Multispectral airborne laser scanning for tree species classification: A benchmark of machine learning and deep learning algorithms

Josef Taher [a],[1],[*], Eric Hyyppä [a],[1], Matti Hyyppä [a],[1], Klaara Salolahti [a],[1], Xiaowei Yu [a],[1], Leena Matikainen [a],[1], Antero Kukko [a], Matti Lehtomäki [a], Harri Kaartinen [a], Sopitta Thurachen [a], Paula Litkey [a], Ville Luoma [b], Markus Holopainen [b], Gefei Kong [c], Hongchao Fan [c], Petri Rönnholm [d], Matti Vaaja [d], Antti Polvivaara [e], Samuli Junttila [e], Mikko Vastaranta [e], Stefano Puliti [f], Rasmus Astrup [f], Joel Kostensalo [g], Mari Myllymäki [h], Maksymilian Kulicki [i,j], Krzysztof Stereńczak [i,k], Raul de Paula Pires [l], Ruben Valbuena [l], Juan Pedro Carbonell-Rivera [m], Jesús Torralba [m], Yi-Chen Chen [n], Lukas Winiwarter [n,o], Markus Hollaus [n], Gottfried Mandlburger [n], Narges Takhtkeshha [p,n], Fabio Remondino [p], Maciej Lisiewicz [k], Bartłomiej Kraszewski [k], Xinlian Liang [q], Jianchang Chen [q], Eero Ahokas [a], Kirsi Karila [a], Eugeniu Vezeteu [a], Petri Manninen [a], Roope Näsi [a], Heikki Hyyti [a], Siiri Pyykkönen [a], Peilun Hu [a], Juha Hyyppä [a]

a Department of Remote Sensing and Photogrammetry, Finnish Geospatial Research Institute FGI, The National Land Survey of Finland, Vuorimiehentie 5, Espoo, FI-02150, Finland
b Department of Forest Sciences, University of Helsinki, Latokartanonkaari 7, Helsinki, FI-00790, Finland
c Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Høgskoleringen 7A, Trondheim, NO-7491, Norway
d Department of Built Environment, Aalto University, PO Box 14100, Espoo, FI-00076 AALTO, Finland
e University of Eastern Finland, Faculty of Science, Forestry, and Technology, School of Forest Sciences, P.O. Box 111, Joensuu, FI-80101, Finland
f Division of Forest and Forest Resources, National Forest Inventory, Norwegian Institute for Bioeconomy Research (NIBIO), Høgskoleveien 8, Ås, NO-1433, Norway
g Natural Resources, Natural Resources Institute Finland, Yliopistokatu 6B, Joensuu, FI-80100, Finland
h Bioeconomy and Environment, Natural Resources Institute Finland, Latokartanonkaari 9, Helsinki, FI-00790, Finland
i IDEAS NCBR, ul. Chmielna 69, Warsaw, 00–801, Poland
j Institute of Fundamental Technological Research, Polish Academy of Science, ul. Pawińskiego 5B, Warsaw, 02–106, Poland
k Department of Geomatics, Forest Research Institute IBL, ul. Braci Leśnej 3, Sekocin Stary, 05–090, Poland
l Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden
m Geo-Environmental Cartography and Remote Sensing Group (CGAT), Universitat Politècnica de València, Camíde Vera s/n, Valencia, 46022, Spain
n Department of Geodesy and Geoinformation, Technische Universität Wien, Wiedner Hauptstraße 8, Vienna, 1040, Austria
o Faculty of Engineering Sciences, Universität Innsbruck, Technikerstraße 13, Innsbruck, 6020, Austria
p 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Via Sommarive, 18, Trento, 38123, Italy
q The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430070, China

ARTICLE INFO

ABSTRACT

Climate-smart and biodiversity-preserving forestry demands precise information on forest resources, extending to the individual tree level. Multispectral airborne laser scanning (ALS) has shown promise in automated point cloud processing, but challenges remain in leveraging deep learning techniques and identifying rare tree species in class-imbalanced datasets. This study addresses these gaps by conducting a comprehensive benchmark of deep learning and traditional shallow machine learning methods for tree species classification. For the study, we collected high-density multispectral ALS data (>1000 pts/m²) at three wavelengths using the FGI-developed HeliALS system, complemented by existing Optech Titan data (35 pts/m²), to evaluate the species classification accuracy of various algorithms in a peri-urban study area located in southern Finland. We

* Corresponding author.
E-mail addresses: josef.taher@nls.fi (J. Taher), eric.hyyppa@nls.fi (E. Hyyppä), matti.hyyppa@nls.fi (M. Hyyppä), klaara.salolahti@nls.fi (K. Salolahti), xiaowei.yu@nls.fi (X. Yu), leena.matikainen@nls.fi (L. Matikainen), juha.hyyppa@nls.fi (J. Hyyppä).
1 These authors contributed equally.

Lidar

established a field reference dataset of 6326 segments across nine species using a newly developed browser-based crowdsourcing tool, which facilitated efficient data annotation. The ALS data, including a training dataset of 1065 segments, was shared with the scientific community to foster collaborative research and diverse algorithmic contributions. Based on 5261 test segments, our findings demonstrate that point-based deep learning methods, particularly a point transformer model, outperformed traditional machine learning and image-based deep learning approaches on high-density multispectral point clouds. For the high-density ALS dataset, a point transformer model provided the best performance reaching an overall (macro-average) accuracy of 87.9% (74.5%) with a training set of 1065 segments and 92.0% (85.1%) with a larger training set of 5000 segments. With 1065 training segments, the best image-based deep learning method, DetailView, reached an overall (macro-average) accuracy of 84.3% (63.9%), whereas a shallow random forest (RF) classifier achieved an overall (macro-average) accuracy of 83.2% (61.3%). For the sparser ALS dataset, an RF model topped the list with an overall (macro-average) accuracy of 79.9% (57.6%), closely followed by the point transformer at 79.6% (56.0%). Importantly, the overall classification accuracy of the point transformer model on the HeliALS data increased from 73.0% with no spectral information to 84.7% with single-channel reflectance, and to 87.9% with spectral information of all the three channels. Furthermore, we studied the scaling of the classification accuracy as a function of point density and training set size using 5-fold cross-validation of our dataset. Based on our findings, multispectral information is especially beneficial for sparse point clouds with 1–50 pts/m^2. Furthermore, we observed that the classification error follows a power law $\varepsilon(m) \approx m^{-\alpha}$ as a function of the training set size m, and the classification error of the point transformer reduced significantly faster with increasing training set size compared to RF.

1. Introduction

Knowledge of the species-specific size distribution of forest stands is crucial for effective forest management planning and optimization of the wood supply chain. This information allows for decisions on tailored silvicultural treatments, such as harvesting schedules, thinning operations, and regeneration strategies. Reliable information on tree species also contributes to sustainable forest management, ensuring a continuous supply of high-quality timber products while maintaining ecological integrity. Namely, the diversity of tree species plays an important role in shaping forest ecosystems, and it affects factors such as productivity, resilience to disturbances, recovery rates, internal competition among trees, ecosystem health, economic potential and biodiversity. For example in the boreal forest zone, aspen (*Populus tremula* L.) is widely recognized as a keystone species, having an important role in supporting biodiversity (Kouki et al., 2004; Kivinen et al., 2020), since aspen stands often harbor a rich diversity of fungi, lichens, insects, birds, and mammals (Kuusinen, 1994; Angelstam and Mikusiński, 1994; Tikkanen et al., 2006; Remm et al., 2017). Additionally, accurate tree species identification in municipalities is required for achieving the targeted restoration and effective enhancement of biodiversity and canopy cover within urban ecosystems (European Parliament and Council, 2025; Federal Office for Agriculture and Food, 2025; Ministry of Agriculture and Forestry of Finland, 2025).

Although remote sensing techniques, such as satellite imagery and airborne laser scanning (ALS), have shown great promise in forest monitoring (Hyyppä et al., 2000; Næsset et al., 2004; Holmgren and Persson, 2004; Ørka et al., 2009), accurate species classification at the individual tree level remains a challenging task. For example, previous ALS studies in the boreal forest zone have demonstrated a high classification accuracy for the dominant pine and spruce trees (Holmgren and Persson, 2004; Ørka et al., 2012; Yu et al., 2017; Hakula et al., 2023), but the classification accuracy of deciduous trees has remained markedly lower especially for under-represented species (Ørka et al., 2012; Axelsson et al., 2018; Yu et al., 2017). In general, the classification accuracy degrades if the number of species is high (Heinzel and Koch, 2011; Prieur et al., 2021) or if there are closely related species, particularly those with similar spectral and structural characteristics. Therefore, further efforts are needed to improve species classification using remote sensing techniques since accurate species classification will be required to enable multifunctional forestry serving multiple ecosystem services optimally at the same time.

Today, airborne laser scanning is the most prevalent technique for forest inventory in the boreal forest zone, with a strong emphasis on economically efficient forest resource utilization. The state of the art

in species classification using ALS is summarized in several review papers (Chen et al., 2024; Michałowska and Rapiński, 2021; Wang et al., 2018; Koenig and Höfle, 2016; Fassnacht et al., 2016) and a recent benchmarking study using the FOR-species20K open dataset by Puliti et al. (2025). In general, the accuracy of species classification is affected by several factors, including the point cloud quality (e.g. point density, noise, geometric accuracy affected by beam footprint), additional features available for classification (intensity at one or multiple wavelengths, echo types, waveform information), the chosen classification method (machine learning or deep learning), and the difficulty of the studied forest area (number and similarity of species, tree density, size distribution of trees). Until recently, ALS-based species classification has been implemented using hand-crafted features and traditional shallow supervised machine learning (ML) classifiers, such as random forests (RF) (Ørka et al., 2012; Dalponte et al., 2012; Yu et al., 2017), support vector machines (SVM) (Ørka et al., 2012; Dalponte et al., 2012, 2014) or discriminant analysis (Holmgren and Persson, 2004; Ørka et al., 2009; Heinzel and Koch, 2011). Typical features for species classification have included point cloud metrics representing the geometry of a tree (e.g. crown shape, tree height, density and height distributions of points, trunk hits, echo ranks, point counts), backscatter and intensity statistics (e.g. intensity/amplitude, reflectance and waveform distributions), and temporal changes in the canopy (e.g. the difference of intensity or average canopy height between leaf-on and leaf-off seasons) (Holmgren and Persson, 2004; Ørka et al., 2009; Fassnacht et al., 2016; Koenig and Höfle, 2016; Lin and Hyyppä, 2016; Yu et al., 2017). Throughout the manuscript, we refer to traditional shallow machine learning classifiers as *machine learning* approaches following the convention in several recent papers in tree species classification (Xi et al., 2020; Mäyrä et al., 2021; Hell et al., 2022). In contrast, we use the term *deep learning* method to refer to deep neural networks that learn their features.

During the past few years, deep learning (DL) approaches have been shown to outperform traditional shallow machine learning methods in species classification (Xi et al., 2020; Liu et al., 2021). As a result, several deep learning methods have been recently proposed for species classification both using ALS data (Hell et al., 2022; Marinelli et al., 2022; Fan et al., 2023; Lin et al., 2024) and terrestrial laser scanning (TLS) data (Seidel et al., 2021; Liu et al., 2022a). Deep learning approaches can be divided into image-based 2D methods (Seidel et al., 2021; Marinelli et al., 2022) and point-based 3D methods (Xi et al., 2020; Liu et al., 2021; Fan et al., 2023; Lin et al., 2024). In image-based DL methods, the point cloud is projected into multiple views from different perspectives, which are given as an input for a 2D convolutional neural network (CNN), such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017), or YOLO (Jocher et al., 2023). On

the other hand, the input of point-based methods is the point cloud itself with potential additional features. Typical neural net architectures of point-based methods include, e.g., PointNet (Qi et al., 2017a), PointNet++ (Qi et al., 2017b), dynamic graph CNN (DGCNN) (Wang et al., 2019b), and point transformer (Zhao et al., 2021). Deep learning methods require large amounts of training data to reach their potential, which has resulted in a need for large open datasets enabling the development and benchmarking of new deep-learning-based classification techniques. To this end, Puliti et al. (2025) recently presented FOR-species20K, a 33 species open dataset consisting of 20 000 trees collected using terrestrial, mobile and unmanned aerial vehicle based laser scanning with the goal to facilitate the development of sensor-agnostic classification approaches using only the point cloud geometry. Furthermore, Puliti et al. (2025) presented a comparison of seven sensor-agnostic deep learning methods concluding that the studied image-based deep learning models marginally outperformed the point-based methods.

Backscatter intensity (or reflectance), even though hard to standardize across different sensors, has been successfully applied for species classification from the beginning of ALS research. Holmgren and Persson (2004) showed that a combination of geometric and intensity features led to an overall accuracy of 95% for the discrimination of spruce and pine. Since then intensity and geometric features have been used in many studies on tree species classification. Ørka et al. (2009) demonstrated a classification accuracy of 88% between Norway spruce and birch for about 350 trees, and Korpela et al. (2010) achieved an accuracy of 88%–90% for Scots pine, Norway spruce, and birch across 13000 trees. Importantly, multiple studies have demonstrated that the classification accuracy obtained with geometric features can be significantly improved with the addition of intensity features (Axelsson et al., 2018; Yu et al., 2017; Lin and Hyyppä, 2016; Suratno et al., 2009; Shi et al., 2018a; Hakula et al., 2023; Ørka et al., 2009). For example, Ørka et al. (2009), Yu et al. (2017), and Hakula et al. (2023) reported 11, 9.9 and 13.5 percentage point improvements, respectively, to the overall classification accuracy when introducing single-channel intensity features to the classification. Motivated by the success of single-channel intensity features, the integration of ALS data with passive multispectral or even hyperspectral data has also been studied to further increase the classification accuracy (Holmgren et al., 2008; Ørka et al., 2012; Dalponte et al., 2014; Deng et al., 2016; Kamińska et al., 2018, 2021; Zhong et al., 2022; Quan et al., 2023; Lisiewicz et al., 2025).

Multispectral ALS using multiple laser wavelengths provides a further alternative to enhance the accuracy of species classification with potential to meet the increasing requirements for biodiversity reporting, climate change adaptation, and sustainable resource management, for example. Originally, multispectral laser scanning was developed to increase the capacity of ALS point clouds for automated object recognition (Kaasalainen et al., 2007; Hyyppä et al., 2013). Multispectral data also provides additional information for tree species classification since different tree species have a different reflectance response as a function of wavelength (Hovi et al., 2017). During the past decade, multispectral ALS has been shown to improve species classification accuracy in multiple studies (Yu et al., 2017; Axelsson et al., 2018; Amiri et al., 2019; Prieur et al., 2021; Rana et al., 2022; Hakula et al., 2023; Wang et al., 2024), with some studies observing a reduction of the classification error by up to $\sim$ 30% compared to the use of single-channel features (Axelsson et al., 2018; Amiri et al., 2019; Hakula et al., 2023). Despite the promise of deep learning for semantic segmentation on multispectral data (Reichler et al., 2024; Oinonen et al., 2024; Ruoppa et al., 2025), deep learning approaches for species classification have been studied to date in only a single paper focusing on multispectral ALS data with a point density of 30 pts/m^2 (Wang et al., 2024). Furthermore, the paper by Hakula et al. (2023) is the only study utilizing high-density (> 1000 pts/m^2) multispectral ALS data for species classification. In Hakula et al. (2023), the use of multispectral

ALS data reduced the overall classification error by 65% compared to the use of single-channel geometric features and by 31% compared to single-channel geometric and intensity features for the main tree species, including spruce, pine and birch. The use of deep learning methods for species classification on high-density multispectral ALS data has not been studied previously.

Consequently, there is a need for more in-depth research on multispectral airborne laser scanning for tree species classification, especially in the presence of a wide variety of species and with high-density point clouds. Thus, the primary goal of our study is to carry out an international benchmark of tree species classification methods on multispectral ALS data. To facilitate this advancement, we undertook the following actions:

- We collected high-density ALS data (1300 pts/m^2) using a novel multispectral scanner system in a study area located in southern Finland, where we have previously collected multispectral ALS data with a point density of 35 pts/m^2 using the Optech Titan system (Fernandez-Diaz et al., 2016). In the following, the two datasets are labeled as dense data and sparse data, respectively. We selected a peri-urban study area since it offers a sufficient variety of tree species in a single study area for our benchmarking study.
- We established a high-quality field reference dataset comprising 6326 segments and nine species, including aspen, that is vital for biodiversity in the boreal forests. The field reference dataset was carefully acquired using a crowdsourcing tool facilitating the collection of large datasets required by deep learning models.
- We shared a training dataset of the dense and sparse multispectral ALS data with the international scientific community to organize a benchmarking study on tree species classification using multispectral ALS data. Thirteen teams participated in the benchmarking study submitting in total nine machine learning methods, 13 point-based deep learning methods (including their variants), and four image-based deep learning methods.
- We further investigated the decisive factors affecting the classification accuracy, including the classification method, availability of spectral information (no intensity vs single-channel intensity vs multi-channel intensity), size of the training dataset, point density, segmentation quality, crown class, and similarity between the species.

2. Materials and methods

2.1. Study area

The study area (center point approximately 60.1462°N 24.6587°E) is located in Espoonlahti in the city of Espoo, directly 20 km west of the center of Helsinki, on the southern coast of Finland. The study area shown in Fig. 1 consists of two sub-regions, A and B, both of which were used for the current study. The study area can be characterized as a peri-urban site since it is partially covered by forests and partially by suburban neighborhoods. The area was selected due to the large diversity of tree species growing in the area compared to typical managed Finnish boreal forests with only a handful of species. On the study area, a total of 20–30 different tree species can be found, including natural and planted trees both in suburban environment and in forests, as the area comprises different residential areas, public buildings, a sports park and some recreational and non-recreational, unmanaged forest areas. The forests are natural forests and typically either pine-dominated rocky forests or mixed forests with coniferous and deciduous species. The mixed forests are relatively dense, while the rocky forests are sparser. Tree species classified in the study included pine (*Pinus sylvestris*), spruce (*Picea sp.*), birch (*Betula sp.*), maple (*Acer platanoides*), aspen (*Populus tremula*), rowan (*Sorbus sp.*), oak (*Quercus robur*), linden (*Tilia sp.*) and alder (*Alnus sp.*), see more information in Section 2.4. Pine, spruce, birch, aspen, rowan and alder are typically occurring natural tree species in the area, while most maples, oaks and lindens are planted trees in the environment.
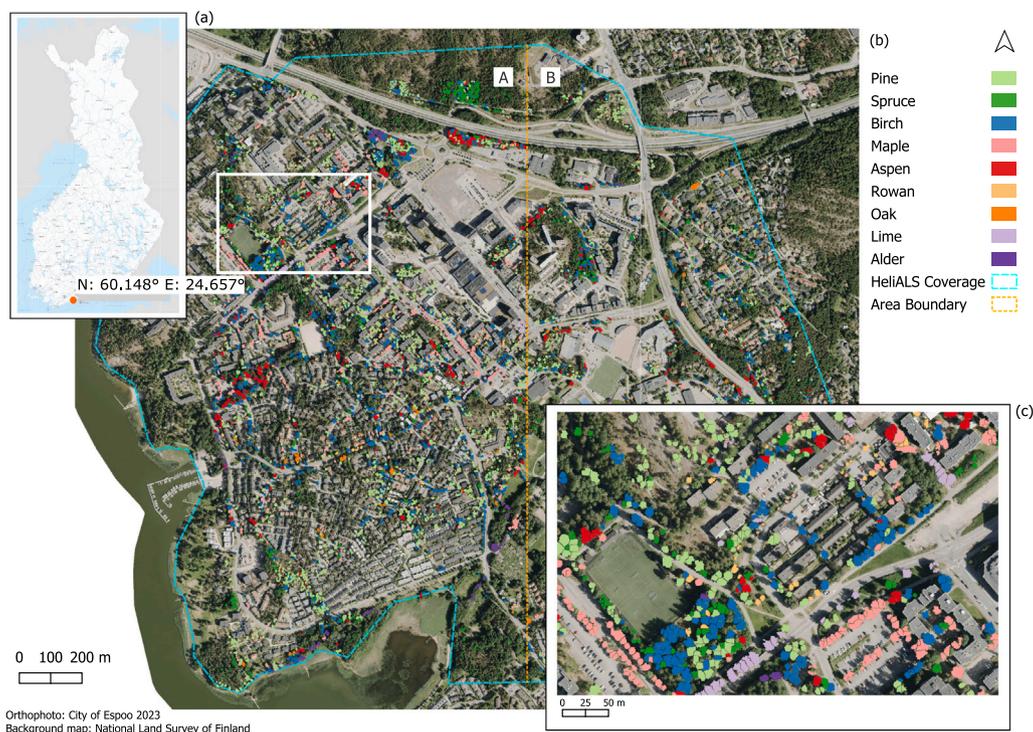
Fig. 1. (a) The Espoonlahti study area is located in southern Finland. (b) Orthophoto of the Espoonlahti study area together with collected tree segments colored by the species in the reference dataset. The coverage of the HeliALS data is shown with the light blue line. Areas A and B are separated by the orange line. (c) A close-up of the orthophoto shown in (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Measurement systems and acquisition of laser scanning data

Measurement systems used for ALS data acquisition included HeliALS and Optech Titan. The HeliALS system (see Fig. 2) has been developed at the Finnish Geospatial Research Institute (FGI), while Optech Titan was the first operational multispectral ALS system and launched by Teledyne Optech (Ontario, Canada) in 2014.

The HeliALS dataset was acquired on July 20 and 28, 2023, from an altitude of 100 m with leaf-on conditions. A helicopter, flying at 14 m/s, carried a system of three Riegl drone laser scanners operating in infrared, near-infrared and green wavelengths, namely a VUX-1HA (λ = 1550 nm), a miniVUX-1DL (λ = 905 nm), and a VQ-840-G (λ = 532 nm). Positioning was achieved using a NovAtel (LITEF) ISA-100C inertial measurement unit (IMU), a NovAtel PwrPak7 Global Navigation Satellite System (GNSS) receiver, and a NovAtel (Vexxis) GNSS-850 antenna. The initial performance of the HeliALS system was presented in Hakula et al. (2023), which was the first study focusing on tree species classification from high-density multispectral ALS data. After the publication, the green channel data have been optimized by increasing the receiver aperture of the VQ-840-G scanner from 3 mrad to 6 mrad. The data were processed to point clouds at the FGI. Further specifications of the HeliALS system and data acquisition are provided in Table 1.

The Optech Titan system operates in three channels with infrared (λ = 1550 nm), near-infrared (λ = 1064 nm) and green wavelengths (λ = 532 nm), respectively. The data were acquired on June 14, 2016, when trees were in full leaf (Karila et al., 2019). The acquisition was carried out in cooperation with TerraTec Oy (Helsinki, Finland). A fixed-wing aircraft and a flying altitude of 700 m were used. Data from the three channels were obtained as three separate point clouds. After basic processing by TerraTec Oy, the data were delivered to the FGI in the Finnish ETRS-TM35FIN coordinate system with height values in the Finnish N2000 height system. Further specifications of the Optech Titan system and data acquisition are presented in Table 2.

Fig. 2. Photograph of the scanner arrangement of the HeliALS system, showing the VUX-1HA scanner at λ = 1550 nm, miniVUX-1DL scanner at λ = 905 nm, and VQ-840-G scanner at λ = 532 nm.

2.3. Pre-processing of laser scanning data

The HeliALS triple-wavelength data were processed into point clouds using RiProcess software (version 1.9.0, Riegl GmbH, Austria). GNSS-IMU trajectories for the two flights were computed in Waypoint Inertial Explorer (8.90, NovAtel Inc., Canada) using a virtual base station from the Trimnet service (Geotrim Oy, Finland) and tightly coupled kinematic post-processing. Boresight calibrations between the IMU and each scanner were solved using RiProcess functionalities that use planar surfaces for solving the rotation angle estimates.

The data from each scanner were loaded to TerraScan (Terrasolid Ltd., Finland), and for each point the closest points within 20 cm were sought from the two other channels to form the reflectance triplets that were then stored in the RGB color fields in the LAZ file structure. Thus, points that do not have close-by counterparts on either one

Table 1
Specifications of the HeliALS measurement system.

Characteristics	Value		
Altitude AGL	100 m		
Flight speed	14 m/s		
Channel-specific characteristics	Channel 1	Channel 2	Channel 3
Scanner name	VUX-1HA	miniVUX-1DL	VQ-840-G
Laser wavelength	1550 nm	905 nm	532 nm
Beam divergence	0.5 mrad	0.5×1.6 mrad	1 mrad
Beam diameter at ground	5 cm	5×16 cm	10 cm
Range accuracy	5 mm	15 mm	20 mm
Direction	15 deg	Nadir	Nadir
Field of view	180°	46° (conical)	40° (conical)
Pulse repetition rate	1017 kHz	100 kHz	200 kHz
Resulting point density	581 pts/m^2	175 pts/m^2	519 pts/m^2

Table 2
Specifications of the Optech Titan measurement system.

Characteristics	Value		
Altitude AGL	700 m		
Flight speed	~70 m/s		
Strip width	510 m		
Lateral overlap	30%		
Channel-specific Characteristics	Channel 1	Channel 2	Channel 3
Laser wavelength	1550 nm	1064 nm	532 nm
Beam divergence	0.35 mrad	0.35 mrad	0.7 mrad
Beam diameter at ground	25 cm	25 cm	50 cm
Direction	3.5° forward	nadir	7° forward
Field of view	40°	40°	40°
Pulse repetition rate	300 kHz	300 kHz	300 kHz
Resulting point density[a]	11 pts/m^2	13 pts/m^2	11 pts/m^2

[a] After cutting points of overlapping flight lines.

or two channels may exhibit only NA (coded as value '1') on the respective RGB fields. The reflectance values in the LAZ were stored as decibel values and exported with RIEGL extrabytes (echo deviation, reflectance, and pulse width). The definition of the reflectance, amplitude and echo deviation attributes for the Riegl scanners are described in Pfennigbauer and Ullrich (2010). Eventually, the full point cloud dataset was split into 200 m square tiles and delivered for the benchmark in 1.2 LAZ format with ETRS-TM35FIN coordinates and ellipsoidal elevation values.

