remote sensing

MDPI

Article

Evaluating Tree Species Mapping: Probability Sampling Validation of Pure and Mixed Species Classes Using Convolutional Neural Networks and Sentinel-2 Time Series

Tobias Schadauer [1,*], Susanne Karel [1], Markus Loew [1], Ursula Knieling [1], Kevin Kopecky [1], Christoph Bauerhansl [1], Ambros Berger [1], Stephan Graeber [1] and Lukas Winiwarter [2,3]

[1] Department of Forest Inventory, Austrian Research Centre for Forests (BFW), Seckendorff-Gudent-Weg 8, 1130 Vienna, Austria; susanne.karel@bfw.gv.at (S.K.); markusloew@gmx.at (M.L.); ursula.knieling@bfw.gv.at (U.K.); kevin.kopecky@bfw.gv.at (K.K.); christoph.bauerhansl@bfw.gv.at (C.B.); ambros.berger@bfw.gv.at (A.B.); stephan.graeber@bfw.gv.at (S.G.)

[2] Research Unit Photogrammetry, Department of Geodesy and Geoinformation, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria; lukas.winiwarter@uibk.ac.at

[3] Department of Basic Sciences in Engineering Sciences, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria

* Correspondence: tobias.schadauer@bfw.gv.at

Abstract: The accurate large-scale classification of tree species is crucial for the monitoring, protection, and management of the Earth's invaluable forest ecosystems. Numerous previous studies have recognized the suitability of satellite imagery, particularly Sentinel-2 imagery, for this task. In this study, we utilized a dense phenology Sentinel-2 time series, which offered consistent data across multiple granules, to map tree species across the entire forested area in Austria. Aiming for the classification scheme to more accurately represent actual forest conditions, we included mixed tree species and sparsely populated classes (classes with sparse canopy cover) alongside pure tree species classes. To enhance the training data for the mixed and sparse classes, synthetic data creation was employed. Autocorrelation has significant implications for the validation of thematic maps. To investigate the impact of spatial dependency on validation data, two methods were employed at numerous split and buffer distances: spatial split validation and a validation method based on a buffered ground reference probability samples provided by the National Forest inventory (NFI). While a random training data holdout set yielded 99% accuracy, the spatial split validation resulted in 74% accuracy, emphasizing the importance of accounting for spatial autocorrelation when validating with holdout sets derived from polygon-based training data. The validation based on NFI data resulted in 55% overall accuracy, 91% post-hoc pure class accuracy, and 79% accuracy when confusions in phenological proximity were disregarded (e.g., spruce–larch confused with spruce). The significant differences in accuracy observed between spatial split and NFI validation underscore the challenge for polygon-based training data to capture ground reference forest complexity, particularly in areas with diverse forests. This hardship is further accentuated by the pure class accuracy of 91%, revealing the substantial impact of mixed stands on the accuracy of tree species maps.

Keywords: large scale; forest diversity; satellite; ground reference forest data; spatial autocorrelation; synthetic training data

check for updates

Citation: Schadauer, T.; Karel, S.; Loew, M.; Knieling, U.; Kopecky, K.; Bauerhansl, C.; Berger, A.; Graeber, S.; Winiwarter, L. Evaluating Tree Species Mapping: Probability Sampling Validation of Pure and Mixed Species Classes Using Convolutional Neural Networks and Sentinel-2 Time Series. *Remote Sens.* **2024**, *16*, 2887. https://doi.org/10.3390/rs16162887

Academic Editors: Jan Komarek, Marlena Kycko and Iñigo Molina

Received: 21 June 2024
Revised: 26 July 2024
Accepted: 30 July 2024
Published: 7 August 2024

1. Introduction

Forests are important for a wide range of ecological, economic, and social reasons. They play a vital role in maintaining the balance of our planet's ecosystems and support biodiversity. They provide climate and water regulation and serve as natural shields, protecting against natural hazards such as soil erosion, floods, landslides, rock falls, and avalanches. Forests offer significant economic benefits, including the production of timber and non-timber forest products, as well as the growth of industries such as tourism and recreation. Overall, forests are essential for preserving the health of our planet, combating climate change and ensuring the well-being of human societies [1–9].

Earth observation (EO) data provide valuable advantages for understanding forests, including large-scale coverage, timeliness, multi-sensor data fusion, non-invasiveness, and cost-effectiveness. The data enable the creation of detailed forest maps with high spatial and temporal resolution, providing insights into forest structure, health, and dynamics. Such insights are crucial in the face of increasing environmental challenges. Climate change is altering habitat suitability for many tree species [10–13], making forests more vulnerable to various threats. For example, pests like *Ips typographus* and pathogens such as *Diplodia pinea* are causing increasing damage to forests in Austria and other regions [10,14–17]. EO data facilitate the creation of detailed maps of species distribution and abundance [18], which are valuable for monitoring changes in forest composition, identifying areas at risk, and informing management strategies. Such information can aid in developing targeted approaches to mitigate the impacts of climate change and biological threats on forest ecosystems [19]. Within the realm of EO data, the availability of the Sentinel-2 (S2) EO mission data has revolutionized the field of remote sensing for tree species classification. With its broad spectral coverage and high spatial and temporal resolution, the S2 data provide an excellent basis for research on tree species classification using remote sensing [18]. Many studies have successfully combined S2 data with machine learning techniques to classify pure tree species stands on small to medium-sized areas at the pixel level, demonstrating the potential of this data source for advancing our understanding of forest ecosystems [20–34].

Multiple studies have highlighted the advantages of using vegetation indices in tree species classification [23,27,30,32–34]. Additionally, incorporating digital terrain model (DTM) data has been shown to improve the accuracy of classification models [13,18,20,28–30]. Multi-temporal imagery, which captures seasonal changes in vegetation cover, has also positively impacted tree species classification models [22–25,30,34].

Expanding the study areas introduces new challenges, particularly when multiple Sentinel-2 granules (the tiling units of S2 products, see [35]), are needed, due to the difficulty of obtaining temporally matched scenes across the entire area. Large-scale studies are scarce. However, for instance, the authors of [33] used a dense phenology time series to acquire cohesive data across multiple granules, while in [36], monthly composites were generated to map tree species across Germany's entire forested area.

The earlier mentioned studies and many others have laid a strong foundation for further research on tree species classification. However, forest structures are often heterogeneous, consisting of multiple tree species mixed in small areas. Consequently, a single S2 pixel can encompass signals from multiple tree species, and the training and validation data consisting only of pure classes do not accurately represent the forest on the mapped area. While very few studies have addressed the issue of mixed pixels, the authors of [37] modelled a vector of the proportions of tree species on a pixel basis. Additionally, Waser et al. [38] found significant differences between validation using training data holdout sets in comparison with data from an independent source, especially in areas with a high degree of mixture.

In the evaluation of ecological maps produced by machine learning techniques, spatial autocorrelation, which is inherent to the data used, poses a central challenge. As highlighted in a review by [39], neglecting the underlying structure in input data during standard cross-validation can lead to the significant overestimation of a model's predictive power. The study suggested that a thoughtful blocking strategy can mitigate this issue; however, it may also lead to an overestimation of interpolation errors. The study conducted by [40] convincingly illustrated this validation data issue in a large-scale aboveground forest biomass case study. A recent study by [41] compared validation approaches, including validation based on independent probability sampling (IPS), standard (random) cross-validation (CV), and spatial cross-validation (SCV), when the training data were collected in clustered and non-clustered random patterns. The results indicated that CV performed comparably to IPS in non-clustered cases, while SCV tended to overestimate the root mean squared error (RMSE). However, with clustered training data, as is common in tree species classification studies, CV tended to underestimate the RMSE. Few tree species classification studies have addressed this issue; however, ref. [30] implemented a spatial cross-validation scheme by dividing the area covered by the reference dataset into squares of 10 by 10 km.

Previous studies have primarily focused on traditional machine learning methods such as random forest, support vector machine, or extreme gradient boosting to classify tree species. The recent research studies by [31] has shown that deep learning and convolutional neural networks (CNNs) outperform traditional techniques. Bolyn et al. [37] successfully applied U-Nets, spatially aware CNN structures, originally developed by [42] for pixelwise image segmentation, to tree species classification. In addition to capturing spatial features, CNN structures can also be leveraged to analyze the temporal dimension of the Sentinel-2 data. A review by [43] highlighted a CNN structure specifically designed for time series classification, featuring one-dimensional convolutions and residual connections [44], as the top-performing deep learning approach across various time series classification tasks. Furthermore, Xi et al. [31] found that a CNN structure based on one-dimensional convolutions outperformed other approaches in their tree species classification study.

We employ neural networks with one-dimensional convolutions and residual connections, along with dense phenology index time series, statistical parameters, DTM, and vegetation height data to predict tree species on a national scale at the S2 pixel level. Our study's overarching objective is to bring large-scale tree species classification closer to the ground reality of forests, a perspective that has received limited attention in previous studies. This is accomplished by enhancing the capture of the forests' complexity and variability in both training and validation.

The primary objectives of our study are as follows:

(1) to expand the set of pure tree species classes by incorporating mixed classes—each consisting of two pairwise different species—and sparse classes with low canopy cover into training and validation;

(2) to evaluate the generated tree species maps through a probability sample-based validation, with a specific emphasis on investigating spatial autocorrelation.

We introduce the following innovative ideas to tree species classification:

- the use of a dense time series from multi-annual S2 data developed by [36], providing phenology index data at the S2 pixel level that are consistent across S2 granules;
- the inclusion of mixed species classes in training and validation data, specific validation metrics for mixed and pure species classes, and training data synthesis;
- a hybrid neural network architecture tailored specifically for the combination of time series and non-time series features, based on [43];
- A spatial split and an independent probability sample-based validation including a comprehensive and innovative spatial autocorrelation analysis for both.

2. Materials and Methods

To commence this section, we offer a concise overview, including a workflow diagram (see Figure 1), along with the motivations driving the subsequent sections. In Section 2.1, we introduce the study area followed by a delineation of the forested area in Section 2.2. Sections 2.3–2.5 delve into the datasets covering the entire study area, which serve as input features for the classification models.

Given that the 10 by 10 m Sentinel-2 pixels often contain more than a single species, mapping tree species faces additional challenges. Following a brief definition of some necessary terms (Section 2.6), we explore the intricacies of classifying not only pure but also mixed and sparse classes and define the classes we aim to separate within the scope of this study (Section 2.7). Moving forward, we address the labeling of data for training and validation purposes. Section 2.8 presents the details of training data acquisition. However, these data present certain challenges, primarily because mixed-class data contain pixels with single species as well as mixed-species pixels. In Section 2.9, we propose the creation of synthetic data to tackle this issue by averaging multiple pixels. After the creation of the training data, Section 2.10 outlines how the national forest inventory (NFI) data were utilized to generate a validation dataset. However, a significant concern arises regarding spatial autocorrelation in training and validation data. Mishandling this issue can lead to overly optimistic estimates of classification accuracies. In Section 2.11, we discuss the strategies to analyze and address spatial autocorrelation, aiming for independence of validation samples and thereby improving the reliability and accuracy of our validation procedures. In Section 2.12, we detail the neural network architecture utilized in this study, while Section 2.13 provides an overview of the training process for the models. Subsequently, in Section 2.14, we explore the validation metrics applied and introduce accuracy measures specifically tailored for the validation of pure and mixed classes. Furthermore, we outline the different validation datasets and explain how the validation results are gathered over multiple training runs. Finally, in the last part of this section (Section 2.15), we present the training data and model configurations employed in our study. The software implementation for data analysis, preprocessing, and other methods was carried out using Python 3.11.5.

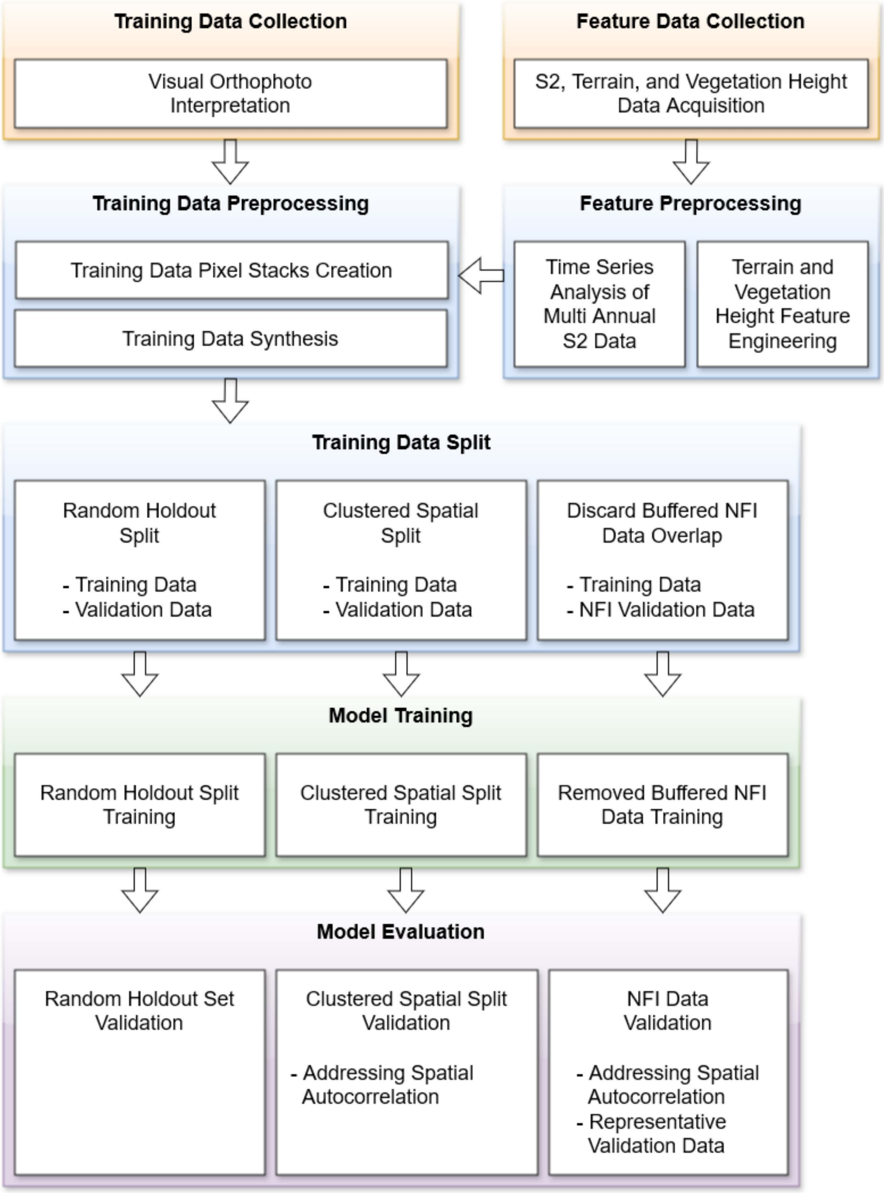
Training Data Collection
Visual Orthophoto Interpretation
Feature Data Collection
S2, Terrain, and Vegetation Height Data Acquisition
Training Data Preprocessing
Training Data Pixel Stacks Creation
Training Data Synthesis
Feature Preprocessing
Time Series Analysis of Multi Annual S2 Data
Terrain and Vegetation Height Feature Engineering
Training Data Split
Random Holdout Split
- Training Data
- Validation Data
Clustered Spatial Split
- Training Data
- Validation Data
Discard Buffered NFI Data Overlap
- Training Data
- NFI Validation Data
Model Training
Random Holdout Split Training
Clustered Spatial Split Training
Removed Buffered NFI Data Training
Model Evaluation
Random Holdout Set Validation
Clustered Spatial Split Validation
- Addressing Spatial Autocorrelation
NFI Data Validation
- Addressing Spatial Autocorrelation
- Representative Validation Data

Figure 1. Tree species mapping workflow.

