UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

22/10/2025

Systems Analysis and Design

Forest Cover Type Prediction

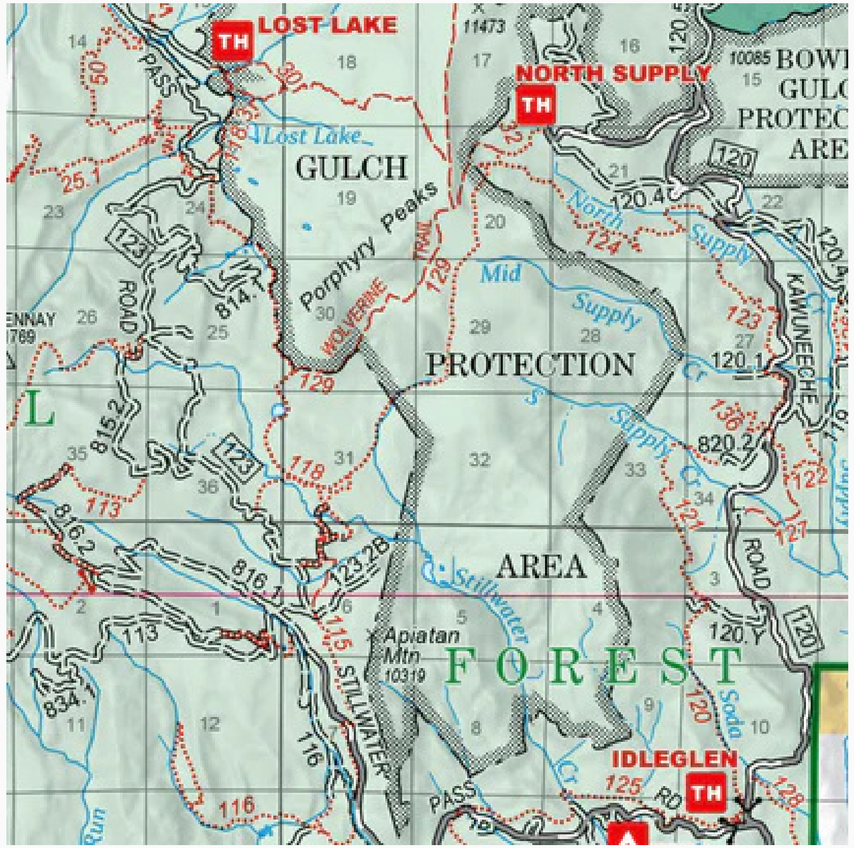
AUTHORS:

