Received: 21 August 2024 | Accepted: 18 December 2024

DOI: 10.1111/2041-210X.14503

Methods in Ecology and Evolution

RESEARCH ARTICLE

Benchmarking tree species classification from proximally sensed laser scanning data: Introducing the FOR-species20K dataset

Stefano Puliti[1] | Emily R. Lines[2] | Jana Müllerová[3] | Julian Frey[4] | Zoe Schindler[4] | Adrian Straker[5] | Matthew J. Allen[2] | Lukas Winiwarter[6] | Nataliia Rehush[7] | Hristina Hristova[7] | Brent Murray[8] | Kim Calders[9] | Nicholas Coops[8] | Bernhard Höfle[10] | Liam Irwin[8] | Samuli Junttila[11] | Martin Krůček[12] | Grzegorz Krok[13] | Kamil Král[12] | Shaun R. Levick[14] | Linda Luck[15,16] | Azim Missarov[12] | Martin Mokroš[17] | Harry J. F. Owen[2] | Krzysztof Stereńczak[13] | Timo P. Pitkänen[18] | Nicola Puletti[19] | Ninni Saarinen[11] | Chris Hopkinson[20] | Louise Terryn[9] | Chiara Torresan[21] | Enrico Tomelleri[22] | Hannah Weiser[9] | Rasmus Astrup[1]

Correspondence
Stefano Puliti
Email: stefano.puliti@nibio.no

Funding information
Austrian Science Fund, Grant/Award Number: J4672; UK Research and Innovation, Grant/Award Number: EP/S022961/1 and MR/T019832/1; Norges Forskningsråd, Grant/Award Number: 309671; Research Council of Finland, Grant/Award Number: 346382 and 357906; European Cooperation in Science and Technology, Grant/Award Number: CA20118; European Commission, Grant/Award Number: 101039795; Narodowe Centrum Badań i Rozwoju, Grant/Award Number: BIOSTRATEG1/267755/4/NCBR/2015; Deutsche Forschungsgemeinschaft, Grant/Award Number: FR 4404/1-1

Handling Editor: Pietro Milanesi

Abstract

1. Proximally sensed laser scanning presents new opportunities for automated forest ecosystem data capture. However, a gap remains in deriving ecologically pertinent information, such as tree species, without additional ground data. Artificial intelligence approaches, particularly deep learning (DL), have shown promise towards automation. Progress has been limited by the lack of large, diverse, and, most importantly, openly available labelled single-tree point cloud datasets. This has hindered both (1) the robustness of the DL models across varying data types (platforms and sensors) and (2) the ability to effectively track progress, thereby slowing the convergence towards best practice for species classification.
2. To address the above limitations, we compiled the FOR-species20K benchmark dataset, consisting of individual tree point clouds captured using proximally sensed laser scanning data from terrestrial (TLS), mobile (MLS) and drone laser scanning (ULS). Compiled collaboratively, the dataset includes data collected in forests mainly across Europe, covering Mediterranean, temperate and boreal biogeographic regions. It includes scattered tree data from other continents, totaling over 20,000 trees of 33 species and covering a wide range of tree sizes and forms. Alongside the release of FOR-species20K, we benchmarked seven

For affiliations refer to page 15.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2025 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

leading DL models for individual tree species classification, including both point cloud (PointNet++, MinkNet, MLP-Mixer, DGCNNs) and multi-view 2D-based methods (SimpleView, DetailView, YOLOv5).
3. 2D Image-based models had, on average, higher overall accuracy (0.77) than 3D point cloud-based models (0.72). Notably, the performance was consistently >0.8 across scanning platforms and sensors, offering versatility in deployment. The top-scoring model, DetailView, demonstrated robustness to training data imbalances and effectively generalized across tree sizes.
4. The FOR-species20K dataset represents an important asset for developing and benchmarking DL models for individual tree species classification using proximally sensed laser scanning data. As such, it serves as a crucial foundation for future efforts to classify accurately and map tree species at various scales using laser scanning technology, as it provides the complete code base, dataset, and an initial baseline representative of the current state-of-the-art of point cloud tree species classification methods.

KEYWORDS
biodiversity, deep learning, lidar, point cloud classification, remote sensing, single-tree inventory

1 | INTRODUCTION

In recent years, there has been a substantial push towards automating the retrieval of key forest variables from various remotely sensed data, with laser scanning technology providing the most detailed and accurate 3D information (Calders et al., 2020). Laser scanning has demonstrated exceptional capabilities in capturing forest structure, with platforms suited to various applications—from large-scale forest inventories using airborne laser scanning (ALS) to smaller in-situ surveys with laser scanning from uncrewed aerial vehicles (ULS), terrestrial laser scanning (TLS) and mobile laser scanning (MLS).

Over the past decade, the value of laser scanning for increased understanding of the structure and function of forests has been widely acknowledged (Calders et al., 2020; Disney, 2019; Krůček et al., 2019; Lines, Fischer, et al., 2022; Malhi et al., 2018). The classification of tree species from laser scanning data is a crucial task for forest monitoring, management, and assessment of forest functions and ecosystem services. Species information is key for estimating carbon storage and sequestration, growing stock volume (i.e. species-specific allometries), wood quality and properties, demography and dynamics, biodiversity and habitat quality, properties which form the basis for successful forest management and conservation (Lines, Allen, et al., 2022). Various individual tree species classifiers have been proposed (see Table 1) across all laser scanning modalities, (i.e. TLS, MLS, ULS and ALS). Recently, deep learning (DL) models such as PointNet++ and multi-view image classifiers have mostly replaced shallower classifiers like Random Forests (Marinelli et al., 2022; Xi et al., 2020).

A review of 19 studies (assessed in July 2024; see Table 1) on the current state of tree species classification from individual tree point clouds identified several gaps:

- Geographic, species and ecosystem diversity in studies: Most studies have been conducted in mature and relatively uniform and simple forest types, predominantly in China and Finland. These included managed plantation forests in temperate and boreal climates with a limited number of species, and limited variation in stand conditions and corresponding diversity of intra-specific crown shapes. Only the studies by Liu, Chen, et al. (2022) and Allen et al. (2023) included complex forest structures with multi-layered canopies with intermingled crowns.
- Dataset size and diversity: Except for the early work by Guan et al. (2015) and Hovi et al. (2016), most studies used small (averaging around 1500 trees) and homogeneous datasets, limiting the transferability to new data. Dataset-specific factors—such as scanning platform, sensor and protocol—significantly affect data structure and quality (e.g. occlusion, resolution, noise) (Lines, Fischer, et al., 2022), which can strongly impact model weights and limit their applicability to particular forest types and specific scanning conditions.
- Platform and sensor-agnostic models: Most studies use single data modalities (ALS, ULS, MLS, or TLS), limiting understanding of cross-platform performance. Recent work by Krisanski et al. (2021) and Wielgosz et al. (2024) demonstrated the potential of using sensor- and platform-agnostic models for forest point cloud segmentation tasks, suggesting benefits for tree species classification. Such models could streamline the

TABLE 1 Summary of the reviewed literature in the domain of individual tree species classification using laser scanning data.

	Dataset					Methods and results	
Reference	Platform	Country	Forest type	No. trees	No. species	Specific classifier	Overall accuracy
Puttonen et al. (2011)	MLS	Finland	Urban	133	10	SVM	0.65
Guan et al. (2015)	MLS	China	Urban	52,013	10	SVM	0.85
Hovi et al. (2016)	ALS	Finland	Boreal	13,560	3	QDA	0.84–0.91
Zou et al. (2017)	TLS	China	Plantation	NA	NA	Voxel CNN	0.93–0.95
Mizoguchi et al. (2017)	TLS	Japan	Temperate	NA	2	2D CNN (bark depth images)	0.85–0.91
Åkerblom et al. (2017)	TLS	Finland	Boreal	1010	3	KNN	0.97
						MLR	0.95
						SVM	0.97
Terryn et al. (2020)	TLS	UK	Temperate broadleaved	788	5	RF, KNN, MLR, SVM	0.82
Xi et al. (2020)	TLS	Canada and Finland	Boreal and temperate	771	9	RF	0.91
						Voxel CNN (ResNet-50,	0.86
						Inception-ResNet-v2)	0.93
						3D CNN (PointNet++ with leaf-wood information)	0.96
Liu et al. (2021)	TLS	China (Mongolia)	NA	1200	2	3D CNN (PointNet)	0.92
	ULS						0.89
Seidel et al. (2021)[b]	TLS	Germany and US	Temperate	690	7	Multi-view 2D CNN (LeNet5; 10 images per tree)	0.86
						3D CNN (PointNet)	–
Lv et al. (2021)	ULS	China	Temperate broadleaved	NA	4	3D CNN (PointNet++ with hand-crafted features)	0.87
Chen et al. (2021)	TLS/ULS	China	Plantation, temperate broadleaved	1000	2	PCTSCN	0.89–0.94
						3D CNN (PointNet)	0.83–0.89
						3D CNN (PointNet++)	0.89–0.92
						Voxel CNN (VoxNet)	0.81–0.85
						2D CNN (ResNet101)	0.88–0.92
Liu, Chen, et al. (2022)	MLS	China	Boreal, temperate, subtropical	1312	8	3D CNN (PointNet++)	>0.95
Liu, Huang, et al. (2022)	TLS	China	Temperate	526	7	3D CNN (PointNet)	0.28
						3D CNN (PointNet++)	0.88
						3D CNN (PointMLP)	0.84
Marinelli et al. (2022)	ALS	Italy	Mountainous temperate	1216	7	SVM	0.64
						3D CNN (PointNet++)	0.67
						Multi-view 2D CNN (8 images per tree)	0.83

(Continues)

TABLE 1 (Continued)

Reference	Dataset					Methods and results	
	Platform	Country	Forest type	No. trees	No. species	Specific classifier	Overall accuracy
Allen et al. (2023)[c]	TLS	Spain	Mediterranean	2478	5	Multi-view 2D CNN (SimpleView; 6 images per tree)	0.81
Fan et al. (2023)	ALS	China		548	11	3D CNN (PointNet++)	0.92
Hakula et al. (2023)	ALS-HD	Finland	Boreal	5500	4	RF	86.6
Lin et al. (2023)[a]	ALS	Kenya	Tropical savanna	4000	6	Multi-view 2D CNN (SimpleView; 6 images per tree)	0.7
						PCT	0.72

Note: The acronyms for the species classifiers are as follows: support vector machines (SVM), quadratic discriminant analysis (QDA), convolutional neural network (CNN), k nearest neighbour (KNN), random forest (RF), multinomial linear regression (MLR), point cloud transformer (PCT).