2.1. Study Area

The republic of Austria is a landlocked country in Central Europe situated between latitudes 46° and 49°N and longitudes 9° and 18°E with a total land area of 83,879 km^2. More than 70% of the country's federal territory comprises mountains, including sections of the Central Eastern Alps. The altitude ranges between 114 and 3798 m above sea level. Austria predominantly lies within the cool/temperate climate zone, with an oceanic type climate prevailing in the western and northern regions, characterized by humid westerly winds. In the east, the climate is mostly Pannonian–continental, with low precipitation, hot summers, and cold winters [45–47]. The study area encompassed the entire forested area of the federal territory of Austria, accounting for 47.9% of the total land area (according to the national forest area definition, which differs from the Forest and Agriculture Organization of the United Nations definition by including other wooded lands). The Austrian forest is diverse with over 60 tree species. The species mainly occurring are spruce *(Picea abies*) 41.8%, beech (*Fagus sylvatica*) 9.5%, larch *(Larix decidua*) 4.4%, white pine (*Pinus sylvestris*) 3.4%, fir (*Abies alba*) 2.2%, mountain pine (*Pinus mugo*) 2.2%, oak (*Quercus*) 1.7%, green alder (*Alnus viridis*) 0.9%, arolla pine (*Pinus cembra*) 0.7%, and black pine (*Pinus nigra*) 0.4%. According to the specific National Forest inventory (NFI) evaluations, the forest area is composed of approximately 32% pure stands (where a single species comprises more than 90%), 47% mixed stands, and 21% areas without trees (temporary unstocked forest areas and shrub areas). The tree coverage is composed of 71% dense (no gaps in the canopy) and 29% sparse canopies.

2.2. Forest Area Map

The Austrian Research Center for Forests, Natural Hazard, and Landscape (BFW) routinely publishes a vectorized forest area map derived through a semi-automated process that utilizes aerial photography data, which are sourced from a flight campaign with updates every three years. The derived forest area map was rasterized and aligned with the S2 granules.

2.3. S2 Phenology Features

Our study uses an interpolated dense S2 time series with seamless S2 granule transitions as phenology features. S2 is a passive multi-spectral imaging mission with 13 spectral bands and a revisit frequency of 5 days at the equator. The S2 products are made freely available via the Copernicus Data Space Ecosystem (https://dataspace.copernicus.eu/, accessed on 19 July 2024) by the European Space Agency. The time series analysis tool (TSA) developed by [36] produces a time series for spectral indices (Table 1) on a single pixel level. The TSA utilizes Level-1C multi annual data and discards areas affected by snow, clouds, and cloud shadows. Additionally, a threshold-based filter is used to exclude outliers. The filtered data points, spanning over multiple years, are combined to create a synthetic single-year phenology dataset. Next, a Savitzky–Golay trajectory [48] is fitted to this dataset to obtain a smooth modeled main phenology course (MPC) on a single pixel level.

Table 1. Spectral indices based on S2 bands B2 (490 nm), B3 (560 nm), B4 (665 nm), and B8 (842 nm) (see [35]) used as phenology features.

Spectral Indices	Acronym	Formula	References
Atmospherically Resistant Vegetation Index	ARVI	$\frac{B8-2*B4+B2}{B8+2*B4+B2}$	[49,50]
Band 8 Near Infra-Red	BNIR	$\frac{B8}{5500}$	Developed within this study
Dark Area Vegetation Near Infra-Red	DAVNIR	$\frac{B8-B4}{B8+B4} * \log(B8) * \frac{log(B2+B3+B4)}{60}$	Developed within this study

Table 1. *Cont.*

Spectral Indices	Acronym	Formula	References
Green Normalized Difference Vegetation Index	GNDVI	$\frac{B8-B3}{B8+B3}$	[51]
Green Share	GREEN_SHARE	$\frac{B3}{B2+B3+B4}$	Developed within this study
Visual Reflectance Absence Index	VRAI	$1-\frac{B2+B3+B4}{7000}$	Developed within this study

In addition, the vitality and seasonality metrics were extracted from the MPC. Basic MPC statistics and a percentile transition analysis (PTA) as presented by [52] were used to derive the descriptive metrics such as greening day of the year (DOY), defoliation DOY, the vegetation period, maximum/minimum value and the corresponding DOY, and first reach, last pass, and a set of percentile values.

The MPC and PTA metrics were calculated for the spectral indices ARVI (Atmospherically Resistant Vegetation Index) [49,50] (with γ = 1), BNIR (Band 8 Near Infra-Red), DAVNIR (Dark Area Vegetation Near Infra-Red), GNDVI (Green Normalized Difference Vegetation Index) [51], GREEN_SHARE (Green Share), and VRAI (Visual Reflectance Absence Index) using data from the years 2017–2021 (Table 1). The BNIR, DAVNIR, GREEN_SHARE, and VRAI indices were developed specifically for this study.

The phenology input features for this study's classification models included the MPC between DOY 100 and DOY 270 at a temporal resolution of 5 days for each of the indices listed in Table 1, as well as the phenological metrics listed in Appendix A Table A1.

2.4. Digital Terrain Model Features

Point clouds from airborne laser scanning (ALS) were filtered to last returns only, before interpolating them to a DTM of cell size 1 m. The resulting DTM dataset was provided by the Austrian Federal Ministry of Agriculture, Forestry, Regions and Water Management (BML). These data were collected in a decentralized manner by individual federal states during flight campaigns spanning from 2003 to 2018. A slope map, along with southness (see Equation (1)) and eastness (see Equation (2)) of the slope aspect were derived from the DTM. These data were resampled to a 10 m resolution, aligned with the S2 granules, and used as additional input data for this study's classification models.

$$southness =: \begin{cases} \frac{\alpha}{180}, & 0 \leq \alpha \leq 180 \\ \frac{360-\alpha}{180}, & 180 < \alpha \leq 360 \end{cases} \tag{1}$$

$$eastness =: \begin{cases} \frac{\alpha+90}{180}, & 0 \leq \alpha \leq 90 \\ \frac{270-\alpha}{180}, & 90 < \alpha \leq 270 \\ \frac{\alpha-270}{180}, & 270 < \alpha \leq 360 \end{cases} \tag{2}$$

2.5. Normalized Digital Surface Model

A digital surface model (DSM) with a spatial resolution of one meter was generated from the digital aerial imagery via image matching [53] using the ApplicationsMaster 13.1 (now Trimble Inpho) software, in particular the Match-T DSM Commander module, by Trimble, Inc. (Westminster, CO, USA) [54]. The generated 3D point clouds were subjected to outlier detection and subsequently used to create raster models by selecting the maximal value per cell at one meter resolution. Subtracting the DTM from the DSM generated the normalized digital surface model (NDSM), which contains object heights measured from the ground level.

The mean, standard deviation, 5-percentile, 95-percentile, and range (maximum–minimum) of the NDSM were computed on the 10 m S2 resolution. These resulting data were also used as additional input features for the tree species classification models.

2.6. Definitions—Feature Stacks and Feature Space

For each S2 pixel in the study area, a concatenation of input features resulted in a feature stack. These feature stacks (features) consisted of 210 phenology index time series features, 684 phenology metric features, 4 DTM features, and 5 NDSM features. Each of these feature stacks can be viewed as a vector in a 903-dimensional vector space, referred to as the feature space. The component of such a vector corresponding to a specific feature is referred to as the feature dimension. The statistical distribution of features in the feature space for a given dataset is referred to as distribution. For instance, the training data distribution pertains to how the feature stacks are distributed in the feature space, encompassing all the pixels within the training data.

2.7. Pure, Mixed, and Sparse Classes

Due to the considerable spatial diversity of forests and the 10 m resolution of S2 pixels, a single pixel may contain the spectral information from multiple tree species [37]. To address this issue, we introduced the use of mixed classes, which are composites of two species (e.g., spruce–beech), to improve the classification scheme's ability to represent the complexity and spatial variety in the forest.

We included eight pure classes for the most common tree species, namely spruce, larch, black pine, white pine, mountain pine, beech, oak, and green alder. To account for the remaining deciduous species, the "other deciduous" class was introduced. Furthermore, the eleven mixed classes spruce–fir, spruce–larch, spruce–white pine, spruce–arolla pine, larch–arolla pine, spruce–beech, spruce–deciduous, larch–deciduous, white pine–oak, white pine–deciduous, and black pine–deciduous and the four sparse classes, spruce sparse, larch sparse, white pine sparse, and deciduous sparse were added to the classification scheme. Finally, we included a low vegetation class for vegetation below 4 m in height, resulting in a total of 25 classes in the classification scheme (see Table 2).

Table 2. Training pixels and areas per class.

Class	Number S2 Pixels	Area (km^2)	Number Areas
Spruce	62,066	6.2	391
Spruce–fir	12,590	1.3	98
Spruce–larch	103,181	10.3	341
Spruce–white pine	26,933	2.7	180
Spruce–arolla pine	4129	0.4	55
Spruce–beech	21,918	2.2	201
Spruce–deciduous	27,411	2.7	148
Spruce sparse	7426	0.7	86
Larch	24,193	2.4	237
Larch–arolla pine	16,885	1.7	91
Larch–deciduous	10,269	1.0	102
Larch sparse	11,811	1.2	129
White pine	25,643	2.6	222
White pine–oak	9763	1.0	27
White pine–deciduous	22,315	2.2	113
White pine sparse	8972	0.9	20
Black pine	27,639	2.8	43

Table 2. *Cont.*

Class	Number S2 Pixels	Area (km^2)	Number Areas
Black pine–deciduous	8537	0.9	48
Mountain pine	8274	0.8	205
Beech	16,729	1.7	115
Oak	31,917	3.2	104
Green alder	1310	0.1	44
Other deciduous	40,052	4.0	138
Deciduous sparse	6895	0.7	63
Low vegetation	31,160	3.1	404
Sum	568,018	56.8	3605

2.8. Training Data Labeling

Labeled training areas were obtained by visually interpreting orthophotos, identifying single-class areas, and marking them with polygons. These areas met the following quality requirements:

- spatial disjointedness from the NFI validation dataset (see Section 2.10);
- at least 90% target class composition;
- a minimum size of 3000 m^2;
- for mixed classes, homogeneous mixing of both tree species;
- for sparse classes, homogeneous mixing of canopy cover and ground-level area.

It is worth mentioning that, in several cases, the NFI-VD was used as a visual baseline on the orthophotos to identify nearby patches of the same species, introducing spatial dependency between the training data and the NFI-VD.

The labeled training dataset was obtained by intersecting the labeled training data areas with the S2 aligned features, resulting in a feature stack for each labeled pixel. Any pixels containing empty (no-data) values in their feature stack were discarded.

2.9. Synthetic Training Data

Mixed class training areas typically include both pure pixels, containing a single species (a constituent of the mixed class), and mixed pixels. These pure pixels negatively impact the pixel label quality, as they are labeled with the mixed class. To address this issue we applied training data synthesis [55] to all mixed and sparse classes. New synthetic pixels were created by randomly selecting two pixels from a training area and taking their mean in every feature dimension. This process was repeated until the original number of pixels of the area was reached.

2.10. NFI Validation Data

The Austrian NFI conducts a systematic grid-based field inventory campaign on a six-year cycle gathering data from approximately 22,000 observational plots on about 200 parameters [56]. These observation plots span an area of around 300 m^2 (corresponding to a circle with a radius of 9.77 m) each and are organized into around 5500 clusters. These clusters are evenly spaced with a distance of 3.89 km in a grid that spans the entire federal territory of Austria. Each cluster comprises four plots arranged in a square with a side length of 200 m. The top-level vegetation on each plot is characterized by parameters for the predominant species, secondary species, and potential admixtures.

To create a labeled NFI validation dataset (NFI-VD) for this study, the plots situated entirely within the forest area were selected, totaling around 9200. These plots were labeled based on the predominant and secondary species parameters. Following the tree species classes defined in this study (Section 2.7), approximately 200 plots that did not fit the classification scheme were discarded. Subsequently, the radius of the remaining 9000 plots was extended to a full 10 m (approx. 314 m^2). The resulting dataset was then overlayed with the 10 m resolution S2 tiles (see [35]; the center of a pixel needed to be within the circular observation plot), resulting in approximately 27,000 class-labeled pixels.

It should be noted that the NFI top-level vegetation description does not differentiate between black pine and white pine. Therefore, these two classes were merged into the "pine" class for NFI validation purposes. Furthermore, the class "low vegetation" was not present in the NFI-VD-class scheme.

2.11. Investigating and Addressing Autocorrelation in Training and Validation Data

Autocorrelation significantly impacts the thematic map validation due to the similarity of spatially close points, leading to overestimated predictor accuracies. To address this issue, in the following sections, we created clustered training data spatial splits and buffered the NFI validation data, removing training data within the buffered areas.

2.11.1. Clustered Spatial Splits

To mitigate the spatial autocorrelation between the training data and the training data holdout set, a random clustered spatial split (CSS) strategy ([41]) on training data holdout sets was employed. The training data were initially organized into clusters, so that for each class, two distinct clusters were at least a variable split distance apart from each other. For a spatial split holdout set, clusters were randomly sampled until at least 5 percent of all training data pixels were added to the holdout set. If the holdout set contained all classes, it was retained, otherwise it was discarded, and the process was restarted. To gain an insight into the relationship between the split distance and holdout set accuracy, this iterative procedure was performed for 10 different random seeds and split distances ranging from 125 to 5000 m, in increments of 125 m.