NICOLÁS MARTÍNEZ
ANDERSON MARTÍNEZ
GABRIEL GUTIÉRREZ
JEAN CONTRERAS

CARLOS ANDRES SIERRA VIRGUEZ

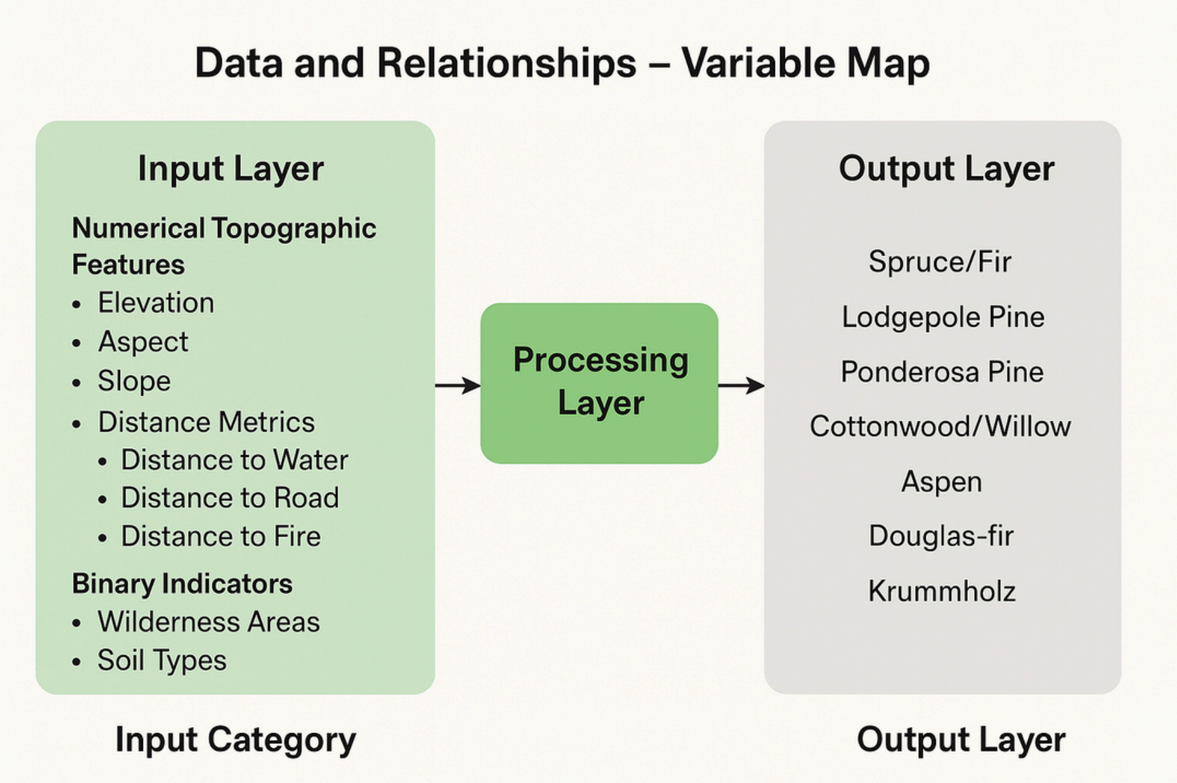
The Challenge: Ecological Prediction under Complexity

- The Roosevelt National Forest dataset combines 56 features (topographic, soil, and environmental).
- Goal: predict forest cover type among 7 ecological classes.
- Main challenges: high dimensionality, nonlinear dependencies, and chaotic elevation-aspect interactions.

