

# EXPLORING INTERPRETABILITY IN DEEP LEARNING-BASED FOREST CLASSIFICATION - SEPARATING SIMILAR FOREST CLASSES USING SENTINEL-2 TIME SERIES

PREPRINT, COMPILED OCTOBER 15, 2026

Maximilian Kirsch <sup>‡</sup> and Christine Preisach <sup>‡</sup>

<sup>‡</sup>Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, Germany

## ABSTRACT

The ecological insurance theory posits that diverse forests are more resilient and better able to adapt to changing environmental conditions. Increasing drought and heat stress in Central Europe promotes mixed forests for maintaining ecosystem services. However, a diverse species distribution complicates remote sensing-based species mapping, as distinguishing pure from mixed species compositions at the pixel level remains challenging. We evaluated the potential of recurrent neural networks (RNNs) for tree-species classification from Sentinel-2 time series and investigated the spectral-temporal features underlying class discrimination, with particular emphasis on separating pure from mixed Norway spruce pixels. We compared RNN classification models with Random Forest (RF) and Extreme Gradient Boosting (XGBoost) and used Integrated Gradients (IG) to quantify feature attributions across spectral bands and time. The best-performing RNN achieved a Macro-F1 score of 0.851 on the test data, outperforming XGBoost by 0.012 and Random Forest by 0.103. Beyond classification performance, the attribution analysis revealed distinct spectral-temporal patterns associated with the separation of pure and mixed Norway spruce pixels. Attribution values increased particularly after approximately day of year (DOY) 125, while B11 showed the strongest contribution during periods of increasing spectral divergence between the two classes. These findings show that recurrent models can provide competitive classification performance while also revealing the phenological periods and spectral regions that are most relevant for distinguishing closely related forest compositions. Combining time-series classification with explainability therefore provides additional insight into the spectral-temporal mechanisms underlying tree-species discrimination and may support more detailed mapping of mixed forest structures.

**Keywords** Tree Species Classification · SITS · Deep-Learning · Explainability · XAI

## 1 INTRODUCTION

Understanding the spatial distribution of tree species provides a basis for decision-making in forest management, biodiversity conservation, and climate protection [1]. Repeated monitoring of forest composition can further reveal changes caused by natural disturbances, harvesting, and regrowth, thereby supporting forest inventories and long-term forest management [2, 3]. Under increasingly frequent heat, drought, and disturbance events [4], forests in Central Europe need to be more resilient and adapted to change [5]. The ecological insurance theory posits that greater species diversity can stabilize ecosystem functioning to environmental fluctuations [6]. Hence, the *New EU Forest Strategy for 2030* of the European Commission explicitly promotes the creation and maintenance of functionally diverse, mixed-species forests [7]. Consequently, mapping approaches must also reliably represent mixed species compositions. Satellite-based forest classification has been applied across a wide range of spatial scales, from site-level studies to global forest mapping [8, 9]. Sentinel-2 is particularly well suited for this purpose because it provides freely available multi-spectral imagery at high revisit frequency and spatial resolutions of up to 10 m. Early applications largely relied on single-date imagery. For example, Immitzer et al. [10] evaluated Sentinel-2 data for crop and tree species classification in Central Europe using a supervised Random Forest (RF) classifier and achieved an overall accuracy of 65% for tree species classification. Their results further highlighted the importance of the red-edge and short-wave infrared (SWIR) bands. However, single-date ob-

servations capture only one stage of the seasonal development and may therefore provide limited discriminatory information for species with similar spectral characteristics at a given phenological stage. With the launch of Sentinel-2B in 2017, the increased observation frequency facilitated the use of Satellite Image Time Series (SITS) for vegetation and forest classification [11]. Exploiting this temporal information can improve species discrimination by capturing phenological differences in seasonal spectral trajectories. Grabska et al. [12], for example, compared single-date and multi-temporal Sentinel-2 imagery for the classification of nine tree species in mixed forests and reported an increase in overall mapping accuracy of 5–10% when using time-series data, with combinations of spring and autumn observations being particularly informative. In tree species classification, Sentinel-2 SITS are commonly analyzed using machine-learning approaches, with tree-based ensemble methods such as RF and Extreme Gradient Boosting (XGBoost) being frequently applied (e.g., [12, 13, 14, 15, 16]). These methods are widely used because of their strong predictive performance, scalability to large study areas, and comparatively high interpretability [17]. However, SITS contain sequential spectral information whose temporal dependencies can also be modeled directly using deep learning approaches. The continued operation of Sentinel-2 has progressively expanded the available image archive, providing increasingly long time series that also facilitate the application of such sequence-based deep-learning approaches (e.g., [18, 17]).

\*correspondence: maximilian.kirsch@h-ka.de

Despite their increasing relevance for ecology and management, distinguishing mixed forests from pure stands using remote sensing data is challenging [19]. In medium-resolution satellite images, such as Sentinel-2 (10m), pixels are likely to contain several individual trees, and the proportion of each species within the pixels can vary substantially, which is reflected in classification performance. Blickensdörfer et al. [13], for example, mapped the main tree species in Germany using combined Sentinel-1 and Sentinel-2 time series and integrated exogenous features such as temperature and topographic features. While they reported high F1 scores for the five dominant species in pure stands, ranging from 72% to 97%, the inclusion of mixed stands reduced accuracy by 4-14 percentage points. At the same time, most forestry inventory, planning, and management tools have been developed for monocultures [5] and tree species classification studies prominently focus on pure stands (e.g., [14, 20]). This is a limitation, especially in Germany, as mixed forests account for 79% of Germany's forest area in 2022 [21]. Disregarding mixed stands in forest mapping products can therefore substantially misrepresent forest composition. Given the difficulty of distinguishing mixed from pure forest classes, classification performance alone provides limited insight into which spectral and temporal characteristics enable their separation. Research on explainability in remote sensing-based tree species classification has predominantly focused on tree-based methods, partly for their comparatively high interpretability (e.g., [14, 22, 15]; [17]). Immitzer et al. [22] investigated the spectral variables and acquisition dates most relevant for classifying individual tree species using RF models trained on combinations of Sentinel-2 bands and spectral indices across 18 scenes. Their results highlighted the red and short-wave infrared bands, while observations acquired between April and August contributed most strongly to classification performance. Similarly, Wegler et al. [14] assessed feature contributions in RF and XGBoost models using SHAP values and model-based feature importance and identified the short-wave infrared and red-edge bands among the most influential features. While Malkanthi et al. [23] also addressed explainability in deep learning-based tree species classification, their analysis focused on exogenous variables rather than on the spectral-temporal information contained in the SITS. Consequently, although previous studies have provided valuable insights into the spectral and temporal features relevant for tree species classification, the spectral-temporal decision processes of deep learning models remain insufficiently understood. This limitation is particularly relevant for closely related classes, such as pure and mixed Norway spruce, for which understanding the wavelengths and phenological periods driving class separation may provide insights beyond overall classification performance.

To address this gap in SITS-based tree species classification, we propose an explainability workflow for multiclass classification that identifies the spectral-temporal features most relevant for class discrimination while remaining applicable to deep learning models. We particularly investigate the separation of pure and mixed Norway spruce pixels and additionally examine deciduous tree species with similar phenological patterns. The study addresses the following research questions:

**RQ1:** How do recurrent deep learning models compare with established tree-based ensemble models in classifying tree species and distinguishing pure from mixed forest compositions using one-year Sentinel-2 time series?

**RQ2:** Can attribution techniques applied to deep learning models provide insight into the spectral bands and phenological periods relevant for class discrimination, particularly between closely related forest classes?

## 2 MATERIALS AND METHODS

### 2.1 Dataset and Study Area

We used the publicly available Sentinel-2 tree species time-series dataset for Germany, published in 2025 by Freudenberg et al. [20]. The dataset combines bottom-of-atmosphere reflectance time series from analysis-ready Sentinel-2 imagery with tree species labels derived from the German National Forest Inventory. It covers individual tree observations across Germany from 2015 to October 2022 and includes 48 tree species and three species groups. We selected the four dominant tree species or species groups in Germany: Norway spruce (*Picea abies*), Scots pine (*Pinus sylvestris*), European beech (*Fagus sylvatica*), and oak, with the latter comprising sessile oak (*Quercus petraea*) and English oak (*Quercus robur*) [24]. Furthermore, we split the Norway spruce class into healthy and disturbed Norway spruce, as well as mixed pixels that comprise Norway spruce and other tree species (mainly beech) in mixed stands. Additional classes were deliberately excluded, as the remaining tree species in the dataset by Freudenberg et al. [20] were underrepresented for the study's specific use case of classification and separability analysis. An overview of the data basis and the spatial distribution of the dataset locations is shown in Figure 1.

The selected dataset comprised pure and mixed Norway spruce time series, as our objective was to determine whether the models could differentiate between classes that are very closely related. Most tree-species maps in Germany focus on pure stands because mixed pixels introduce substantial classification uncertainty. Their spectral signatures tend to resemble those of the dominant tree species within the pixel, which can lead to mixed pixels being misclassified as pure stands (see Figure 2).

We subsequently filtered the selected data and applied the pre-processing steps described in Section 3.1.

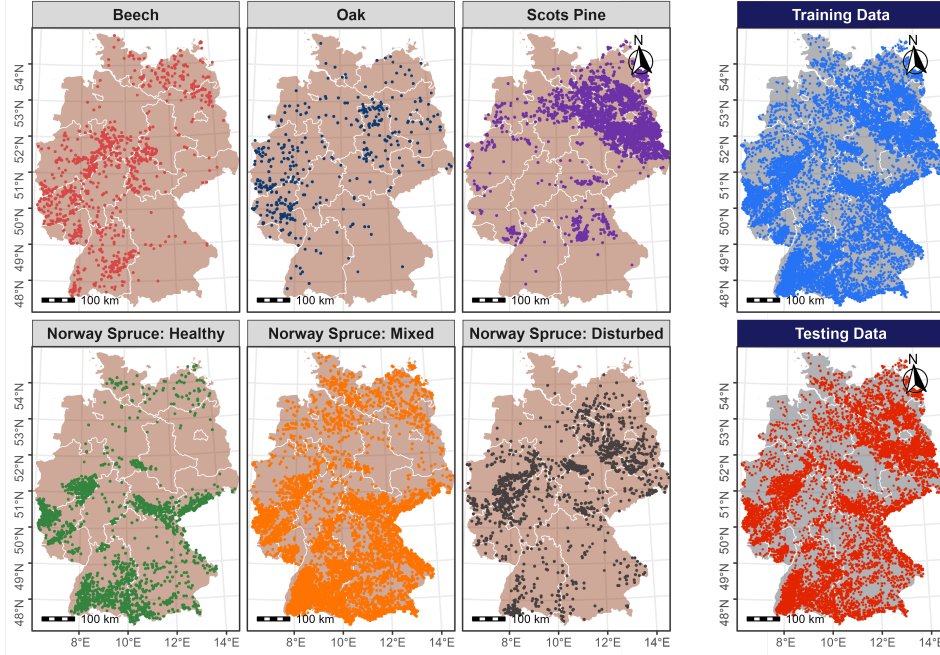


Figure 1: *Data distribution*: spatial distribution of sample locations used in this study, grouped by species.