Preprocessing steps for the Optech Titan data included some basic processing steps carried out by TerraTec Oy, intensity range (R) correction according to R^2 (Matikainen et al. 2017), cutting of overlap points and removal of an additional cross line (using TerraScan). Multispectral information for this dataset is available in the form of three range-corrected intensity channels. Reflectance information is not available. The point densities measured after preprocessing were 11 points/m^2, 13 points/m^2 and 11 points/m^2 for the Channels 1, 2 and 3, respectively. The preprocessed set of data was provided as 'original' data for the benchmarking.

The original point clouds from the Optech Titan data were also further classified into ground, vegetation, building and noise by using LAStools software (rapidlasso GmbH, Germany). Ground points were used to create a digital terrain model and for normalizing the point clouds by subtracting the ground elevation from the height of the laser points.

Tree segmentation was carried out based on the Optech Titan data. The resulting segments were also applied for tree species classification with the HeliALS data to eliminate the variability of segment boundaries from the comparison. Given the fact that plenty of deciduous trees are growing in the study area, we used a varying window watershed-based method for the tree detection first depicted in Kaartinen et al. (2012). First, we created a canopy height model (CHM) from the first returns of the vegetation points of Channels 1 and 2 with 0.5 m pixel

size. The CHM was then smoothed by a Gaussian filter with a varying window size, followed by a local maximum filtering to identify the tops of trees. The window size varied according to the CHM height of the target as follows: 3 pixels for 0–7 m height, 5 pixels for 7–20 m height, 7 pixels for 20–30 m height, and 9 pixels for greater than 30 m height. Finally, the marker-controlled watershed algorithm was applied to delineate the crowns of trees using detected tree tops as seeds (Kaartinen et al., 2012). Output of the tree segmentation is a raster image with a unique label for each segment, i.e. the pixels with the same label form one segment. These images together with the original and normalized point clouds were delivered to the partners for species classification.

2.4. Benchmark dataset

The contestants of the benchmark competition received a training set containing segmented point clouds, the segment labels and the ground-truth species classes. The training set consisted of 1065 segments. The participating methods were benchmarked with a test set containing 5261 segments. The dataset was split into training and test sets using random sampling for a given number of segments per species. The full dataset was not available at the time of constructing the training set because the collection of the field reference data was still ongoing. Therefore, no detailed stratification approaches were used for splitting the dataset into training and test sets. Fig. 3 visualizes the nine tree species included in the dataset, the number of segments per species in the training and test sets, and also the species-wise tree height distributions for the entire dataset. See also the distribution of mean intensity for the studied tree species at the three wavelengths of the HeliALS and Optech Titan systems in Appendix A. The distributions showcase the advantage of multispectral data in providing spectral information at multiple wavelengths.

2.5. Reference data collection

A browser-based crowdsourcing tool was developed to help collecting the ground-truth species data in the field. The crowdsourcing tool is a web application that visualizes the study area on a 2D background map as shown in Fig. 4. The web application shows the pre-processed segment boundaries on the map and surveyors can annotate segments with the correct species and also add notes if needed. The crowdsourcing application enables surveyors to locate themselves on the study area with GNSS. To aid positioning when GNSS accuracy is low, a true orthophoto map from the City of Espoo (2024) was visualized as the background map. The 5-cm-resolution orthophoto map had been photographed in the summer of 2021. To help identifying the correct trees on the study area, the canopy height model of the study area was also added to the map as an additional map layer.

Reference data was collected in the field using the crowdsourcing application by the employees of the FGI in 2023 and 2024. The process was guided by several key principles. First, participants were introduced to the data collection tools before starting the field work to establish a baseline for consistency and reduce initial misclassification errors. Second, we made improvements to the user interface of the crowdsourcing application during the field campaign based on our learnings in the field to further ease the collection of the ground-truth data. Third, we implemented an expert verification process to ensure data quality and reduce inherent subjectivity: The initial species classifications made by the crowd were subsequently revisited on the field and verified by experts using the crowdsourcing tool. To help determine which segments needed more attention in the verification phase, we computed for each segment the proportion of classification methods that predicted an incorrect species and prepared an associated map. In the case of minority species, segments with multiple incorrect predictions were typically correctly labeled in the field. For the more common species, some segments with multiple incorrect predictions
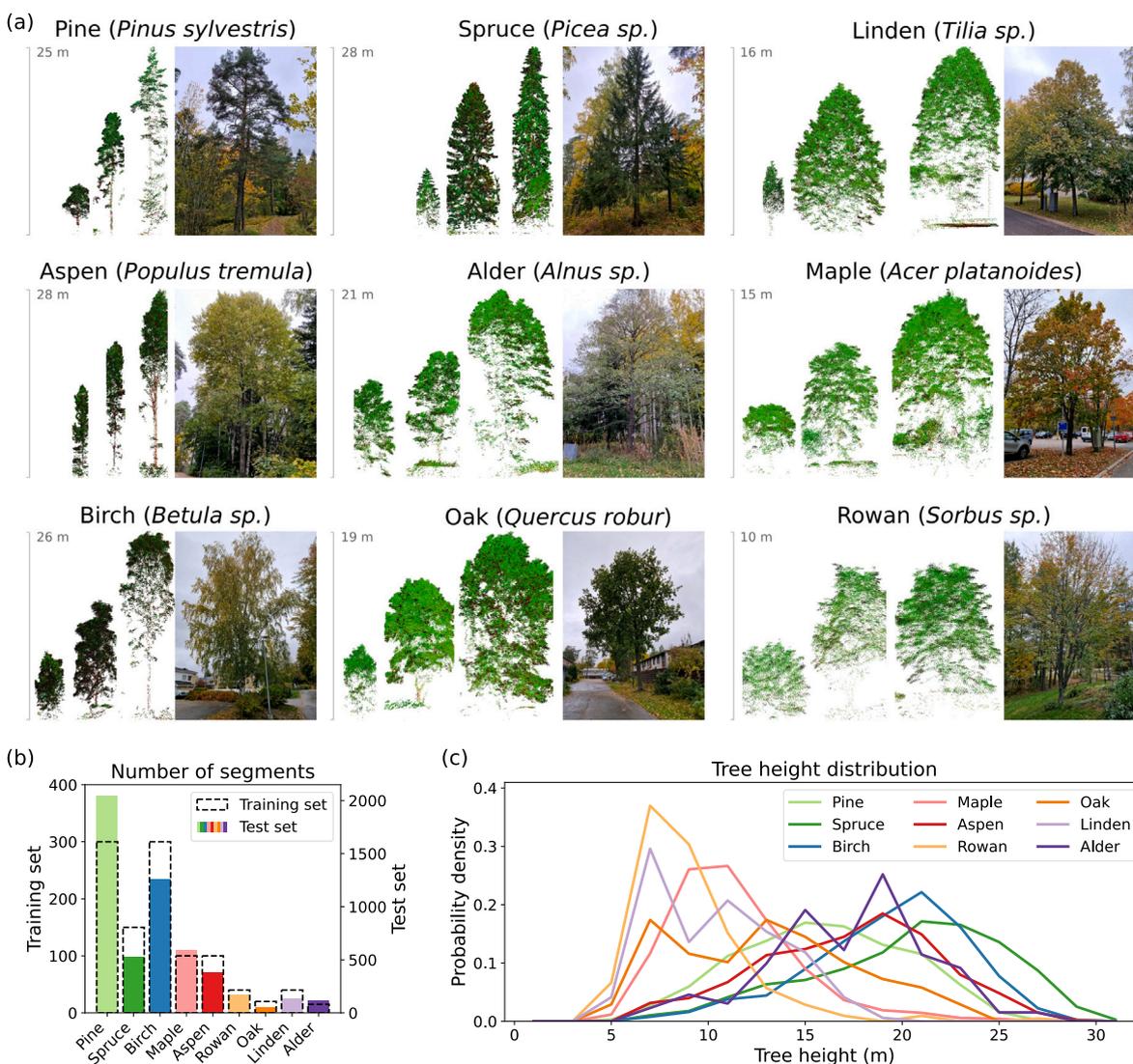
Fig. 3. (a) Example multispectral point clouds and photographs of the nine tree species found in the dataset. The red, green and blue color channels of the point clouds represent the laser pulse return intensities at the wavelengths $\lambda_1 = 532$ nm, $\lambda_2 = 905$ nm and $\lambda_3 = 1550$ nm, respectively. The height scale varies between the visualized point clouds. Photographs of the representative species instances were taken during the autumn of 2024. (b) Number of segments per species in the training and test sets. The dashed lines represent the training set (left y-axis) and the solid bars correspond to the test set (right y-axis). (c) Species-wise tree height distributions for the entire dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

had been incorrectly labeled on the first round of field measurements, and they were subsequently corrected during the on-site verification. This two-step approach — using the crowd for broad coverage and experts for final validation — helped to reduce subjectivity.

In addition to field data collection using the crowdsourcing application, some planted deciduous roadside trees were annotated using the open tree database of City of Espoo (2024), and some clear instances of spruces were labeled using the 5-cm-resolution true orthophoto due to the unique shape of the spruces.

2.6. Categories of tree segments

After labeling the segments and verifying the labels, the segments were categorized by manually inspecting profiles generated from the HeliALS segments. The applied categories were 'Single tree', 'Large tree with undergrowth', 'Many trees of the same species', 'Many trees of many species', and 'Tree section'. In the case of 'Single trees', there was only one tree visible in the profile image of the segment. A segment belongs to the 'Large tree with undergrowth' category if there was one

taller tree and possibly some significantly smaller trees or bushes at most one third of the height of the larger tree in the segment. 'Many trees of the same species' category corresponds to segments which contained multiple trees, each having the same species. If there were multiple trees with different species in the same segment, the segment was categorized into the 'Many trees of many species' category and the ground-truth species was recorded according to the species of the tallest tree within the segment. Segments with many trees where a single prominent tree species could not be identified were not included in the analysis. 'Tree section' corresponds to segments that contained only a part of a tree. Fig. 5 presents the profile category distribution by species for the test set. For each species class, the majority of the segments belong to the 'Single tree' category. However, it is worth mentioning that in the study area, alders typically grow in places with lots of other alders and, therefore, 30% of segments labeled as alder belong to the 'Many trees of the same species' category. Furthermore, 22% of maples and oaks belong to the 'Tree section' category because of over-segmentation of their wide canopies.
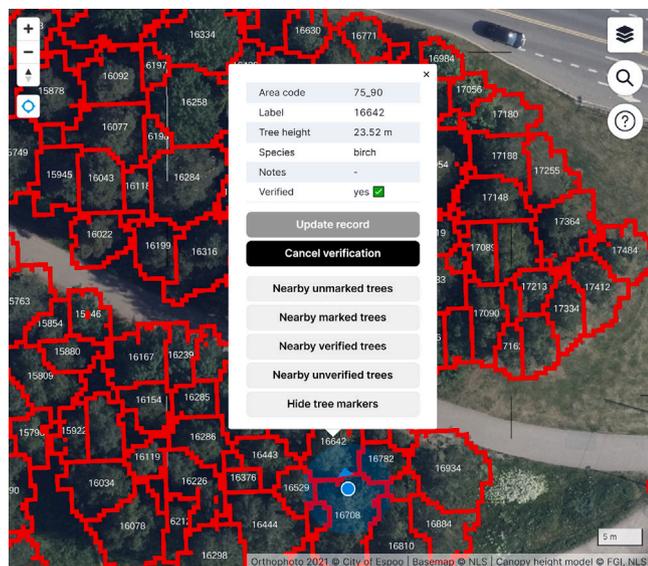
Fig. 4. Screenshot of the crowdsourcing application used to collect ground-truth species data. In this example, the application shows a true orthophoto map, segment boundaries (red lines), the surveyor's location (blue circle) based on GNSS positioning, and the information of the nearest segment together with possible actions for the surveyor. The true orthophoto map in the background was obtained from the City of Espoo (2024).

The segments were also categorized to different crown classes using alternative criteria based on tree heights and distances to neighboring trees as inspired by Wang et al. (2019a). Segments were considered 'Isolated' if within 8 m radius, there were only segments with a height of at most half of the height of the 'Isolated' segment. Segments were considered 'Dominant' if the segment was not 'Isolated' and it was the tallest segment within 8 m radius. If there was at least one segment at least 5 m taller than a given segment within 6 m radius, the segment was assigned to the 'Smaller tree next to larger tree' category. 'Roadside' segments were detected using the open tree database of the City of Espoo (2024). All other segments were assigned to the 'Co-dominant' category.

3. Benchmark contest and species classification algorithms

3.1. Overview of the benchmark contest

The benchmark dataset was released at the Second ISPRS GEOBENCH workshop, "Evaluation and Benchmarking of Sensors, Systems and Geospatial Data in Photogrammetry and Remote Sensing," held in Krakow, Poland, on October 23–24, 2023. The initial submission deadline at the end of January 2024 was extended to the summer of 2024. This extension was necessary to allow for thorough checking of the reference data and to ensure accurate and consistent contributions.

Participating partners were asked to submit the following: (1) a description of their applied methods, including relevant citations if previously published; (2) an Excel spreadsheet containing the tree segment number, classified species number, and sub-region code (A or B) for each segment (study area was divided into two sub-regions); and (3) if modified segments (i.e., divided or merged) were used, a raster image of the final segment numbers, matching the original segment number raster in cell size and location.

FGI also developed classification methods that were submitted for benchmarking. Therefore, the analysis of the classification results was performed by separate people at the FGI.

3.2. Participating methods

Submissions to the benchmarking competition included point-based (3D) deep learning methods, profile-based (2D) deep learning methods and shallow machine learning methods. The following sections describe the participating methods in detail. We adopt the following naming convention for the participating methods: ⟨organization name⟩-⟨method details⟩-⟨classifier category⟩-⟨2D or 3D-based method⟩. As an exception, the variants of the DetailView method do not fit this naming convention as discussed in Section 3.2.2.

3.2.1. Point-based (3D) deep learning methods

We summarize the studied point-based deep learning methods in Table 3, where we present relevant details for each of the methods, including information on point cloud pre-processing, format of the input data, model architecture, training, and inference. We refer to the point-based methods also as 3D deep learning methods.

In this work, our comparison includes the following 3D deep learning methods: FGI-PointTransformer-DL-3D and its variants, FGI-DGCNN-DL-3D, FGI-PointNet-DL-3D, FGI-Point2Vec-DL-3D, TUW-PointNet++-DL-3D, and FBK-PointNet++-DL-3D. The studied methods utilize various point-based deep learning models, including Point Transformer (Zhao et al., 2021) for FGI-PointTransformer-DL-3D, dynamic graph convolutional net (Wang et al., 2019b) for FGI-DGCNN-DL-3D, PointNet (Qi et al., 2017a) for FGI-PointNet-DL-3D, Point2Vec (Zeid et al., 2023) for FGI-Point2Vec-DL-3D, and PointNet++ (Qi et al., 2017b) for TUW-PointNet++-DL-3D and FBK-PointNet++-DL-3D. For most of the methods, nearest neighbor interpolation is used to represent the multispectral point cloud in a format, where each point contains the radiometric information from all of the three channels. As an exception, TUW-PointNet++-DL-3D does not perform nearest neighbor interpolation. Instead, each point is assigned a single intensity value, and three flag variables indicate the source channel of the intensity.

There were differences in the used radiometric features and echo properties between the different models in the case of the HeliALS dataset that included the intensity, amplitude, reflectance, echo deviation and an echo return number for each point. FGI-PointTransformer-DL-3D, FGI-DGCNN-DL-3D, FGI-PointNet-DL-3D, and FGI-Point2Vec-DL-3D utilized all of the radiometric features and echo properties, implying that each 3D point was assigned 13 additional radiometric and echo features. TUW-PointNet++-DL-3D only utilized the intensities and FBK-PointNet++-DL-3D only considered the reflectances in the model input. For FGI-PointTransformer-DL-3D, we also studied different combinations of the included radiometric and echo features in order to evaluate their effect on the classification accuracy. In FGI-PointTransformer-VUX-DL-3D, we utilized the geometric information from the three scanners and radiometric and echo features only from the VUX scanner. As a further variant, FGI-PointTransformer-VUXReflectance-DL-3D only utilized the reflectance of the VUX scanner while disregarding all other radiometric and echo features. Finally, FGI-PointTransformer-GeometryOnly-DL-3D utilized no radiometric or echo information from any of the scanners, thus serving as a baseline when no radiometric information is available.

When it comes to the format of the input data, most of the software implementations of the studied architectures, such as DGCNN, PointNet++ and PointNet, required a fixed point count per segment. The input point count ranges from 500 points per segment for FBK-PointNet++-DL-3D to 8192 points per segment for FGI-PointTransformer-DL-3D, FGI-DGCNN-DL-3D, FGI-PointNet-DL-3D, and TUW-PointNet++-DL-3D. Importantly, the software implementation of the Point Transformer architecture does not set constraints to the point count, but we still used a fixed point count to obtain a fair comparison between the different architectures. Before feeding the points to the model, FGI-PointTransformer-DL-3D, FGI-DGCNN-DL-3D, FGI-PointNet-DL-3D normalize the coordinates of the input point cloud into
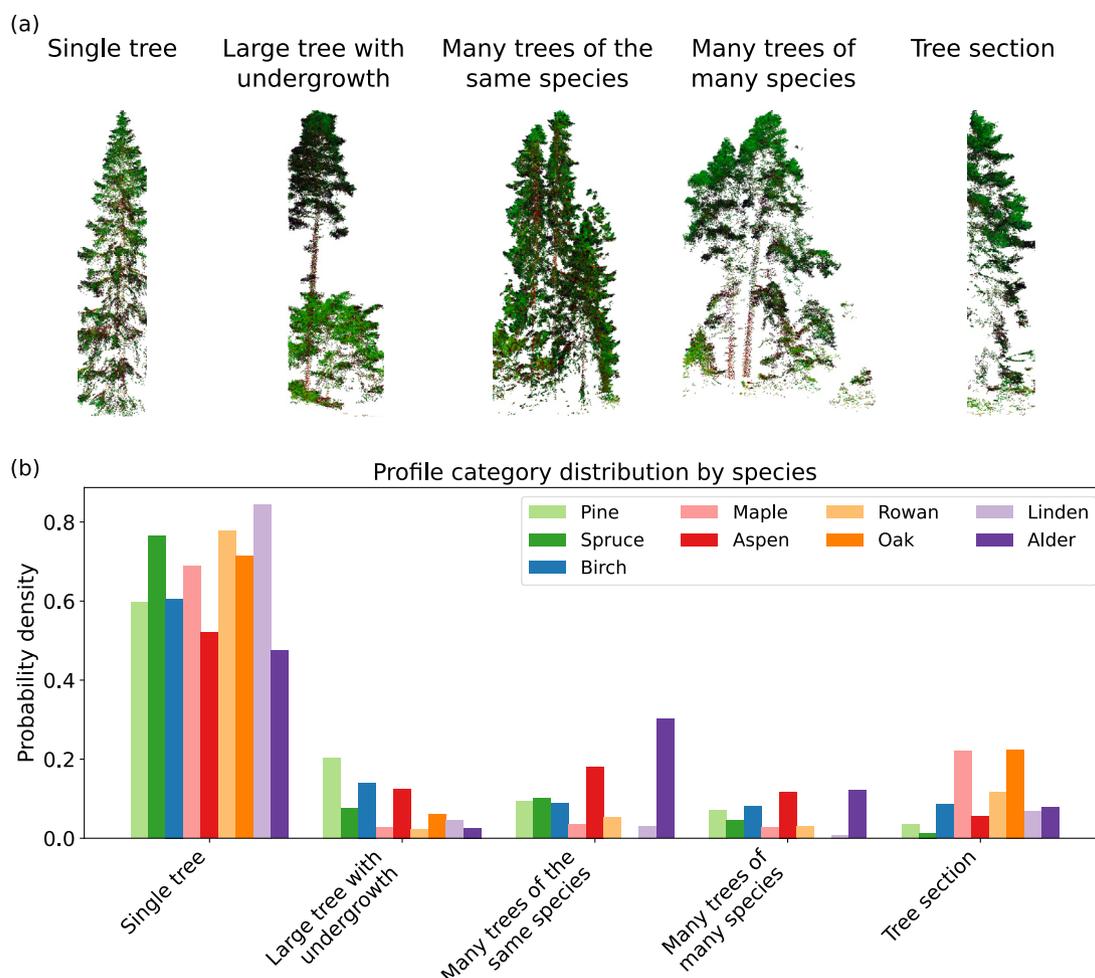
Fig. 5. (a) Point cloud projections of segment instances belonging to the different profile categories. (b) Profile category distribution by species for the test set.

a unit sphere, whereas TUW-PointNet++-DL-3D and FBK-PointNet++-DL-3D center the coordinates, hence preserving the scale.

Some of the studied methods attempt to mitigate the class imbalance of the training set. FGI-PointTransformerWeighted-DL-3D, FGI-Point2Vec-DL-3D and TUW-PointNet++-DL-3D utilize a weighted loss function to provide a higher weight to minority species, such as oak, alder and rowan. FBK-PointNet++-DL-3D instead augments the training dataset by generating geometric augmentations for rowan and linden, and geometric and spectral augmentations for alder and oak. On the other hand, TUW-PointNet++-DL-3D and the methods by FGI use augmentations to all segments during training without any specific focus on the minority species. The methods by FGI apply small random translation, scaling, and jittering of coordinates, as well as small random scaling and jittering of other attributes. TUW-PointNet++-DL-3D generates six rotational augmentations around the z-axis for each segment.

At inference time, the methods by FGI utilize an ensemble voting strategy based on five trained models obtained with different random initializations. TUW-PointNet++-DL-3D and FBK-PointNet++-DL-3D directly utilize the prediction of the trained model without any ensemble voting strategies.

3.2.2. Profile-based (2D) deep learning methods

We summarize the profile-based, i.e., 2D deep learning methods in Table 4, where we present relevant details for each of the methods, including information on point cloud pre-processing, format of the input data, model architecture, training, and inference.

Our comparison of 2D deep learning approaches includes four methods: SLU-YOLOv8-DL-2D, DetailView-DL-2D, SA-Detailview-DL-2D, and

NTNU-ConvNeXt-T-DL-2D. We note that SA-Detailview-DL-2D represents a sensor agnostic (SA) deep learning model that has been previously trained on the FOR-species20K dataset (Puliti et al., 2024) without any re-training on the provided training dataset. All the other models, including Detailview-DL-2D, have been trained on the training set provided as a part of the benchmarking competition. When it comes to the used spectral features, NTNU-ConvNeXt-T-DL-2D utilized radiometric information from all of the three channels. SLU-YOLOv8-DL-2D took into account both geometry and spectral information from VUX and mini-VUX scanners, while ignoring the data of the green channel. In contrast, DetailView utilized only the geometry of the multispectral point cloud without any radiometric information.

The studied 2D deep learning methods are based on various convolutional neural net backbones, including YOLOv8 (Jocher et al., 2023) for SLU-YOLOv8-DL-2D, DenseNet-201 (Huang et al., 2017) for DetailView, and ConvNeXt-T (Liu et al., 2022b) for NTNU-ConvNeXt-T-DL-2D. DetailView and NTNU-ConvNeXt-T-DL-2D convert the segment-wise point clouds into multiple depth image projections from different angles. On the other hand, SLU-YOLOv8-DL-2D projects the point clouds into multiple RGB images, where the RGB values of each pixel are based on the mean intensities of VUX, mini-VUX and their combination. As their input, all the models take between four and seven 2D projections from different viewing angles. The resolutions of the projections range from 128 × 128 pixels for NTNU-ConvNeXt-T-DL-2D to 256 × 256 pixels for DetailView and 160 × 320 pixels for SLU-YOLOv8-DL-2D. NTNU-ConvNeXt-T-DL-2D rescales the point cloud projection to 256 × 256 pixels before the input layer of the model. All of the methods use side-view projections corresponding to different

Table 3

Summary of point-based (3D) deep learning methods for species classification. For each method, the table briefly describes point cloud pre-processing steps, input data format of the deep learning model, architecture of the model, training strategies, and inference strategies.

