Remote Sensing for Forest Recovery: Proposal Report

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

2025-05-09

Table of contents

1 Abstract

Monitoring afforestation progress across hundreds of remote and ecologically diverse sites in Canada poses significant challenge, particularly due to the weak spectral signals from newly planted trees with sparse canopies in early growth stages. This project seeks to address two key research questions: (1) Can satellite-derived vegetation indices and site-level data be used to accurately predict tree survival over time in large-scale afforestation programs? and (2) What modeling approaches are most effective for this task? Using data from Forest Canada, we train and evaluate a suite of machine learning models—including logistic regression, random forests, gradient boosting, and, if time permits, deep learning architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models. Addressing these questions is critical for monitoring afforestation efforts, which are esssential for climate mitigation, biodiversity conservation, and ecosystem restoration.

2 Introduction

Monitoring the success of large-scale afforestation initiatives is a critical yet complex challenge especially for early growth stages. Young trees often produce weak spectral signals due to sparse canopies, making them difficult to detect using traditional remote sensing methods (University of British Columbia Master of Data Science Program 2025). As part of Canada's 2 Billion Trees program—which aims to plant two billion trees across Canadian provinces by 2031—Natural Resources Canada must track survival rates across hundreds of ecologically diverse and often remote planting sites (Natural Resources Canada 2021).

This problem is crucial because the ecological and climate benefits of afforestation—such as carbon reduction, and biodiversity restoration can only be realized if newly planted trees survive and grow.

In this study, we aim to investigate two main research questions:

- Can satellite-derived vegetation indices and site-level data be used to accurately predict tree survival over time in large-scale afforestation programs?
- Which modelling approach is most effective, and how long after planting is needed before accurate survival predictions can be made?

To address these qestions, this study leverages satellite-derived vegetation indices and site-level data to train machine learning models. By evaluating multiple modelling approaches—including logistic regression, random forests, and deep learning architectures—we aim to determine which techniques provide the most accurate predictions of tree survival rate to support the mission of supporting sustainable forest management and addressing climate change.

3 Data Description

The dataset used in this study combines field-measured survival rates for more than 2,500 afforested sites collected by Forest Canada (University of British Columbia Master of Data Science Program 2025) with remote sensing data obtained from the Harmonized Landsat Sentinel-2 (HLS) project (USGS 2024).

The combination of time-series satellite data and field-measured survival rate allows us to investigate how spectral signals change with survival rates and support the model development for survival prediction based on satellite data.

3.1 Features Description

In the data, each site is divided into one or more pixels. Each row in the dataset represents a pixel-level satellite observation at a given time, with its corresponding site-level (Table 1) and pixel-level (Table 2) features.

Table 1: Summary of site-level features. The site-level features provide spatial, temporal and ecological information associated with each afforested site, including the site ID, area, previous land use, afforestation information, species type, and our target: field-measured survival rates from Year 1 to Year 7.

Category	Column Name	Description
Identifier	`ID`	Site ID
Spatial	`Area_ha`	Area of the Site (hectares)
	`prevUse`	Previous Land Use of the Site
Temporal	`PlantDt`	Planting Date
	`Season`	Planting Year
	`AsssD_1` to `AssD_7`	Date of Field Survival Assessment (Years 1 to 7)
Ecological	`SpcsCmp`	Species Composition of Site
	`Type`	Species Type (Conifer, Deciduous, Mixed)
	`Planted`	Number of Trees Planted (Initial Field Record)
	`NmbrPlO`	Number of Trees Originally Planted
	`NmbrPlR`	Number of Trees Replanted
	`NmbrPlT`	Total Number of Trees Planted (`NmbrPlO` + `NmbrPlR`)
Target	`SrvvR_1` to `SrvvR_7`	Field Measured Survival Rate (Years 1 to 7)

Table 2: Summary of pixel-level features. The pixel-level features include the pixel ID, the capture date of the satellite data, and our primary predictor: the spectral indices.

Category	Column Name	Description
Identifier	`PixelID`	Pixel ID
Temporal	`ImgDate`	Image Date of the Remote Sensing Data
	`Year`	Image Year of the Remote Sensing Data
	`DOY`	Image Day of Year of the Remote Sensing Data
Spectral Indicies	`NDVI`, `SAVI`, `MSAVI`, `EVI`, `EVI2`, `NDWI`, `NBR`, `TCB`, `TCG`, `TCW`	See Table 3 for details.

3.2 Spectral Indicies

In this study, our primary predictor is the spectral indices. Table 3 describes the spectral indices used in this study.

Table 3: Description and evaluation of spectral indices available in the dataset.(Zeng et al. 2022; Baig et al. 2014; Landsat Missions; Mondal 2011; n.d.)