Abbreviation: NA, not available.

[a]Open code.

[b]Open data.

[c]Open data and code.

automation of 3D forest data processing and reduce the need for new training datasets with emerging sensor/platform combinations.

- Accessibility of data and code: Strikingly, out of the nineteen studies reviewed, only Seidel et al. (2021) and Allen et al. (2023) made their datasets publicly available and only Allen et al. (2023) and Lin et al. (2023) made the code open. In a realm where accessibility to high-resolution laser scanning data is limited by costs of and accessibility to sensors, the lack of open data greatly limits the potential for innovation.
- Lack of and automation: The ability for laser scanning data to operationally replace traditional ground surveys has been limited by the lack of effective individual tree segmentation methods, which still require significant manual work. Latest advances in forest 3D scene panoptic segmentation (Wielgosz et al., 2024; Xiang et al., 2024) open up new opportunities for streamlining laser scanning forest surveys.

While deep learning (DL) shows promise for individual tree species classification, navigating the landscape of existing models remains challenging and universally applicable models have not yet emerged, limiting the deployment in real-world applications. A key challenge is the difficulty in objectively comparing model performances, as many models are validated on internal test data, often leading to inflated accuracy metrics that fail to reflect real-world conditions. The availability of public benchmark datasets, along with open model weights and DL-ready datasets, is invaluable for advancing the field, enabling objective model development, and helping users understand the strengths and limitations of various approaches. Public leaderboards that compare model performance on these benchmarks are essential for users seeking reliable, automated tree species classification models for forest monitoring and ecosystem service assessments.

This paper addresses the need for improved tree species classification by collating, standardizing, and publishing the FOR-species20K dataset, a large, diverse, individual tree point cloud DL-ready benchmark dataset for platform- (airborne and terrestrial) and sensor- (consumer and survey grade) agnostic tree species classification tasks. We also establish a baseline by benchmarking current state-of-the-art DL models, underlining the importance of moving towards platform- and sensor-agnostic models. We analyse the outputs of the benchmarking exercise not only using standard ML approaches, but also by considering the context. For example, we explore the variation in performance by sensor type and species, showing the need for methods to effectively cope with data imbalances, a common challenge in nature where tree species distributions are often uneven. Finally, we publish all our data, including the test-train-validation split, and our model weights, openly with the aim of encouraging future development.

2 | MATERIALS

The FOR-species20K dataset was compiled by combining 25 different datasets, composed of either open datasets or in-kind contributions.

Data were collated based on directly approaching relevant researchers, the authors' networks, and an open call disseminated via multiple channels, including academic newsletters, conference presentations and social media channels. The basic data unit is individual tree point clouds, so to ensure the quality of the dataset, we selected only submissions where the individual tree segmentation was of high quality (i.e. manual segmentation). The final number of trees included in the FOR-species20K dataset was 20,158.

2.1 | Data origin

The large majority (90% of the trees) of data in the FOR-species20K were collected within Europe, with additional scattered data collections from Canada, Australia and New Zealand (see Figure 1).

2.2 | Platforms and sensors

The FOR-species20K dataset is composed of mostly terrestrial laser scanning (TLS) data (70% of the trees), followed by UAV laser scanning (ULS) (22%), and lastly, mobile laser scanning (MLS) data (8%). Likely due to the longer history of forest use of static laser scanners compared to mobile and airborne solutions, the TLS data herein is characterized by a broader variability in terms of sensors (12 different sensors) and scanning protocols, spanning from fixed position single scan to equally spaced scans and including varying designs in between. On the contrary, MLS and ULS data came from only one type of sensor each. Further details on the different datasets can be found in Table 2.

As a result of such data source heterogeneity, the individual tree point clouds included in the FOR-species20K data exhibit a qualitative variation in terms of resolution, completeness (i.e. occlusion) and measurement accuracy (see Figure 2). Compared to multi-scan TLS data, the MLS and ULS point clouds were characterized by typical occlusions towards the top in MLS data or the lower parts of the canopy in ULS data (Schneider et al., 2019), while the single-scan TLS data show both horizontal and vertical occlusion away from the scanner position.

2.3 | Forest types and tree species

The FOR-species20K data cover the three main forest ecoregions in Europe with most of the available trees coming from temperate forests (61%), followed by boreal forests (25%) and Mediterranean forests (7%). Further, a small percentage of trees was from temperate and boreal plantation forests outside Europe (4%) and tropical savannas (3%).

Compiling the FOR-species20K data, we retained only species with at least 50 individuals represented, resulting in 33 species of 19 genera (Figure 3). Whereas mostly focused on European forest tree species, FOR-species20K includes additional species from a more global ecological and climatic spectrum, including tree species from sclerophyll forests in Australia and broadleaf forests in North America. The most common species include *Pinus sylvestris*, with 3296 individuals, *Fagus sylvatica*, with 2482 samples, and *Picea abies*, with 1983 samples. Conversely, some species are much less represented, such as *Quercus robur* (195 individuals), *Abies alba* (119 individuals), *Larix decidua* (94 individuals), with the rarest, *Prunus avium* having just 50 individuals. While heavily imbalanced, this variation in species representation reflects realistic abundance distributions in European forest ecosystems with European dominant tree species being well represented and rarer species less so. FOR-species20K

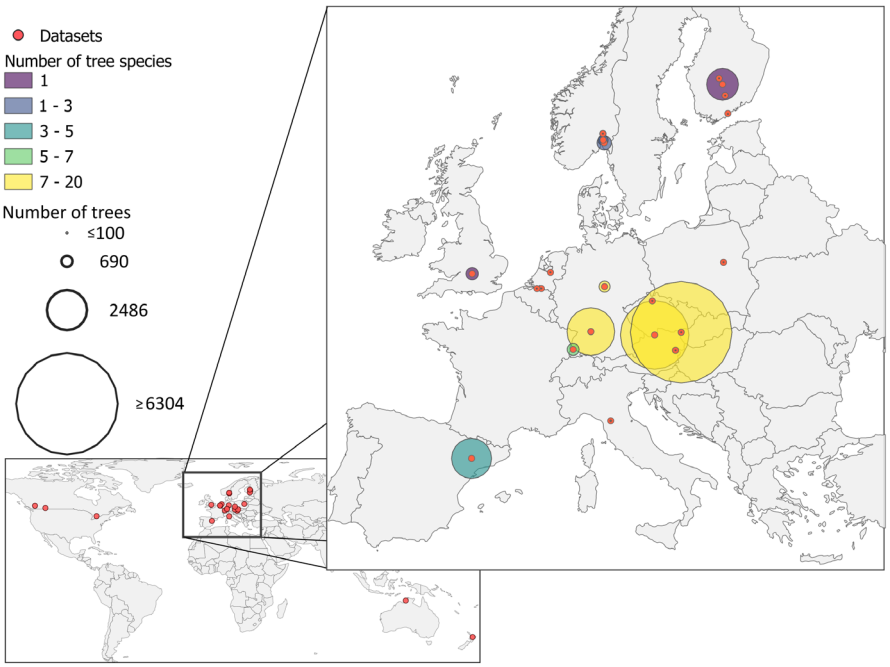


FIGURE 1 Geographic overview of the FOR-species20K dataset.

TABLE 2 Summary of the individual data collections within the FOR-species20K dataset.

Dataset name[a]	Reference	No. trees	No. species	Data type	Sensor	Biogeographic region
BlueCat_a	—	3926	9	TLS	Leica P20	Temperate
Owen_2021	Owen et al. (2021)	2478	5	TLS	Leica HDS6200	Mediterranean
Saarinen_2021	Saarinen et al. (2020); Saarinen, Calders, et al. (2021); Saarinen, Kankare, et al. (2021)	1976	1	TLS	Trimble TX5 3D	Boreal
BlueCat_b	—	1409	3	TLS	Leica P20	Temperate
Calders_2022	Calders et al. (2022)	769	6	TLS	RIEGL VZ-400	Temperate
Xi_2020a	Xi et al. (2020)	661	3	TLS	Optech Ilris HD	Boreal
Seidl_2021	Seidel et al. (2021)	577	7	TLS	Faro Focus 3D 120, Zoller and Fröhlich Imager 5006	Temperate
Frey_2022	—	472	6	TLS	RIEGL VZ-400i	Temperate
Xi_2020_2	Xi et al. (2020)	398	2	TLS	Optech Ilris HD	Temperate
Liang_2018	Liang et al. (2018)	385	3	TLS	Leica HDS6100	Boreal
Luck	—	352	1	TLS	Leica BLK360	Tropical savana
Weiser_2022a	Weiser et al. (2022)	263	11	TLS	RIEGL VZ-400	Temperate
Luke	—	225	3	TLS	Leica P40	Boreal
REMBIOFOR	—	57	3	TLS	FARO Focus 3D X130	Temperate
Junttila	—	51	1	TLS	Leica RTC360	Boreal
VanDeBerge_2021	Van Den Berge et al. (2021)	50	2	TLS	RIEGL VZ-1000	Temperate
Puliti_a	—	895	4	MLS	Geoslam Horizon	Boreal
LAUTx	Tockner et al. (2022)	434	6	MLS	Geoslam Horizon	Temperate
UBC_2022	—	279	2	MLS	Geoslam Horizon	Temperate
Mokros_2022	—	114	1	MLS	Geoslam Horizon	Temperate
Weiser_2022b	Weiser et al. (2022)	2908	15	ULS	RIEGL miniVUX-1UAV	Temperate
Puliti_b	—	621	3	ULS	VUX1-UAV	Boreal
FORinstance_NIBIO	Puliti et al. (2023)	479	4	ULS	RIEGL miniVUX-1UAV	Boreal
FORinstance_SCION	Puliti et al. (2023)	135	1	ULS	RIEGL miniVUX-1UAV	Temperate
FORinstance_CULS	Puliti et al. (2023)	47	1	ULS	RIEGL VUX-1UAV	Temperate

[a]Corresponding to the dataset name in data published in Zenodo (https://zenodo.org/records/13255198).

represents the most comprehensive dataset regarding the number of tree species openly available to date, which is essential for developing and evaluating robust classification models.