Method and data	Pre-processing	Input data	Architecture and training	Inference
FGI-PointTransformer-DL-3D, HeliALS	Removal of noise and ground points. Voxelization with a side length of 5 cm to obtain a more balanced point distribution. Point clouds of the three channels merged and each point assigned radiometric information from the two other channels based on nearest neighbor interpolation. Imputation strategies used if point attributes are missing for some of the channels.	(3+13)D point cloud with 8192 points per segment. Besides coordinates, attributes include intensities, amplitudes, reflectances, and echo deviations for the 3 scanners, and an echo return number. Coordinates normalized to unit sphere, other attributes normalized by scaling (except for the echo return number).	Five randomly initialized Point Transformer models (Zhao et al., 2021) trained from scratch with 95–5 train-validation split. Both ordinary cross-entropy loss (FGI-PointTransformer-DL-3D) and weighted cross-entropy loss (FGI-PointTransformerWeighted-DL-3D) were tested. Augmentations applied at random (small random translation, scaling, and jittering of coordinates, and small random scaling and jittering of other attributes).	Majority voting based on the five trained Point Transformer models.
FGI-PointTransformer-VUX-DL-3D, HeliALS	Same as FGI-PointTransformer-DL-3D but using radiometric and echo information only from the VUX scanner while ignoring radiometric information of the two other scanners.	(3+6)D point cloud with 8192 points per segment. Besides point coordinates, attributes include the intensity, echo deviation, amplitude, reflectance, number of returns, and return number of the VUX scanner. Normalization as for FGI-PointTransformer-DL-3D.	Same as FGI-PointTransformer-DL-3D. Both ordinary and weighted cross-entropy loss were tested.	Same as FGI-PointTransformer-DL-3D.
FGI-PointTransformer-VUXReflectance-DL-3D, HeliALS	Same as FGI-PointTransformer-DL-3D but using only the reflectance from the VUX scanner and ignoring any other radiometric or echo information.	(3+1)D point cloud with 8192 points per segment. Besides coordinates, attributes include the reflectance of the VUX scanner. Normalization as for FGI-PointTransformer-DL-3D.	Same as FGI-PointTransformer-DL-3D. Both ordinary and weighted cross-entropy loss were tested.	Same as FGI-PointTransformer-DL-3D.
FGI-PointTransformer-GeometryOnly-DL-3D, HeliALS	Same as FGI-PointTransformer-DL-3D but using only the geometry of the combined point cloud.	3D point cloud with 8192 points per segment. Coordinates normalized to unit sphere.	Same as FGI-PointTransformer-DL-3D with ordinary cross-entropy loss.	Same as FGI-PointTransformer-DL-3D.
FGI-DGCNN-DL-3D, HeliALS	Same as FGI-PointTransformer-DL-3D.	Same as FGI-PointTransformer-DL-3D.	Five randomly initialized dynamic graph convolutional neural net (DGCNN) models (Wang et al., 2019b) trained from scratch with 90–10 train-validation split and cross-entropy loss. Same augmentations as FGI-PointTransformer-DL-3D.	Majority voting based on the five trained DGCNN models.
FGI-PointNet-DL-3D, HeliALS	Same as FGI-PointTransformer-DL-3D.	Same as FGI-PointTransformer-DL-3D.	Five randomly initialized PointNet models (Qi et al., 2017a) trained from scratch with 90–10 train-validation split and cross-entropy loss. Same augmentations as FGI-PointTransformer-DL-3D.	Majority voting based on the five trained PointNet models.
FGI-Point2Vec-DL-3D, HeliALS	Same as FGI-PointTransformer-DL-3D, HeliALS.	(3+13)D point cloud with 4096 points per segment. Besides coordinates, attributes include intensities, amplitudes, reflectances, and echo deviations for the 3 scanners, and an echo return number. Coordinates normalized to unit sphere, other attributes z-score normalized.	Point2Vec model (Zeid et al., 2023) in fully-supervised classification setup. Trained from scratch using 80–20 train-validation split. Geometric augmentations applied at random (scaling, translation and rotation in 30° degree steps around the z-axis). Weighted cross-entropy loss with label smoothing.	Majority voting based on five inference runs with a single model.
TUW-PointNet++-DL-3D, HeliALS	Point clouds of the three channels combined by introducing three flag variables indicating the source channel of each point.	(3+4)D point cloud with 8192 points per segment. Besides coordinates, attributes include the three flag variables (one of which is set to one), and the intensity value. Coordinates centered, but not scaled to preserve height.	PointNet++ model (Qi et al., 2017b) with two PointNet layers and a final multilayer perceptron layer trained from scratch with 67–33 train-validation split. Geometric augmentations by generating 6 rotations around the z-axis for each segment. Weighted loss function to mitigate class imbalance. The model extends Ensemble PointNet++ (Winiwarter, 2024; Puliti et al., 2025) to multispectral data.	Prediction of the trained model.

Table 3 (*continued*).

Method and data	Pre-processing	Input data	Architecture and training	Inference
FBK-PointNet++-DL-3D, HeliALS and Optech Titan	Removal of noise and ground points. Point clouds of the three channels were merged and each point was assigned radiometric information from the three channels based on nearest neighbor interpolation.	(3+3)D point cloud with 500 points per segment. Besides coordinates, attributes include the reflectance for each channel. Coordinates centered, but not scaled. Reflectances normalized.	PointNet++ model (Qi et al., 2017b) trained from scratch with 80–20 train-validation split. Geometric and spectral augmentations for alder and oak to triple the number of examples. Geometric augmentations for rowan and linden to double the number of examples.	Prediction of the trained model.
FGI-PointTransformer-DL-3D, Optech Titan	Removal of noise and ground points. Point clouds of the three channels merged and each point assigned intensity information from the two other channels based on nearest neighbors.	(3+4)D point cloud with 2048 points per segment. Besides coordinates, attributes include intensities for the 3 scanners, and an echo return number. Normalization as for the HeliALS dataset.	Same as FGI-PointTransformer-DL-3D for the HeliALS dataset using ordinary cross-entropy loss.	Same as FGI-PointTransformer-DL-3D for the HeliALS dataset.
FGI-PointNet-DL-3D, Optech Titan	Same as FGI-PointTransformer-DL-3D, Optech Titan.	Same as FGI-PointTransformer-DL-3D, Optech Titan.	Same as FGI-PointNet-DL-3D, HeliALS.	Majority voting based on the five trained PointNet models.
FGI-DGCNN-DL-3D, Optech Titan	Same as FGI-PointTransformer-DL-3D, Optech Titan.	Same as FGI-PointTransformer-DL-3D, Optech Titan.	Same as FGI-DGCNN-DL-3D, HeliALS.	Majority voting based on the five trained DGCNN models.

rotations around the z-axis. Furthermore, DetailView utilizes a top-view projection, a bottom-view projection, and a projection to a close-up of the trunk. In addition to the side-, top- and bottom-view projections, NTNU-ConvNeXt-T-DL-2D includes a front-view depth image and 20 additional global features mostly based on the intensities of the three channels. While most of the methods utilize orthographic projections, NTNU-ConvNeXt-T-DL-2D follows the approach by Goyal et al. (2021) and performs the projection using a pinhole camera model (perspective projection).

To generate more training examples, DetailView utilizes point cloud random subsampling and rotations around the z-axis as augmentations. SLU-YOLOv8-DL-2D utilizes four rotational augmentations around the z-axis per segment, while NTNU-ConvNeXt-T-DL-2D applies random translation and scaling in addition to random rotations around the z-axis. At inference time, SLU-YOLOv8-DL-2D and DetailView use an ensemble voting strategy to obtain the final prediction, whereas NTNU-ConvNeXt-T-DL-2D directly uses the output of the model for prediction. DetailView performs the prediction for 50 different augmentations of the segment increasing the required amount of computation.

3.2.3. Machine learning methods

We summarize the studied machine learning classifiers in Table 5, where we present relevant details for each of the methods, including information on point cloud pre-processing, feature extraction, classifier type, model training, and inference.

Our comparison includes the following methods: IBL-BalancedRF-ML, UEF-RF/LGBM/SVM-ML, IDEAS-RF-ML, FGI-RF-ML, LUKE-MultiRF-ML, UPV-GB-ML, and Aalto-RF-ML. Most of the studied methods, such as IBL-BalancedRF-ML, UEF-RF-ML, IDEAS-RF-ML, FGI-RF-ML, LUKE-MultiRF-ML, and Aalto-RF-ML, utilize a random forest classifier or its variant for the species prediction. IBL-BalancedRF-ML applies a balanced RF procedure (Chen et al., 2004) to up-sample the number of training examples for minority species, whereas LUKE-MultiRF-ML utilizes multiple RF models for the species prediction. In addition to the RF classifiers, our comparison includes a support vector machine (SVM) used by UEF-SVM-ML, and variants of gradient boosting (GB) techniques used by UEF-LGBM-ML and UPV-GB-ML.

All of the studied methods utilize hand-crafted geometric and radiometric features from the three channels, though the number of features and details vary from one method to another. For example, UEF-RF/LGBM/SVM-ML utilizes only a single geometric feature, whereas UPV-GB-ML evaluates 384 geometric features before dimensionality reduction. Apart from IDEAS-RF-ML and UPV-GB-ML, all the methods use

features combining intensities from multiple channels. Furthermore, several methods, such as UEF-RF/LGBM/SVM-ML, IDEAS-RF-ML, and FGI-RF-ML, include intensity histograms or intensity statistics as a function of height as their features. Laser echo properties are considered in the features by UEF-RF-ML, FGI-RF-ML, and LUKE-MultiRF-ML. Apart from LUKE-MultiRF-ML and Aalto-RF-ML, the studied methods apply some dimensionality reduction strategy to determine the most important features. After dimensionality reduction, the number of features varies between 18 and 143 for the different methods.

The majority of the methods do not utilize any strategy for species imbalance mitigation or training data augmentation. IBL-BalancedRF-ML, LUKE-MultiRF-ML, and Aalto-RF-ML use over-sampling strategies for the minority species in an attempt to improve their classification accuracy. At inference time, most of the methods directly use the trained classifier to obtain the prediction. As an exception, LUKE-MultiRF-ML utilizes nine single-species classifiers and one default classifier, which is used for the classification task if the results of the single-species classifiers are not consistent.

3.3. Accuracy analysis

The benchmarking results were evaluated with five main evaluation metrics: overall accuracy, precision, recall, F_1 score and macro-average accuracy. Additionally, confusion matrices were computed for a subset of the competition results.

The overall accuracy (OA) is defined as the proportion of correct classifications among all samples

$$\text{OA}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \mathbb{1}(\hat{y}_i = y_i), \tag{1}$$

where $\hat{y}_i$ is the predicted species of the ith segment, y_i is the corresponding ground-truth species, $\mathbb{1}(\cdot)$ is an indicator function and n_{samples} is the total number of ground-truth tree segments.

The species-wise precision, i.e., user's accuracy, and recall, i.e., producer's accuracy, are given as

$$\text{Precision}_s = \frac{\text{TP}_s}{\text{TP}_s + \text{FP}_s}, \tag{2}$$

$$\text{Recall}_s = \frac{\text{TP}_s}{\text{TP}_s + \text{FN}_s}, \tag{3}$$

where the true positives TP_s, false positives FP_s and false negatives FN_s can be obtained from the confusion matrix for each tree species class

Table 4

Summary of profile-based (2D) deep learning methods for species classification. For each method, the table briefly describes point cloud pre-processing steps, input data format of the deep learning model, architecture of the model, training strategies, and inference strategies.

Method and data	Pre-processing	Input data	Architecture and training	Inference
SLU-YOLOv8-DL-2D, HeliALS	For each segment, points from VUX and mini-VUX were merged and the point cloud was projected to four 2D RGB images corresponding to different rotations around the z-axis (0°, 45°, 90°, or 135°). Green channel was not used.	$4 \times$ 2D RGB images (160 $\times$ 320 pixels) per segment. For each pixel, the RGB values correspond to a mean mini-VUX intensity (NIR), mean VUX intensity (SWIR), and a combination of the two, respectively.	YOLOv8 (Jocher et al., 2023) pretrained on the ImageNet dataset (Deng et al., 2009) was finetuned over 15 training epochs. The RGB images for rotation angles 0°, 45°, and 90° were used in the training set, while the RGB images for the angle 135° formed the validation set.	Trained YOLOv8 model predicted the species for each of the four RGB images generated from a single segment, and the results were aggregated to obtain the final prediction.
DetailView-DL-2D, HeliALS	Removal of noise and ground points. Voxelization with a side length of 5 cm to obtain a more balanced point distribution. Point clouds of the three channels merged. Intensity and reflectance ignored.	$(4 + 1 + 1 + 1) \times$ 2D depth images (256 $\times$ 256 pixels) per segment. Intensity information was not used. Out of the seven depth images, four were from the sides, one from the top, one from the bottom, and one representing a close-up between $z = 1.0$ m and $z = 1.5$ m.	DetailView neural net model (Puliti et al., 2025; Frey and Schindler, 2024) based on DenseNet-201-instances (Huang et al., 2017) trained from scratch over 24 epochs using 90–10 train-validation split. The batch size was set twice as large and the learning rate a factor of 10 smaller compared to Puliti et al. (2025). Augmentations included point cloud random subsampling and rotations around the z-axis.	Predicted class probabilities averaged across 50 augmentations of each segment.
SA-Detailview-DL-2D, HeliALS	Point clouds of mini-VUX (NIR) and VUX (SWIR) merged. Removal of ground points. The green channel was ignored.	Same as DetailView-DL-2D.	DetailView neural net model (Puliti et al., 2025; Frey and Schindler, 2024) based on DenseNet-201-instances (Huang et al., 2017). The model had been previously trained on the FOR-species dataset with 33 different species and 98-2 training-validation split using same augmentations as DetailView-DL-2D. **The model was NOT re-trained using the provided training set.**	Predicted class probabilities averaged across 50 augmentations of the segment. If predicted tree species was not present in the current dataset, it was reclassified to a species in the dataset based on the largest relative accuracy in the confusion matrix.
NTNU-ConvNeXt-T-DL-2D, Optech Titan	Point clouds of the three scanners merged. For each segment, the point cloud was converted to 6 depth images corresponding to 4 different rotations around the z-axis and 2 projections along the negative and positive z-axis following Goyal et al. (2021). Furthermore, a front-view depth image was obtained between $z = 1.0$ m and $z = 2.5$ m.	$(6 + 1) \times$ 2D depth images (128 $\times$ 128 pixels that was rescaled before input layer to 256 $\times$ 256 pixels) per segment. Furthermore, 20 additional global features were computed for each segment based on the tree height and intensities of the 3 channels following Yu et al. (2017). Single-channel intensity features included, e.g., minimum, maximum, and moments.	Neural net with 3 input branches corresponding to the 6 depth images from the sides, the front-view depth image, and the global features. ConvNeXt-T (Liu et al., 2022b) provided the backbone for the two image branches, and a multi-layer perceptron block for the global features. Training-validation split of 80–20 with data augmentations consisting of random rotations around the z-axis, translations and scaling.	Prediction of the trained model.

s. In this work, the recall metric is also referred to as the species-wise accuracy.

The F_1 score for a single species class is defined as

$$F_1^{(s)} = 2 \cdot \frac{\text{Precision}_s \cdot \text{Recall}_s}{\text{Precision}_s + \text{Recall}_s}. \qquad (4)$$

Furthermore, macro F_1 is the arithmetic mean over all the species-wise F_1 scores.

In an attempt to evaluate the performance of the studied methods in the presence of substantially unbalanced dataset, we used the macro-average accuracy. The macro-average accuracy (also known as macro-average recall) is given as:

$$\text{Macro-average accuracy} = \frac{1}{|S|} \sum_{s \in S} \frac{\text{TP}_s}{\text{TP}_s + \text{FN}_s} \qquad (5)$$

where $|S|$ denotes the total number of tree species in the benchmarking dataset.

Moreover, we used the classification error as a performance metric when appropriate: the overall classification error refers to the complement of the overall accuracy, i.e.,

$$\varepsilon_{\text{OA}} = 1 - \text{OA}, \qquad (6)$$

where ε_{OA} denotes the overall classification error. Similarly, the macro-average classification error is the complement of the macro-average accuracy.

We further estimated the uncertainty of overall accuracy and macro-average accuracy for each of the classification methods using a bootstrapping approach. Based on the test set of our dataset, we generated 2000 bootstrapping datasets of 5261 tree segments using random sampling with replacement. For each classification method, we recalculated the accuracy metrics for each bootstrapping dataset, resulting in a bootstrapping distribution of 2000 accuracy values. We derived a 95% confidence interval (CI) by taking the 2.5th and 97.5th percentiles of the distribution. The estimated confidence intervals reflect the uncertainty arising from the finite test set, but do not include, e.g., randomness in the model training.

Table 5

Summary of machine learning methods for species classification. For each method, the table briefly describes point cloud pre-processing steps, features of the classifier, type of the classifier, and training strategies.

Method and data	Pre-processing and features	Input data	Classifier and training
IBL-BalancedRF-ML, Optech Titan	Segments with under- or over-segmentation errors were identified and removed, which resulted in 660 training segments. Following Kamińska et al. (2021) and Lisiewicz et al. (2025), features were computed, including geometric features based on the merged point cloud, and intensity features based on individual channels and their combinations. For most features, calculations were performed for points situated above half the height of each segment.	A feature vector of 30 elements after dimensionality reduction. Variables based on combinations of different channels had the highest importance.	Random forest model was trained with a balanced RF approach (Chen et al., 2004), up-sampling minority species. To ensure stable results and mitigate overfitting, 5-fold cross-validation was repeated 20 times.
UEF-RF/LGBM/SVM-ML, Optech Titan	Following Holmgren and Persson (2004), Ørka et al. (2009), Ørka et al. (2012), and Hovi et al. (2016), intensity features were computed, including single-channel moments as a function of height for each echo type, and two multi-channel features known as the normalized difference water index and conifer index (Trier et al., 2018). The only geometric feature was the 95th percentile of the z coordinate.	A feature vector of 143 elements after dimensionality reduction and 259 elements prior to dimensionality reduction. In the dimensionality reduction, correlation coefficient was evaluated for each pairs of features, and one of the features in the pair was removed if the correlation coefficient exceeded 0.9.	Three ML methods were tested for the classification, including random forest, light gradient boosting machine (LGBM), and support vector machine (SVM) using radial basis function (RBF) kernel. 80–20 train–test split used. Grid search and cross validation were utilized to optimize hyperparameters. No augmentation or species imbalance mitigation.
IDEAS-RF-ML, Optech Titan	Features were extracted separately from the full tree point cloud (with the bottom 2 m excluded) and crown points (isolated with a 1 m buffer around CHM). In total, 122 features were computed across both point clouds, including point count, height distribution metrics, intensity statistics per each channel, and features based on correlations between height and intensity.	A feature vector of 122 elements before dimensionality reduction and 90 elements after dimensionality reduction.	Random forest classifier trained using 80–20 train–test split. Optimal number of features and hyperparameters were selected using a grid search by optimizing the F_1 score. No augmentation or species imbalance mitigation.
FGI-RF-ML, HeliALS and Optech Titan	92 features computed for each segment before dimensionality reduction. Following Lehtomäki et al. (2016), Yu et al. (2017), and Hakula et al. (2023), features included geometric features using the channel with most points, intensity statistics including, e.g., normalized intensity histograms for each channel, proportions of laser pulse return types (e.g. first of many), and histogram of local descriptors.	A feature vector of 92 elements before dimensionality reduction and 20–40 elements after dimensionality reduction.	Random forest classifier trained based on out-of-bag prediction. During training, feature selection and model optimization were performed using a grid search based on the out-of-bag prediction. No augmentation or species imbalance mitigation.
LUKE-MultiRF-ML, Optech Titan	36 features computed for each segment, including, e.g., moments and quantiles of the z-coordinate and the radial distance from the center of the segment, average intensities of the channels, average intensities of the first and other echoes for each channel, and features combining intensities of two channels.	A feature vector of 36 elements.	10 random forest models trained using 5-fold cross validation, including a default model classifying segments into the nine species and nine single-species models. Synthetic minority over-sampling technique (Chawla et al., 2002) used to balance the training set with the oversampling parameter set to 10 for pine, spruce, and birch, and to 15 for the rest. If exactly one of the single-species classifiers predicted the segment to belong to its species at inference time, the segment was assigned to this species. Otherwise, the species of the segment was assigned based on the default model.

3.4. Scaling analysis of classification error

In addition to benchmarking the classification methods on the fixed training and test set introduced in Section 2.4, we also studied the scaling of the classification error as a function of training set size and point density. We conducted these experiments using the deep learning and machine learning models achieving the highest overall accuracy on the HeliALS dataset in the benchmark contest. In the scaling studies, we excluded 88 segments from the original training set of the benchmarking competition due to non-optimal profile images, sparse point clouds, or building intersection within the segment boundaries. In addition, the species label of 18 segments from the original training set was corrected due to errors detected during the on-site verification. Compared to the test set of the benchmark contest, the dataset for scaling studies

Table 5 (*continued*).

Method and data	Pre-processing and features	Input data	Classifier and training
UPV-GB-ML, Optech Titan	Point clouds were filtered to remove ground points and outliers. In total, 384 geometric features and 48 single-channel intensity features (16 for each channel) were computed. Single-channel intensity features included, e.g., min, mean, standard deviation and different percentiles. Similar statistical metrics were evaluated for geometrical neighborhood features, all included in the Class3Dp software (Carbonell-Rivera et al., 2024), such as, number of returns, eigenvalues of the covariance matrix, and features based on West et al. (2004), Rusu (2010), and Demantké et al. (2012).	A feature vector of 18 elements after dimensionality reduction. The dimensionality reduction was based on ranking the features based on their permutation importance and then evaluating the learning curve. The learning curve plateaued after 18 features.	Gradient boosting (GB) machine used for classification. The training was performed using 5-fold cross validation. No augmentation or species imbalance mitigation.
Aalto-RF-ML, Optech Titan	34 features computed for each segment, including e.g., intensity statistics for each channel based on points within the top 3 m of the segment, crown diameter within and below the top 3 m of the segment, the apex angle of the tree, and features combining intensities of two channels following Hakula et al. (2023).	A feature vector of 34 elements.	Random forest classifier trained using 80–20 training validation split. To mitigate class imbalance, new feature vectors were simulated for each species with fewer than 125 examples using the class-wise mean and standard deviation of the feature vector (Rönnholm et al., 2022). After augmentation, each class had at least 125 training examples.

included 19 additional segments. In total, the dataset for scaling law studies consisted of 6257 tree segments. Below, this dataset is called the scaling analysis dataset. Importantly, these modifications to the dataset did not significantly affect the classification results. For example, FGI-PointTransformer-DL-3D obtained an overall accuracy of 87.9% when trained on the original training set of 1065 segments in the benchmark competition. The same model obtained an overall accuracy of 88.5% when trained on 1000 segments from the scaling analysis dataset.

In the scaling studies, we estimated the overall and macro-average classification errors of the selected classification models using 5-fold cross-validation. The folds were selected using stratified random sampling to ensure that each fold follows the same species distribution. For the scaling experiments with varying training set size, we utilized 5-fold cross-validation on subsets of the scaling analysis dataset, thus effectively creating training sets ranging from approximately 250 segments to approximately 5000 segments. Importantly, each fold in the subsets had a similar species distribution as the original dataset as a result of another round of stratified random sub-sampling from the folds previously generated from the full dataset.

For the scaling experiments as a function of point density, we randomly sub-sampled the HeliALS tree segments to a desired point density. We varied the point density from 10^{-1} pts/m^2 representing an extremely sparse point cloud to around 1000 pts/m^2 representing a dense ALS point cloud. We evaluated the classification performance using 5-fold cross-validation on the entire sub-sampled scaling analysis dataset.

4. Results and discussion

In this section, we evaluate the performance of the studied species classification methods on the multispectral data collected with the HeliALS and Optech Titan systems. First in Section 4.1, we present the results of the benchmarking contest and discuss key factors affecting the classification accuracy, such as the use of a deep learning or a machine learning model, model architecture and training strategy. In Section 4.2, we analyze the effect of multispectral information on classification accuracy. Section 4.3 discusses the relation between classification accuracy and species. In Section 4.4, we analyze the scaling of classification accuracy for selected deep learning and machine learning models as a function of training set size and point density to study the impact of these factors in more detail and to highlight the differences

of deep learning and machine learning methods. Subsequently, we investigate the effect of segmentation quality and crown class on species classification in Section 4.5. Finally, we provide further discussion and compare our results to previous studies in Section 4.6.

4.1. International benchmarking of species classification methods

The species classification methods introduced in Section 3.2 were benchmarked using overall accuracy, macro-average accuracy, species-wise precision and recall as the key metrics. We compare the overall and macro-average accuracy for the methods using the HeliALS data in Fig. 6(a) and for the methods using the Optech Titan data in Fig. 6(b). In general, the methods using the HeliALS data reached a higher overall and macro-average accuracy compared to methods using the Optech Titan data, which can be attributed to higher point density and geometric accuracy of the point cloud provided by the HeliALS system, see Section 2.2.

For the HeliALS dataset, FGI-PointTransformer-DL-3D reached the highest overall accuracy of 87.9%, and the corresponding method with a weighted loss function, FGI-PointTransformerWeighted-DL-3D, achieved the highest macro-average accuracy of 74.5%. In addition to the point transformer model, the other best performing methods on the HeliALS dataset were also point-based deep learning methods, with FGI-DGCNN-DL-3D and FGI-PointNet-DL-3D reaching an overall (macro-average) accuracy of 87.0% (67.3%) and 85.9% (64.2%), respectively. The best image-based deep learning method, DetailView-DL-2D, reached an overall (macro-average) accuracy of 84.3% (63.9%) and the only machine learning model, FGI-RF-ML, had an overall (macro-average) accuracy of 83.2% (61.3%). Based on the confidence intervals from bootstrapping, the overall accuracy of FGI-PointTransformer-DL-3D was within the confidence intervals of FGI-PointTransformerWeighted-DL-3D and FGI-DGCNN-DL-3D, suggesting that the difference between these methods is not statistically significant. The total width of the 95% confidence intervals for the overall accuracy ranged between 1.7% and 2.7% across the methods using the HeliALS dataset.