Type	Index	Description	Evaluation
Vegetatio Index	Normalized Difference Vegetation Index (NDVI)	Measures vegetation greenness and health by comparing near-infrared (NIR) and red reflectance.	Sensitive to vegetation changes, but tends to saturate under dense vegetation.
	Soil-Adjusted Vegetation Index (SAVI)	Adjusted NDVI that reduces background soil influence and corrects for soil brightness.	Less likely to saturate under dense vegetation, suitable for areas with sparse vegetation.
	Modified Soil-Adjusted Vegetation Index (MSAVI)	Improved SAVI that minimises soil background influence.	Compared to SAVI, more sensitive to vegetation changes in areas with sparse vegetation. Suitable for monitoring plant greenness and health for young trees or sparsely vegetated regions.

Type	Index	Description	Evaluation
	Enhanced Vegetation Index (EVI)	Measures vegetation greenness using blue, red, and NIR bands to correct for atmospheric and canopy background influences.	Suitable for areas with dense vegetation.
	Two-band Enhanced Vegetation Index (EVI2)	Similar to EVI, but only uses red and NIR bands.	Suitable when the blue band is not available.
	Tasseled Cap Greenness (TCG)	Measures vegetation greenness using a tasselled cap transformation of spectral bands.	Less sensitive to vegetation changes compared to NDVI
Water Index	Normalized Difference Water Index (NDWI)	Measures moisture content by comparing NIR and shortwave infrared (SWIR) reflectance.	Mostly used for identifying water bodies, but can be useful in monitoring vegetation growth as the water content in trees increases with time. Sensitive to changes in vegetation water content, as well as soil and atmospheric background noises.
	Tasseled Cap Wetness (TCW)	Measures soil and vegetation moisture using a tasselled cap transformation of spectral bands.	Compared to NDWI, TCB is less sensitive to atmospheric noises, but it is also less sensitive to changes in vegetation content.
Fire Index	Normalized Burn Ratio (NBR)	Identify burned areas and measure burn severity using NIR and SWIR bands.	NBR has been used to monitor vegetation recovery rate after fire, which parallels afforestation survival monitoring. Although it is not designed for monitoring vegetation health, it may still be useful for prediction.
Surface Reflectance	Tasseled Cap Brightness (TCB)	Measures soil brightness using a tasselled cap transformation of spectral bands.	Not designed for monitoring vegetation health, useful for detecting soil exposure and changes in vegetation cover.

3.3 Exploratory Insights

To better understand the data, we conducted exploratory data analysis (EDA) and identified 4 key data quality and structural issues that should be addressed before model training. First, there are out-of-range values in the survival rate and spectral indices columns, where the spectral indices should range between -1 to 1 and survival rates should not exceed 100%. We will remove these outliers during data cleaning.

Secondly, we noticed severe class imbalance in `SpcsCmp`. There are over 300 categories in `SpcsCmp`, making it impractical to use directly in modelling. Likewise, >80% of values in `prevUse` correspond to the class `AG` (agriculture land). We will not be using these columns during modelling to avoid overfitting.

Thirdly, there is strong seasonality in spectral indices (Figure 1), where the signal peaks during summer, and drops during winter. This will be something we need to address during model development.

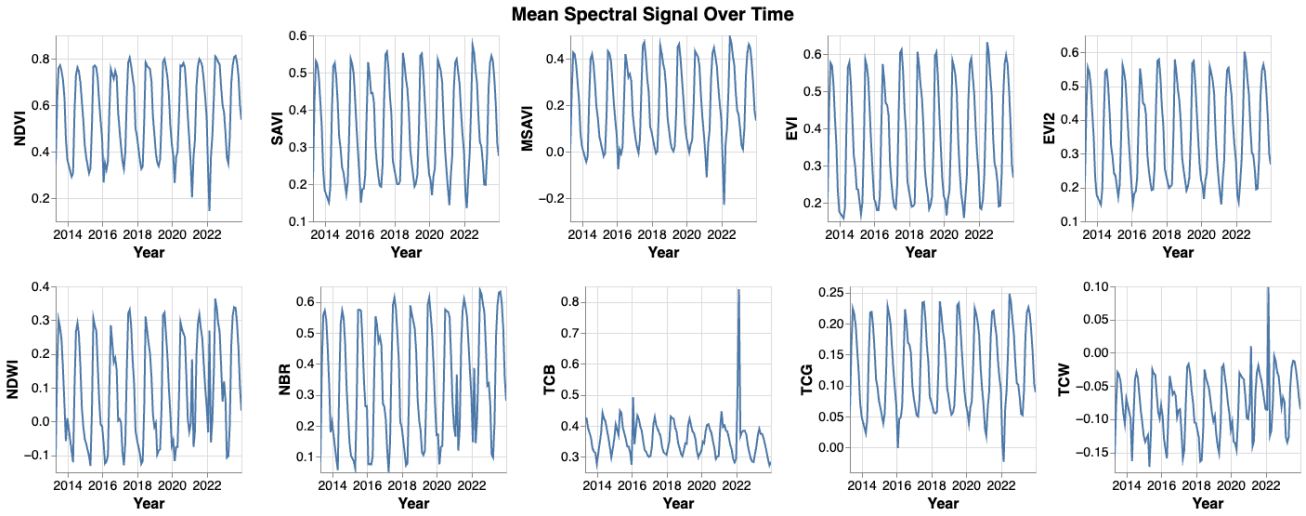

Figure 1: Plot showing seasonality in spectral indices. The seasonal fluctuation is consistent with known vegetation cycles. The signal peaks during summer, when canopy cover is densest and drops sharply in winter when trees shed their leaves.

