Chair of Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich

Tree Species Classification

Data Science in Earth Observation

Pei-Ling Song (03798081)

Hongyu Jiang (03804823)

Meng-Ju Hsieh (03797997)

Hoi-Wang Lo (03797896)

Kit-Lung Chan (03797955)

Project Report for 25SOSE Data Science in Earth Observation Course

Master of Science

at the School of Engineering and Design

of the Technical University of Munich

Supervised by

Prof. Dr. Xiaoxiang Zhu

Dr. Muhammad Shahzad

Dr. Andrés Camero Unzueta

Mr. Yang Mu

Submitted on

July 21, 2025

Abstract

The aim of this project is to classify dominant tree species across Germany by leveraging satellite image time series. Forest management practices benefit from knowledge of species distribution, which supports site-adapted tree selection and climate-resilient forest planning. Tree species labels in the TreeSatAI dataset are organised hierarchically (leaf type, genus, species), allowing us to develop classification methods that exploit this structure to improve model performance. We integrate Sentinel-2 multi-spectral imagery (10 spectral bands, 5 vegetation indices) into a data processing pipeline that extracts monthly composites (March–October 2022) and generates 5×5 pixel patches for each reference point. We trained and evaluated five classification models: Random Forest, XGBoost, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer architectures.Hierarchical labels (L1, L2) significantly improved accuracy both traditional machine learning and deep learning models(CNN, RNN). Furthermore, data augmentation with rotation and translation enhanced the generalisation ability of deep learning models. Among all methods, XGBoost achieved the highest test accuracy (83.6%)

Keywords:

Tree Species Classification, Sentinel-2, Remote Sensing, Machine Learning, Deep Learning, Random Forest, XGBoost, CNN, RNN, Transformer, TreeSatAI, Earth Observation

Contribution

Pei Ling Song: Exploratory Data Analysis, Random Forest Code, Data curation, Methodology(feature combination and time selection),Hierarchy analysis, Validation and Visualization, Writing – review & editing.

Hongyu Jiang: XGBoost Code, Data Curation, Methodology(feature combination and time selection), Hierarchy analysis, Validation and Visualization, Writing – review & editing.

Meng Ju Hsieh: CNN Code, Data Curation, Methodology(Spatial & Spectral Data Augmentation, Hierarchical Data Enhancement) , Validation and Visualization, Writing – review & editing.

Hoi Wang Lo: RNN code, Methodology(time and patches), Data extraction, and Visualization, Writing – review & editing.

Kit Lung Chan: Data Acquisition, Transformer-based Method Development, Optimization approaches, Validation(test), Conclusion, Writing – review & editing.

Contents

Introduction & Exploratory Data Analysis (EDA)....1

Baseline Establishment....1

Sample Size and Missing Data Analysis....1

Random Forest....2

Baseline Establishment....2

Model Optimization and Tuning....2

Feature Engineering Comparison....2

Performance Evaluation....2

XGBoost....3

Data processing....3

Model Training and Evaluation....3

Evaluation on Independent Test Dataset....3

Optimization....4

CNN (Convolutional Neural Network)....4

Baseline Establishment....5

Model Optimization and Tuning....5

....6

Performance Evaluation....6

RNN (Recurrent Neural Network)....7

Baseline Establishment....7

Data Augmentation....7

Model Optimization and Tuning....7

Performance Evaluation....8

Transformer-based Method....9

Data Acquisition and Preparation....9

Data Preprocessing and Augmentation....9

Model Architecture....9

Model Optimization Strategy....9

Network Optimization and Training Result:....10

Conclusion....11

Crossing Comparison of Different Approaches....11

Bibliography....12

List of Figures

Figure 1: EDA - Tree Species Sample Distribution in Germany....1
Figure 2: EDA - Tree Species Sample Size and Missing Data Percentage 1
Figure 3: EDA - Tree Species Sample Distribution in Germany....1
Figure 4: Random Forest - Confusion Matrix of Rnadom Forest....2
Figure 5: XGBoost - Training History....3
Figure 6: XGBoost - Confusion Matrix of Training Model....3
Figure 7: CNN – 3D CNN Structure Design....5
Figure 8: CNN - Training Accuracy of the Hierarchical Model 6
Figure 9: CNN - Species Classification Accuracy of the Hierarchical Model 6
Figure 10: CNN - Relationship Between Sample Size and Accuracy....6
Figure 11: RNN - Model Performance against the parameter efficiency 7
Figure 12: RNN - Training and Validation Accuracy plot of GRU-5p....7
Figure 13: RNN - Class Accuracy of GRU-12m on Test Set....8
Figure 14: RNN - Class Accuracy of GRU-hier on Test Set 8
Figure 15: Transformer - Based Method 9

