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

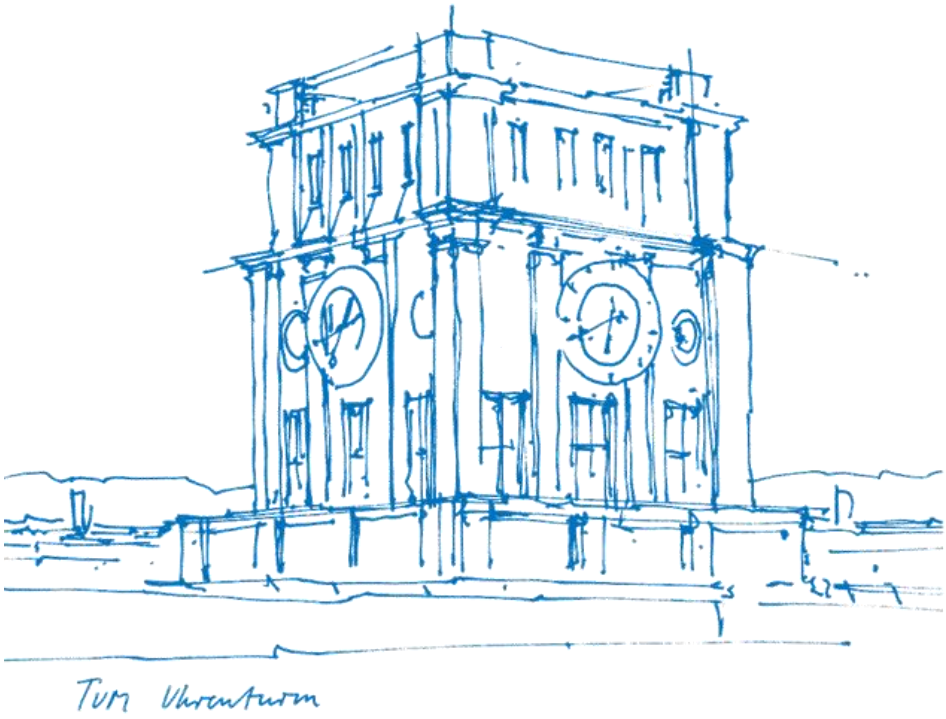
Pei-Ling Song (03798081)
Hongyu Jiang (03804823)
Meng-Ju Hsieh (03797997)
Hoi-Wang Lo (03797896)
Kit-Lung Chan (03797955)

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

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.

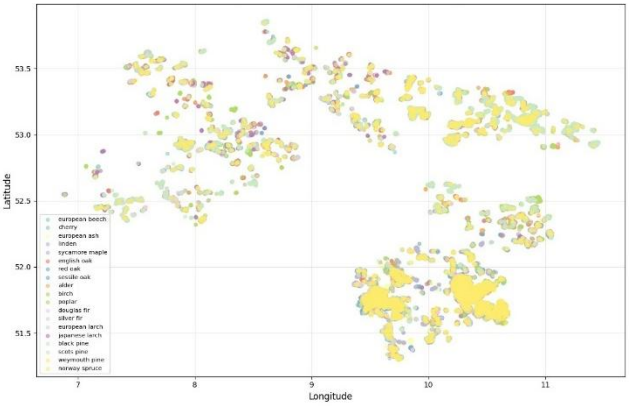


Figure 1: EDA - Tree Species Sample Distribution in Germany

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.

