Geo-information Science and Remote Sensing

Thesis Report GIRS-2025-08

EVALUATING DEEP FEATURES FROM PRE-TRAINED TIME-SERIES DEEP LEARNING MODEL FOR PIXEL-LEVEL TREE SPECIES CLASSIFICATION ON FOREST INVENTORY

Takayuki Ishikawa

February 27, 2025

WAGENINGEN UNIVERSITY & RESEARCH

Evaluating Deep Features from Pre-trained Time-Series Deep Learning Model for Pixel-level Tree Species Classification on Forest Inventory

Takayuki Ishikawa

Registration number 1372351

Supervisors:

Marc Russwurm
Carmelo Bonannella

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

February 27, 2025
Wageningen, The Netherlands

Thesis code number: GRS-80436
MSc Thesis report: GIRS-2025-08
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Contents

List of Abbreviations

BA	Basal Area
DW	Dynamic World
ERA5	European Centre for Medium-Range Weather Forecasts Re Analysis v5
GEE	Google Earth Engine
FAO	Food and Agriculture Organization
FRA	Forest Resource Assessment
GHG	Greenhouse Gas
GRD	Ground Range Detected
MAE	Masked Auto Encoder
MLP	Multi Layer Perceptron
NDVI	Normalized Difference Vegetation Index
NFI	National Forest Inventory
RF	Random Forest
RS	Remote Sensing
SAR	Synthetic-Aperture Radar
SITS	Satellite Image Time Series
S1	Sentinel-1
S2	Sentinel-2
SRTM	Shuttle Radar Topography Mission
Presto	Pretrained Remote Sensing Transformer

1 Abstract

Forests play a vital role in mitigating climate change and providing essential ecosystem services such as carbon sequestration. National Forest Inventory (NFI)s serve as the primary source of forest information, providing crucial tree species distribution data for carbon storage estimation and biodiversity assessments. However, maintaining these inventories requires labor-intensive on-site campaigns by forestry experts to identify and document tree species. Remote sensing approaches, particularly when combined with machine learning, offer opportunities to update NFIs more frequently and at larger scales. While the use of Satellite Image Time Series (SITS) has proven effective for distinguishing tree species through seasonal canopy reflectance patterns, current approaches rely primarily on Random Forest (RF) classifiers with hand-designed features and phenology-based metrics.

Recent advances in Artificial Intelligence (AI) through deep neural networks offer a complementary strategy, using learned features from annotated data in an end-to-end fashion. However, while these approaches show superior results compared to traditional methods, they typically require large annotated datasets and substantial computational resources for training—requirements that are particularly challenging for NFIs. Pre-trained deep learning models address these limitations by leveraging unlabeled data through self-supervised pre-training and are freely available for use. For example, Masked Auto Encoder (MAE) enables networks to learn meaningful features by reconstructing masked input time series without requiring ground-truth annotations. These pre-trained models can then be efficiently fine-tuned with smaller labeled datasets for specific classification tasks.

In this work, we investigate two key research questions: "To what extent do deep features extracted from a fine-tuned pre-trained model improve tree species classification accuracy in NFIs compared to traditional harmonic and medoid seasonal composite predictors?" and "What is the effect of domain-specific second-stage pre-training on tree species classification accuracy?" We evaluate these questions using three datasets: the Dutch NFI data (an unbalanced set of 1,479 pure species plots grouped into seven species classes), the Dutch NFI data (an unbalanced set of 1,462 pure species plots with thirteen species classes) and the Francini dataset (a balanced set of 13,790 pure species plots with seven classes). We extracted time-series data from Sentinel-1 (S1), Sentinel-2 (S2) and European Centre for Medium-Range Weather Forecasts Re Analysis v5 (ERA5) satellites data (January-December 2020) and Shuttle Radar Topography Mission (SRTM) data using Google Earth Engine. Comparing deep features from the fine-tuned Pretrained Remote Sensing Transformer (Presto) model against the current state-of-the-art approach using the same RF classifier framework, we found that Presto-derived features substantially outperformed traditional hand-crafted harmonic and seasonal medoid features. Interestingly, additional pre-training on unlabeled Dutch forest time series data did not yield further accuracy improvements.

Our experiments demonstrate that fine-tuning pre-trained deep learning foundation models offers a cost-efficient approach for large-scale tree species classification in NFIs, despite the limited benefits of second-stage pre-training. By leveraging openly available satellite data and pre-trained models, this approach significantly improves classification accuracy compared to traditional methods and can effectively complement existing forest inventory processes. The results highlight the potential of pre-trained deep learning models for enhancing the efficiency and scale of forest monitoring applications.

2 Introduction

Forests play a significant role in mitigating climate change, adopting disaster prevention strategies, and providing ecosystem services, including sequestering carbon dioxide (CO_2), providing wood materials, and serving as a source of biodiversity (Tomppo et al., 2010; FAO, 2020; Francini et al., 2024). Tree species diversity improves productivity of materials and resistance to natural disturbance (Jactel et al., 2017). Therefore, monitoring and tracking records of forests including spatial distribution of tree species are necessary for a sustainable forest management.

National Forest Inventory (NFI)s are the primary source of information for various purposes such as sustainable forest management, industry investment planning, biodiversity monitoring, and Greenhouse Gas (GHG) accounting (Tomppo et al., 2010; Bonannella, 2024). The Paris Agreement also requires the submission of a national carbon inventory, including carbon removals and reductions from forest lands (UNFCCC, 2015). Spatial tree species distribution information plays an important role in NFIs for various applications such as carbon storage estimation, forest management, and biodiversity assessments (Hermosilla et al., 2022; Blickensdörfer et al., 2024). Additionally, detailed tree species information is essential for national reports to the Forest Resource Assessment (FRA) of the Food and Agriculture Organization (FAO) and Forest Europe (M. Schelhaas et al., 2014).

One of the key challenges in managing forest inventory is frequency and scale. Traditional inventory methods are based on sample-based field measurements conducted every 5 to 10 years (Tomppo et al., 2010). Current climate change and land-use changes due to economic growth cannot be captured in a timely manner using these traditional methods (Bonannella, 2024).

Remote sensing helps improve and update NFIs (Francini et al., 2024; Hermosilla et al., 2022). Satellite sensors capture information about Earth every few days to several weeks, allowing us to monitor Earth's condition more frequently and extensively, even in remote areas. However, the global coverage and complex data volumes in remote sensing make human visual interpretation impractical for large-scale analysis.

Machine learning techniques have achieved significant improvements in Earth observations, including tree species classification, particularly when using multiple timestamps of images over large areas (Blickensdörfer et al., 2024; Francini et al., 2024; Hermosilla et al., 2022). The RF algorithm is one of the most popular machine learning algorithms for tree species classification due to its robustness, interpretability, and ability to handle high-dimensional data (Breiman, 2001). However, RF models require well-designed input features engineering, and this selection of appropriate features is crucial for model performance (Heaton, 2016). These choices depend on domain knowledge and target area characteristics such as climate and tree species variety, and often fail to include all necessary features (Ahlswede et al., 2022).

In the context of NFIs, several state-of-the-art machine learning models utilizing RF have emerged (Hermosilla et al., 2022; Blickensdörfer et al., 2024; Francini et al., 2024). These current methods for tree species classification rely on country-specific knowledge for input features, parameter settings, and high-quality data. However, scaling these methods to other countries at a national level presents challenges due to cost constraints and data availability limitations. Furthermore, deep learning models, which generally require large datasets to enhance performance, have remained largely unexplored for national-level tree species classification, primarily due to the limited size of available training datasets. These constraints collectively hinder the transferability of existing methods across different regions or countries.

Deep learning models such as transformer architectures (Vaswani et al., 2023) have been recently introduced for forest monitoring, including tree species classification. This adoption is driven by increasing interest in multimodal and time-series data fusion in Remote Sensing (RS), enabled by the availability of big data and advancements in deep learning models (J. Li et al., 2022). Deep learning models can capture complex patterns in input data and create deep features that can be used for downstream tasks with classifiers and regressors including RF (Basu et al., 2015). While recent studies have achieved success in regional-scale tree species classification using high-quality labeled data, significant gaps remain in large-scale classifications (Fassnacht et al., 2016) for NFIs due to limited labeled data availability and high computational cost for training.

Freely available pre-trained models, trained on large unlabeled datasets containing millions of pixels or images, have emerged as powerful tools for various downstream tasks. These models can achieve comparable or superior accuracy to traditional state-of-the-art machine learning approaches through fine-tuning without computationally expensive pre-training (Bommasani et al., 2022). Self-supervised learning, where models are trained without labels, has gained particular attention in RS applications (Wang et al., 2022). Using self-supervised learning as a model backbone with fine-tuning on limited labeled data has demonstrated significant accuracy improvements (Yu et al., 2022), particularly in time-series analysis tasks, while requiring less inductive bias (Dosovitskiy et al., 2021). The rapid growth of both labeled and unlabeled datasets for RS (Gorelick et al., 2017; Ahlswede et al., 2023) has enabled the development of various pre-trained models for tasks including tree classification (Lu et al., 2024). However, research comparing performance between fine-tuned pre-trained models and traditional approaches for tree species classification in NFIs remains limited.

The second stage of pre-training, where pre-trained models are re-trained on domain-specific data and/or specific tasks, has been shown to improve performance in downstream tasks such as Natural Language Processing (Gururangan et al., 2020) and image-based object identification (Ma et al., 2023). However, the impact of domain-specific pre-training on pixel-level tree species classification in NFIs using unlabeled tree plot data has not been explored.

This research aims to evaluate the effectiveness of deep features from freely available fine-tuned pre-trained models compared with existing feature engineering, and explore the impact of domain-specific pretraining for tree species classification through two key questions:

- RQ1: To what extent do deep features extracted from the fine-tuned time-series pre-trained model improve pixel-level tree species classification accuracy in NFIs when compared to traditional harmonic and medoid seasonal composite predictors (Francini et al., 2024)?
- RQ2: Can additional domain-specific second-stage pretraining with unlabeled forest plots data in the Netherlands improve accuracy for pixel-level tree species classification?

The overall process flow in this study is described in figure 1.