2.4 | Dataset variability in tree size and crown architecture

The FOR-species20K dataset captures significant variation in tree height (Figure 3) and crown architecture (examples in Figure 4) encompassing a broad range of tree developmental stages under a variety of growth conditions.

Coniferous species cover a slightly wider height range (min = 0.3 m; mean = 20.4 m; max = 56.3 m, standard deviation = 8.2 m) than broadleaf species (min = 0.3 m; mean = 11.4 m; max = 42.1 m; standard deviation = 8.7 m). This variation in tree height also varies significantly according to species. For dominant European tree species, such as *P. abies*, *F. sylvatica*, and *P. sylvestris*, the dataset covered the full spectrum of forest developmental stages from young saplings to mature trees (see Figure 3). For rarer species such as *P. avium* the availability of data from only a few trees, often from the same stand, resulted in a rather uniform distribution of tree height (see Figure 3). Another source of variability in the FOR-species20K dataset is the significant intraspecific variation in crown architecture (see some visual examples in Figure 4), which is tightly linked to tree growth and competition under different growing conditions. Overall, the FOR-species20K dataset is highly diverse, encompassing a broad range of tree morphological variations. This diversity is a crucial feature for robust model training and evaluation in tree classification tasks.

2.5 | Data split

To benchmark different classifiers, the full dataset was split into development (90% of the trees; 17,707 trees) and test set (10% of

the trees or 2254 trees). The development set is composed of the individual tree point clouds and corresponding species labels and is meant to be used for both model training and validation. In the test set, the tree species labels were withheld from the participants to ensure a fair competition/unbiased evaluation.

To address the challenge of benchmarking the classifiers' ability to handle largely imbalanced data in terms of tree species and size and data type, the test dataset was selected based on a stratified random sampling aimed at creating an artificially balanced dataset according to the following logic:

- Balancing tree species: to avoid overrepresentation of the dominant species, we set a maximum threshold of $n = 100$ trees from each species
- Balancing tree height: to ensure a more balanced representation in terms of tree size and tree developmental stages, we defined a strata A consisting of $i = 20$ tree height bins of 2.8 width.
- Balancing data platforms: to ensure a more homogeneous representation of the different platforms (ULS, MLS, and TLS) in the test data, we defined a strata B, where each statum is defined by the combination of a height bin from strata A with groups defined by the different platforms (i.e. A_i_TLS, A_i_MLS, A_i_ULS), resulting in a varying size j of strata depending of the availability of trees from different platforms across different height bins and species.
- Sampling: for each species and strata B, we randomly sampled a number n_B of trees, where n_B is defined as the ratio between n and the j number of strata in B, thus ensuring the selection of a maximum of n trees for each species. If a species had fewer trees

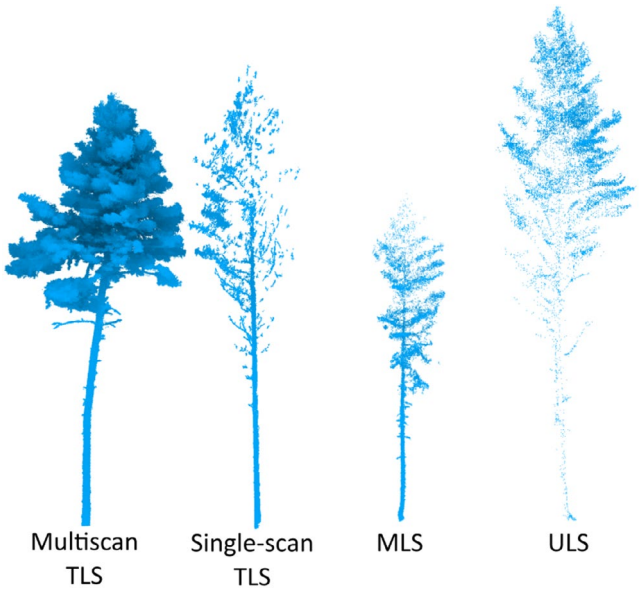


FIGURE 2 Example variation of the point cloud quality across platforms and different scanning protocols (TLS single- or multi-scan, MLS and ULS).

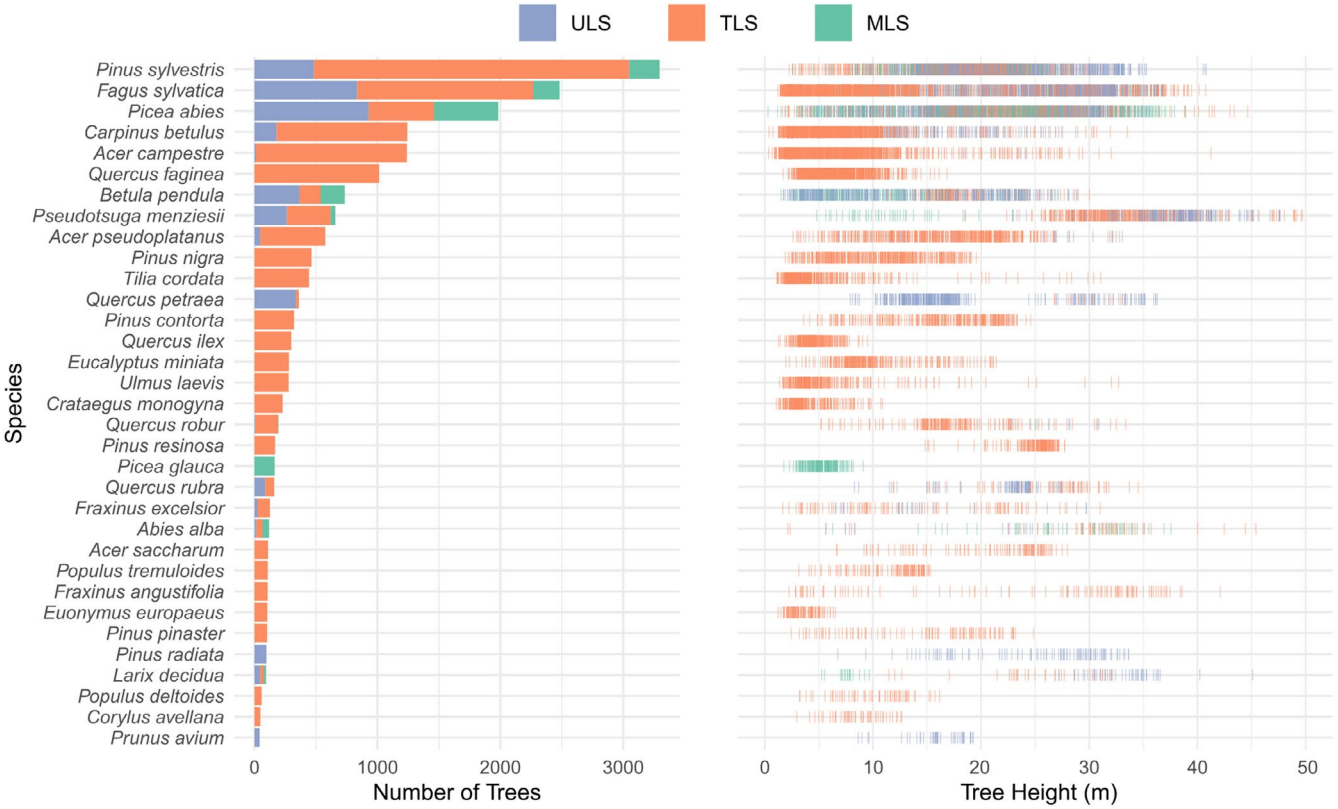


FIGURE 3 Tree species frequency and height distributions across data types in the FOR-species20K training data, split by sensor type.

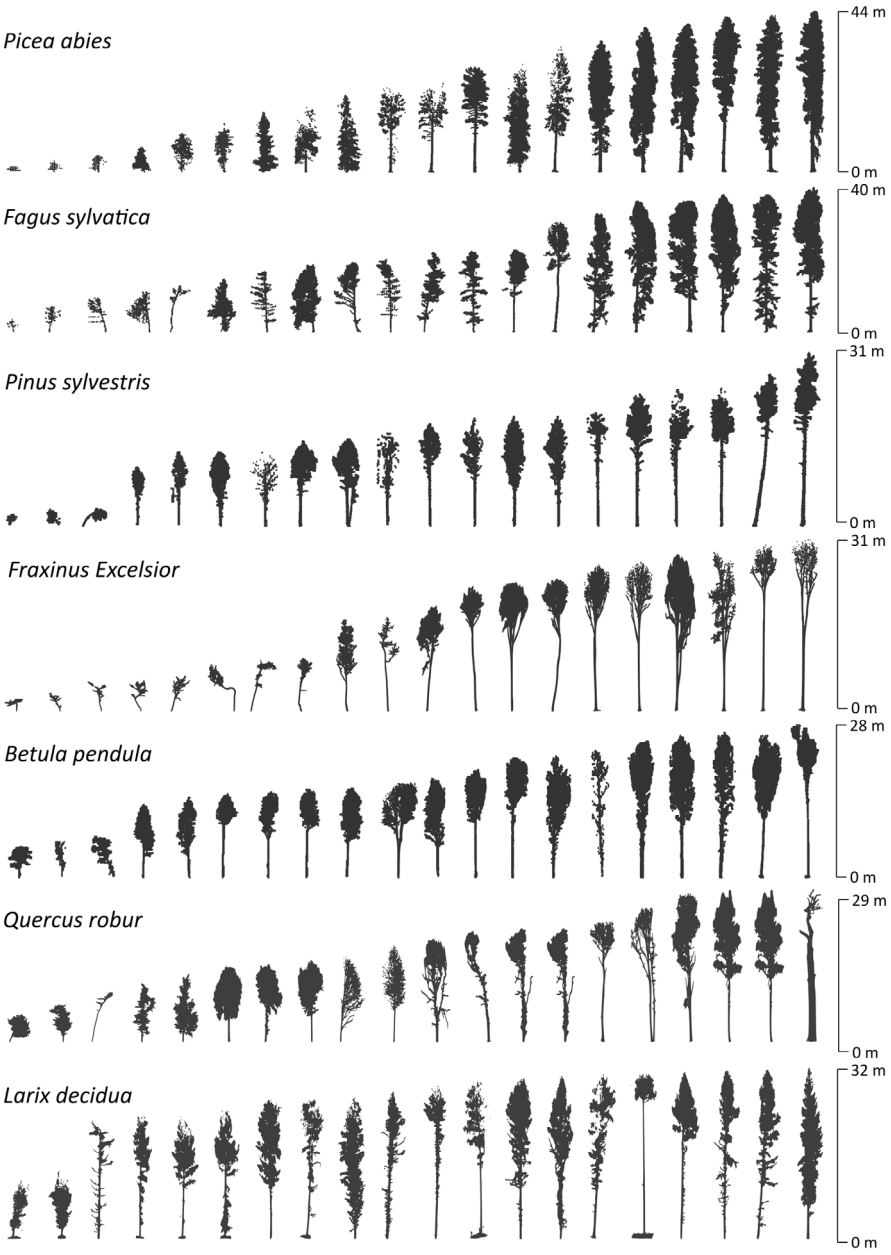


FIGURE 4 Examples of the variety of tree size and crown shape present in the FOR-species20K data for a selection of common European tree species.

in a specific stratum than the target n_B, all available trees in that stratum were included in the sample.

