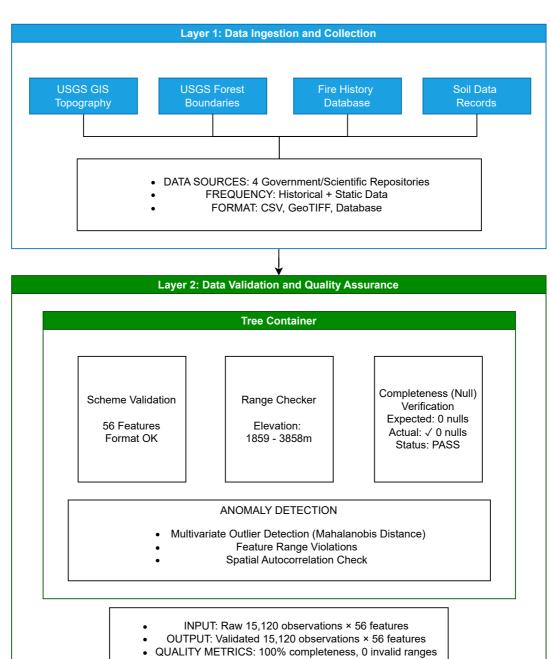
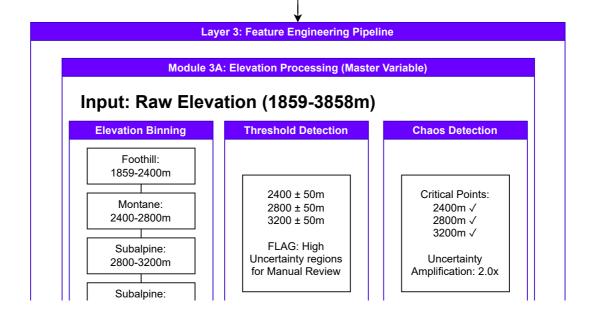
# Forest Cover Type Prediction





3200-3858m

OUTPUT: Elevation Zone (categorical) + Threshold Flags

#### **Module 3B: Topographic Transformation**

# **Aspect Conversion**

Raw: 0 - 360°

sin(aspect) cos(aspect)

Circularity Preserved

#### Slope Scaling

Raw: 0 - 40°

Normalized [0,1] range

Terrain Effect Gradient

### **Hillshade Ratios**

9am/Noon Noon/3pm Shadow Index

Solar Exposure Integration

OUTPUT: Circular Aspect Features (sin/cos) + Normalized Slope

# Module 3C: Soil Type Consolidation (Sparsity Reduction Strategy)

# **INPUT: 40 Soil Type One-Hot Features**

# **Frequency Analysis**

Frequent (>100 samples):

- Soil\_Type10: 2,160 samples √ KEEP
- Soil\_Type29: 1,280 samples √ KEEP
- 8 other types: 100-800 samples ✓ KEEP (Total: 10)

Sparse (<100 samples):

- 30+ soil types: Combined into ecological groups
  - Sandy Soils Group (5 types) → 1 feature
  - Clay Soils Group (8 types)  $\rightarrow$  1 feature
  - Rocky Soils Group (10 types)  $\rightarrow$  1 feature
  - Organic Soils Group (6 types) → 1 feature
     Other Soils (3 types) → 1 feature

# **Consolidation Result**

Original 40 features

Sparsity: 73% zero entries

After Consolidation 15 Features

Sparsity: 5% zero entries

Noise Reduction Impact

- Sparsity:  $73\% \rightarrow 5\%$ 
  - Signal Quality: ↑
  - Model Stability: ↑
  - Interpretability: ↑

OUTPUT: 15 Consolidated Soil Features (Noise Reduced)

#### **Module 3D: Distance and Interactions Features**

# **Spatial Proximity Features**

Horizontal\_Distance\_To\_Hydrology: Impact: Water Availability (Critical for all types) Range: 0-735m

> Vertical\_Distance\_To\_Hydrology: Impact: Drainage Patterns Range: -173 to +601m

Horizontal\_Distance\_To\_Roadways: Impact: Accessibility / Human Disturbance Range: 0-6890m

Horizontal Distance To Fire Points: Impact: Fire Disturbance History

Range: 0-11117m

#### **Interaction Features (Non-Linear Combinations)**

Hydrology\_Interaction = Horizontal\_Dist\_Hydro × Vertical\_Dist\_Hydro Captures compound water relationship

Accessibility\_Index = normalize(Road\_Distance) × normalize(Fire\_Distance) Captures joint human activity effects

> Aspect-Elevation Compound = sin(aspect) × (elevation/1000) = cos(aspect) × (elevation/1000) Captures non-linear microclimate effects