For most of the methods, the macro-average accuracy was significantly lower than the overall accuracy due to the inherent difficulty of predicting the minority species, such as alder or oak. The point transformer model clearly outperformed the other methods in the prediction of the minority species, which resulted in a significant improvement of the macro-average accuracy, especially when using the weighted loss

(a)

(b)
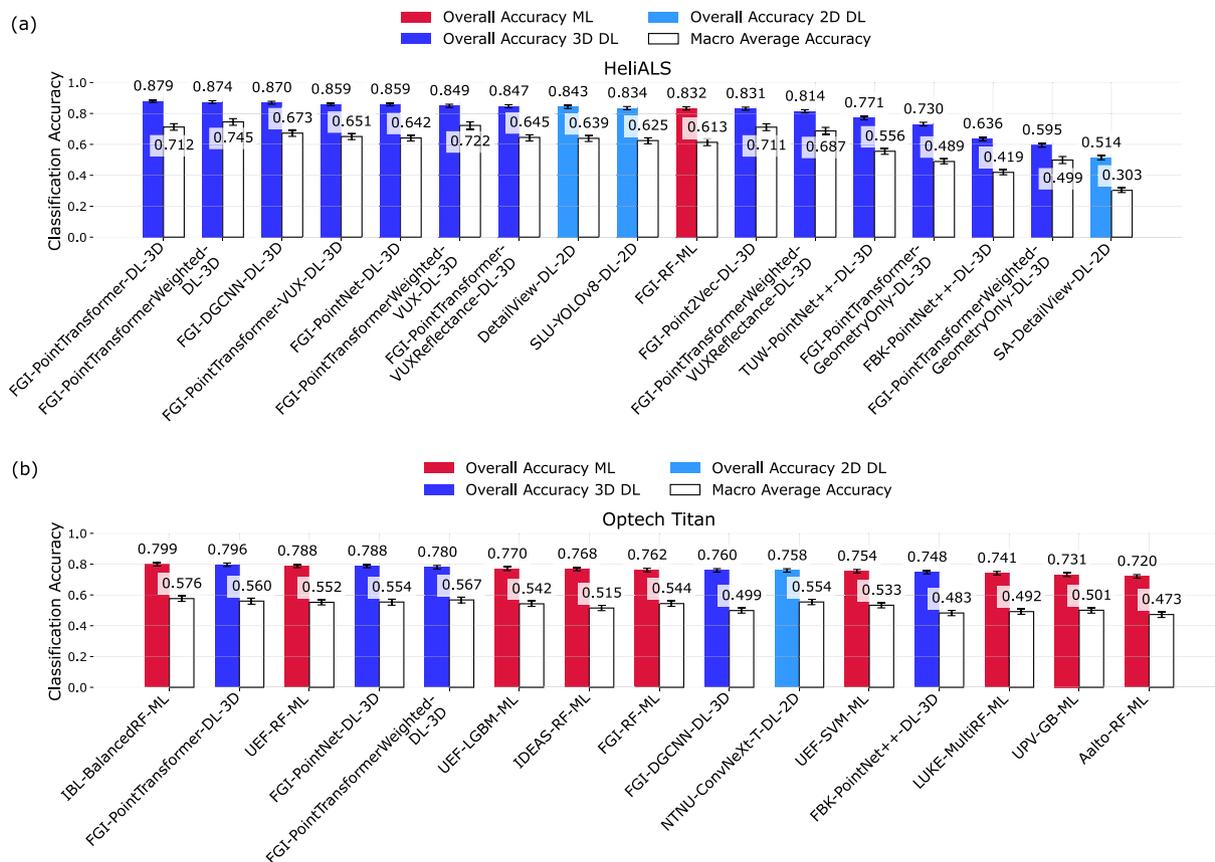
Fig. 6. (a) Comparison of overall (filled bar) and macro-average (outline) accuracy for methods using HeliALS data, with point-based DL methods shown with dark blue color, image-based DL methods shown with light blue color, and machine learning methods shown with red color. The error bars represent 95% confidence intervals based on a bootstrapping approach, see Section 3.3. SA-DetailView-DL-2D was not retrained on the training set of the current study, and the model parameters were directly transferred from Puliti et al. (2025). When neglecting aspens, rowans and alders not present in the FOR-species20K dataset (Puliti et al., 2025), SA-Detailview-DL-2D reached an overall accuracy of 58.8% and a macro-average accuracy of 45.4%. (b) Same as (a) but for the methods using the Optech Titan dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

function. FGI-PointTransformerWeighted-DL-3D achieved the highest macro-average accuracy even when taking the finite confidence intervals into account. In general, the total width of the 95% confidence intervals for the macro-average accuracy was 3.1–4.7% for the methods using the HeliALS data. Since the macro-average accuracy weighs each class equally, rare classes contribute a higher variance, leading to a wider confidence interval compared to overall accuracy.

For the Optech Titan dataset, the highest classification accuracy was reached by a random forest classifier, IBL-BalancedRF-ML, which achieved an overall accuracy of 79.9% and a macro-average accuracy of 57.6%. The point transformer method, FGI-PointTransformer-DL-3D, ranked second with an overall accuracy of 79.6% and a macro-average accuracy of 56.0%. For a larger training set size, the deep learning method is expected to outperform the machine learning model as discussed in Section 4.4. Based on the confidence intervals from bootstrapping, the overall accuracy of IBL-BalancedRF-ML was within the confidence intervals of FGI-PointTransformer-DL-3D, UEF-RF-ML, and FGI-PointNet-DL-3D, implying that the difference between these methods is not statistically significant. The total width of the 95% confidence intervals ranged between 2.1–2.4% for the overall accuracy and between 3.6–4.2% for the macro-average accuracy across the methods using the Optech Titan dataset.

On the Optech Titan dataset, most of the participating methods were RF classifiers or their variants. The best and worst-performing RF models had a difference of 7.9 (10.3) percentage points in terms of the overall (macro-average) accuracy implying that the performance

of the various RF approaches was relatively consistent. This suggests that RF is a robust and reliable method, providing reasonable accuracy across different implementations. Based on the submissions by UEF, a RF classifier also outperformed both LGBM and SVM classifiers, though the difference was only a few percentage points.

It is difficult to draw exhaustive conclusions on the factors causing the performance differences between the studied RF classifiers. However, the method descriptions in Section 3.2.3 suggest that the best-performing RF classifiers often utilized multi-channel intensity and echo features as well as various histogram features, such as intensity histograms. Interestingly, intensity features appear to be more important than geometric features since the second best RF method, UEF-RF-ML, only utilized a single geometric feature, but still managed to reach a good performance thanks to various intensity and echo features. When it comes to training strategies, the best performing model, IBL-BalancedRF-ML, cleaned the training dataset by removing under- or over-segmented trees, and applied a balanced RF procedure (Chen et al., 2004) to up-sample the number of training segments for minority species. However, LUKE-MultiRF-ML and Aalto-RF-ML also applied strategies to mitigate class imbalance but they were outperformed by several methods that did not use any strategy for class imbalance mitigation.

Some RF methods, such as FGI-RF-ML, had trouble evaluating all of their features on certain segments due to missing point attributes for some of the channels, thus hampering their prediction. On sparse tree segments, the computation of complicated geometric features may
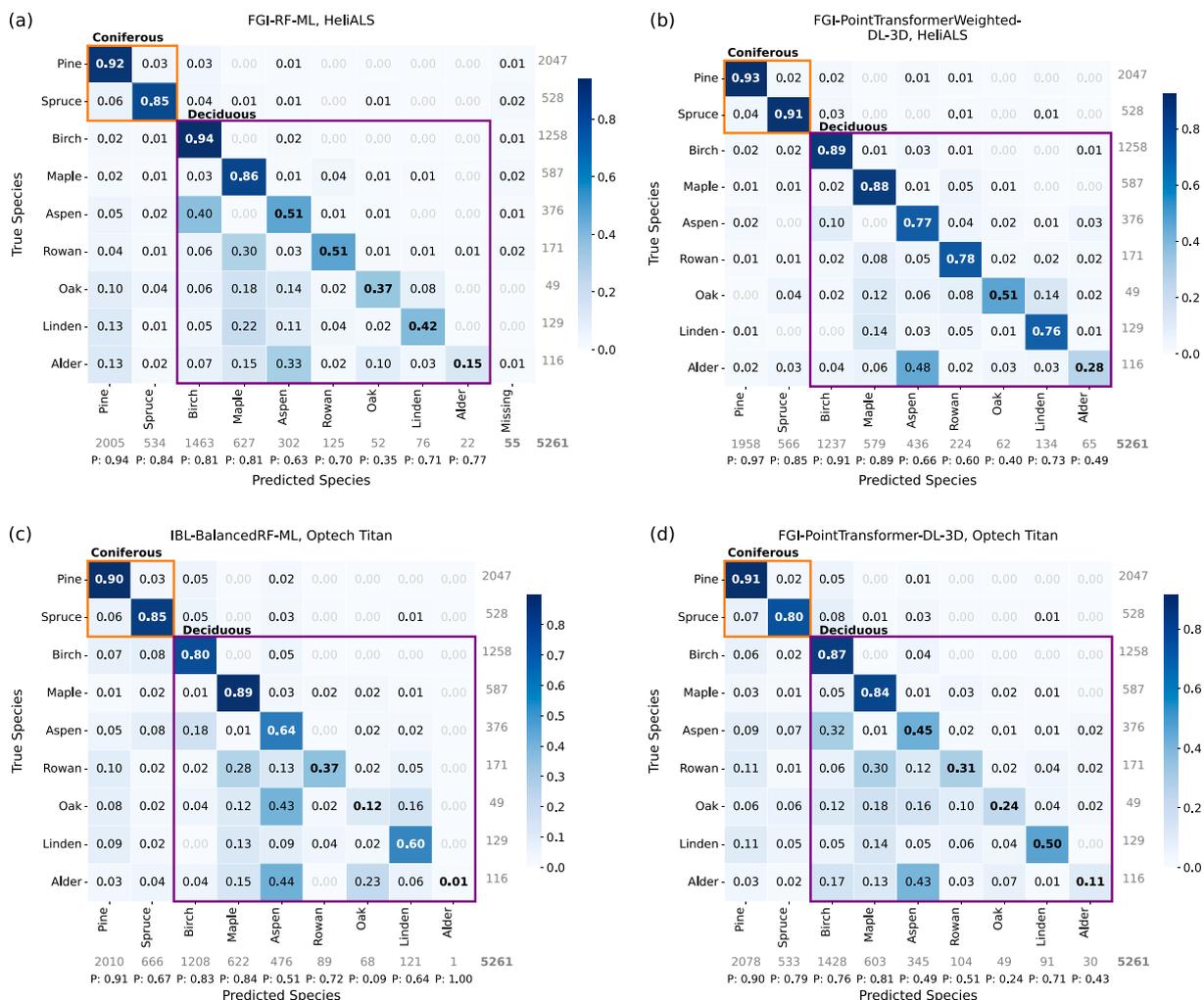
Fig. 7. (a) Normalized confusion matrix of FGI-RF-ML obtaining the highest macro F_1 score (0.64) among the machine learning models using HeliALS data. FGI-RF-ML produced 55 missing predictions. (b) Normalized confusion matrix of FGI-PointTransformerWeighted-DL-3D obtaining the highest macro F_1 score (0.73) among the deep learning models using HeliALS data. (c) Normalized confusion matrix of IBL-BalancedRF-ML obtaining the highest macro F_1 score (0.57) among the machine learning models using Optech Titan data. (d) Normalized confusion matrix of FGI-PointTransformer-DL-3D obtaining the highest macro F_1 score (0.58) among the deep learning models using Optech Titan data. All confusion matrices were row-normalized to enable comparison across species regardless of the class size. Each cell count was divided by the sum of its true-class row, and thus, each row sums to 1. Diagonal values represent species-wise recalls. Precisions are displayed below the matrix. Row and column sums are shown to the right and bottom of the matrix in gray.

also not always be possible, which may lead to further missing feature values. Thus, sufficiently simple geometric features should be preferred on sparse ALS data. Imputation strategies may also be useful if the point attributes are missing for some of the channels on certain tree segments due to differences in the sensing directions of the different channels.

The spread of the overall and macro-average accuracy metrics was lower on the Optech Titan dataset compared to HeliALS dataset. Perhaps surprisingly, the worst-performing methods using the Optech Titan data were better than the worst-performing methods using the HeliALS data, for which most of the submitted methods were deep learning models. We suspect that the larger spread in the results for the HeliALS dataset is partly caused by the deep learning models having a higher risk of overfitting unless necessary precautions are implemented, such as using a sufficient number of (randomized) augmentations or early stopping. Importantly, the best performing deep learning methods, such as FGI-PointTransformer-DL-3D, FGI-DGCNN-DL-3D, FGI-PointNet-DL-3D, and DetailView-DL-2D applied a large number of different augmentations during training, including point cloud random subsampling, rotations around the z-axis, small random scaling and jittering of coordinates and other features. The worst-performing

point-based deep learning method applied augmentations only for the minority species to mitigate class imbalance. Furthermore, the most successful methods typically applied some ensemble strategy at inference time, such as majority voting based on five separately trained classifiers (FGI-PointTransformer-DL-3D) or averaged predictions across a large number of rotational augmentations (DetailView-DL-2D). These observations suggest that deep learning methods have potential for high classification accuracy but care is needed in the choice of training and inference strategies.

Using the HeliALS data, we also tested a sensor-agnostic method, SA-Detailview-DL-2D. The method had been previously trained on the FOR-species20K dataset (Puliti et al., 2025), and we did not re-train it on the current dataset as explained in Section 3.2.2. The method did not utilize any spectral information either. On the full test set, SA-Detailview-DL-2D achieved an overall (macro-average) accuracy of 51.4% (30.3%) whereas the overall (macro-average) accuracy was 58.8% (45.4%) when neglecting aspen, rowan, and alder not present in the FOR-species20K dataset. However, the performance of the DetailView model was greatly improved by training the model from scratch using the training dataset of our study. This resulted in an
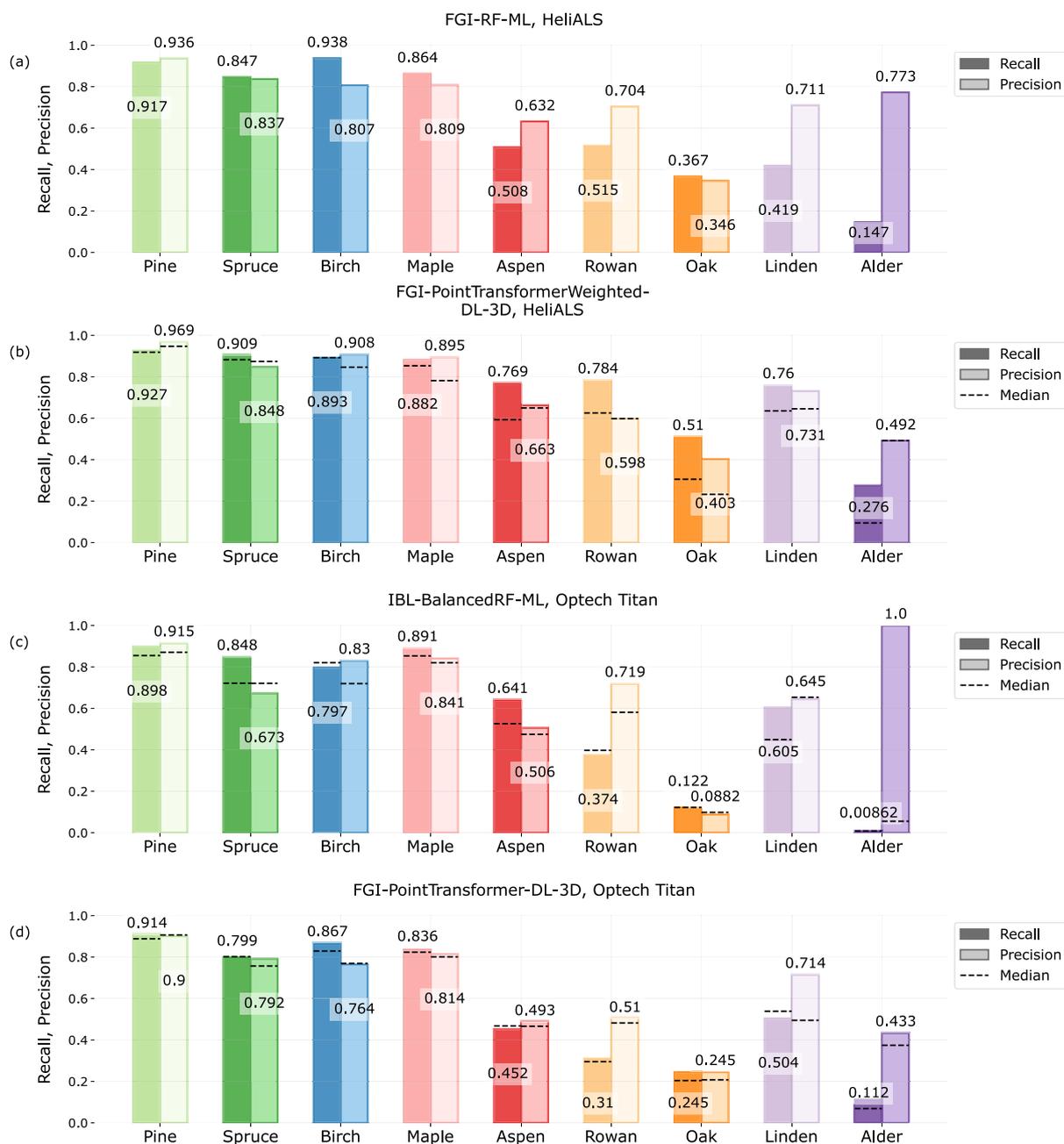
Fig. 8. Precision and recall for the best machine learning and deep learning methods for HeliALS (panels (a) and (b)) and Optech Titan (panels (c) and (d)) datasets. Best methods were chosen based on the macro-average F_1 score. Median values for precision and recall (dashed horizontal lines) were calculated across the methods utilizing the same dataset and approach (DL/ML). (a) Precision and recall for FGI-RF-ML utilizing HeliALS data (Macro $F_1 = 0.64$) (colored bars). (b) Precision and recall for the best DL method, FGI-PointTransformerWeighted-DL-3D, (Macro $F_1 = 0.73$) (colored bars) and the median across other DL methods (dashed lines) on the HeliALS dataset. (c) Precision and recall for the best ML method (Macro $F_1 = 0.57$) (colored bars) and the median across other ML methods (dashed lines) on the Optech Titan data. (d) Precision and recall for the best DL method (Macro $F_1 = 0.58$) (colored bars) and the median across other DL methods on the Optech Titan data (dashed lines).

overall (macro-average) accuracy of 84.3% (63.9%) for DetailView-DL-2D. The finding suggests that the model architecture itself is promising for species classification, but the model trained on the global FOR-species20K dataset did not generalize well enough for our ALS dataset. Possible reasons behind the poor out-of-domain performance of SA-Detailview-DL-2D are the relatively small proportion of high-density ALS data compared to TLS and MLS data in the FOR-species20K training dataset, the impact of segmentation artifacts, and the poor representativeness of the species of this study in the FOR-species20K training dataset.

4.2. Effect of multispectral information

To investigate the importance of spectral information on species classification, we compared variants of the best-performing Point Transformer model in scenarios, where no intensity information was available (FGI-PointTransformer-GeometryOnly-DL-3D), reflectance information from a single channel was available (FGI-PointTransformer-VUXReflectance-DL-3D), all intensity and echo information from a single channel was available (FGI-PointTransformer-VUX-DL-3D), and all multispectral information from the three channels was available (FGI-PointTransformer-DL-3D). Using geometric information only, we

obtained an overall (macro-average) accuracy of 73.0% (48.9%), which was further improved to 84.7% (64.5%) with a single-channel reflectance, to 85.9% (65.1%) with all single-channel information, and to 87.9% (71.2%) with spectral information from all the three channels. Thus, single-channel spectral information reduced the overall error by 48% and macro-average error by 32%. The multispectral information further decreased the overall error by 14% and macro-average error by 17% compared to using all single-channel spectral information. Importantly, the macro-average classification error could be reduced by a further 11% by switching to use a weighted loss function, which enabled a more accurate classification of the minority species without sacrificing the overall classification accuracy.

4.3. Species-wise analysis

To understand differences in the classification accuracy between different species, we present confusion matrices of the best-performing deep learning and machine learning models for both the HeliALS and Optech Titan datasets in Figs. 7(a)–(d). See Fig. D.1 in Appendix D for the corresponding unnormalized confusion matrices. Furthermore, Figs. 8(a)–(d) illustrate the species-wise precision and recall for the best methods on both datasets together with the median of the machine learning or deep learning approaches. Based on Figs. 7(a)–(d), we observe that pines, spruces, birches, and maples are reliably classified on both datasets across the models. The other deciduous trees representing minority species (see Fig. 3) are associated with significantly lower species-wise recall and precision. Based on the confusion matrices, we observe that aspens are often classified as birches, rowans as maples, lindens as maples, alders as aspens, and oaks as all other deciduous trees across the different models and datasets. The difficulty in predicting these species arises partly due to their low frequency in the benchmarking dataset and partly due to a rather similar visual appearance as further discussed in Section 4.6. For the Optech Titan dataset, we observe no clear qualitative difference between the predictions of the best deep learning and the best machine learning model.

The dense HeliALS dataset enables a more accurate classification of the minority species. Importantly, the best deep learning method, FGI-PointTransformerWeighted-DL-3D, clearly outperforms the best machine learning method FGI-RF-ML in the prediction of minority species on the HeliALS dataset, which highlights the promise of deep learning for species classification on dense ALS point clouds. FGI-PointTransformerWeighted-DL-3D achieved a recall of 76.9% and a precision of 66.3% for aspen that has an important role in the biodiversity of boreal forests as discussed in Section 1. Using Optech Titan data, IBL-BalancedRF-ML achieved a recall of 64.1% and a precision of 50.6% for aspen. See Appendix D for a full list of normalized confusion matrices for the methods not included in Figs. 7(a)–(d).

4.4. Scaling laws of classification error

A significant aspect of machine learning is the influence of training set size on the classification accuracy. Another major aspect is the influence of specific properties of the input data, such as point cloud density. In the following sections we provide detailed analysis of both aforementioned factors.

4.4.1. Scaling laws with respect to training set size

In this section, we study the effect of the training set size on the species classification accuracy using FGI-PointTransformer-DL-3D as a representative deep learning method and FGI-RF-ML as a representative machine learning method. See Section 3.4 for details on the implementation of the scaling analysis. We illustrate the overall and macro-average accuracy of the studied models as a function of the training set size for HeliALS and Optech Titan datasets in Figs. 9(a) and (c), respectively. For the HeliALS data, we observe that the deep learning model outperforms the machine learning model already at

the training set size of 250 segments. As the number of segments in the training set increases, the difference in the performance of the DL and ML models widens up. For the largest considered training set sizes, the macro-average accuracy of the deep learning method nearly catches up with the overall accuracy of the machine learning method. For a training set size of 5000 segments, FGI-PointTransformer-DL-3D reached an overall (macro-average) accuracy of 92.0% (85.1%), whereas FGI-RF-ML achieved an overall (macro-average) accuracy of 86.8% (69.6%).

For small training sets of the Optech Titan dataset, the ML method outperforms the DL method in terms of both overall and macro-average accuracy. Similarly to the case of the HeliALS data, the accuracy of the DL method improves more rapidly as the size of the training set is increased. Importantly, there is a crossover point at around 700 training segments, above which the overall and macro-average accuracy of the DL model outperform those of the ML model. For the largest considered training set size of 5000 segments, FGI-PointTransformer-DL-3D reached an overall (macro-average) accuracy of 83.7% (66.7%), whereas FGI-RF-ML achieved an overall (macro-average) accuracy of 79.5% (60.4%).

To gain more quantitative insight into the scaling behavior of species classification, we show the overall and macro-average classification error as a function of the training set size in a logarithmic scale in Figs. 9(b) and (d) for the HeliALS and Optech Titan data, respectively. Similarly to neural scaling laws for language and vision tasks (Hestness et al., 2017; Kaplan et al., 2020; Bahri et al., 2024), we observe a power law scaling of the classification error ϵ following the model

$$\epsilon(m) = Am^{-\alpha}, \tag{7}$$

where A is a multiplier, m denotes the training set size, and α describes the speed of convergence. We observe that the power law scaling applies both for the point transformer model representing deep learning methods and the RF classifier representing machine learning methods. For the HeliALS data in Fig. 9(b), we notice no clear deviation from the power law fit across the studied range of training set sizes spanning an order of magnitude. For the Optech Titan dataset, the macro-average classification error of the ML model appears to slowly deviate from the power law behavior at large training set sizes, though even larger training set sizes would be needed to verify this observation. Importantly, the classification error of the DL method converges approximately twice faster with increasing training set size compared to the ML method. For example on the HeliALS dataset, the DL method is characterized by $\alpha = 0.256$ ($\alpha = 0.353$) regarding the overall (macro-average) error, whereas the ML method reaches $\alpha = 0.14$ ($\alpha = 0.159$). Moreover, we observe that the DL method has a higher α and converges faster on the HeliALS dataset than on the Optech Titan dataset, implying that DL methods may provide a larger advantage on dense ALS point clouds with high geometric accuracy.