Finally, we observed strong collinearity among most spectral indices (Figure 2), suggesting the need for dimensionality reduction to prevent overfitting.

Our EDA also revealed some interesting trends in the data. From Figure 3, we can see that the relationship between spectral signals and survival rates varies significantly by species type, suggesting 'Type' could be a viable feature for our model.

Additionally, a positive correlation between spectral indices and tree age was observed in Figure 4, except for TCB. Minimal changes in spectral signals are observed between ages 1 to

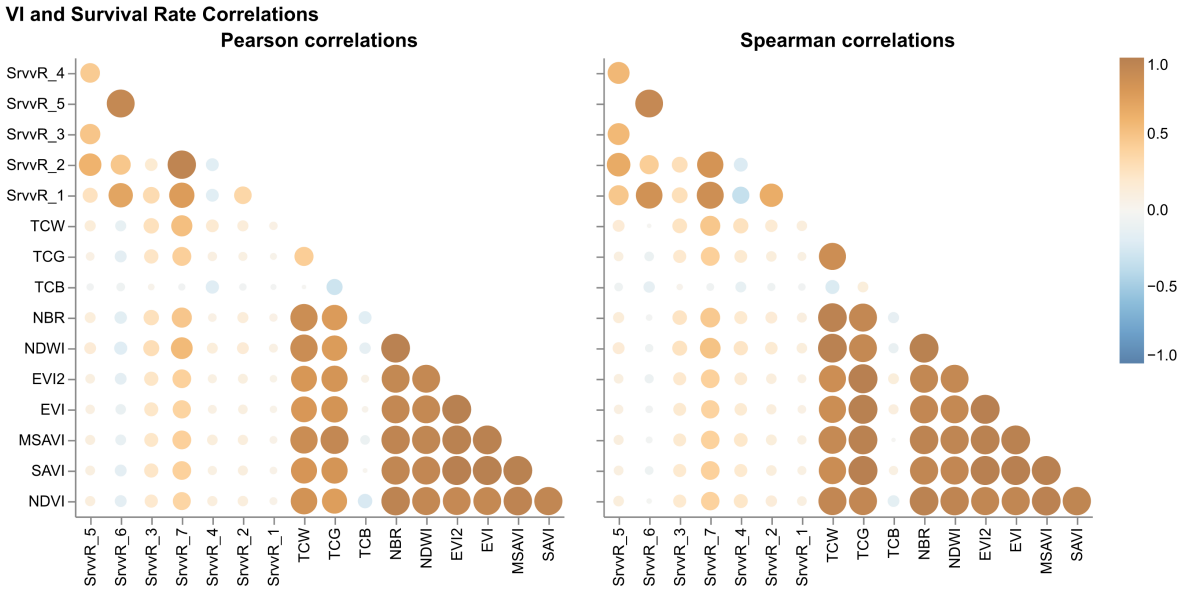

Figure 2: Correlation plot showing strong collinearity between vegetation indices. Notice the exception of TCB, which behaved differently as it is designed to detect soil exposure instead of vegetation. Notably, the correlation between vegetation indices and survival rates is much larger in Year 7 compared to previous years. This again highlights the challenges in early-stage survival rate prediction.