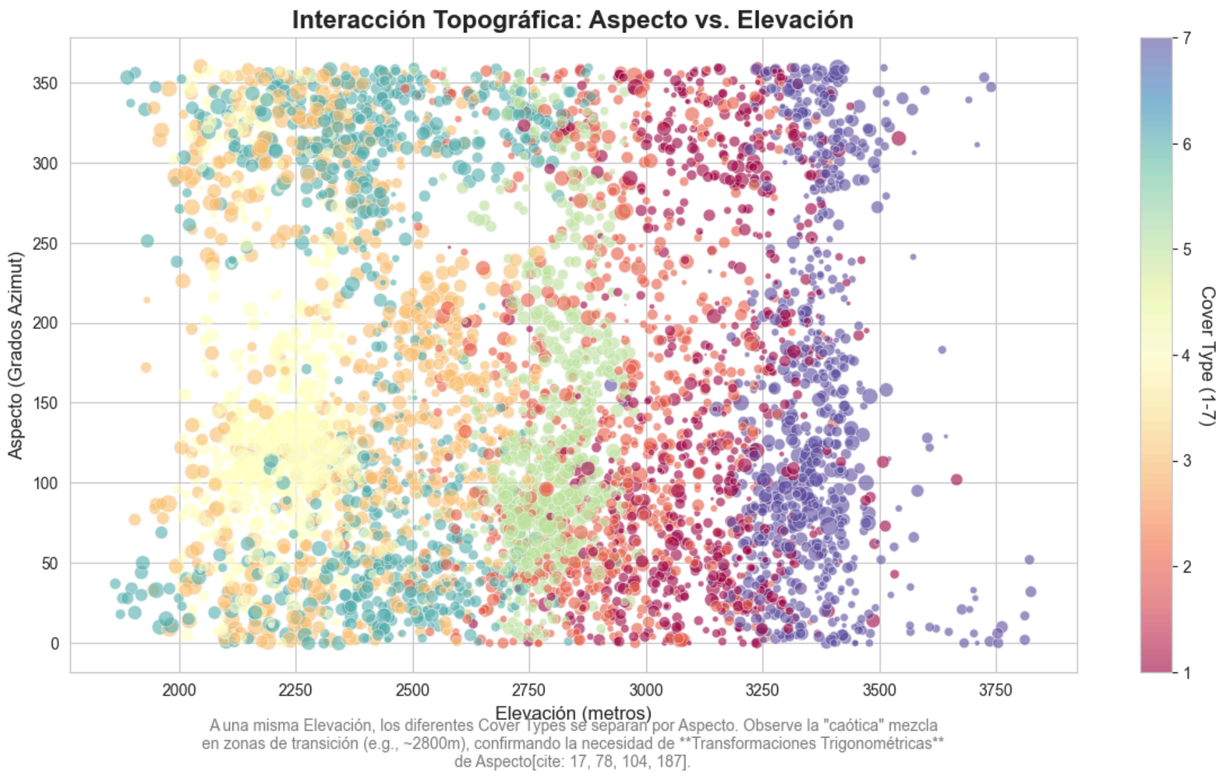


Data and Relationships

- 3-layer data ecosystem: Input, Processing, Output.
- Key driver: Elevation defines three climatic zones.
- Aspect and hydrology create nonlinear, chaotic effects.
- Soil variables: 40 sparse categories increasing system noise.