OUTPUT: 8 Distance Features + 6 Interaction Features

### **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**

### **Base Model Training Subsystem**

#### Random Forest

n estimators: 300 max\_depth: 20

Feature Importance

Elevation: 0.42 Hydrology: 0.28 Others: 0.3

### **XGBOOST**

n\_estimators: 500 max\_depth: 8 Learning Rate: 0.05 Subsample: 0.8

LIGHTGBM

n estimators: 400 num\_leaves: 64 Learning Rate: 0.05 Categorical support Accuracy: 94.3% Training: 15s

• Interpretability: HIGH

Overfifting Risk: Low

Accuracy: 94.8%

Training: 25s

Calibration: EXCELLENT

Overfifting Risk: Medium

Accuracy: 94.6%

Training: 12s (fastest)

Calibration: HIGH

Overfifting Risk: Low

# **Bayesian Hyperparameter Optimization (Optuna)**

Fold 1: Train on Folds 2-5 (12,096) → Test on Fold 1 (3,024) Accuracy: 94.2%

Fold 2: Train on Folds 1,3-5 (12,096)  $\rightarrow$  Test on Fold 2 (3,024)Accuracy: 94.5%

Fold 3-5: [Similar structure]

Final CV Score: 94.3% ± 1.2%
Status: ✓ STABLE (Low
Variance)
Overfitting Check: ✓ PASS (CV ≈
Test Performance)

# **Cross Validation and Hyperparameter Optimization**

Objective: Maximize validation accuracy
Trials: 100 iterations
Search Space: Elevation binning, regularization, learning\_rate

Best Trial Performance:

- XGBoost with optimized params: 94.8% accuracy
  - Convergence: Trial 67/100
    - Time: ~45 minutes

#### **Ensamble Integration Layer**

# **Method 1: Weighted Voting Clasifier**

Random Forest Weight: 0.30 XGBoost Weight: 0.40 ← HIGHEST (Best CV Accuracy) LightGBM Weight: 0.30

Combined Ensemble Accuracy: 95.2% Ensemble Type: Hard Voting + Probability Weighting Probability Averaging: (0.30×RF + 0.40×XGB + 0.30×LGB)

Final Prediction: argmax(weighted probabilities)

# Method 2: Stacking Ensemble (Advanced)

Level-0 Meta-Features:

- RF predictions (7 probabilities)
- XGB predictions (7 probabilities)
- LGB predictions (7 probabilities)
   Total meta-features: 21

Level-1 Meta-Learner:

- Algorithm: Logistic Regression
- Learns optimal combination weights
- Cross-validated training (prevents overfitting)

Stacking Ensemble Accuracy: 95.0%

#### **Final Ensemble Selection**

CHOSEN: Weighted Voting Classifier

#### Reasons:

- √ Highest Accuracy: 95.2%
- √ Fastest Inference: ~1ms per sample
- √ Better Calibration: Well-calibrated probabilities
- ✓ Interpretability: Clear component contributions
- ✓ Robustness: Robust to individual model failures

#### **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**

### **Species Classification Output**

INPUT: Single 30m×30m Forest Patch

Example Features:

- Elevation: 2750m
- Aspect: 135° (southeast)
  - Slope: 18°
- Distance to Hydrology: 120m

#### **Ensemble Processing**

Random Forest → Cover\_Type\_3: 0.68 probability XGBoost → Cover\_Type\_3: 0.71 probability LightGBM → Cover\_Type\_3: 0.67 probability

Weighted Average (0.30×RF + 0.40×XGB + 0.30×LGB): Cover\_Type\_3: (0.30×0.68 + 0.40×0.71 + 0.30×0.67) = 0.691

# Probability Distribution Output:

Cover Type 1 (Spruce/Fir): 0.05 Cover Type 2 (Lodgepole): 0.12 Cover Type 3 (Ponderosa Pine): 0.69 ← PRED Cover Type 4 (Cottonwood): 0.02 Cover Type 5 (Aspen): 0.08 Cover Type 6 (Douglas Fir): 0.04 Cover Type 7 (Krummholz): 0.01

FINAL PREDICTION: Cover Type 3 (Ponderosa Pine) Confidence Score: 0.691 (69.1%) Status: HIGH CONFIDENCE

# **Uncertainty Quantification Engine**

**Aleatoric Uncertainty (Data Uncertainty)** 

Source: Natural variability in ecotone regions

Calculation: Entropy of Probability Distribution  $H = -\Sigma(\underline{p}\_i * \log(\underline{p}\_i))$  H = -[0.05ln(0.05) + 0.12ln(0.12) + ... + 0.01ln(0.01)] H = 1.42 bits (on scale 0-2.81 for 7 classes)