The power law behavior indicates that the classification accuracy may be further improved by increasing the training set size. According to our results, the benefits of deep learning methods become more apparent on large datasets consisting of thousands of trees, whereas machine learning methods may provide sufficient performance on small datasets of up to a few hundred trees. However, the improvements obtained by increasing the training set size obey a law of diminishing returns. Namely, the relative reduction of the error obtained by increasing the training set size from 1000 segments to 10000 segments is expected to be equivalent to the reduction obtained by increasing the training set size from 100 to 1000 segments. Extrapolating the fitted scaling laws on the HeliALS data, we estimate that reaching a macro-average accuracy above 90% across the nine species would require 14 000 training segments using the DL method and up to 4.9×10^6 segments using the ML approach, highlighting the advantages of deep learning on large datasets. For the Optech Titan data, we similarly estimate that 12×10^6 segments would be required to reach a macro-average accuracy above 90% with the deep learning method. However, these estimates
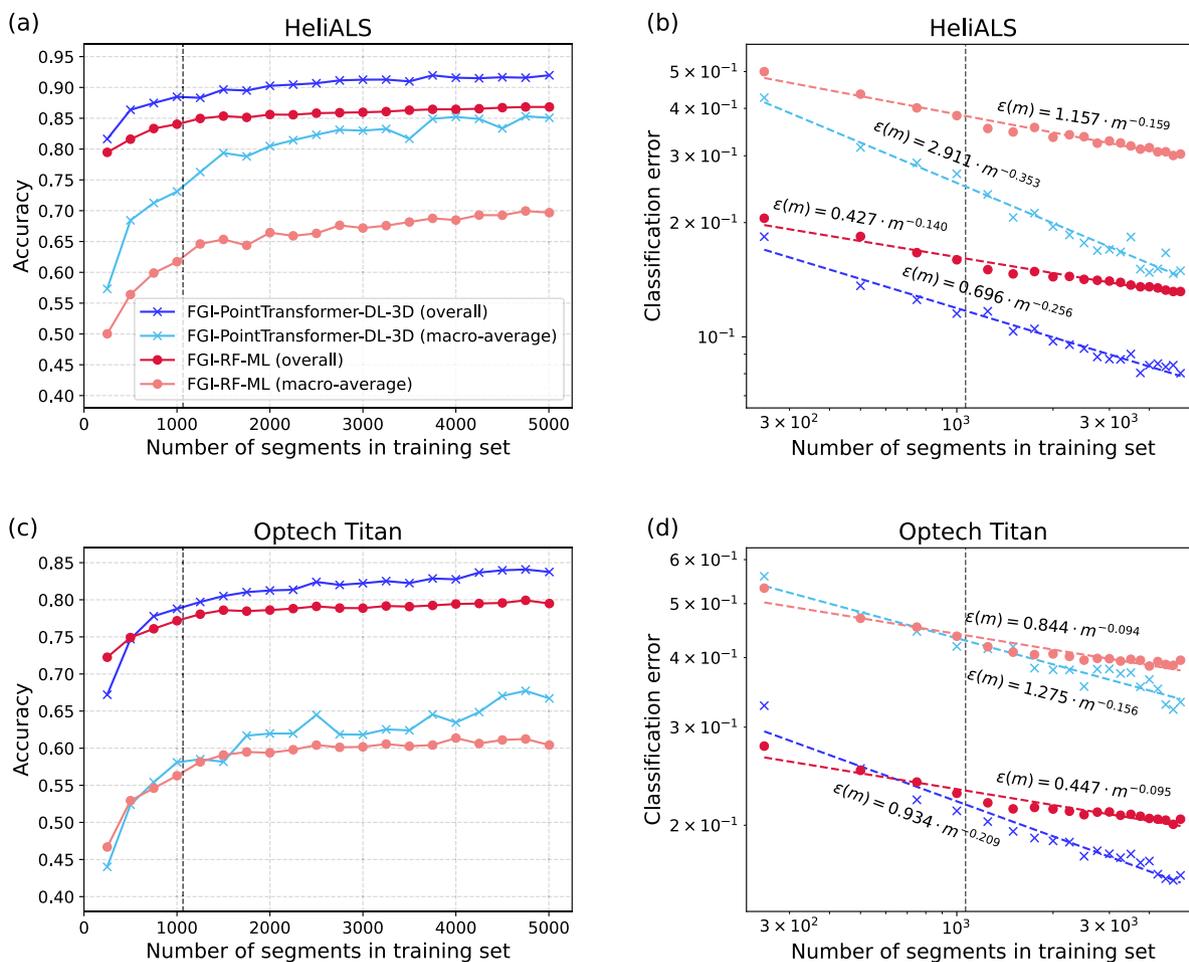
Fig. 9. (a) Overall accuracy (dark color) and macro-average accuracy (light color) of FGI-PointTransformer-DL-3D (blue) and FGI-RF-ML (red) as a function of the number of training segments for HeliALS data. (b) Classification error corresponding to (a) as a function of training set size in logarithmic scale, together with power law fits given by Eq. (7). (c) Same as (a) but using the Optech Titan dataset. (d) Same as (b) but using the Optech Titan dataset. In all panels, the gray vertical lines denote the size of the training set in the benchmark competition. For the DL method, the training set was further divided into training and validation sets for hyperparameter tuning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

should be regarded as rough order of magnitude estimates, and we note that the transferability of the deep learning methods should be studied further as discussed in Section 4.6. The good fit of the power law model also highlights the high quality of our field reference dataset, namely, any potential classification errors in our field reference dataset must be significantly rarer than the lowest achieved error of 8.0% since we observe no clear deviation from the power law scaling.

We present the species-wise classification error as a function of the number of training examples for the DL model in Fig. 10(a) and for the ML model in Fig. 10(b). Using the DL model, we observe a power law scaling of the classification error for each of the species. However, the minority species, such as oak, linden, and alder, have a significantly higher classification error compared to the majority species, such as pine, birch, spruce, and maple, for a given number of training segments. Pine trees are the easiest species for the classification, and a classification error of around 6% is reached with only 100 pine trees in the training set. Using the RF model, the classification error of the majority species appears to stagnate at large training set sizes, whereas the classification error of the minority species continues to improve with increased training set size, with the exception of oak that is hardly learned by the RF model.

4.4.2. Scaling laws with respect to point density

We also study the classification error as a function of point density by randomly sub-sampling the segments to a desired point density as

illustrated in Fig. 11(a) and explained in Section 3.4. In Figs. 11(b)–(d), we compare the classification error as a function of point density for FGI-PointTransformer-DL-3D representing a baseline DL model, FGI-PointTransformer-GeometryOnly-DL-3D using only geometric information, a variant of the baseline DL model using only 25% of the training set, and FGI-RF-ML representing a ML model. We observe that the baseline DL model exhibits two regimes of convergence. At low point densities below 10 pts/m^2, the error of the baseline DL model exhibits rapid convergence following a power law

$$\varepsilon(\sigma) = A\sigma^{-\alpha}, \qquad (8)$$

where σ denotes the point density per area. The overall (macro-average) classification error of the baseline DL model approaches 10.5% (22.7%) already at a point density of 10 pts/m^2. For point densities above 10 pts/m^2, both the overall and macro-average error continue to decrease with increasing point density but with a significantly lower rate. In this regime, the classification error potentially follows another power law, but with different parameters.

The geometry-only DL model exhibits 2–4 times higher overall and macro-average classification error across the studied point densities compared to the baseline DL model utilizing multi-spectral information as shown in Fig. 11(b). Interestingly, the geometry-only DL model shows a slower convergence of the classification error at low point densities compared to the baseline DL model, but the power law scaling
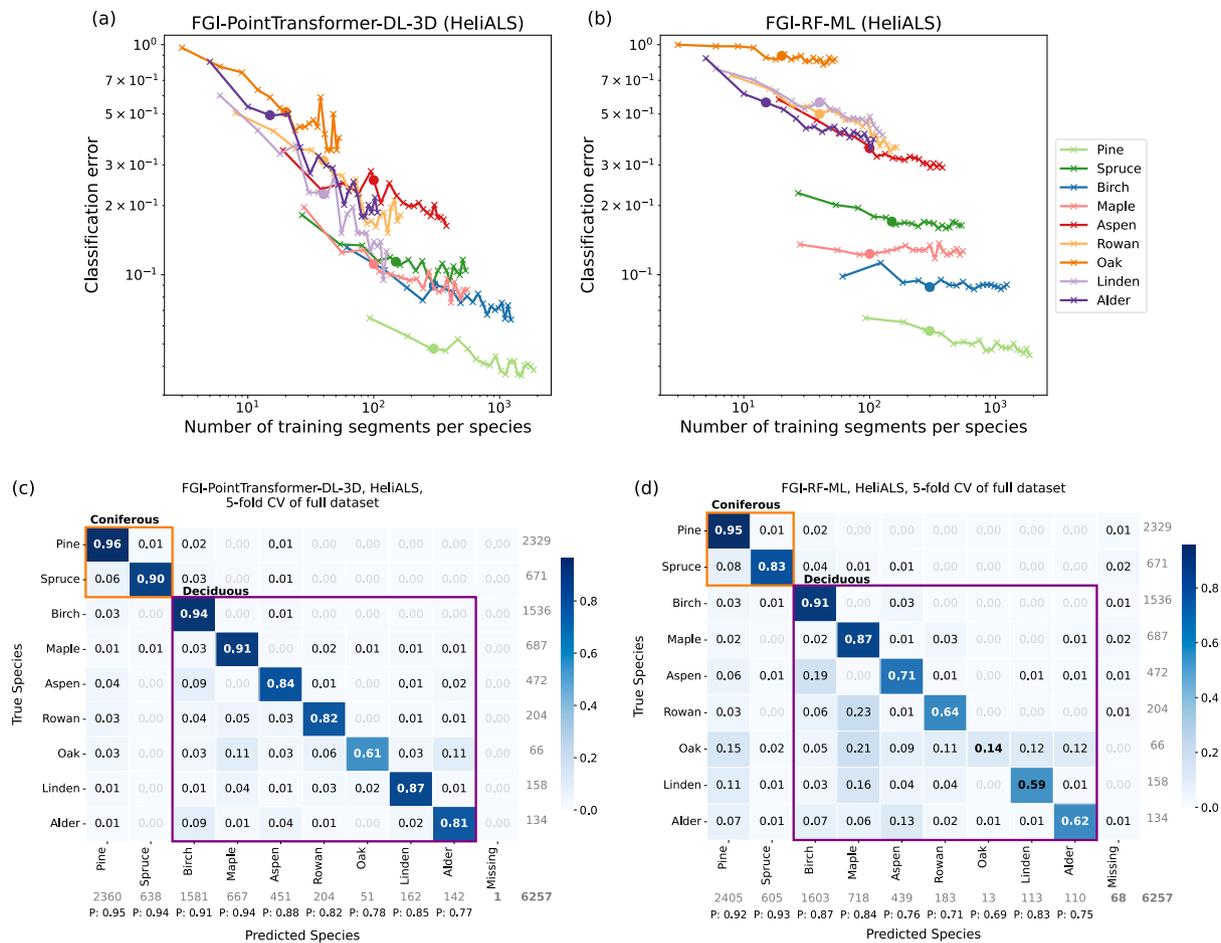
Fig. 10. Species-wise classification errors of (a) FGI-PointTransformer-DL-3D and (b) FGI-RF-ML as a function of the number of training segments in logarithmic scale for HeliALS data. The circle markers denote the number of training segments per species in the benchmark competition. For the DL method, the training set was further divided into training and validation sets for hyperparameter tuning. (c) and (d) show the row-normalized confusion matrices of FGI-PointTransformer-DL-3D (macro $F_1 = 0.86$) and FGI-RF-ML (macro $F_1 = 0.73$) for HeliALS data when using 5-fold cross-validation on the full ground-truth species dataset (5000 training segments).

extends across a larger range of point densities up to ≥ 100 pts/m². Thus, the geometry-only DL model begins to catch up with the baseline model for point densities above 10 pts/m². This also suggests that the improvement provided by the multi-spectral information is the largest for point densities around 10 pts/m², which highlights the advantage of multi-spectral information especially on sparse ALS datasets. For the largest considered point density of around 500 pts/m², the overall classification error of the geometry-only DL model still remains a factor of two higher than that of the baseline DL model using multispectral information. The observed results confirm similar findings from a multitude of applications relating to multispectral- and hyperspectral laser scanning (Takhtkeshha et al., 2024).

In Fig. 11(c), we compare the classification error of the baseline DL model when using the full training set and when using one fourth of the training set. We see that the classification error reduces more rapidly as a function of the point density when using the larger training set, especially in the regime of low point density. However, classification error begins to deviate from the power law scaling around 10 pts/m² for both training set sizes.

In Fig. 11(d), we compare the classification error of the DL model and the ML model as a function of the point density. For a low point density below 5 pts/m², the error of the ML model rapidly increases with decreasing point density. For point densities below 1 pts/m², some of the features used by the RF classifier can no longer be computed. In contrast, the DL model still reaches a decent performance in this regime of point densities. Above a point density of 10 pts/m², both the

ML model and DL model show a slow but steady improvement of the classification error with increasing point density.

4.5. Effect of segmentation quality on classification accuracy

In this section, we study the classification accuracy across different profile categories and tree crown classes introduced in Section 2.6, see also Fig. 5. We show the overall classification accuracy for the best deep learning method and the best machine learning method in the benchmarking study together with the corresponding median values across the five profile categories using the HeliALS data in Fig. 12(a) and using the Optech Titan data in Fig. 12(b). For both the HeliALS and Optech Titan datasets, the highest overall accuracy is reached for the categories 'Single tree' and 'Large tree with undergrowth', which cover a large majority of the dataset as shown in Fig. 5(b). This is expected since these profile categories represent ideal segments with no nearby growing trees of the same or different species. On both datasets, the worst classification accuracy is obtained for the profile category 'Many trees of many species' representing a segment, in which the largest tree is accompanied with one or multiple smaller trees of different species. Nevertheless, the best deep learning model using the HeliALS dataset is able to predict the species of the largest tree correctly in such segments with an overall (macro-average) accuracy of 75.5% (64.9%).

For the HeliALS dataset, we observe that the best DL method consistently outperforms the best ML method across the profile categories
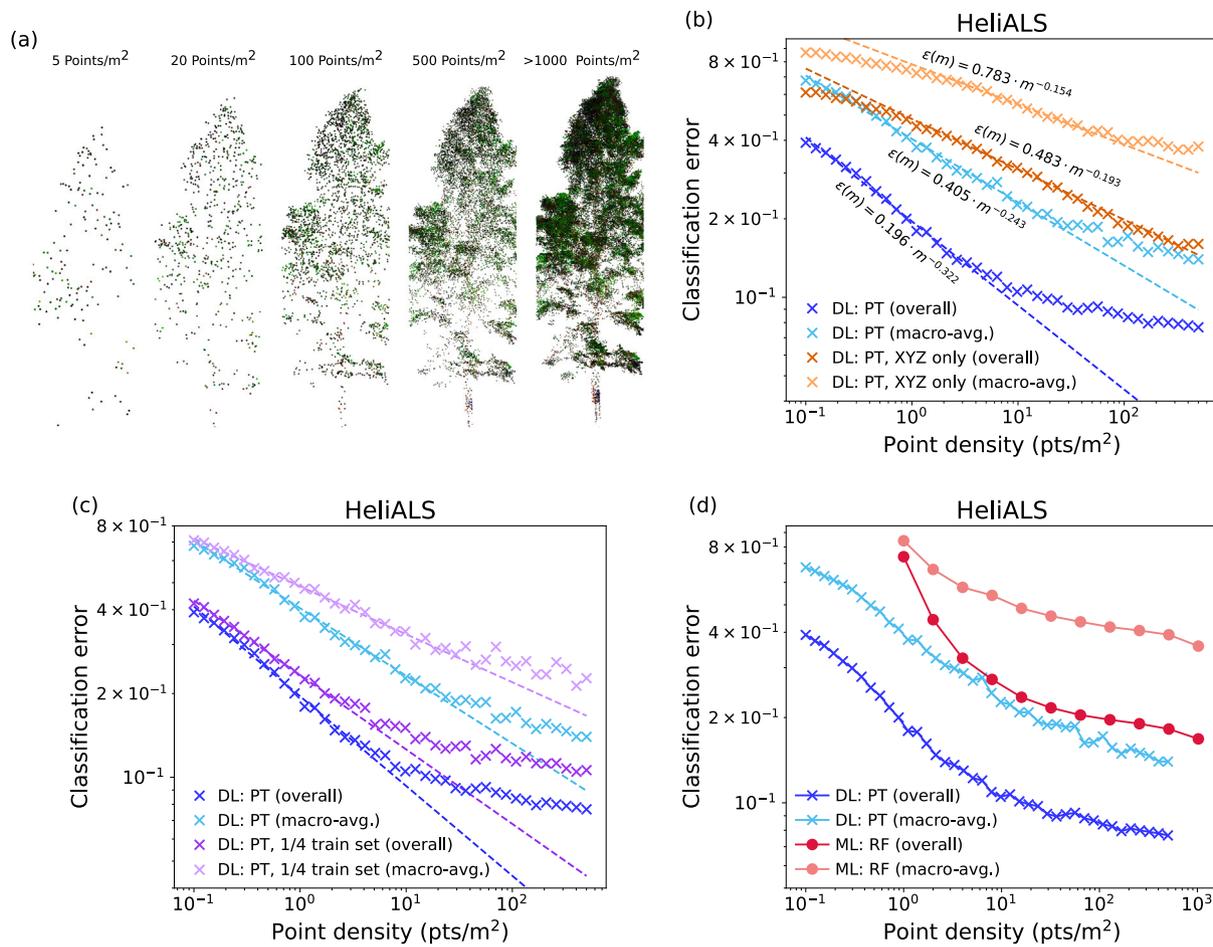
Fig. 11. (a) Point cloud projections of an example segment with varying point density obtained as a result of sub-sampling of the HeliALS point cloud. The red, green and blue color channels represent the laser pulse return intensities at the wavelengths $\lambda_1 = 532$ nm, $\lambda_2 = 905$ nm and $\lambda_3 = 1550$ nm, respectively. (b) Overall (dark color) and macro-average (light color) classification errors as a function of point density for FGI-PointTransformer-DL-3D (PT) when using all features (blue markers) and when using only geometric features (orange markers) on the HeliALS dataset. The dashed lines show fits of Eq. (8). (c) Overall (dark color) and macro-average (light color) classification errors as a function of point density for FGI-PointTransformer-DL-3D when trained with the full HeliALS training set (blue markers) and when trained with 25% of the training set (purple markers). (d) Overall (dark color) and macro-average (light color) classification errors as a function of point density for FGI-PointTransformer-DL-3D (blue markers) and FGI-RF-ML (RF, red markers) when trained with the full HeliALS training set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
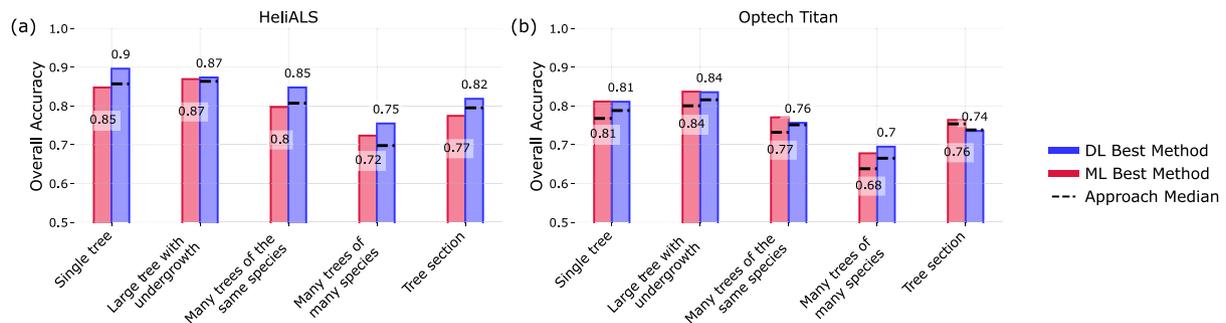
Fig. 12. Overall accuracy by profile category for the best machine learning and deep learning methods for (a) HeliALS and (b) Optech Titan datasets. Dashed lines indicate median values of overall accuracy across the methods utilizing the same dataset and approach type (DL/ML). The best method was determined by macro F_1 score, which is the unweighted average of the F_1 scores across all species. (a) For the HeliALS dataset, the best deep learning method is FGI-PointTransformerWeighted-DL-3D (Macro $F_1 = 0.73$) and the best machine learning method is FGI-RF-ML (Macro $F_1 = 0.64$). (b) For the Optech Titan dataset, the best deep learning method is FGI-PointTransformer-DL-3D (Macro $F_1 = 0.58$) and the best machine learning method is IBL-BalancedRF-ML (Macro $F_1 = 0.57$).

as shown in Fig. 12(a). When considering the macro-average accuracy (see Appendix C), the improvement provided by the DL method becomes more apparent especially for the most challenging category 'Many trees of many species'. For this category, the best DL method

reaches a macro-average accuracy of 64.9%, while the corresponding quantity for the best ML method is only 45.5%. For the Optech Titan dataset, we observe that the best DL and ML methods perform equivalently across the different profile categories as shown in Fig.
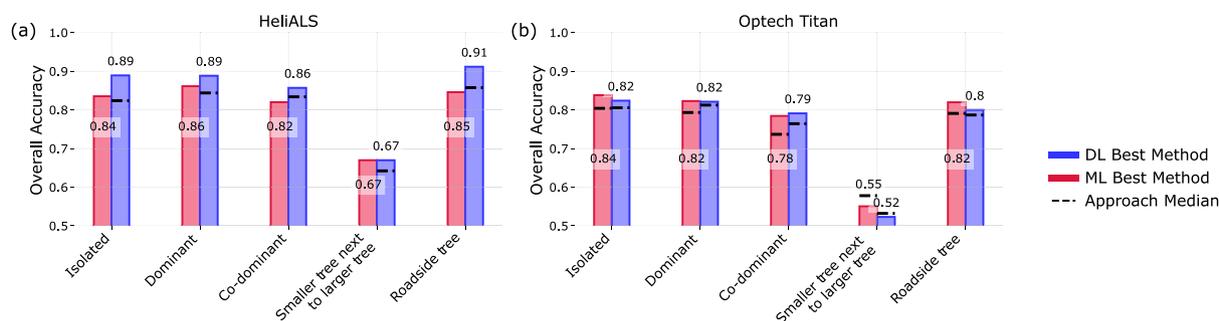
Fig. 13. Overall accuracy by crown class category for the best machine learning and deep learning methods for (a) HeliALS and (b) Optech Titan datasets. Dashed lines indicate median values of overall accuracy across the methods utilizing the same dataset and approach type (DL/ML). The best method was determined by macro F_1 score. (a) For the HeliALS dataset, the best deep learning method is FGI-PointTransformerWeighted-DL-3D (Macro $F_1 = 0.73$) and the best machine learning method is FGI-RF-ML (Macro $F_1 = 0.64$). (b) For the Optech Titan dataset, the best deep learning method is FGI-PointTransformer-DL-3D (Macro $F_1 = 0.58$) and the best machine learning method is IBL-BalancedRF-ML (Macro $F_1 = 0.57$).

12(b), though the DL model provides a small improvement for the category 'Many trees of many species', while the roles are reversed for the category 'Tree section'.

When it comes to the crown class categories, we observe that the overall classification accuracy is the highest and essentially equivalent for the classes 'Isolated', 'Dominant', and 'Roadside trees' across the different deep learning and machine learning models on both datasets as shown in Figs. 13(a) and (b). For the crown class 'Co-dominant', the overall accuracy is a few percentage points lower compared to the above mentioned crown classes. The classification accuracy is significantly lower for the category 'Smaller Tree next to larger tree'. For example, the best DL method on the HeliALS data reaches an overall accuracy of only 67% for this category, which is over 20 percentage points lower than the overall accuracy of around 90% for the crown classes 'Isolated', 'Dominant', and 'Roadside trees'. These observations are consistent with prior research (e.g., Hakula et al., 2023), where suppressed trees, characterized by their small size and proximity to larger neighbors, had an overall accuracy 20 percentage points lower than that of dominant trees.

To the best of our knowledge, the effect of segmentation quality on species classification has not been investigated in detail in prior literature since the segmentation has often been improved manually to ensure near-ideal tree segments, such as in Puliti et al. (2025). However, ALS-based tree species estimation across large areas will necessitate automatic segmentation, which will unavoidably lead to imperfect segmentation quality, especially for small or suppressed trees. In our study, the segmentation was based on a traditional watershed approach applied to the sparse Optech Titan dataset. In the future, segmentation methods based on deep learning, such as ForAINet (Xiang et al., 2024), SegmentAnyTree (Wielgosz et al., 2024), and TreeLearn (Henrich et al., 2024), may improve the segmentation quality especially on dense ALS data, where the high point density may contain cues for improved segmentation beyond traditional approaches. Further work should also study whether it is advisable to use the same segmentation method for both the training phase and the prediction phase in an operational setting as opposed to using a high-quality, labor-intensive manual segmentation for the training phase and an automatic segmentation for the prediction phase. According to our hypothesis, it is better to use the same segmentation method throughout the process to ensure that the training and prediction data come from the same distribution.

4.6. Further discussion

Our study compares favorably against recent UAV-based studies employing state-of-the-art laser scanning systems and high-quality passive hyperspectral imaging systems. For example, Quan et al. (2023) used a Random Forest classifier with leaf-on data acquired from a Riegl VUX-1UAV laser scanner and a MicroCASI 1920 hyperspectral imager to classify 11 species (2 coniferous, 9 broad-leaved), achieving a macro-average accuracy of 69.2% on a dataset with 904 trees using 90/10 train–test split. Zhong et al. (2022) employed a Support Vector Machine to classify 5 species (1 coniferous, 4 broad-leaved) using data from a RIEGL miniVUX-1UAV and a Resonon Pika L hyperspectral sensor. They achieved an overall accuracy of 89.2% and a macro-average accuracy of 84.5% on a dataset with 880 trees using a 60/40 train–test split. In our study, the best-performing model, FGI-PointTransformer-DL, reaches an overall accuracy of 92.0% and a macro-average accuracy of 85.1% across the nine studied species when using the full HeliALS dataset with 6257 segments and 80/20 train–test split, see Fig. 9(a). Thus, our best result surpasses the performance of these recent state-of-the-art studies on UAV-based species classification.