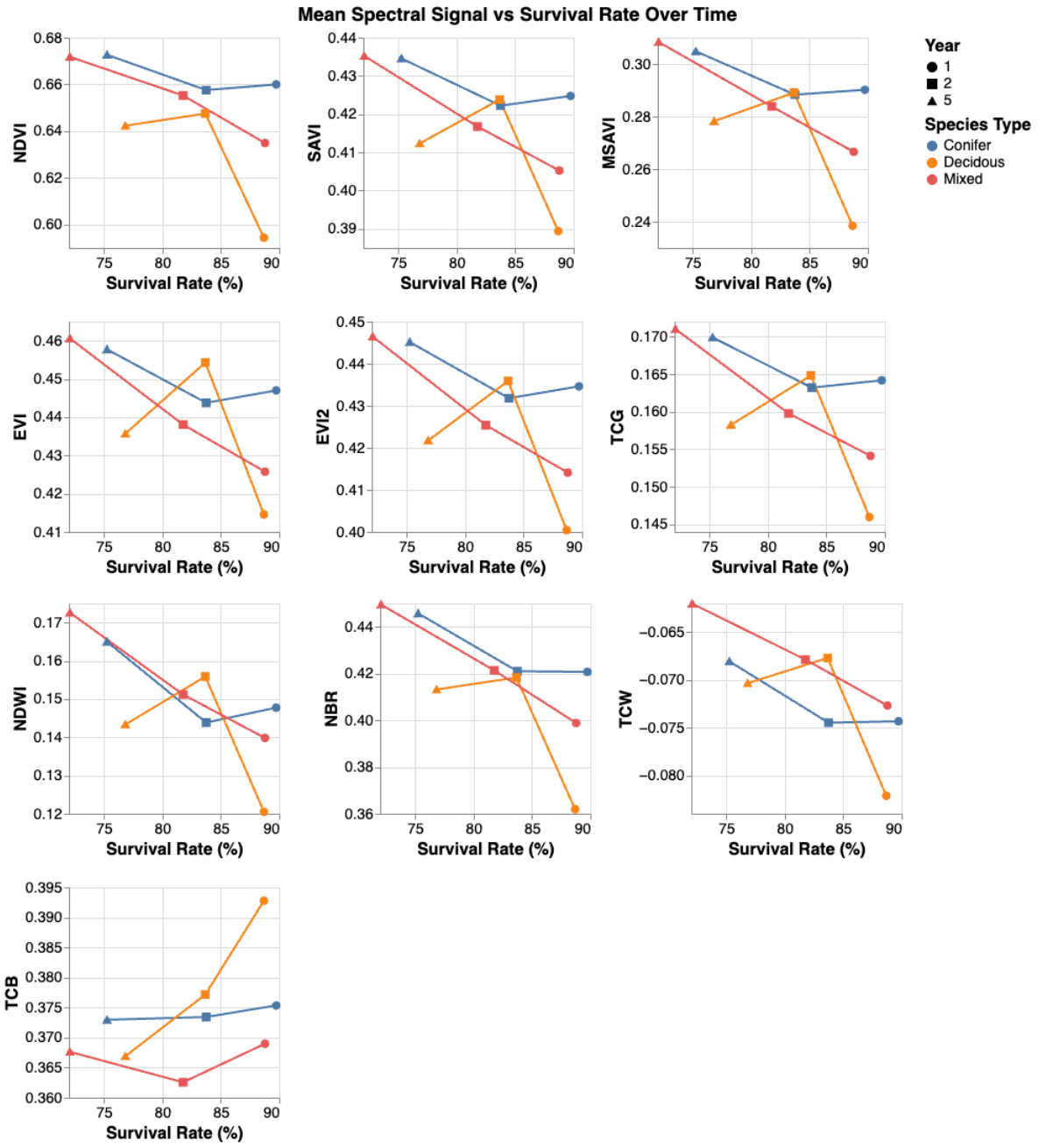

Figure 3: Plot showing mean survival rate and vegetation index signals for different species types across Years 1, 2 and 5. There is a significant difference in the relationship between survival rate and spectral signals for different species types. Conifers show a weaker signal response to changes in survival rate. Deciduous shows the strongest response during the first two years. Mixed type shows a linear relationship between survival rate and spectral signal.

4, suggesting potential challenges in predicting survival rates for younger trees.

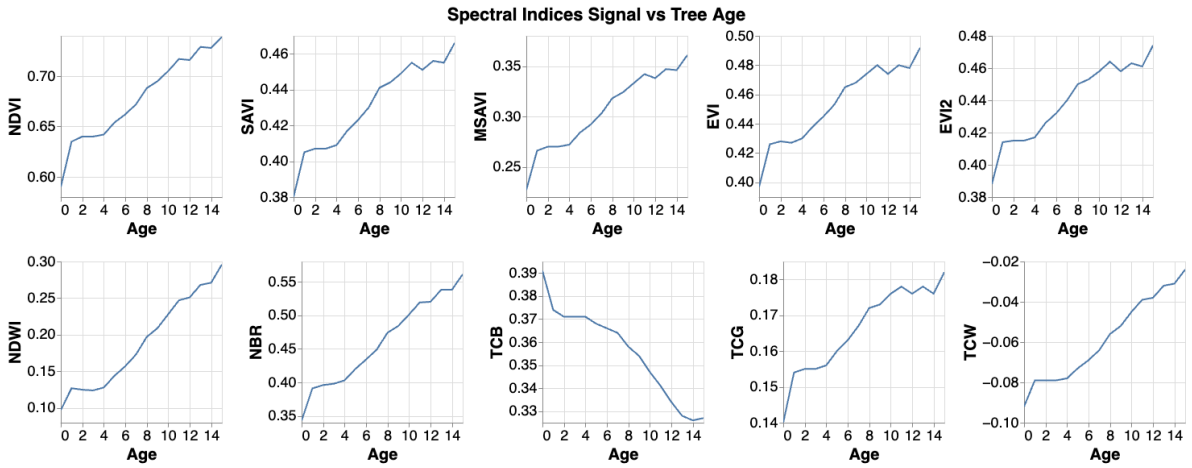

Figure 4: Plot showing mean spectral signal by tree age. Except for TCB, the spectral signal increases with age. A negative relationship was observed for TCB due to lower surface brightness from canopy cover.