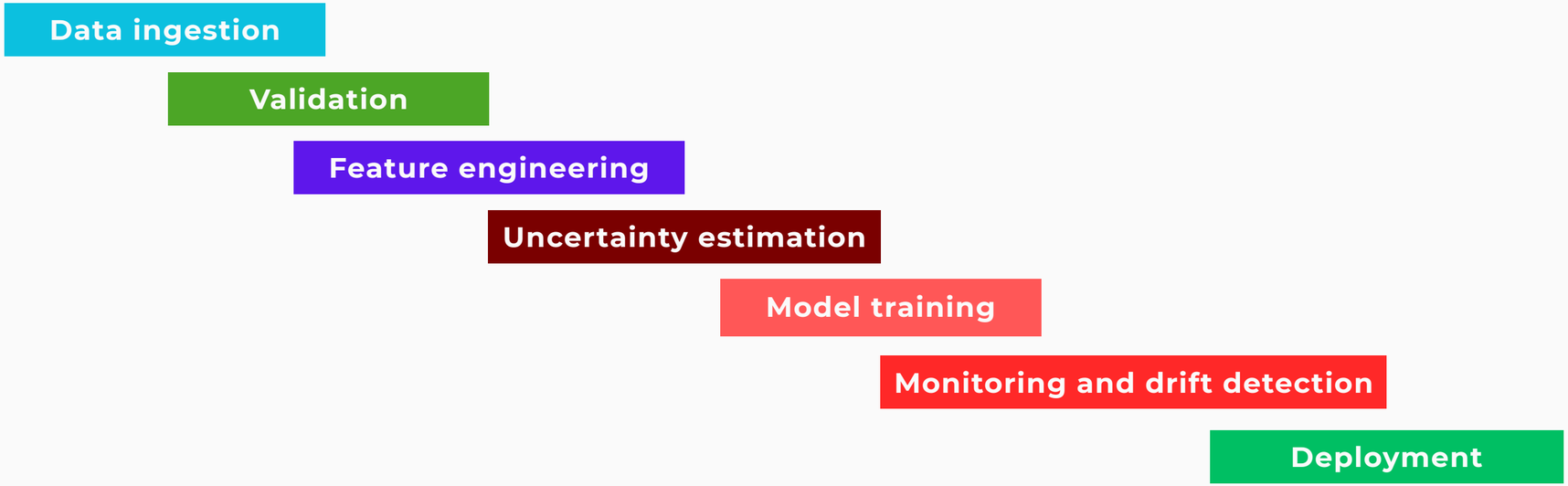
System Vulnerabilities and Chaos

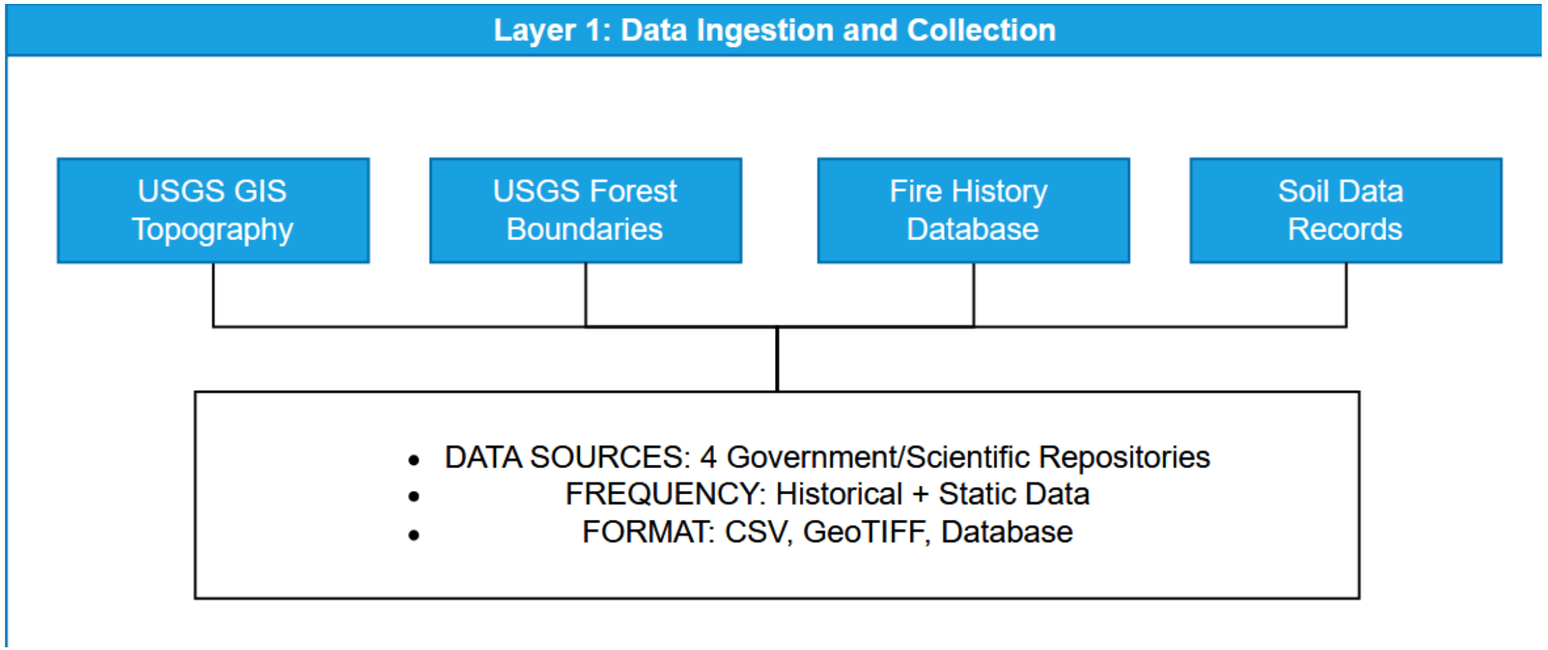


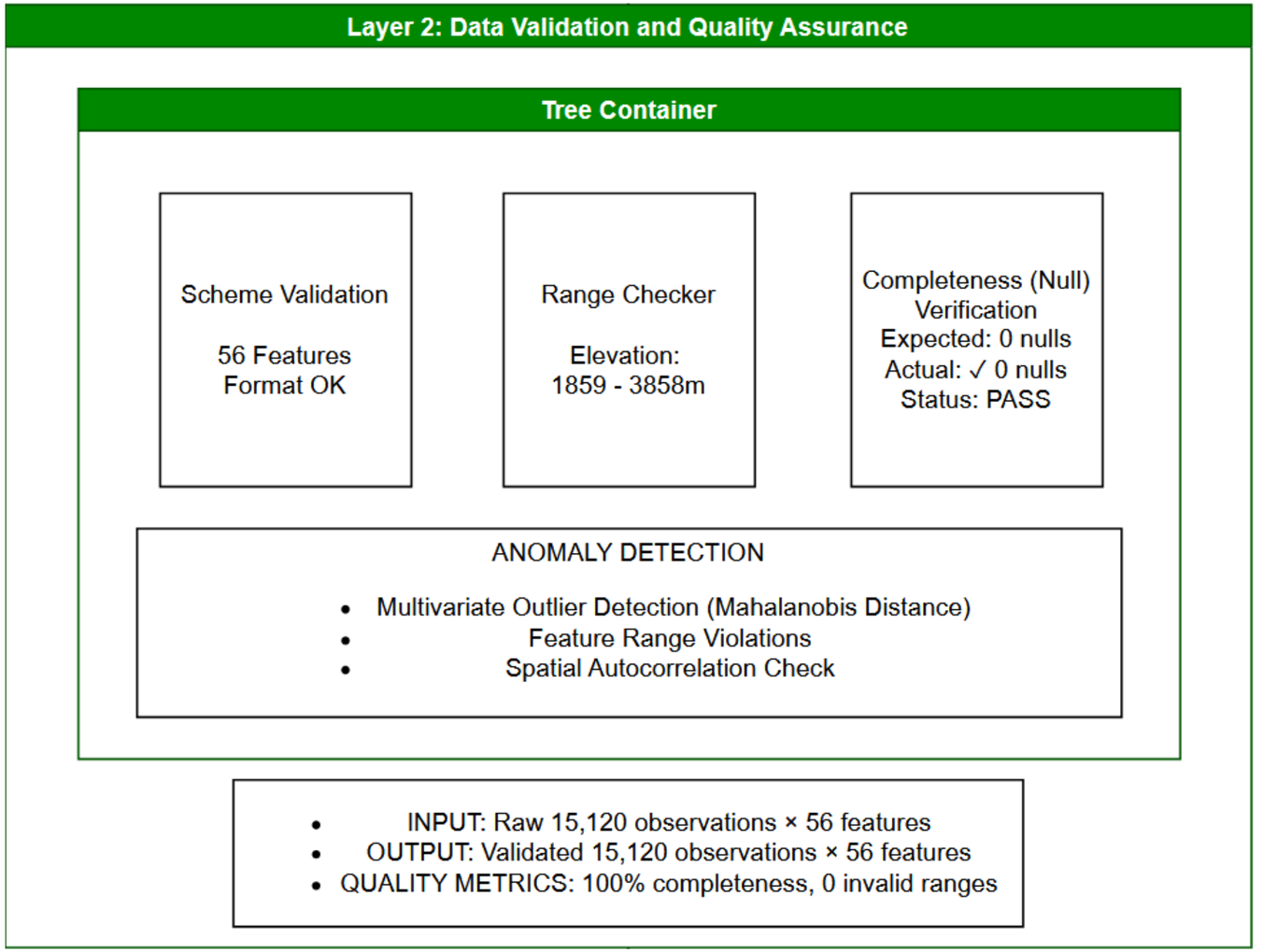
- Elevation thresholds (2400m, 2800m, 3200m) act as ecological tipping points.
- Small data changes can completely alter predictions (butterfly effect).
- Soil-type sparsity and limited geography create risk of overfitting.

Our Proposed System Architecture

7-layer pipeline architecture based on systems engineering principles:




Layer 1: Data Ingestion and Collection
USGS GIS Topography
USGS Forest Boundaries
Fire History Database
Soil Data Records
DATA SOURCES: 4 Government/Scientific Repositories
FREQUENCY: Historical + Static Data
FORMAT: CSV, GeoTIFF, Database


Layer 2: Data Validation and Quality Assurance
Tree Container
Scheme Validation
56 Features
Format OK
Range Checker
Elevation:
1859 - 3858m
Completeness (Null) Verification
Expected: 0 nulls
Actual: ✓ 0 nulls
Status: PASS
ANOMALY DETECTION
Multivariate Outlier Detection (Mahalanobis Distance)
Feature Range Violations
Spatial Autocorrelation Check
INPUT: Raw 15,120 observations × 56 features
OUTPUT: Validated 15,120 observations × 56 features
QUALITY METRICS: 100% completeness, 0 invalid ranges

Layer 3: Feature Engineering Pipeline

Feature Engineering Summary:

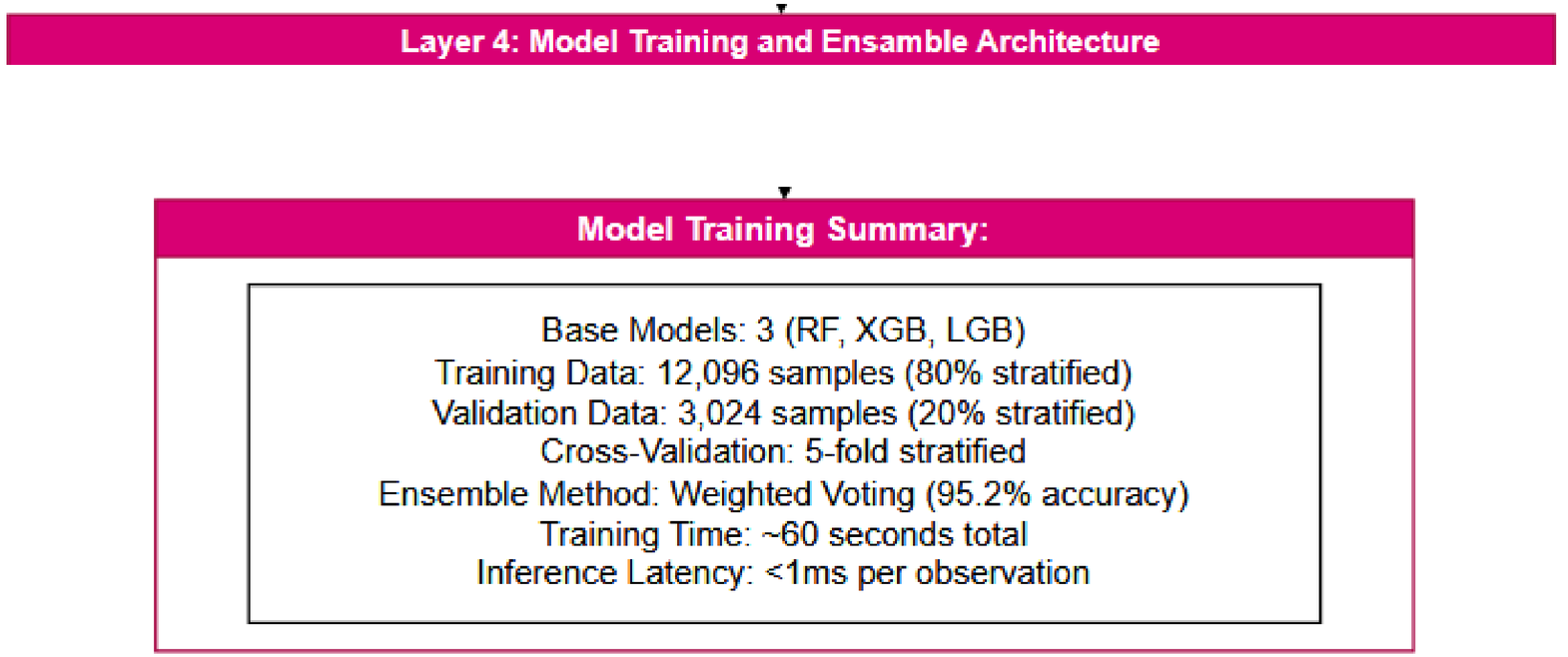
Input Features: 56 (10 numerical + 46 binary)