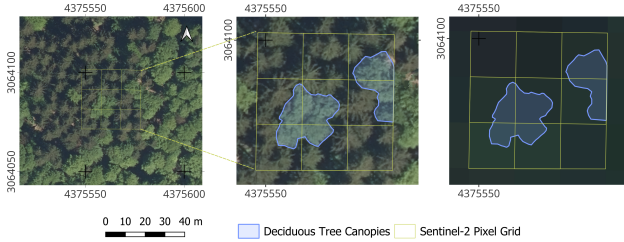


Figure 2: *Mixed Sentinel-2 pixels*. Illustration of Sentinel-2 pixels containing pure Norway spruce stands and mixed stands. *A* shows an orthophoto of a forest structure including Norway spruce and varying deciduous trees. *B* shows a subset of the same area including the outlines of deciduous trees within a  $3 \times 3$  Sentinel-2 pixel grid. *C* shows the corresponding Sentinel-2 pixels averaged over one year (2021).

## 2.2 Models

In this study, we selected two ensemble methods, XGBoost and RF, as well as four deep learning approaches. XGBoost and RF were selected because they have proven to be very successful in the field of SITS forest classification [14, 15, 13]. For the deep learning models, we developed an architecture that utilizes Recurrent Neural Networks (RNNs), which are particularly well suited to sequential data because they can model temporal dependencies within multivariate time series [25]. Their potential for SITS data has also been explored and validated (e.g., [26, 27]). We selected Long Short-Term Memory (LSTM) [28] and Gated Recurrent Unit (GRU) [29] to extract spectral-temporal information from Sentinel-2 time series and embed it into a latent feature space. Their schematic functionality is illustrated in Figure 3.

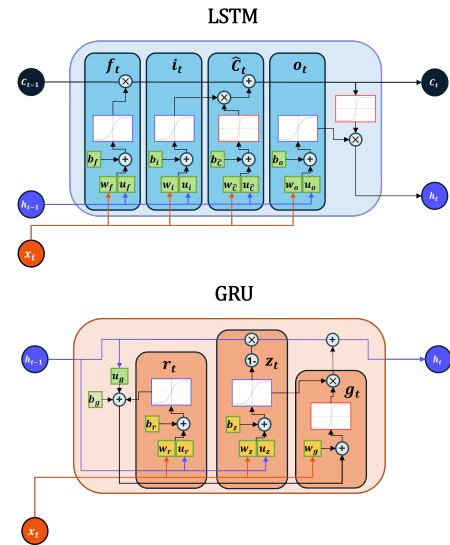


Figure 3: *LSTM and GRU architectures*. At time step  $t$ ,  $x_t$  is the multivariate input. The LSTM unit uses forget ( $f_t$ ), input ( $i_t$ ), and output ( $o_t$ ) gates with candidate cell state  $\tilde{C}_t$ . The GRU unit uses update ( $z_t$ ) and reset ( $r_t$ ) gates with candidate activation  $g_t$ . The LSTM carries forward long- and short-term memory separately as cell state  $C$  and hidden state  $h$ , while GRU only forwards one combined state. Blue and red frames indicate sigmoid ( $\sigma$ ) and tanh activations, respectively. The recurrent weights  $w, u$  and biases  $b$  are learned.

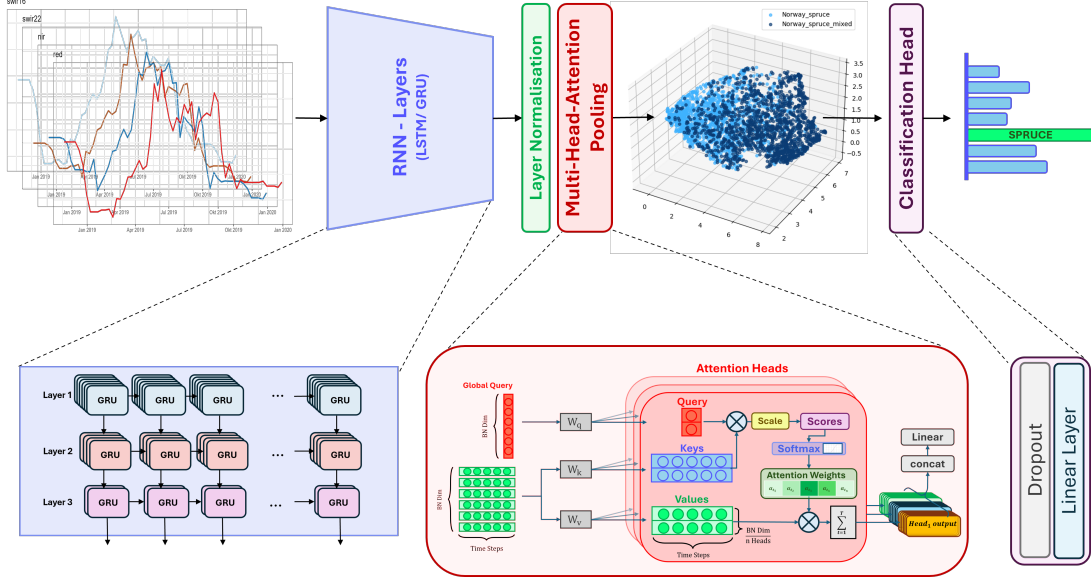


Figure 4: *GRU-MHA Classifier Architecture*. The RNN output is projected into key and value representations, while a learnable global query provides the query representation. After splitting the projected representations into multiple attention heads, scaled dot-product attention is computed between the global query and the keys of all time steps. The resulting attention scores are normalized using softmax and used to calculate a weighted sum of the value vectors, yielding one pooled representation per head. These are connected and used as inputs for the classification head.

The equations for the gated units and the forward pass of the LSTM and GRU are provided in Appendix A.1.

The general model architecture follows an RNN-encoder design, specifically the encoder described by Kirsch et al. [26]. The encoder embeds the spectral-temporal information into a higher-dimensional hidden representation before progressively reducing it to retain the most relevant information. We adapted this architecture for classification by replacing the decoder with a classification head, which maps the learned feature representation to the target classes. In addition, we implemented an optional Multi-head Attention (MHA)-based pooling mechanism, inspired by the Pooling by Multihead Attention (PMA) approach proposed by Lee et al. [30], prior to the classification head. Unlike classification approaches that rely solely on the final hidden state, the attention pooling mechanism incorporates information from all RNN time steps. A learnable query attends to the temporal latent representations, assigning attention weights to the individual time steps and aggregating them into a single fixed-dimensional representation that is subsequently passed to the classification head. The architecture of an RNN model with MHA pooling is illustrated in Figure 4.

We incorporated MHA pooling to allow the model to weight the temporal representations produced by the recurrent encoder. Because time steps contribute unequally to classification, the MHA pooling enables the model to learn time steps containing particularly discriminative information. By using multiple attention heads, different heads can learn complementary weighting patterns [31], potentially capturing information associated with distinct phenological periods. We therefore expected MHA pooling to produce a more informative latent-space representation without increasing the dimensionality of the latent representation. Consequently, we expected it to improve classification,

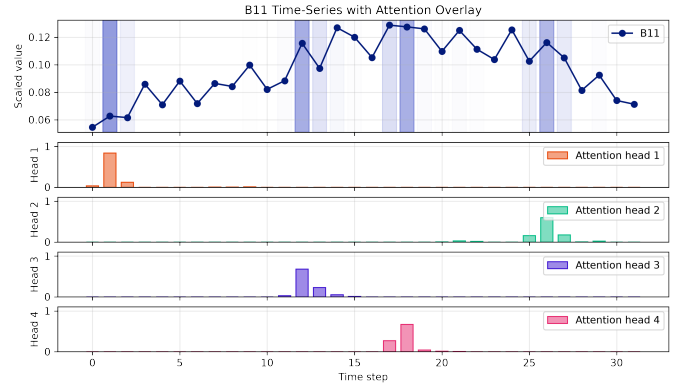


Figure 5: *Multi-head attention weighting*. The top panel shows the temporal profile of spectral band 11 (SWIR 1), which is included for visualization purposes. The panels below show the attention weights assigned to each time step by the individual attention heads. The blue bars in the top panel represent the attention weights aggregated across all heads for each time step. The attention mechanism operates on the latent representations rather than directly on the spectral band values.

particularly for classes whose spectral differences are most pronounced during specific periods of the growing season. This mechanism is schematically illustrated in Figure 5. For simplicity, attention is shown for a single spectral band, although in the model it is applied to the full embedded representation of all bands at each time step.

Because the time series are incomplete and irregularly spaced, we included the day of year (DOY) as a scaled continuous input feature. This provided the recurrent models with explicit

calendar-time information and relaxed the assumption that consecutive observations are equally spaced. Because each sequence covered a single calendar year and did not span the December–January transition, cyclic encoding of DOY was not required.

### 2.3 Feature Importance

Deep learning models are often considered black boxes because they provide limited insight into the factors driving their predictions [32, 33]. There are several attribution methods that assign a scalar contribution value to each input feature, aiming to quantify how strongly each feature influenced the network’s prediction [34]. These methods can make it difficult to determine whether observed inaccuracies originate from the model itself or from the attribution method [35]. We selected *Integrated Gradients* (IG), introduced by Sundararajan et al. [35], as the attribution method for this study. IG is a widely used and theoretically grounded attribution method for explaining predictions of differentiable models [36, 37]. It has also been shown to provide informative explanations for multivariate time-series models [38].

Integrated Gradients is a gradient-based attribution method that assigns an attribution value to each input feature, that is, to each spectral band at each time step. For each input feature, IG integrates the gradients of the model output along a linear interpolation path between a predefined baseline and the observed multivariate time series [35]. The resulting attribution values quantify how each feature contributes to the difference between the model output for the observed input and that for the baseline. Consequently, the resulting attribution matrix represents the contribution of each spectral-band-time-step combination to the prediction of the respective target class relative to the baseline input.

$$\text{IG}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (1)$$

Here  $x$  denotes the input sample and  $x'$  denotes the corresponding baseline input. The term  $\frac{\partial F(x)}{\partial x_i}$  is the gradient of the model output  $F(x)$  along the  $i^{\text{th}}$  dimension. In practice, the integral is approximated using a Riemann sum with  $m$  steps.

$$\text{IG}_i^{\text{approx}}(x) ::= (x_i - x'_i) \times \sum_{k=1}^m \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i} \times \frac{1}{m} \quad (2)$$

The authors suggest comparing observations with a zero baseline to identify the input features most relevant to a model’s classification decision [35]. By applying IG to multiple observations and averaging the results, a more robust estimate of feature importance can be obtained.