4 Data Engineering

Figure 5 outlines the proposed data engineering pipeline for this analysis.

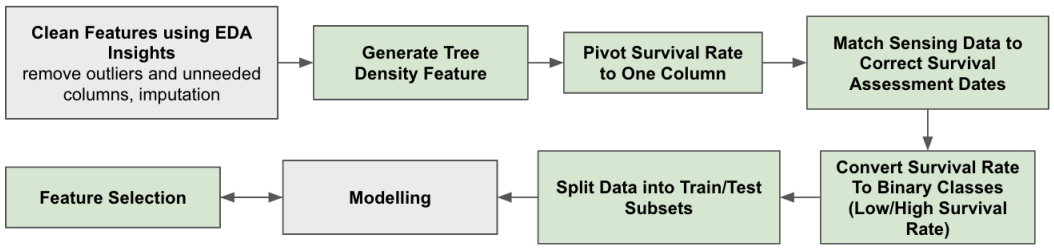

Figure 5: A flowchart of the data engineering process of the model pipeline. Green boxes indicate key data engineering steps, while grey boxes represent other essential steps in the pipeline. For example, modeling must occur in tandem with feature selection, as understanding how each feature impacts model performance is necessary when selecting the optimal features for the final model.

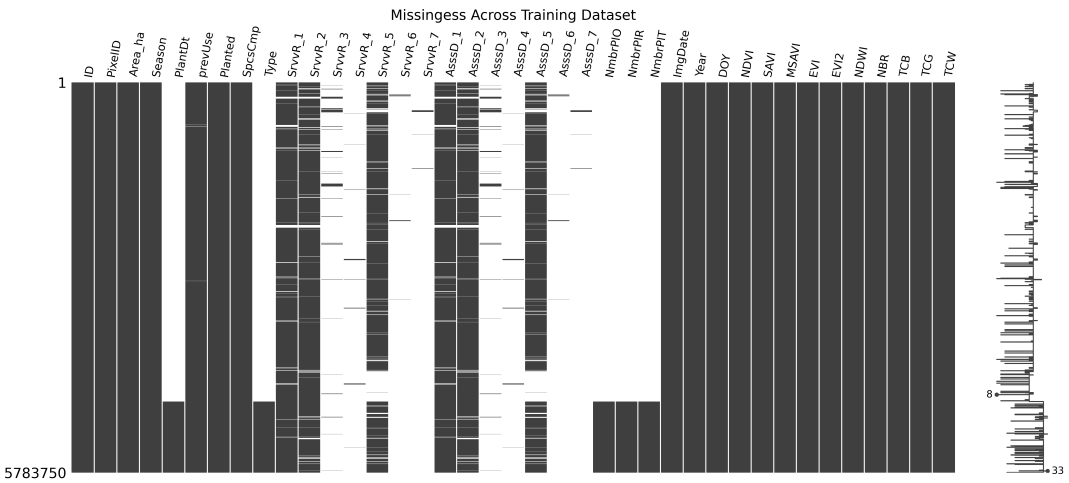

Figure 6: A plot visualizing missing record patterns across the dataset. Each column corresponds to a column in the dataset, and grey-coloured rows indicate non-missing entries.

4.1 Missing Data

Figure 6 illustrates significant missing data in the columns `PlantDt`, `Type`, `NmbrPIO`, `NmbrPIR`, and `NmbrPIT`. `PlantDt`, `NmbrPIO`, `NmbrPIR`, and `NmbrPIT` pertain to sites where replanting has occurred and can be excluded as they fall outside the project scope. The `Type` column can be fully imputed through string processing of the `SpcsCmp` column. There is a direct correspondence between missingness in survival rate and assessment time, allowing for easier tracking of temporal dependence across survival records.

4.2 Feature Engineering

Minimal feature engineering will be conducted, as the primary focus of the analysis is on remote sensing data. However, we aim to experiment with tree density (number of trees per unit area) as a predictor, which can be derived as the quotient of the `Planted` and `Area_ha` columns.

4.3 Data Pivoting

Most machine learning models require the input data to contain just one target column. We will pivot the seven target columns into a single column, tracking temporality using column names and assessment dates. We will then remove rows with missing survival rates and those with mismatching assessment and imaging dates.

4.4 Conversion to Classifier Problem

Since the survival rates are given as percentage proportions, they will be converted to binary classes to simplify the analysis and emphasize high-risk sites. Given that most survival rates range between 70% and 100%, considerations of usefulness and class imbalance are essential when determining an appropriate threshold.