3 | METHODS

3.1 | Data science competition

Using the above-described dataset, a data science competition was launched in November of 2022 and lasted until June 2023 with the intention of benchmarking some of the most common classifiers available in the literature, including some of the latest model architectures. The competition was advertised in the same wide variety of channels as our call for data. Competitors tested seven different DL methods, including methods operating either directly on 3D point clouds or on 2D images obtained from projecting point clouds. Table 3 and the following subsections provide a high-level understanding of the benchmarked methods, their similarities, and their differences. Further implementation details can be found in Appendix S1.

3.1.1 | Point-cloud-based methods

PointAugment and DGCNN

This method combined PointAugment (Li et al., 2020), a generative adversarial network (GAN), and the Dynamic Graph Convolutional Neural Network (DGCNN; Wang et al., 2019). PointAugment enhances the point clouds to address species imbalance by generating more complex yet similar shapes. DGCNN, known for its ability to create and modify graph connections, aggregates these relationships for classification. Point clouds were down-sampled to 4096

TABLE 3 Summary of the main characteristics of the benchmarked DL architectures, including the input data, how the development data was split into training (train) and validation (val) data, the type of augmentation techniques adopted within the model, and specific characteristics related to the inference, or prediction step.

Name	Data input	Train/val data split	Augmentation	Inference
DGCNN + PointAugment	3D point clouds (4096 pts. tree^{-1})	Tree size and tree species stratified 90% train; 10% val	PointAugment + down-sampling, noise, rotation	Best model applied to test data
Ensemble PointNet++	3D point clouds (8192 pts. tree^{-1})	Same split as above	6-fold rotation (z-Axis), random sampling	Ensemble classifiers, max avg. prob.
MinkNet	Voxelized point clouds (>200 and <16,384 pts. tree^{-1})	Tree size and platform type stratified 90% train; 10% val	Random rotation along Z-axis	Majority voting after 50 rotations
PointMixer	3D point clouds (4098 pts. tree^{-1})	Tree size and platform type stratified 90% train; 10% val	None	Soft voting with 100 iterations
SimpleView	6×2D projected images (512×512 pixels)	Random 90% train; 10% val	None	Accuracy on validation set
DetailView	7×2D projected depth images (256×256 pixels)	Weighted random sampler for the training data (98%), furthest distance sampling for the validation data (2%)	Point cloud random subsampling and rotation; Image flip	Averaged probabilities from 50 runs
YOLOv5	4×2D projected images (600×800 pixels)	Tree species stratified 90% train; 10% val	YOLOv5 augmentations	Weighted mean of class probabilities

points per tree, and manual augmentation was applied to balance the species. A species-wise stratified random subsample of 10% of the training trees was withheld from training and used as validation data for hyperparameter tuning and identification of the best model over the entire training period. The code for this method can be found in the repository stored in the submission_data_code/code/PointAugment_DGCNN.zip file.

Ensemble PointNet++

PointNet++ (Qi et al., 2017) was utilized to extract features directly from 3D point clouds through three sets of subsampling and grouping operations. Point clouds were down-sampled to 8192 points per tree, and augmentations involved rotational augmentation (6-fold around the z-Axis) and random point sampling, improving prediction quality by using an ensemble of 10 classifiers. The approach tested different configurations, finding that repeated random sampling and rotation significantly enhanced accuracy. The training involved various epochs based on the ensemble configuration and was done using the same split as the one used in the PointAugment and DGCNN method above. The code for this method can be found in the repository stored in the submission_data_code/code/Ensemble_PointNet.zip file.

MinkNet

MinkNet (Choy et al., 2019; https://github.com/NVIDIA/MinkowskiEngine) employs 3D sparse convolutions on voxelized point clouds using the Minkowski Engine framework. The model was calibrated with a 90% train, 10% validation split stratified by tree size and platform type and considering only trees with >200 points, using a voxel size of 0.1 m. Training involved 250 epochs with data augmentation through random rotation along the Z-axis, resulting in a robust classification performance. During inference on the test data, the model was applied to a sample 50 times while each time the point cloud was randomly rotated along the Z-axis. The final species was assigned based on majority voting (the most common label was chosen). The code for this implementation can be found in the repository stored in the submission_data_code/code/MinkNet.zip file.

PointMixer

PointMixer (Choe et al., 2022) blends features within and between point sets, making it effective for tree species classification. The training set was stratified to ensure a representative sample across different tree sizes and platform types. Using Farthest Point Sampling to 4098 points per tree, the input points were selected and classified through soft voting. The training was set to 300 epochs and the best model (epoch 269) was selected using the validation overall accuracy. Inference was done using soft voting over 100 iterations with 10 votes per iteration. The predicted probabilities for class labels were summed up and the final class label was chosen based on the highest sum value. The code for this implementation can be found in the repository stored in the submission_data_code/code/PointMixer.zip file.

3.1.2 | Image-based methods

All of the tested image-based methods relied on multi-view approaches, whereby the species classification is seen as an image classification task applied to different images of the same tree generated by projecting the point clouds onto 2D planes from different viewpoints.

SimpleView

SimpleView (Goyal et al., 2021) uses a multi-view approach with six orthogonal camera projections of point clouds. Six projected images,

coloured by depth, were used to train a ResNet-18 backbone (He et al., 2016). Previously applied to the tree species classification task by Allen et al. (2023), in this version, we modified the method to include larger images (from 256×256 pixels to 512×512 pixels), and down-sampling of the point clouds to 16,384 points. The model was trained on a simple random sample of 90% of the trees, and to account for data imbalance, the best model was selected as the one maximizing the validation (10% of the trees) balanced accuracy rather than the overall accuracy as in the original implementation. The code for this implementation can be found in the repository stored in the submission_data_code/code/SimpleView.zip file.

DetailView

DetailView, builds upon SimpleView and incorporates dataset balancing by species, tree size, and platform through weighted random sampling. Further, DetailView uses a DenseNet-201 backbone (Iandola et al., 2014) and adds top and bottom views and a high-resolution projection of the trunk to leverage bark structure for classification. Only point clouds with at least 100 points were used for training. Image size was set to 256×256 pixels and augmentations included both point cloud (random subsampling and rotations) and image methods (flip) which were applied directly within the model. Further, tree size was included in the classification. The final predictions were obtained by averaging the 50 predicted probabilities per class and choosing the respective class with the maximum average probability. The code for this implementation can be found in the repository stored in the submission_data_code/code/DetailView.zip file.

YOLOv5

The general concept of this approach is based on the application of a modified YOLOv5 architecture (Jocher et al., 2022) and the use of four side view orthographic projections (600×800 pixels) coloured by point count. We randomly split the training data set into 90% for training and 10% for validation. This split was conducted per species to ensure a representation of all species in both data sets. The final classification model was trained for 46 epochs using a stochastic gradient descent optimizer, a batch size of 128, and using default hyperparameters in the official YOLOv5 release. The model with the highest overall accuracy on the validation set was selected as the best model. The final tree species predictions were obtained by calculating a weighted average of the 20 predicted class probabilities per tree using weights of 0.5, 0.35, 0.05, 0.05 and 0.05 for the highest five predicted class probabilities per side view image, respectively and selecting the class with the highest weighted average probability. The code for this method can be found in the repository stored in the submission_data_code/code/YOLOv5zip file.

3.2 | Benchmarking and metrics

The presented models were benchmarked in relation to their ability to classify tree species in the unseen test data. The ground truth (GT) and predicted tree (PT) species were used to generate the confusion matrices based on which we obtained the counts for the true positives (TP), false positives (FP), and false negatives (FN) required to compute a selection of commonly used metrics for the evaluation of classification tasks. The following metrics were used to evaluate the species-wise and global model's predictive accuracy:

$$\text{Overall accuracy (OA)} = \frac{\text{TP}}{\text{GT}}, \tag{1}$$

$$\text{Species accuracy } (\text{SA}_s) = \frac{\text{TP}_s}{\text{GT}_s}, \tag{2}$$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3}$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}. \tag{5}$$

Further, to address the ability of the model to get a fully agnostic understanding of the task we evaluated the overall accuracy for the different data types (TLS, MLS, and ULS), and performance across different tree sizes.

For future benchmarking against the FOR-species20K test data, we established a Codabench (Xu et al., 2022) benchmarking page which will be made openly available in the future.

4 | RESULTS AND DISCUSSION

4.1 | Benchmarking

The leaderboard for the results of the data science competition is shown in Table 4. The best-performing model by overall accuracy (OA), recall and F1-score was DetailView, while the best-performing model by precision was YOLOv5, with DetailView a close second. We therefore interpret DetailView as the best-performing model overall. We found a notable performance disparity between multi-view image-based methods (average OA = 77.8%) and point cloud methods (average OA = 72.1%), with the worst-performing image-based method outperforming the best-performing point cloud method.