Because we focused on classes that were difficult to distinguish, we also used representative baselines derived from the respective comparison classes. These class-specific baselines allowed us to identify the spectral bands and time steps that contributed most strongly to distinguishing the observed class from the class used as the baseline.

## 3 METHODOLOGY

In this section, we describe the methodological framework used for forest classification and model interpretation. We first describe the data preprocessing steps, then the training and inference procedures for the recurrent neural network and ensemble models. We then introduce the evaluation procedure and the Integrated Gradients approach used to analyze the spectral and temporal relevance of the model inputs. To support reproducibility, the code used in this study for data preprocessing, model development, evaluation, and Integrated Gradients analysis is publicly available on GitHub<sup>2</sup>.

### 3.1 Data Preprocessing

Many recent studies use multi-year satellite image time series for forest classification [14, 10, 18]. In contrast, we focused on single-season sequences because the objective of this study is to identify phenologically relevant time periods and spectral regions for separating similar forest structures, rather than to optimize an operational tree species mapping model.

The time series were therefore divided into yearly sequences that can be generalized across annual phenological developments. This reduced the risk of the models learning patterns specific to individual extreme events, such as the exceptionally hot and dry summer of 2018. Furthermore, this approach also increased the number of available training examples. Additionally, we excluded observations before 2017, because Sentinel-2B started in June 2017, leading to more observations per year [20]. The train-test split was performed prior to temporal segmentation to avoid data leakage. This ensured that time series segments of the same pixel from different years were strictly assigned to either the training or the test dataset and did not appear in both datasets simultaneously. We used a standard 80:20 train-test split. To reduce class imbalance in the training data, the number of samples per class was limited to at most twice the sample size of the smallest class:

$$n_c \leq 2 \cdot \min_{k \in C} n_k \quad \forall c \in C \quad (3)$$

where  $n_c$  denotes the number of samples in class  $c$ ,  $C$  denotes the set of all classes,  $k$  is an index over the classes in  $C$ , and  $\min_{k \in C} n_k$  denotes the sample size of the smallest class. Down-sampling was performed after the data were divided into yearly subsequences. This allowed us to retain a spatially and temporally diverse subset of the highly represented classes, as samples were randomly selected from different pixels and years, thereby increasing both spatial and temporal diversity.

Because the deep learning models used sequential inputs, the original time-series structure could be retained. In contrast, the ensemble models required tabular input data. For these models, the time series were transformed into a wide-format representation, where each spectral band at each time step was represented by a separate predictor variable. Using daily time steps would have resulted in  $366 \text{ days} \times 10 \text{ spectral bands} = 3,660$  predictors. This would have produced a very high-dimensional dataset, increasing the risk of overfitting and leading to a large number of missing values, as the time series contained, on average, only

<sup>2</sup>GitHub repository: <https://github.com/Max25832/TreeClassXAI>

Table 1: Final Sequences Sample Size

| Species                 | Description     | Samples |       |
|-------------------------|-----------------|---------|-------|
|                         |                 | Train   | Test  |
| <i>Fagus sylvatica</i>  | pure stand      | 13131   | 6828  |
| <i>Quercus spp.</i>     | pure stand      | 7738    | 2244  |
| <i>Pinus sylvestris</i> | pure stand      | 16702   | 32484 |
| <i>Picea abies</i>      | pure stand      | 15989   | 19688 |
| <i>Picea abies</i>      | mixed stand     | 16965   | 40511 |
| <i>Picea abies</i>      | disturbed stand | 7767    | 2230  |
|                         | raw soil        | 15875   | 17992 |

approximately 24 observations per year. To reduce dimensionality and obtain a more compact representation, we aggregated the data to monthly median values. Remaining gaps were imputed by forward filling. As a result, the tabular dataset used for the ensemble models contained 120 predictors.

### 3.2 Model Training and Inference

Hyperparameter tuning for both ensemble models was performed on the training set using five-fold cross-validation and predefined tuning grids. The evaluated hyperparameters are provided in Appendix Table 6. For the deep learning models, the training set was further split into training and internal validation subsets using an 80:20 split. The internal validation subset was used for checkpoint selection during training, while the hold-out test set was excluded from training, validation, and hyperparameter tuning. After evaluation, test-set Macro-F1 was used to select the model configurations for the feature importance analysis described below. All RNN-based models were trained for 80 epochs using the Adam optimizer with a learning rate of  $1 \times 10^{-3}$ . For the loss function, we used a weighted combination of cross-entropy loss and a Macro-F1-based loss, with weights of 0.7 for cross-entropy loss and 0.3 for the Macro-F1-based loss. We additionally applied early stopping during training, using Macro-F1 on the validation set as the monitoring metric. Both recurrent architectures, GRU and LSTM, were trained with and without multi-head attention pooling. All RNN-based models used a bidirectional variant, in which the input sequence was processed in both forward and backward temporal directions. We tested bottleneck dimensions of 32, 64, and 128. The initial embedding dimension was set to four times the corresponding bottleneck dimension (see Appendix Table 6).

After training, all models were evaluated on the held-out test set using accuracy, balanced accuracy, and Macro-F1. Balanced accuracy and Macro-F1 were included to account for class imbalance in the dataset.

### 3.3 Application to a Mixed-Forest Area

The dataset authors note that it is intended primarily for model training and is less suitable for independent validation [20]. To account for this and to illustrate the practical application of the trained model, we visually selected an additional mixed-forest area in Thuringia from orthorectified aerial images<sup>3</sup>. We

<sup>3</sup>Orthophotos: <https://geoportal.thueringen.de/gdi-th/download-offene-geodaten/download-luftbilder-und-orthophotos>.

extracted Sentinel-2 time series for each pixel within the area using Google Earth Engine and classified them with the best-performing trained model. The resulting pixel-wise predictions were used to generate a species-composition map and to estimate the area assigned to each forest class. We used orthophotos from 2023 for visual confirmation and classified the years 2022 and 2023. For each pixel, we selected the year with the higher classification confidence to generate the final map.

### 3.4 Feature Importance Analysis

To investigate model explainability, we applied IG to quantify the contribution of individual spectral bands and time steps to each classification. For each of the main tree species, we initially computed attributions relative to a zero baseline. We randomly selected 2,000 multivariate time series per class from the test set and restricted the analysis to samples that were correctly classified by the respective model. This was done to ensure that the resulting attribution patterns reflected successful model decisions. The path integral in IG was approximated using a Riemann sum with  $m = 30$  interpolation steps between the baseline and the observed input. Furthermore, we compared observed time series against class-specific baselines derived from other classes to identify the spectral and temporal features that contributed most strongly to distinguishing the observed class from the respective baseline class. In this context, an individual input feature corresponds to the value of a single spectral band at a single time step. Because IG requires the baseline and input to have the same shape, we had to account for the varying sequence lengths. We therefore constructed a median baseline time series for each selected baseline species by calculating the median separately for each spectral band and day of year across all available observations. For each observed sequence, we then selected only those values from the median baseline that corresponded to the day of year present in that sequence, thereby ensuring temporal alignment between the input and its baseline. This process is illustrated for a single spectral band of one observation in Figure 6.

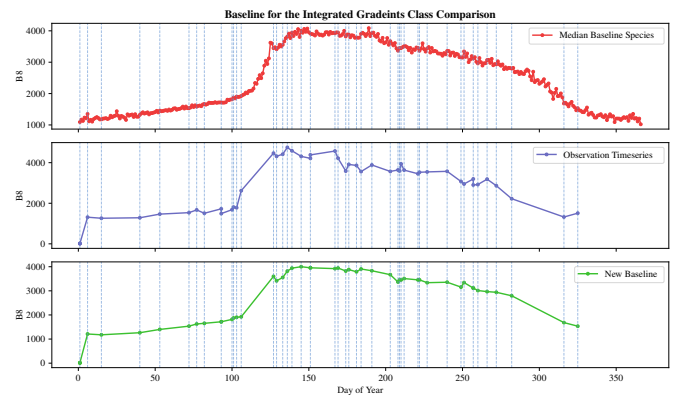


Figure 6: Class-specific baseline derivation for IG attribution. For each observation, a baseline with the same temporal structure was derived from the median time series of the comparison class. The example shows an oak observation (blue) and the corresponding beech baseline (green) for band 11, derived from the median of all beech observations (red) and matched to the observation time steps.

For the feature importance analysis, we selected the model configurations based on the highest test-set Macro-F1 scores. We did not compare models that differed only in embedding dimensions. In contrast, we separately analyzed models with and without multi-head attention pooling because these architectures aggregate the temporal representations differently and may therefore differ in attribution patterns. We compared all classes against the zero baseline. In addition, pairwise class-specific baseline comparisons were performed between *pure Norway spruce* and *mixed Norway spruce* and between *oak* and *beech*. Each pairwise comparison was conducted in both directions, with each class serving once as the observed class and once as the baseline class.

## 4 RESULTS

To evaluate model performance, we compared RNN configurations with different recurrent architectures, input and bottleneck dimensions, and either MHA pooling or no pooling against the ensemble models. Performance was assessed primarily using the Macro-F1 score, supported by additional classification metrics. We further applied IG to analyze the spectral and temporal patterns contributing to correct class predictions and conducted detailed class-comparison analyses for Norway spruce versus mixed Norway spruce and oak versus beech.

### 4.1 Classification Performance

Based on the Macro-F1 score, the GRU model with multi-head attention and a bottleneck dimension of 128 achieved the best overall performance, with a Macro-F1 score of 0.851. It additionally reached an accuracy of 0.854 and a balanced accuracy of 0.933. The GRU-128 model without attention performed similarly but achieved a slightly lower Macro-F1 score of 0.847, together with an accuracy of 0.853 and the same balanced accuracy of 0.933. Among the baseline models, XGBoost achieved the highest Macro-F1 score of 0.839 and an accuracy of 0.851, whereas the Random Forest showed the lowest overall performance. For both recurrent architectures, larger bottleneck dimensions generally corresponded to higher Macro-F1 scores, while the effect of multi-head attention was comparatively small. MHA did not improve the Macro-F1 performance of the LSTM models, but increased Macro-F1 by 0.004 for the GRU models. These results are shown in Table 2.