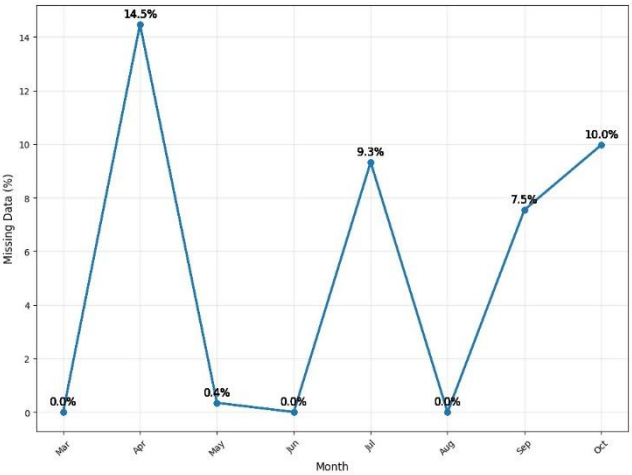


Figure 2: EDA - Missing Data Percentage by Month

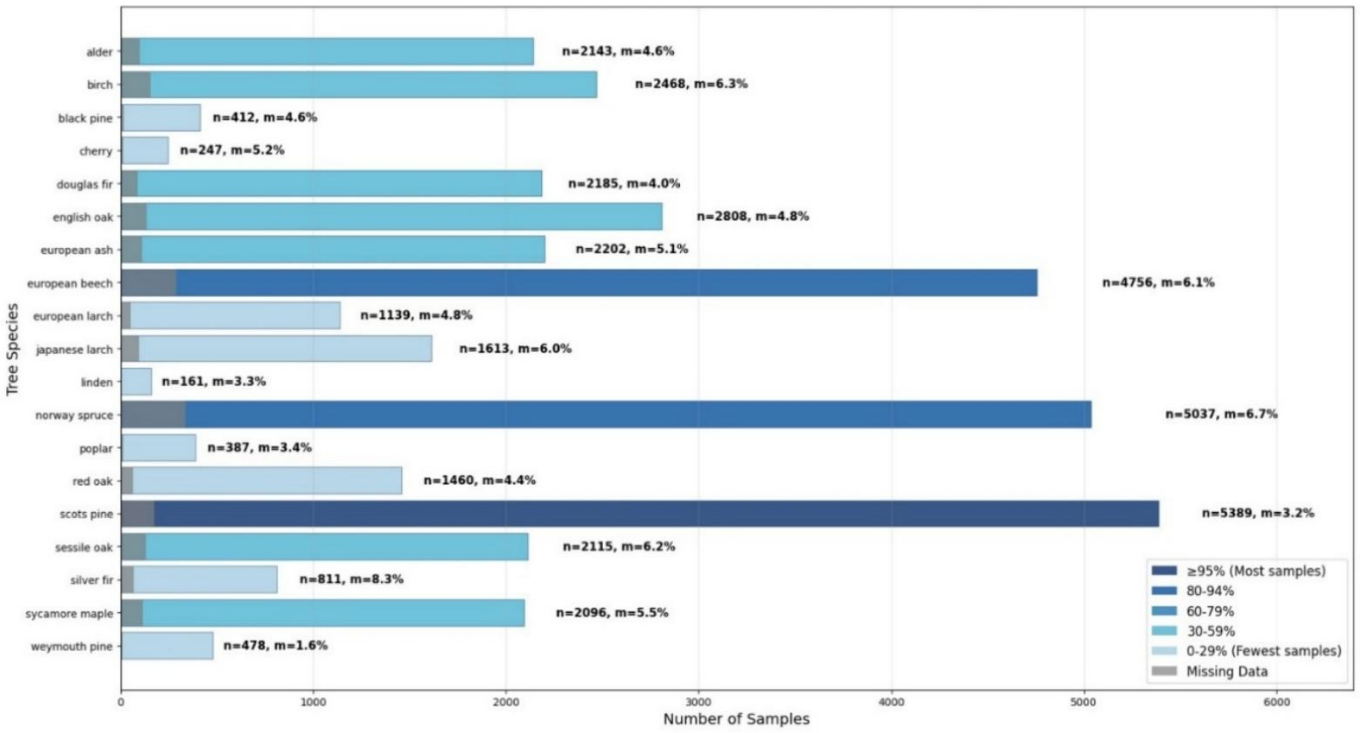


Figure 3: EDA - Tree Species Sample Size and Missing Data Percentage

CNN (Convolutional Neural Network)

Baseline Establishment

Architecture and Input Tensor Design

This study adopts a three-layer 3D CNN architecture. The input tensor has a dimension of (15, 8, 5, 5), corresponding to 15 spectral features (B2–B12 + NDVI), 8 temporal steps, and 5×5 spatial slices. The number of filters is 32, 64, and 128, respectively, with BatchNorm, MaxPooling, and Dropout (0.25/0.25/0.5). The Adam optimizer is used (lr = 0.001), with a batch size of 32.

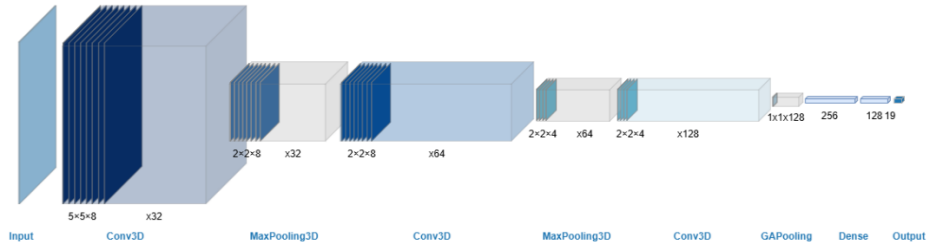


Figure 7: CNN - 3D CNN Structure Design

Validity	Samples	Percentage
0-20%	0	0.0%
20-40%	0	0.0%
40-60%	28	0.1%
60-80%	3,639	9.6%
80-100%	**34,240**	**90.3%**