The performance disparity between image and point cloud methods may be due to the current technology readiness level of the two methods, with image-based approaches benefitting from the matured feature extraction capabilities of CNNs and traditional image processing techniques such as extensive data augmentation and the simplification of data through 2D projections. These techniques simplify the input data structure by making it uniform (pixel grid) and allow for the projection of many more points than allowed by point cloud methods thus allowing for a

TABLE 4 Leaderboard for the data science competition described in this study, ordered by overall accuracy.

Tested method	Input data	Overall accuracy (%)	Precision (%)	Recall (%)	F1 (%)
DetailView	Image	**79.5**	82.3	**76.7**	**78.0**
YOLOv5	Image	77.9	**84.2**	75.0	77.3
SimpleView	Image	76.2	76.9	75.5	75.6
Ensemble PointNet++	Point cloud	75.6	78.2	73.5	74.9
MinkNet	Point cloud	73.7	79.9	70.6	72.3
PointMixer	Point cloud	71.1	74.4	65.5	71.1
PointAugment + DGCNN	Point cloud	68.3	72.5	65.7	70.3

Note: The bold values represent the best results on the FOR-species20K test dataset for each metric.

more efficient data compression and digestion by the model. Another potential advantage of multi-view CNNs is that the final prediction is derived from an ensemble of predictions across different projections, which may enhance the accuracy and reliability of the results.

In contrast, point cloud-based methods encountered challenges stemming from the sparse and unstructured nature of the 3D data. Although architectures like PointNet++ and DGCNN are designed to tackle these problems, they are relatively recent developments and continue to face difficulties related to high computational demands and, therefore, the need to heavily subsample the original point clouds. Such information loss is likely one of the causes of the poorer performance of the point cloud methods. On the other hand, we expect point cloud methods to have a larger potential for improvements over time, since they are starting from a lower baseline. Thus, although we found that image-based methods represent the current state-of-the-art in terms of classification performance, it is likely that point cloud architectures will attract more development and may even outperform 2D methods in the future.

The confusion matrix of the best-performing method, DetailView (Figure 5), shows that, on average, the overall accuracy for coniferous species (87.4%) was larger than for the broadleaved species (71.3%). Such a result is likely influenced by the larger number of broadleaf species (22 species) compared to only 11 coniferous species, as it is typical in classification tasks that performance drops when the number of classes increases. In addition, broadleaved trees typically develop more plastic and asymmetric crowns which substantially increase the intra-specific variation in tree structure and thus the ability of finding common representative features within a species. Conversely, the better performance for coniferous trees may be seen as surprising when taking into account that 7 of the 11 species are from the same genus (*Pinus*). While we found examples of intra-genus confusion (e.g. between *P. pinaster* and *P. nigra* or between *Q. faginea* and *Q. ilex*), these were not noticeably more frequent or severe compared to the confusion amongst species in different genera.

We examined the influence of data imbalances across tree species and datasets on classification accuracy (Figure 6) and found that neither the quantity of data points nor the number of datasets within each species significantly affected the model F1-score (with Pearson's correlation coefficients being −0.03 and 0.06, respectively). While this finding may have been partly driven by a degree of similarity between the development and test trees, it suggests that the DetailView model effectively compensates for data imbalances.

In addition to the above results, a visual analysis of a sample of the incorrectly classified trees (see examples in Figure 7) suggests that the model's performance may influenced by the lack of crown architectural clues (e.g. branches and branching patterns) in small trees. These trees often belong to lower layers in the canopy, where significant competition strongly affects the shape and complexity of tree crown architecture. Thus, the model's performance issues seem to stem more from these architectural complexities and incomplete crown data rather than the sheer volume of data available.

4.2 | Accuracy by acquisition platform

The assessment of the platform-agnostic capabilities of the tested models (see Figure 8) revealed that for the top five methods, the overall accuracy >70% for all platforms. OA was highest for all methods for MLS data, and DetailView performed best for MLS and ULS data, but was outperformed by 0.1% for TLS data by YOLOv5.

The above result shows that not only is tree species classification reliable from each platform, but that all models performed well regardless of the laser scanning data source. This capability is particularly relevant for streamlining the adoption of tree species classifiers in operational settings. By unifying under a single model, users can avoid the complexity of adopting different processing pipelines for the autonomous characterization of forest environments using proximal laser scanned data. Additionally, this ensures consistency in outputs across different data modalities, which is crucial as multiple technologies are increasingly integrated into forest information systems.

Although the experimental results (see Figure 7) generally show higher predictive accuracy for MLS data, these values are likely inflated because MLS data represent only a small portion of the FOR-species20K dataset, both by the number of sample trees and species diversity (just 8 species have MLS data, vs. 30 for TLS; Figure 3).

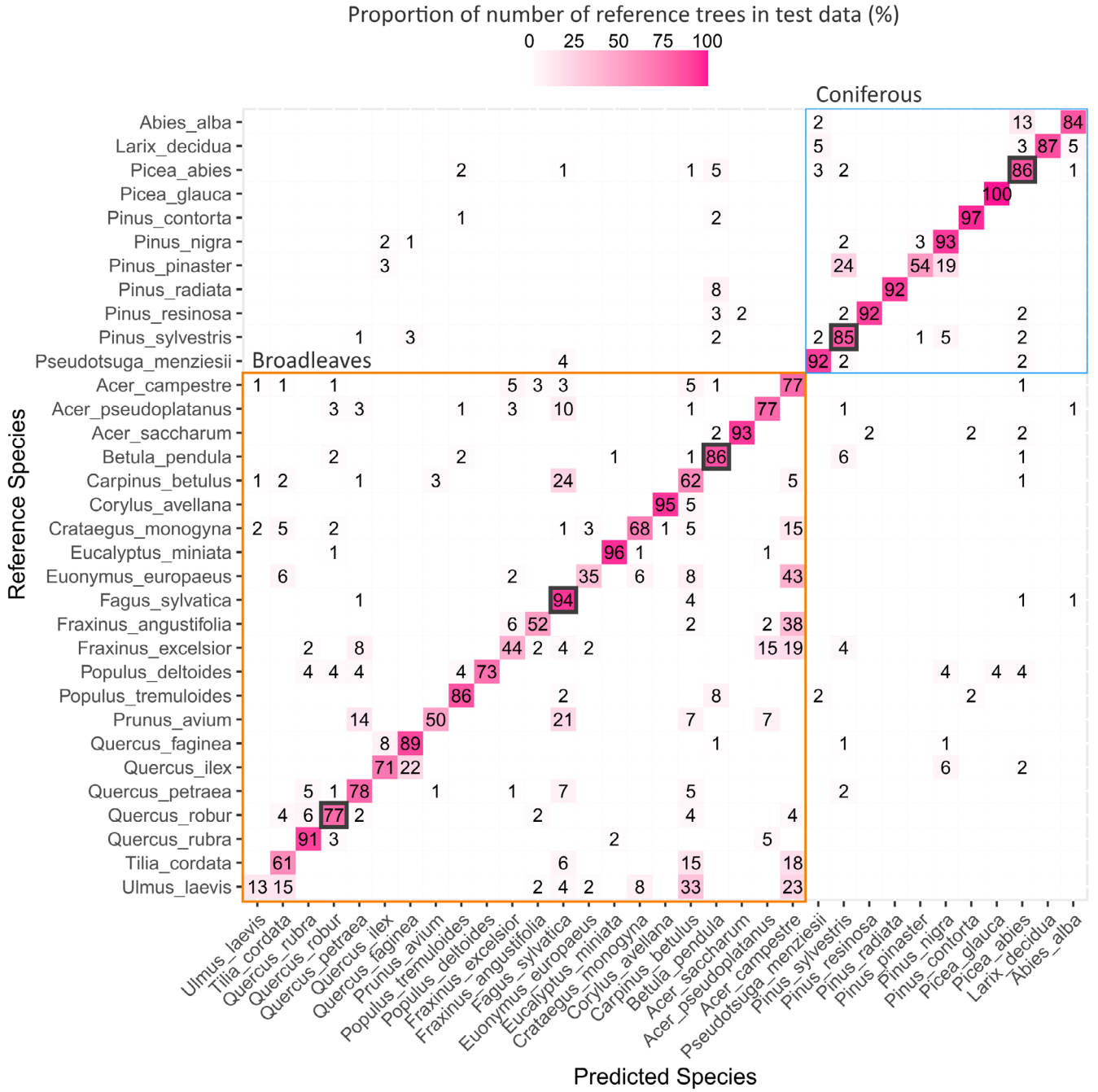


FIGURE 5 Normalized confusion matrix (%) computed using the DetailView method on the withheld test data. The confusion matrix is ordered alphabetically by coniferous and broadleaved species to allow within- and cross-genus comparison. The boxes with black outline highlight the performance on some of the most abundant species in European forests (in bold).

Interestingly, for the sparser ULS data, the PointNet++ ensemble and the Minknet models outperformed YOLOv5 and SimpleView. This suggests that point cloud-based methods may better leverage sparser point representations and face challenges with very dense datasets, or alternatively image methods may struggle where there is less information in the data. Point cloud-based methods often reduce computational demands by subsampling the point cloud, which consequently reduces the richness of the information. The superior performance of DetailView on the rarer MLS and ULS data might be explained by the weighted random sampler used in DetailView (see description in Section 3.1.2), which allows balancing the training and validation data specifically based on the platform type.

4.3 | Accuracy by tree size

The analysis of the impact of tree size on DetailView's overall accuracy showed that for trees taller than around 8 m the accuracy remained relatively stable regardless of tree height (Figure 9). However, accuracy was substantially lower for small trees (<5 m in height), falling as low as 25% for trees shorter than 2 m.