The confusion matrices in Figure 7 revealed that the main classification challenge was distinguishing Norway spruce from mixed Norway spruce. Across all models, approximately one-fifth of the samples in each class were confused with the other class. The green confusion matrices are normalized by true class and report the proportion of samples in percent; therefore, the diagonal entries correspond to class-specific sensitivity. For the GRU-MHA model, 21.6% of Norway spruce samples were classified as mixed Norway spruce, while 21.2% of mixed Norway spruce samples were assigned to Norway spruce. Accordingly, class-specific sensitivity was comparatively low for Norway spruce (76.3%) and mixed Norway spruce (72.3%). The remaining classes achieved substantially higher sensitivities. The GRU-MHA model reached 97.2% for Scots pine, 96.3% for beech, 90.5% for disturbed Norway spruce, 93.7% for oak, and 97.8% for soil. Compared with the corresponding GRU model

Table 2: Comparison of Model Configurations

| Model   |           |         | Performance metrics |              |              |
|---------|-----------|---------|---------------------|--------------|--------------|
| RNN     | Attention | BN Dim. | Acc                 | BA           | Macro-F1     |
| LSTM    | MHA       | 32      | 0.832               | 0.909        | 0.825        |
|         |           | 64      | 0.827               | 0.920        | 0.814        |
|         |           | 128     | 0.848               | 0.930        | 0.843        |
|         | None      | 32      | 0.821               | 0.912        | 0.813        |
|         |           | 64      | 0.830               | 0.923        | 0.830        |
|         |           | 128     | 0.850               | 0.931        | 0.843        |
| GRU     | MHA       | 32      | 0.832               | 0.917        | 0.821        |
|         |           | 64      | 0.840               | 0.924        | 0.835        |
|         |           | 128     | <b>0.854</b>        | <b>0.933</b> | <b>0.851</b> |
|         | None      | 32      | 0.825               | 0.915        | 0.813        |
|         |           | 64      | 0.828               | 0.925        | 0.832        |
|         |           | 128     | 0.853               | <b>0.933</b> | 0.847        |
| XGBoost |           |         | 0.851               | 0.925        | 0.839        |
| RF      |           |         | 0.792               | 0.890        | 0.748        |

\* BN Dim. = Model Bottleneck Dimension; Acc = accuracy; BA = balanced accuracy.

without MHA, the addition of MHA reduced the proportion of oak samples misclassified as beech from 4.5% to 2.6%, while the reverse confusion increased slightly from 1.3% to 1.8%. This indicates that MHA produced only small, class-specific changes in the confusion patterns. The mutual confusion between pure and mixed Norway spruce remained the dominant source of classification error.

The class-specific evaluation metrics support the patterns observed in the confusion matrices. Norway spruce achieved the lowest F1 scores, ranging from 0.614 to 0.691. The GRU-128 model with multi-head attention achieved the highest F1 score (0.691) and sensitivity (0.763), while the highest specificity was 0.919. For mixed Norway spruce, F1 scores ranged from 0.697 to 0.784, with two recurrent models tying for the highest score. The maximum sensitivity and specificity were 0.729 and 0.943, respectively. For both classes, specificity exceeded sensitivity, indicating that false-negative rates were higher than false-positive rates. The remaining classes achieved higher F1 scores. The best scores were obtained by the GRU-128 model with multi-head attention for disturbed areas (0.797), the GRU-128 model without attention for Scots pine (0.968), and the LSTM-128 model without attention for beech (0.909). XGBoost achieved the highest scores for oak (0.836) and soil (0.983), although the recurrent models performed similarly. Overall, soil and Scots pine had the highest class-specific F1 scores, whereas Norway spruce had the lowest. Random Forest generally underperformed compared with XGBoost and the recurrent models. These results are presented in detail in Table 3.

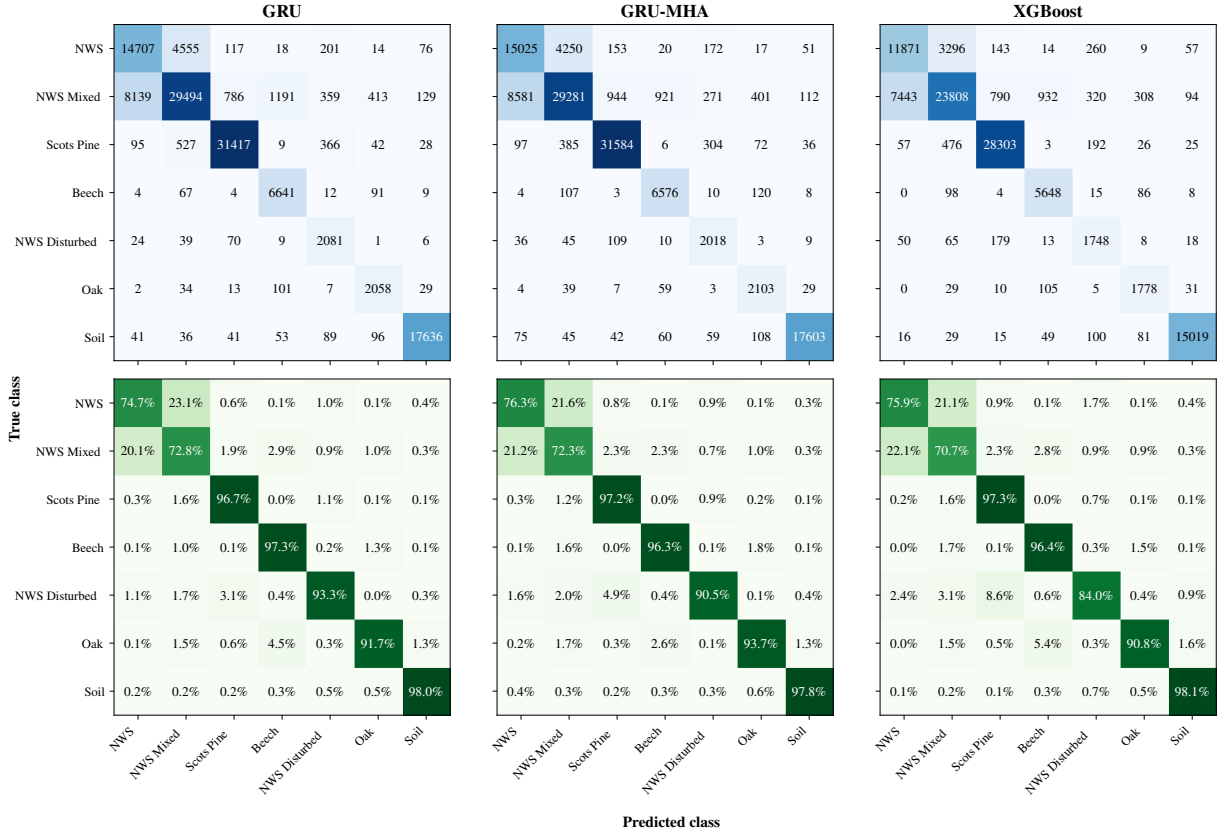


Figure 7: Confusion matrices for the two best-performing RNN models, GRU with and without attention, and the XGBoost model

Table 3: Per-class comparison of model configurations

| Tree species        | Metric      | RF    | XGBoost      | LSTM-128     |              | GRU-128      |              |
|---------------------|-------------|-------|--------------|--------------|--------------|--------------|--------------|
|                     |             |       |              | MHA          | No Att.      | MHA          | No Att.      |
| Norway spruce       | F1          | 0.614 | 0.677        | 0.677        | 0.684        | <b>0.691</b> | 0.689        |
|                     | Sensitivity | 0.704 | 0.759        | 0.731        | 0.742        | <b>0.763</b> | 0.747        |
|                     | Specificity | 0.895 | 0.914        | 0.917        | 0.918        | 0.914        | <b>0.919</b> |
| Norway spruce mixed | F1          | 0.697 | 0.774        | 0.775        | 0.782        | <b>0.784</b> | <b>0.784</b> |
|                     | Sensitivity | 0.610 | 0.707        | 0.717        | <b>0.729</b> | 0.723        | 0.728        |
|                     | Specificity | 0.932 | <b>0.943</b> | 0.934        | 0.933        | 0.940        | 0.935        |
| Scots pine          | F1          | 0.935 | 0.967        | 0.966        | 0.962        | 0.967        | <b>0.968</b> |
|                     | Sensitivity | 0.935 | 0.973        | <b>0.973</b> | 0.964        | 0.972        | 0.967        |
|                     | Specificity | 0.975 | 0.985        | 0.985        | 0.985        | 0.986        | <b>0.988</b> |
| Beech               | F1          | 0.814 | 0.895        | 0.891        | <b>0.909</b> | 0.908        | 0.894        |
|                     | Sensitivity | 0.905 | 0.964        | 0.966        | 0.949        | 0.963        | <b>0.973</b> |
|                     | Specificity | 0.981 | 0.989        | 0.988        | <b>0.992</b> | 0.991        | 0.988        |
| Oak                 | F1          | 0.646 | <b>0.836</b> | 0.835        | 0.809        | 0.830        | 0.830        |
|                     | Sensitivity | 0.825 | 0.908        | 0.924        | <b>0.938</b> | 0.937        | 0.917        |
|                     | Specificity | 0.986 | <b>0.995</b> | 0.995        | 0.993        | 0.994        | 0.995        |
| Soil                | F1          | 0.971 | <b>0.983</b> | 0.982        | 0.980        | 0.982        | 0.982        |
|                     | Sensitivity | 0.963 | <b>0.981</b> | 0.980        | 0.978        | 0.978        | 0.980        |
|                     | Specificity | 0.996 | 0.997        | 0.997        | 0.997        | <b>0.998</b> | 0.997        |
| Disturbed           | F1          | 0.556 | 0.741        | 0.776        | 0.776        | <b>0.797</b> | 0.779        |
|                     | Sensitivity | 0.772 | 0.840        | 0.913        | 0.921        | 0.905        | <b>0.933</b> |
|                     | Specificity | 0.979 | 0.991        | 0.992        | 0.992        | <b>0.993</b> | 0.991        |

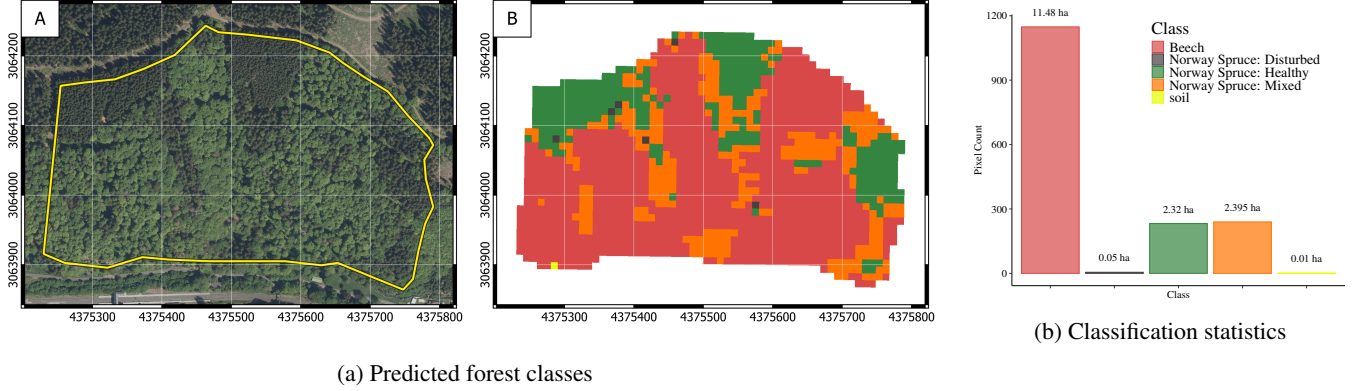


Figure 8: Map with distribution statistics of predicted forest classes.

