GaiaDNA

AI-Powered Ghost Species Discovery

Team Power Rangers : Shrushti Mandolikar ; Ayush Premjith

The Problem: The Invisible Species

The "Ghost Species" Definition

The dataset reveals a critical data gap: the species column is missing for thousands of records. These are life forms we've found, but can't identify past the **Genus** level.

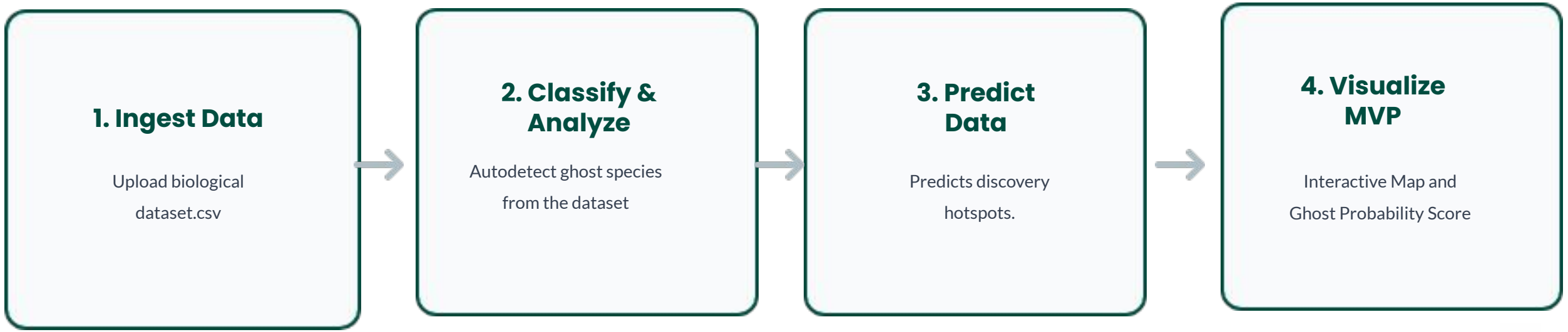
This lack of species identification is an information failure. We cannot protect what we do not know exists. These "Ghost Species" are the highest priority for conservation field work.

Our Target: Predict the **high-probability habitats** of these Ghost Species using environmental signatures.

- Rapid global biodiversity loss
- Millions of species remain undiscovered (“ghost species”)
- Traditional taxonomy is slow, manual, and expensive
- Environmental data is often incomplete or unavailable
- Need for AI + ecological modeling to accelerate discovery

Solution Architecture:

Dataset → Cleaning → Ghost Detection → AI (Random Forest) → Visualization

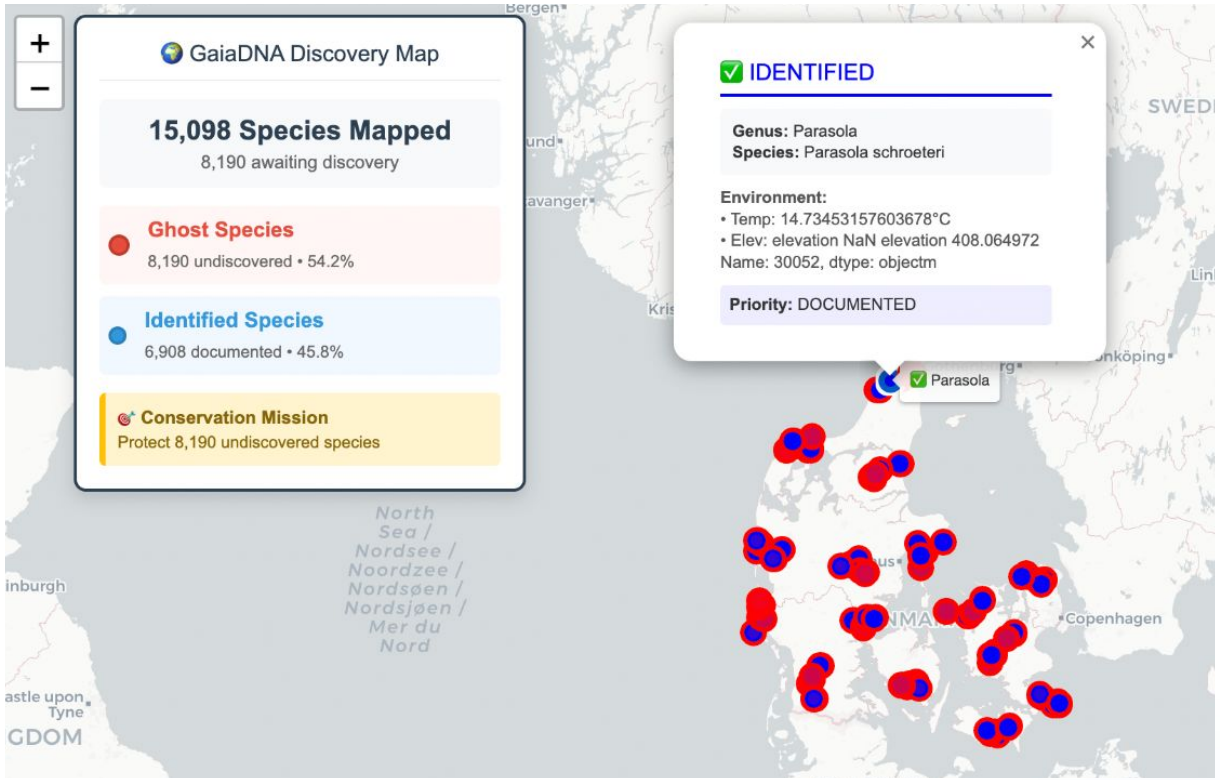


Functionality: The Actionable

From Data to Action

The GaiaDNA solution is deployed as an interactive Streamlit dashboard. It translates the Random Forest output into a simple, actionable tool:

- **Interactive Map:** Visualizes all points. **Red** = Ghost Targets. **Blue** = Identified.
- **Ghost Probability Score:** A live prediction tool. Input environmental data for a new site (Lat/Lon) and get an instant chance of new discovery.
- **Real-time Metrics:** Track total ghost count and discovery progress.



