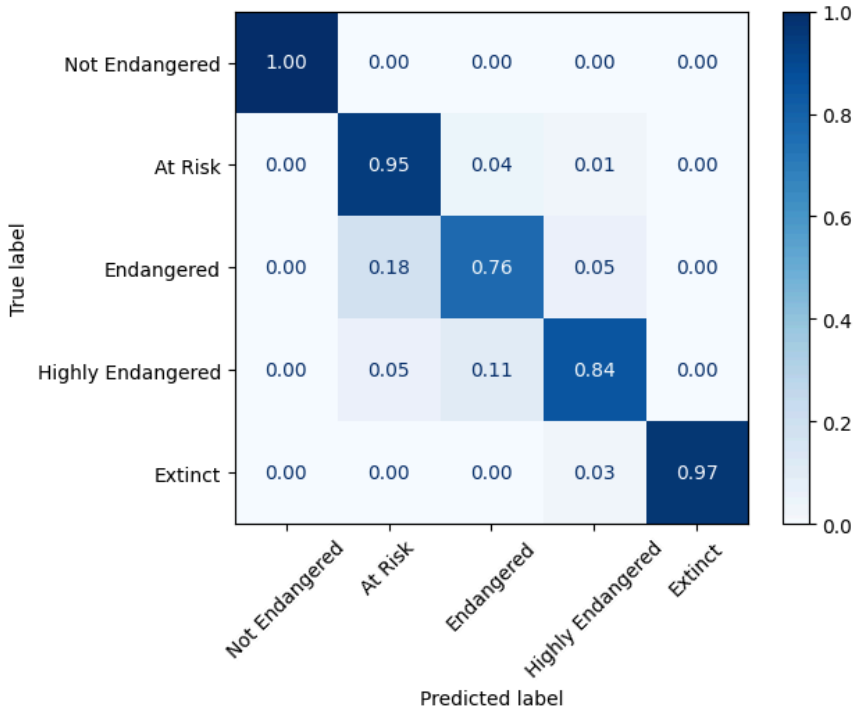
1. Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., & Hua, X. (2022). Global predictors of language endangerment and the future of linguistic diversity. Nature Ecology & Evolution, 6(2), 163–173. https://doi.org/10.1038/s41559-021-01604-y

2. Dwivedi, P. (2020). Predicting language endangerment: A machine learning approach. In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 147–153). https://doi.org/10.1109/ICCCNT49239.2020.9225576

Appendix :

Appendix A: Gradient Boosting Model - Confusion Matrix

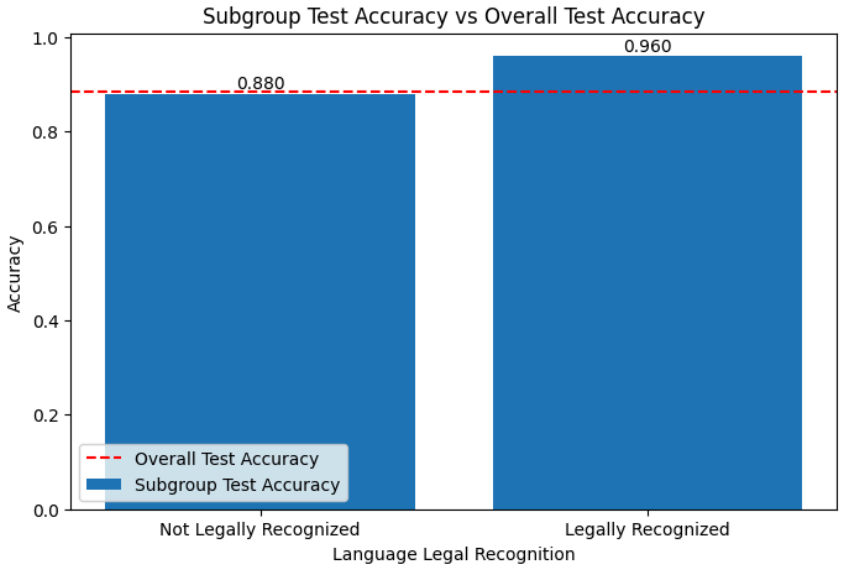
The normalized confusion matrix reveals a clear performance pattern across classes that aligns with expectations. We observe that the model performs best on the most distinct, "extreme" categories, Not Endangered and Extinct, while accuracy decreases as output labels approach the middle of the endangerment scale. Endangered, the central level, is most frequently misclassified, with errors primarily occurring between neighboring classes. This suggests that "middle" labels may have greater feature overlap, making adjacent categories harder for the model to distinguish.