#### 4.2 Performance in a Mixed-Forest Area

To additionally assess model performance in a mixed-forest area, we applied the GRU-MHA-128 model to single-year time series from a randomly selected site in Thuringia. The resulting classification map is shown in Fig. 8.

The model classified 11.48 ha as beech, 2.32 ha as pure Norway spruce, 2.40 ha as mixed Norway spruce, 0.05 ha as disturbed forest, and 0.01 ha as open soil. Because pixel-level labels were not available, we used an orthorectified aerial image as a qualitative visual reference. The visual comparison indicated reasonable agreement between the predicted classes and the aerial image, providing an additional qualitative assessment of model performance outside the original dataset.

#### 4.3 Model Explainability

To improve the explainability of the deep learning model, we applied IG in two ways. First, individual class attributions were calculated using a zero baseline. Second, we directly compared Norway spruce with mixed Norway spruce and oak with beech by using the respective opposing class as the baseline. The results of these analyses are presented in the following sections.

##### 4.3.1 Class-Specific Integrated Gradients

Hereafter, spectral bands are referred to by their band identifiers, as defined in Table 4.

Table 4: Sentinel-2 Level-2A bands used in this study.

| Band Identifier | Band Name  | Central Wavelength (nm) |
|-----------------|------------|-------------------------|
| B2              | Blue       | 492.7                   |
| B3              | Green      | 559.8                   |
| B4              | Red        | 664.6                   |
| B5              | Red Edge 1 | 704.1                   |
| B6              | Red Edge 1 | 740.5                   |
| B7              | Red Edge 1 | 782.8                   |
| B8              | NIR        | 832.8                   |
| B8A             | Narrow NIR | 864.7                   |
| B11             | SWIR 1     | 1613.7                  |
| B12             | SWIR 2     | 2202.4                  |

\* NIR = near infrared; SWIR = short-wave infrared.

The feature attribution analysis revealed distinct temporal and spectral patterns among the classes. For beech, attribution values peaked around DOY 120 in bands B3, B5, B6, and B11. Oak showed a similar temporal pattern, but with lower attribution values in bands B3–B6 and an earlier peak in B12, approximately between DOY 110 and 125. For Scots pine, the IG analysis showed high B11 attribution values between approximately DOY 130 and 190, whereas B6 exhibited moderate attribution values throughout the year. Pure Norway spruce and mixed Norway spruce showed similar attribution patterns, with high values in B6 around DOY 150. For mixed Norway spruce, the B6 attributions were concentrated between approximately DOY 140 and 175. In contrast, for pure Norway spruce, they extended from approximately DOY 120 to 220. Both Norway spruce and Scots pine showed elevated attribution values between approximately DOY 100 and 180. However, Scots pine also exhibited elevated attributions toward the end of the year. The disturbed Norway spruce class showed no pronounced temporal attribution pattern. Attribution values were broadly similar across bands, except for B2, B7, and B12.

Overall, the IG analysis revealed consistent differences in attribution magnitude among the spectral bands. Across classes, B2 and B12 generally showed the lowest attribution values, whereas B6 and B11 showed the highest attributions for correct species classifications. The results are visualized in Figure 9.

##### 4.3.2 Integrated Gradients for Class Comparison

The IG analysis focused on the distinction between pure and mixed Norway spruce stands and between beech and oak. Because beech and oak were classified with comparatively high performance, the following analysis focuses on the attribution patterns for pure and mixed Norway spruce, while also examining differences between beech and oak.

In the comparison between pure and mixed Norway spruce, B6 and B11 showed the highest overall attribution values. Most other bands exhibited comparable cumulative attributions across the year, whereas B2 and B12 consistently showed lower values. B11 exhibited elevated attributions throughout much of the year, with a pronounced peak between approximately DOY 140 and 180. Attributions for B6 increased after approximately DOY 125 and remained elevated until around DOY 290. The visible bands B2, B3, and B4 and the red-edge band B5 showed comparatively

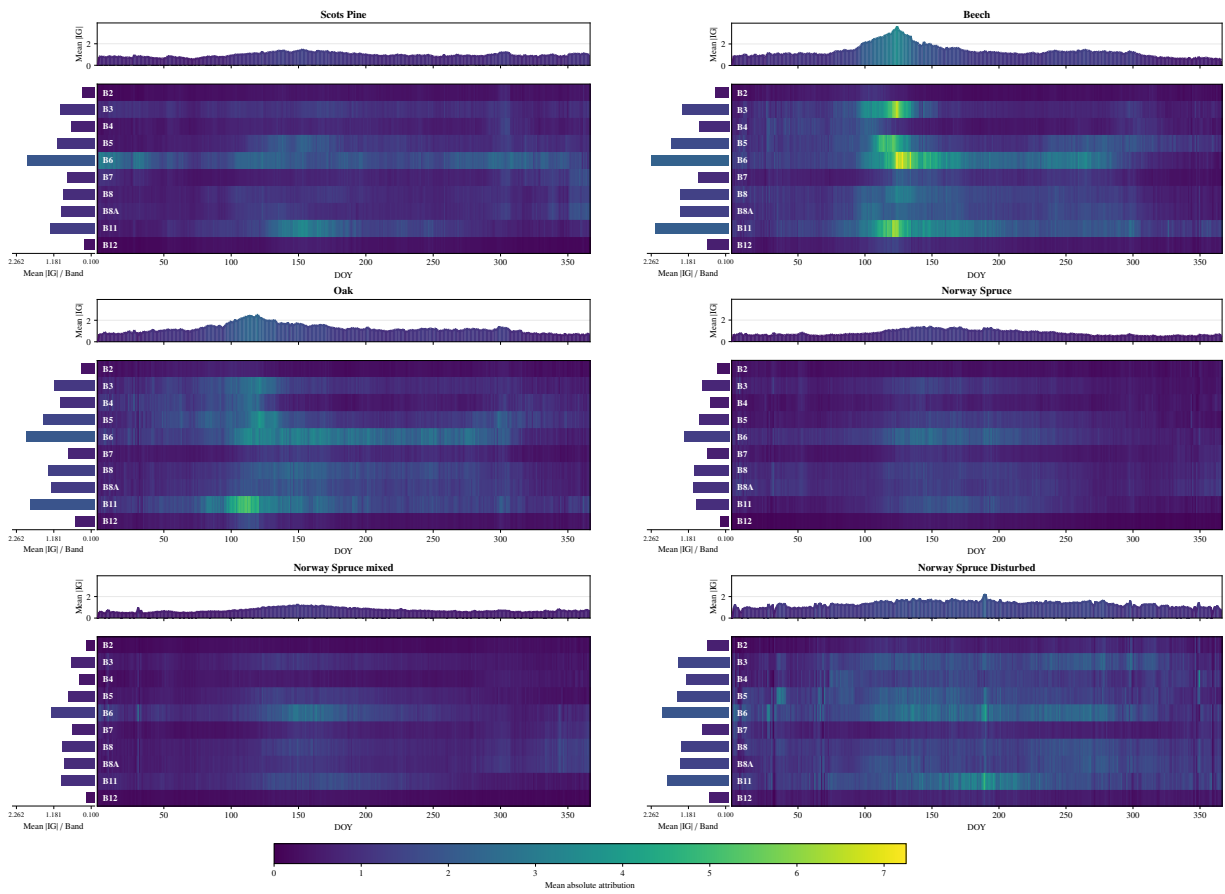


Figure 9: Integrated Gradients attribution results for the six investigated classes against a zero-baseline.

low but persistent attributions throughout the year. Bands B7, B8, and B8A followed a temporal pattern similar to that of B6, although their attribution magnitudes were generally lower. Overall, the highest attribution values for distinguishing pure from mixed Norway spruce occurred between approximately DOY 125 and 225. This comparison is visualized in Figure 10.

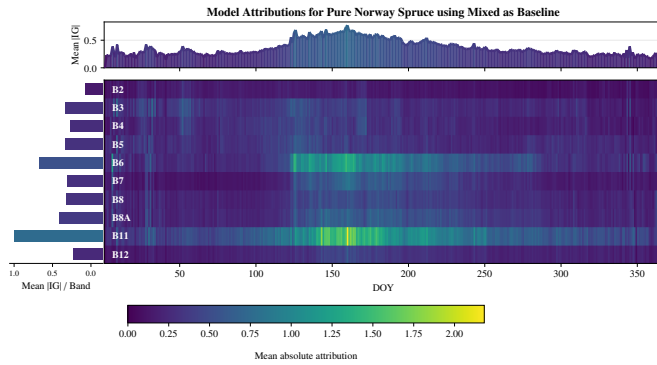


Figure 10: Attribution map for the separation of pure and mixed Norway spruce.

For beech versus oak, B3, B5, and B6 showed attribution patterns similar to those observed for beech relative to the zero baseline. However, B6 exhibited elevated attributions over a broader period, approximately from DOY 110 to 140. The highest attribution values were observed for B6 and B11, particularly between approximately DOY 110 and 150. This comparison is visualized in Figure 11.

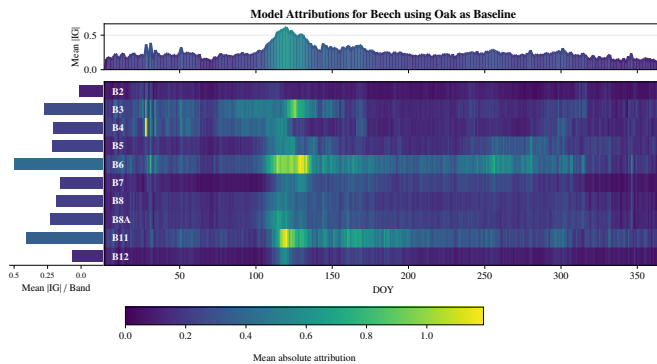


Figure 11: Attribution Map for the separation of beech and oak

## 5 DISCUSSION

This study demonstrates the value of recurrent neural networks and attribution-based analysis for classifying and interpreting annual multispectral time series. The RNN models outperformed XGBoost, highlighting their ability to capture temporal dependencies in the data. Combined with Integrated Gradients, they also enabled a more fine-grained interpretation of relevant spectral bands and time periods than feature-importance measures from tree-based models, which are typically aggregated over broader temporal intervals. In addition, the class-comparison attribution analysis provided insights beyond the conventional zero-baseline approach by revealing shifts in the relative impor-

tance of individual spectral bands. These findings show that directly comparing classes can improve the interpretation of spectral and temporal differences between closely related forest classes.

