ML Lab Research Proposal — Species Extinction Risk Predictor

Team B: Science/Medical Applications
Date: 6-1-26

1. Problem Statement

Over one million species are currently threatened with extinction, yet most have never been formally assessed for risk. Traditional ecological surveys are too slow and expensive to monitor them all. We are building an AI system that predicts how endangered a species is by analyzing satellite habitat imagery and climate trends — catching species sliding toward extinction before any human fieldwork would.

2. Proposed Approach

We are using a multimodal neural network that combines satellite land-cover images with climate time-series data to predict a species' IUCN extinction risk category. Think of it like giving the AI two different sets of eyes — one watching how forests disappear from space, and another tracking how temperature and rainfall have shifted over 30 years. We chose this over traditional single-source models because habitat loss alone is a weak predictor; it is the interaction between shrinking habitat and worsening climate that determines whether a species survives. This approach lets us screen thousands of unmonitored species at once, something no team of ecologists could do manually.

3. Dataset

We are using three free, publicly available sources: species risk labels from the **IUCN Red List API**, satellite land-cover imagery from **NASA Earthdata**, and 30-year climate records from **WorldClim**. Species location data comes from **GBIF**, which tracks over 2 billion wildlife observations globally. We will start by matching each species' known geographic range to its corresponding satellite

and climate data, then clean and normalize everything so all inputs are on the same scale. The heavy satellite extraction will be handled through Google Earth Engine's free academic access, avoiding the need to store raw imagery locally.

4. Evaluation Metrics

We will know the project is a success if the model achieves a **Macro F1-Score above 0.80** across all five IUCN risk classes — Least Concern, Near Threatened, Vulnerable, Endangered, and Critically Endangered. We will also measure **AUC-ROC above 0.92** to ensure the model correctly ranks species by true risk level. Most importantly, we will check whether the satellite regions the model focuses on actually correspond to known deforestation hotspots, validating that the AI is learning real ecological signals rather than geographic coincidences.

5. Compute Requirements

Training will require a GPU with at least **12GB of VRAM** (such as an RTX 3080 or a free Colab A100 instance) and is expected to take approximately **24–36 hours**. Key libraries include PyTorch for the neural network, Rasterio and GeoPandas for handling geographic data, and the Google Earth Engine API for satellite tile extraction. No exotic hardware is needed beyond a standard deep learning setup.

6. Expected Deliverables

At the end of two weeks we will have a trained model that predicts extinction risk for any species given its location and ecological traits. We will also produce an interactive world map visualizing our predictions, highlighting unassessed species the model flags as high risk — essentially a conservation triage list. Finally, we will deliver a Jupyter notebook documenting our full pipeline, results, and visualizations.

7. Risk & Mitigation

The biggest risk is class imbalance, since most assessed species are Least Concern, which could cause the model to ignore rarer high-risk classes. We will

address this using stratified sampling and a focal loss function that forces the model to pay extra attention to endangered cases. If satellite data extraction runs too slowly, we will pre-cache a fixed set of 10,000 species in Week 1 and scope down from there. Finally, since unassessed species have no ground truth, we will validate strictly on the held-out 20% of IUCN-labeled species and present unassessed predictions only as a demo output, not a formal metric.