4.5 Train-Test Splitting

Splitting the dataset into training and testing subsets is necessary to prevent data leakage; the model must only be trained on the training data, and the test data cannot be used until the very last stage of model performance evaluation. This ensures that the performance on the test data is a valid estimate of the model's performance in deployment. For many machine learning problems, this can be done by randomly dividing the given data into two subsets. However, the hierarchical structure of the data, as depicted in Figure 7, requires a more thoughtful methodology.

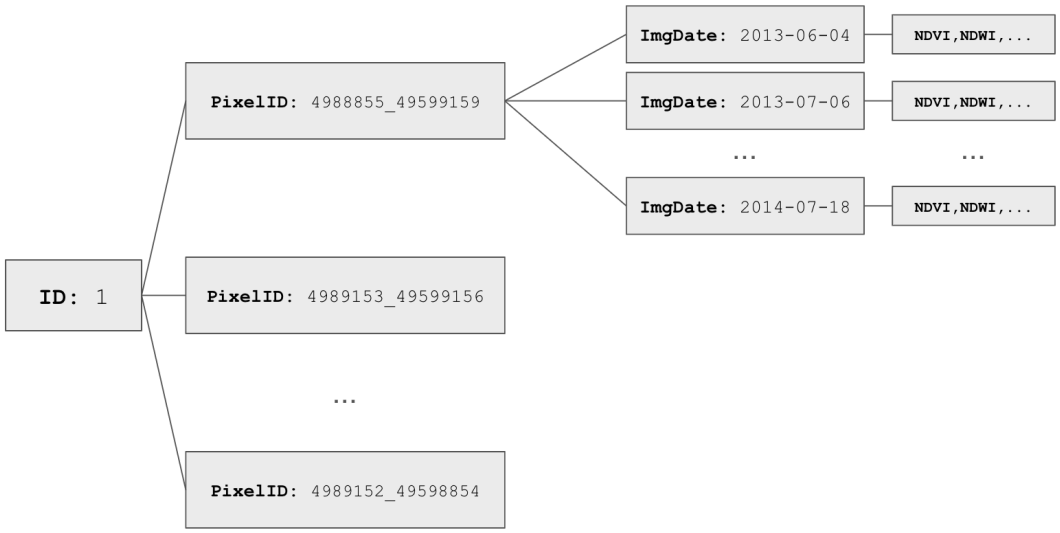

Figure 7: A flowchart depicting the hierarchical structure of the remote sensing data. Unique afforestation sites (which are characterized by the `ID` column) may contain one or more pixels (identified by the `pixelID` column), and each pixel has multiple records of sensing data throughout time, often in monthly intervals (time of imaging is marked by the `ImgDate` column). Since the data is stored in a tabular format, rows correspond to one remote sensing record, for one pixel, at one point in time.

We perform the train-test split by site ID to ensure pixels and time step records for a particular site appear in only one of the two subsets. This approach allows us to fully capture temporal

fluctuations in sensing data via sequential modeling later in the analysis (see Section 5 for further details).

4.6 Feature Selection

As outlined in Section 3.3, collinearity among remote sensing features indicates the need for feature selection; reducing the number of vegetation indices needed by the model may greatly decrease model variance, training time, and computational costs. Table 4 outlines several potential feature selection methods. Additionally, domain knowledge based on spectral index characteristics given by Zeng et al. (2022) may also be leveraged in this process.

Table 4: Comparison of Feature Selection Methods, including **Recursive Feature Elimination** and **Permutation Importance** available in scikit-learn (Pedregosa et al. 2011), as well as methods such as **SHAP Values** (Lundberg and Lee 2017) and **Bayesian Model Averaging** (Hoeting et al. 1998).

Method	Description	Advantages	Disadvantages
Recursive Feature Elimination (RFE)	Iteratively fits a model and removes the least important feature at each step, based on model-derived importance metrics.	Simple to implement within a model pipeline; suitable for tree-based models that provide feature importance metrics.	Computationally expensive for large feature sets; may overlook complex feature interactions; tree-based feature importance can be difficult to interpret.
SHAP Values	Utilizes Shapley values from game theory to quantify each feature's contribution to individual predictions.	Offers detailed insights into feature contributions; captures feature interactions; supported by visualization tools.	Computationally intensive with large datasets; requires understanding of underlying statistical concepts.
Bayesian Model Averaging (BMA)	Combines predictions from multiple models, weighted by their posterior probabilities, to account for model uncertainty.	Considers model uncertainty; suitable for comparing multiple candidate models with varying architectures.	High computational cost; relies on strong assumptions or approximations (e.g., BIC); limited direct support in Python and may be time-consuming to implement manually.

