Remote Sensing for Forest Recovery: Final Report

Benjamin Frizzell Zanan Pech Mavis Wong Hui Tang
Piotr Tompalski Alexi Rodríguez-Arelis

2025-06-16

Table of contents

---

1 Executive Summary

Problem. Field surveys at Year-7 are costly and delayed; need forecasts for 80% survival threshold to guide timely interventions.

The **Canada Forest Service** requires an early-warning system to flag high-risk sites for proactive intervention.

Approach.
- Merged multispectral rasters (2010–2023) with planting and Year-7 survival survey tables.
- Two modeling phases:
1. **Static classifiers** (Logistic Regression, Random Forest, XGBoost) on aggregated features.
2. **Temporal RNNs** (GRU, LSTM) on annual spectral sequences.
- Employed grouped 5-fold cross-validation by site ID; optimized for F1 score.

Key Findings.
- **Random Forest** (80% threshold) achieved the highest F1 = 0.521.
- **Gradient Boosting** F1 = 0.411; **Logistic Regression** F1 = 0.515.
- **GRU** sequence model (site + spectral features) reached F1 = 0.440 and Recall = 0.569, outperforming LSTM (F1 = 0.368).

Deliverables.
- Reproducible Makefile targets (`make data`, `make train_models`, `make train_rnn`, `make evaluate`).
- Versioned model artifacts in `models/80` and evaluation plots in `results/80`.
- Quick-start Jupyter notebook for batch scoring.

Recommendations.
- **Advisory use only:** Validate flagged sites with targeted surveys before field action.
- **Data enhancement:** Expand ground surveys in under-sampled ecozones.
- **Future exploration:** Prototype a CNN–LSTM hybrid to jointly learn spatial and temporal patterns.
- **Deployment:** Develop a Streamlit dashboard for non-technical planners.

2 Introduction

Early canopy survival surveys (Year 7) are critical for assessing reforestation success, yet they are time-consuming and expensive. The **Canada Forest Service** seeks a data-driven early-warning system to identify sites at high risk of low survival before costly field surveys occur.

2.1 Data Overview

- **Multispectral Imagery (2010–2023):** Annual raster stacks of 10 spectral indices per site, capturing vegetation health over time.

- **Planting Records:** Approximately 11,000 geolocated planting sites with attributes such as planting density, species type, and planting year.

- **Survival Surveys:** Year-7 canopy cover measurements, binarized as survival (80%) or failure (<80%).

2.2 Refined Objectives

1. **Static Classification:** Train and evaluate Logistic Regression, Random Forest, and XGBoost models on aggregated spectral and site features to predict seven-year survival.

2. **Temporal Modeling:** Develop sequence models (GRU and LSTM) that leverage annual spectral time series for improved early forecasting.

3. **Performance Comparison:** Compare static and sequence pipelines using grouped 5-fold cross-validation by site ID, optimizing for F1 score to balance precision and recall.

4. **Reproducibility:** Package the entire workflow into a Makefile and Jupyter notebooks, enabling non-technical stakeholders to run data preparation, model training, and evaluation with minimal setup.

3 Data Science Methods

3.1 Phase 1: Static Models

- **Preprocessing:** Drop IDs/dates, scale spectral indices & density, one-hot encode species type.
- **Models & Hyperparameters:**
  - Logistic Regression: C (regularization strength), penalty=l2
  - Random Forest: n_estimators, max_depth
  - XGBoost: learning_rate (eta), max_depth, reg_alpha, reg_lambda

- **Tuning & Validation:** Randomized search (50–300 iterations) with grouped 5-fold CV (by site ID), optimizing F1 score.