List of Tables

Table 1: Random Forest - Model Performance Across Configurations 2
Table 2: XGBoost - Traing_Model Performance Across Configuration 4
Table 3: XGBoost - Test_Model Performance Across Configurations 4
Table 4: XGBoost - 4 Best Months Combination Accuracy....4
Table 5: CNN - Data Validity Distribution 5
Table 6: CNN - Model Performance Across Training Configurations....5
Table 7: RNN - model results....7
Table 8: Transformer - based- Network Tuning and Training Result 10
Table 9: Conclusion – Accuracy Comaprison of Different Model....11

Introduction & Exploratory Data Analysis (EDA)

Baseline Establishment

The TreeSatAI dataset contains 37,907 labeled samples across northern Germany, with coordinates referenced in EPSG:25832 (UTM zone 32N). Each sample includes hierarchical labels at three levels: L1 (leaf type), L2 (genus), and L3 (species), as well as multi-temporal Sentinel-2 features. The spectral information consists of 10 bands and 5 vegetation indices, collected monthly during the growing season (March–October), resulting in a high-dimensional, multi-temporal feature space.

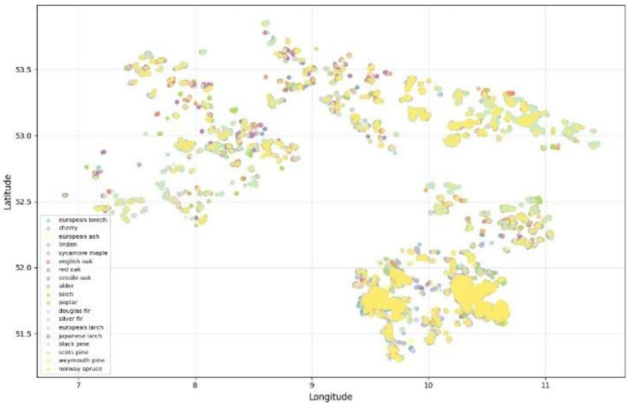


Figure 1: EDA - Tree Species Sample Distribution in Germany

Sample Size and Missing Data Analysis

Sample counts vary significantly between species, with Scots Pine having the highest representation (n=5,389) and Linden. the lowest (n=161), resulting in a class imbalance ratio of approximately 33.5:1. At the L1 level, broadleaf samples (n=20,843) slightly outnumber needleleaf samples (n=17,064, while at the L2 level, Oak and Pine dominate.

Temporally, the missing data is concentrated in April (14.5%), July (9.3%), and October (10.0%). This pattern likely reflects seasonal variations in cloud cover and imagery quality. Overall, missing data averaged below 5% per species, but challenges from class imbalance and incompleteness require attention in later preprocessing and modelling.