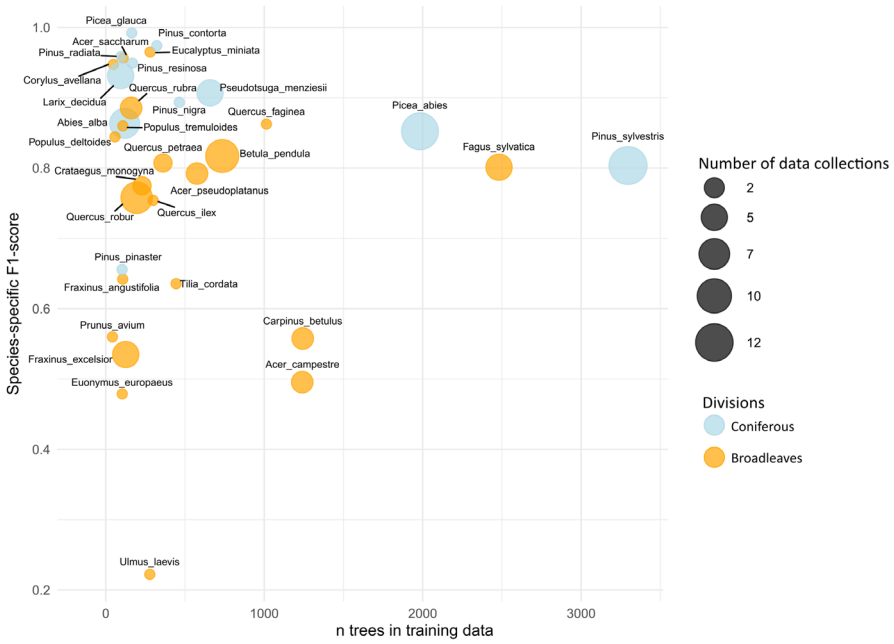


FIGURE 6 Bubble plot representing the relationship between the sample size in the training data and the F1-score obtained in the test data for each species. The size of the bubbles represents the number of datasets that contributed to each of the species. The plot was computed using the results from the best-performing method (DetailView).



FIGURE 7 Examples of misclassified trees for some tree species of forestry relevance in Europe, comparing ground truth (GT) and predicted (Pred.) species using the DetailView method.

4.4 | Domain/application evaluation

Although the primary aim of the FOR-species20K dataset was to objectively benchmark different species classifiers, rather than providing a complete tree species database for training models for operational tree species prediction, we examined three DetailView's key potential application scenarios in mature forest (tree height >5 m) in three European biogeographic regions, including boreal, hemiboreal, and temperate regions (Table 5). Unfortunately, due to the lack of enough sample trees for Mediterranean species such scenario was here omitted.

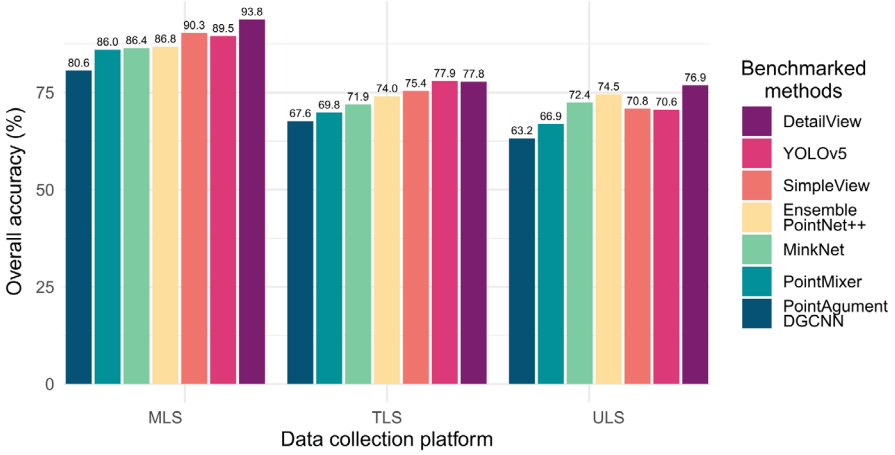


FIGURE 8 Overall accuracy by benchmarked method for the tree laser scanning data collection platform: Mobile (MLS), terrestrial (TLS), and uncrewed aerial vehicle (ULS).

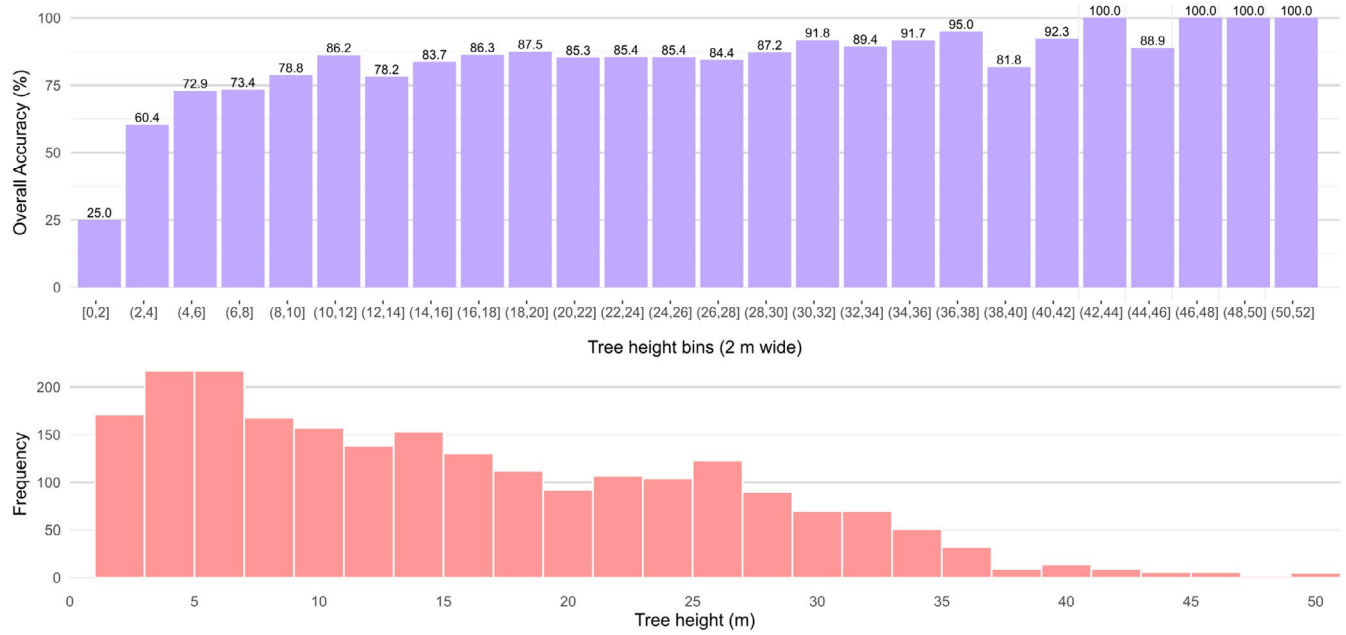


FIGURE 9 DetailView's overall accuracy (%) categorized by tree size with tree heights binned into 2-m intervals.

TABLE 5 DetailView's overall accuracy for three European biogeographic regions with increasing species diversity and across different input data types.

Biogeographic region	Available tree species	Input data type	No. test trees	Overall accuracy (%)
Boreal	*Picea abies*, *Pinus sylvestris*, *Betula pendula*	ULS	87	90.8
		TLS	100	87.0
		MLS	90	88.9
Hemiboreal	*Picea abies*, *Pinus sylvestris*, *Betula pendula*, *Quercus robur*, *Fagus sylvatica*, *Fraxinus excelsior*, *Acer pseudoplatanus*, *Corylus avellana*	ULS	167	79.0
		TLS	280	84.3
		MLS	138	92.8
Temperate	*Picea abies*, *Pinus sylvestris*, *Betula pendula*, *Quercus robur*, *Fagus sylvatica*, *Fraxinus excelsior*, *Acer pseudoplatanus*, *Corylus avellana*, *Pseudotsuga menziesii*, *Abies alba*, *Larix decidua*, *Carpinus betulus*, *Quercus petraea*, *Prunus avium*, *Populus tremuloides*, *Crataegus monogyna*, *Tilia cordata*	ULS	367	75.7
		TLS	517	82.8
		MLS	179	92.7

While valid within the context of the FOR-species20K data, we found that DetailView performs reliably across different biomes, consistently achieving high overall accuracy (>75%) with all data platforms in all ecoregions. Accuracy was highest (87%–90%) in species-poor biomes, such as boreal forests, and decreased as tree species richness increased, as seen in the transition from predominantly coniferous boreal forests to mixed broadleaved temperate forests.

While this study demonstrates DetailView's potential for diverse ecoregions, deploying it in user-friendly operational systems remains a priority. As noted in the introduction, post-processing of laser scanning data remains a bottleneck for many users due to its computational demands and complexity. Recent advancements in deep learning, such as end-to-end tree segmentation frameworks like ForAInet (Xiang et al., 2024) or SegmentAnyTree (Wielgosz et al., 2024), can streamline workflows by automating tree instance extraction from unstructured point clouds. These developments suggest that tools like DetailView could be incorporated into "one-button" solutions, reducing pre-processing needs and making advanced point cloud analytics more accessible to non-specialist ecologists and forest managers. As deep learning architectures evolve, such pipelines will bridge technical complexity with usability, supporting scalable forest ecosystem monitoring and management.

4.5 | Limitations

While these results are promising, users must be aware that these results are likely inflated due to the use of a test dataset split from the same pool of data as the training dataset and thus are assumed to have similar data properties and tree morphology. We suggest users keep in mind the following issues when deploying any model trained on FOR-species20K to new datasets:

- Open-set recognition: The species list is incomplete even for Europe, leading to incorrect predictions for species unseen to the model and likely impacting the real-world performance in diverse forests encompassing species beyond those available in the FOR-species20K data. Therefore, the models published here should only be applied to data with known species diversity within the FOR-species20K diversity.
- Differences in quality of the tree segmentation: FOR-species20K is composed of very high-quality tree segmentation, and thus it is unclear to what extent the model can be applied to poorly segmented trees. This is particularly relevant for complex forest structures where individual trees are difficult to accurately segment due to multiple canopy layers and overlapping crowns.
- Species representation on different platforms or sensors: Some species have been recorded by a single platform only, and our dataset uses a relatively small range of sensors (Table 2), and model performance on other platforms or sensors is untested.