3.2 Phase 2: Temporal Models

- **Sequence Preparation:**
  - Aggregate per-site yearly spectral indices into variable-length sequences.
  - Engineer time features: log-transformed Δt, seasonality via sine/cosine of Day-of-Year.
- **Models & Hyperparameters:**
  - GRU: hidden_size=32, num_layers=1, dropout=0.2, lr=1e-3, optimizer=Adam
  - LSTM: same architecture and training settings.
- **Training & Evaluation:**
  - Trained for 25 epochs with early stopping (patience=5).
  - Batch size=64, padded sequences, tracked sequence lengths.

3.3 Metrics & Ethics

- **Primary Metrics:** Precision, Recall, F1 (primary).
- **Cross-Validation:** GroupKFold to prevent spatial leakage.
- **Stakeholder Impact:** Emphasize recall to minimize false negatives (missed high-risk sites).

4 Data Product & Results

Deliverables
- **CLI Pipeline:** `make data`, `make train_models`, `make train_rnn`, `make evaluate`
- **Model Artifacts:** Pickled models under `models/<threshold>/`
- **Evaluation Plots:** ROC / PR curves in `results/<threshold>/`
- **Quick-start Notebook:** `notebooks/data_product_quickstart.ipynb`

4.1 Phase 1: Static Models Performance (80 % Threshold)

Model	Precision	Recall	F1
Gradient Boosting	**0.485**	0.403	0.411
Random Forest	0.427	**0.668**	**0.521**
Logistic Regression	0.435	0.630	0.515

4.2 Phase 2: Sequence Models Performance (80 % Threshold)

Model (Features)	Precision	Recall	F1
LSTM (Site + Sat)	0.367	0.369	0.368
GRU (Site + Sat)	**0.359**	**0.569**	**0.440**

The GRU with site + satellite features is the best performing RNN configuration.

> **Interpretation:**
> * Classical static models plateau around F1 0.52, with Random Forest achieving the highest F1.
> * GRU significantly improves Recall on high-risk sites, while LSTM offers modest gains.
> * Temporal models increase false positives, suggesting a CNN–LSTM hybrid could better capture spatial–temporal patterns.

5 Conclusions & Recommendations

Problem Recap.
Year-7 survival field surveys by the Canada Forest Service are time-consuming and expensive, delaying actionable insights on reforestation success.

Methodological Summary.
- **Phase 1:** Static classifiers (Logistic Regression, Random Forest, XGBoost) on aggregated spectral and site features.
- **Phase 2:** Temporal RNNs (GRU, LSTM) on annual spectral time series, with and without site features.

Key Outcomes.
- At the **80 % survival threshold**, the best static model (Random Forest) achieved **F1 = 0.521** and **Recall = 0.668**, outperforming Gradient Boosting (F1 = 0.411) and Logistic Regression (F1 = 0.515).
- The **GRU** sequence model (site + spectral features) delivered **F1 = 0.440** and improved recall from 0.403 → 0.569 (**+16 pp**) relative to XGBoost, demonstrating the value of temporal information.
- **LSTM** yielded F1 = 0.368 and was less stable, confirming GRU as the more efficient RNN choice for this dataset.

Limitations.
- **Class Imbalance & Thresholding:** The high-survival class dominates; even at an 80 % cut-off we observe skewed precision/recall trade-offs.
- **Data Gaps:** Irregular satellite acquisition dates and missing rasters introduce noise.
- **Model Complexity vs. Interpretability:** RNNs capture temporal patterns but are less transparent than tree-based methods.

Recommendations.
1. **Advisory Monitoring:** Use model outputs as early-warning flags; validate with targeted ground surveys before intervention.
2. **Data Enrichment:** Increase sampling in under-represented ecozones and improve temporal coverage of satellite inputs.
3. **Model Enhancements:** Prototype a CNN–LSTM hybrid to jointly learn spatial context and temporal dynamics.
4. **Operationalization:** Develop a lightweight dashboard (e.g., Streamlit) for CFS planners to review flagged high-risk sites and track model confidence.

Tip

Rendering Instructions

```
cd reports/"final report"
conda activate mds-afforest-dev
quarto render report.qmd
open report.pdf
```