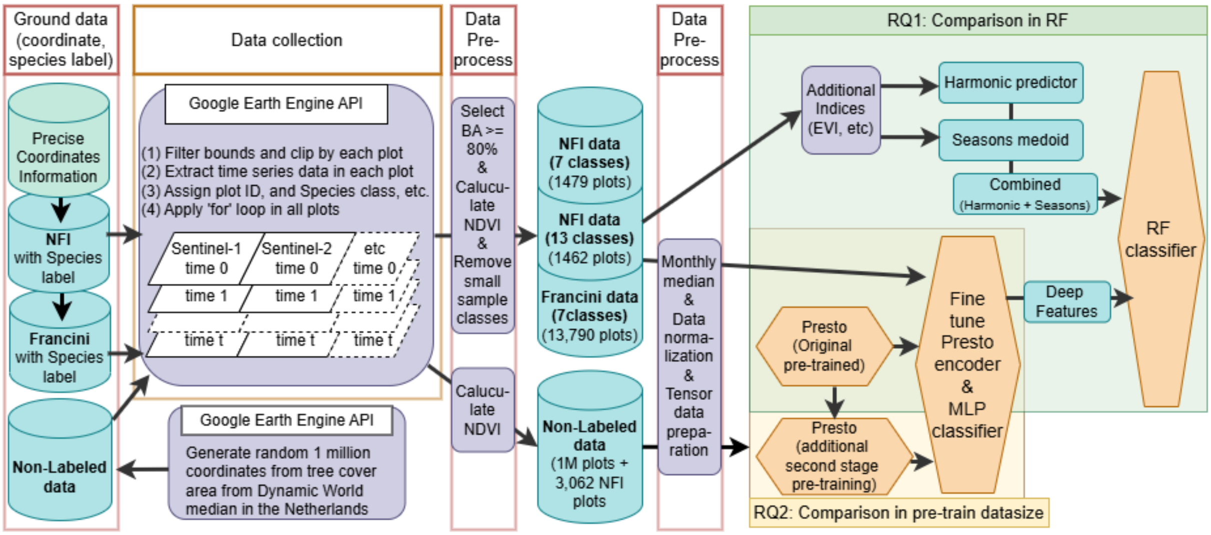

Ground data (coordinate, species label)
Data collection
Data Pre-process
Data Pre-process
RQ1: Comparison in RF
Precise Coordinates Information
NFI with Species label
Francini with Species label
Non-Labeled data
Google Earth Engine API
(1) Filter bounds and clip by each plot
(2) Extract time series data in each plot
(3) Assign plot ID, and Species class, etc.
(4) Apply 'for' loop in all plots
Sentinel-1 time 0
Sentinel-2 time 0
etc time 0
time 1
time 1
time 1
time t
time t
time t
Google Earth Engine API
Generate random 1 million coordinates from tree cover area from Dynamic World median in the Netherlands
Select BA >= 80% & Calcu-late NDVI & Remove small sample classes
Calcu-late NDVI
NFI data (7 classes) (1479 plots)
NFI data (13 classes) (1462 plots)
Francini data (7classes) (13,790 plots)
Non-Labeled data (1M plots + 3,062 NFI plots
Monthly median & Data norma-lization & Tensor data prepa-ration
Additional Indices (EVI, etc)
Harmonic predictor
Seasons medoid
Combined (Harmonic + Seasons)
RF classifier
Presto (Original pre-trained)
Fine tune Presto encoder & MLP classifier
Deep Features
Presto (additional second stage pre-training)
RQ2: Comparison in pre-train datasize

Figure 1: *Overall process flow from data collection to model comparison.*

3 Data

3.1 Study area

This study focuses on forests in the Netherlands, which were described in the annual GHG accounting report (Arets et al., 2023). According to the latest National Forest Inventory 7, forests covered 363,801 ha in 2021, corresponding to 11% of the land use in the Netherlands (M. J. Schelhaas et al., 2022) and comprising approximately 36.4 million 10x10m pixels.

3.2 National Forest Inventory (NFI) data

Ground truth data were collected through field measurements at 3,062 plots for the Dutch National Forest Inventory 6 between 2012 and 2013. Each plot contained a circular area with a variable radius (5 to 20 m) to ensure inclusion of at least 20 trees (M. J. Schelhaas et al., 2022). These plots correspond to 1 to 16 pixels at a 10×10m resolution.

Due to privacy considerations, the precise coordinates of the plots were obtained under a confidentiality agreement with the Dutch government, with the requirement that the data be discarded after project completion. Prior to data disposal, we utilized the plot center coordinates from the total 3,062 pixel-level data points to extract satellite data at 10x10 m pixel resolution from Google Earth Engine (GEE) (Gorelick et al., 2017).

The original NFI data contains 19 dominant tree species classes, which we aggregated into 7 classes to enable comparison with Francini et al., 2024. We selected plots where a single dominant species represented more than 80% of the Basal Area (BA), which is the cross-sectional area of trees at breast height. The *Castanea spp* dominant species class has no samples at this threshold. This selection process yielded 1,479 data points for the aggregated 7 classes classification task.

In addition, we noticed that six dominant species classes have less than 10 samples, which may reduce model performance (Kang et al., 2017), then removed these classes resulting in 1,462 samples with 13 classes for the dominant species classification task.

3.3 Francini data

We also evaluated our methods using the dataset from Francini et al., 2024, which contains 13,790 data points evenly distributed across the aggregated classes (1,970 points per class). This dataset originated from the same NFI data but was augmented with additional labeled data points through visual interpretation of satellite imagery.

3.4 Non-labeled data for pre-training

For the second stage of pre-training, we randomly collected 1 million points coordinates (2.75% of total forest pixels in the Netherlands) sampled from forest areas identified in the Netherlands using the Dynamic World (DW) land classification map (Brown et al., 2022). NFI coordinates with 3,062 plots were attached to this data and total number of non-labeled data is 1,003,062 points. This data does not have species labels, but it is used to pre-train the model.

3.5 Ground data overview

The overview of all ground data used in this study is shown in table 1.

Table 1: Overview of ground data used in this thesis.

Data Type	No. Samples	No. Classes	Notes
NFI (7 classes)	1,479	7	Aggregated to 7 classes
NFI (13 classes)	1,462	13	Classes with less than 10 samples removed
Francini data	13,790	7	Augmented with additional labeled data from NFI data
Non-labeled data	1,003,062	-	Used for pre-training

In table 2, the number of samples for aggregated 7 species groups and 13 dominant species in the NFI dataset, and for aggregated 7 species groups in Francini data is shown.

In figure 2, the distribution of the species classes in the NFI dataset, Francini dataset, and non-labeled data is visualized.

3.6 Satellite data sources

This study implements a comprehensive multi-source RS approach that integrates satellite imagery with environmental data to characterize forest vegetation both spatially and temporally. The primary data sources required as model inputs comprise:

- **S1 Synthetic-Aperture Radar (SAR) Ground Range Detected (GRD) Data:** 6-day revisit frequency, providing VH and VV polarizations
- **S2 Multispectral Data:** Top of Atmosphere Reflectance (Level 1C) with 5-10 day revisit frequency, incorporating 10 spectral bands and Normalized Difference Vegetation Index (NDVI):
  - B2-B8, B8A (blue, green, red, redE1-E4, NIR)
  - B11, B12 (SWIR1, SWIR2)

Table 2: Species count and grouping in NFI and Francini datasets.

Aggregated Group	National Forest Inventory (NFI) data				**Francini data**
	No. samples **(7 classes)**	Dominant Species	No. samples ≥ Basal Area (BA) 80% (19 classes)	No. samples ≥ BA 80% small samples excluded **(13 classes)**	No. samples **(7 classes)**
Pinus	603	*Pinus sylvestris*	513	513	1,970
		Pinus pinaster	1	–	
		Other *Pinus*	89	89	
Larix	56	*Larix* spp	56	56	1,970
Quercus	288	*Quercus robur petraea*	255	255	1,970
		Other *Quercus*	33	33	
Beech	58	*Fagus* spp	58	58	1,970
Populus	72	*Populus* spp	72	72	1,970
Other Broadleaves	242	*Alnus* spp	30	30	1,970
		Betula spp	58	58	
		Fraxinus spp	40	40	
		Castanea spp	0	–	
		Carpinus spp	3	–	
		Abies spp	2	–	
		Robinia pseudoacacia	7	–	
		Other broadleaved	102	102	
DarkConifer	160	*Pseudotsuga menziesii*	90	90	1,970
		Picea spp	66	66	
		Other conifers	4	–	
Total	1,479		1,479	1,462	13,790

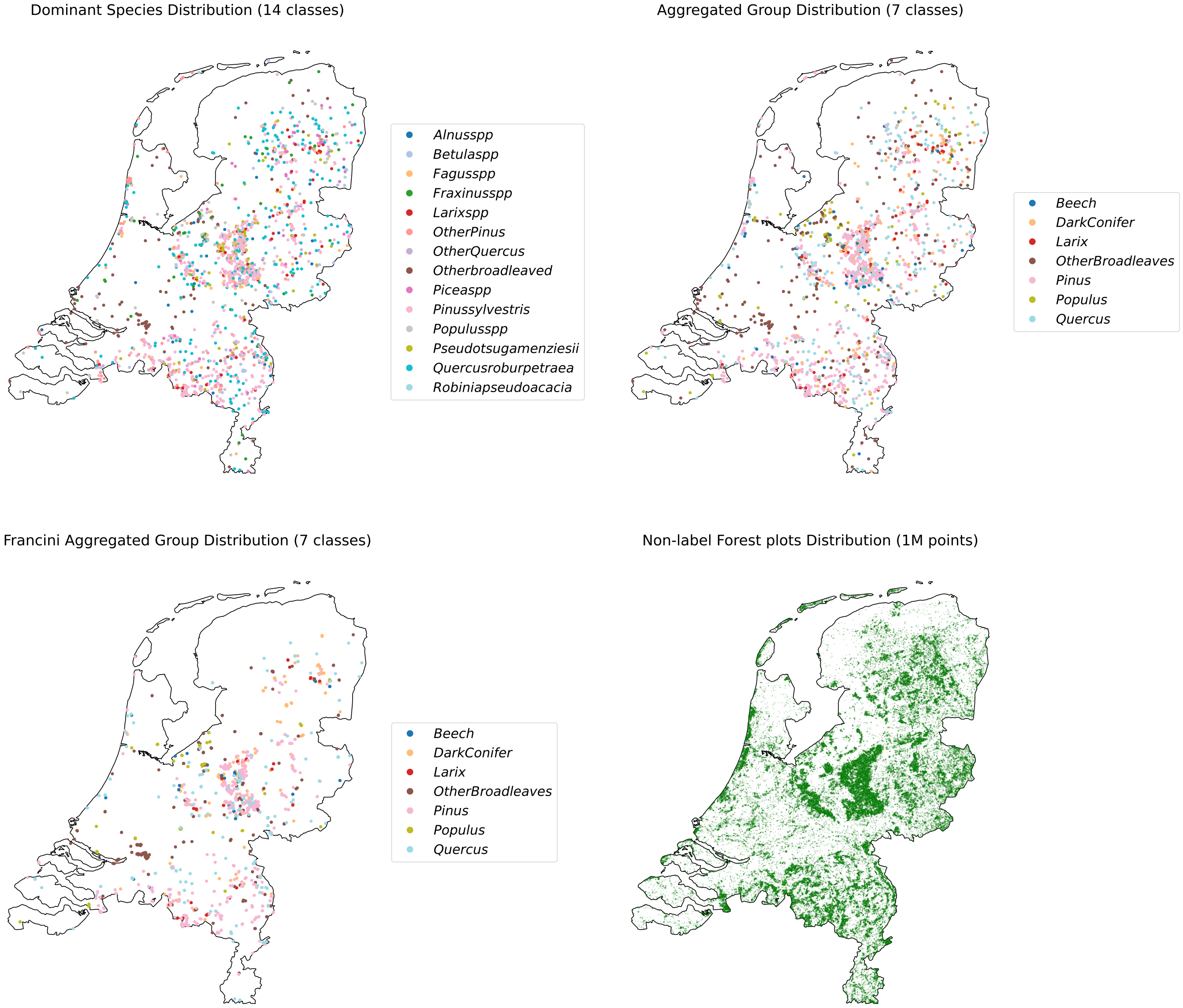


Figure 2: *Top left: 13 dominant species distribution in NFI data. Top right: Aggregated 7 species groups distribution in NFI data. Bottom left: Aggregated 7 species groups distribution in Francini data. Bottom right: Forest plots distribution in non-labeled data.*

  - NDVI computed during preprocessing
- **ERA5 Climate Data:** Monthly measurements (Muñoz Sabater, 2024):
  - 2m temperature (temperature of air at 2m above the surface)
  - Total precipitation
- **SRTM Terrain Data:** static measurement (Farr et al., 2007):
  - Elevation
  - Slope

3.7 Data collection

Since mono-temporal data cannot adequately capture seasonal leaf phenology for tree species classification, we utilized time-series data spanning January through December 2020. This temporal range aligns with both the Presto model requirements (Tseng et al., 2024) and previous Dutch tree species classification research (Francini et al., 2024). For S2 data, cloud masking was performed using the S2 cloud probability dataset (Pasquarella et al., 2023), excluding pixels with cloud cover probability exceeding 65% (Francini et al., 2024).