2.11.2. Buffered NFI Validation Data

To address the autocorrelation between the training data and NFI-VD (see Section 2.8), a buffering strategy on NFI plots was employed. A direct implementation of this approach, given the dense grid of NFI plots, would have led to the removal of a substantial amount of training data. To mitigate this, we opted for a staged approach, as follows:

- The NFI-VD plots were divided into 10 folds, denoted as $VF_1, \ldots, VF_{10}$, with classes almost evenly distributed amongst them. This division was accomplished by first determining the number of plots n_{cl} for each class cl. For each fold VF_i, random observational plots (not already assigned to another fold) were selected. If the class of a selected plot cl was not yet represented by more than $\frac{n_{cl}}{10}$ in VF_i, the plot was added to the fold. Additionally, the other plots from the same NFI cluster (each comprising four NFI plots) were included in the fold.
- For each NFI-VD fold, training areas contained within any buffered NFI plot from the respective fold were eliminated. This process yielded 10 sets of training data, denoted as $TD_1, \ldots, TD_{10}$, each corresponding to an NFI validation data fold.
- For each set of training data TD_i, a model m_{TD_i} was trained.
- Subsequently, each model m_{TD_i} was validated using the corresponding NFI validation data fold VF_i.
- Finally, the validation results over all folds were evaluated.

To explore the relationship between the buffer distance and NFI-VD accuracy, split distances ranging from 250 to 20,000 m with increments of 250 m (from 250 to 5000 m) and increments of 2500 m (from 7500 to 20,000 m) were chosen. Based on these parameters, two series of experiments were conducted. The first involved discarding training data based on buffer distance, as described earlier. In the second series, the discard rate was maintained at a constant level up to a buffer distance of 15,000 m. This was achieved by discarding training areas within the current buffer distance individually for each split and class and then randomly discarding training areas until the discard percentage for the current split and class matched the discard rate at 15,000 m buffer.

2.12. Neural Network Architecture

In this sub-section, we draw upon the foundational definitions in neural network literature, particularly the insights provided by [57], to establish a common understanding of the key concepts. Our study adopted a hybrid neural network architecture that combined the ResNet architecture proposed by [44], with downscaled numbers of filter channels for the convolutional layers, and a multilayer perceptron (MLP). The neural network models were implemented within the scope of this study using the PyTorch library (version 2.0.1), particularly the torch.nn module for constructing and training deep learning architectures.

The time series classification ResNet architecture applied in this study consists of three consecutive blocks each comprising multiple convolutional, normalization, and activation layers. The convolutional layers scan the time series data within a moving window extracting temporal patterns. The normalization layers standardize the output of the convolutional layers, improving training stability. Lastly, the activation layers introduce non-linearity to the network, enabling it to learn complex (non-linear) relationships within the data. With each block, a residual connection allows for the direct propagation of data from the input to the block, to the stage immediately preceding the final activation function within the block.

Figure 2 depicts the detailed flow of data in this study's MLP-ResNet hybrid structure. The phenology time series data for each index are fed into the first ResNet block. Within each block, one-dimensional convolutional layers are applied with parameters described as "conv(8 × 1, 6, 16)" where "8 × 1" represents the filter size (8 wide and 1 high), "6" represents the number of input channels, and "16" represents the number of output channels. Each convolutional layer is followed by a batch normalization layer with a batch size of 128 (which is the same for all batch normalization layers), and a ReLU activation function. After the third block, a pooling layer averages along the time-axis resulting in 32 output features of the ResNet. These 32 output features together with the non-time series input features are then fed into an MLP consisting of five liner layers (input and output dimensions given within the brackets), each followed by a batch normalization layer and a LeakyReLu [58] activation function, except for the very last one, which is trailed by a softmax layer, resulting in class probabilities.

2.13. Neural Network Training

The training data were divided into 95% training data (~540,000 pixels) and 5% holdout set validation data (~28,000 pixels). When training models to be validated with NFI-VD, the holdout set was a simple random sample, otherwise it was selected as a CSS, as described in Section 2.11.1. Before training and inference, the data were standardized by $(x - \mu)/\sigma$, where x is the feature vector, μ is the vector of feature means, and σ is the vector of feature standard deviations. Both μ and σ were calculated using forest data of the complete study area. Following rigorous experimentation to optimize the training parameters, this study employed the following settings: conducting neural network training over nine epochs, employing a batch size of 128, and configuring the learning rate to 0.00025 (see [57] for definitions). During training, the performance of the model was evaluated on a holdout set to detect any signs of overfitting with respect to the training data. As a training criterion, the cross-entropy loss was used along with the Adam optimizer [59] and a learning rate scheduler to reduce the learning rate during loss-plateaus (see [57] for definitions). The models were trained on a NVIDIA GeForce RTX 3070 Ti (Nvidia Corp., Santa Clara, CA, USA) with training times ranging from 10 to 30 min. Inferring the models on the complete study area took about 20 h.

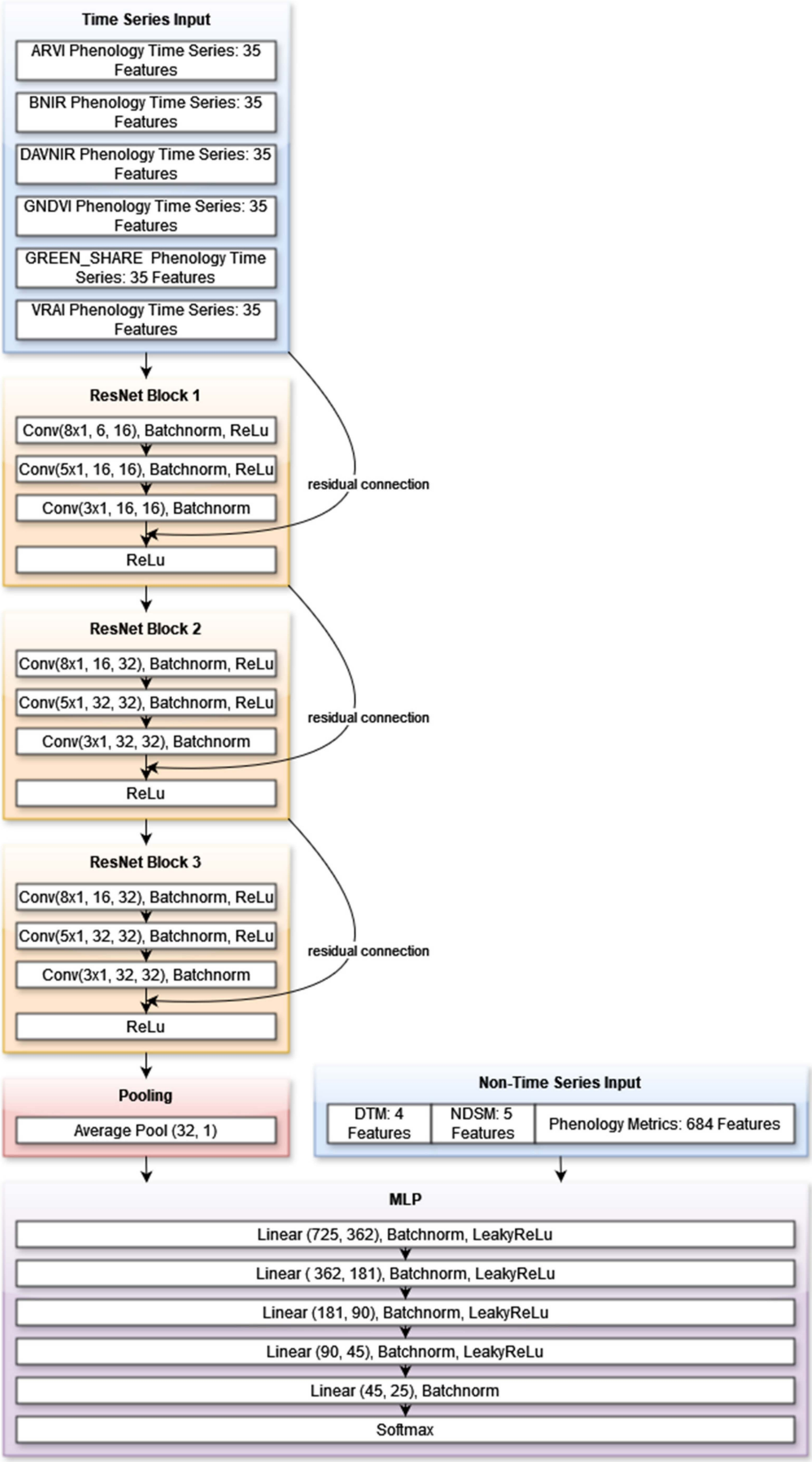
Time Series Input
ARVI Phenology Time Series: 35 Features
BNIR Phenology Time Series: 35 Features
DAVNIR Phenology Time Series: 35 Features
GNDVI Phenology Time Series: 35 Features
GREEN_SHARE Phenology Time Series: 35 Features
VRAI Phenology Time Series: 35 Features
ResNet Block 1
Conv(8x1, 6, 16), Batchnorm, ReLu
Conv(5x1, 16, 16), Batchnorm, ReLu
Conv(3x1, 16, 16), Batchnorm
ReLu
residual connection
ResNet Block 2
Conv(8x1, 16, 32), Batchnorm, ReLu
Conv(5x1, 32, 32), Batchnorm, ReLu
Conv(3x1, 32, 32), Batchnorm
ReLu
residual connection
ResNet Block 3
Conv(8x1, 16, 32), Batchnorm, ReLu
Conv(5x1, 32, 32), Batchnorm, ReLu
Conv(3x1, 32, 32), Batchnorm
ReLu
residual connection
Pooling
Average Pool (32, 1)
Non-Time Series Input
DTM: 4 Features
NDSM: 5 Features
Phenology Metrics: 684 Features
MLP
Linear (725, 362), Batchnorm, LeakyReLu
Linear (362, 181), Batchnorm, LeakyReLu
Linear (181, 90), Batchnorm, LeakyReLu
Linear (90, 45), Batchnorm, LeakyReLu
Linear (45, 25), Batchnorm
Softmax

Figure 2. MLP-ResNet hybrid schematic architecture.

2.14. Tree Species Map Validation

A multitude of validation metrics were computed to assess the resulting tree species maps, as follows:

- Overall accuracy (OA) for the NFI-VAL and NFI-weighted OA (NFI-w-OA) during training (the holdout set OA was weighted by NFI-Class-Distribution).
- Overall misclassification score (OMS) for the NFI-VAL: A score calculated based on the severity, judged by the phenological similarity, of all misclassifications in the confusion matrix. A score of 1.00 is the best possible value and means that all predictions are correct. See Appendix B for details.
- Prediction in close phenological proximity (PCPP) for the NFI-VAL: Predictions where at least one of the involved classes is correctly predicted. Examples include pixels predicted as spruce–fir but validated as spruce, pixels predicted as spruce–larch but validated as larch–arolla pine, pixels predicted as pine–oak but validated as oak, and pixels predicted as spruce–beech but validated as spruce–deciduous.
- Deciduous and coniferous confusions (DCC) for the NFI-VAL: Confusions between pure coniferous and pure deciduous classes. Examples include pixels predicted as larch but validated as oak and pixels predicted as beech but validated as spruce–larch.
- Post hoc pure class overall accuracy (POA), determined by eliminating all non-pure-class entries from the confusion matrix.
- Post hoc mixed class overall accuracy (MOA), calculated by eliminating all non-mixed-class entries from the confusion matrix.
- Macro-averaged (each class was weighted equally) F1 score (MAF1).
- F1 scores, producer and user accuracies, and misclassification scores on a class level.
- Confusion matrices for splits as well as aggregated confusion matrices over multiple splits.

The OA for the NFI validation, as a measure, closely aligns with the NFI-w-OA observed during training, given that the NFI validation data are a representative sample of the Austrian forest. In all validation methods, the sparse classes were merged with their respective base class—spruce-sparse with spruce, for instance.

The classification models were assessed using three different data sources: random training data holdout sets, CSS (see Section 2.11.1) training data holdout sets, and NFI-VD (see Section 2.11.2).

2.14.1. Random Holdout Set Validation

The randomly selected pixels from the entire training dataset formed a random training data holdout set for model validation.

2.14.2. Clustered Spatial Split Validation

To assess a model configuration with CSS training data holdout sets (CSS-VAL), 10 separate CSS were generated. Models were trained and validated in three distinct runs for each CSS fold. Subsequently, the means and standard deviations of validation metrics and an aggregated confusion matrix were computed from these 30 validations. See Section 2.11.1 for details on CSS.

2.14.3. NFI Data Validation

The merging of black and white pine classes (see Section 2.10) as well as the combination of sparse and base classes (see the beginning of this section) led to an NFI validation (NFI-VAL) scheme, comprising 19 classes.

To validate a model configuration with NFI-VAL, a model was trained on each of the 10 reduced training datasets, as detailed in Section 2.11.2, the results were aggregated, and means and standard deviations were calculated.

To address the spatial inaccuracies of NFI-VD as well as the alignment of the S2 data, nine positional variants were considered for each NFI plot. One variant was centered precisely at the given coordinates, while for four variants the position was shifted one

S2 pixel to the north, east, south, and west, respectively. For the remaining four variants, the position was shifted by one Sentinel-2 pixel in two contiguous directions, resulting in diagonal shifts. Out of these nine positional variants, the ones with the most pixels within the forest area were selected, and then the one with the best validation match was considered for the final validation result.

2.15. Training Data and Model Configurations

In this section, the data and model architecture configurations employed in the following results section are delineated. These configurations were selected after an extensive series of experiments that explored various aspects, including varying features, incorporating synthetic data, adjusting model capacities, and modifying training and architecture parameters. From this series, we aim to spotlight the configurations that yielded the most noteworthy results.

The base model's configuration encompassed all features outlined in Sections 2.3–2.5. It integrated the synthetic training data for the mixed and sparse classes (see Section 2.9), and the model's architecture is as described in Section 2.12. The training data were standardized employing means and standard deviations computed from all forest data, and the training used the parameters detailed in Section 2.13.