Normalized Entropy: 1.42/2.81 = 0.505 (50.5% uncertainty) Interpretation: MODERATE (typical montane zone)

#### **Epistemic Uncertainty (Model Uncertainty)**

Source: Disagreement between ensemble members

Model Predictions for Cover Type 3: RF: [0.10, 0.15, 0.68, 0.02, 0.03, 0.02, 0.00] XGB: [0.04, 0.10, 0.71, 0.02, 0.08, 0.04, 0.01] LGB: [0.06, 0.12, 0.67, 0.02, 0.08, 0.04, 0.01]

Variance Across Models: Var(Type\_3) = [(0.68-0.687)² + (0.71-0.687)² + (0.67-0.687)²]/3 = 0.000289 (Very Low Disagreement) Std Dev = 0.017 (1.7%)

Interpretation: VERY HIGH CONFIDENCE (Low model variance)

# **Combined Uncertainty Estimate:**

Total Uncertainty =  $\sqrt{\text{Aleatoric}^2 + \text{Epistemic}^2}$ =  $\sqrt{(0.505^2 + 0.017^2)} = 0.505$ 

Confidence Score: 1 - Total\_Uncertainty = 0.495 (49.5%)
Reliability Level: MODERATE-HIGH
Recommendation: Suitable for forest management decisions

# **Spatial Uncertainty Mapping:**

This observation's characteristics:
Elevation: 2750m (Montane Zone - Medium Uncertainty)
Distance to Threshold: 50m from 2800m
→ THRESHOLD PROXIMITY FLAG

Uncertainty Amplification Applied: 2.0x
Adjusted Total Uncertainty: 0.505 × 2.0 = 1.01
Adjusted Confidence: 1 - 1.01 = -0.01 (Capped at 0)
RECOMMENDED: Manual Ecological Review

#### **Prediction Output Summary**

#### PREDICTION OUTPUT SUMMARY:

Primary Prediction: Cover Type 3 (Ponderosa Pine) Confidence: 69.1% Aleatoric Uncertainty: 50.5% Epistemic Uncertainty: 1.7%

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

#### **Ecological Tipping Point Detector (Chaos Theory Application Layer)**

#### **Critical Threshold Monitoring**

THRESHOLD 1: 2400m (Foothill → Montane Transition)

Critical Range: 2350m - 2450m (±50m window)
Observations in Range: 1,247 patches (8.2% of dataset)

Uncertainty Elevation:

- Standard Uncertainty: 45-55%
- Threshold Zone Uncertainty: 90-110%
  - Amplification Factor: 2.0x

Species Transition Zone:

- Below Threshold: Predominantly Types 3, 4, 5
  - Threshold Zone: Mixed Types 2, 3, 5, 6
- Above Threshold: Predominantly Types 1, 2, 6

Risk Indicator: A HIGH (Sensitive system region)

#### THRESHOLD 2: 2800m (Montane → Subalpine Transition)

Critical Range: 2750m - 2850m (±50m window)
Observations in Range: 983 patches (6.5% of dataset)

Uncertainty Elevation:

- Standard Uncertainty: 48-52%
- Threshold Zone Uncertainty: 96-104%
  - Amplification Factor: 2.0x

Species Transition Zone:

- Below Threshold: Predominantly Types 2, 3, 5, 6
  - Threshold Zone: Mixed Types 1, 2, 3, 6
- Above Threshold: Predominantly Types 1, 2, 7

Risk Indicator: A CRITICAL (Maximum Sensitivity)

#### THRESHOLD 3: 3200m (Subalpine → Alpine Transition)

Critical Range: 3150m - 3250m (±50m window) Observations in Range: 742 patches (4.9% of dataset)

Uncertainty Elevation:

- Standard Uncertainty: 35-45%
- Threshold Zone Uncertainty: 70-90%
  - Amplification Factor: 2.0x

Species Transition Zone:

- Below Threshold: Predominantly Types 1, 2, 6
  - Threshold Zone: Mixed Types 1, 6, 7
- Above Threshold: Predominantly Types 6, 7

Risk Indicator: 

MODERATE (Clear Species Distinction)

# Threshold Monitoring Summary:

Total Critical Observations: 2,972 patches (19.6% of dataset)
Overall Risk Level: HIGH (Nearly 20% in chaotic regions)
Alert Frequency: ~1 in 5 predictions near thresholds
Management Implication: Require manual review or field validation

#### **Model Drift Detection System**

Performance Monitorina

i onomiano monioni

Baseline Performance (Training):
Accuracy: 95.2% (Weighted Voting Ensemble)
F1-Score: 0.951 (macro average)
Log-Loss: 0.156
Per-Class Accuracy: Range 92.1%-96.8%