Geographic coordinates were utilized both for satellite data extraction from the GEE archive and as input features for the model. Raw time-series data was downloaded as individual CSV files for each plot location across all datasets (NFI, Francini, and non-labeled data), ensuring direct correspondence between remote sensing observations and ground truth measurements.

3.8 Data preprocessing

3.8.1 Monthly median value

The model requires monthly median composite values for each band in S1, S2 and ERA5. Therefore, we calculated monthly median values for each band in S1 and S2 except monthly ERA5 data. After getting monthly median values for S1 and S2, NDVI was calculated from S2 bands to capture vegetation dynamics.

In figure 3, the timeseries plot of mean NDVI values for each aggregated 7 tree species classes in the NFI dataset is shown. This plot shows difficulty in distinguishing tree species classes based on mean NDVI values.

3.8.2 Data normalization

Data normalization follows established protocols for each data source in the model (Tseng et al., 2024). S1 backscatter values were normalized to the [-25,25] range to standardize the input features. S2 surface reflectance values were scaled by a factor of 10000, following standard practice. ERA5 climate data underwent unit conversion, with temperature converted to Celsius. SRTM elevation data was scaled by 2000 and slope by 50 to bring all features into comparable numerical ranges, which is crucial for stable model training. Precise normalization was applied to each data source as described in table 3.

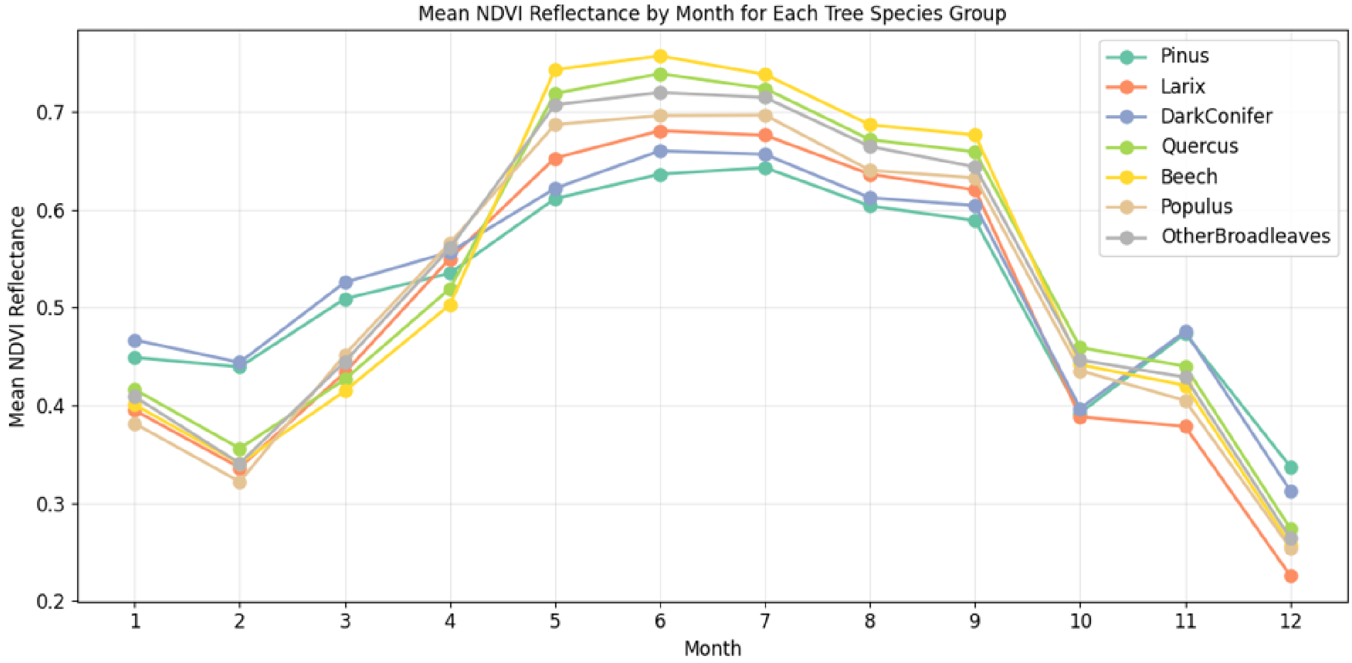


Figure 3: Mean NDVI values for each month and each aggregated tree species class in the NFI dataset.

Table 3: Data normalization parameters defined by original Presto and value ranges in our data by source and variable type.

Data Source	Band/Variable	Before Range	Normalization	After Range
Sentinel-1	VH, VV	-31 to 17 dB	Shifted by $+25$, divided by 25	-0.24 to 1.68
Sentinel-2	All bands	10 to 15,769	Divided by 10,000	0.00 to 1.58
ERA5 monthly	Temperature	278 to 295 K (6 to 23 C)	Shifted by -272.15, divided by 35	0.17 to 0.66
	Precipitation	0.008 to 0.208 m	Divided by 0.03	0.27 to 6.67
SRTM	Elevation	-27 to 331 m	Divided by 2,000	-0.01 to 0.17
	Slope	0.0 to 39.3°	Divided by 50	0.00 to 0.79

3.8.3 Tensor data preparation

All monthly normalized satellite data and tree species information are merged based on plot id and month, which are stored in a single csv file. Based on this data, we prepare the several tensor data for the Presto model input as table 4. Due to the complexity of entry points to original Presto code which expect data coming from Google Cloud storage, we manually prepare the tensor data for the model input.

The bands tensor data is prepared as a 3D tensor with dimensions [plot, month (12), feature]. The feature dimension is 17 for S1, S2, ERA5 data and SRTM data. Additional tensors for latitude and longitude coordinates (latlons), start-month index set to 0 which indicates time-series data starting from January 2020 (month), and categorical tree species information (label) are prepared.

With regard to missing value handling and masking strategy for self-supervised pre-training in bands tensor, the method implements two masking approaches: First, we create initial masks (mask tensor) for missing or zero values, which are used to ignore some data during pre-training, finetuning and feature extraction. The method applies band group-specific masking where if any value in a group (e.g., RGB group in S2 bands) is missing, the entire group is masked. These band groups are defined by Presto. Second, pre-training masks for self-supervised learning are generated using a Presto-native masking strategy with a 75% mask ratio, including group bands, random timesteps, and chunk timesteps masking strategy. This ratio is applied uniformly across bands and timesteps to create a training mask for the model. The process culminates in the creation of several key tensors by combining masks using these two types of masks (combined mask), input tensor (X), which contains the original bands data with masked values set to zero, and target tensor (Y) which contains only the masked values for reconstruction.

In addition, dynamic world tensor with all tree classes value (1) is prepared because it is necessary for model input, even though dynamic world cross entropy losses are not used during pre-training. Based on dynamic world mask (mask dw), dw input tensor (dw x) and dw target tensor (dw y) are prepared.

This tensorization pipeline ensures consistent handling of multi-source earth observation data while preparing it for pre-training and fine-tuning.

Table 4: Tensor Outputs for Presto

Tensor	Shape	Description
bands tensor	(B, T, F)	Original data tensor containing all band values, where B is batch size, $T = 12$ timesteps, and $F = 17$ features (bands)
dynamic world	(B, T)	Dynamic World classification tensor with values $\{1, 9\}$, where 1 represents tree class and 9 represents masked values
latlons	$(B, 2)$	Geographical coordinates tensor containing latitude and longitude
labels	(B)	Target labels tensor containing class indices
month	(B)	Temporal indicator tensor (currently set to zeros)
mask tensor	(B, T, F)	Boolean tensor indicating naturally missing or invalid values
mask dw	(B, T)	Boolean tensor for Dynamic World masking strategy
combined mask	(B, T, F)	Boolean tensor combining natural missingness and masking strategy
x	(B, T, F)	Input tensor with masked values set to zero
y	(B, T, F)	Target tensor containing only the masked values
dw x	(B, T)	Input tensor with Dynamic World masked values set to zero
dw y	(B, T)	Target tensor containing only the Dynamic World masked values

Where: B = Batch size (number of samples), T = Number of timesteps (12 months), F = Number of features (17 bands)

4 Methods

4.1 Deep learning

4.1.1 Pre-trained model overview and Presto architecture

Several pre-trained models in RS have been developed since 2021, trained on different datasets and designed for various downstream tasks such as scene classification, semantic segmentation, object detection, and change detection (Lu et al., 2024). The accessibility of a model and its pre-trained weights is important for model selection due to limitations in computational power and time, and the choice of pre-trained datasets and expected downstream tasks is crucial for tree species classification using time-series data such as S2. Major models are trained by self-supervised learning using MAE as described in Figure 4 (He et al., 2021). This methods does not require labeled data, which is beneficial for RS tasks where labeled data is scarce.

We adopted the Presto model (Tseng et al., 2024) for its demonstrated capabilities in handling multi-temporal and pixel-based satellite data. This model is based on transformer structures (Vaswani et al., 2023) and was trained on 21.5 million pixel time-series with 12-month contiguous intervals. Each month's composite satellite data included S1, S2 and its NDVI, ERA5, DW, and SRTM (Figure 5), extracted between 2020-01-01 and 2021-12-31. The advantages of this model include the ability to handle freely available multi-source and multi-temporal data even when some sources (e.g., ERA5) or temporal data (e.g., S2 values in November) are missing, and computational efficiency when processing large areas.

Deep features (encodings/ embeddings) with 128 dimension values were extracted from encoders of best fine-tuned models with input data in figure 6. This features were used for tree species classification with RF and Multi Layer Perceptron (MLP) classifier (RQ1).