To gain further insights into the classification errors discussed in Section 4.3, we visualize the clustering of different tree species in the feature space in Figs. 14(a) and (b) using input features of FGI-RF-ML and the deep features of FGI-PointTransformer-DL-3D, respectively. To project the high-dimensional features into a two-dimensional plane, we use t-distributed stochastic neighbor embedding (t-SNE) (Hinton and Roweis, 2002) with perplexity values ranging from 15 to 30 to have a clear separation of the main clusters. In Fig. 14(a) showing the clusters for the RF features, we observe that pines, spruces, birches and maples mostly form separate clusters in line with the high classification accuracy of these species. Visually, these species have clearly distinct characteristics, and they are easy to recognize in the field based on their shapes and needles or leaves, see Fig. 3. However, we note that a small set of spruces is located in the middle of the feature space between pine and alder clusters, separate from other spruces. A closer inspection of individual trees in this cluster showed that these spruces were typically growing on yards, roadsides or other places in the built-up environment, and they had a narrow shape different from large spruces growing in a forest.

For the remaining five species, there is significant inter-species overlap in the feature space. Clusters for aspen and alder are located next to each other with a partial overlap, which seems natural due to the similar appearance of these species. Furthermore, aspens also have overlap with birches in the feature space. Rowans, oaks and lindens are overlapping and close to other deciduous species, especially maples. This might be related to similarities in the size and shape of these species. Naturally, these observations are well-aligned with our remarks based on the confusion matrices presented in Section 4.3. The separation of deciduous species in the feature space and hence their classification accuracy may be affected by the season of data acquisition. For example, Lisiewicz et al. (2025) observed an increase in the classification accuracy of oak and aspen using ALS data collected during leaf-off season, whereas maple and linden were more accurately classified using leaf-on ALS data. A combination of features derived from leaf-off and leaf-on data has been shown to improve the species
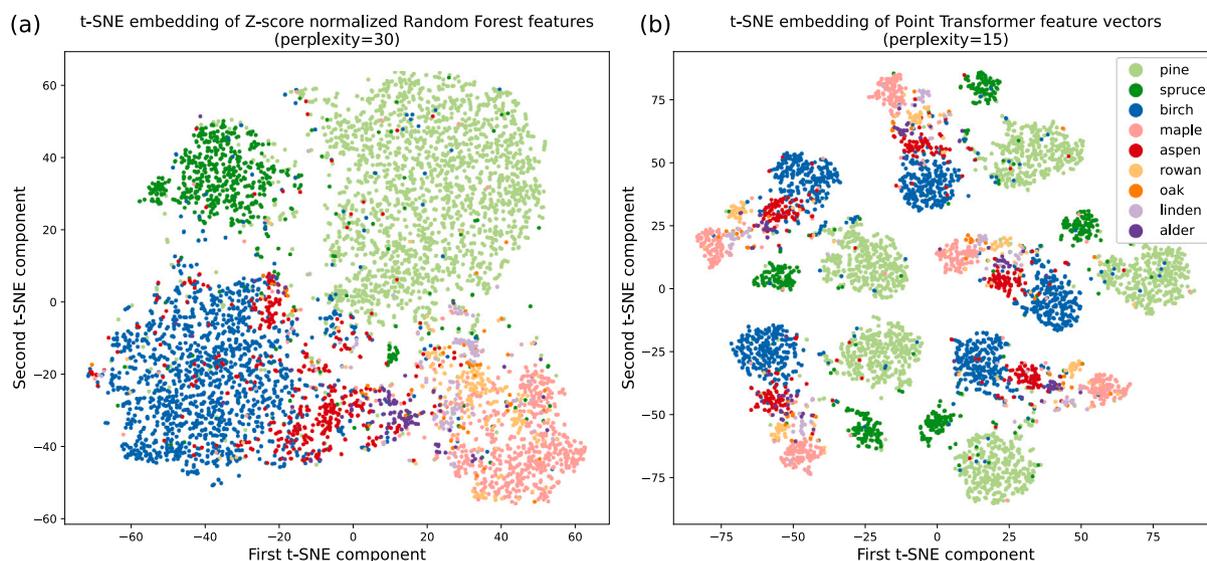
Fig. 14. A comparison of t-distributed stochastic neighbor embedding (t-SNE) visualizations for Z-score normalized FGI-RF-ML input features $\mathbf{f}_{RF} \in \mathbb{R}^{30}$ in (a) and FGI-PointTransformer-DL-3D deep features $\mathbf{f}_{PT} \in \mathbb{R}^{512}$ in (b). The data samples were obtained from the full reference dataset, and the colors represent the species in the field reference.

classification accuracy in several studies (Kim et al., 2009; Shi et al., 2018b; Kamińska et al., 2021; Lisiewicz et al., 2025).

In the feature space of the point transformer model shown in Fig. 14(b), each species is separated into five clusters for a reason that is not well understood. Generally, the separability of the species seems slightly better in the feature space of the point transformer model than in the RF feature space as expected from the higher classification accuracy. However, the relations between nearby species in the feature space of the point transformer model are overall similar to those of the RF model. Importantly, we believe that feature space visualizations may also be useful for the verification of large reference datasets in future studies on species classification. Namely, reference trees located within a wrong cluster may potentially correspond to a mistake in the reference dataset due to, e.g., a positioning error during field measurements. Thus, such trees may require a further verification in the field to ensure the correctness of their species.

As briefly discussed in Section 4.1, DetailView-DL-2D trained on our dataset without any spectral information achieved a comparable classification accuracy as the Point Transformer model augmented with single-channel (VUX) reflectance. Hence, further studies should investigate incorporating single-channel or multi-channel spectral information in the DetailView model, which holds potential to improve the classification accuracy especially for the minority species. Importantly, retraining of the DetailView model was necessary to achieve competitive performance on the ALS data of the current study. This conclusion seems to be valid even when accounting for the different species composition between our study and that of Puliti et al. (2025), which we tested by re-training the classification head while using frozen encoder parameters of the sensor-agnostic model from Puliti et al. (2025). Such an approach resulted in an overall (macro-average) accuracy of 78.7% (56.7%), which is still significantly lower compared to the overall (macro-average) accuracy of 84.3% (63.9%) for the fully retrained model.

Based on our studies with varying point densities in Section 4.4.2, the influence of single- and multi-channel spectral information is most pronounced at low point densities of around 10 pts/m^2. However, we still observe a significant improvement in the classification accuracy when using the HeliALS data (>1000 pts/m^2) instead of the Optech Titan data (35 pts/m^2), especially when it comes to the macro-average accuracy that is 14 percentage points higher for the best-performing model of the HeliALS data. Our studies with a varying point density

rely on artificial subsampling of the HeliALS point cloud. In practical implementations, the reduced point density is often a byproduct of a higher flight altitude, which may reduce the geometric accuracy of the point cloud due to an increased beam width. Thus, our results in Fig. 11 may not perfectly describe a scenario, in which the flight altitude controls the point density. It is also important to note the disparity in the availability of point cloud data with densities similar to HeliALS or Optech Titan systems. The Optech Titan data exhibits a density comparable to that used in municipal mapping applications in Finland, typically ranging from 40 to 60 points/m^2. Conversely, the HeliALS data density aligns more closely with drone-based LiDAR surveys, indicating its suitability for smaller, more localized studies, e.g. for forest holding.

In principle, our results demonstrate precision and recall values suitable for operational mapping of aspen populations at the individual tree level, reaching $F_1 = 0.71$ for the best performing model on the HeliALS data and $F_1 = 0.57$ for the best-performing model on the Optech Titan data. However, the proportion of aspen trees was higher on our study area compared to typical boreal forests, and therefore, a lower F_1 score may be achieved in operational use for an equivalent training set size. For example, Toivonen et al. (2024) reached an F_1 score of 0.44 for mapping large aspens using ALS (7.2 pts/m^2) and aerial imagery on sample plots with an aspen proportion of around 6%, whereas the F_1 score reduced to 0.21 on sample plots with an aspen proportion of around 0.1%. Nevertheless, our results suggest that a higher point density and multispectral information improve the classification accuracy of aspen.

Importantly, the point density of nation-wide ALS surveys is approaching that of Optech Titan data, while the ALS surveys of several cities and municipalities already utilize denser point clouds than those collected with Optech Titan in this study. In most of the current nation- and city-level ALS surveys, spectral information is available only at a single wavelength. When it comes to species classification across large areas, such as entire countries, we also expect that the models using single-channel or multi-channel spectral information may not generalize as well to new scanners compared to models relying only on the point cloud geometry. Furthermore, multispectral ALS data may suffer from a reduced point density or a complete lack of points for certain channels across some parts of the mapped area due to, e.g., different fields of view of the scanners, which can be mitigated by a careful planning of the flight path. In the HeliALS dataset, a

small part of the tree segments ($\sim$ 1%) suffered from a complete lack of points from miniVUX or VQ840G scanners, which led some methods to develop imputation strategies for assigning the radiometric information to each point. Thus, the intensity information offers great advantages for improving the accuracy of species classification, but care is needed to ensure data coverage of all channels and consistent intensity calibration across large areas.

When it comes to mapping large areas, the computational cost of species classification becomes also a relevant performance metric even though species classification for forest inventory is typically a post-processing activity without strict runtime requirements. To assess the scalability of the studied classification methods, we compare the run time of model training and inference in Appendix B for three selected methods including FGI-RF-ML representing a shallow machine learning method, FGI-PointTransformer-DL-3D representing a 3D deep learning method, and DetailView-DL-2D representing a 2D deep learning method. We found that the training time of the RF model was the shortest, but interestingly, the inference time of FGI-PointTransformer-DL-3D was 177 times shorter than that of the RF model and 52 times shorter than that of the DetailView on the HeliALS data. For the RF model, the feature computation dominated the total runtime, whereas the training and inference times after the feature computation were quick. For more details, see Appendix B and Fig. B.1.

Multispectral data also opens up new possibilities for the use of echo classes in species classification. After the first echo, the LiDAR signal is affected by a variety of factors, such as transmission, absorption losses, and multiple scattering, which alter the measured return intensity. The return intensities in such multiple scattering events are affected differently for each wavelength, and the effect may vary by species due to different canopy geometries, thus providing additional information for species classification. For example, features related to intermediate echos were found to contribute to the classification performance of UEF-RF-ML, though features related to the first echoes were the most important ones as they correspond to the purest return intensities. Features related to last echos provided no added value to the classifier.

More studies are also needed to investigate the transferability of deep learning methods to new datasets that may also cover large areas. In Section 4.4.1, we observed that the classification error followed a power law scaling as a function of training set size, which, in principle, enables us to estimate the required number of training segments for a desired classification accuracy when using a given model. However, we also noted that the DetailView model performed significantly worse in this study when trained on the FOR-species20K dataset compared to the training set of the current study. Thus, it remains to be verified whether a similar power law scaling of the classification error would be observed if the models trained on the current dataset were transferred to new areas, though in the same forest zone. The gap between deep learning and machine learning models may narrow down if the faster convergence of deep learning models is partially due to overfitting to the current dataset.

We also point out that it is challenging to collect accurate reference data of individual trees due to limitations in GNSS signal reception under dense canopies and the difficulty of visually correlating ground observations with a map representation of the area. Furthermore, limited coverage of trees in open tree databases of cities necessitates considerable amount of field work and verification in detailed studies. Therefore, the crowdsourcing tool introduced in Section 2.5 was developed to this need. It allowed flexible collection of reference data by a large group of people. Improvements were made to the tool based on the first experiences in the field work, making the reference data collection easier and more reliable as the study progressed. As a result of these improvements, the tool is easy to use. Nevertheless, the persons collecting the reference data still need some skills and practice, e.g. to recognize the tree species and to understand the correspondence between one's position in the terrain and the map in complicated environment with several trees around and some uncertainty in GNSS

positioning. Furthermore, one must understand the contents and nature of spatial datasets visible in the tool including the height of trees, segments that can include several trees or part of a tree, and possible differences between the data and the current situation due to cut and newly planted trees. Due to challenges related to the reference data collection, a separate verification phase was implemented to improve the quality of the reference data.

5. Conclusion

This research represents a thorough benchmark of machine and deep learning methodologies for tree species classification, leveraging multispectral ALS data. By employing both high-density HeliALS data (> 1000 pts/m^2) and lower-density Optech Titan data (35 pts/m^2), we have provided a comprehensive comparison of various algorithmic approaches. For the benchmark study, we established a training set of 1065 tree segments and a test set of 5261 segments representing nine species in southern Finland. In our benchmark, point-based deep learning methods outperformed both image-based deep learning and machine learning methods on the high-density HeliALS data, while the best point-based deep learning methods performed equivalently with machine learning methods on the lower-density Optech Titan data. On the HeliALS data, a Point Transformer model topped the list with an overall (macro-average) accuracy of 87.9% (74.5%). On the Optech Titan dataset, a random forest classifier reached the highest overall (macro-average) accuracy of 79.9% (57.6%), though three other methods were within error bars in terms of the overall accuracy.

Additionally, we observed a remarkable improvement in the classification accuracy using single-channel and multi-channel intensity features. On the HeliALS data, the overall (macro-average) classification error of the Point Transformer model was reduced by 48% (32%) when incorporating single-channel spectral information compared to using no spectral information. By incorporating multispectral information of three channels, the overall (macro-average) classification error was further reduced by 14% (17%). As explained in Section 4.4.2, the multi-channel spectral features appear to have the largest impact at low point densities (~ 10 pts/m^2) according to our studies with varying point densities obtained by subsampling the HeliALS data. At higher point densities, it seems that deep learning approaches can effectively leverage the richer structural information, thus partially compensating for a lack of intensity data. With high-density multispectral data, deep learning approaches also substantially improve the macro-average classification accuracy, i.e. the detection of more rare species. For example, the macro-average accuracy of the Point Transformer model is 13 percentage points higher than that of a random forest classifier on the HeliALS dataset used for the benchmarking competition.

Furthermore, we show that the accuracy improvement provided by deep learning methods increases with training set size as discussed in Section 4.4.1. We observe that the classification error obeys a power law $\varepsilon(m) = Am^{-\alpha}$ with respect to the training set size both for a point transformer model and for a random forest classifier. Importantly, the scaling exponent α is approximately twice higher for the deep learning model leading to a drastically faster convergence of the classification error. Based on our analysis, the studied deep learning method outperforms the random forest classifier for training set sizes exceeding a few hundred segments on both datasets. Using the fitted scaling laws on the HeliALS data, we predict that a macro-average accuracy of 90% would require 14 000 training segments with the point transformer model and 4.9×10^6 training segments with the random forest classifier, thus highlighting the benefit of deep learning approaches on large datasets. The development of the crowdsourced field reference dataset, along with the innovative browser-based crowdsourcing tool, represent a significant contribution to the scientific community aiding the collection of large datasets required to scale up these methods to larger forest areas. These resources will facilitate future research and development in this domain.

The ability to accurately identify tree species at the individual tree level is essential for sustainable municipality and multifunctional forest management practices. Future research should focus on further refining deep learning architectures for multispectral ALS data, and addressing the challenges associated with identifying rare species. More detailed studies are also needed on the effects of different preprocessing steps and data augmentation methods on the performance of deep learning methods. Integration of tree species classification to direct measurements of stem curve and volume using a high-resolution airborne laser scanning system, as depicted in Hyyppä et al. (2022), would be further step in developing a fully automatic reference collection technique. Additionally, investigating the transferability of these models across different geographic regions and sensor platforms is crucial for broader adoption. From the practical point of view, it is also important to pay attention to datasets that are available nationwide or over large areas.

CRediT authorship contribution statement

Josef Taher: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Eric Hyyppä:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matti Hyyppä:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Klaara Salolahti:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xiaowei Yu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Leena Matikainen:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antero Kukko:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization. **Matti Lehtomäki:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Harri Kaartinen:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization. **Sopitta Thurachen:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paula Litkey:** Writing – review & editing, Software, Resources, Investigation, Data curation, Conceptualization. **Ville Luoma:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Markus Holopainen:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Gefei Kong:** Writing – review & editing, Software, Methodology, Investigation. **Hongchao Fan:** Writing – review & editing, Software, Methodology, Investigation. **Petri Rönnholm:** Writing – review & editing, Software, Methodology, Investigation. **Matti Vaaja:** Writing – review & editing, Software, Methodology, Investigation. **Antti Polvivaara:** Writing – review & editing, Software, Methodology, Investigation. **Samuli Junttila:** Writing – review & editing, Software, Methodology, Investigation. **Mikko Vastaranta:** Writing – review & editing, Software, Methodology, Investigation. **Stefano Puliti:** Writing – review & editing, Software, Methodology, Investigation. **Rasmus Astrup:** Writing – review & editing, Software, Methodology, Investigation. **Joel Kostensalo:** Writing – review & editing, Software, Methodology, Investigation. **Mari Myllymäki:** Writing – review & editing, Software, Methodology, Investigation. **Maksymilian Kulicki:** Writing – review & editing, Software, Methodology, Investigation. **Krzysztof Stereńczak:** Writing – review & editing, Software, Methodology, Investigation. **Raul de Paula Pires:** Writing – review & editing, Software, Methodology, Investigation. **Ruben Valbuena:** Writing – review & editing, Software, Methodology, Investigation. **Juan Pedro Carbonell-Rivera:** Writing – review & editing, Software, Methodology, Investigation. **Jesús Torralba:** Writing – review & editing, Software, Methodology, Investigation. **Yi-Chen Chen:** Writing – review & editing, Software, Methodology, Investigation. **Lukas Winiwarter:** Writing – review & editing, Software, Methodology, Investigation. **Markus Hollaus:** Writing – review & editing, Software, Methodology, Investigation. **Gottfried Mandlburger:** Writing – review & editing, Software, Methodology, Investigation. **Narges Takhtkeshha:** Writing – review & editing, Software, Methodology, Investigation. **Fabio Remondino:** Writing – review & editing, Software, Methodology, Investigation. **Maciej Lisiewicz:** Writing – review & editing, Software, Methodology, Investigation. **Bartłomiej Kraszewski:** Writing – review & editing, Software, Methodology, Investigation. **Xinlian Liang:** Writing – review & editing, Software, Methodology, Investigation. **Jianchang Chen:** Writing – review & editing, Software, Methodology, Investigation. **Eero Ahokas:** Writing – review & editing, Resources, Data curation. **Kirsi Karila:** Writing – review & editing, Resources, Data curation. **Eugeniu Vezeteu:** Writing – review & editing, Resources, Data curation. **Petri Manninen:** Writing – review & editing, Resources, Data curation. **Roope Näsi:** Writing – review & editing, Resources, Data curation. **Heikki Hyyti:** Writing – review & editing, Resources, Data curation. **Siiri Pyykkönen:** Writing – review & editing, Resources, Data curation. **Peilun Hu:** Writing – review & editing, Resources, Data curation. **Juha Hyyppä:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing 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.

Acknowledgments

Fig. A.1. Distribution of mean intensity for the three channels of the HeliALS system for the studied nine species across all segments of our dataset. Here, blue corresponds to the VQ-840-G scanner, green to the miniVUX-1DL scanner, and red to the VUX-1HA scanner. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
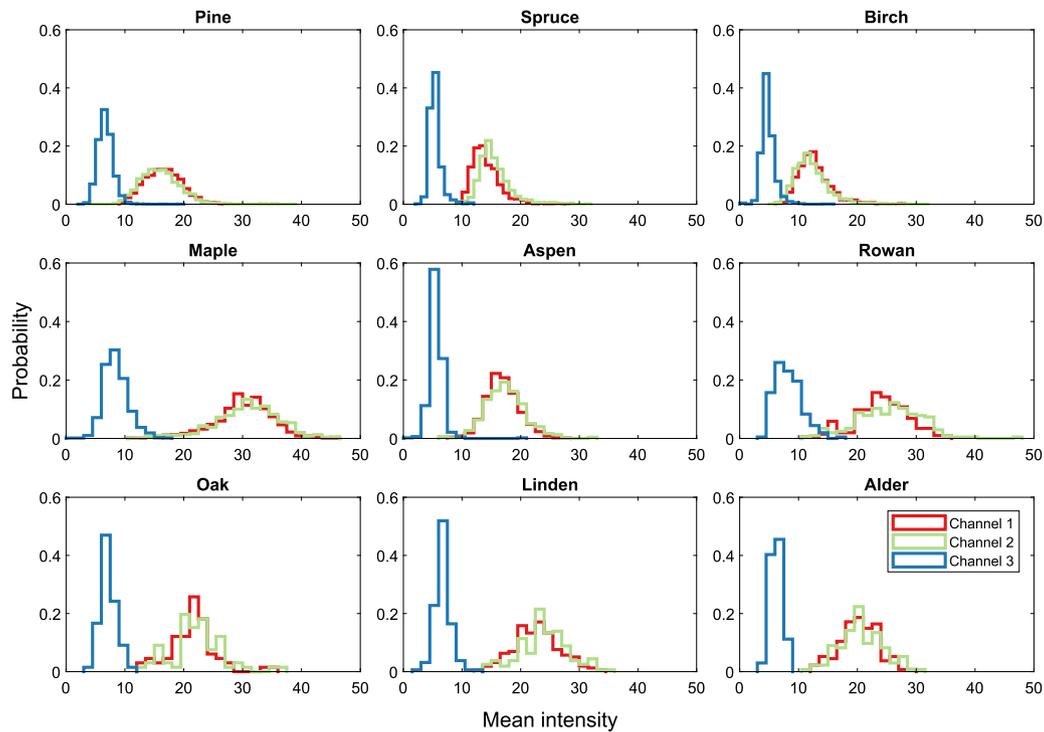
Fig. A.2. Distribution of mean intensity for the three channels of the Optech Titan system for the studied nine species across all segments of our dataset. Here, blue corresponds to $\lambda = 532$ nm, green to $\lambda = 1064$ nm, and red to $\lambda = 1550$ nm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
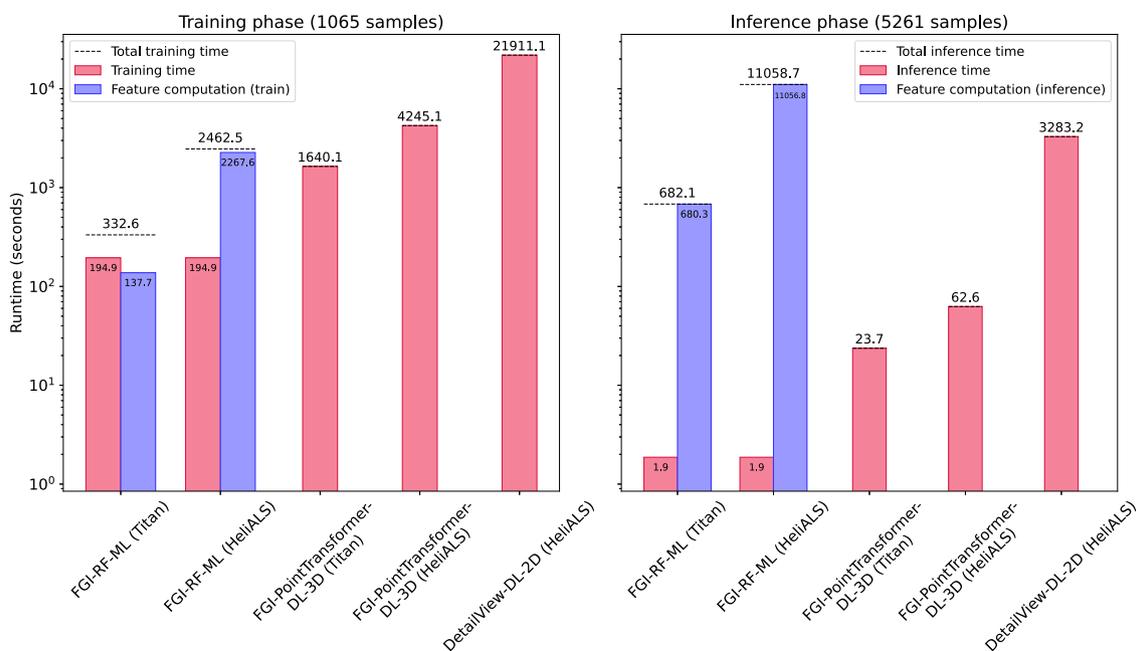
Fig. B.1. Comparison of running times between three tree species classification methods in different datasets. The training set size was 1065 tree segments and the test set size was 5261 tree segments. The dashed horizontal line represents the total running time including feature computation and classification. For deep learning based methods, the feature computation is inherent to the computations and has not been separately reported.

Table B.1
Hardware specifications of the two computers used for runtime measurements.

Specification	Computer 1	Computer 2
CPU	Intel Core i7-9850H (6 cores, 2.6 GHz)	Intel Xeon(R) Gold 6244 (16 cores, 3.6 GHz)
RAM	32 GB	512 GB
GPU	not utilized	NVIDIA RTX A6000 (48 GB)
Operating System	Windows 10	Ubuntu 22.04.5 LTS

Appendix A. Distribution of mean intensity for the studied tree species

Multispectral data provides a benefit for tree species classification since different tree species have a different reflectance response as a function of wavelength (Hovi et al., 2017). Figs. A.1 and A.2 illustrate the species-wise distribution of mean intensity for the three channels of the HeliALS system and the Optech system, respectively. For example, Fig. A.1 shows that the mean intensity of the 3rd channel ($\lambda = 532$ nm) of the HeliALS data differs significantly between pine and birch, while the difference of mean intensities is smaller for these two species at the other two wavelengths. Overall, the variations in the intensity distributions between different species showcase the advantage of multispectral ALS for tree species classification.