5 | CONCLUSIONS

This study, addressing the need for benchmarked individual tree species classifiers, is part of a broader effort to automate forest ecosystems characterization using proximal laser scanning data. In this context, DL tree species classification models are a key component in larger DL pipelines that also includes individual tree segmentation. While FOR-species20K represents an important starting point in bringing together the scientific community to build critical data infrastructure for benchmarking and developing species classification from proximal sensed laser scanning data, enabling the development of the next generation of species classifiers requires a larger community effort to expanding the database. Here, the focus should be on extending the dataset with more tree species, increasing the sample size for poorly represented species and their seedlings, as well as increasing the number of trees from underrepresented scanning approaches (i.e. MLS and ULS), and tree size classes.

AUTHOR CONTRIBUTIONS

Stefano Puliti conceived the ideas and designed methodology; Julian Frey, Kim Calders, Louise Terryn, Nicholas Coops, Bernhard Höfle, Liam Irwin, Samuli Junttila, Martin Krůček, Grzegorz Krok, Kamil Král, Shaun R. Levick, Linda Luck, Azim Missarov, Martin Mokroš, Harry J. F. Owen, Krzysztof Stereńczak, Timo P. Pitkänen, Nicola Puletti, Ninni Saarinen, Chris Hopkinson, Enrico Tomelleri, and Hannah Weiser collected the data; Stefano Puliti, Julian Frey, Zoe Schindler, Adrian Straker, Matthew J. Allen, Lukas Winiwarter, Nataliia Rehush, Hristina Hristova, and Brent Murray analysed the data; Stefano Puliti, Emily Lines, Jana Müllerová led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

AFFILIATIONS

[1]Division of Forest and Forest Resources, National Forest Inventory, Norwegian Institute for Bioeconomy Research (NIBIO), Ås, Norway; [2]Department of Geography, University of Cambridge, Cambridge, UK; [3]Department of Geoinformatics, Jan Evangelista Purkyně University in Ústí nad Labem, Ústí nad Labem, Czech Republic; [4]Chair of Forest Growth and Dendroecology, University of Freiburg, Freiburg, Germany; [5]Faculty of Forest Sciences and Forest Ecology, Burckhardt-Institute, Forest Inventory and Remote Sensing, Georg-August-Universität Göttingen, Göttingen, Germany; [6]Research Unit Photogrammetry, Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria; [7]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland; [8]Department of Forest Resources Management, Faculty of Forestry, University of British Columbia, Vancouver, British Columbia, Canada; [9]Q-ForestLab, Department of Environment, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium; [10]3DGeo Research Group, Institute of Geography, Heidelberg University, Heidelberg, Germany; [11]School of Forest Sciences, University of Eastern Finland, Joensuu, Finland; [12]Department of Forest Ecology, Silva Tarouca Research Institute, Průhonice, Czech Republic; [13]Department of Geomatics, Forest Research Institute, Raszyn, Poland; [14]Land and Water, Commonwealth Scientific and Industrial Research Organization (CSIRO), Winnellie, Northern Territory, Australia; [15]Research Institute for the Environment and Livelihoods, Charles Darwin University, Casuarina, Northern Territory, Australia; [16]Section 1.4 Remote Sensing and Geoinformatics, German Research Centre for GeoSciences (GFZ), Potsdam, Germany; [17]Department of Geography, University College of London,

London, UK; [18]Natural Resources Institute Finland (Luke), Helsinki, Finland; [19]CREA-FL, Council for Agricultural Research and Economics, Research Centre for Forestry and Wood, Arezzo, Italy; [20]Department of Geography, University of Lethbridge, Lethbridge, Alberta, Canada; [21]National Research Council—Institute of BioEconomy, San Michele all'Adige (TN), Italy and [22]Free University of Bolzano, Faculty of Agricultural, Environmental and Food Sciences, Bolzano, Italy

ACKNOWLEDGEMENTS

This work was supported by the COST Action 3DForEcoTech (CA20118). This work is part of the Center for Research-based Innovation SmartForest: Bringing Industry 4.0 to the Norwegian forest sector (NFR SFI project no. 309671, smartforest.no). ERL and HJFO were funded by a UKRI Future Leaders Fellowship awarded to E.R.L. (MR/T019832/1). MJA was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (EP/S022961/1). We acknowledge the technical support and compute time at the Vienna Scientific Cluster VSC-5 for parts of the Ensemble-PointNet++ results. LW was funded in part by the Austrian Science Fund (FWF) [J4672]. The contribution of the DetailView model was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project FR 4404/1-1. NS was supported by the Academy of Finland through UNITE Flagship (357906) and Scan4rest Research Infrastructure (346382). KC was funded by the European Union (ERC-2021-STG Grant agreement No. 101039795). ET was funded by INEST—PNRR (Italian National Plan for Recovery and Resilience), Project id, ECS00000043. REMBIOFOR dataset was funded by National Centre for Research and Development in Poland under the BIOSTRATEG programme (grant agreement number BIOSTRATEG1/267755/4/NCBR/2015), project REMBIOFOR 'Remote sensing-based assessment of woody biomass and carbon storage in forests'. KK, AM and MK were funded by INTER-COST project LUC23023. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the respective funding agencies which can therefore not be held responsible for them.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14503.

DATA AVAILABILITY STATEMENT

The FOR-species20K data are publicly available via https://zenodo.org/records/13255198 (Puliti et al., 2024). The code available for each of the benchmarked methods can be accessed permanently via https://zenodo.org/records/14650218 (Puliti et al., 2025) Updates to the code will be accessible via https://github.com/stefp/FOR-species20K. Researchers who wish to benchmark their methods against the withheld test data from this study can submit their results through the Codabench benchmark available here: https://www.codabench.org/competitions/3667/.

ORCID

Stefano Puliti https://orcid.org/0000-0003-4624-8987
Emily R. Lines https://orcid.org/0000-0002-5357-8741
Jana Müllerová https://orcid.org/0000-0001-7331-3479
Zoe Schindler https://orcid.org/0000-0003-2972-1920
Adrian Straker https://orcid.org/0009-0009-0851-1888
Matthew J. Allen https://orcid.org/0000-0001-6877-6591
Lukas Winiwarter https://orcid.org/0000-0001-8229-1160
Nataliia Rehush https://orcid.org/0000-0003-4966-1945
Brent Murray https://orcid.org/0000-0003-3053-9448
Kim Calders https://orcid.org/0000-0002-4562-2538
Nicholas Coops https://orcid.org/0000-0002-0151-9037
Bernhard Höfle https://orcid.org/0000-0001-5849-1461
Liam Irwin https://orcid.org/0000-0001-8582-6490
Samuli Junttila https://orcid.org/0000-0001-8276-9259
Martin Krůček https://orcid.org/0000-0002-3137-2259
Grzegorz Krok https://orcid.org/0000-0003-3198-0294
Kamil Král https://orcid.org/0000-0002-3848-2119
Shaun R. Levick https://orcid.org/0000-0003-4437-9174
Linda Luck https://orcid.org/0000-0002-4260-3366
Azim Missarov https://orcid.org/0000-0001-9881-6531
Martin Mokroš https://orcid.org/0000-0002-8559-5757
Harry J. F. Owen https://orcid.org/0000-0002-4294-1728
Krzysztof Stereńczak https://orcid.org/0000-0002-9556-0144
Timo P. Pitkänen https://orcid.org/0000-0002-2275-5338
Nicola Puletti https://orcid.org/0000-0002-2142-959X
Ninni Saarinen https://orcid.org/0000-0003-2730-8892
Chris Hopkinson https://orcid.org/0000-0002-3998-4778
Louise Terryn https://orcid.org/0000-0001-8405-2788
Enrico Tomelleri https://orcid.org/0000-0001-6546-6459
Hannah Weiser https://orcid.org/0000-0003-3256-7311
Rasmus Astrup https://orcid.org/0000-0003-2988-9520

REFERENCES

Åkerblom, M., Raumonen, P., Mäkipää, R., & Kaasalainen, M. (2017). Automatic tree species recognition with quantitative structure models. *Remote Sensing of Environment*, *191*, 1–12.

Allen, M. J., Grieve, S. W., Owen, H. J., & Lines, E. R. (2023). Tree species classification from complex laser scanning data in Mediterranean forests using deep learning. *Methods in Ecology and Evolution*, *14*, 1657–1667.

Calders, K., Adams, J., Armston, J., Bartholomeus, H., Bauwens, S., Bentley, L. P., Chave, J., Danson, F. M., Demol, M., Disney, M., Gaulton, R., Krishna Moorthy, S. M., Levick, S. R., Saarinen, N., Schaaf, C., Stovall, A., Terryn, L., Wilkes, P., & Verbeeck, H. (2020). Terrestrial laser scanning in forest ecology: Expanding the horizon. *Remote Sensing of Environment*, *251*, 112102.

Calders, K., Verbeeck, H., Burt, A., Origo, N., Nightingale, J., Malhi, Y., Wilkes, P., Raumonen, P., Bunce, R. G. H., & Disney, M. (2022). Laser scanning reveals potential underestimation of biomass carbon in temperate forest. *Ecological Solutions and Evidence*, *3*, e12197.

Chen, J., Chen, Y., & Liu, Z. (2021). Classification of typical tree species in laser point cloud based on deep learning. *Remote Sensing*, *13*, 4750.

Choe, J., Park, C., Rameau, F., Park, J., & Kweon, I. S. (2022). Pointmixer: Mlp-mixer for point cloud understanding. In *European conference on computer vision* (pp. 620–640). Springer.

Choy, C., Gwak, J., & Savarese, S. (2019). 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3075–3084). https://openaccess.thecvf.com/content_CVPR_2019/papers/Choy_4D_Spatio-Temporal_ConvNets_Minkowski_Convolutional_Neural_Networks_CVPR_2019_paper.pdf

Disney, M. (2019). Terrestrial LiDAR: A three-dimensional revolution in how we look at trees. *New Phytologist*, *222*, 1736–1741.

Fan, Z., Wei, J., Zhang, R., & Zhang, W. (2023). Tree species classification based on PointNet++ and airborne laser survey point cloud data enhancement. *Forests*, *14*, 1246.

Goyal, A., Law, H., Liu, B., Newell, A., & Deng, J. (2021). Revisiting point cloud shape classification with a simple and effective baseline. In *International conference on machine learning* (pp. 3809–3820). PMLR. https://proceedings.mlr.press/v139/goyal21a

Guan, H., Yu, Y., Ji, Z., Li, J., & Zhang, Q. (2015). Deep learning-based tree classification using mobile LiDAR data. *Remote Sensing Letters*, *6*, 864–873.