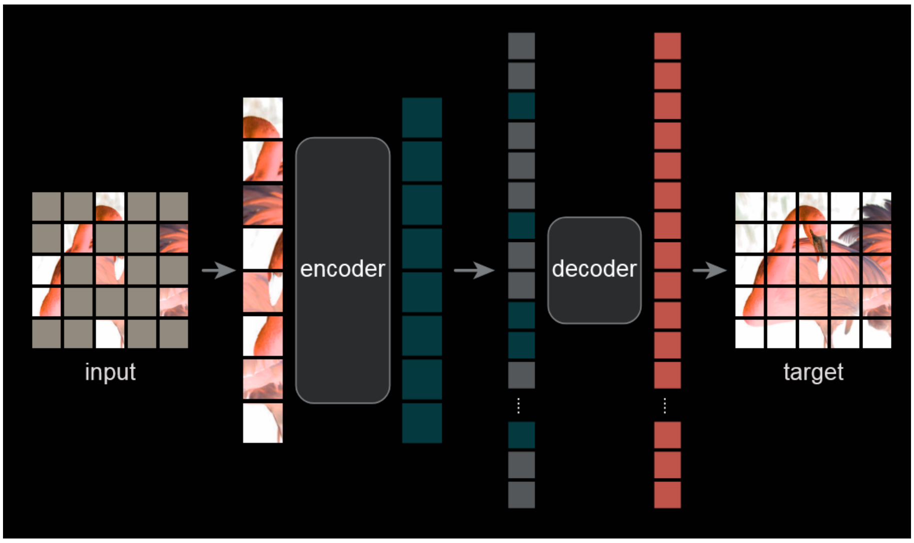


Figure 4: Original Masked Auto Encoder (MAE) Structure (He et al., 2021) for Vision transformer (Dosovitskiy et al., 2021). Encoder block learns the representation of the masked input data, and decoder block reconstructs the original input data from the representation.

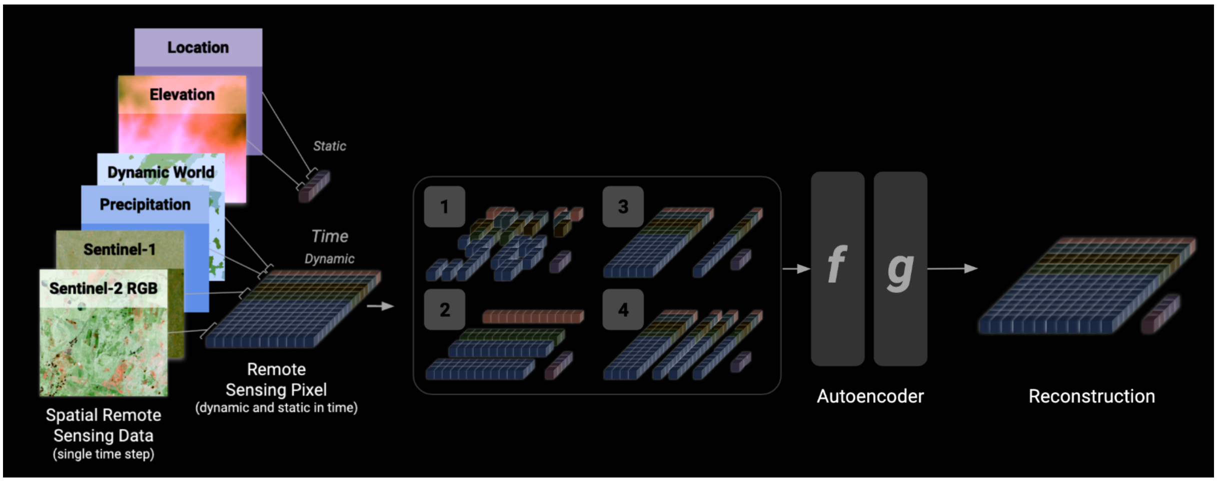


Figure 5: In Pretrained Remote Sensing Transformer (Presto), the masked auto encoder (f) learns spatial time series representations from masked multi-source and multi-temporal data. The decoder (g) reconstructs the masked-out part of the input. During this self-supervised learning process with non-Labeled data, the encoder (f) learns the data representation, which can be used for various downstream tasks.

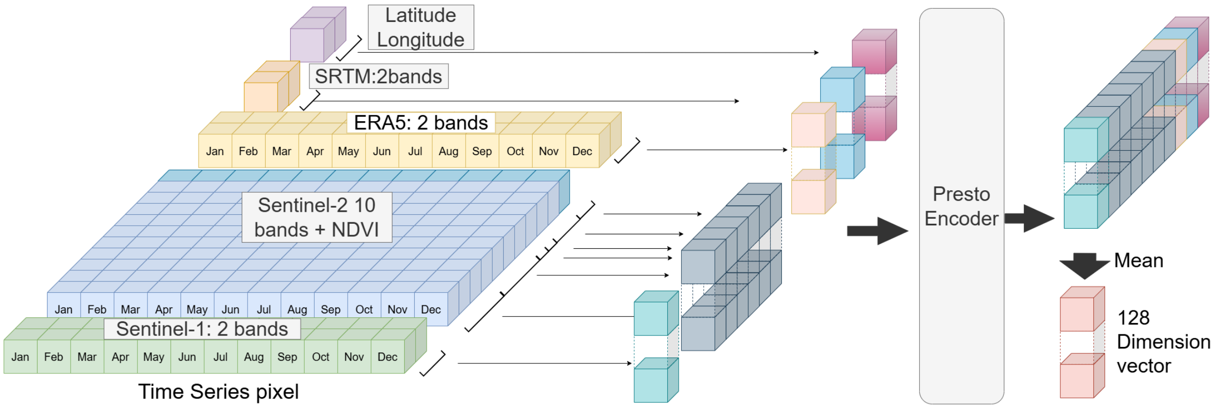

Latitude
Longitude
SRTM:2bands
ERA5: 2 bands
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Sentinel-2 10 bands + NDVI
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Sentinel-1: 2 bands
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Time Series pixel
Presto Encoder
Mean
128 Dimension vector

Figure 6: Presto encoder feature extraction for downstream tasks.

4.1.2 Pre-training (RQ2)

Pre-training parameters were selected as same as the original Presto model to make it comparable. We used an AdamW optimizer, a cosine annealing schedule for our learning rate, with a maximum learning rate of 0.001 at the 2nd epoch. We applied a weight decay of 0.05, and β of (0.9, 0.95). Several pre-train datasize were evaluated while 3,062 samples which extracted from NFI coordinates are consistently added. Therefore, 53062, 103062, 203062, 403062, 1003062 samples were used for pre-training. 2% of dataset is used for validation as original Presto allocate 1 data container out of 59 data containers for its first pre-training. Validation data was used to decide the best model based on the lowest validation loss.

With regard to batch size, too small a batch size could lead to local minima and cause instability during pre-training (Z. Li et al., 2020). In addition, although larger batch sizes result in better pre-trained models for a fixed amount of iterations (Vaessen & Leeuwen, 2024), too large batch size will lead to the model stucking in a sharp minima in a limited pre-training size (Keskar et al., 2017). Considering Presto pre-training which has 4,096 batch size and 20 epochs, we used same batch size and epochs for different different pre-train dataset size to ensure consistency with original pre-training.

The model employs dynamically scaled batch sizes and steps based on dataset size in table 5.

Table 5: Pre-training parameters for different dataset sizes

Dataset Size	Batch Size	Steps per Epoch	Total Epochs	Total Steps
53,032	4,096	13	20	260
103,062	4,096	25	20	500
203,062	4,096	50	20	1,000
403,062	4,096	98	20	1,960
1,003,062	4,096	245	20	4,900

In the original Presto, the loss calculation combines mean squared error reconstruction loss for band values and DW cross entropy loss for DW land class. In our setting, since all land class is designated as tree (class 1) in DW, we omitted the DW cross entropy loss component.

We run the model with 20 epochs 5 times each to ensure the model's stability and reproducibility.

The model was trained on a single NVIDIA A100 GPU with 80GB memory. The training process was monitored, and the best model was selected based on the lowest validation loss.

4.1.3 Fine-tuning (RQ1, RQ2)

For fine-tuning the pre-trained models, the encoder block from the pre-trained model was reused and trained with newly attached Multi Layer Perceptron (MLP) classifier for the downstream task. This fine-tuning model is trained on the 1,479 samples for 7 aggregated groups classification task and the 1,462 samples for the 13 dominant species classification task in both NFI data and Francini data. Pre-trained encoder was set as trainable in this process.

Based on the current state-of-the-art classifier of MLP on tree species classification (Mouret et al., 2024), this classifier architecture consists of a 3-layer MLP with no dropout (rate: 0.0). The layer configuration comprises 1024 nodes in the first layer, 512 nodes in the second layer, and 256 nodes for the last layer, with batch normalization and ReLU activation applied throughout. The training parameters are configured as follows: a learning rate of 0.0001 is used with a weight decay of 0.00746. The model is trained for a maximum of 100 epochs using a batch size of 64. The training process employs cross-entropy loss for optimization. The best fine-tuned models and its accuracy were selected based on best validation loss.

4.2 Classic machine learning with RF (RQ1)

4.2.1 Random Forest Classifier

We used a scikit-learn RandomForestClassifier with 500 estimators, maximum features set to the square root of the number of features, which are the same setting as the RF in Francini et al., 2024.

4.2.2 Feature Engineering

For comparison with Francini et al., 2024 method, two feature engineering are applied. Although the original method used only S2, we found that the combination of S1 and S2 further enhance accuracy in NFI data based on the results in table 6. Therefore, total 12 bands from S1 and S2 are selected and additional 7 indices are calculated from S2 including NDVI, Normalized Burn Ratio (NBR), Enhanced Vegetation Index (EVI), and Tasseled Cap transformations: Brightness (TCB), Wetness (TCW), Greenness (TCG), Angle (TCA). Based on these 19 bands, two feature engineering approaches were implemented:

Harmonic Features : Seven harmonic metrics (β_0, β_1, β_2, β_3, A, ϕ, and $RMSE$) were calculated for each band (total 133 bands): β_0 is the constant, β_1 is the time coefficient, and β_2 and β_3 are the frequency sine and cosine coefficients, respectively. To fit these four coefficients and to select P_t, the pixel p harmonic values at time t, we used a least squares regression to fit Eq. 1. A is the amplitude of the harmonic curve on the y-axis (Eq. 2), ϕ is the phase of the curve on the x-axis to the origin (Eq. 3) and $RMSE$ is root mean square error between P_t and the actual pixel values X_t (Eq. 4).

$$P_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi\omega t) + \beta_3 \sin(2\pi\omega t) \tag{1}$$

$$A = \sqrt{\beta_2^2 + \beta_3^2} \tag{2}$$

$$\phi = \arctan(\beta_3/\beta_2) \tag{3}$$

Table 6: Classification accuracy comparison across different feature combinations and satellite data sources in NFI data (in %). Bold values indicate best performance for each label type.