### 5.1 Comparative Model Performance

The class-specific results indicate that classification performance was primarily determined by the separability of the spectral-temporal classes rather than by the architecture of the model. The main source of error across all configurations was the mutual confusion between pure and mixed Norway spruce, with approximately one-fifth of the observations misclassified between these two classes. This persistent confusion is likely due to the strong overlap in multispectral and temporal characteristics between the two classes. Pure stands tend to occupy more distinct regions of spectral-temporal feature space, whereas mixed pixels are strongly influenced by their species composition and form a gradual spectral continuum. The difficulty of separating these classes therefore appears to be constrained not only by model capacity but also by the information content of the Sentinel-2 time series and the definition of the classes themselves. In contrast, Scots pine, beech, and soil were consistently classified with high accuracy across model architectures, further supporting the interpretation that class separability was the dominant factor determining classification success. Disturbed Norway spruce was also separated more clearly from healthy spruce than pure and mixed healthy spruce were from each other. Disturbance alters canopy condition and therefore produces stronger deviations in spectral-temporal trajectories, which may produce a more distinctive spectral-temporal signal than differences in species composition within otherwise healthy spruce stands.

Despite these class-dependent limitations, the recurrent models achieved the highest overall classification performance, with the GRU using multi-head attention and a bottleneck dimension of 128 reaching the highest Macro-F1 score of 0.851, while Random Forest consistently performed worst. However, the improvement over XGBoost was relatively small. This suggests that, although explicit sequence modeling provides an advantage for annual multispectral time-series classification, much of the relevant temporal information can also be captured through engineered or temporally aggregated features. The benefit of recurrent modeling was particularly evident for disturbed Norway spruce, where the recurrent models achieved substantially higher F1 scores and sensitivity than XGBoost, indicating that explicit modeling of temporal dependencies may be especially useful for classes characterized by changes in canopy condition over time. In contrast, improvements for most healthy tree-species classes were small, and XGBoost remained competitive or even performed best for some well-separated classes.

Larger bottleneck dimensions generally improved the performance of both recurrent architectures, suggesting that sufficient latent capacity may be more important than the choice between LSTM and GRU or the inclusion of attention. Although larger bottleneck dimensions may improve classification by preserving more information, they also increase the risk of overfitting. This trade-off should be considered when selecting the latent representation size. The funnel-shaped encoder progressively compresses the hidden representation and may therefore encourage the model to retain the most relevant information. However,

the present experiments cannot determine whether the funnel architecture itself is beneficial or whether the observed gains primarily result from the increased latent dimensionality and overall model capacity. A direct comparison with a non-funneling architecture of similar size should be considered in future work.

MHA had only a limited effect on predictive performance. It produced a marginal improvement for the GRU-128 model and did not improve the Macro-F1 score of the LSTM-128 model, suggesting that the recurrent hidden states already encoded most of the temporal information required for classification. Nevertheless, MHA pooling improved the Macro-F1 score across all GRU configurations, with somewhat larger gains for models with smaller bottleneck dimensions, and produced minor class-specific improvements. Attention may therefore compensate for limited latent capacity by suppressing less informative temporal representations. However, its main value in this study lies in improving the temporal interpretation of the model by emphasizing relevant time steps and making their contribution more apparent in the attribution analysis.

Overall, the performance gains over XGBoost were class-dependent and moderate. Latent-space size had a clearer influence on performance than MHA, while the remaining errors were driven mainly by limited spectral-temporal separability rather than by the selected model architecture. The main advantage of the RNNs, however, lies in their temporally structured representation, which supports a more detailed interpretation of phenologically relevant periods.

## 5.2 Explainability

The class-comparison IG analysis provided insights beyond the zero-baseline analysis. Relative to a zero baseline, B6 showed the highest attributions for both pure and mixed Norway spruce. However, when each class was evaluated against a baseline derived from the other class, B11 became the most strongly attributed band. Thus, the features supporting individual class predictions differed from those used to distinguish the two closely related classes. The direct comparison also revealed a seasonal pattern that was not apparent in the zero-baseline results. Attributions for B6, B7, B8 and B8A increased after approximately DOY 125, broadly coinciding with leaf development in Central European broadleaf species [39]. The differences become clearer when the median reflectance time series of pure and mixed Norway spruce are compared (see Figure 12). The median B6 reflectance trajectories of the two classes begin to diverge around DOY 100 and remain separated for most of the growing season. However, the highest attribution values during this period occur in B11 rather than B6. In particular, B11 attributions increased when the difference between the B6 trajectories was most pronounced. B11 reflectance also remained consistently higher for the mixed class throughout the year, with a peak in late spring. These results suggest that the model did not rely on individual bands independently but rather on their combined spectral-temporal patterns when distinguishing pure from mixed Norway spruce stands.

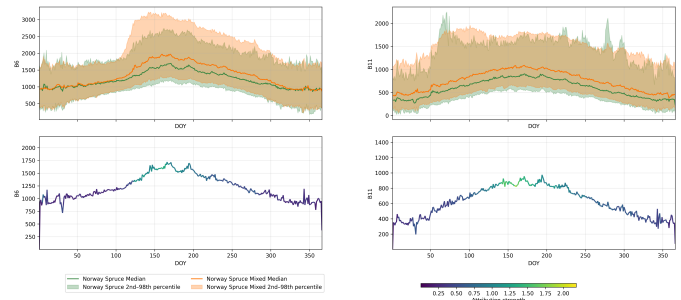


Figure 12: *Separation between pure and mixed Norway spruce.* The top row shows the median test-set time series with confidence bands. The bottom row displays Integrated Gradients attributions along the Norway spruce median time series, highlighting time steps relevant for distinguishing the two species.

This result is particularly relevant because pure and mixed Norway spruce showed the highest degree of mutual misclassification across the evaluated models. Although the generally high relevance of red-edge and short-wave infrared bands is consistent with previous remote-sensing research [14, 22], the class-comparison temporal attribution patterns provide more detailed insight into how the models distinguished these closely related forest classes.

Although the phenological patterns of beech and oak appeared more similar than those of pure and mixed Norway spruce, the models distinguished the two broadleaf species with high accuracy. However, the attribution patterns did not correspond directly to the periods with the largest visible differences between the median time series. Instead, the highest attributions occurred mainly between DOY 100 and 150, coinciding with the leaf-unfolding period. Although the median time series showed little separation between beech and oak during this period, the model appears to have used subtle spectral-temporal differences to distinguish the two species (Figure 13).

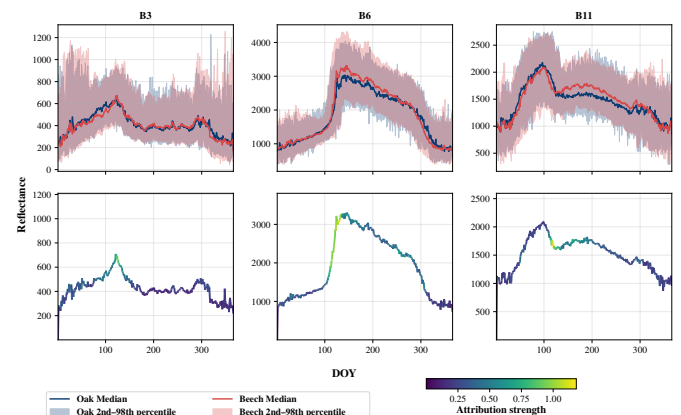


Figure 13: *Separation between Beech and Oak.* The top row shows the median test-set time series with confidence bands. The bottom row displays Integrated Gradients attributions along the beech median time series, highlighting time steps relevant for distinguishing the two species.

### 5.3 Potential Applications in Forestry

We expected the spectral-temporal patterns of species to vary severely between years, potentially reducing the effectiveness of single-year time series. In particular, extreme summers such as 2018 and 2022 in Germany were expected to alter spectral patterns and increase misclassification. However, the proportion of false predictions did not increase noticeably in these years (Fig. 14).

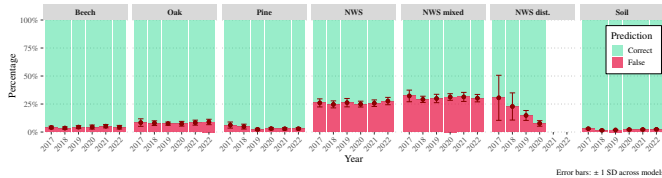


Figure 14: *False Predictions over years*: Here the mean percentage  $\pm$  SD of false predictions for each class is visualized for every observed year across all RNN models. Norway spruce is abbreviated to NWS. The disturbed Norway spruce class has no observations for the years 2021 and 2022, as there was no data where the disturbance year was after 2020.

Across all RNN models, fewer annual observations generally reduced classification performance. The impact of observation count on correct classification percentage for pure and mixed Norway spruce over all RNN models is visualized in Figure 15. This offers potential for future studies, as the launch of Sentinel-2C increased yearly observations in Europe by about 33% and reduced the average datagap length [40], which suggests better interpretability with our approach.

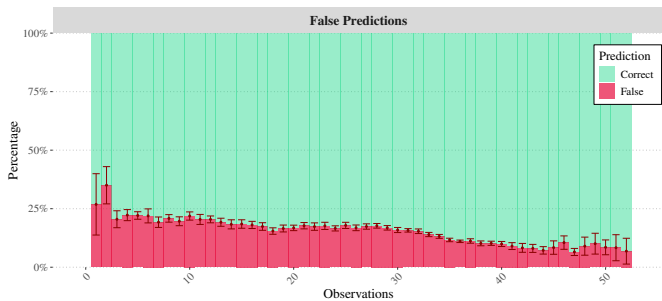


Figure 15: *Mean percentage of false predictions ( $\pm$  SD) for pure and mixed Norway spruce combined, grouped by the number of time steps per time series.*

A direct comparison with studies using multi-year time series for similar forest classes in Germany is difficult, as these studies often included a larger number of classes [14, 13, 18]. This increases the potential for misclassification, particularly among broadleaf species such as beech and oak. At the same time, other studies did not distinguish between pure and mixed Norway spruce stands, which likely contributed to higher Macro-F1 scores for Norway spruce. The classification tasks therefore differ substantially, and a robust comparison would require further targeted studies. Nevertheless, the strong classification performance achieved here indicates that the appropriate input length is task-specific and that single-year time series should not be

excluded by default. For detailed forest mapping, longer multi-year time series may provide higher classification performance [14, 13, 18]. In contrast, shorter time series may be preferable when lightweight operational models or energy-efficient training and inference are priorities. Previous studies have shown that reducing the number of input observations can improve computational and energy efficiency [41]. In such a setting, reducing the number of input features also becomes important, as it further contributes to better energy efficiency [41]. The IG analysis indicated that bands B2 and B12 contributed little to the model predictions and could therefore be tested for removal to improve efficiency. Reducing the input dimensionality may also allow similar performance with smaller hidden representations and fewer trainable parameters, thereby lowering computational demand and associated emissions. However, this would require further evaluation. Identifying the temporal and spectral observations that contribute most strongly to distinguishing pure and mixed stands provides useful information for forest composition mapping. The insights of the explainability analysis could guide the development of future models that better separate pure and mixed stands by focusing on the most informative spectral bands and seasonal periods. This may also enable feature and time-step reduction, supporting the development of more lightweight and computationally efficient models.