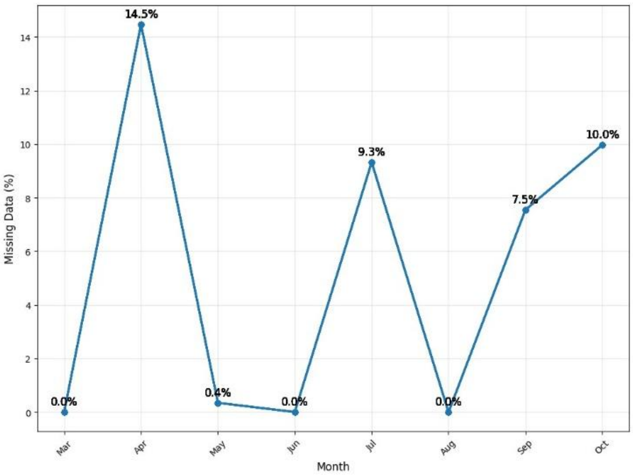


Figure 2: EDA - Missing Data Percentage by Month

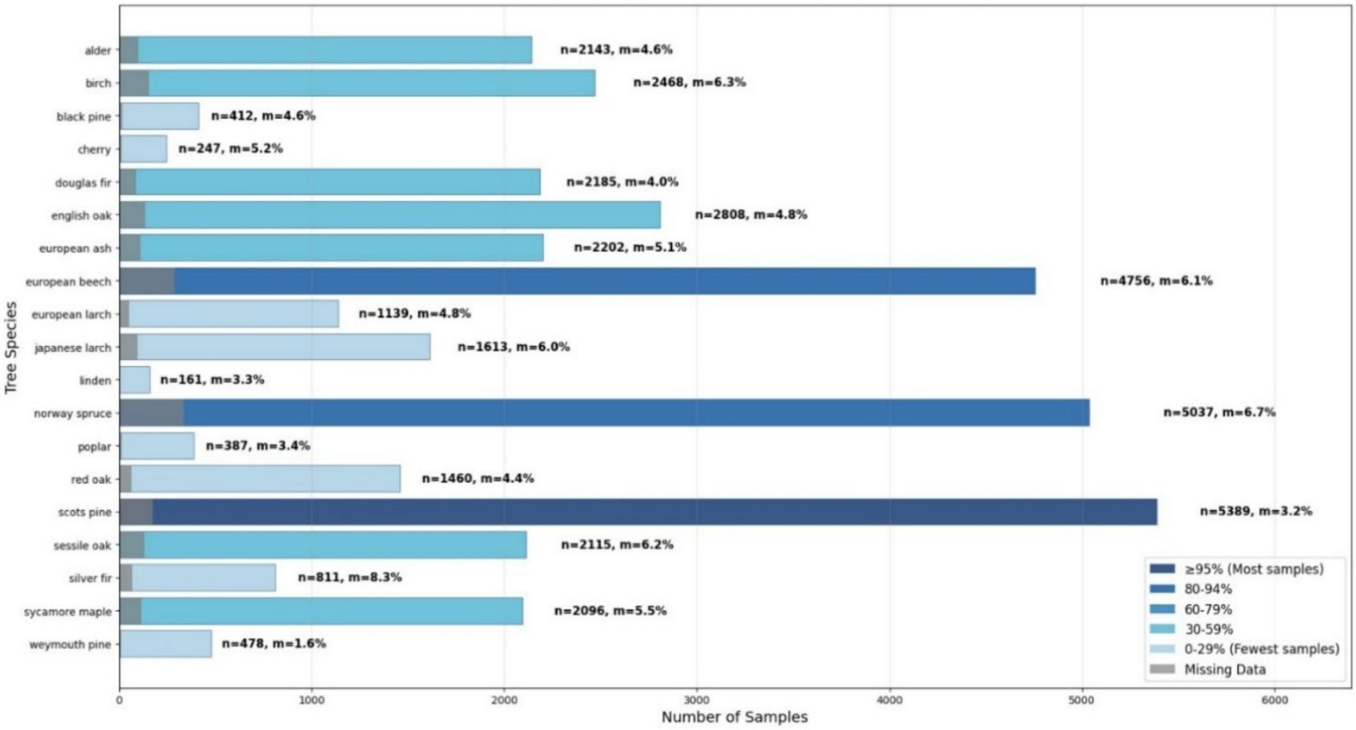


Figure 3: EDA - Tree Species Sample Size and Missing Data Percentage

Random Forest

Baseline Establishment

Architecture and Input Design

The baseline model uses Random Forest (RF) as the classifier. Input data are structured as tensors with dimensions (15, 8, 5, 5), representing 15 spectral features (B2–B12, NDVI), 8 temporal steps (monthly samples), and 5×5 spatial patches. All patches are flattened into 25-pixel vectors and concatenated across time and spectral dimensions to form a high-dimensional feature matrix.

Preprocessing and Baseline Performance Analysis

Each spectral and vegetation index column was flattened to handle missing values or anomalous entries, replacing them with zero vectors when necessary. The resulting flattened vectors were concatenated to form a high-dimensional feature matrix with approximately 3000 features. The baseline Random Forest model achieved an accuracy of 63.55% and a Macro F1 score of 0.5086 on the test dataset.

Model Optimization and Tuning

Parameter Tuning

Grid search was used for hyperparameter tuning. Key adjustments included increasing the number of trees to 500 for stability and setting both the minimum samples per split and minimum leaf samples to 2, enabling the model to better learn minority class patterns under class imbalance. After optimization, the model's accuracy rising to 65.93% and Macro F1 increasing to 0.5986.