Label	Features	Sentinel-1		Sentinel-2		S1-S2 Combined	
		Band	Acc.	Band	Acc.	Band	Acc.
Dominant Species (13 classes)	Seasonal (S)	S1	51.03	S2	64.03	S1-S2	64.52
	Harmonic (H)	S1	55.04	S2	63.79	S1-S2	64.64
	All (S+H)	S1	57.72	S2	66.34	S1-S2	**66.71**
Group (7 classes)	Seasonal (S)	S1	55.60	S2	70.18	S1-S2	71.93
	Harmonic (H)	S1	60.19	S2	71.39	S1-S2	71.52
	All (S+H)	S1	62.75	S2	72.87	S1-S2	**74.22**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_t - X_t)^2} \tag{4}$$

where t is time, $\omega = 1$ (in years) is the frequency and indicates one cycle per unit of time, initial guess values before fitting four coefficients were set as $\beta_0 = 0.1$, $\beta_1 = 0.1$, $\beta_2 = 0.4$, $\beta_3 = 0.4$.

Seasonal medoid Features : Medoid compositing selects pixel values from a specific time point whose reflectance values are most similar to the median values of the entire image collection, rather than simply using the median values themselves. Season-wise medoid representative values were selected for each band, resulting in 19 bands for each of the 4 seasons (total 76 bands): winter (January, February, December), spring (March-May), summer (June-August), and autumn (September-November).

4.3 Model validation

4.3.1 Validation metrics

Model performance was evaluated using a comprehensive set of established accuracy metrics: per-class precision and recall, F1 scores, confusion matrices, and overall accuracy. While overall accuracy provides a general assessment of classification performance, it exhibits reduced sensitivity to class imbalance issues that are prevalent in ecological datasets. Therefore, the F1 score, which represents the harmonic mean of precision and recall, was employed as a supplementary metric to provide a more balanced evaluation of model performance when addressing the inherent class imbalances in tree species distribution data. These metrics are widely used in tree species classification studies (Goutte and Gaussier, 2005; Hermosilla et al., 2022; Francini et al., 2024; Blickensdörfer et al., 2024) and provide comprehensive assessment of classification performance across all species classes.

Precision : Precision is calculated by the ratio of true positive (TP) predictions to the total predicted positives (true positives (TP) + false positives (FP)).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

Recall : The recall is calculated by the ratio of true positives (TP) predictions to the total actual positives (true positives (TP) + false negatives (FN)).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

F1 Score : The F1 score combines both precision and recall and gives a balanced metric to assess the model's accuracy. The F1 score is particularly useful for us since we have a class imbalance.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Confusion Matrix : Confusion matrices have been used to visualize the accuracy statistics per classification class in the format shown below.

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Overall Accuracy: Calculates the proportion of correctly classified instances for both True Positives (TP) and True Negatives (TN) to the total number of predictions.

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

4.3.2 Train-test split

Train-test split ratio is a critical factor in model performance on downstream tasks. If dataset size is large, 80/20 or higher ratios are commonly used (Rácz et al., 2021; Joseph, 2022). Based on our imbalanced dataset and previous research (Blickensdörfer et al., 2024), a 70/30 (approximately 2:1 ratio) train-test split was applied for fine-tuning and RF classification to increase model performance and reduce standard deviation on tree species classification over large areas for NFI. During fine-tuning and subsequent RF classification for features extracted from the fine-tuned model, the same training and test data were used. This practice ensures that extracted features are not trained on test data, which would cause data leakage. By adopting a class-per-split strategy, each class is equally split into train and test data at this ratio.

In remote sensing machine learning, ensuring that the train-test split strategy does not introduce spatial autocorrelation is crucial for obtaining unbiased model performance estimates (Karasiak et al., 2022). Our NFI data accommodates a sampling strategy which selects each plot with a density of 1 point per 100ha (M. Schelhaas et al., 2014). This means that each random point is designated per square kilometer. Additionally, the Francini data removed adjacent pixels with a minimum distance of 15m (Hermosilla et al., 2022; Francini et al., 2024). These sampling strategies ensure that the train-test split does not introduce strong spatial autocorrelation in the data; therefore, no additional train-test split strategy was applied in this study.

4.3.3 Model replication

To evaluate variability of model performance, each model was evaluated five times with different random seeds, which affect to train-test split and model initialization. The mean and standard deviation of the accuracy metrics were calculated for each model. This approach provides a more robust evaluation of model performance.

5 Results

5.0.1 Research Question 1: Deep features performance

Comparison with current state-of-the-art models

We compared the performance of our deep features model with current state-of-the-art models using harmonic and medoid predictors. Table 7 shows that the deep features extracted from the pre-trained Presto model consistently outperformed the harmonic and medoid predictors across all dataset types by a substantial margin, even when using the same Random Forest (RF) classifier.

Table 7: Classification performance comparison between deep features and harmonic+medoid features (in %). Each model was run five times and the mean and standard deviation (±) are reported. Bold values indicate the best performance between deep features extracted from Presto and harmonic+medoid features per dataset type.

	Dataset type	Data size	Random Forest	
			Deep feat. (Tseng et al., 2024)	Harm.+med. (Francini et al., 2024)
Overall Accuracy	NFI data (7 classes)	1479	**74.62 ± 1.93**	70.58 ± 1.78
	NFI data (13 classes)	1462	**66.77 ± 1.51**	62.71 ± 1.51
	Francini data (7classes)	13,790	**95.28 ± 0.60**	84.31 ± 0.99
F1 Score	NFI data (7 classes)	1479	**60.54 ± 3.08**	51.57 ± 3.36
	NFI data (13 classes)	1462	**46.84 ± 2.65**	38.54 ± 2.14
	Francini data (7 classes)	13,790	**95.27 ± 0.60**	84.26 ± 0.99

For the NFI dataset with 7 classes (n=1,479), models trained with deep features achieved an overall accuracy of 74.62% (±1.93), compared to 70.58% (±1.78) for the harmonic+medoid approach—an improvement of approximately 4 percentage points. The difference was even more pronounced when examining F1 scores, where deep features yielded 60.54% (±3.08) versus 51.57% (±3.36) for traditional features, representing an improvement of nearly 9 percentage points.

When the classification task became more challenging with the 13-class NFI dataset (n=1,462), deep features maintained their advantage, achieving an overall accuracy of 66.77% (±1.51) compared to 62.71% (±1.51) for traditional features. Similarly, the F1 score showed a substantial improvement from 38.54% (±2.14) with traditional features to 46.84% (±2.65) with deep features—a gain of more than 8 percentage points. This indicates that the deep learning approach handles the increased complexity of multi-class classification more effectively.

The most substantial performance differential was observed in the Francini dataset (n=13,790), which contained a considerably larger training sample size. In this dataset, deep features achieved a remarkable overall accuracy of 95.28% (±0.60), representing an improvement of nearly 11 percentage points over the traditional features' accuracy of 84.31% (±0.99). The F1 scores demonstrated comparable improvements, with deep features achieving 95.27% (±0.60) compared to 84.26% (±0.99) for traditional features.

The confusion matrices in Figure 7 provide deeper insights into the classification performance differences between deep features and traditional features approaches across all tested datasets. These visualizations reveal important patterns in how each model handles the complex task of forest type classification.

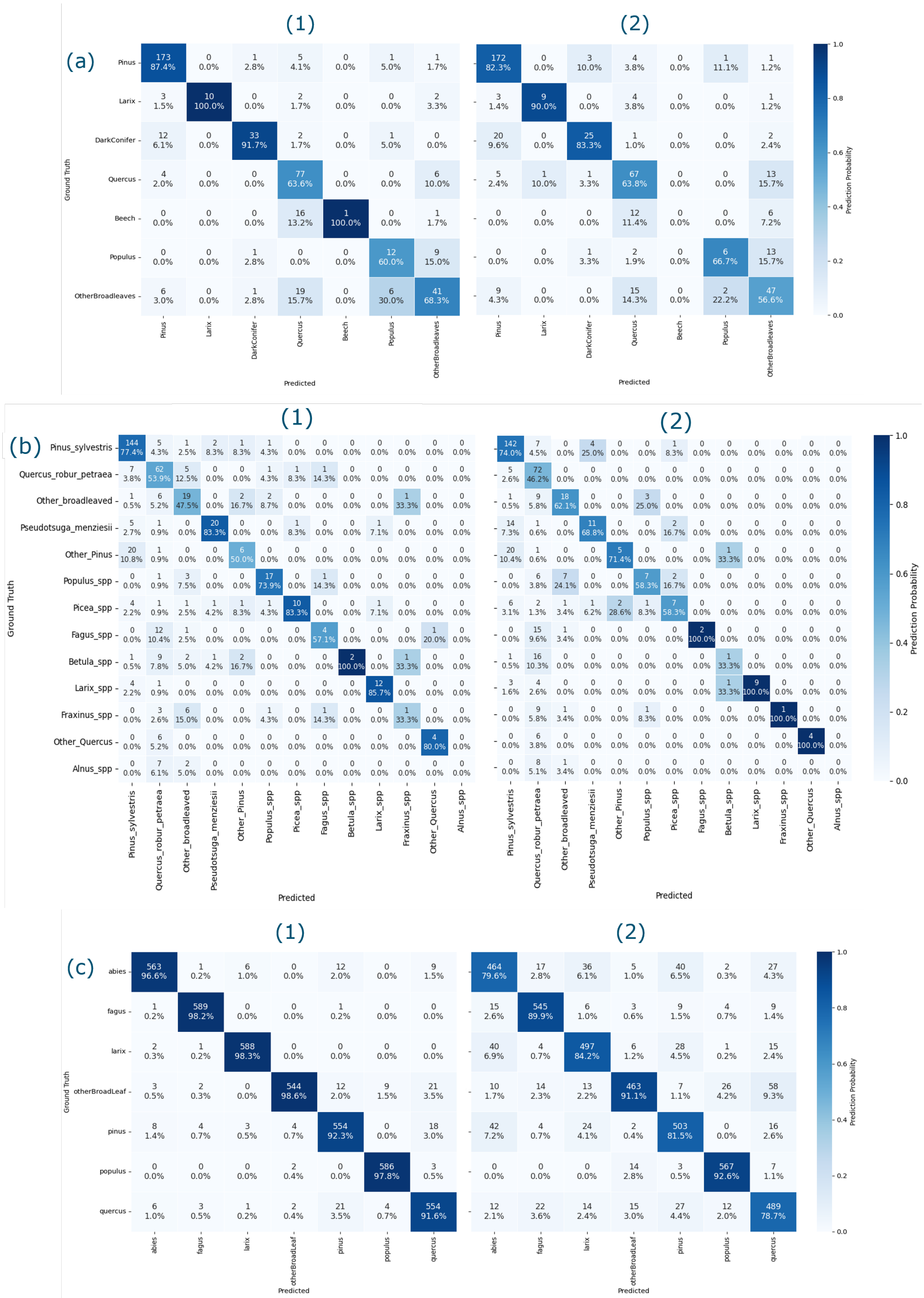


Figure 7: *Confusion matrices for the best performing models. Columns: (1) RF classifier for deep features, (2) RF classifier for harmonic and medoid features (Francini et al., 2024). Rows: (a) NFI data (7 classes), (b) NFI data (13 classes), (c) Francini data (7 classes). In each cell, the first value shows the number of samples predicted, and the second value indicates the percentage of true positive samples among all true positive and false positive samples per class (precision).*

In the 7-class NFI dataset, the deep features model generally outperforms traditional features, particularly for coniferous species groups. However, for deciduous classes, results are mixed: *Quercus* (63.6%) and *Populus* (60.0%) show slightly lower precision than traditional features (63.8% and 66.7%, respectively). Both models struggle with the Other Broadleaves class, often confusing it with *Quercus* and *Populus*.