The subsequent models utilized the base model's configuration, with specific variations as explicitly described (see Table 3 for architecture details), as follows:

- The no_syn model used the raw training data without any synthetic data for mixed or sparse classes.
- The res_*xx*_*yy*_*yy* models were built with a with *xx* filters in the first block of the residual network and *yy* filters in the second and third blocks. Additionally, the sizes of the trailing MLP layers were adjusted.

Table 3. Model architecture configurations.

Model	Resnet Blocks Filters	MLP Layers (in/out Dimensions)	Parameters
res_64_128_128	64, 128, 128	821, 410, 205,102, 51, 25	955,324
base	16, 32, 32	725, 362, 181, 90, 45, 25	344,140
res_8_16_16	8, 16, 16	709, 177, 44, 25	143,445
res_8_16_16s	8, 16, 16	709, 70, 25	60,168
no_syn	16, 32, 32	725, 362, 181, 90, 45, 25	383,584

The final map was created by ensembling the 10 NFI fold models. It determined the predicted class for each pixel based on the majority vote of the models. In case of a tie, the tiebreaker relied on the sum of classes probabilities (see Section 2.12).

3. Results

3.1. Clustered Spatial Split Distance Analysis

As described in Section 2.11.1, three models were trained for each of the ten spatial splits, covering split distances ranging from 125 to 5000 m with an increment of 125 m. Figure 3a illustrates the mean NFI-weighted overall accuracy (NFI-w-OA) and the mean MAF1 across the spatial split distances. The results reveal a clear decline in accuracies up to 3000 m, followed by a stabilization around 4000 m, particularly in the macro-averaged F1 score (MAF1). Figure 3b shows the standard deviations in the holdout set size relative to training data for the individual classes and splits, demonstrating a clear upward trend with increasing split distance. This trend indicates that the holdout set share per class and split becomes increasingly unequal as the split distance grows.

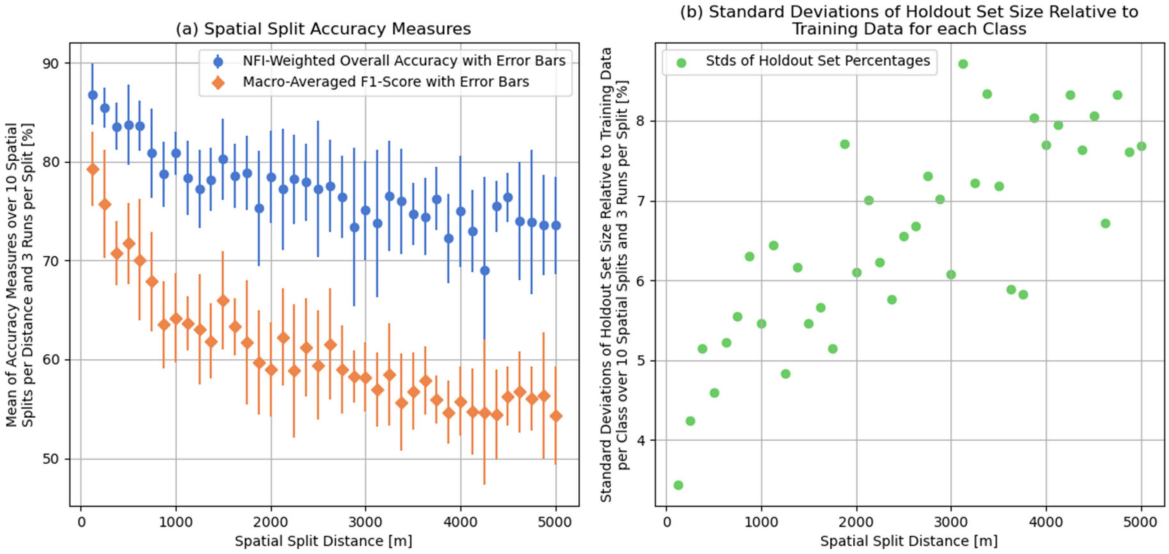

Figure 3. Analysis of accuracy measures over spatial split distances.

3.2. NFI Validation Buffer Distance Analysis

To analyze the effects of the buffer distance on the NFI-VAL results, two series of experiments for buffer distances ranging from 0 to 20,000 m with varying increments were conducted (see Section 2.14.3).

The results of the first series, where the training data discard was determined solely by the buffers are depicted in Figure 4. Figure 4a shows the accuracy results for overall accuracy (OA) and MAF1 per buffer distance, while Figure 4b illustrates the amount of training data discard due to the buffering of the NFI plots per buffer distance. After a small drop from 0 to 250 m, the accuracy results remain steady until approximately 5000 m buffer distance, with a training data loss of up to about 30%. From 7500 m onwards, accuracies start dropping significantly, while the training data discard rises to 95%.

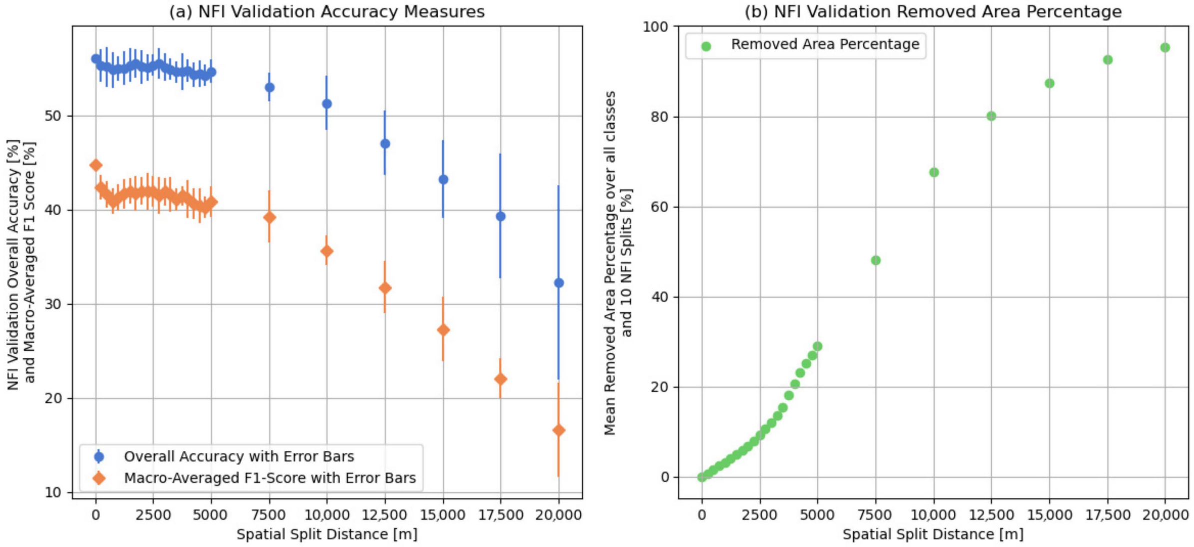

Figure 4. National Forest inventory data validation (NFI-VAL) buffer distance analysis.

The results for the second series, where the amount of training data discarded was kept constant up to 15,000 m, show steady accuracies until the 10,000 m buffer distance and a slight drop from 10,000 to 15,000 m. The standard deviations for the accuracy measures are higher than in the first series. The data for 17,500 and 20,000 m were not affected by the method for this second series, and the accuracy values are in accordance with that (Figure 5).

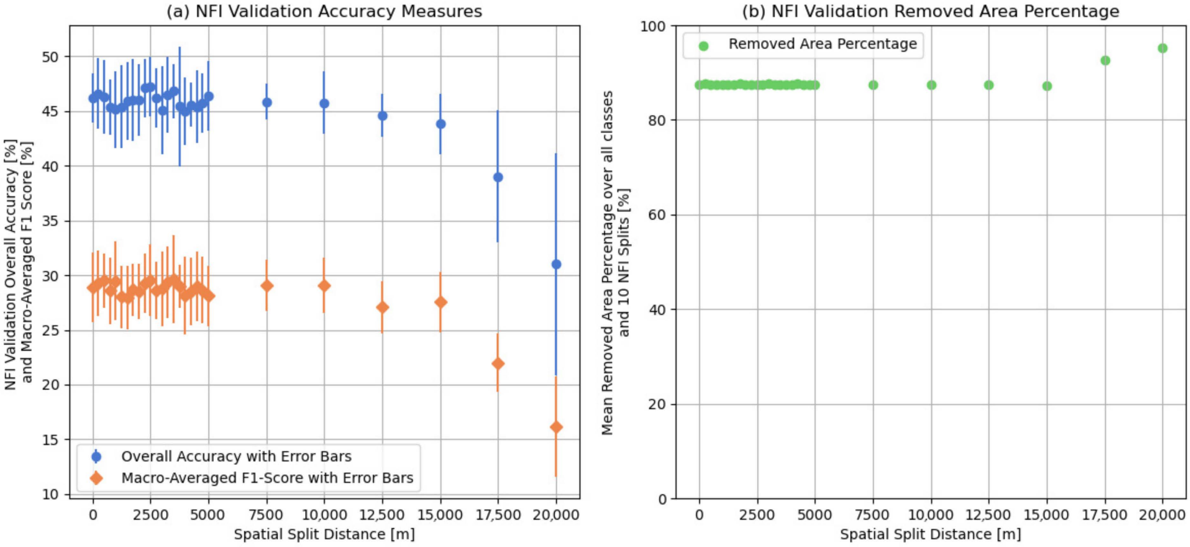

Figure 5. National Forest inventory data validation (NFI-VAL) buffer distance analysis with training data discard kept constant up to 15,000 m.

3.3. Models

In this section, we present the validation results of the set of models described in Section 2.15. During the spatial split and NFI fold training, the configurations mostly achieved 99% NFI-weighted overall accuracy (NFI-w-OA) and 99% NFI-w-OA on a random holdout set for NFI fold training. The exceptions were the no_syn, res_8_16_16 models, which reached 98%, and the res_8_16_16s model with 96%. The results of the clustered spatial split validation (CSS-VAL) are presented in Table 4.

Table 4. Clustered spatial split validation.

Model	Number of Parameters	NFI-w-OA [%] ± std	MAF1 [%] ± std
base	344,140	**73.8 ± 5.4**	55.0 ± 3.5
res_64_128_128	955,324	73.3 ± 5.5	**55.1 ± 3.4**
res_8_16_16	143,445	72.4 ± 5.3	54.3 ± 3.3
res_8_16_16s	60,168	72.0 ± 5.2	54.4 ± 3.4
no_syn	383,584	63.8 ± 6.7	49.8 ± 4.0

The best values for each metric are highlighted in bold.

In the CSS-VAL assessment, the base model outperformed (measured by the NFI-w-OA) the other models by reaching an NFI-weighted overall accuracy (NFI-w-OA) of 73.8% (±5.4) and a macro-averaged F1 score (MAF1) of 55.0% (±3.5). The integration of the synthetic training data, as illustrated by comparing the base model with the no_syn model, leads to a significant enhancement in the classifier accuracy. Specifically, we observed an increase of 10% in NFI-w-OA and 5.2% in MAF1. Model capacity variations for the base model exhibited a minor downward trend at the lowest parameter counts. Interestingly, the parameter reduction had a more pronounced effect on NFI-w-OA compared to MAF1.

The accuracy measures of the NFI-VAL are presented in Table 5.

Table 5. National Forest inventory data validation (NFI-VAL) accuracy measures.

Model	**OA [%] ± std**	**MAF1 [%] ± std**	**OMS ± std**	**PCPP [%] ± std**	**DCC [%] ± std**	**POA [%] ± std**	**MOA [%] ± std**
base	**55.3 ± 1.8**	42.0 ± 1.3	**1.63 ± 0.04**	79.4 ± 1.4	1.5 ± 0.3	**90.7 ± 1.3**	**64.6 ± 4.2**
res_8_16_16	55.1 ± 2.1	**42.2 ± 1.8**	1.66 ± 0.04	79.4 ± 1.4	1.7 ± 0.4	89.7 ± 1.6	62.7 ± 3.8
res_8_16_16s	54.8 ± 1.5	41.5 ± 1.5	1.68 ± 0.01	79.3 ± 1.2	1.7 ± 0.4	89.5 ± 0.9	62.1 ± 2.7
no_syn	47.7 ± 1.1	40.6 ± 1.5	1.84 ± 0.03	**81.4 ± 1.1**	**1.2 ± 0.3**	89.8 ± 1.4	58.4 ± 3.2

The best values for each column are highlighted in bold.

In the NFI-VAL assessment, the base model outperformed the other models, achieving an overall accuracy (OA) of 55.3% (±1.8), MAF1 of 42.0% (±1.3), and an overall misclassification score (OMS) of 1.63 (±0.04) (see Section 2.14). The model demonstrated an accuracy of 79.4% (±1.4) when disregarding confusions in phenological proximity (PCPP), such as spruce–larch confused with spruce–beech (see Section 2.14) and the confusion percentage between deciduous and coniferous classes (DCC) was 1.5% (±0.3). Additionally, the post-hoc pure class overall accuracy (POA) reached 90.7% (±1.4), while the post-hoc mixed class accuracy (MOA) was 64.6% (±4.2). The confusion matrix and additional accuracy measures (Appendix A Tables A2–A4) reveal additional noteworthy observations: Within the pure classes, larch and pine demonstrate lower F1 scores compared to other pure classes, and mixed classes exhibit lower scores than pure classes.

The integration of synthetic training data, as demonstrated by comparing the base model with the no_syn model, significantly enhanced the classifier's predictive power. Specifically, the OA increased by 7.6%, the MAF1 by 1.4%, the OMS was improved by 0.21 points, the POA by 0.9%, and the MOA by 6.2%. However, it is noteworthy that the PCPP and the DCC slightly worsened. The reduction in model capacity (res_8_16_16 and res_8_16_16s) showed a subtle downward trend in the accuracy metrics, predominantly reflected in the OMS.

The tree species map over the area of the Austrian federal territory is presented in Figure 6.