Appendix B: Gradient Boosting Model - Subgroup Evaluations

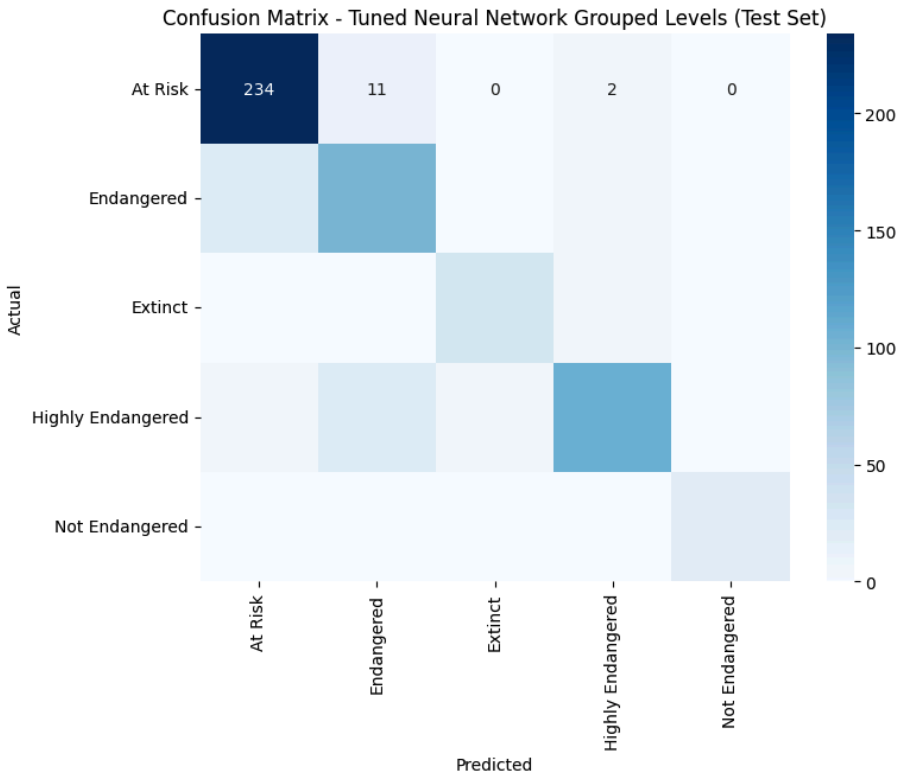
The following graph compares the model performance on legally vs. non-legally recognized languages, where "legally recognized" includes all languages that are "official" or "national" in one or more countries. The substantial

difference in test accuracy suggests that the model generalized better to legally-recognized languages, potentially due to better data availability, higher data quality, and reduced variability compared to non-recognized languages.



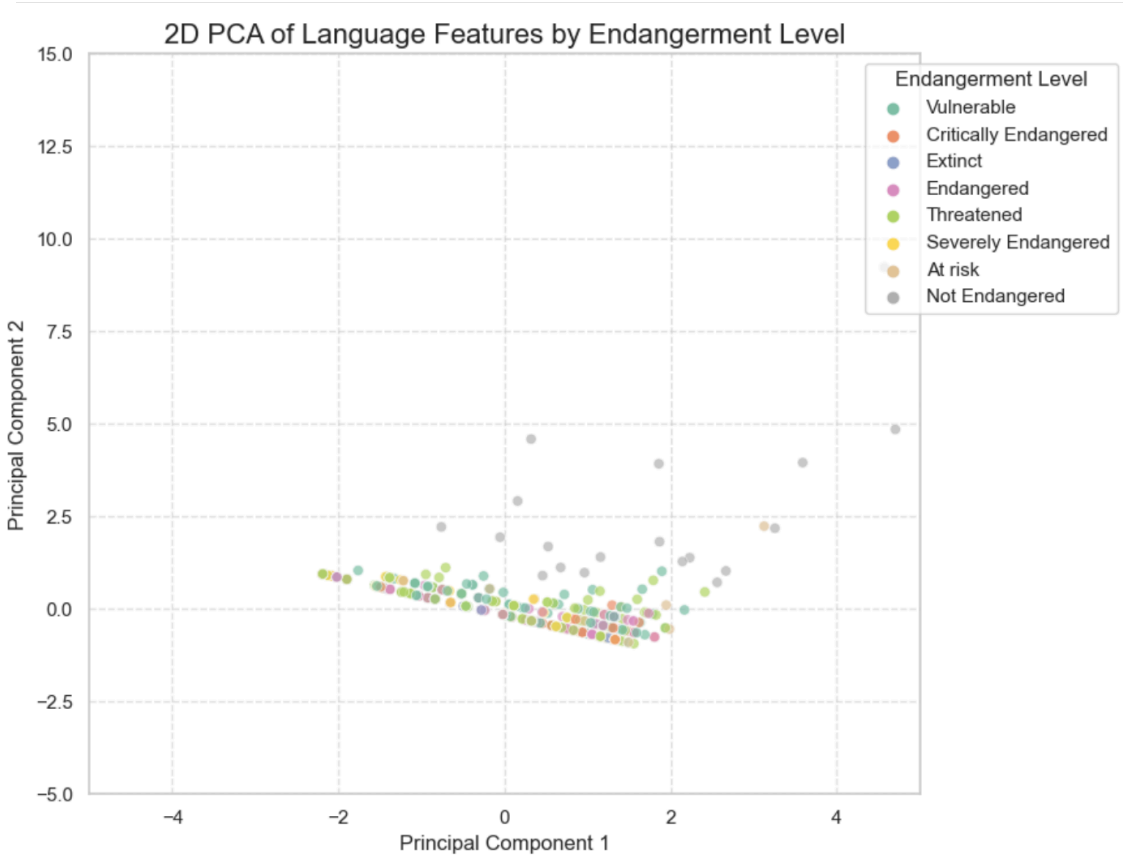
Appendix C: Confusion Matrix Performance for Tuned and Grouped Neural Network Model