With 13 classes, classification becomes more challenging, yet deep features improve precision across multiple species: *Pinus sylvestris* (77.4% vs. 74.0%), *Quercus robur/petraea* (53.9% vs. 46.2%), *Pseudotsuga menziesii* (83.3% vs. 68.8%), *Populus spp.* (73.9% vs. 58.3%), *Picea spp.* (83.3% vs. 58.3%), and *Betula spp.* (100% vs. 33.3%).

For the Francini dataset, deep features achieve near-perfect classification, minimizing confusion between classes. Precision remains consistently high across all forest types, highlighting the robustness of deep feature representations.

The improved class separation in the deep features model suggests that these features effectively encode species-specific temporal and spectral signatures that traditional handcrafted features cannot capture as effectively, especially with a large and balanced training dataset.

Comparison of MLP and RF classifiers with Deep Features

We also evaluated the Multi Layer Perceptron (MLP) classifier for tree species classification (Mouret et al., 2024) on the same deep features extracted from the pre-trained Presto model (Table 8).

Table 8: Classification performance comparison between MLP and RF classifiers with deep features (in %). Each model was run five times and the mean and standard deviation (±) are reported. Bold values indicate the best performance between MLP and RF classifier per dataset type.

	Dataset type	Data size	Results MLP classifier	 RF classifier
Overall Accuracy	NFI data (7 classes)	1479	**77.13 ± 1.17**	74.62 ± 1.93
	NFI data (13 classes)	1462	**67.99 ± 1.33**	66.77 ± 1.51
	Francini data (7 classes)	13,790	**98.21 ± 0.20**	95.28 ± 0.60
F1 Score	NFI data (7 classes)	1479	**68.24 ± 2.85**	60.54 ± 3.08
	NFI data (13 classes)	1462	**52.45 ± 2.25**	46.84 ± 2.65
	Francini data (7 classes)	13,790	**98.20 ± 0.20**	95.27 ± 0.60

The MLP classifier consistently outperformed the RF classifier across all datasets when using the same deep features. For the NFI dataset with 7 classes, the MLP achieved an overall accuracy of 77.13% (±1.17) compared to 74.62% (±1.93) for the RF with a difference of approximately 2.5 percentage points. The performance gap was even more pronounced in terms of F1 score, with the MLP achieving 68.24% (±2.85) versus 60.54% (±3.08) for RF, representing a substantial improvement of nearly 8 percentage points. For the more challenging 13-class NFI dataset, the MLP classifier maintained its advantage with an overall accuracy of 67.99% (±1.33) compared to 66.77% (±1.51) for the RF classifier. Similarly, the F1 score showed the MLP's superior performance with 52.45% (±2.25) versus 46.84% (±2.65) for RF. The largest dataset, Francini data with 7 classes, demonstrated the same pattern, with the MLP achieving an outstanding overall accuracy of 98.21% (±0.20) compared to

95.28% (±0.60) for RF. The corresponding F1 scores were 98.20% (±0.20) for theMLPand 95.27% (±0.60) for RF.

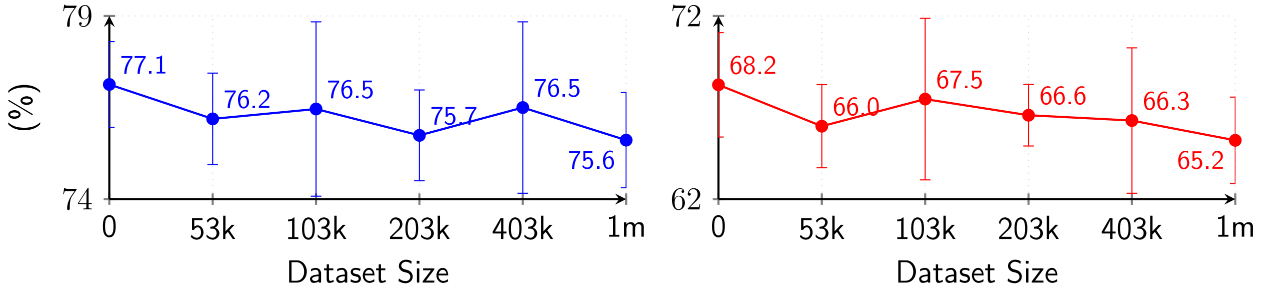
Notably, the MLP classifier not only achieved higher mean performance values but also demonstrated more consistent results across repeated runs, as evidenced by the lower standard deviations in most cases. These results indicate that while both classifiers can effectively leverage deep features, the MLP classifier with pre-trained Presto offers superior performance for forest type classification tasks across datasets of varying complexity and size.

5.0.2 Research Question 2: The effect of the second stage of pre-training and its dataset size

The performance of models incorporating a second stage of pre-training is presented in table 9. The trends in model performance across varying pre-training dataset sizes are visualized in figures 8, 9, and 10.

Table 9: Effect of additional pre-training dataset size on MLP classification accuracy (in %). Each model was run five times and the mean and standard deviation (±) are reported. Bold values indicate the best performance among different dataset sizes per dataset type.

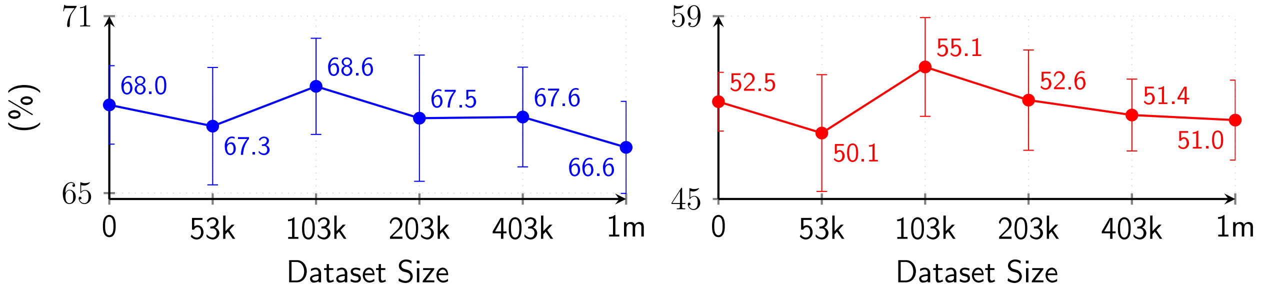
Metric	Dataset Size	MLP		
		NFI data (7 classes)	NFI data (13 classes)	Francini data (7 classes)
Overall Accuracy	0	**77.13 ± 1.17**	67.99 ± 1.33	98.21 ± 0.20
	53,062	76.19 ± 1.25	67.27 ± 1.99	**98.34 ± 0.21**
	103,062	76.46 ± 2.38	**68.62 ± 1.63**	98.24 ± 0.13
	203,062	75.74 ± 1.24	67.54 ± 2.14	98.26 ± 0.34
	403,062	76.50 ± 2.34	67.58 ± 1.69	98.27 ± 0.11
	1,003,062	75.61 ± 1.30	66.55 ± 1.56	98.32 ± 0.16
F1 Score	0	**68.24 ± 2.85**	52.45 ± 2.25	98.20 ± 0.20
	53,062	65.98 ± 2.27	50.05 ± 4.47	**98.34 ± 0.21**
	103,062	67.46 ± 4.41	**55.11 ± 3.78**	98.23 ± 0.13
	203,062	66.58 ± 1.68	52.57 ± 3.84	98.26 ± 0.35
	403,062	66.29 ± 3.97	51.43 ± 2.75	98.27 ± 0.11
	1,003,062	65.21 ± 2.36	51.04 ± 3.06	98.31 ± 0.17



(a) Overall Accuracy

(b) F1 Score

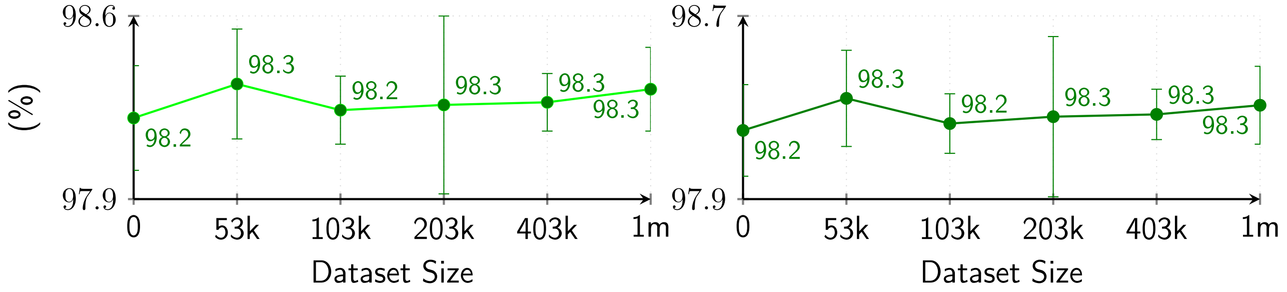
Figure 8: **NFI data (7 classes)** accuracy and F1 score in different dataset sizes. Values are mean of 5 runs and the error bars represent the standard deviation of the results.



(a) Overall Accuracy

(b) F1 Score

Figure 9: **NFI data (13 classes)** accuracy and F1 score in different dataset sizes. Values are mean of 5 runs and the error bars represent the standard deviation of the results.



(a) Overall Accuracy

(b) F1 Score

Figure 10: **Francini data (7 classes)** accuracy and F1 score in different dataset sizes. Values are mean of 5 runs and the error bars represent the standard deviation of the results.

For the NFI dataset with 7 classes, the highest overall accuracy (77.13%) and F1 score (68.24%) were observed when no additional pre-training data was used (0 samples). Increasing the pre-training dataset size to 403,062 samples yielded the second-best accuracy (76.50%), but did not surpass the baseline model, with a large variability of 2.34% standard deviation. The best model with no additional pre-training still has a large standard deviation (1.17% and 2.85% in accuracy and F1 score respectively) sufficient to overlap with other models except the model trained with 1M samples in terms of its score. This suggests that for the 7-class NFI dataset, additional pre-training data did not consistently enhance model performance and might even introduce some level of performance degradation.

In contrast, the NFI dataset with 13 classes (more fine-grained classification) showed different behavior. The best overall accuracy (68.62%) was achieved with 103,062 pre-training samples, representing a modest improvement over the baseline (67.99%). Similarly, the F1 score peaked at 55.11% with the same pre-training dataset size, compared to 52.45% without pre-training. However, each standard deviation is high enough to cover the difference between the best and the baseline model. This indicates that the benefits are not significant even though moderate amounts of additional pre-training data could lead to slight improvement for more complex classification tasks with more classes.