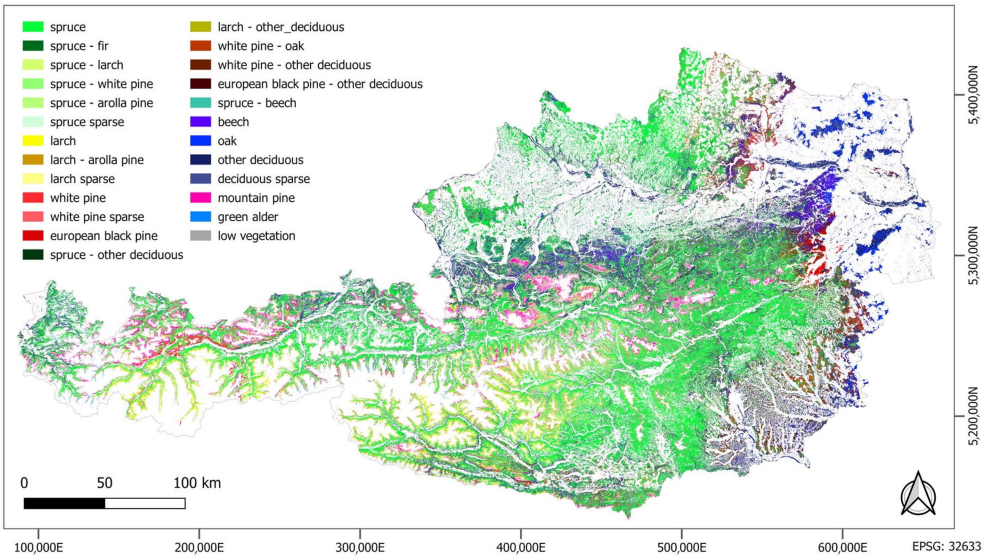

Figure 6. Tree species map over the entire forest in the study area.

Figures 7–11 display a color-infrared (CIR) orthophoto, provided by the Austrian Federal Ministry of Agriculture, Forestry, Regions and Water Management (BML), on the left, with the tree species map overlaid on top of it on the right.

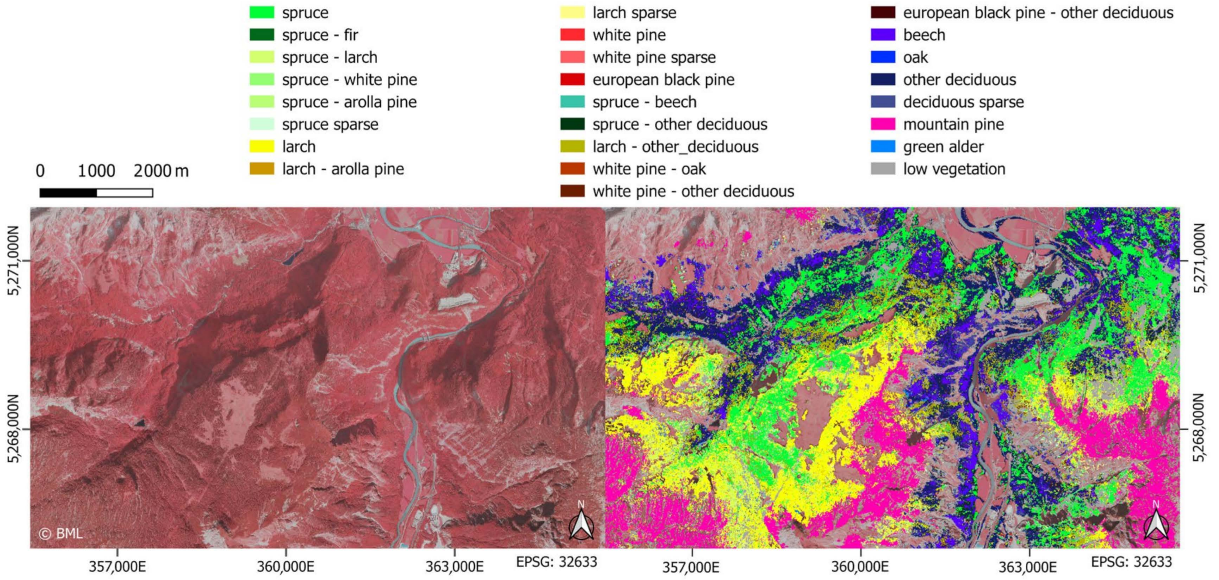

Figure 7. Tree species map on an intermediate zoom level.

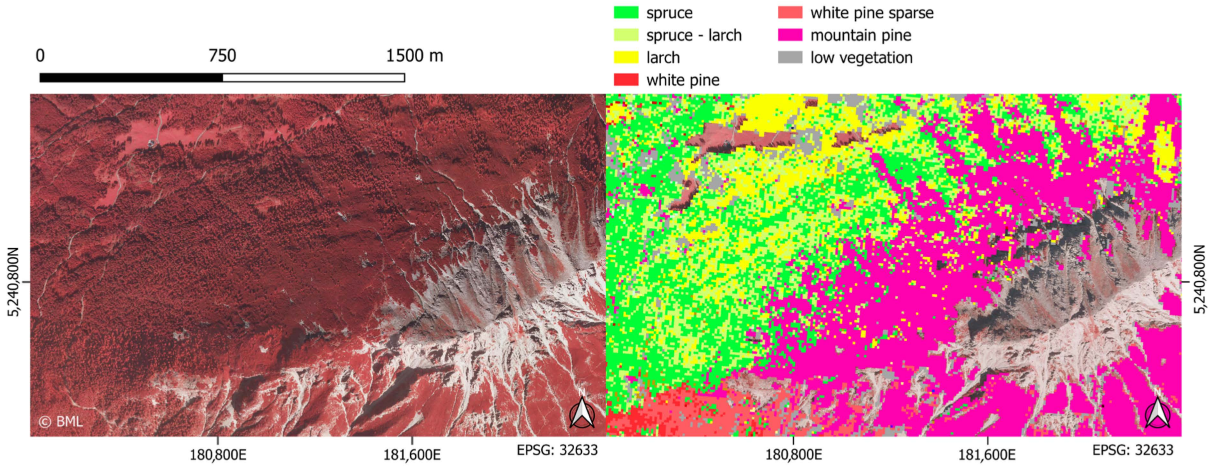

Figure 8. Tree species map intermediate–close zoom level: predominantly mountain pine, white pine, spruce, and larch.

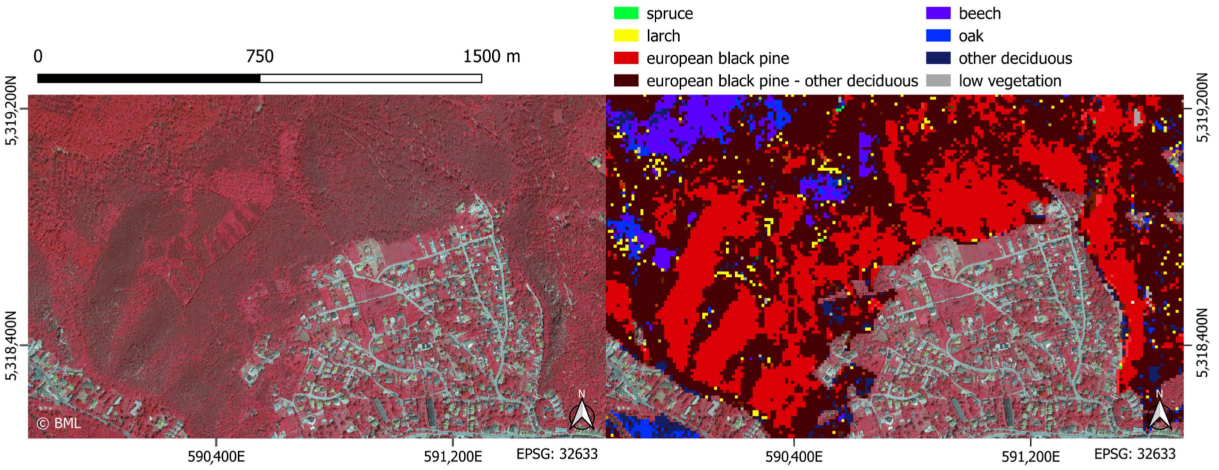

Figure 9. Tree species map intermediate–close zoom level: predominantly black pine, black pine–other deciduous, and beech.

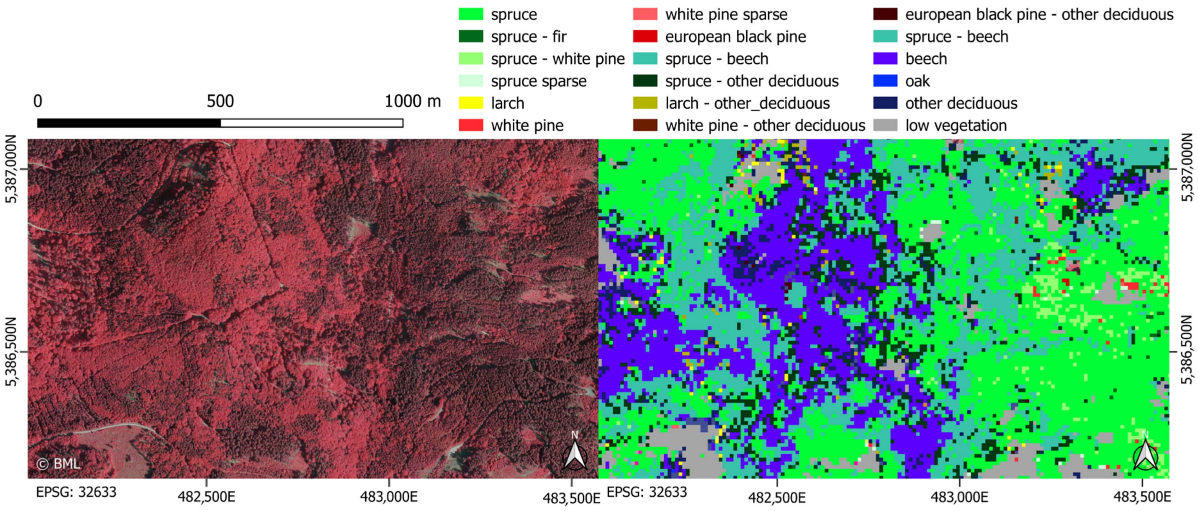

Figure 10. Tree species map intermediate–close zoom level: predominantly spruce and beech.

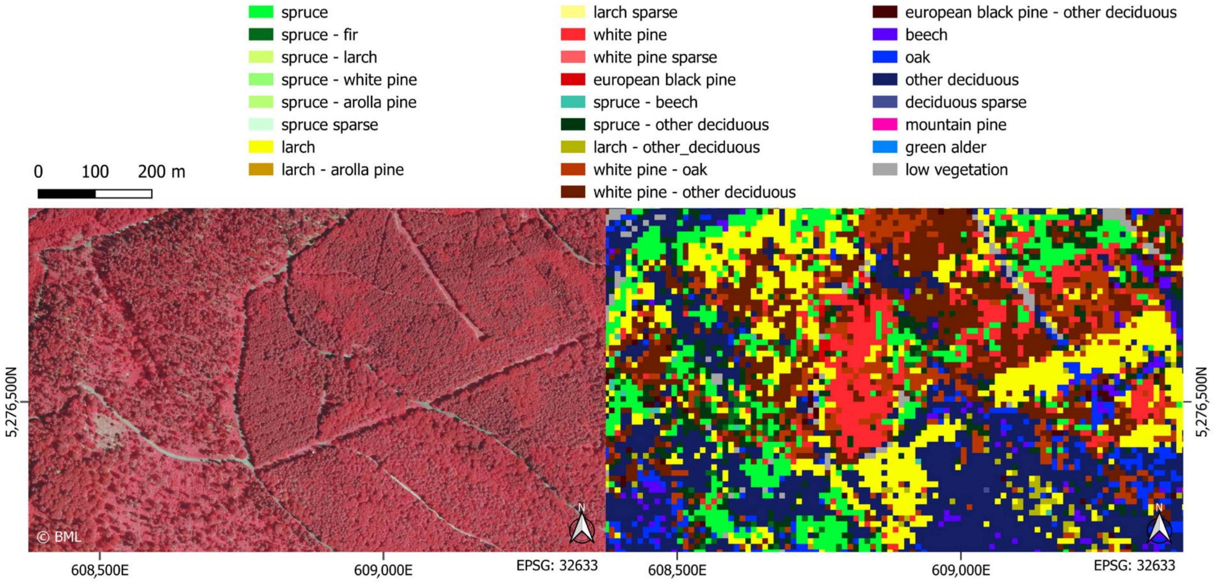

Figure 11. Tree species map intermediate–close zoom level: typical mixed forest.

An online-accessible and user-friendly version of the resulting tree species map from this publication can be found on the Austrian National Forest Inventory's Homepage [60]: https://www.waldinventur.at/?x=1486825&y=6059660&z=7.75968&r=0&l=1111#/map/1/mBaumartenkarte/Bundesland/erg9 (accessed on 25 January 2024). It provides a simplified and interactive representation of the map, allowing users to easily explore the data.

4. Discussion

In this study, the classification of tree species was conducted across a large area spanning 40,178 km^2 of forests. By utilizing a dense phenology time series extracted from freely available Sentinel-2 (S2) imagery, our methodology demonstrates its suitability for large-scale applications. To align with the scale of this study, a comprehensive training dataset consisting of about 570,000 S2 pixels (~57 km^2), spread across the entire study area, was collected via orthophoto interpretation. The inclusion of mixed classes in this study's classification scheme offers a robust alternative to the approach suggested by [37] to align training and validation data more closely with real forest conditions. The utilization of synthetic data resulted in significant improvements in the model's predictive power.

The challenge of spatial autocorrelation typically remained inadequately addressed in large-scale tree species classification. Our study proposes methods to assess the influence of spatial dependencies specifically from the perspective of map validation. In addition to employing a clustered spatial split validation (CSS-VAL), we utilize a validation method based on a ground truth probability sample (NFI-VAL). Through the analysis of national forest inventory data-based validation (NFI-VAL) results across various buffer distances, we assert that this validation approach closely approximates the true independence of the validation data. Comparing different validation approaches emphasized the importance of accounting for spatial autocorrelation and accurately modeling the complexity of the forest. The post-hoc pure class accuracy of 91%, given the large study area and the meticulous validation method, compares well with other studies focusing exclusively on pure classes.

4.1. Mixed Species Classes, Training Data Labeling, and Training Data Synthesis

The significant disparity between the post-hoc pure class accuracy (POA) and the overall accuracy (OA) in the NFI-VAL underscores the importance of accounting for mixed species stands in tree species mapping validation, particularly in regions with high forest diversity. From a training perspective, including mixed species information into the training data is essential for modeling diverse forest areas. Understanding the interactions and overlaps between different species within the same spatial context can enhance the generalization capabilities and predictive performance of the models. However, a major challenge arises. While it is feasible to label training areas with proportions of large numbers of different species, the quality of labels at the single-pixel level deteriorates, because mixtures are typically not spatially homogeneous. We believe that mixed species classes, consisting of two different species, represent a natural progression in complexity from considering only pure species stands, which is common in large-scale tree species mapping.