Permutation Importance	Measures feature importance by randomly shuffling each feature and observing the impact on model performance.	Model-agnostic; easy to implement and interpret; available in libraries like scikit-learn.	Sensitive to feature correlation; may not accurately reflect importance in the presence of multicollinearity.

5 Modelling Techniques

Figure 8 provides a brief outline of the proposed modelling plan.

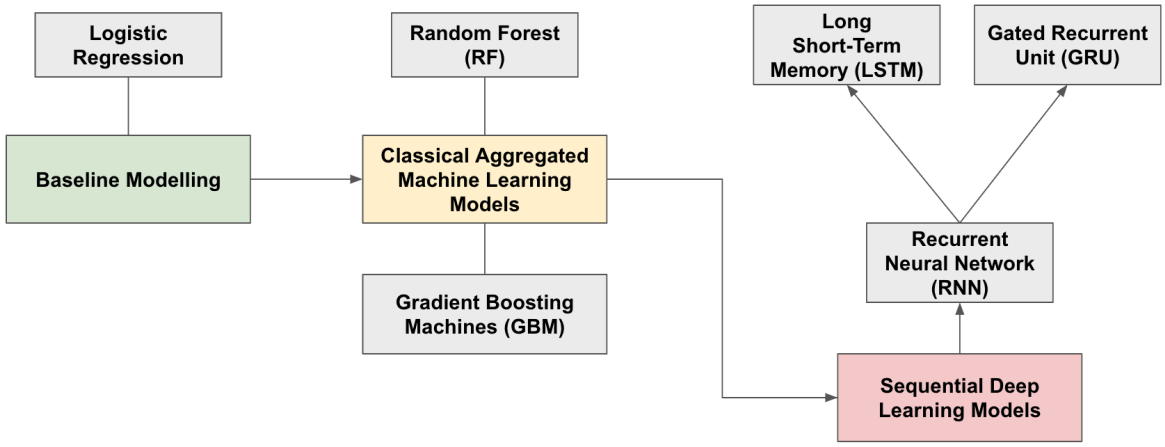

Figure 8: A flowchart depicting the planned modelling during the analysis. Models increase in complexity from left to right and will be implemented in this fashion. It is possible that not all models listed here will be implemented; for example, there may be no need to implement RNNs if the aggregated tree models perform adequately, or if time does not permit their implementation.

We will begin modeling with a baseline **Logistic Regression**, followed by ensemble tree-based methods: **Random Forest** and **Gradient Boosting Machines**. These models were selected based on their demonstrated effectiveness in utilizing vegetation indices to predict tree mortality in similar studies (Bergmüller and Vanderwel 2022). If time permits, we will implement sequential deep learning models such as **Recurrent Neural Networks (RNNs)**, **Long Short-Term Memory (LSTM)** networks, and/or **Gated Recurrent Units (GRUs)**. These models are capable of capturing complex non-linear relationships and may leverage temporal patterns in vegetation indices more effectively (Paszke et al. 2019). A detailed comparison of all models is provided in Table 5.

Table 5: Descriptions of the various modelling techniques that are to be implemented in this analysis.

Model	Description	Advantages	Disadvantages
Logistic Regression	A generalized linear model for binary classification. The log-odds of class membership are modeled as a linear function of input features.	Simple and interpretable; outputs class probabilities; feature importance is directly interpretable via model coefficients.	Limited flexibility for modeling complex patterns; better suited as a benchmark model to be surpassed by more complex models.
Random Forest (RF)	An ensemble of rule-based decision trees are trained on bootstrapped subsets of data. Each tree is overfitted, but aggregation via majority vote reduces variance.	Strong prior performance on remote sensing data; naturally handles nonlinearities and interactions; training is easily parallelizable.	Ignores temporal structure in data; can be memory-intensive and slower to predict on large datasets.
Gradient Boosting Machine (GBM)	An ensemble method where trees are added sequentially, each one correcting errors of its predecessor. Unlike RFs, boosting relies on underfit learners and weighted combinations.	High predictive accuracy; handles complex patterns well; libraries like LightGBM and XGBoost offer efficient training.	Training is slow, sequential and cannot be fully parallelized; like RFs, does not incorporate temporal dependencies.
Recurrent Neural Network (RNN)	A type of neural network designed for sequential data. It processes inputs step-by-step, with hidden states carrying information forward in time.	Captures temporal dependencies; suitable for modeling changes in vegetation indices through time.	Suffers from vanishing gradients, making it difficult to learn long-term dependencies; may require extensive, complex data manipulation to implement; slow to train and requires large datasets.
Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM)	Variants of RNNs that include gating mechanisms (GRU) or memory cells (LSTM) to better manage long-term dependencies in sequences.	May be useful when modeling long-term temporal and/or seasonal patterns in vegetation indices.	The most computationally intensive models of this selection; requires large training datasets; Unclear if long-term differences in vegetation indices will be important to capture for prediction.