Configuration	Features	Acc (%)	Marco F1	Weighted F1
Baseline	3000	63.55	0.5086	0.6220
Parameter Tuning	3000	65.93	0.5986	0.65
Only Spectral Bands	2000	65.01	0.5894	0.6386
Spectral + Vegetation Indices	3000	65.80	0.5928	0.6472
Spectral + Vegetation + L1, L2	3002	72.72	0.6536	0.7185
Reduced Time Series + L1, L2	1502	73.70	0.6396	0.7270

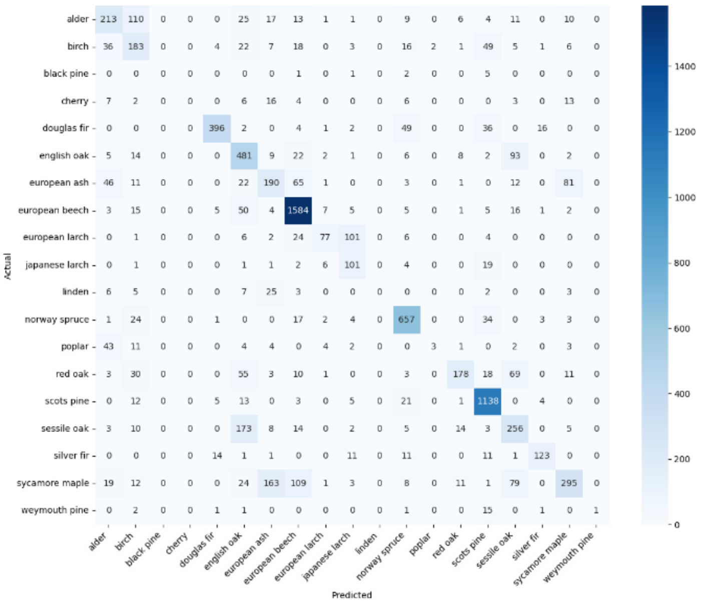
Table 1: Random Forest - Model Performance Across Configurations

Feature Engineering Comparison

The evaluation of four feature configurations revealed that adding vegetation indices to spectral bands slightly improved accuracy from 65.01% to 65.80%. Incorporating hierarchical labels (L1: leaf type, L2: genus) in the Spectral + Vegetation + L1, L2 setup significantly enhanced performance, achieving 72.72% accuracy and a Macro F1 of 0.6536. The Reduced Time Series + L1, L2 configuration showed the highest accuracy (73.70%) and weighted F1 (0.7270), despite reducing feature dimensions by half. However, its Macro F1 was slightly lower (0.6396), indicating less balanced performance across classes.

Performance Evaluation

The final Random Forest model achieved an overall accuracy of 70.36% and a Macro F1 score of 0.4758 on the test dataset. The classification report indicates strong performance for dominant classes such as European beech (precision: 0.84, recall: 0.93, F1: 0.88) and Scots pine (precision: 0.85, recall: 0.95, F1: 0.89), reflecting the model's ability to correctly identify frequent species. However, minority classes like black pine, cherry, and linden showed poor or zero recall and F1 scores, highlighting persistent challenges in classifying underrepresented species. The weighted average F1 score of 0.69 demonstrates that the model performs well when accounting for class distribution, but the lower Macro F1 indicates imbalances in predictive performance across all classes.



	alder	birch	black pine	cherry	douglas fir	english oak	european ash	european beech	european larch	japanese larch	linden	norway spruce	poplar	red oak	scots pine	sessile oak	silver fir	sycamore maple	weymouth pine
alder	213	110	0	0	0	25	17	13	1	1	0	9	0	6	4	11	0	10	0
birch	36	183	0	0	4	22	7	18	0	3	0	16	2	1	49	5	1	6	0
black pine	0	0	0	0	0	0	0	1	0	1	0	2	0	0	5	0	0	0	0
cherry	7	2	0	0	0	6	16	4	0	0	0	6	0	0	0	3	0	13	0
douglas fir	0	0	0	0	396	2	0	4	1	2	0	49	0	0	36	0	16	0	0
english oak	5	14	0	0	0	481	9	22	2	1	0	6	0	8	2	93	0	2	0
european ash	46	11	0	0	0	22	190	65	1	0	0	3	0	1	0	12	0	81	0
european beech	3	15	0	0	5	50	4	1584	7	5	0	5	0	1	5	16	1	2	0
european larch	0	1	0	0	0	6	2	24	77	101	0	6	0	0	4	0	0	0	0
japanese larch	0	1	0	0	0	1	1	2	6	101	0	4	0	0	19	0	0	0	0
linden	6	5	0	0	0	7	25	3	0	0	0	0	0	0	2	0	0	3	0
norway spruce	1	24	0	0	1	0	0	17	2	4	0	657	0	0	34	0	3	3	0
poplar	43	11	0	0	0	4	4	0	4	2	0	0	3	1	0	2	0	3	0
red oak	3	30	0	0	0	55	3	10	1	0	0	3	0	178	18	69	0	11	0
scots pine	0	12	0	0	5	13	0	3	0	5	0	21	0	1	1138	0	4	0	0
sessile oak	3	10	0	0	0	173	8	14	0	2	0	5	0	14	3	256	0	5	0
silver fir	0	0	0	0	14	1	1	0	0	11	0	11	0	0	11	1	123	0	0
sycamore maple	19	12	0	0	0	24	163	109	1	3	0	8	0	11	1	79	0	295	0
weymouth pine	0	2	0	0	1	1	0	0	0	0	0	1	0	0	15	0	1	0	1