The training data were labeled through an iterative process that took model evaluation and consistency metrics into account. Additionally, we identified areas of relatively low probability for the predicted class and sought to locate training areas within these regions to improve the model’s weak spots. Labeling the mixed data classes presents inherent difficulties. One main challenge is the reduced separability of different classes in the feature space, as classes containing the same species are naturally closer together. Furthermore, when vectorizing training areas, achieving a consistent mixture in every S2 pixel of a mixed class’s two constituent species can be challenging. To add to that, training areas that exhibit too much homogeneity reduce the variety in the training data and fail to represent real forest conditions, resulting in a trade-off between label quality and training data representativity. In our study, mixed class training areas typically contain both pure and mixed pixels (i.e., a spruce–beech labeled polygon contains pure spruce, pure beech pixels, and spruce–beech mixed pixels), and up to 10% species impurity was permitted. This label inconsistency has a negative impact on class separability; however, synthetic training data can help to address the issue.

The introduction of synthetic data for mixed classes demonstrated significant increases in CSS and NFI validation. Intuitively, this improvement can be attributed to the synthesis (averaging of pixels) counteracting the occurrence of pixels containing only a single (constituent) species (of the target class) in training areas.

Synthetic data in the context of machine learning are not a novel concept; see the recent survey by [61], to give one example. The synthetic training data approach was based on the simplified model that describes the spectral signature of multi-species pixels as a linear combination of the constituents.

4.2. Neural Network Architecture

In a review study by [44], a residual neural network (ResNet) architecture employing one-dimensional convolutions and specifically designed for time series classification, emerged as the best performing deep learning approach across various time series classification tasks. Furthermore, a study by [31] found that a neural network structure based on one-dimensional convolutions outperformed other approaches in their tree species classification study. Our study's architecture is based on the ResNet architecture presented in [44] and specifically tailored to address the intricacies of our classification task, accommodating both phenology time series features and non-time series features such as phenology metrics, DTM, and NDSM. We opted against using spatially aware architectures like U-Net [43]. Due to our training data's clustered nature and homogeneity, the data featured more homogeneous pixel neighborhoods compared to actual forest conditions. Therefore, we believe the addition of neighborhood relations to the models' input data would exacerbate the existing mismatch between the training and real-world data.

4.3. Autocorrelation Analysis

The autocorrelation analysis of the CSS-VAL depicted in Figure 3a reveals an anticipated trend: as the split distances increase and the autocorrelation decreases, the decline in accuracy levels off. Figure 3b highlights a discernible trend of increasing variation in holdout set ratios relative to training data for individual classes and splits. We believe this trend is particularly pronounced in classes with fewer training samples and more spatially concentrated distributions. As the split distance grows, individual clusters expand, making it increasingly challenging to maintain a consistent holdout set ratio across multiple splits and classes. The growing variation in holdout set ratio explains the larger standard deviation in accuracy with growing distances, exhibited in Figure 3a. These results, particularly those in Figure 3a, suggest quasi-spatial independence within the set of training data polygons at a split distance of 4000 m.

As the collection of training data was not independent of the NFI-VD, we utilized a buffered NFI-VAL to examine the spatial dependency. Figure 4 illustrates a slight drop at 250 m, a plateau up to 5000 m, and a decline in accuracies with increasing buffer distances from 7500 m onwards, especially notable when a substantial portion of training data was excluded due to the expanded buffer sizes. To further test the spatial autocorrelation between the training and NFI-VAL data, we initiated an experiment where the discard of training data was kept constant up to a buffer distance of 15,000 m (see Section 2.11.2), prompted by the significant drop in accuracies observed when the training data were discarded. The results in Figure 5 showed stabilized accuracies for up to 10,000 m buffer distances and slight drops at 12,500 and 15,000 m. Overall, these results provide evidence of the NFI-VD's quasi-spatial independence even at relatively small buffer distances, affirming the suitability of the NFI-VAL as a measure for the model's predictive power.

It is interesting to note the significant differences in autocorrelation distances between the CSS and NFI-VAL datasets. We attribute this discrepancy to the differing spatial distributions: the CSS dataset reflects a sample of the training data's spatial pattern, whereas the NFI-VAL dataset represents a systematic probability sample of the entire study area. The training data have clusters of labeled polygons with varying densities, reflecting the nature of the training data collection process. These clusters might capture specific localized patterns and result in stronger spatial dependencies. In contrast, the evenly distributed NFI-VAL dataset does not exhibit these localized patterns, leading to different autocorrelation characteristics.

Wadoux et al. [41] highlighted that due to spatial autocorrelation, the accuracy of predictors for thematic maps, trained on clustered training data, may be overestimated when validated with randomly selected data from the same training set. Consistent with this claim, our study demonstrates a substantial decline in accuracy when comparing random holdout set validation and clustered spatial split validation (CSS-VAL, see Section 4.4). Our approach for analyzing spatial autocorrelation, based on model validation, seamlessly incorporates a high number of input dimensions, unlike more traditional methods such as semi-variograms, and directly quantifies the impact of spatial autocorrelation on the model's accuracy. However, it is important to note that these approaches are contingent upon the characteristics of the data (both input and validation) and the model itself and the specific results may not generalize well to different datasets. Therefore, we advise against directly applying the resulting split and buffer distances to other studies.

4.4. Validation

Our study underscores a well-known principle: the importance of independent and representative ground reference data in validating tree species maps. This finding is consistent with results from various studies across different machine learning disciplines, including [39–41]. While we did not encounter large-scale tree species classification studies addressing this specific issue, research such as [38] on the classification of dominant leaf types has identified notable differences in validation outcomes when comparing validation data from the training distribution with data from an independent distribution, especially in areas with a high degree of mixture.

Despite its utility, employing NFI-VD presents certain challenges and limitations. Spatial inaccuracies arising from differences between the geolocation of NFI plots and S2 images can significantly impact validation results. As reported by [62], the long-term performance for unrefined S2 products is close to 11 m or better at 95% confidence. Since the activation of the global refinement in August 2021, the absolute geolocation error is better than 7.1 m for S2A and 5.6 m for S2B at 95% confidence. Therefore, the data used in our study, spanning from 2017 to 2021, are mostly affected by an 11 m or better geolocation error at 95% confidence. Given the small plot areas (two to four S2 pixels), even a shift by one pixel significantly impacts the validation results. To mitigate this issue, we implemented a validation with shift variants, which offers greater stability under these spatial inaccuracies (see Section 2.14.3). Additionally, the classes provided by the NFI-VD do not entirely align with the classification scheme. Specifically, the white and black pine classes could not be distinguished in the NFI-VD, and the low vegetation class could only be assigned in the confusion matrix when it was misclassified. Furthermore, plots labeled as mixed classes by the NFI-VD may contain single pixels containing only a single constituent species. These cases are not correctly represented in the NFI-VD, as full plots are being labeled instead of single pixels. For example, an NFI plot labeled as spruce–beech may intersect with three S2 pixels, of which two contain spruce and beech while one contains only spruce. Even if the model correctly predicts the pixel with only spruce as spruce, it is still considered a misclassification. This discrepancy prompted us to introduce an overall misclassification score (OMS) and a prediction in close phenological proximity (PCPP) metric to model the phenological differences and similarities between species, and pure and mixed classes. These metrics allow for a better interpretation of model performance and guide development.

4.5. Results and Model Performance

The validation results for the best performing model (base) exhibit a tremendous gap between the random holdout set validation (99% NFI-weighted overall accuracy (NFI-w-OA)) and the CSS-VAL (74% NFI-w-OA). We believe this decrease of about 25% can be attributed to the spatial autocorrelation inherent in the (training) data. To maintain sufficient label quality on the single S2 pixel level, the training areas were selected to be homogeneous in species distribution. However, this led to increased pixel similarity within training areas, especially after training data synthesis. This pixel similarity within training polygons explains the drop from 99% NFI-w-OA to about 87% NFI-w-OA at a 125 m split distance, where clusters mainly consist of single polygons. The further decline of 13% to a 4000 m split distance can be attributed to both the spatial autocorrelation inherent to the data, independent of training area homogeneity, and the similarity of homogeneous training areas in spatial proximity. These findings highlight the absolute necessity of accounting for spatial autocorrelation in tree species map validation, particularly when working with clustered training data.

The comparison of CSS-VAL NFI-w-OA and NFI-VAL overall accuracy (OA) results for the base model shows a vast decline of about 20% in OA (given that the NFI-VAL is naturally NFI-weighted, these metrics compare well). We attribute this decline to the feature distribution disparity between the training data and ground truth forests. While homogeneous polygons provide a feasible way of labeling training data, in diverse forest situations where individual pixels can differ greatly in close spatial proximity, they fail to capture real forest complexity. We believe this argument is further supported by the high post-hoc pure class overall accuracy (POA) of 91% and the higher F1 scores for pure classes (see Appendix A Table A4), showcasing that in simple forest compositions, the model is much more capable of generalizing from training to inference data than in situations with species mixtures.

The overall misclassification score (OMS) provides a valuable tool for comparing the quality of generated maps, taking the intricacies of mixed class confusions into account. The base model's prediction in the close phenological proximity (PCPP) metric result of 79% provides an important context to the OA of 55%. A total of 24% of the confusions occur in close phenological proximity to the target class when one of the species in the mixed classes matches, but the other does not. The deciduous coniferous confusions (DCC) of 1.5% remain low, recognizing the significant phenological differences between coniferous and deciduous species, especially during leafing-out and leaf fall. Table A4 in Appendix A reveals interesting details on the class level. In F1 scores, pure classes—apart from larch and pine—clearly and expectedly outperform mixed classes. The misclassification scores paint a different picture. Here, spruce, spruce–coniferous mixed (except for spruce–arolla pine, a class with very little training data), mountain pine, and green alder are the best performing classes. This reflects the results in Appendix A Table A2, where most of the confusions for the spruce and spruce–coniferous classes happen amongst themselves. Too little training data for individual classes can negatively impact their performance, as shown by classes such as spruce–arolla pine and pine–oak. However, when classes are less difficult to separate, such as mountain pine and green alder (due to their low height), given small amounts of training data, they can still perform well.

Figures 8–10 showcase diverse age structures and mixtures, predominantly involving two species in spatial proximity, as identifiable in the CIR orthophoto. The model demonstrates proficient handling of these scenarios. In Figure 11, a characteristic mixed forest is depicted, featuring a diverse array of tree species within a confined spatial scale. Furthermore, the presence of sparsely populated regions, where different ground vegetation types introduce noise to the signal from the actual forest canopy, poses significant challenges for the model.

The comparison of the base and no_syn model's results exhibits profound improvements in OA, OMS, and post-hoc mixed class overall accuracy (MOA). The POA was barely affected, as the pure class training data were not synthesized. The 2% decrease in PCPP is noteworthy: at the PCPP level, confusions between pure classes and mixed species classes containing the respective pure class are not considered. The idea behind synthetic training data was to reduce labeling errors in mixed class training areas when pure species pixels were present. The PCPP not improving with training data synthesis supports our theory on mixed class label quality. The slight decrease emphasizes that the homogeneous training data decrease the models' generalization capabilities. Overall, the considerable accuracy improvements resulting from training data synthesis highlight that mixed class label quality was improved and justify the resulting reduction in training data representativity.

To test for the potential overfitting of our models, we conducted an extensive series of experiments using models with varying capacities and presented a few key results. Investigating the results of the complete series, the NFI-VAL and CSS-VAL showed no signs of overfitting, but rather a slight downward trend in accuracy with decreasing model capacity, as is reflected by the published results.

5. Conclusions

This study underscores the imperative need to approach real forest complexity in modeling and validation while accounting for spatial autocorrelation in the validation process. Our findings reveal immense disparities between the random training data holdout set and clustered spatial split validation, highlighting the importance of considering spatial factors in reliable tree species map evaluation. Additionally, we observed a substantial decline in accuracy when validating with an independent probability sample and a major accuracy increase when only pure species classes were considered, emphasizing the need to account for real forest complexity both in validation and modelling.

In our study, we introduced several innovative methods. We incorporated mixed species classes into the classification scheme, allowing us to better capture the diversity of forests on a Sentinel-2 pixel level. We implemented training data synthesis for mixed species classes, significantly improving accuracy results, and developed validation metrics tailored to the intricacies of mixed species class validation. Our in-depth analysis of spatial autocorrelation solidified the independent probability sample-based validation approach, investigating its susceptibility to spatial biases.

However, this study is not without its limitations. The geolocation accuracy of Sentinel-2 imagery and NFI plots poses challenges in the evaluation process. Furthermore, the vectorized training data introduce a trade-off between label quality and data representativity, reflecting the challenges of accurately modeling real forest conditions.

Tree species classification over large areas presents a multitude of intriguing research challenges. We believe that many of these challenges can be addressed by leveraging and adapting methods from deep learning research. With great curiosity, we eagerly look forward to exploring how these cutting-edge techniques can revolutionize tree species classification over large areas.

Author Contributions: Conceptualization, T.S. and L.W.; data curation, T.S. and S.K.; formal analysis, T.S.; investigation, T.S.; methodology, T.S., S.K., M.L. and L.W.; project administration, T.S.; resources, S.K., M.L., U.K., K.K. and C.B.; software, T.S.; validation, T.S. and S.K.; visualization, T.S.; writing—original draft, T.S.; writing—review and editing, T.S., A.B., S.G. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Federal Ministry of Agriculture, Forestry, Regions and Water Management [grant numbers BMLRT/III-2021-M6 (FAI.35), BMLRT/III-2021-M4/2 (FAI.2)] and the Austrian Climate Research Program [grant number ACRP14–Austria Fire Futures–KR21KB0K00001].

Data Availability Statement: The forest area map data, phenology-index time series data, training data, and tree species map presented in this study as well as the software source code are not readily available and will only be shared for non-commercial use. Requests to access the datasets should be directed to the corresponding author. The NDSM data presented in this study are not available and can only be provided to forest owners due to legal reasons. The DTM data presented in this study data presented in this study are not available. The data were obtained from the Federal Ministry of Agriculture, Forestry, Regions and Water Management. The NFI validation data presented in this study are not available to assure for the representativity of the sample of the NFI to remain uncompromised.