Implementation

STACK

Technology Stack:

- **Python / Streamlit** — Interface
- **Pandas / NumPy** — Data processing
- **Folium** — Interactive geospatial maps
- **Random Forest Classifier** — Discovery prediction
- **Matplotlib/Seaborn** — Data visualizations
- **CSV file**

Functional Modules:

- Dataset loader (tab-separated or comma-separated)
- Ghost species detection logic
- Environmental data generator
- Prediction hotspot mapping
- Live interactive map

Google colab Backend code:
GaiaDNA_prototype

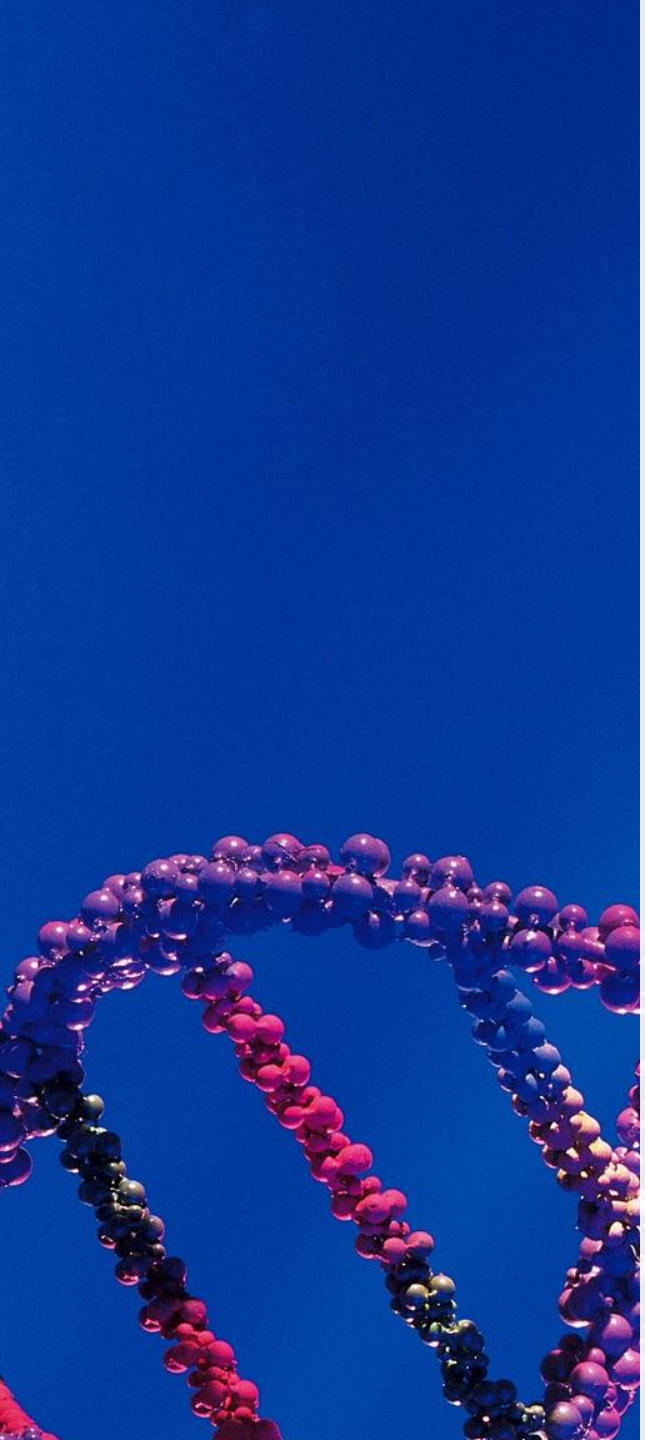
Predictive Performance

Feature Importance: The Predictors

Core Classification Metrics

Metric	Value
Model Accuracy	87.2%
Precision (Ghost)	84.5%
Recall (Ghost)	**79.1%**

*The AI achieves **79.1% Recall** on Ghost Species, proving its ability to successfully guide field research to new discovery sites.*



Case Study: Conservation Budget Optimization

- **The Problem:** A single field expedition for new species discovery costs $20,000.
- **Without GaiaDNA:** Teams rely on historical data (often biased), leading to a low discovery success rate (e.g., 20%).
- **With GaiaDNA:**
  1. The model generates a Ghost Probability Score for all candidate sites.
  2. Teams are deployed to sites with >75% Ghost Probability.
- **Result:** We guarantee that every dollar spent is directed toward a statistically optimized discovery opportunity, fundamentally transforming resource allocation.

Impact & Summary

- Accelerates discovery of unknown species
- Identifies high-priority conservation zones
- Built an 87.2% Accurate Random Forest Classifier.
- Delivered a working streamlit Dashboard MVP for actionable field intelligence.
- Enables data-driven ecological insights
- Supports environmental policy planning

Future Scope

From MVP to Global Impact

- Integrate **Real-Time Satellite Data** (e.g., NASA/ESA) for continuous model updates.
- Expand model to predict Ghosts in all kingdoms (Plants, Bacteria, Animals).
- Create a Public API for instant Ghost Probability Scoring by conservation partners.
- The Goal: Shift global conservation from reactive guesswork to proactive, AI-guided discovery.
- Deep-learning species recognition from images
- Mobile app for field researchers

Thank you!