### 5.4 Limitations

The training dataset included observations from different locations and years, allowing the models to capture spatial and interannual variability. However, this variability was limited to the period from 2017 to 2022. In addition, the dataset providers noted that the dataset is well suited for model training but less suitable for independent validation [20]. The reported classification performance should therefore be interpreted with caution. Furthermore, Spatial and temporal autocorrelation may have influenced the performance estimates despite measures taken to prevent data leakage. To avoid any data leakage a leave-one-region-out cross-validation approach would be most appropriate. In this setup the data is spatially divided into regions and the model is trained on all but one region. The left out region is used for testing. This is repeated iteratively until each region has served once as the test set. We partly addressed these limitations by applying the model to an independent mixed-forest region. However, this provided only a limited visual assessment of model performance and cannot replace evaluation on a larger and more diverse independent test dataset. Another limitation is that the study region is limited to Germany and the results may not transfer directly to other forest types, climatic conditions or management systems. We selected species classes based on the four predominant tree species in Germany, with different health stages and structures for Norway spruce, which also sets us apart from other German studies in terms of our class definitions and study design. This makes it difficult to draw direct comparisons with previous studies regarding model performance, particularly with regard to the separability of pure and mixed spruce stands, as other studies have not focused on this approach. Finally, the potential efficiency gains from shorter time series, feature reduction, and smaller model architectures were not measured directly. Their effects on computational demand, energy consumption, and classification performance therefore require further evaluation.

## 6 CONCLUSION

This study investigated the classification and interpretation of annual Sentinel-2 time series for forest species mapping, with particular emphasis on the discrimination of spectrally and phenologically similar classes. Recurrent deep learning models were compared with established tree-based ensemble methods, while their spectral-temporal decision processes were examined using Integrated Gradients. The best-performing model, GRU-MHA-128, achieved a Macro-F1 score of 0.851, corresponding to only a modest improvement of 0.012 over XGBoost. The primary aim of this study, however, was not to optimize an operational tree species mapping approach, but to investigate which spectral bands and temporal periods contribute to model predictions and class separation.

Combining recurrent models with Integrated Gradients enabled attribution for each individual spectral band for each available timestep without prior temporal aggregation. In addition to conventional class-wise attributions, using observations from an opposing class as the baseline allowed us to identify the features driving class separation, rather than only those contributing to the prediction of an individual class. For pure and mixed Norway spruce, the contributions increased for B6, B7, B8, and B8A after approximately DOY 125, while B11 showed particularly strong attribution around DOY 100. For beech and oak, B6 and B11 and the period between approximately DOY 110 and 150 were especially relevant for class separation. The analysis further demonstrated that features contributing strongly to the prediction of an individual class are not necessarily identical to those driving the distinction between similar classes.

The spectral temporal attribution patterns identified for pure and mixed Norway spruce broadly corresponded with known phenological processes, particularly the development of broadleaf vegetation within mixed stands. The explainability analysis therefore provides a physically plausible interpretation of the model process, although it did not reveal substantially different spectral-temporal relationships beyond those expected from existing phenological knowledge. Still, explainable AI implications like IG improve the understanding of the discrimination process and can inform future features or data design.

The increased number of annual observations offered by Sentinel-2C motivates repeating the proposed workflow with denser time series in future work, as more observations improved classification performance in this study. Additional temporal information may also reveal spectral-temporal relationships that are currently difficult to resolve. Furthermore, moving beyond categorical pure-versus-mixed labels toward estimating species proportions or subpixel forest composition could provide a more continuous representation of mixed stands and allow a more detailed investigation of the spectral response to increasing species admixture. Overall, this study demonstrates that recurrent deep learning models combined with class-comparative attribution provide a framework for investigating spectral-temporal decision processes in forest classification, particularly where classes are spectrally and phenologically similar.

## REFERENCES

- [1] Lihui Zhong, Zhengquan Dai, Panfei Fang, Yong Cao, and Leiguang Wang. A Review: Tree Species Classification Based on Remote Sensing Data and Classic Deep Learning-Based Methods. *Forests*, 15(5):852, May 2024. ISSN 1999-4907. doi: 10.3390/f15050852. URL <https://www.mdpi.com/1999-4907/15/5/852>.
- [2] Johannes Stoffels, Sebastian Mader, Joachim Hill, Willy Werner, and Godehard Ontrup. Satellite-based stand-wise forest cover type mapping using a spatially adaptive classification approach. *European Journal of Forest Research*, 131(4):1071–1089, July 2012. ISSN 1612-4677. doi: 10.1007/s10342-011-0577-2. URL <https://doi.org/10.1007/s10342-011-0577-2>.
- [3] Jun Tan, Jing Li, Tianyue Ma, Xingguang Yan, and Ziyue Huo. Leveraging Sentinel-1/2 time series and deep learning for accurate forest tree species mapping. *Frontiers in Forests and Global Change*, 8:1599510, July 2025. ISSN 2624-893X. doi: 10.3389/ffgc.2025.1599510. URL <https://www.frontiersin.org/articles/10.3389/ffgc.2025.1599510/full>.
- [4] S. I. Seneviratne, X. Zhang, M. Adnan, W. Badi, C. Dereczynski, A. Di Luca, S. Ghosh, I. Iskandar, J. Kossin, S. Lewis, F. Otto, I. Pinto, M. Satoh, S. M. Vicente-Serrano, M. Wehner, and B. Zhou. Weather and Climate Extreme Events in a Changing Climate. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1513–1766. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.013.
- [5] Jürgen Bauhus, David I. Forrester, and Hans Pretzsch. Mixed-Species Forests: The Development of a Forest Management Paradigm. In Hans Pretzsch, David I. Forrester, and Jürgen Bauhus, editors, *Mixed-Species Forests*, pages 1–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 9783662545515 9783662545539. doi: 10.1007/978-3-662-54553-9\_1. URL [http://link.springer.com/10.1007/978-3-662-54553-9\\_1](http://link.springer.com/10.1007/978-3-662-54553-9_1).
- [6] Shigeo Yachi and Michel Loreau. Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *Proceedings of the National Academy of Sciences*, 96(4):1463–1468, February 1999. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.96.4.1463. URL <https://pnas.org/doi/full/10.1073/pnas.96.4.1463>.
- [7] European Commission. Joint Research Centre. *Climate change impacts and adaptation in Europe: JRC PESETA IV final report*. Publications Office, LU, 2020. URL <https://data.europa.eu/doi/10.2760/171121>.
- [8] Anna Iglseder, Markus Immitzer, Alena Dostálová, Andreas Kasper, Norbert Pfeifer, Christoph Bauerhansl, Stefan Schöttl, and Markus Hollaus. The potential of combining satellite and airborne remote sensing data for habitat classification and monitoring in forest landscapes. *International Journal of Applied Earth Observation and Geoinformation*, 117:103131, March 2023. ISSN 15698432. doi:

- 10.1016/j.jag.2022.103131. URL <https://linkinghub.elsevier.com/retrieve/pii/S1569843222003193>.
- [9] Michele Martone, Paola Rizzoli, Christopher Wecklich, Carolina González, José-Luis Bueso-Bello, Paolo Valdo, Daniel Schulze, Manfred Zink, Gerhard Krieger, and Alberto Moreira. The global forest/non-forest map from TanDEM-X interferometric SAR data. *Remote Sensing of Environment*, 205:352–373, February 2018. ISSN 00344257. doi: 10.1016/j.rse.2017.12.002. URL <https://linkinghub.elsevier.com/retrieve/pii/S0034425717305795>.
- [10] Markus Immitzer, Francesco Vuolo, and Clement Atzberger. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sensing*, 8(3):166, February 2016. ISSN 2072-4292. doi: 10.3390/rs8030166. URL <https://www.mdpi.com/2072-4292/8/3/166>.
- [11] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing*, 12(14):2291, July 2020. ISSN 2072-4292. doi: 10.3390/rs12142291. URL <https://www.mdpi.com/2072-4292/12/14/2291>.
- [12] Ewa Grabska, Patrick Hostert, Dirk Pflugmacher, and Katarzyna Ostapowicz. Forest Stand Species Mapping Using the Sentinel-2 Time Series. *Remote Sensing*, 11(10):1197, May 2019. ISSN 2072-4292. doi: 10.3390/rs11101197. URL <https://www.mdpi.com/2072-4292/11/10/1197>.
- [13] Lukas Blickensdörfer, Katja Oehmichen, Dirk Pflugmacher, Birgit Kleinschmit, and Patrick Hostert. National tree species mapping using Sentinel-1/2 time series and German National Forest Inventory data. *Remote Sensing of Environment*, 304:114069, April 2024. ISSN 00344257. doi: 10.1016/j.rse.2024.114069. URL <https://linkinghub.elsevier.com/retrieve/pii/S0034425724000804>.
- [14] Marco Wegler, Patrick Kacic, Frank Thonfeld, Stefanie Holzwarth, Niklas Jaggy, Ursula Gessner, and Claudia Kuenzer. Tree species from space: a new product for Germany based on Sentinel-1 and -2 time series. *International Journal of Remote Sensing*, 46(16):6075–6108, August 2025. ISSN 0143-1161, 1366-5901. doi: 10.1080/01431161.2025.2530236. URL <https://www.tandfonline.com/doi/full/10.1080/01431161.2025.2530236>.
- [15] Pan Liu, Chunying Ren, Zongming Wang, Mingming Jia, Wensen Yu, Huixin Ren, and Chenzhen Xia. Evaluating the Potential of Sentinel-2 Time Series Imagery and Machine Learning for Tree Species Classification in a Mountainous Forest. *Remote Sensing*, 16(2):293, January 2024. ISSN 2072-4292. doi: 10.3390/rs16020293. URL <https://www.mdpi.com/2072-4292/16/2/293>.
- [16] Abdulhakim M. Abdi and Fan Wang. Mapping forest tree species and their uncertainty using Earth observation and National Forest Inventory data: towards operational monitoring in Sweden. *International Journal of Remote Sensing*, 47(7):2912–2943, April 2026. ISSN 0143-1161, 1366-5901. doi: 10.1080/01431161.2026.2625513. URL <https://www.tandfonline.com/doi/full/10.1080/01431161.2026.2625513>.
- [17] Florian Mouret, David Morin, Milena Planells, and Cécile Vincent-Barbaroux. Tree Species Classification at the Pixel Level Using Deep Learning and Multispectral Time Series in an Imbalanced Context. *Remote Sensing*, 17(7):1190, March 2025. ISSN 2072-4292. doi: 10.3390/rs17071190. URL <https://www.mdpi.com/2072-4292/17/7/1190>.
- [18] Maximilian Freudenberg, Sebastian Schnell, Marko Henning, and Paul Magdon. Consistent Multi-Temporal National Scale Tree Species Maps for Germany 2018 – 2024, 2025. URL <https://www.ssrn.com/abstract=5705254>.
- [19] Haidi Abdullah, Andrew K. Skidmore, Roshanak Darvishzadeh, and Marco Heurich. Timing of red-edge and shortwave infrared reflectance critical for early stress detection induced by bark beetle (*Ips typographus*, L.) attack. *International Journal of Applied Earth Observation and Geoinformation*, 82:101900, October 2019. ISSN 15698432. doi: 10.1016/j.jag.2019.101900.
- [20] Maximilian Freudenberg, Sebastian Schnell, and Paul Magdon. A Sentinel-2 machine learning dataset for tree species classification in Germany. *Earth System Science Data*, 17(2):351–367, February 2025. ISSN 1866-3516. doi: 10.5194/essd-17-351-2025. URL <https://essd.copernicus.org/articles/17/351/2025/>.
- [21] Ernährung und Heimat Bundesministerium für Landwirtschaft. Bundeswaldinventur 2022, 2022. URL <https://www.thuenen.de/de/themenfelder/waelder/die-bundeswaldinventur/ergebnisse-der-bundeswaldinventur-2022>.
- [22] Markus Immitzer, Martin Neuwirth, Sebastian Böck, Harald Brenner, Francesco Vuolo, and Clement Atzberger. Optimal Input Features for Tree Species Classification in Central Europe Based on Multi-Temporal Sentinel-2 Data. *Remote Sensing*, 11(22):2599, November 2019. ISSN 2072-4292. doi: 10.3390/rs11222599. URL <https://www.mdpi.com/2072-4292/11/22/2599>.
- [23] A.M.C. Malkanthi, R.M.K.T. Rathnayaka, B.T.G.S. Kumara, and D.R. Welikanna. Forest Cover Classification with Explainable AI: A Comparative Analysis. In *2025 9th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, pages 1–6, Colombo, Sri Lanka, November 2025. IEEE. ISBN 9798331561239. doi: 10.1109/SLAAI-ICAI68534.2025.11318453. URL <https://ieeexplore.ieee.org/document/11318453/>.
- [24] Thünen-Institut. Bundeswaldinventur 2022 – Ergebnisdatenbank: Vorrat nach Land und Baumartengruppe, 2024. URL <https://bwi.info/inhalt1.3.aspx?Text=3.03%20Baumartengruppe&prRolle=public&prInv=BWI2022&prKapitel=3.03>.
- [25] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson Series in Artificial Intelligence. Pearson, Hoboken, NJ, fourth edition edition, 2020. ISBN 978-0-13-461099-3.
- [26] Maximilian Kirsch, Jakob Wernicke, Pawanjeet Singh Datta, and Christine Preisach. Early Detection of Forest