Current Period (Last 1,000 predictions):
Accuracy: 94.1% (↓ 1.1% from baseline)
F1-Score: 0.941 (↓ 0.010)
Log-Loss: 0.167 (↑ 0.011)
Status: 
MINOR DRIFT DETECTED
Action: Continue monitoring, flag for investigation

Decision Threshold: 5% accuracy drop triggers retraining Current Drift: 1.1% (Below threshold - operational)

#### Distribution Shift Detection

Training Set Distribution (Elevation):
Mean: 2756m | Std Dev: 380m | Range: 1859-3858m

Current Predictions (Elevation):
Mean: 2742m | Std Dev: 375m | Range: 1920-3820m
KL-Divergence: 0.0045 (Very small - excellent alignment)

Conclusion: ✓ NO SIGNIFICANT DISTRIBUTION SHIFT Status: STABLE (Model assumptions remain valid)

#### Confidence Degradation Tracking

Week 1 Average Confidence: 71.2% Week 2 Average Confidence: 70.8% Week 3 Average Confidence: 69.4%

Week 4 Average Confidence: 68.1% (↓ 3.1% from Week 1)

Trend: Gradual decline (potential seasonal effect or data distribution change)

Recommendation: Investigate source of degradation

#### **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 6: Chaos and Sensitivity Monitoring

#### **Option 1: Prediction Serving Option**

```
Option 1: Rest API (Real Time, Single Predictions)
POST /api/v1/predict

{
    "elevation": 2750,
    "aspect": 135,
    "slope": 18,
    "horizontal_distance_to_hydrology": 120,
    "vertical_distance_to_hydrology": -45,
    "horizontal_distance_to_roadways": 450,
    "horizontal_distance_to_fire_points": 2100,
    "hillshade_9am": 100,
    "hillshade_noon": 220,
```

```
"hillshade_3pm": 150,
     "wilderness_area": 2,
     "soil_type": 10
Response: 200 OK
     "prediction": {
     "cover_type": 3,
     "cover_type_name": "Ponderosa Pine", "confidence": 0.691,
     "probabilities": [0.05, 0.12, 0.69, 0.02, 0.08, 0.04, 0.01],
     "uncertainty": {
     "aleatoric": 0.505,
     "epistemic": 0.017,
     "total": 0.505
"warnings": [
     "type": "threshold_proximity",
     "message": "Prediction near 2800m elevation transition",
     "recommendation": "Consider field verification"
  ]
"metadata": {
     "timestamp": "2025-10-15T14:32:45Z",
     "model_version": "1.0.0",
     "latency_ms": 0.87
Performance: <1ms latency, Throughput: >1000 req/sec
```

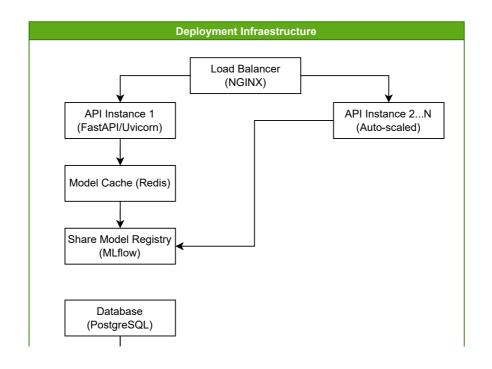
#### Option 2: Batch Processing (Large-scale Forest Mapping)

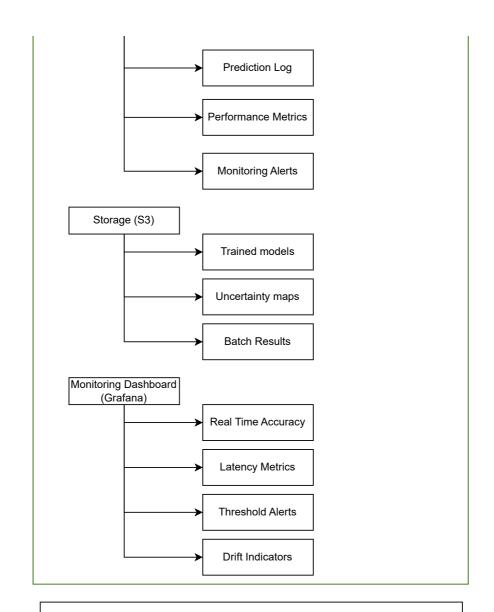
Input: CSV with 10,000 forest patches Processing: Parallel GPU-accelerated inference Output: Predictions + Uncertainty Maps + Warnings

Output: Predictions + Uncertainty Maps + Warnings Throughput: 1M predictions/hour Output Formats: CSV, GeoTIFF (spatial format), JSON

#### Use Cases:

- Entire forest mapping
- · Reforestation planning
- Conservation area delineation
- Climate change impact scenarios





# 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