Appendix B. Comparison of algorithm running times

The computational cost of the FGI-RF-ML, the FGI-PointTransformer-DL-3D and the DetailView-DL-2D methods was assessed to determine their scalability to larger datasets. The comparison was performed separately for model training phases and inference phases. The training set contained 1065 samples, while the test set comprised 5261 samples. In addition, the running times of the FGI-RF-ML and the FGI-PointTransformer-DL-3D were reported both for the Optech Titan dataset and for the HeliALS dataset. The hardware specifications of the experiments are listed in Table B.1. The FGI-RF-ML method was run on computer 1, and the FGI-PointTransformer-DL-3D and the DetailView-DL-2D methods on computer 2.

The results of the running time comparison have been visualized in Fig. B.1. In the training phase, the computational cost of the FGI-RF-ML and the FGI-PointTransformer-DL-3D methods are quite similar. However, the DetailView-DL-2D method has a substantially longer training time than either of the previous methods. The FGI-RF-ML method is the fastest to train and can be trained in less than six minutes on our Optech Titan dataset.

During inference, the FGI-PointTransformer-DL-3D method performs the best in terms of running time. When using the HeliALS dataset, the FGI-RF-ML method is approximately 177 times slower than the FGI-PointTransformer-DL-3D method. Similarly, the inference speed of the DetailView-DL-2D method is approximately 52 times slower when compared to the FGI-PointTransformer-DL-3D method. The FGI-RF-ML method uses a majority of the computation time for the calculation of the point cloud features (99,98% of the total time for the HeliALS dataset and 99,74% of the total time for the Optech Titan dataset), whereas the prediction process of the RF classifier is relatively fast. It can be hypothesized that a more efficient implementation of the feature calculation process could substantially increase the inference speed of FGI-RF-ML.

The implementation details of the inference process influence the results to a certain degree. For example, the FGI-PointTransformer-DL-3D was trained separately for five times and the five model instances were used to infer the tree species on the test dataset using a majority voting scheme. Likewise, the DetailView-DL-2D method performed the inference for 50 randomly augmented versions of each test set sample obtaining the prediction by selecting the species with the maximum cumulative prediction probability. When these factors are taken into consideration, the FGI-PointTransformer-DL-3D method is approximately five times faster than the DetailView-DL-2D method.

Appendix C. Macro-average accuracy by profile category and tree crown category

Figs. C.1 and C.2 present the macro-average accuracies by profile category and by crown class category, respectively, for the best methods according to the macro F_1 score. The figures have been made for comparison against Figs. 12 and 13, which present the overall accuracies for the same categories. Notice the different scaling of the y-axis compared to Figs. 12 and 13.

Fig. C.1. Macro-average accuracy by profile category for the best machine learning and deep learning methods for HeliALS (Panel A) and Optech Titan (Panel B) datasets. Median values of macro-average accuracy were calculated across the methods that utilize the same dataset and approach type (DL/ML). The best method was determined by macro F_1 score. (a) For the HeliALS dataset, the best deep learning method is FGI-PointTransformerWeighted-DL-3D (Macro F_1 = 0.73) and the best machine learning method is FGI-RF-ML (Macro F_1 = 0.64). (b) For the Optech Titan dataset, the best deep learning method is FGI-PointTransformer-DL-3D (Macro F_1 = 0.58) and the best machine learning method is IBL-BalancedRF-ML (Macro F_1 = 0.57).

Fig. C.2. Macro-average accuracy by crown class category for the best machine learning and deep learning methods for HeliALS (Panel A) and Optech Titan (Panel B) datasets. Median values of macro-average accuracy were calculated across the methods that utilize the same dataset and approach type (DL/ML). The best method was determined by macro F_1 score. (a) For the HeliALS dataset, the best deep learning method is FGI-PointTransformerWeighted-DL-3D (Macro F_1 = 0.73) and the best machine learning method is FGI-RF-ML (Macro F_1 = 0.64). (b) For the Optech Titan dataset, the best deep learning method is FGI-PointTransformer-DL-3D (Macro F_1 = 0.58) and the best machine learning method is IBL-BalancedRF-ML (Macro F_1 = 0.57).

Table D.1
Confusion matrix: FBK-PointNet++-DL-3D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.70**	0.06	0.06	0.13	0.03	0.00	0.01	0.00	0.00	2047
Spruce	0.07	**0.70**	0.09	0.09	0.03	0.00	0.01	0.00	0.00	528
Birch	0.03	0.03	**0.79**	0.09	0.05	0.00	0.00	0.00	0.00	1258
Maple	0.10	0.02	0.09	**0.50**	0.14	0.08	0.01	0.01	0.04	587
Aspen	0.08	0.03	0.28	0.14	**0.42**	0.01	0.01	0.01	0.02	376
Rowan	0.08	0.00	0.09	0.37	0.16	**0.22**	0.01	0.04	0.03	171
Oak	0.02	0.02	0.31	0.29	0.20	0.08	**0.00**	0.02	0.06	49
Linden	0.13	0.01	0.02	0.25	0.20	0.05	0.00	**0.33**	0.01	129
Alder	0.08	0.03	0.30	0.08	0.35	0.00	0.02	0.03	**0.11**	116
Count	1635	554	1397	906	488	114	29	76	62	5261

Table D.2
Confusion matrix: FGI-Point2Vec-DL-3D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.92**	0.02	0.02	0.00	0.01	0.01	0.01	0.00	0.01	2047
Spruce	0.05	**0.90**	0.03	0.01	0.01	0.00	0.01	0.00	0.00	528
Birch	0.03	0.02	**0.81**	0.02	0.07	0.02	0.01	0.01	0.02	1258
Maple	0.01	0.01	0.01	**0.85**	0.02	0.03	0.04	0.02	0.01	587
Aspen	0.02	0.01	0.14	0.01	**0.61**	0.05	0.03	0.04	0.09	376
Rowan	0.02	0.01	0.01	0.05	0.08	**0.69**	0.04	0.01	0.09	171
Oak	0.02	0.00	0.00	0.10	0.02	0.08	**0.71**	0.02	0.04	49
Linden	0.00	0.00	0.01	0.08	0.04	0.11	0.05	**0.66**	0.05	129
Alder	0.02	0.01	0.03	0.08	0.34	0.04	0.22	0.00	**0.25**	116
Count	1961	541	1131	573	415	217	150	136	137	5261

Appendix D. Confusion matrices for all methods and confusion matrices for best methods with raw counts

Fig. D.1 shows the raw confusion matrices of counts for the best performing machine learning and deep learning methods on the HeliALS and Optech Titan datasets. The corresponding normalized confusion matrices are shown in Fig. 7.

Tables D.1–D.8 show the row-normalized confusion matrices of the deep learning methods using the HeliALS dataset, with the bolded diagonal entries representing the per-class recall. Furthermore, Tables D.9–D.13 show the normalized confusion matrices of the deep learning methods using the Optech Titan dataset, and Tables D.14–D.21 show the normalized confusion matrices of the machine learning methods using the Optech Titan dataset.

Table D.3
Confusion matrix: FGI-DGCNN-DL-3D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.94**	0.02	0.02	0.00	0.01	0.00	0.00	0.00	0.00	2047
Spruce	0.05	**0.89**	0.04	0.01	0.00	0.01	0.00	0.01	0.00	528
Birch	0.02	0.01	**0.93**	0.01	0.02	0.00	0.00	0.00	0.00	1258
Maple	0.01	0.01	0.02	**0.90**	0.01	0.03	0.01	0.01	0.00	587
Aspen	0.03	0.01	0.20	0.03	**0.66**	0.02	0.01	0.03	0.01	376
Rowan	0.02	0.01	0.04	0.18	0.05	**0.63**	0.02	0.05	0.00	171
Oak	0.04	0.06	0.02	0.29	0.08	0.08	**0.33**	0.10	0.00	49
Linden	0.02	0.00	0.03	0.20	0.05	0.01	0.02	**0.67**	0.00	129
Alder	0.02	0.02	0.16	0.20	0.36	0.03	0.08	0.04	**0.09**	116
Count	2014	524	1364	650	358	155	47	132	17	5261

(a) FGI-RF-ML, HeliALS — Coniferous / Deciduous, True Species (rows) × Predicted Species (columns)

True \ Predicted	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Missing	Sum
Pine	1877	54	71	3	13	2	5	5	1	16	2047
Spruce	33	447	20	7	5	1	3	0	0	12	528
Birch	24	15	1180	4	23	1	0	2	0	9	1258
Maple	9	4	15	507	6	22	6	4	2	12	587
Aspen	18	7	150	0	191	3	3	1	1	2	376
Rowan	7	2	10	51	5	88	2	2	1	3	171
Oak	5	2	3	9	7	1	18	4	0	0	49
Linden	17	1	6	29	14	5	3	54	0	0	129
Alder	15	2	8	17	38	2	12	4	17	1	116
Sum	2005	534	1463	627	302	125	52	76	22	55	5261
P	0.94	0.84	0.81	0.81	0.63	0.70	0.35	0.71	0.77		

(b) FGI-PointTransformerWeighted-DL-3D, HeliALS

True \ Predicted	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Sum
Pine	1897	50	38	4	26	14	7	9	2	2047
Spruce	19	480	15	1	2	4	2	4	1	528
Birch	27	21	1123	11	42	15	5	4	10	1258
Maple	3	7	12	518	6	29	8	2	2	587
Aspen	8	1	39	1	289	16	7	2	13	376
Rowan	1	1	4	13	8	134	3	4	3	171
Oak	0	2	1	6	3	4	25	7	1	49
Linden	1	0	0	18	4	6	1	98	1	129
Alder	2	4	5	7	56	2	4	4	32	116
Sum	1958	566	1237	579	436	224	62	134	65	5261
P	0.97	0.85	0.91	0.89	0.66	0.60	0.40	0.73	0.49	

(c) IBL-BalancedRF-ML, Optech Titan

True \ Predicted	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Sum
Pine	1839	61	95	1	39	5	3	4	0	2047
Spruce	30	448	26	2	16	2	1	3	0	528
Birch	84	102	1003	6	57	2	4	0	0	1258
Maple	5	12	8	523	16	9	11	3	0	587
Aspen	17	31	66	2	241	1	9	9	0	376
Rowan	17	3	3	48	23	64	4	9	0	171
Oak	4	1	2	6	21	1	6	8	0	49
Linden	11	3	0	17	12	5	3	78	0	129
Alder	3	5	5	17	51	0	27	7	1	116
Sum	2010	666	1208	622	476	89	68	121	1	5261
P	0.91	0.67	0.83	0.84	0.51	0.72	0.09	0.64	1.00	

(d) FGI-PointTransformer-DL-3D, Optech Titan

True \ Predicted	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Sum
Pine	1871	40	101	7	17	5	1	5	0	2047
Spruce	35	422	44	5	15	2	2	3	0	528
Birch	81	24	1091	4	51	2	2	1	2	1258
Maple	17	8	28	491	7	20	10	4	2	587
Aspen	35	26	120	3	170	6	5	3	8	376
Rowan	18	2	11	51	21	53	4	7	4	171
Oak	3	3	6	9	8	5	12	2	1	49
Linden	14	6	7	18	6	8	5	65	0	129
Alder	4	2	20	15	50	3	8	1	13	116
Sum	2078	533	1428	603	345	104	49	91	30	5261
P	0.90	0.79	0.76	0.81	0.49	0.51	0.24	0.71	0.43	

Fig. D.1. (a) Unnormalized confusion matrix of FGI-RF-ML obtaining the highest macro F_1 score (0.64) among the machine learning models using HeliALS data. FGI-RF-ML produced 55 missing predictions. (b) Unnormalized confusion matrix of FGI-PointTransformerWeighted-DL-3D obtaining the highest macro F_1 score (0.73) among the deep learning models using HeliALS data. (c) Unnormalized confusion matrix of IBL-BalancedRF-ML obtaining the highest macro F_1 score (0.57) among the machine learning models using Optech Titan data. (d) Unnormalized confusion matrix of FGI-PointTransformer-DL-3D obtaining the highest macro F_1 score (0.58) among the deep learning models using Optech Titan data. All confusion matrices show raw counts. Precisions are displayed below the matrix. Row and column sums are shown to the right and bottom of the matrix in gray.

Table D.4
Confusion matrix: FGI-PointNet-DL-3D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.94**	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	2047
Spruce	0.06	**0.87**	0.05	0.02	0.00	0.00	0.00	0.00	0.00	528
Birch	0.02	0.01	**0.93**	0.00	0.02	0.00	0.00	0.00	0.00	1258
Maple	0.01	0.01	0.04	**0.89**	0.02	0.02	0.01	0.01	0.00	587
Aspen	0.06	0.01	0.24	0.03	**0.59**	0.02	0.01	0.04	0.00	376
Rowan	0.04	0.01	0.05	0.16	0.06	**0.63**	0.02	0.03	0.01	171
Oak	0.04	0.06	0.08	0.35	0.08	0.08	**0.20**	0.10	0.00	49
Linden	0.04	0.00	0.02	0.24	0.05	0.02	0.00	**0.64**	0.00	129
Alder	0.03	0.02	0.10	0.33	0.33	0.04	0.02	0.05	**0.08**	116
Count	2037	519	1386	669	331	161	25	121	12	5261

Table D.5
Confusion matrix: FGI-PointTransformer-DL-3D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.94**	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	2047
Spruce	0.05	**0.88**	0.04	0.01	0.01	0.01	0.00	0.01	0.00	528
Birch	0.03	0.01	**0.92**	0.01	0.02	0.01	0.00	0.00	0.00	1258
Maple	0.01	0.01	0.03	**0.89**	0.02	0.03	0.01	0.00	0.00	587
Aspen	0.02	0.01	0.15	0.01	**0.78**	0.02	0.01	0.00	0.01	376
Rowan	0.01	0.01	0.02	0.10	0.05	**0.76**	0.02	0.02	0.01	171
Oak	0.04	0.04	0.08	0.14	0.06	0.12	**0.39**	0.12	0.00	49
Linden	0.01	0.01	0.01	0.19	0.03	0.03	0.01	**0.71**	0.02	129
Alder	0.03	0.01	0.17	0.11	0.47	0.02	0.03	0.03	**0.14**	116
Count	2014	524	1324	600	428	192	39	113	27	5261

Data availability

We have made available the Espoonlahti dataset for tree species classification in Zenodo: https://doi.org/10.5281/zenodo.17077255. The dataset includes both HeliALS and Optech Titan point clouds for the studied tree segments together with ground-truth species labels. The published Zenodo dataset will be described in more detail in a forthcoming data manuscript.

Table D.6

Confusion matrix: SA-Detailview-DL-2D, HeliALS.

Reference	Predicted species										
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Missing	Count
Pine	**0.50**	0.13	0.23	0.05	0.00	0.00	0.08	0.00	0.00	0.00	2047
Spruce	0.05	**0.87**	0.07	0.00	0.00	0.00	0.01	0.00	0.00	0.01	528
Birch	0.01	0.11	**0.83**	0.02	0.00	0.00	0.02	0.00	0.00	0.00	1258
Maple	0.01	0.07	0.37	**0.28**	0.00	0.00	0.26	0.00	0.00	0.00	587
Aspen	0.03	0.08	0.73	0.03	**0.00**	0.00	0.11	0.00	0.00	0.01	376
Rowan	0.01	0.24	0.54	0.15	0.00	**0.00**	0.05	0.00	0.00	0.00	171
Oak	0.04	0.04	0.65	0.02	0.00	0.00	**0.24**	0.00	0.00	0.00	49
Linden	0.26	0.20	0.19	0.10	0.00	0.00	0.24	**0.00**	0.00	0.01	129
Alder	0.03	0.03	0.46	0.20	0.00	0.00	0.28	0.00	**0.00**	0.01	116
Count	1122	1008	2251	375	2	0	484	0	0	19	5261

Table D.7

Confusion matrix: SLU-YOLOv8-DL-2D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.90**	0.01	0.03	0.01	0.02	0.00	0.01	0.01	0.00	2047
Spruce	0.05	**0.89**	0.03	0.00	0.02	0.00	0.00	0.01	0.00	528
Birch	0.02	0.01	**0.93**	0.01	0.03	0.00	0.01	0.00	0.00	1258
Maple	0.03	0.01	0.03	**0.83**	0.02	0.04	0.03	0.01	0.00	587
Aspen	0.06	0.03	0.23	0.02	**0.59**	0.01	0.06	0.01	0.00	376
Rowan	0.07	0.00	0.04	0.20	0.03	**0.60**	0.02	0.03	0.00	171
Oak	0.20	0.02	0.06	0.24	0.04	0.06	**0.31**	0.06	0.00	49
Linden	0.04	0.01	0.02	0.30	0.07	0.05	0.02	**0.48**	0.00	129
Alder	0.35	0.03	0.08	0.23	0.11	0.06	0.03	0.02	**0.09**	116
Count	2008	527	1383	643	340	161	91	96	12	5261

Table D.8

Confusion matrix: TUW-PointNet++-DL-3D, HeliALS.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.85**	0.03	0.05	0.02	0.03	0.01	0.00	0.01	0.00	2047
Spruce	0.04	**0.76**	0.10	0.02	0.07	0.00	0.00	0.01	0.00	528
Birch	0.04	0.02	**0.87**	0.00	0.05	0.00	0.01	0.00	0.00	1258
Maple	0.02	0.00	0.02	**0.76**	0.02	0.06	0.07	0.03	0.01	587
Aspen	0.06	0.03	0.24	0.01	**0.57**	0.01	0.03	0.05	0.01	376
Rowan	0.06	0.00	0.03	0.41	0.03	**0.38**	0.02	0.06	0.01	171
Oak	0.06	0.02	0.10	0.20	0.18	0.12	**0.18**	0.12	0.00	49
Linden	0.04	0.00	0.01	0.16	0.03	0.12	0.09	**0.53**	0.02	129
Alder	0.04	0.03	0.09	0.11	0.48	0.01	0.09	0.07	**0.09**	116
Count	1862	499	1382	612	474	149	112	147	24	5261

Table D.9

Confusion matrix: FBK-PointNet++-DL-3D, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.87**	0.03	0.09	0.00	0.01	0.00	0.00	0.00	0.00	2047
Spruce	0.06	**0.81**	0.10	0.01	0.01	0.00	0.00	0.00	0.01	528
Birch	0.07	0.04	**0.84**	0.00	0.03	0.00	0.00	0.00	0.00	1258
Maple	0.04	0.02	0.05	**0.77**	0.02	0.06	0.02	0.01	0.02	587
Aspen	0.10	0.14	0.28	0.02	**0.38**	0.03	0.02	0.01	0.04	376
Rowan	0.13	0.01	0.07	0.27	0.13	**0.22**	0.03	0.12	0.02	171
Oak	0.04	0.10	0.12	0.27	0.24	0.04	**0.14**	0.02	0.02	49
Linden	0.12	0.06	0.09	0.15	0.17	0.15	0.00	**0.26**	0.02	129
Alder	0.05	0.08	0.28	0.07	0.33	0.07	0.03	0.03	**0.06**	116
Count	1999	626	1487	563	311	119	37	74	45	5261

Table D.10

Confusion matrix: FGI-DGCNN-DL-3D, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.93**	0.02	0.03	0.00	0.01	0.00	0.00	0.01	0.00	2047
Spruce	0.11	**0.79**	0.06	0.00	0.01	0.00	0.00	0.02	0.00	528
Birch	0.12	0.03	**0.80**	0.01	0.04	0.00	0.00	0.00	0.00	1258
Maple	0.09	0.02	0.03	**0.66**	0.02	0.02	0.02	0.14	0.00	587
Aspen	0.16	0.08	0.25	0.01	**0.44**	0.01	0.02	0.03	0.00	376
Rowan	0.29	0.02	0.03	0.16	0.05	**0.09**	0.06	0.30	0.00	171
Oak	0.24	0.04	0.04	0.16	0.20	0.00	**0.12**	0.18	0.00	49
Linden	0.15	0.03	0.01	0.12	0.02	0.02	0.01	**0.64**	0.00	129
Alder	0.07	0.05	0.07	0.15	0.57	0.02	0.03	0.04	**0.00**	116
Count	2320	537	1240	478	335	33	43	275	0	5261

Table D.11

Confusion matrix: FGI-PointNet-DL-3D, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.90**	0.02	0.05	0.01	0.01	0.01	0.00	0.00	0.00	2047
Spruce	0.05	**0.84**	0.07	0.00	0.01	0.01	0.00	0.01	0.00	528
Birch	0.06	0.04	**0.84**	0.01	0.04	0.00	0.00	0.00	0.00	1258
Maple	0.02	0.02	0.04	**0.82**	0.02	0.07	0.02	0.01	0.00	587
Aspen	0.06	0.09	0.28	0.02	**0.49**	0.02	0.02	0.02	0.02	376
Rowan	0.08	0.01	0.06	0.29	0.10	**0.38**	0.01	0.08	0.00	171
Oak	0.02	0.04	0.12	0.18	0.20	0.12	**0.20**	0.10	0.00	49
Linden	0.05	0.02	0.02	0.19	0.12	0.12	0.05	**0.43**	0.00	129
Alder	0.02	0.03	0.09	0.13	0.58	0.03	0.03	0.03	**0.08**	116
Count	1999	592	1360	616	384	153	35	103	19	5261

Table D.12

Confusion matrix: FGI-PointTransformerWeighted- DL-3D, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.88**	0.03	0.06	0.00	0.02	0.00	0.00	0.00	0.00	2047
Spruce	0.04	**0.80**	0.07	0.01	0.04	0.00	0.01	0.02	0.00	528
Birch	0.06	0.05	**0.82**	0.01	0.06	0.00	0.00	0.00	0.00	1258
Maple	0.02	0.01	0.03	**0.82**	0.05	0.04	0.02	0.01	0.00	587
Aspen	0.09	0.08	0.18	0.02	**0.58**	0.02	0.02	0.01	0.01	376
Rowan	0.09	0.01	0.04	0.31	0.19	**0.28**	0.01	0.05	0.02	171
Oak	0.06	0.02	0.06	0.18	0.31	0.04	**0.24**	0.08	0.00	49
Linden	0.09	0.03	0.01	0.13	0.12	0.12	0.02	**0.57**	0.01	129
Alder	0.00	0.03	0.05	0.14	0.57	0.01	0.09	0.01	**0.10**	116
Count	1971	599	1297	607	508	88	53	114	24	5261

Table D.13

Confusion matrix: NTNU-ConvNeXt-T-DL-2D, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.84**	0.02	0.09	0.02	0.02	0.00	0.00	0.01	0.00	2047
Spruce	0.05	**0.73**	0.15	0.01	0.03	0.00	0.00	0.02	0.00	528
Birch	0.06	0.04	**0.81**	0.01	0.05	0.00	0.01	0.01	0.00	1258
Maple	0.01	0.00	0.02	**0.91**	0.03	0.02	0.01	0.01	0.00	587
Aspen	0.06	0.06	0.18	0.05	**0.48**	0.01	0.07	0.06	0.02	376
Rowan	0.03	0.01	0.02	0.43	0.08	**0.36**	0.01	0.06	0.00	171
Oak	0.08	0.02	0.08	0.24	0.20	0.02	**0.20**	0.14	0.00	49
Linden	0.04	0.02	0.01	0.22	0.05	0.05	0.01	**0.60**	0.00	129
Alder	0.02	0.02	0.08	0.21	0.42	0.02	0.10	0.09	**0.05**	116
Count	1855	505	1384	743	411	95	70	179	19	5261

Table D.14

Confusion matrix: Aalto-RF-ML, Optech Titan.

Reference	Predicted species										
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Missing	Count
Pine	**0.83**	0.02	0.13	0.00	0.00	0.00	0.00	0.01	0.00	0.00	2047
Spruce	0.09	**0.67**	0.22	0.01	0.00	0.01	0.00	0.00	0.00	0.00	528
Birch	0.08	0.02	**0.88**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1258
Maple	0.02	0.00	0.04	**0.84**	0.00	0.06	0.03	0.00	0.01	0.00	587
Aspen	0.13	0.08	0.48	0.01	**0.00**	0.03	0.10	0.07	0.12	0.00	376
Rowan	0.08	0.02	0.05	0.32	0.00	**0.40**	0.07	0.05	0.02	0.00	171
Oak	0.02	0.08	0.16	0.08	0.00	0.29	**0.16**	0.08	0.12	0.00	49
Linden	0.14	0.05	0.18	0.09	0.00	0.11	0.09	**0.33**	0.01	0.00	129
Alder	0.06	0.03	0.10	0.17	0.00	0.16	0.26	0.09	**0.14**	0.00	116
Count	1957	467	1744	594	0	177	121	113	87	1	5261

Table D.15

Confusion matrix: FGI-RF-ML, Optech Titan.