Acknowledgments: We extend our sincere appreciation to the European Space Agency and the European Commission for their invaluable contribution in granting free access to Sentinel-2 data, enabling our research. We are deeply grateful to the Federal Ministry of Agriculture, Forestry, Regions and Water Management, as well as the Austrian Climate Research Program, for their partial funding of our work. In addition, we would like to extend our heartfelt appreciation to Florian Aschauer for his valuable suggestions and insights in reviewing the script. We are grateful for his contributions to improving the quality of our work. We would also like to express our gratitude to the three anonymous reviewers for their constructive feedback and helpful comments, which greatly contributed to the refinement of this publication. During the preparation of this work, the authors used ChatGPT 3, 3.5, and 4o in order to *improve readability and language.* After using ChatGPT 3, 3.5, and 4o the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table A1. Main phenology course (MPC) statistics.

Name	Description
MPC_increm_abs *	Increment from one DOY to the next
DEFOLIATION_doy	DOY in [245:330] where MPC_increm_abs is minimal
DEFOLIATION_start	DOY of last local maximum before DEFOLIATION_doy
DEFOLIATION_end	DOY of first local minimum after DEFOLIATION_doy that is below 25th-MPC-percentile
DEFOLIATION_duration	DEFOLIATION_end-DEFOLIATION_start
DEFOLIATION_doy_adj	Mean (DEFOLIATION_start, DEFOLIATION_end)
DEFOLIATION_gradient_median	Median (MPC_increm_abs[DEFOLIATION_start, DEFOLIATION_end])
DEFOLIATION_gradient_min	Min(MPC_increm_abs[245:330])
GREENING_doy	DOY in [90:182] where MPC_increm_abs is maximal
GREENING_start	DOY of last local minimum before GREENING_doy
GREENING_end	First local maximum after GREENING_doy that is above 75th-MPC-percentile
GREENING_doy_adj	Mean (GREENING_start, GREENING_end)
GREENING_duration	GREENING_end-GREENING_start
GREENING_gradient_median	Median(MPC_increm_abs[GREENING_start, GREENING_end])
GREENING_gradient_max	Max(MPC_increm_abs[90:182])
DP_max	Maximum value from data points
DP_ampl	(DP_max-MOD_mean)/MOD_mean * 100
MOD_ALL_nDP	Number of data points before outlier filter
MOD_MP_nDP	Number of data points used for modelling after filtering and modelling period adaption
MOD_n_years	Number of years for modelling
MOD_max	Maximum of MPC
MOD_max_doy	DOY of maximum of MPC
MOD_min	Minimum of MPC
MOD_min_doy	DOY of minimum of MPC
MOD_mean	Mean of MPC
MOD_median	Median of MPC
MOD_percx	x-th percentile of MPC
MOD_range_max_min	MOD_max-MOD_min
MOD_range_p75_p25	MOD_perc75-MOD_perc25
MOD_range_p90_p20	MOD_perc90-MOD_perc10
MOD_sd	Std (differences (model, data points)) in modeling period
MOD_ampl_max	(MOD_max-MOD_mean)/MOD_mean * 100
MOD_ampl_p75	(MOD_per75-MOD_mean)/MOD_mean * 100
MOD_ampl_p90	(MOD_per90-MOD_mean)/MOD_mean * 100
MOD_dp_dev_all_abs	Median (diffs(model, data points)) in modeling period
MOD_dp_dev_neg_abs	Median (non-positive differences (model, data points)) in modeling period

Table A1. *Cont.*

Name	Description
MOD_dp_dev_pos_abs	Median (non-negative differences (model, data points)) in modeling period
MOD_dp_dev_all_rel	Median (differences (model, data points)/MPC values * 100) in modeling period
MOD_dp_dev_neg_rel	Median (non-positive differences (model, data points)/MPC values * 100) in modeling period
MOD_dp_dev_pos_rel	Median (non-negative differences (model, data points)/MPC values * 100) in modeling period
MTC	Second biggest number of days above 0.5 perc in a row
MTC_startdoy	Start DOY of MTC
PTA_x_firstreach	DOY when percentile x is reached for the first time
PTA_x_lastpass	DOY when percentile x is passed from above for the last time
PTA_x_n_above	Number of MPC values above percentile x
PTA_x_n_transition	Number of times MPC values transition above percentile x
PTA_x_value	Value of percentile x
PTC	Maximum of number of days above 0.65 perc in a row
PTC_startdoy	start DOY for PTC
VP_start	PTA_0.6_firstreach
VP_end	VP_start + LBG
VPL	PTA_0.6_lastpass-PTA_0.6_firstreach
VEGPERIOD_length	DEFOLIATION_doy-GREENING_doy
VEGPERIOD_length_adj	DEFOLIATION_doy_adj-GREENING_doy_adj
VA (Vegetation-Abundance-Index)	Mean (MPC values $\geq$ 0.5 Percentile)
TD (Temporal-Dispersion)	(days above 0.75 percentile) * mean (MPC values above 0.75 perc-0.75)
LGB (Length of growing biomass)	Number of days where MPC is above threshold (threshold = (0.95 perc + 0.05 perc)/2

* intermediate results. The " MOD_percx" statistics were calculated for x in {10, 25, 75, 90}. The "PTA_x" statistics were calculated for x in {0.05, 0.1, . . ., 0.95}.

Table A2. Confusion matrix.

	Spruce	Spruce-Fir	Spruce-Larch	Spruce-Pine	Spruce-Arolla Pine	Larch	Larch-Arolla Pine	Pine	Spruce-Beech	Spruce-Other Deciduous	Larch-Other Deciduous	Pine-Oak	Pine-Other Deciduous	Beech	Oak	Other Deciduous	Mountain Pine	Green Alder	Low Vegetation
Spruce	8168	613	922	638	45	24	15	47	436	365	16	5	25	29	9	64	3	0	0
Spruce-fir	34	222	14	5	0	0	0	0	18	13	0	0	0	1	0	3	0	0	0
Spruce-larch	362	28	957	22	25	30	47	0	20	25	1	0	0	0	0	4	5	1	0
Spruce-pine	25	8	4	186	0	0	0	2	6	6	0	0	1	0	0	1	0	0	0
Spruce-arolla pine	33	0	13	0	17	0	4	0	2	0	0	0	0	0	0	0	0	0	0
Larch	100	13	176	5	16	208	54	3	44	21	12	1	5	14	2	11	9	6	0
Larch-arolla pine	64	1	60	3	16	7	61	0	0	3	0	0	0	0	0	0	3	1	0
Pine	91	11	52	328	1	5	5	308	26	33	5	21	57	6	5	12	2	0	0
Spruce-beech	180	67	60	46	0	2	0	1	733	108	8	0	9	56	0	33	0	0	0
Spruce-other deciduous	147	78	45	51	0	0	0	6	342	510	23	2	29	115	10	94	1	0	0
Larch-other deciduous	37	7	26	10	0	3	0	0	143	29	41	2	3	71	2	30	2	1	0
Pine-oak	1	0	0	6	0	0	0	5	7	8	0	40	9	0	1	3	0	0	0
Pine-other deciduous	32	9	11	57	0	0	0	28	85	47	10	55	409	45	16	27	0	0	0
Beech	16	4	3	3	0	0	0	0	181	16	34	1	9	1008	26	57	2	0	0
Oak	2	0	1	2	0	0	0	2	19	24	11	21	8	54	375	97	0	0	0
Other deciduous	99	22	40	29	0	6	0	6	325	142	44	33	56	292	84	1457	0	6	0
Mountain pine	19	0	16	2	3	10	7	0	0	2	0	0	0	1	0	0	406	1	0
Green alder	0	0	7	0	0	2	1	0	0	1	1	0	0	1	0	1	2	82	0
Low vegetation	870	94	422	84	22	50	20	12	310	419	21	2	17	106	13	197	24	55	0

Table A3. Overall accuracy measures.

Accuracy Measure	Value
Overall accuracy [%]	55.33 ±1.8
Post-hoc pure class accuracy [%]	90.73 ± 1.3
Post-hoc mixed class accuracy [%]	64.64 ± 4.2
Macro F1 score [%]	42.6 ± 1.3
Overall misclassification score	1.63 ± 0.04
Level 1 misclassifications [%]	69.63
Up to Level 2 misclassifications [%]	79.39
Up to Level 3 misclassifications [%]	84.55
Level 4 misclassifications [%]	3.95
Level 5 misclassifications [%]	1.53
Level 0 misclassifications [%]	9.97

Table A4. Class accuracy measures.

	Spruce	Spruce-Fir	Spruce-Larch	Spruce-Pine	Spruce-Arolla Pine	Larch	Larch-Arolla Pine	Pine	Spruce-Beech	Spruce-Other Deciduous	Larch-Other Deciduous	Pine-Oak	Pine-Other Deciduous	Beech	Oak	Other Deciduous	Mountain Pine	Green Alder	Low Vegetation
F1 scores [%]	75	30	44	22	16	40	28	44	37	32	13	30	56	64	65	62	88	65	0
Misclassification scores	1.30	1.74	1.54	1.67	2.08	2.04	2.41	2.02	2.12	1.97	3.28	2.43	2.00	1.98	2.10	2.06	1.28	1.42	
Producer accuracy [%]	79.46	18.86	33.83	12.59	11.72	59.94	28.50	73.33	27.18	28.78	18.06	21.86	64.21	56.03	69.06	69.68	88.45	53.59	
User accuracy [%]	71.50	71.61	62.67	77.82	24.64	29.71	27.85	31.82	56.25	35.10	10.07	50.00	49.22	74.12	60.88	55.17	86.94	83.67	0.00
Producer overall misclassification score	1.19	1.98	1.57	2.01	2.14	1.46	2.04	1.65	2.38	1.73	2.84	2.69	1.73	2.24	1.94	1.57	1.16	1.08	
Producer Level 1 misclassifications [%]	83.88	70.94	72.64	77.99	42.75	70.6	53.73	73.81	39.86	39.16	32.60	51.91	70.65	56.03	69.06	69.68	88.45	53.59	
Producer up to Level 2 misclassifications [%]	87.06	86.32	80.49	90.65	71.03	71.46	77.56	81.67	64.44	70.25	57.71	74.86	88.55	59.14	69.24	76.9	88.45	53.59	
Producer up to Level 3 misclassifications [%]	89.73	88.44	82.89	91.33	84.82	82.70	90.18	93.57	73.15	70.70	61.23	77.05	89.96	78.43	89.50	84.31	93.24	57.51	
Producer Level 4 misclassifications [%]	0.68	1.36	0.39	0.68	0.00	0.58	0.00	1.67	15.35	5.64	29.52	21.86	7.38	12.84	5.16	1.72	0.65	0.65	
Producer Level 5 misclassifications [%]	1.14	2.21	1.80	2.30	0.00	2.31	0.47	1.90	0.00	0.00	0.00	0.00	0.00	2.83	2.95	4.54	0.87	5.88	
Producer Level 0 misclassifications [%]	8.46	7.99	14.92	5.69	15.17	14.41	9.35	2.86	11.49	23.65	9.25	1.09	2.67	5.89	2.39	9.42	5.23	35.95	
User overall misclassification score	1.42	1.50	1.5	1.34	2.03	2.63	2.78	2.40	1.86	2.21	3.71	2.17	2.26	1.73	2.26	2.54	1.40	1.76	0.00
User Level 1 misclassifications [%]	90.92	82.58	88.34	89.12	72.47	62.57	31.05	65.7	64.54	62.22	17.93	61.25	62.70	74.12	60.88	55.17	86.94	83.67	0.00
User up to Level 2 misclassifications [%]	97.93	98.71	99.34	99.58	100.00	64.28	65.75	73.76	95.93	90.78	32.43	76.25	76.18	87.43	64.29	64.33	86.94	83.67	0.00
User up to Level 3 misclassifications [%]	98.71	98.71	99.67	99.58	100.00	85.14	98.17	91.01	97.23	90.92	68.06	95.00	86.41	93.53	88.8	78.79	99.15	85.71	0.00
User Level 4 misclassifications [%]	0.40	0.00	0.00	0.00	0.00	10.14	1.37	6.61	2.76	9.08	31.94	5.00	13.6	4.41	10.06	13.56	0.43	2.04	0.00
User Level 5 misclassifications [%]	0.89	1.29	0.33	0.42	0	4.71	0.46	2.38	0.00	0.00	0.00	0.00	0.00	2.06	1.14	7.65	0.43	12.24	0.00
User Level 0 misclassifications [%]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
Producer sum	10,280	1177	2829	1477	145	347	214	420	2697	1772	227	183	637	1799	543	2 091	459	153	0
User sum	11,424	310	1527	239	69	700	219	968	1303	1453	407	80	831	1360	616	2 641	467	98	2738

Appendix B

Misclassification Score

To account for the complexity of class matches in validation introduced by mixed classes, a set of misclassification levels and a misclassification score were developed. Firstly, the classes were partitioned into five meta-classes:

- Pure coniferous consisting of spruce, larch, white pine, black pine, and mountain pine.
- Pure deciduous consisting of beech, oak, green alder, and other deciduous.
- Mixed coniferous consisting of spruce–white pine, spruce–larch, spruce–fir, spruce–arolla pine, and larch-arolla pine.
- Mixed coniferous-deciduous consisting of spruce–beech, spruce–deciduous, white pine–oak, white pine–deciduous, black pine–deciduous, and larch-deciduous.
- Low vegetation.

Next, seven misclassification levels were established to model misclassification severity, considering matches and confusions between meta-classes and predicted species. Confusions between meta-classes, excluding low vegetation, were treated more severely than confusions within meta-classes. The misclassification levels are defined as follows:

- Level 1: An exact match between the predicted and the validated class.
- Level 2: Confusion within mixed coniferous or between pure coniferous and mixed coniferous, where one species of the predicted class matches the validation. Examples include predicted spruce–fir but validated as spruce and predicted spruce–larch but validated as larch-arolla pine.
- Level 3: Confusion between pure coniferous or pure deciduous and mixed coniferous-deciduous, between mixed coniferous and mixed coniferous-deciduous or within mixed coniferous-deciduous, where one species of the predicted class matches the validation. Examples include predicted spruce but validated as spruce–beech, predicted pine–oak but validated as oak, predicted spruce–white pine but validated as spruce–deciduous and predicted spruce–beech but validated as spruce–deciduous.
- Level 4: Confusion within pure or mixed coniferous, between pure and mixed coniferous, within pure deciduous, within mixed coniferous-deciduous, where no species of the predicted class matches the validation. Examples include predicted spruce but validated as larch, predicted larch-arolla pine but validated as spruce, predicted larch-arolla pine but validated as spruce–white pine, predicted beech but validated as oak and predicted spruce–beech but validated as black pine–deciduous.
- Level 5: Confusion between pure or mixed coniferous and mixed coniferous-deciduous or between pure deciduous and mixed coniferous-deciduous, where no species of the predicted class matches the validation. Examples include predicted larch-arolla pine but validated as spruce–oak and predicted oak but validated as spruce–deciduous.
- Level 6: Confusion between pure or mixed coniferous and pure deciduous. Examples include predicted larch but validated as oak and predicted beech but validated as spruce–larch.
- Level 0: Confusion between any meta-class and the low vegetation class received distinct handling because the low vegetation class was not included in the NFI-VD data. Therefore, only user confusions could be calculated for these cases.