The confusion matrix shows strong performance on the "At Risk" class, with 234 out of 247 species correctly classified (recall = 0.95). "Not Endangered" was perfectly predicted, with all 18 species correctly labeled. Performance was lower for "Highly Endangered," with a recall of 0.74 and 38 misclassified species, mostly as "At Risk" or "Endangered." This suggests difficulty distinguishing between mid-spectrum categories. Overall, the model performs best on the most distinct classes at either end of the risk scale.



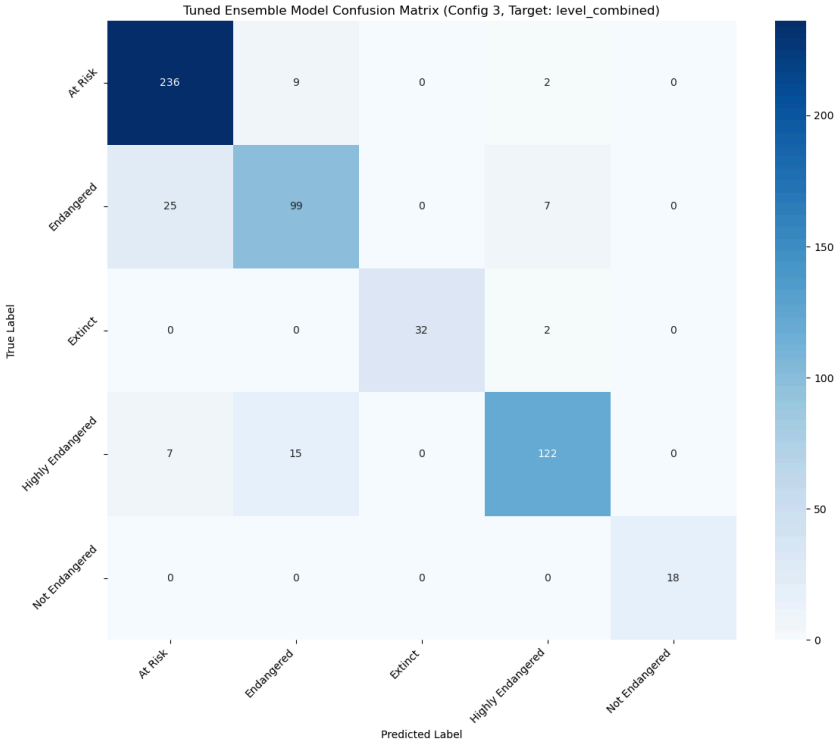
Appendix D: Principal Component Analysis of Language Features by Endangerment Level

The PCA plot shows that "Not Endangered" languages tend to cluster separately from the rest, often falling outside the dense central area. In contrast, endangered and extinct languages overlap significantly and cluster near the origin. This suggests that features like speaker count, urbanization, internet use, and country count together do not strongly differentiate between levels of endangerment. Instead, they mainly separate non-endangered languages from the rest.



Appendix E: Confusion Matrix for Ensemble Voting Model with Hyperparameter Tuning

The confusion matrix for the tuned voting ensemble model reveals high performance across the Not Endangered and Extinct categories, with 100% of the test inputs categorized correctly. The model struggled the most with identifying key differences between Endangered, Highly Endangered and At Risk models, exhibiting lower overall precision and recall, likely due to difficulties in identifying distinct boundaries between the categories.


Tuned Ensemble Model Confusion Matrix (Config 3, Target: level_combined)
True Label
Predicted Label
At Risk
Endangered
Extinct
Highly Endangered
Not Endangered
236
9
0
2
0
25
99
0
7
0
0
0
32
2
0
7
15
0
122
0
0
0
0
0
18
200
150
100
50
0