Hakula, A., Ruoppa, L., Lehtomäki, M., Yu, X., Kukko, A., Kaartinen, H., Taher, J., Matikainen, L., Hyyppä, E., & Luoma, V. (2023). Individual tree segmentation and species classification using high-density close-range multispectral laser scanning data. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *9*, 100039.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://ieeexplore.ieee.org/document/7780459

Hovi, A., Korhonen, L., Vauhkonen, J., & Korpela, I. (2016). LiDAR waveform features for tree species classification and their sensitivity to tree-and acquisition related parameters. *Remote Sensing of Environment*, *173*, 224–237.

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., & Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint*, arXiv:1404.1869. https://doi.org/10.48550/arXiv.1404.1869

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Yifu, Z., Wong, C., & Montes, D. (2022). ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*. https://doi.org/10.5281/zenodo.3908559

Krisanski, S., Taskhiri, M. S., Gonzalez Aracil, S., Herries, D., & Turner, P. (2021). Sensor agnostic semantic segmentation of structurally diverse and complex forest point clouds using deep learning. *Remote Sensing*, *13*, 1413.

Krůček, M., Trochta, J., Cibulka, M., & Král, K. (2019). Beyond the cones: How crown shape plasticity alters aboveground competition for space and light—Evidence from terrestrial laser scanning. *Agricultural and Forest Meteorology*, *264*, 188–199.

Li, R., Li, X., Heng, P.-A., & Fu, C.-W. (2020). Pointaugment: An auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6378–6387). https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_PointAugment_An_Auto-Augmentation_Framework_for_Point_Cloud_Classification_CVPR_2020_paper.pdf

Liang, X., Hyyppä, J., Kaartinen, H., Lehtomäki, M., Pyörälä, J., Pfeifer, N., Holopainen, M., Brolly, G., Francesco, P., Hackenberg, J., Huang, H., Jo, H.-W., Katoh, M., Liu, L., Mokroš, M., Morel, J., Olofsson, K., Poveda-Lopez, J., Trochta, J., … Wang, Y. (2018). International benchmarking of terrestrial laser scanning approaches for forest inventories. *ISPRS Journal of Photogrammetry and Remote Sensing*, *144*, 137–179.

Lin, H., Nazari, M., & Zheng, D. (2023). *Pctrees—3d point cloud tree species classification using airborne lidar images*. Association for the Advancement of Artificial Intelligence. www.aaai.org

Lines, E. R., Allen, M., Cabo, C., Calders, K., Debus, A., Grieve, S. W., Miltiadou, M., Noach, A., Owen, H. J., & Puliti, S. (2022). AI applications in forest monitoring need remote sensing benchmark datasets. In *2022 IEEE international conference on big data (big data)* (pp. 4528–4533). IEEE. https://arxiv.org/abs/2212.09937

Lines, E. R., Fischer, F. J., Owen, H. J. F., & Jucker, T. (2022). The shape of trees: Reimagining forest ecology in three dimensions with remote sensing. *Journal of Ecology*, *110*, 1730–1745.

Liu, B., Chen, S., Huang, H., & Tian, X. (2022). Tree species classification of backpack laser scanning data using the PointNet++ point cloud deep learning method. *Remote Sensing*, *14*, 3809.

Liu, B., Huang, H., Tian, X., & Ren, M. (2022). Individual tree species classification using the pointwise MLP-based point cloud deep learning method. *Environmental Sciences Proceedings*, *22*, 19.

Liu, M., Han, Z., Chen, Y., Liu, Z., & Han, Y. (2021). Tree species classification of LiDAR data based on 3D deep learning. *Measurement*, *177*, 109301.

Lv, Y., Zhang, Y., Dong, S., Yang, L., Zhang, Z., Li, Z., & Hu, S. (2021). A convex hull-based feature descriptor for learning tree species classification from ALS point clouds. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Malhi, Y., Jackson, T., Patrick Bentley, L., Lau, A., Shenkin, A., Herold, M., Calders, K., Bartholomeus, H., & Disney, M. I. (2018). New perspectives on the ecology of tree structure and tree communities through terrestrial laser scanning. *Interface Focus*, *8*, 20170052.

Marinelli, D., Paris, C., & Bruzzone, L. (2022). An approach based on deep learning for tree species classification in LiDAR data acquired in mixed forest. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Mizoguchi, T., Ishii, A., Nakamura, H., Inoue, T., & Takamatsu, H. (2017). Lidar-based individual tree species classification using convolutional neural network. In *Videometrics, range imaging, and applications XIV* (pp. 193–199). SPIE. https://doi.org/10.1117/12.2270123.short

Owen, H., Lines, E., & Flynn, W. (2021). Data from: Terrestrial laser scanning (TLS) data on tree crown morphology and neighbourhood competition from both Cuellar and Alto Tajo, Spain. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.0k6djhb0m

Puliti, S., Frey, J., Schindler, Z., Straker, A., Allen, M. J., Winiwarter, L., Rehush, N., Hristova, H., & Murray, B. (2025). For-species code. *Zenodo*. https://zenodo.org/records/14650218

Puliti, S., Lines, E., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M. J., Lukas, W., Rehush, N., Hristova, H., Murray, B., Calders, K., Terryn, L., Coops, N., Höfle, B., Junttila, S., Krucek, M., Krok, G., Král, K., … Astrup, R. (2024). FOR-species20K dataset. *Zenodo*. https://zenodo.org/records/13255198

Puliti, S., Pearse, G., Surový, P., Wallace, L., Hollaus, M., Wielgosz, M., & Astrup, R. (2023). For-instance: A UAV laser scanning benchmark dataset for semantic and instance segmentation of individual trees. *arXiv preprint*, arXiv:2309.01279. https://doi.org/10.48550/arXiv.2309.01279

Puttonen, E., Jaakkola, A., Litkey, P., & Hyyppä, J. (2011). Tree classification with fused mobile laser scanning and hyperspectral data. *Sensors*, *11*, 5158–5182.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, *30*, 1–10.

Saarinen, N., Calders, K., Kankare, V., Yrttimaa, T., Junttila, S., Luoma, V., Huuskonen, S., Hynynen, J., & Verbeeck, H. (2021). Data for investigating structural complexity of individual Scots pine trees. *Zenodo*. https://zenodo.org/records/4419878

Saarinen, N., Kankare, V., Huuskonen, S., Hynynen, J., Bianchi, S., Yrttimaa, T., Luoma, V., Junttila, S., Holopainen, M., Hyyppä, J., & Vastaranta, M. (2021). Point clouds from terrestrial laser scanning from crowns of individual Scots pine trees. *Zenodo*. https://doi.org/10.5281/zenodo.5783404

Saarinen, N., Kankare, V., Pyörälä, J., Yrttimaa, T., Liang, X., Wulder, M. A., Holopainen, M., Hyyppä, J., & Vastaranta, M. (2020). Point cloud data from terrestrial laser scanning for stem volume modelling of scots pine trees. *Zenodo*. https://doi.org/10.5281/zenodo.3712900

Schneider, F. D., Kükenbrink, D., Schaepman, M. E., Schimel, D. S., & Morsdorf, F. (2019). Quantifying 3D structure and occlusion in dense tropical and temperate forests using close-range LiDAR. *Agricultural and Forest Meteorology*, *268*, 249–257.

Seidel, D., Annighöfer, P., Thielman, A., Seifert, Q. E., Thauer, J.-H., Glatthorn, J., Ehbrecht, M., Kneib, T., & Ammer, C. (2021). Predicting tree species from 3D laser scanning point clouds using deep learning. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.635440

Terryn, L., Calders, K., Disney, M., Origo, N., Malhi, Y., Newnham, G., Raumonen, P., & Verbeeck, H. (2020). Tree species classification using structural features derived from terrestrial laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *168*, 170–181.

Tockner, A., Gollob, C., Ritter, T., & Nothdurft, A. (2022). Lautx-individual tree point clouds from austrian forest inventory plots. *Zenodo*. https://zenodo.org/records/6560112

Van Den Berge, S., Vangansbeke, P., Calders, K., Vanneste, T., Baeten, L., Verbeeck, H., Krishna Moorthy, S. P., & Verheyen, K. (2021). Biomass expansion factors for hedgerow-grown trees derived from terrestrial LiDAR. *Bioenergy Research*, *14*, 561–574.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, *38*, 1–12.

Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Fassnacht, F. E., & Höfle, B. (2022). Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests. *Earth System Science Data*, *14*, 2989–3012.

Wielgosz, M., Puliti, S., Xiang, B., Schindler, K., & Astrup, R. (2024). SegmentAnyTree: A sensor and platform agnostic deep learning model for tree segmentation using laser scanning data. *Remote Sensing of Environment*, *313*, 114367.

Xi, Z., Hopkinson, C., Rood, S. B., & Peddle, D. R. (2020). See the forest and the trees: Effective machine and deep learning algorithms for wood filtering and tree species classification from terrestrial laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *168*, 1–16.

Xiang, B., Wielgosz, M., Kontogianni, T., Peters, T., Puliti, S., Astrup, R., & Schindler, K. (2024). Automated forest inventory: Analysis of high-density airborne LiDAR point clouds with 3D deep learning. *Remote Sensing of Environment*, *305*, 114078.

Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.-W., Yao, Q., Zhao, H., & Guyon, I. (2022). Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, *3*, 100543.

Zou, X., Cheng, M., Wang, C., Xia, Y., & Li, J. (2017). Tree classification in complex forest point clouds based on deep learning. *IEEE Geoscience and Remote Sensing Letters*, *14*, 2360–2364.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1: Detailed description of the benchmarked tree species classification models.

How to cite this article: Puliti, S., Lines, E. R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M. J., Winiwarter, L., Rehush, N., Hristova, H., Murray, B., Calders, K., Coops, N., Höfle, B., Irwin, L., Junttila, S., Krůček, M., Krok, G., Král, K., … Astrup, R. (2025). Benchmarking tree species classification from proximally sensed laser scanning data: Introducing the FOR-species20K dataset. *Methods in Ecology and Evolution*, *00*, 1–18. https://doi.org/10.1111/2041-210X.14503