- Disturbances in Homogeneous Stands – Deep Learning-Based Monitoring of *Picea Abies*. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:22300–22316, 2025. ISSN 1939-1404, 2151-1535. doi: 10.1109/JSTARS.2025.3603094. URL <https://ieeexplore.ieee.org/document/11141690/>.
- [27] Yun-Long Kong, Qingqing Huang, Chengyi Wang, Jingbo Chen, Jiansheng Chen, and Dongxu He. Long Short-Term Memory Neural Networks for Online Disturbance Detection in Satellite Image Time Series. *Remote Sensing*, 10(3):452, March 2018. ISSN 2072-4292. doi: 10.3390/rs10030452.
- [28] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.8.1735.
- [29] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL <http://aclweb.org/anthology/D14-1179>.
- [30] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks, 2018. URL <https://arxiv.org/abs/1810.00825>.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. URL <https://arxiv.org/abs/1706.03762>.
- [32] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, 2017. URL <https://arxiv.org/abs/1708.08296>.
- [33] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 16(1):45–74, January 2024. ISSN 1866-9956, 1866-9964. doi: 10.1007/s12559-023-10179-8. URL <https://link.springer.com/10.1007/s12559-023-10179-8>.
- [34] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Gradient-Based Attribution Methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 169–191. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6\_9. URL [https://doi.org/10.1007/978-3-030-28954-6\\_9](https://doi.org/10.1007/978-3-030-28954-6_9).
- [35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, 2017. URL <https://arxiv.org/abs/1703.01365>.
- [36] Anupama Jha, Joseph K. Aicher, Matthew R. Gazzara, Deependra Singh, and Yoseph Barash. Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biology*, 21(1):149, December 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02055-7. URL <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02055-7>.
- [37] Soumya Sanyal and Xiang Ren. Discretized Integrated Gradients for Explaining Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.805. URL <https://aclanthology.org/2021.emnlp-main.805>.
- [38] Lorenzo Tronchin, Ermanno Cordelli, Lorenzo Ricciardi Celsi, Daniele Maccagnola, Massimo Natale, Paolo Soda, and Rosa Sicilia. Translating Image XAI to Multivariate Time Series. *IEEE Access*, 12:27484–27500, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3366994. URL <https://ieeexplore.ieee.org/document/10439172/>.
- [39] B. Schieber, R. Janík, and Z. Snopková. Phenology of four broad-leaved forest trees in a submountain beech forest. *Journal of Forest Science*, 55(1):15–22, January 2009. ISSN 12124834, 1805935X. doi: 10.17221/51/2008-JFS. URL <http://jfs.agriculturejournals.cz/doi/10.17221/51/2008-JFS.html>.
- [40] Katarzyna Ewa Lewińska, Olivier Hagolle, Ferran Gascon, Patrick Griffiths, Ulf Leser, and Patrick Hostert. Benefits of the Sentinel-2 mission triplet constellation – one year on. *Remote Sensing Applications: Society and Environment*, 44:102278, December 2026. ISSN 23529385. doi: 10.1016/j.rsase.2026.102278. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352938526004118>.
- [41] Roberto Verdecchia, Luis Cruz, June Sallou, Michelle Lin, James Wickenden, and Estelle Hotellier. Data-Centric Green AI An Exploratory Empirical Study. In *2022 International Conference on ICT for Sustainability (ICT4S)*, pages 35–45, Plovdiv, Bulgaria, June 2022. IEEE. ISBN 9781665482868. doi: 10.1109/ICT4S55073.2022.00015. URL <https://ieeexplore.ieee.org/document/9830097/>.

Table 5: Comparison of model configurations

| Model             | RF    | XGBoost | LSTM  |         | GRU          |              |
|-------------------|-------|---------|-------|---------|--------------|--------------|
|                   |       |         | MHA   | No Att. | MHA          | No Att.      |
| Accuracy          | 0.792 | 0.851   | 0.848 | 0.850   | <b>0.854</b> | 0.853        |
| Balanced accuracy | 0.890 | 0.925   | 0.930 | 0.931   | <b>0.933</b> | <b>0.933</b> |
| Macro F1          | 0.748 | 0.839   | 0.843 | 0.843   | <b>0.851</b> | 0.847        |

## A APPENDIX

### A.1 Equations for LSTM and GRU

$$f_t = \sigma(w_f x_t + u_f h_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma(w_i x_t + u_i h_{t-1} + b_i) \quad (5)$$

$$o_t = \sigma(w_o x_t + u_o h_{t-1} + b_o) \quad (6)$$

$$\tilde{C}_t = \tanh(w_{\tilde{C}} x_t + u_{\tilde{C}} h_{t-1} + b_{\tilde{C}}) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (8)$$

$$h_t = o_t \odot \tanh(C_t) \quad (9)$$

Where:

- The *forget gate* is denoted  $f_t$ , the *input gate*  $i_t$ , the *cell candidate gate*  $\tilde{C}_t$  and the *output gate*  $o_t$ .
- The input weights are represented as  $w$ , the recurrent weights for the Hidden State as  $u$  and the biases are denoted as  $b$ .
- $h$  is the hidden state,  $C$  is the cell state and  $x$  is the input data at time step  $t$ .
- $\odot$  denotes the point-wise multiplication of two vectors.
- $\sigma$  is the sigmoid activation function,  $\tanh$  is the tanh activation function.

For the GRU unit the gates are defined by the following functions [29]:

$$r_t = \sigma(w_r x_t + b_r + w_r h_{t-1} + b_r) \quad (10)$$

$$z_t = \sigma(w_z x_t + b_z + w_z h_{t-1} + b_z) \quad (11)$$

$$g_t = \tanh(w_g x_t + b_g + r_t \odot (w_g h_{t-1} + b_g)) \quad (12)$$

$$h_t = (1 - z_t) \odot g_t + z_t \odot h_{t-1} \quad (13)$$

Where:

- The *reset gate* is denoted by  $r_t$ , the *update gate* by  $z_t$ , and the *candidate hidden state* by  $g_t$ .

The sigmoid activation function is calculated as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (14)$$

The tanh activation function is calculated as:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (15)$$

### A.2 Additional Tables

Table 6: Hyperparameter search spaces for Random Forest and XGBoost models

| Model         | Hyperparameter       | Tested values            |
|---------------|----------------------|--------------------------|
| Random Forest | n_estimators         | 100, 250, 500, 750       |
|               | max_depth            | 3, 6, 10, 15             |
|               | min_samples_split    | 2, 5, 10                 |
|               | min_samples_leaf     | 1, 2, 5                  |
|               | max_features         | sqrt, log2, 0.5, 0.7     |
|               | bootstrap            | True                     |
|               | max_samples          | 0.5, 0.7, 0.9, None      |
|               | class_weight         | None, balanced           |
| XGBoost       | max_depth            | 3, 6, 10, 15             |
|               | learning_rate        | 0.01, 0.1, 0.2, 0.3, 0.4 |
|               | subsample            | 0.5, 0.6, 0.7, 0.8, 0.9  |
|               | colsample_bytree     | 0.5, 0.6, 0.7, 0.8, 0.9  |
|               | colsample_bylevel    | 0.5, 0.6, 0.7, 0.8, 0.9  |
|               | n_estimators         | 100, 250, 500, 750       |
|               | reg_alpha            | 0.1, 0.001, 0.00001      |
|               | reg_lambda           | 0.1, 0.001, 0.00001      |
| RNN-Models    | RNN                  | LSTM, GRU                |
|               | Attention            | MHA, None                |
|               | Bi-directional       | True                     |
|               | Bottleneck dimension | 32, 64, 128              |
|               | Hidden Dimension     | Bottleneck dimension×4   |
|               | dropout              | 0.2                      |