For the Francini dataset with 7 classes, which contained significantly more labeled and class-balanced samples (13,790 compared to approximately 1,480 for the NFI datasets), the effect of pre-training was minimal. The highest overall accuracy (98.34%) was achieved with 53,062 pre-training samples, only marginally higher than the baseline (98.21%). This suggests that when sufficient labeled data is available for the target task, the benefits of additional pre-training become negligible.

Across all datasets, the results indicate that there is no clear linear relationship between pre-training dataset size and classification performance. Rather, each dataset exhibits a different optimal pre-training size, with diminishing or even negative returns beyond certain thresholds. In addition, the standard deviations across repeated runs remain relatively high across all models trained with different dataset sizes, suggesting no significant difference between the models.

6 Discussion

Our results demonstrate that deep features extracted from the pre-trained Presto model more effectively capture spatial and temporal information from satellite time series compared to traditional methods using harmonic and medoid predictors. Furthermore, the MLP classifier with fine-tuned pre-trained Presto significantly enhanced the classification accuracy of tree species in The Netherlands. As shown in Table 13, the integration of S1 and S2 data yielded optimal performance for Dutch NFI classification, while supplementary environmental data from ERA5 and SRTM had minimal impact on classification accuracy. This approach is readily transferable to other countries' NFI systems and is computationally efficient due to Presto's lightweight architecture, which enables fine-tuning on a single GPU or CPU (Tseng et al., 2024). The computational efficiency is evidenced by the Presto pre-training process, which processed 21.5 million pixel time-series in only 2 hours 12 minutes per epoch (43 hours 15 minutes for 20 epochs), suggesting that large-scale pixel-label mapping could be feasible for other countries, including The Netherlands with its 36.4 million 10×10m pixels.

The study also highlights the importance of high-quality and abundant training data for improving tree species classification accuracy. The difference in performance between the NFI and Francini datasets highlights the impact of class imbalance on classification accuracy. The NFI dataset contained varying numbers of samples per class, while the Francini dataset was more balanced. This suggests that addressing class imbalance could further improve classification performance, especially for regions with limited ground reference data, focusing collection of samples for poorly classified species groups than broad data collection efforts. Following the approach of Francini et al., 2024, supplementing NFI data with additional samples through visual interpretation of high-resolution satellite imagery, possibly using tools like Collect Earth Online (Saah et al., 2019), could further enhance model performance.

Regarding the second research question, our results indicate that additional pre-training of Presto on non-labeled data did not significantly improve downstream accuracy in tree species classification. This finding contrasts with observations from Natural Language Processing (Gururangan et al., 2020) and image-based object identification tasks (Ma et al., 2023), where significant improvements were achieved through domain-specific and task-specific second-stage pre-training.

The lack of significant improvement in our pixel-level tree species classification task can be attributed to four key factors:

1. **Limited Data Variability:** Our non-labeled datasets were restricted to forest pixels in The Netherlands with only one year of temporal coverage, potentially constraining the model's ability to learn useful features.

2. **Task Complexity:** Our pixel-level tree species classification task may not have been sufficiently complex to benefit from additional pre-training. The pre-trained Presto model likely already possessed adequate capacity to capture the spatial and temporal information needed to differentiate between 7 or 13 classes.

3. **Methodological Differences:** Variations between our methodology for pixel extraction from GEE and the original Presto implementation may have affected model performance. While our current extraction method using GEE is more accessible than the OpenMapFlow package (Zvonkov et al., 2023), differences in cloud and shadow treatment could have introduced inconsistencies between first and second-stage pre-training data. Such differences can potentially degrade model accuracy, as irrelevant data has been shown to impair performance (Ma et al., 2023).

4. **Pre-training Parameters:** The pre-training settings, including learning rate, weight decay, band selection, and mask ratio for MAE, may not have been optimal. As noted by (Z. Li et al.,

2020), batch size and learning rate requirements can vary significantly with pre-training dataset size.

These findings emphasize the importance of carefully considering pre-training configurations, including dataset extraction methodology, alignment with existing frameworks, and data quality and diversity in pixel-level additional pre-training.

The enhanced classification accuracy achieved through our methodology has significant implications for forest ecosystem monitoring and management. Improved tree species mapping accuracy directly contributes to several critical areas such as biomass estimation, biodiversity assessment, and forest management planning. However, the observed sensitivity to training data quality and quantity highlights a critical consideration: the continued need for investment in high-quality reference datasets for remote sensing applications in forestry. This emphasizes the importance of maintaining and expanding systematic forest inventory programs, particularly in regions where limited ground truth data may constrain the application of advanced classification methods.

7 Conclusion

This study examined the efficacy of deep learning approaches for tree species classification in National Forest Inventories, yielding several significant findings. First, deep features extracted from pre-trained Presto models consistently outperformed traditional methods employing harmonic and medoid predictors across varying dataset sizes and classification complexities. This performance advantage was particularly evident with larger and more balanced training datasets, as demonstrated by the results from the Francini dataset, indicating superior scalability of deep learning approaches with increased high-quality data availability. Second, the MLP classifier demonstrated superior performance compared to RF in our experimental setting, corroborating previous findings (Mouret et al., 2024). Collectively, these results demonstrate the potential for implementing computationally efficient, large-scale tree species mapping within NFI systems.

Our investigation into enhancing model performance through second-stage pre-training with domain-specific data yielded an important insight: additional pre-training did not provide significant improvements over the original pre-trained model. This unexpected finding contrasts with results from other domains and highlights the unique challenges of transfer learning in pixel-level remote sensing applications. The outcome emphasizes that successful transfer learning in forestry applications requires careful consideration of data temporal coverage, pre-training methodology, and the complexity of the target classification task.

This research significantly advances our understanding of deep learning applications in forestry through two key contributions. First, we have demonstrated the effectiveness of fine-tuning pre-trained deep learning models for tree species classification, establishing a computationally efficient approach for improving NFI accuracy. Second, we have identified important limitations in domain-specific pre-training approaches at the pixel level, providing valuable insights for future applications. Our findings underscore that while pre-trained models offer powerful feature extraction capabilities, their successful adaptation to specific forestry applications requires careful consideration of data quality, pre-training strategies, and methodological alignment. These insights lay the groundwork for future developments in automated forest inventory systems and broader applications in environmental monitoring.

8 Use of generative AI statement

This research utilized several artificial intelligence tools to enhance the quality of writing and code implementation. The following AI assistants were employed:

- OpenAI's ChatGPT 3.5
- GitHub Copilot (powered by GPT-4)
- Anthropic's Claude 3.5 Sonnet

These tools were specifically used for:

1. **Writing Enhancement:** Grammar correction, academic writing style improvement, and clarity of expression in the thesis manuscript.
2. **Code Development:**
    - Debugging assistance and error resolution
    - Code optimization and refactoring
    - Generation of unit tests
3. **Document Formatting:** Assistance in converting tabular data to LaTeX format

All AI-assisted content was carefully reviewed and validated for academic appropriateness. The core research methodology, analysis, and conclusions were developed independently of these tools.

9 Acknowledgments

I would like to express my sincere gratitude to all those who supported me in the completion of this MSc thesis. I am particularly indebted to my main supervisor, Dr. Marc Russwurm, for his invaluable guidance and insightful advice throughout this research especially on deep learning domain. Special thanks to my co-supervisor, Dr. Carmelo Bonannella, for his patient mentorship and support especially on forestry domain. Their dedication to securing time for discussions and maintaining an open, positive, and constructive atmosphere was instrumental to this work's success.

I extend my appreciation to Dr. Saverio Francini for generously providing additional datasets and detailed information about his previous research, which significantly contributed to this study. His collaboration enhanced the scope and depth of this investigation.

I am grateful to my colleagues in the GRS group at Wageningen University for creating an inspiring research environment. My heartfelt thanks go to my family and friends, both in Japan and Wageningen, whose unwavering support and encouragement were invaluable throughout this journey.

References

Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B., & Kleinschmit, B. (2022, September). TreeSatAI Benchmark Archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing. https://doi.org/10.5194/essd-2022-312

Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B., & Kleinschmit, B. (2023). *TreeSatAI Benchmark Archive*: A multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data*, *15*(2), 681–695. https://doi.org/10.5194/essd-15-681-2023

Arets, E. J. M. M., Baren, S. A. v., Hendriks, C. M. J., Kramer, H., Lesschen, J. P., & Schelhaas, M. J. (2023). Greenhouse gas reporting of the LULUCF sector in the Netherlands : Methodological background, update 2023. https://doi.org/10.18174/588942

Basu, S., Ganguly, S., Mukhopadhyay, S., DiBiano, R., Karki, M., & Nemani, R. (2015). DeepSat: A learning framework for satellite imagery. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–10. https://doi.org/10.1145/2820783.2820816

Blickensdörfer, L., Oehmichen, K., Pflugmacher, D., Kleinschmit, B., & Hostert, P. (2024). National tree species mapping using Sentinel-1/2 time series and German National Forest Inventory data. *Remote Sensing of Environment*, *304*, 114069. https://doi.org/10.1016/j.rse.2024.114069

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022, July). On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258

Bonannella, C. (2024). Spatiotemporal modeling of vegetation dynamics in a changing environment: Combining earth observation and machine learning. https://doi.org/10.18174/655208

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. *Scientific Data*, *9*(1), 251. https://doi.org/10.1038/s41597-022-01307-4

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, June). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

FAO. (2020). *Global Forest Resources Assessment 2020: Main report*. Retrieved October 1, 2024, from https://doi.org/10.4060/ca9825en

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., & Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics*, *45*(2). https://doi.org/10.1029/2005RG000183

Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., Straub, C., & Ghosh, A. (2016). Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*, *186*, 64–87. https://doi.org/10.1016/j.rse.2016.08.013

Francini, S., Schelhaas, M.-J., Vangi, E., Lerink, B., Nabuurs, G.-J., McRoberts, R. E., & Chirici, G. (2024). Forest species mapping and area proportion estimation combining Sentinel-2 harmonic predictors and national forest inventory data. *International Journal of Applied Earth Observation and Geoinformation*, *131*, 103935. https://doi.org/10.1016/j.jag.2024.103935

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In D. E. Losada & J. M. Fernández-Luna (Eds.), *Advances in Information Retrieval* (pp. 345–359). Springer. https://doi.org/10.1007/978-3-540-31865-1_25

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021, November). Masked Autoencoders Are Scalable Vision Learners. Retrieved October 7, 2024, from https://arxiv.org/abs/2111.06377v3

Heaton, J. (2016). An empirical analysis of feature engineering for predictive modeling. *SoutheastCon 2016*, 1–6. https://doi.org/10.1109/SECON.2016.7506650

Hermosilla, T., Bastyr, A., Coops, N. C., White, J. C., & Wulder, M. A. (2022). Mapping the presence and distribution of tree species in Canada's forested ecosystems. *Remote Sensing of Environment*, *282*, 113276. https://doi.org/10.1016/j.rse.2022.113276

Jactel, H., Bauhus, J., Boberg, J., Bonal, D., Castagneyrol, B., Gardiner, B., Gonzalez-Olabarria, J. R., Koricheva, J., Meurisse, N., & Brockerhoff, E. G. (2017). Tree Diversity Drives Forest Stand Resistance to Natural Disturbances. *Current Forestry Reports*, *3*(3), 223–243. https://doi.org/10.1007/s40725-017-0064-1

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: An ASA Data Science Journal*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Kang, Q., Chen, X., Li, S., & Zhou, M. (2017). A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification. *IEEE Transactions on Cybernetics*, *47*(12), 4263–4274. https://doi.org/10.1109/TCYB.2016.2606104

Karasiak, N., Dejoux, J.-F., Monteil, C., & Sheeren, D. (2022). Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, *111*(7), 2715–2740. https://doi.org/10.1007/s10994-021-05972-1

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017, February). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. https://doi.org/10.48550/arXiv.1609.04836

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102926. https://doi.org/10.1016/j.jag.2022.102926

Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D., & Gonzalez, J. E. (2020). Train large, then compress: Rethinking model size for efficient training and inference of transformers. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 5958–5968.