Reference	Predicted species										
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Missing	Count
Pine	**0.86**	0.02	0.11	0.00	0.01	0.00	0.00	0.00	0.00	0.00	2047
Spruce	0.05	**0.74**	0.16	0.01	0.02	0.00	0.00	0.01	0.00	0.00	528
Birch	0.09	0.05	**0.79**	0.00	0.04	0.00	0.00	0.00	0.00	0.02	1258
Maple	0.01	0.01	0.03	**0.85**	0.03	0.03	0.03	0.00	0.00	0.01	587
Aspen	0.10	0.09	0.15	0.01	**0.61**	0.01	0.01	0.01	0.00	0.00	376
Rowan	0.10	0.01	0.02	0.25	0.11	**0.44**	0.02	0.04	0.00	0.02	171
Oak	0.08	0.08	0.04	0.12	0.37	0.16	**0.12**	0.02	0.00	0.00	49
Linden	0.10	0.02	0.00	0.15	0.13	0.07	0.03	**0.48**	0.00	0.02	129
Alder	0.03	0.05	0.08	0.09	0.53	0.03	0.14	0.04	**0.00**	0.00	116
Count	1976	544	1382	593	446	129	61	89	0	41	5261

Table D.16
Confusion matrix: LUKE-MultiRF-ML, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.83**	0.06	0.06	0.01	0.01	0.00	0.02	0.00	0.00	2047
Spruce	0.09	**0.72**	0.14	0.01	0.02	0.00	0.01	0.01	0.00	528
Birch	0.06	0.04	**0.82**	0.00	0.06	0.00	0.01	0.00	0.01	1258
Maple	0.04	0.02	0.01	**0.85**	0.03	0.02	0.02	0.01	0.01	587
Aspen	0.10	0.06	0.23	0.01	**0.51**	0.01	0.02	0.02	0.05	376
Rowan	0.09	0.02	0.03	0.51	0.13	**0.17**	0.01	0.05	0.00	171
Oak	0.10	0.18	0.10	0.24	0.16	0.00	**0.12**	0.04	0.04	49
Linden	0.12	0.01	0.02	0.33	0.13	0.08	0.01	**0.32**	0.00	129
Alder	0.17	0.03	0.08	0.12	0.40	0.02	0.09	0.02	**0.08**	116
Count	1935	587	1346	691	411	62	87	81	61	5261

Table D.17
Confusion matrix: IDEAS-RF-ML, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.89**	0.01	0.08	0.00	0.02	0.00	0.00	0.00	0.00	2047
Spruce	0.11	**0.64**	0.23	0.01	0.02	0.01	0.00	0.00	0.00	528
Birch	0.08	0.02	**0.84**	0.01	0.04	0.00	0.00	0.00	0.00	1258
Maple	0.03	0.00	0.03	**0.87**	0.03	0.03	0.01	0.00	0.00	587
Aspen	0.13	0.05	0.26	0.01	**0.53**	0.01	0.01	0.01	0.00	376
Rowan	0.19	0.01	0.02	0.24	0.09	**0.40**	0.01	0.05	0.00	171
Oak	0.16	0.06	0.08	0.16	0.29	0.04	**0.08**	0.12	0.00	49
Linden	0.16	0.02	0.00	0.16	0.19	0.07	0.01	**0.40**	0.00	129
Alder	0.06	0.00	0.16	0.14	0.52	0.04	0.05	0.03	**0.00**	116
Count	2108	416	1471	618	429	116	25	78	0	5261

Table D.18
Confusion matrix: UEF-LGBM-ML, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.86**	0.03	0.09	0.00	0.01	0.00	0.00	0.01	0.00	2047
Spruce	0.05	**0.81**	0.10	0.01	0.01	0.00	0.00	0.01	0.00	528
Birch	0.09	0.04	**0.82**	0.00	0.04	0.00	0.00	0.00	0.00	1258
Maple	0.03	0.02	0.02	**0.87**	0.02	0.03	0.02	0.01	0.00	587
Aspen	0.14	0.10	0.21	0.01	**0.49**	0.02	0.01	0.01	0.02	376
Rowan	0.11	0.01	0.04	0.29	0.09	**0.37**	0.04	0.06	0.00	171
Oak	0.20	0.00	0.10	0.24	0.22	0.08	**0.10**	0.04	0.00	49
Linden	0.09	0.02	0.02	0.19	0.07	0.02	0.02	**0.56**	0.00	129
Alder	0.13	0.08	0.09	0.14	0.40	0.05	0.10	0.01	**0.01**	116
Count	2014	595	1389	629	355	105	46	113	15	5261

Table D.19
Confusion matrix: UEF-RF-ML, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.89**	0.01	0.09	0.00	0.01	0.00	0.00	0.00	0.00	2047
Spruce	0.06	**0.81**	0.09	0.01	0.02	0.00	0.00	0.01	0.00	528
Birch	0.11	0.03	**0.82**	0.00	0.03	0.00	0.00	0.00	0.00	1258
Maple	0.02	0.00	0.02	**0.89**	0.02	0.03	0.01	0.00	0.00	587
Aspen	0.20	0.07	0.14	0.00	**0.57**	0.01	0.02	0.01	0.00	376
Rowan	0.12	0.01	0.02	0.32	0.06	**0.42**	0.01	0.05	0.00	171
Oak	0.14	0.00	0.10	0.20	0.20	0.16	**0.12**	0.06	0.00	49
Linden	0.14	0.02	0.00	0.22	0.10	0.08	0.00	**0.45**	0.00	129
Alder	0.09	0.01	0.07	0.10	0.53	0.04	0.15	0.00	**0.00**	116
Count	2131	529	1334	640	385	119	40	83	0	5261

Table D.20
Confusion matrix: UEF-SVM-ML, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.84**	0.02	0.11	0.00	0.02	0.00	0.01	0.01	0.00	2047
Spruce	0.06	**0.71**	0.18	0.00	0.03	0.00	0.00	0.01	0.00	528
Birch	0.09	0.04	**0.82**	0.00	0.04	0.00	0.00	0.00	0.00	1258
Maple	0.01	0.01	0.04	**0.83**	0.04	0.05	0.03	0.00	0.00	587
Aspen	0.12	0.07	0.23	0.01	**0.55**	0.02	0.01	0.00	0.00	376
Rowan	0.04	0.01	0.01	0.35	0.10	**0.48**	0.00	0.01	0.00	171
Oak	0.04	0.02	0.08	0.29	0.31	0.16	**0.08**	0.02	0.00	49
Linden	0.06	0.01	0.03	0.19	0.12	0.08	0.02	**0.49**	0.00	129
Alder	0.03	0.02	0.10	0.07	0.64	0.03	0.09	0.02	**0.00**	116
Count	1924	504	1478	607	460	147	55	86	0	5261

Table D.21
Confusion matrix: UPV-GB-ML, Optech Titan.

Reference	Predicted species									
	Pine	Spruce	Birch	Maple	Aspen	Rowan	Oak	Linden	Alder	Count
Pine	**0.84**	0.04	0.09	0.01	0.02	0.00	0.00	0.00	0.00	2047
Spruce	0.05	**0.70**	0.17	0.01	0.04	0.00	0.01	0.01	0.01	528
Birch	0.09	0.07	**0.78**	0.01	0.04	0.00	0.00	0.00	0.00	1258
Maple	0.03	0.02	0.03	**0.83**	0.01	0.06	0.01	0.01	0.00	587
Aspen	0.10	0.14	0.23	0.00	**0.45**	0.02	0.03	0.01	0.02	376
Rowan	0.11	0.04	0.05	0.25	0.13	**0.39**	0.02	0.02	0.00	171
Oak	0.20	0.12	0.12	0.14	0.14	0.12	**0.14**	0.00	0.00	49
Linden	0.19	0.02	0.07	0.14	0.08	0.10	0.05	**0.36**	0.00	129
Alder	0.13	0.16	0.19	0.07	0.23	0.11	0.10	0.00	**0.01**	116
Count	1964	644	1401	597	360	145	66	66	18	5261

References

Amiri, N., Krzystek, P., Heurich, M., Skidmore, A., 2019. Classification of tree species as well as standing dead trees using triple wavelength ALS in a temperate forest. Remote. Sens. 11 (22), 2614.

Angelstam, P., Mikusiński, G., 1994. Woodpecker assemblages in natural and managed boreal and hemiboreal forest—a review. In: Annales Zoologici Fennici. JSTOR, pp. 157–172.

Axelsson, A., Lindberg, E., Olsson, H., 2018. Exploring multispectral ALS data for tree species classification. Remote. Sens. 10 (2), 183.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., Sharma, U., 2024. Explaining neural scaling laws. Proc. Natl. Acad. Sci. 121 (27), e2311878121.

Carbonell-Rivera, J.P., Estornell, J., Ruiz, L.Á., Crespo-Peremarch, P., Almonacid-Caballer, J., Torralba, J., 2024. Class3Dp: A supervised classifier of vegetation species from point clouds. Environ. Model. Softw. 171, 105859.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chen, J., Liang, X., Liu, Z., Gong, W., Chen, Y., Hyyppä, J., Kukko, A., Wang, Y., 2024. Tree species recognition from close-range sensing: A review. Remote Sens. Environ. 313, 114337.

Chen, C., Liaw, A., Breiman, L., et al., 2004. Using random forest to learn imbalanced data. Univ. Calif. Berkeley 110 (24), 1–12.

City of Espoo, 2024. Espoo WMS and WFS open data resources. Retrieved 10.6.2024, URL: https://kartat.espoo.fi/opendata/.

Dalponte, M., Bruzzone, L., Gianelle, D., 2012. Tree species classification in the southern alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data. Remote Sens. Environ. 123, 258–270.

Dalponte, M., Ørka, H.O., Ene, L.T., Gobakken, T., Næsset, E., 2014. Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data. Remote Sens. Environ. 140, 306–317.

Demantké, J., Vallet, B., Paparoditis, N., 2012. Streamed vertical rectangle detection in terrestrial laser scans for facade database production. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. 1, 99–104.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Deng, S., Katoh, M., Yu, X., Hyyppä, J., Gao, T., 2016. Comparison of tree species classifications at the individual tree level by combining ALS data and RGB images using different algorithms. Remote. Sens. 8 (12), 1034.

European Parliament and Council, 2025. Regulation (EU) 2024/1991 of the European parliament and of the council of 24 june 2024 on nature restoration. Retrieved 6.10.2025, URL: https://eur-lex.europa.eu/eli/reg/2024/1991/oj/eng.

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 (6), 1246.

Fassnacht, F.E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L.T., Straub, C., Ghosh, A., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sens. Environ. 186, 64–87.

Federal Office for Agriculture and Food, 2025. The national programme concept for the conservation and sustainable utilization of forest genetic resources in the federal Republic of Germany. Retrieved 6.10.2025, URL: https://www.genres.de/en/sector-specific-portals/trees-and-shrubs/national-programme.

Fernandez-Diaz, J.C., Carter, W.E., Glennie, C., Shrestha, R.L., Pan, Z., Ekhtari, N., Singhania, A., Hauser, D., Sartori, M., 2016. Capability assessment and performance metrics for the titan multispectral mapping lidar. Remote. Sens. 8 (11), 936.

Frey, J., Schindler, Z., 2024. Detailview on github. Retrieved 27.1.2025, URL: https://github.com/JulFrey/DetailView.

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. PMLR, pp. 3809–3820.

Hakula, A., Ruoppa, L., Lehtomäki, M., Yu, X., Kukko, A., Kaartinen, H., Taher, J., Matikainen, L., Hyyppä, E., Luoma, V., et al., 2023. Individual tree segmentation and species classification using high-density close-range multispectral laser scanning data. ISPRS Open J. Photogramm. Remote. Sens. 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.

Heinzel, J., Koch, B., 2011. Exploring full-waveform LiDAR parameters for tree species classification. Int. J. Appl. Earth Obs. Geoinf. 13 (1), 152–160.

Hell, M., Brandmeier, M., Briechle, S., Krzystek, P., 2022. Classification of tree species and standing dead trees with lidar point clouds using two deep neural networks: pointcnn and 3dmfv-net. PFG–J. Photogramm. Remote. Sens. Geoinf. Sci. 90 (2), 103–121.

Henrich, J., van Delden, J., Seidel, D., Kneib, T., Ecker, A.S., 2024. TreeLearn: A deep learning method for segmenting individual trees from ground-based LiDAR forest point clouds. Ecol. Inform. 84, 102888.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y., 2017. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Hinton, G.E., Roweis, S., 2002. Stochastic neighbor embedding. Adv. Neural Inf. Process. Syst. 15.

Holmgren, J., Persson, Å., 2004. Identifying species of individual trees using airborne laser scanner. Remote Sens. Environ. 90 (4), 415–423.

Holmgren, J., Persson, Å., Söderman, U., 2008. Species identification of individual trees by combining high resolution LiDAR data with multi-spectral images. Int. J. Remote Sens. 29 (5), 1537–1552.

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 Sens. Environ. 173, 224–237.

Hovi, A., Raitio, P., Rautiainen, M., 2017. A spectral analysis of 25 boreal tree species. Silva Fenn. 51 (4), 7753.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Hyyppä, J., Hyyppä, H., Inkinen, M., Engdahl, M., Linko, S., Zhu, Y.H., 2000. Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. Forest Ecol. Manag. 128 (1–2), 109–120.

Hyyppä, J., Jaakkola, A., Chen, Y., Kukko, A., Kaartinen, H., 2013. Unconventional LIDAR mapping from air, terrestrial and mobile. In: Proceedings of the Photogrammetric Week. Wichmann/VDE Verlag Berlin Germany, pp. 205–214.

Hyyppä, E., Kukko, A., Kaartinen, H., Yu, X., Muhojoki, J., Hakala, T., Hyyppä, J., 2022. Direct and automatic measurements of stem curve and volume using a high-resolution airborne laser scanning system. Sci. Remote. Sens. 5, 100050.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLO (version 8.0.0). Retrieved 27.1.2025, URL: https://github.com/ultralytics/ultralytics.

Kaartinen, H., Hyyppä, J., Yu, X., Vastaranta, M., Hyyppä, H., Kukko, A., Holopainen, M., Heipke, C., Hirschmugl, M., Morsdorf, F., et al., 2012. An international comparison of individual tree detection and extraction using airborne laser scanning. Remote. Sens. 4 (4), 950–974.

Kaasalainen, S., Lindroos, T., Hyyppa, J., 2007. Toward hyperspectral lidar: Measurement of spectral backscatter intensity with a supercontinuum laser source. IEEE Geosci. Remote. Sens. Lett. 4 (2), 211–215.

Kamińska, A., Lisiewicz, M., Stereńczak, K., 2021. Single tree classification using multi-temporal ALS data and CIR imagery in mixed old-growth forest in Poland. Remote. Sens. 13 (24), 5101.

Kamińska, A., Lisiewicz, M., Stereńczak, K., Kraszewski, B., Sadkowski, R., 2018. Species-related single dead tree detection using multi-temporal ALS data and CIR imagery. Remote Sens. Environ. 219, 31–43.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Karila, K., Matikainen, L., Litkey, P., Hyyppä, J., Puttonen, E., 2019. The effect of seasonal variation on automated land cover mapping from multispectral airborne laser scanning data. Int. J. Remote Sens. 40 (9), 3289–3307.

Kim, S., McGaughey, R.J., Andersen, H.-E., Schreuder, G., 2009. Tree species differentiation using intensity data derived from leaf-on and leaf-off airborne laser scanner data. Remote Sens. Environ. 113 (8), 1575–1586.

Kivinen, S., Koivisto, E., Keski-Saari, S., Poikolainen, L., Tanhuanpää, T., Kuzmin, A., Viinikka, A., Heikkinen, R.K., Pykälä, J., Virkkala, R., et al., 2020. A keystone species, European aspen (Populus tremula L.), in boreal forests: Ecological role, knowledge needs and mapping using remote sensing. Forest Ecol. Manag. 462, 118008.

Koenig, K., Höfle, B., 2016. Full-waveform airborne laser scanning in vegetation studies—a review of point cloud and waveform features for tree species classification. Forests 7 (9), 198.

Korpela, I., Ørka, H.O., Maltamo, M., Tokola, T., Hyyppä, J., et al., 2010. Tree species classification using airborne lidar–effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. Silva Fenn. 44 (2), 319–339.

Kouki, J., Arnold, K., Martikainen, P., 2004. Long-term persistence of aspen–key host for many threatened species–is endangered in old-growth conservation areas in Finland. J. Nat. Conserv. 12 (1), 41–52.

Kuusinen, M., 1994. Epiphytic lichen flora and diversity on populus tremula in old-growth and managed forests of southern and middle boreal Finland. In: Annales Botanici Fennici. JSTOR, pp. 245–260.

Lehtomäki, M., Jaakkola, A., Hyyppä, J., Lampinen, J., Kaartinen, H., Kukko, A., Puttonen, E., Hyyppä, H., 2016. Object classification and recognition from mobile laser scanning point clouds in a road environment. IEEE Trans. Geosci. Remote Sens. 54 (2), 1226–1239.

Lin, Y., Hyyppä, J., 2016. A comprehensive but efficient framework of proposing and validating feature parameters from airborne LiDAR data for tree species classification. Int. J. Appl. Earth Obs. Geoinf. 46, 45–55.

Lin, H., Nazari, M., Zheng, D., 2024. PCTreeS: 3D point cloud tree species classification using airborne LiDAR images. arXiv preprint arXiv:2412.04714.

Lisiewicz, M., Kamińska, A., Kraszewski, B., Kuberski, Ł., Pilch, K., Stereńczak, K., 2025. Comprehensive mapping of individual living and dead tree species using leaf-on and leaf-off ALS and CIR data in a complex temperate forest. For.: An Int. J. For. Res. cpaf007.

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.

Liu, B., Huang, H., Su, Y., Chen, S., Li, Z., Chen, E., Tian, X., 2022a. Tree species classification using ground-based LiDAR data by various point cloud deep learning methods. Remote. Sens. 14 (22), 5733.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022b. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

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 Geosci. Remote. Sens. Lett. 19, 1–5.

Mäyrä, J., Keski-Saari, S., Kivinen, S., Tanhuanpää, T., Hurskainen, P., Kullberg, P., Poikolainen, L., Viinikka, A., Tuominen, S., Kumpula, T., et al., 2021. Tree species classification from airborne hyperspectral and LiDAR data using 3D convolutional neural networks. Remote Sens. Environ. 256, 112322.

Michałowska, M., Rapiński, J., 2021. A review of tree species classification based on airborne LiDAR data and applied classifiers. Remote. Sens. 13 (3), 353.

Ministry of Agriculture and Forestry of Finland, 2025. National forest strategy 2035 of Finland. Retrieved 6.10.2025, URL: https://mmm.fi/en/nfs2035.

Næsset, E., Gobakken, T., Holmgren, J., Hyyppä, J., Maltamo, M., Nilsson, M., Olsson, H., Persson, Å., Söderman, U., 2004. Laser scanning of forest resources: the Nordic experience. Scand. J. For. Res. 19 (6), 482–499.

Oinonen, O., Ruoppa, L., Taher, J., Lehtomäki, M., Matikainen, L., Karila, K., Hakala, T., Kukko, A., Kaartinen, H., Hyyppä, J., 2024. Unsupervised semantic segmentation of urban high-density multispectral point clouds. arXiv preprint arXiv:2410.18520.

Ørka, H.O., Gobakken, T., Næsset, E., Ene, L., Lien, V., 2012. Simultaneously acquired airborne laser scanning and multispectral imagery for individual tree species identification. Can. J. Remote Sens. 38 (2), 125–138.

Ørka, H.O., Næsset, E., Bollandsås, O.M., 2009. Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. Remote Sens. Environ. 113 (6), 1163–1174.

Pfennigbauer, M., Ullrich, A., 2010. Improving quality of laser scanning data acquisition through calibrated amplitude and pulse deviation measurement. In: Laser Radar Technology and Applications XV, vol. 7684, SPIE, pp. 463–472.

Prieur, J.F., St-Onge, B., Fournier, R.A., Woods, M.E., Rana, P., Kneeshaw, D., 2021. A comparison of three airborne laser scanner types for species identification of individual trees. Sensors 22 (1), 35.

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., Levick, S.R., Luck, L., Missarov, A., Mokroš, M., Owen, H., Stereńczak, K., Pitkänen, T., Puletti, N., Saarinen, N., Hopkinson, C., Torresan, C., Tomelleri, E., Weiser, H., Astrup, R., 2024. FOR-species20K dataset. http://dx.doi.org/10.5281/zenodo.13255198.

Puliti, S., Lines, E.R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M.J., Winiwarter, L., Rehush, N., Hristova, H., et al., 2025. Benchmarking tree species classification from proximally sensed laser scanning data: Introducing the FOR-species20k dataset. Methods Ecol. Evol.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. 30.

Quan, Y., Li, M., Hao, Y., Liu, J., Wang, B., 2023. Tree species classification in a typical natural secondary forest using UAV-borne LiDAR and hyperspectral data. GIScience Remote. Sens. 60 (1), 2171706.

Rana, P., St-Onge, B., Prieur, J.F., Budei, B.C., Tolvanen, A., Tokola, T., 2022. Effect of feature standardization on reducing the requirements of field samples for individual tree species classification using ALS data. ISPRS J. Photogramm. Remote Sens. 184, 189–202.

Reichler, M., Taher, J., Manninen, P., Kaartinen, H., Hyyppä, J., Kukko, A., 2024. Semantic segmentation of raw multispectral laser scanning data from urban environments with deep neural networks. ISPRS Open J. Photogramm. Remote. Sens. 12, 100061.

Remm, J., Hanski, I.K., Tuominen, S., Selonen, V., 2017. Multilevel landscape utilization of the siberian flying squirrel: Scale effects on species habitat use. Ecol. Evol. 7 (20), 8303–8315.

Rönnholm, P., Wittke, S., Ingman, M., Putkiranta, P., Kauhanen, H., Kaartinen, H., Vaaja, M., 2022. Utilising simulated tree data to train supervised classifiers. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sciences-ISPRS Arch. 43 (B2-2022), 633–639.

Ruoppa, L., Oinonen, O., Taher, J., Lehtomäki, M., Takhtkeshha, N., Kukko, A., Kaartinen, H., Hyyppä, J., 2025. Unsupervised deep learning for semantic segmentation of multispectral LiDAR forest point clouds. arXiv preprint arXiv:2502.06227.

Rusu, R.B., 2010. Semantic 3D object maps for everyday manipulation in human living environments. KI-Künstliche Intell. 24, 345–348.

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. Front. Plant Sci. 12, 635440.

Shi, Y., Skidmore, A.K., Wang, T., Holzwarth, S., Heiden, U., Pinnel, N., Zhu, X., Heurich, M., 2018a. Tree species classification using plant functional traits from LiDAR and hyperspectral data. Int. J. Appl. Earth Obs. Geoinf. 73, 207–219.

Shi, Y., Wang, T., Skidmore, A.K., Heurich, M., 2018b. Important LiDAR metrics for discriminating forest tree species in central Europe. ISPRS J. Photogramm. Remote Sens. 137, 163–174.

Suratno, A., Seielstad, C., Queen, L., 2009. Tree species identification in mixed coniferous forest using airborne laser scanning. ISPRS J. Photogramm. Remote Sens. 64 (6), 683–693.

Takhtkeshha, N., Mandlburger, G., Remondino, F., Hyyppä, J., 2024. Multispectral light detection and ranging technology and applications: a review. Sensors 24 (5), 1669.

Tikkanen, O.-P., Martikainen, P., Hyvärinen, E., Junninen, K., Kouki, J., 2006. Red-listed boreal forest species of Finland: associations with forest structure, tree species, and decaying wood. In: Annales Zoologici Fennici. JSTOR, pp. 373–383.

Toivonen, J., Kangas, A., Maltamo, M., Kukkonen, M., Packalen, P., 2024. Mapping large European aspen (Populus tremula L.) in Finland using airborne lidar and image data. Can. J. Forest Res. 54 (7), 762–773.

Trier, Ø.D., Salberg, A.B., Kermit, M., Rudjord, Ø., Gobakken, T., Næsset, E., Aarsten, D., 2018. Tree species classification in Norway from airborne hyperspectral and airborne laser scanning data. Eur. J. Remote. Sens. 51 (1), 336–351.

Wang, Y., Lehtomäki, M., Liang, X., Pyörälä, J., Kukko, A., Jaakkola, A., Liu, J., Feng, Z., Chen, R., Hyyppä, J., 2019a. Is field-measured tree height as reliable as believed–A comparison study of tree height estimates from field measurement, airborne laser scanning and terrestrial laser scanning in a boreal forest. ISPRS J. Photogramm. Remote Sens. 147, 132–145.

Wang, L., Lu, D., Xu, L., Robinson, D.T., Tan, W., Xie, Q., Guan, H., Chapman, M.A., Li, J., 2024. Individual tree species classification using low-density airborne multispectral LiDAR data via attribute-aware cross-branch transformer. Remote Sens. Environ. 315, 114456.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019b. Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. (Tog) 38 (5), 1–12.

Wang, K., Wang, T., Liu, X., 2018. A review: Individual tree species classification using integrated airborne LiDAR and optical imagery with a focus on the urban environment. Forests 10 (1), 1.

West, K.F., Webb, B.N., Lersch, J.R., Pothier, S., Triscari, J.M., Iverson, A.E., 2004. Context-driven automated target detection in 3D data. In: Automatic Target Recognition XIV, vol. 5426, SPIE, pp. 133–143.

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 Sens. Environ. 313, 114367.

Winiwarter, L., 2024. Ensemble pointnet++ on github. Retrieved 27.1.2025, URL: https://github.com/lwiniwar/Tr3D_species_lwiniwar.

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 J. Photogramm. Remote Sens. 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 Sens. Environ. 305, 114078.

Yu, X., Hyyppä, J., Litkey, P., Kaartinen, H., Vastaranta, M., Holopainen, M., 2017. Single-sensor solution to tree species classification using multispectral airborne laser scanning. Remote. Sens. 9 (2), 108.

Zeid, K.A., Schult, J., Hermans, A., Leibe, B., 2023. Point2vec for self-supervised representation learning on point clouds. In: DAGM German Conference on Pattern Recognition. Springer, pp. 131–146.

Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V., 2021. Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16259–16268.

Zhong, H., Lin, W., Liu, H., Ma, N., Liu, K., Cao, R., Wang, T., Ren, Z., 2022. Identification of tree species based on the fusion of UAV hyperspectral image and LiDAR data in a coniferous and broad-leaved mixed forest in northeast China. Front. Plant Sci. 13, 964769.