Table 5: CNN - Data Validity Distribution

Design Dataset Preprocessing and Baseline Analysis

The data completeness analysis shows that 90.3% of samples have a fill rate over 80%, so 34,240 samples are retained, and filled the rest with 0. After Z-score normalization, stratified sampling (70/15/15) and balanced class weighting are applied. The test accuracy reaches 68.09%, and the model converges steadily at around 67 epochs. Another group using all 37,907 samples with zero-filling achieves a test accuracy of 66.84%, and the validation gap increases, showing that incomplete samples weaken generalization ability. Therefore, the 34k dataset is used as the baseline for subsequent experiments.

Model Optimization and Tuning

Basic Parameter Tuning

Phase	Configurations	Test Acc (%)	Improvement(%)	Overfitting Gap	Epochs	Time(min)
Baseline	37k_70_15_15	66.84	-1.25%	-0.0154	96	20.6
Baseline	34k_70_15_15	68.09	-	-0.0084	67	11.6
Data Split	80_10_10	69.07	0.98%	0.0085	80	19.6
Learning Rate	0.0015	69.25	1.16%	-0.0115	86	22.8
Regularization	dr_light	71.26	3.17%	0.0464	65	19.1
Regularization	dr_heavy	66.12	-2.00%	-0.0357	87	39.1
Spatial Aug	flip_vertical	70.01	1.90%	0.0413	63	34.9
Spatial Aug	flip_horizontal	70.30	2.20%	0.0423	61	23.7
Spatial Aug	Rotation_90	71.99	3.90%	0.0105	66	41.4
Spatial Aug	rotation_zoom	71.70	3.60%	0.0007	65	69.6
Spatial Aug	rotation_shift	**73.31**	5.22%	0.0063	100	160.4
Spectral Aug	SpecNoise	72.34	4.30%	0.0283	71	69.4
Spectral Aug	BrightScale	70.65	2.60%	0.0091	91	87.1
Mixed	rotation_shift + Noise	72.72	4.63%	0.0201	100	95.24
Hierachy	rotation_shift & L1+L2	**89.05**	20.96%	0.0121	90	139.14

Table 6: CNN - Model Performance Across Training Configurations

Learning rate testing shows 0.0015 performs the best (Test Acc = 69.25%, Epoch = 86), better than 0.0005 (68.90%) and 0.002 (67.90%), and is used in subsequent experiments. For split strategy, 80_10_10 performs the most stable (69.07%), better than 70_15_15 and 90_05_05. The latter suffers from greater fluctuations due to too few validation samples. In terms of regularization, light Dropout (0.15/0.3/0.2) improves accuracy to 71.26%, the best so far. Overly strong regularization reduces validation accuracy to 66.12%, reflecting limited learning. Overall, moderate regularization helps improve generalization.

Spatial and Spectral Data Augmentation

This study evaluates spatial augmentation methods including flip, rotation, and rotation_shift. Among them, rotation_shift achieves the highest accuracy (73.31%), outperforming rotation_90 (71.99%) and flip_horizontal (70.30%) by effectively simulating minor viewpoint variations.

For spectral augmentation, adding Gaussian noise (std = 0.01–0.04) to each band yields 72.34% accuracy with improved robustness (gap = 0.0283), while brightness scaling performs slightly lower at 70.65%. Combining rotation_shift with spectral noise achieves 72.72%, suggesting improved stability but no additive gain. Overall, rotation_shift is the best non-hierarchical augmentation strategy.

Hierarchical Data Enhancement

To enhance fine-grained classification, this study introduces the three-level classification of TreeSatAI and uses L1 (leaf type) and L2 (genus) as auxiliary inputs. The model combines the original tensor with embedded vectors (8D and 16D), resulting in a 152-dimensional input. Total parameters slightly increase. While retaining the original CNN backbone, the hierarchical version adds feature fusion and extra fully connected layers, increasing parameters to 237,171 (+4%). Experimental results show the test accuracy increases to 89.05% (+21%), and the gap drops to 0.0055. Results confirm that embedding hierarchical labels significantly improves fine-grained classification and reveals the potential value of biological taxonomy in remote sensing tasks.