Lu, S., Guo, J., Zimmer-Dauphinee, J. R., Nieusma, J. M., Wang, X., VanValkenburgh, P., Wernke, S. A., & Huo, Y. (2024, August). AI Foundation Models in Remote Sensing: A Survey. https://doi.org/10.48550/arXiv.2408.03464

Ma, H., Li, X., Yuan, X., & Zhao, C. (2023). Two-phase self-supervised pretraining for object re-identification. *Knowledge-Based Systems*, *261*, 110220. https://doi.org/10.1016/j.knosys.2022.110220

Mouret, F., Morin, D., Planells, M., & Vincent-Barbaroux, C. (2024, November). Tree species classification at the pixel-level using deep learning and multispectral time series in an imbalanced context. https://doi.org/10.48550/arXiv.2408.08887

Muñoz Sabater, J. (2024, September). ERA5-Land monthly averaged data from 1950 to present. https://doi.org/doi:10.24381/cds.68d2bb30

Pasquarella, V. J., Brown, C. F., Czerwinski, W., & Rucklidge, W. J. (2023). Comprehensive Quality Assessment of Optical Satellite Imagery Using Weakly Supervised Video Learning, 2125–2135. Retrieved October 9, 2024, from https://openaccess.thecvf.com/content/CVPR2023W/EarthVision/html/Pasquarella_Comprehensive_Quality_Assessment_of_Optical_Satellite_Imagery_Using_Weakly_Supervised_CVPRW_2023_paper.html

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules (Basel, Switzerland)*, *26*(4), 1111. https://doi.org/10.3390/molecules26041111

Saah, D., Johnson, G., Ashmall, B., Tondapu, G., Tenneson, K., Patterson, M., Poortinga, A., Markert, K., Quyen, N. H., San Aung, K., Schlichting, L., Matin, M., Uddin, K., Aryal, R. R., Dilger, J., Lee Ellenburg, W., Flores-Anderson, A. I., Wiell, D., Lindquist, E., . . . Chishtie, F. (2019). Collect Earth: An online tool for systematic reference data collection in land cover and use applications. *Environmental Modelling & Software*, *118*, 166–171. https://doi.org/10.1016/j.envsoft.2019.05.004

Schelhaas, M., Clerkx, A. P. P. M., Daamen, W. P., Oldenburger, J. F., Velema, G., Schnitger, P., Schoonderwoerd, H., & Kramer, H. (2014). *Zesde Nederlandse bosinventarisatie : Methoden en basisresultaten* (tech. rep. No. 2545). Alterra. Wageningen. Retrieved September 12, 2024, from https://library.wur.nl/WebQuery/wurpubs/454875

Schelhaas, M. J., Teeuwen, S., Oldenburger, J., Beerkens, G., Velema, G., Kremers, J., Lerink, B., Paulo, M. J., Schoonderwoerd, H., Daamen, W., Dolstra, F., Lusink, M., Tongeren, K. v., Scholten, T., Pruijsten, I., Voncken, F., & Clerkx, A. P. P. M. (2022). Zevende Nederlandse Bosinventarisatie: Methoden en resultaten. https://doi.org/10.18174/571720

Tomppo, E., Gschwantner, T., Lawrence, M., & McRoberts, R. E. (Eds.). (2010). *National Forest Inventories: Pathways for Common Reporting*. Springer Netherlands. https://doi.org/10.1007/978-90-481-3233-1

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., & Kerner, H. (2024, February). Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. https://doi.org/10.48550/arXiv.2304.14065

UNFCCC. (2015). Paris agreement. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

Vaessen, N., & Leeuwen, D. A. v. (2024, February). The Effect of Batch Size on Contrastive Self-Supervised Speech Representation Learning. https://doi.org/10.48550/arXiv.2402.13723

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022). Self-Supervised Learning in Remote Sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 213–247. https://doi.org/10.1109/MGRS.2022.3198244

Yu, A., Liu, B., Cao, X., Qiu, C., Guo, W., & Quan, Y. (2022). Pixel-Level Self-Supervised Learning for Semi-Supervised Building Extraction From Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2022.3207465

Zvonkov, I., Tseng, G., Nakalembe, C., & Kerner, H. (2023). OpenMapFlow: A Library for Rapid Map Creation with Machine Learning and Remote Sensing Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 14655–14663. https://doi.org/10.1609/aaai.v37i12.26713

Appendix

10 Appendix

10.1 Content of the zip file

Table of Content of the zip file that accompanies the thesis report.

- Documentation of what is where in the file (including folder structure; Word, PDF)
- Report (Word, PDF)
- Midterm & Final presentation (PPTX)
- Datasets used and created
- Figures/Maps/Tables
- Scripts /code/exe
- Literature (PDFs of used articles / preferable Endnote)

10.2 Additional results

10.2.1 Results of models using only S1 and S2 data for fine-tuning

Our primary models incorporated all available satellite data as input for both the second stage of pre-training and fine-tuning to extract deep features. In this section, we present results from models using only S1 and S2 data as input for fine-tuning and feature extraction, excluding SRTM and ERA5 data.

Research Question 1: Deep features model performance

Table 10 demonstrates that deep features extracted by pre-trained Presto from only S1 and S2 data outperformed harmonic and medoid features in the RF classifier. These results were comparable to models that utilized all satellite data as input.

Table 10: Classification performance comparison between deep features and harmonic+medoid features (in %). Each model was run five times and the mean and standard deviation (±) are reported. Bold values indicate the best performance between deep features extracted from Presto and harmonic+medoid features per dataset type.

			Random Forest	
	Dataset type	Data size	Deep feat. (Tseng et al., 2024)	Harm.+med. (Francini et al., 2024)
Overall Accuracy	NFI data (7 classes)	1480	**75.47 ± 1.15**	70.58 ± 1.78
	NFI data (13 classes)	1462	**67.77 ± 2.38**	62.71 ± 1.51
	Francini data (7 classes)	13,790	**94.96 ± 0.58**	84.31 ± 0.99
F1 Score	NFI data (7 classes)	1480	**63.83 ± 1.21**	51.57 ± 3.36
	NFI data (13 classes)	1462	**48.24 ± 4.31**	38.54 ± 2.14
	Francini data (7 classes)	13,790	**94.94 ± 0.58**	84.26 ± 0.99

Table 11 shows that the MLP classifier consistently achieved higher accuracy than the RF classifier, even when using only S1 and S2 data as input for the pre-trained Presto model.

Table 11: Classification performance comparison between MLP and RF classifiers with deep features (in %). Each model was run five times and the mean and standard deviation (±) are reported. Bold values indicate the best performance between MLP and RF classifier per dataset type.

			Results	
	Dataset type	Data size	MLP classifier	RF classifier
Overall Accuracy	NFI data (7 classes)	1480	**78.79 ± 1.01**	75.47 ± 1.15
	NFI data (13 classes)	1462	**69.39 ± 1.65**	67.77 ± 2.38
	Francini data (7 classes)	13,790	**98.08 ± 0.09**	94.96 ± 0.58
F1 Score	NFI data (7 classes)	1480	**70.57 ± 2.20**	63.83 ± 1.21
	NFI data (13 classes)	1462	**54.91 ± 4.29**	48.24 ± 4.31
	Francini data (7 classes)	13,790	**98.07 ± 0.09**	94.94 ± 0.58

Research Question 2: Effect of second-stage pre-training and dataset size
As shown in Table 12, no significant improvement was observed after additional pre-training with S1 and S2 data. These results align with those from models that used all satellite data as input.

Table 12: Effect of additional pre-training dataset size on MLP classification accuracy (in %) in S1 and S2 inputs. Each model was run five times and the mean and standard deviation (±) are reported. Bold values indicate the best performance among different dataset sizes per dataset type.

Metric	Dataset Size	MLP		
		NFI data (7 classes)	NFI data (13 classes)	Francini data (7 classes)
Overall Accuracy	0	**78.79 ± 1.01**	**69.39 ± 1.65**	98.08 ± 0.09
	53,062	76.46 ± 1.27	67.86 ± 0.82	98.14 ± 0.11
	103,062	76.68 ± 0.67	67.58 ± 1.25	98.06 ± 0.21
	203,062	77.44 ± 1.88	69.12 ± 1.48	98.06 ± 0.20
	403,062	77.98 ± 1.60	68.94 ± 1.42	98.21 ± 0.37
	1,003,062	77.13 ± 0.39	66.82 ± 0.89	**98.24 ± 0.23**
F1 Score	0	**70.57 ± 2.20**	**54.91 ± 4.29**	98.07 ± 0.09
	53,062	68.59 ± 1.40	52.15 ± 2.54	98.14 ± 0.11
	103,062	68.39 ± 1.01	52.07 ± 3.67	98.06 ± 0.21
	203,062	69.49 ± 3.31	53.99 ± 2.46	98.06 ± 0.20
	403,062	69.41 ± 2.49	53.93 ± 3.78	98.20 ± 0.37
	1,003,062	69.08 ± 2.89	50.62 ± 2.35	**98.24 ± 0.23**

Comparison between all bands and only S1 and S2 bands
Table 13 compares models using all satellite data against those using only S1 and S2 data with MLP

classifier. Interestingly, models with only S1 and S2 data achieved higher accuracy with NFI data, while no significant difference was observed with Francini data.

Table 13: Classification performance comparison between using all available bands versus using only S1 and S2 bands (in %). Each model was run five times with mean and standard deviation (±) reported. Bold values indicate best performance between the two band configurations for each dataset type.

	Dataset type	Data size	MLP classifier	
			All bands	S1S2 only
Overall Accuracy	NFI data (7 classes)	1480	77.13 ± 1.17	**78.79 ± 1.01**
	NFI data (13 classes)	1462	67.99 ± 1.33	**69.39 ± 1.65**
	Francini data (7 classes)	13,790	**98.21 ± 0.20**	98.08 ± 0.09
F1 Score	NFI data (7 classes)	1480	68.24 ± 2.85	**70.57 ± 2.20**
	NFI data (13 classes)	1462	52.45 ± 2.25	**54.91 ± 4.29**
	Francini data (7 classes)	13,790	**98.20 ± 0.20**	98.07 ± 0.09