Misclassifications up to Level 3 coincide with the prediction in close phenological proximity (PCPP) metric defined in Section 2.14. In Level 4, the distinction between coniferous, deciduous and mixed is still correct. Levels 5 and 6 were considered severe errors. Specifically, Level 6 corresponds to the deciduous and coniferous confusions (DCC) metric defined in Section 2.14. To evaluate the performance of the classifiers, each cell in the confusion matrices was assigned a misclassification level based on the rules previously described. User and producer misclassification scores were then computed for each class as a weighted average over the corresponding row and column in the confusion matrix, respectively. The weights were determined by the misclassification level of each cell. Finally, an overall misclassification score OMS was computed as a weighted average over all gradings and classes, with the weights determined by the number of validation samples for each class.

References

1. Baumbach, L.; Hickler, T.; Yousefpour, R.; Hanewinkel, M. High economic costs of reduced carbon sinks and declining biome stability in Central American forests. *Nat. Commun.* **2023**, *14*, 2043. [CrossRef] [PubMed]
2. Berger, F.; Rey, F. Mountain Protection Forests against Natural Hazards and Risks: New French Developments by Integrating Forests in Risk Zoning. *Nat. Hazards* **2004**, *33*, 395–404. [CrossRef]
3. Brang, P.; Schnenberger, W.; Ott, E.; Gardner, B. Forests as Protection from Natural Hazards. In *The Forests Handbook*; Evans, J., Ed.; Blackwell Science Ltd.: Oxford, UK, 2001; Volume 2, pp. 53–81, ISBN 978-0-470-75707-9. [CrossRef]
4. Jim, C.Y.; Chen, W.Y. Assessing the ecosystem service of air pollutant removal by urban trees in Guangzhou (China). *J. Environ. Manag.* **2008**, *88*, 665–676. [CrossRef] [PubMed]
5. Miller, D.C.; Hajjar, R. Forests as pathways to prosperity: Empirical insights and conceptual advances. *World Dev.* **2020**, *125*, 104647. [CrossRef]
6. O'Brien, L.E.; Urbanek, R.E.; Gregory, J.D. Ecological functions and human benefits of urban forests. *Urban For. Urban Green.* **2022**, *75*, 127707. [CrossRef]
7. Sander, H.; Polasky, S.; Haight, R.G. The value of urban tree cover: A hedonic property price model in Ramsey and Dakota Counties, Minnesota, USA. *Ecol. Econ.* **2010**, *69*, 1646–1656. [CrossRef]
8. Teich, M.; Accastello, C.; Perzl, F.; Kleemayr, K. (Eds.) *Protective Forests as Ecosystem-Based Solution for Disaster Risk Reduction (Eco-DRR)*; IntechOpen: London, UK, 2022; ISBN 978-1-83969-325-0. Available online: https://www.intechopen.com/books/10812 (accessed on 1 February 2024).
9. Yang, J.; McBride, J.; Zhou, J.; Sun, Z. The urban forest in Beijing and its role in air pollution reduction. *Urban For. Urban Green.* **2005**, *3*, 65–78. [CrossRef]
10. Patacca, M.; Lindner, M.; Lucas-Borja, M.E.; Cordonnier, T.; Fidej, G.; Gardiner, B.; Hauf, Y.; Jasinevičius, G.; Labonne, S.; Linkevičius, E.; et al. Significant increase in natural disturbance impacts on European forests since 1950. *Glob. Chang. Biol.* **2023**, *29*, 1359–1376. [CrossRef] [PubMed]
11. Seidl, R.; Thom, D.; Kautz, M.; Martin-Benito, D.; Peltoniemi, M.; Vacchiano, G.; Wild, J.; Ascoli, D.; Petr, M.; Honkaniemi, J.; et al. Forest disturbances under climate change. *Nat. Clim. Chang.* **2017**, *7*, 395–402. [CrossRef]
12. Lindner, M.; Maroschek, M.; Netherer, S.; Kremer, A.; Barbati, A.; Garcia-Gonzalo, J.; Seidl, R.; Delzon, S.; Corona, P.; Kolström, M.; et al. Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. *For. Ecol. Manag.* **2010**, *259*, 698–709. [CrossRef]
13. Allen, C.D.; Macalady, A.K.; Chenchouni, H.; Bachelet, D.; McDowell, N.; Vennetier, M.; Kitzberger, T.; Rigling, A.; Breshears, D.D.; Hogg, E.T.; et al. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *For. Ecol. Manag.* **2010**, *259*, 660–684. [CrossRef]
14. Hallas, T.; Steyrer, G.; Laaha, G.; Hoch, G. Two unprecedented outbreaks of the European spruce bark beetle, *Ips typographus* L. (Col., Scolytinae) in Austria since 2015: Different causes and different impacts on forests. *Cent. Eur. For. J.* **2024**, *70*, 1–12. [CrossRef]
15. Hlásny, T.; König, L.; Krokene, P.; Lindner, M.; Montagné-Huck, C.; Müller, J.; Qin, H.; Raffa, K.F.; Schelhaas, M.-J.; Svoboda, M.; et al. Bark Beetle Outbreaks in Europe: State of Knowledge and Ways Forward for Management. *Curr. For. Rep.* **2021**, *7*, 138–165. [CrossRef]
16. Kautz, M.; Meddens, A.J.H.; Hall, R.J.; Arneth, A. Biotic disturbances in Northern Hemisphere forests—A synthesis of recent data, uncertainties and implications for forest monitoring and modelling. *Glob. Ecol. Biogeogr.* **2017**, *26*, 533–552. [CrossRef]
17. Ritzer, E.; Schebeck, M.; Kirisits, T. The pine pathogen *Diplodia sapinea* is associated with the death of large Douglas fir trees. *For. Pathol.* **2023**, *53*, e12823. [CrossRef]
18. Fassnacht, F.E.; Latifi, H.; Stereńczak, K.; Modzelewska, A.; Lefsky, M.; Waser, L.T.; Straub, C.; Ghosh, A. Review of studies on tree species classification from remotely sensed data. *Remote Sens. Environ.* **2016**, *186*, 64–87. [CrossRef]
19. Hallas, T.; Netherer, S.; Pennerstorfer, J.; Karel, S.; Schadauer, T.; Löw, M.; Baier, P.; Bauerhansl, C.; Kessler, D.; Englisch, M.; et al. The Bark Beetle Dashboard—Towards a Holistic Risk Assessment of Ips Typographus. 2024. Available online: https://rgdoi.net/10.13140/RG.2.2.11420.09603 (accessed on 23 July 2024).
20. Immitzer, M.; Vuolo, F.; Atzberger, C. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
21. Bolyn, C.; Michez, A.; Gaucher, P.; Lejeune, P.; Bonnet, S. Forest mapping and species composition using supervised per pixel classification of Sentinel-2 imagery. *BASE* **2018**, *22*, 172–187. [CrossRef]
22. Persson, M.; Lindberg, E.; Reese, H. Tree Species Classification with Multi-Temporal Sentinel-2 Data. *Remote Sens.* **2018**, *10*, 1794. [CrossRef]
23. Puletti, N.; Chianucci, F.; Castaldi, C. Use of Sentinel-2 for forest classification in Mediterranean environments. *Ann. Silvic. Res.* **2018**, *42*, 32–38. [CrossRef]
24. Wessel, M.; Brandmeier, M.; Tiede, D. Evaluation of Different Machine Learning Algorithms for Scalable Classification of Tree Types and Tree Species Based on Sentinel-2 Data. *Remote Sens.* **2018**, *10*, 1419. [CrossRef]

25. Grabska, E.; Hostert, P.; Pflugmacher, D.; Ostapowicz, K. Forest Stand Species Mapping Using the Sentinel-2 Time Series. *Remote Sens.* **2019**, *11*, 1197. [CrossRef]
26. Hościło, A.; Lewandowska, A. Mapping Forest Type and Tree Species on a Regional Scale Using Multi-Temporal Sentinel-2 Data. *Remote Sens.* **2019**, *11*, 929. [CrossRef]
27. Immitzer, M.; Neuwirth, M.; Böck, S.; Brenner, H.; Vuolo, F.; Atzberger, C. Optimal Input Features for Tree Species Classification in Central Europe Based on Multi-Temporal Sentinel-2 Data. *Remote Sens.* **2019**, *11*, 2599. [CrossRef]
28. Grabska, E.; Frantz, D.; Ostapowicz, K. Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the Polish Carpathians. *Remote Sens. Environ.* **2020**, *251*, 112103. [CrossRef]
29. Bjerreskov, K.S.; Nord-Larsen, T.; Fensholt, R. Classification of Nemoral Forests with Fusion of Multi-Temporal Sentinel-1 and 2 Data. *Remote Sens.* **2021**, *13*, 950. [CrossRef]
30. Kollert, A.; Bremer, M.; Löw, M.; Rutzinger, M. Exploring the potential of land surface phenology and seasonal cloud free composites of one year of Sentinel-2 imagery for tree species mapping in a mountainous region. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *94*, 102208. [CrossRef]
31. Xi, Y.; Ren, C.; Tian, Q.; Ren, Y.; Dong, X.; Zhang, Z. Exploitation of Time Series Sentinel-2 Data and Different Machine Learning Algorithms for Detailed Tree Species Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7589–7603. [CrossRef]
32. Zagajewski, B.; Kluczek, M.; Raczko, E.; Njegovec, A.; Dabija, A.; Kycko, M. Comparison of Random Forest, Support Vector Machines, and Neural Networks for Post-Disaster Forest Species Mapping of the Krkonoše/Karkonosze Transboundary Biosphere Reserve. *Remote Sens.* **2021**, *13*, 2581. [CrossRef]
33. Hemmerling, J.; Pflugmacher, D.; Hostert, P. Mapping temperate forest tree species using dense Sentinel-2 time series. *Remote Sens. Environ.* **2021**, *267*, 112743. [CrossRef]
34. Lechner, M.; Dostálová, A.; Hollaus, M.; Atzberger, C.; Immitzer, M. Combination of Sentinel-1 and Sentinel-2 Data for Tree Species Classification in a Central European Biosphere Reserve. *Remote Sens.* **2022**, *14*, 2687. [CrossRef]
35. Delwart, S. ESA SENTINEL-2 User Handbook. 2015. Available online: https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook (accessed on 12 February 2024).
36. Löw, M.; Koukal, T. Phenology Modelling and Forest Disturbance Mapping with Sentinel-2 Time Series in Austria. *Remote Sens.* **2020**, *12*, 4191. [CrossRef]
37. Bolyn, C.; Lejeune, P.; Michez, A.; Latte, N. Mapping tree species proportions from satellite imagery using spectral–spatial deep learning. *Remote Sens. Environ.* **2022**, *280*, 113205. [CrossRef]
38. Waser, L.T.; Rüetschi, M.; Psomas, A.; Small, D.; Rehush, N. Mapping dominant leaf type based on combined Sentinel-1/-2 data—Challenges for mountainous countries. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 209–226. [CrossRef]
39. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
40. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **2020**, *11*, 4540. [CrossRef] [PubMed]
41. Wadoux, A.M.J.-C.; Heuvelink, G.B.M.; De Bruin, S.; Brus, D.J. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol. Model.* **2021**, *457*, 109692. [CrossRef]
42. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
43. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778. [CrossRef]
45. Climate Austria: Average Temperature, Weather by Month & Weather for Austria. Available online: https://en.climate-data.org/europe/austria-4/?utm_content=cmp-true (accessed on 12 February 2024).
46. Klimamittel—ZAMG. Available online: https://www.zamg.ac.at/cms/de/klima/klimauebersichten/klimamittel-1971-2000 (accessed on 12 February 2024).
47. Zampieri, M.; Scoccimarro, E.; Gualdi, S. Atlantic influence on spring snowfall over the Alps in the past 150 years. *Environ. Res. Lett.* **2013**, *8*, 034026. [CrossRef]
48. Savitzky, A.; Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
49. Xue, J.; Su, B. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *J. Sens.* **2017**, *2017*, 1353691. [CrossRef]
50. Kaufman, Y.J.; Tanre, D. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 261–270. [CrossRef]

51. Ahamed, T.; Tian, L.; Zhang, Y.; Ting, K.C. A review of remote sensing methods for biomass feedstock production. *Biomass Bioenergy* **2011**, *35*, 2455–2469. [CrossRef]
52. Qiu, B.; Zou, F.; Chen, C.; Tang, Z.; Zhong, J.; Yan, X. Automatic mapping afforestation, cropland reclamation and variations in cropping intensity in central east China during 2001–2016. *Ecol. Indic.* **2018**, *91*, 490–502. [CrossRef]
53. Mandlburger, G.; Wenzel, K.; Spitzer, A.; Haala, N.; Glira, P.; Pfeifer, N. Improved topographic models via concurrent airborne lidar anddense image matching. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *IV-2/W4*, 259–266. [CrossRef]
54. Trimble Inpho | Office Software. Available online: https://geospatial.trimble.com/products/software/trimble-inpho (accessed on 25 January 2024).
55. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
56. Gschwantner, T.; Berger, A.; Büchsenmeister, R.; Hauk, E. Austria. In *National Forest Inventories*; Vidal, C., Alberdi, I.A., Hernández Mateo, L., Redmond, J.J., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 135–157. [CrossRef]
57. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Adaptive Computation and Machine Learning; The MIT Press: Cambridge, MA, USA, 2016; 775p.
58. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013.
59. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
60. Austrian National Forest Inventory—Tree Species Map. 2024. Available online: https://www.waldinventur.at/?x=1486825&y=6059660&z=7.75968&r=0&l=1111#/map/1/mBaumartenkarte/Bundesland/erg9 (accessed on 25 January 2024).
61. Figueira, A.; Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **2022**, *10*, 2733. [CrossRef]
62. S. Clerc & MPC Team. L1C Data Quality Report. 2022. Available online: https://sentinel.esa.int/documents/247904/685211/Sentinel-2_L1C_Data_Quality_Report (accessed on 20 June 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.