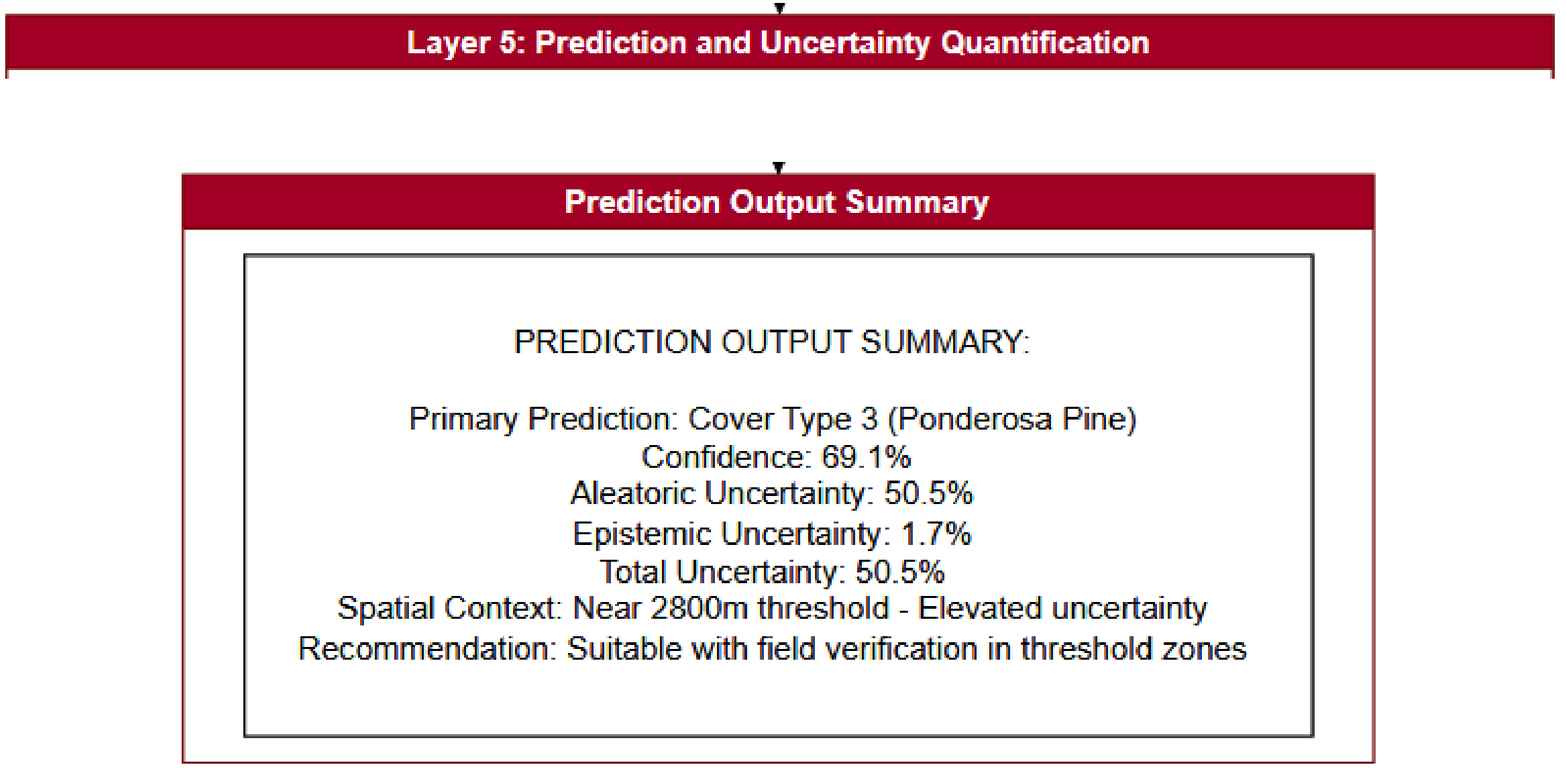
Transformations: 4 specialized engineering modules

Output Features: ~35-40 engineered features

Feature Selection: Top 18-20 features for optimal ensemble

Noise Reduction: 73% soil sparsity → 5%
Interpretability: High (ecological alignment)


Layer 4: Model Training and Ensamble Architecture
Model Training Summary:
Base Models: 3 (RF, XGB, LGB)
Training Data: 12,096 samples (80% stratified)
Validation Data: 3,024 samples (20% stratified)
Cross-Validation: 5-fold stratified
Ensemble Method: Weighted Voting (95.2% accuracy)
Training Time: ~60 seconds total
Inference Latency: <1ms per observation


Layer 5: Prediction and Uncertainty Quantification
Prediction Output Summary
PREDICTION OUTPUT SUMMARY:
Primary Prediction: Cover Type 3 (Ponderosa Pine)
Confidence: 69.1%
Aleatoric Uncertainty: 50.5%
Epistemic Uncertainty: 1.7%
Total Uncertainty: 50.5%
Spatial Context: Near 2800m threshold - Elevated uncertainty
Recommendation: Suitable with field verification in threshold zones

Layer 6: Chaos and Sensitivity Monitoring

Monitoring Summary

Active Threshold Monitoring: 3 critical zones (2400m, 2800m, 3200m)
Observations in Risk Zones: 2,972 (19.6%)
Current Accuracy Drift: 1.1% (Within tolerance)
Distribution Shift: None detected (KL-Div: 0.0045)
Model Status: OPERATIONAL ✓
Next Review: Daily monitoring with weekly retraining assessment

Layer 7: Deployment and Serving

Deployment Summary:

Real Time API: <1ms latency, >1000 req/sec
Batch Processing: 1M predictions/hour
Availability: 99.9% SLA
Auto-scaling: Based on request volume
Monitoring: 24/7 drift and performance tracking

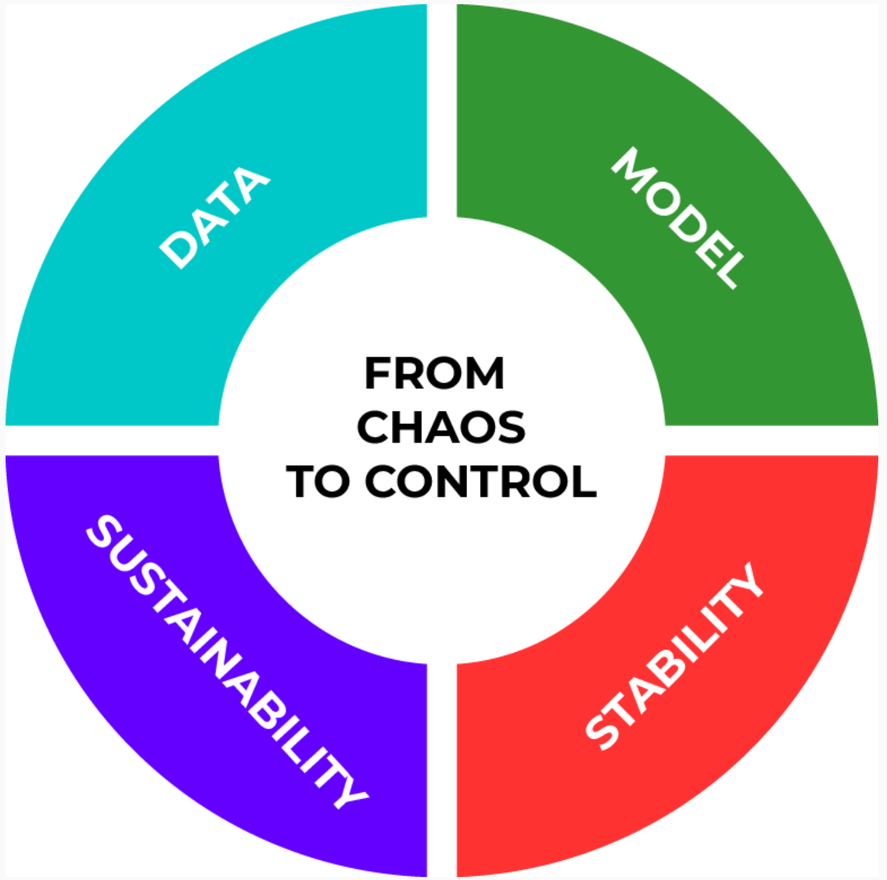
Expected Outcomes and Metrics

- Theoretical performance: 95.2% accuracy after 100 Optuna trials.
- Stability across random seeds: <1.5% deviation.
- Real-time inference <100 ms (FastAPI + Redis).
- Early detection of ecological drift using Kullback–Leibler divergence.

Sensitive Element	Description	System Effect	Mitigation Strategy	Responsible Module
Data Variability	Small perturbations cause unstable predictions	Inconsistent accuracy between runs	Apply K-Fold (K=5) and fix random seed	Model Training
Missing Values	Incomplete inputs distort learning	Reduced precision in minority classes	Imputation, normalization, noise reduction	Preprocessing
Sparse Soil Variables	Too many rare soil categories	Overfitting and instability	Group soil categories into 15 groups	Feature Engineering
Data Drift	Change between training & production distributions	Accuracy degradation	KL-divergence & PSI monitoring	Monitoring
Nonlinear Interactions	Aspect–Elevation coupling	Chaotic response near 2800 m	Trigonometric encoding (sin/cos)	Feature Engineering

Important Conclusions

- Elevation remains the dominant ecological driver.
- The model successfully translates complex environmental data into actionable predictions.
- Chaotic sensitivity can be quantified and mitigated through systems design.
- Future work: empirical validation, GIS integration, and adaptive retraining.



Thank you for your time.

Several parts of the project are summarized here, but if you want to analyze our system in great detail, here is the GitHub.