6 Success Criteria

The success of this project will be evaluated based on the relevance of selected predictors and the performance of key evaluation metrics (see Table 6). A minimum target of **60%**

accuracy has been set for correct predictions, which will serve as the **primary metric** for communicating results to non-technical audiences. However, due to class imbalance, metrics such as **log loss**, **F1 score**, and **Receiver Operating Characteristic (ROC)/Precision-recall curve (PR) curves** will provide essential insights for technical stakeholders.

Table 6: Evaluation metrics used to assess model performance, including their target audiences

Metric	Description	Audience
Accuracy	Proportion of correct predictions over total	General (e.g., government officials)
Log Loss	Prediction probability to measure uncertainty	Technical
F1 Score	Score of precision over recall	Technical
ROC Curve (AUC)	Plots true positive rate vs. false positive rate across thresholds	Technical
PR Curve (AUC)	Plots precision vs. recall across thresholds	Technical

7 Timeline

Date & Time	Deliverable	Description
May 2, 2025	Proposal Presentations	Group oral presentations framing the objective of **using remote sensing to reduce physical site visits**.
May 6, 2025	Proposal Report – Draft to Mentor	Ungraded draft defining research questions and success metrics for **reducing site visits via remote sensing** and **selecting the best modelling approach** for survival prediction.
May 9, 2025	Final Proposal Report	Final version to partner & mentor with a clear plan showing how we'll achieve each tangible objective—**reduce site visits**, **model selection**, **timely survival prediction**, and **feature-importance ranking**.
June 9, 2025	Data Product – Runnable Draft	Draft pipeline & code to mentor, demonstrating preliminary **model-selection** results, **earliest reliable prediction timing**, and initial **feature-importance rankings**.
June 12–13, 2025	Final Presentation	Group presentation of modelling approach & key results: accuracy, threshold selection vs practical reduction of site visits, comparison of models, prediction timeliness, and top predictors.

Date & Time	Deliverable	Description
June 25, 2025	Final Data Product & Report	Final pipeline, data product, and technical report demonstrating fulfillment of all four objectives: **site-visit reduction**, **model selection**, **timeliness**, and **feature importance**.

8 Conclusion

This project addresses the challenge of monitoring tree survival across hundreds of afforestation sites in Canada's 2 Billion Trees program. By leveraging site-level data and satellite-derived spectral indices, we evaluate machine learning and deep learning models to predict survival over time. Our approach spans from interpretable models like logistic regression to advanced methods such as random forests, gradient boosting, and recurrent neural networks. Identifying key predictors is essential for building effective, scalable tools to support national afforestation efforts.

References

n.d. *UP42 Documentation*. https://docs.up42.com/help/spectral-indexes/nbr#:~:text=NBR%20ranges%20between%20%2D1%20and,have%20values%20close%20to%20zero.

Baig, Muhammad Hasan Ali, Lifu Zhang, Tong Shuai, and Qingxi Tong and. 2014. "Derivation of a Tasselled Cap Transformation Based on Landsat 8 at-Satellite Reflectance." *Remote Sensing Letters* 5 (5): 423–31. https://doi.org/10.1080/2150704X.2014.915434.

Bergmüller, Kai O, and Mark C Vanderwel. 2022. "Predicting Tree Mortality Using Spectral Indices Derived from Multispectral UAV Imagery." *Remote Sensing* 14 (9): 2195.

Hoeting, Jennifer A, David Madigan, Adrian E Raftery, and Chris T Volinsky. 1998. "Bayesian Model Averaging." In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, 335:77–83. Citeseer.

Landsat Missions, USGS. "Landsat Surface Reflectance-Derived Spectral Indices." https://www.usgs.gov/landsat-missions/landsat-surface-reflectance-derived-spectral-indices.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4765–74. Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Mondal, Pinki. 2011. "Quantifying Surface Gradients with a 2-Band Enhanced Vegetation Index (EVI2)." *Ecological Indicators* 11 (3): 918–24.

Natural Resources Canada. 2021. "2 Billion Trees Program." https://www.canada.ca/en/campaign/2-billion-trees/2-billion-trees-program.html.

Paszke, Adam, Sam Gross, Francesco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

University of British Columbia Master of Data Science Program. 2025. "Remote Sensing for Forest Recovery." https://pages.github.ubc.ca/mds-2024-25/DSCI_591_capstone-proj_students/proposals/Remote_Sensing_for_Forest_Recovery.html.

USGS. 2024. "HLS Overview." https://lpdaac.usgs.gov/data/get-started-data/collection-overview/missions/harmonized-landsat-sentinel-2-hls-overview.

Zeng, Yelu, Dalei Hao, Alfredo Huete, Benjamin Dechant, Joe Berry, Jing M Chen, Joanna Joiner, et al. 2022. "Optical Vegetation Indices for Monitoring Terrestrial Ecosystems Globally." *Nature Reviews Earth & Environment* 3 (7): 477–93.