Figure 4: Random Forest - Confusion Matrix

Conclusion

Crossing Comparison of Different Approaches

The background of this project is to classify 19 dominant tree species for forests in Germany with different machine learning method, using Sentinel-2 multi-spectral imagery. With the provided training coordinate and tree information, we have obtained the data of Sentinel-2 multi-spectral imagery (15 bands/indexes over 8 months period).

Data were analyzed but we have spotted the data imbalance issue. Some species have over 5,000 samples, but some are less than 500. Such massive difference creates a huge challenge to the model training. As those minorities have insufficient training samples, the trained model shall have bias toward major groups. These minor groups are likely to be misclassified/missed. Hence affect the overall accuracy and individual class accuracy. So, for five well-known machine learning methods, different optimization strategies have been introduced in each method to boost the performance, and their result shown as below.

Model	Best optimization strategy	Train Accuracy	Test Accuracy
Random Forest	Hierarchy (L1, L2)	72.7%	70.4%
XGBoost	Hierarchy (L1, L2)	87.9%	83.6%
CNN	Rotation Shift + Hierarchy (L1, L2)	89.1%	80.3%
RNN	Temporal (12 months) + Hierarchy	88.7%	81.2%
Transformer	Temporal (Selected)+ Rotation-Translation	73.5%	69.3%

Table 9: Conclusion – Accuracy Comaprison of Different Model

For the overall accuracy, the table above reveals that XGBoost, CNN and RNN could archive a relatively high accuracy (~80%) with Hierarchical method applied. For CNN extra Rotation Shift has applied to improve the accuracy. For RNN, 12 months data have been utilized as well. For Random Forest, as the fundemental method, archive 70% accuracy after Hierarchical approach is applied. For Transformer method, only 1-2 percent is improved even SMOTE and Hierarchical approach were tried. This is due to the attention-based algorithm design received less impact from these methods.

For the accuracy of minor species, Transformer-based method (with Rotation-Translation to increase sample), on the other hand, archived the best result. This is due to the attention-based algorithm design as well. Other methods seem to have higher bias toward the major species, hence a low accuracy of minor species.

Future Research

Each model in this study demonstrated unique strengths: XGBoost showed stable generalization, CNN captured spatial features well, RNN handled temporal patterns effectively, and the Transformer was especially promising for minority class recognition. Since no single method can fully address class imbalance and data complexity, we propose adopting **ensemble learning** in future work. By combining multiple models through weighted voting or stacking, it is possible to enhance both overall and minority-class accuracy. This strategy balances performance across classes and leads to a more robust and practical tree species classification system.

Overall, we tested various machine learning method and optimization approaches to improve overall and individual class accuracy. Our results show partial success, despite significant class imbalance in the dataset.

Bibliography

[1] M. O. Turkoglu, S. Kaya, E. Sertel, and U. Alganci, "Crop mapping from image time series: Deep learning with multiscale label hierarchies," *Remote Sensing of Environment*, vol. 264, p. 112603, 2021.

[2] acamero, *Data Science EO - Classification*, GitHub. [Online]. Available: https://github.com/acamero/data-science-eo-classification

[3] acamero, *Data Science EO - Regression*, GitHub. [Online]. Available: https://github.com/acamero/data-science-eo-regression

[4] acamero, *Data Science EO - Segmentation*, GitHub. [Online]. Available: https://github.com/acamero/data-science-eo-segmentation

[5] H. K. J. Kuo, E. Arisoy, A. Emami, and P. Vozila, "Large scale hierarchical neural network language models," in *Proc. INTERSPEECH*, 2012.

[6] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proc. AISTATS*, 2005, pp. 246–252.

[7] B. Zhuang, J. Liu, Z. Pan, H. He, Y. Weng, and C. Shen, "A survey on efficient training of transformers," *arXiv preprint arXiv:2305.00047*, 2023. [Online]. Available: https://arxiv.org/abs/2305.00047