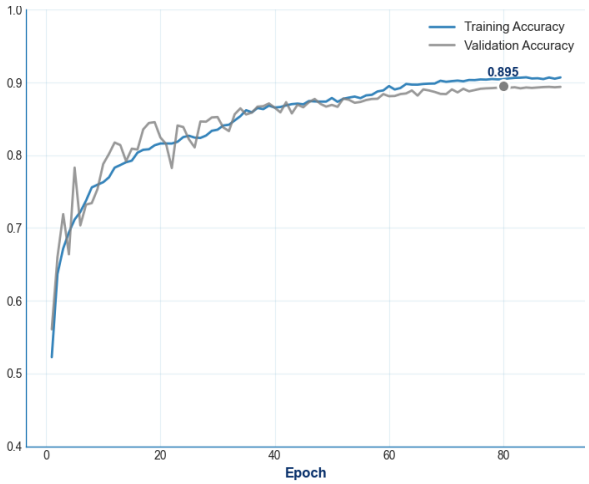


Figure 8: CNN - Training Accuracy of the Hierarchical Model

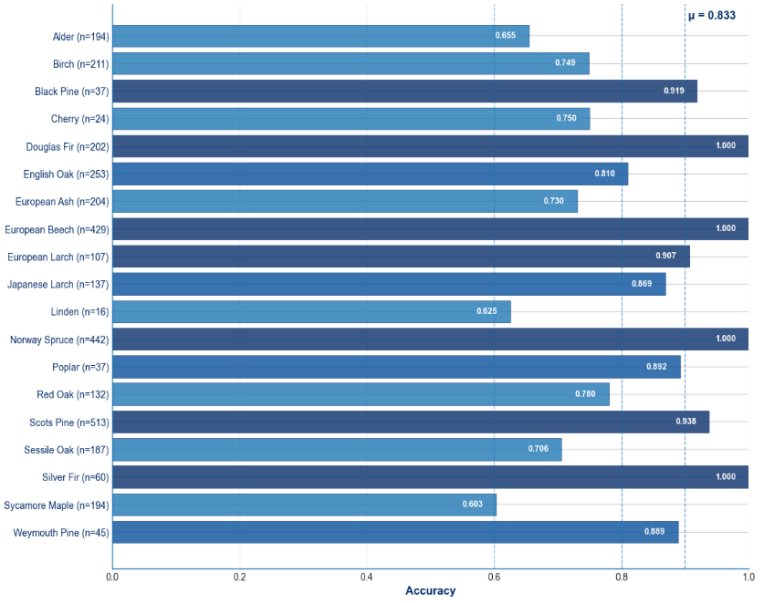


Figure 9: CNN - Species Classification Accuracy of the Hierarchical Model

Performance Evaluation

This study uses an assigned independent test set (6,077 samples, 19 species) for final validation. The hierarchical model achieved **80.32%** accuracy, with 100% identification in European beech, Douglas fir, Norway spruce, and Silver fir. Mid-sized classes like Birch (88.47%), Japanese larch (85.04%), and Scots pine (99.13%) also showed stable results.

The 8.73% drop from validating accuracy (89.05%) is mainly due to severe class imbalance, Zero-accuracy classes (Black pine: n=6; Linden: n=14). A strong correlation between sample size and accuracy (r = 0.621) further validates the model design. Overall, results confirm the hierarchical strategy's effectiveness in well-represented classes.

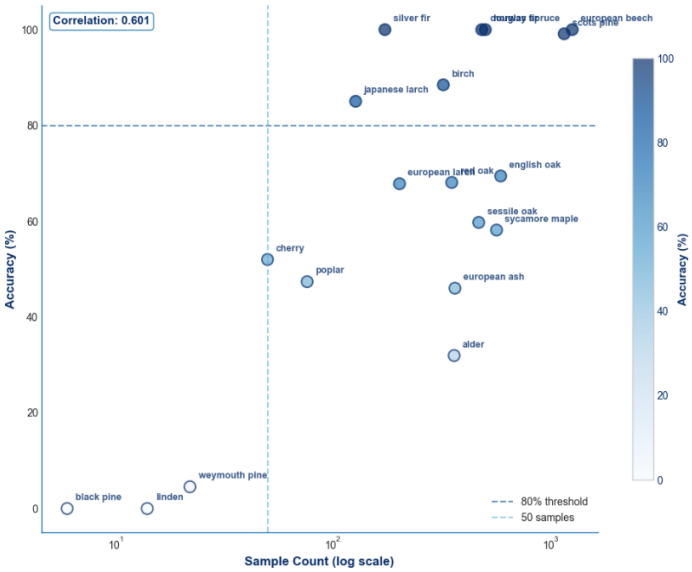


Figure 10: CNN - Relationship Between Sample Size and Accuracy

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.