Smart Agricultural Technology 14 (2026) 102132

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Parameter-efficient vision transformer adaptation for stem quality classification from smartphone forest images

Niccolò Biondi [a], Simone Ricci [b], Niccolò Arati [b], Michela Nocetti [c,*], Giovanni Aminti [c], Pietro Pala [b], Michele Brunetti [c]

[a] Department of Information Engineering and Computer Science, Via Sommarive 9, Trento, 38123, Italy
[b] MICC, Department of Information Engineering, University of Florence, Via Santa Marta 3, Florence, 50139, Italy
[c] CNR-IBE, Institute for Bioeconomy, National Research Council, Via Madonna del Piano 10, Sesto Fiorentino, 50019, Italy

ARTICLE INFO

Keywords:
Digital forestry
Low-rank adaptation (LoRA)
Wood quality
Smartphone image-based classification

ABSTRACT

Assessing wood quality at early stages of the forest-wood value chain remains a significant challenge, as qualitative evaluations are more typically performed after harvesting. This study presents the application of Vision Transformers (ViTs) for the qualitative classification of standing trees at the forest-compartment level, based on images captured using a standard smartphone camera.

A dataset of 460 terrestrial forest images was collected in 31 Douglas fir compartments managed for timber production, with images manually assigned to three stem-quality classes (117 class 1, 243 class 2, 100 class 3). ViT models pre-trained on ImageNet were employed to classify stem quality at both the image and forest-compartment levels. Several adaptation strategies were evaluated, including Full Fine-Tuning, and parameter-efficient fine-tuning based on Low-Rank Adaptation (LoRA). Model performance was assessed using a stratified 10-fold cross-validation, with data splitting performed at the compartment level to ensure spatial independence between training and testing data.

The results demonstrate that ViT-based models can effectively classify stem quality despite limited and imbalanced training data. Among the evaluated strategies, LoRA achieved the highest performance, reaching approximately 0.69 accuracy at the image level and 0.78 accuracy at the compartment level, consistently outperforming both Full Fine-Tuning and baseline approaches. Aggregation of predictions at the compartment level via majority voting further improved robustness and reduced misclassifications compared to image-level predictions.

The proposed approach enables the extraction of qualitative information from low-cost image data and can be readily integrated into digital forest inventory workflows. This development supports more informed forest management and timber commercialization strategies by complementing traditional quantitative metrics with the assessments of wood quality.

1. Introduction

Quality assessment represents a fundamental component of the efficient utilization of any product. Within the forest-wood value chain, such assessments are most commonly conducted during the final phases of the production process, either to enhance market value or to comply with mandatory regulations (e.g., for structural timber in construction [1,2]). However, anticipating the evaluation of material quality up to standing trees in the forest or to roundwood at the log yard offers numerous potential advantages for actors operating both in the forestry sector and in downstream industrial processes.

Access to information on wood quality, in addition to quantity, supports forest management strategies that emphasize the cultivation of value beyond volume [3]. At the same time, it enhances resource use, making it more efficient and economically viable. Consequently, timber harvesting and commercialization can be better aligned with the specific requirements of different industrial end uses. From a processing standpoint, industries could benefit from receiving raw material that is pre-selected, more homogeneous, and better suited to the intended applications, ultimately resulting in higher yields and improved product performance [4–6].

Previous studies on stand or log segregation according to quality have demonstrated its effectiveness in improving the prediction of sawn timber properties and increasing material recovery [7–11].

However, the earlier quality evaluation is implemented along the value chain, the greater the opportunities for informed decision-making.

* Corresponding author.
E-mail address: michela.nocetti@cnr.it (M. Nocetti).

https://doi.org/10.1016/j.atech.2026.102132
Received 15 February 2026; Received in revised form 18 April 2026; Accepted 18 April 2026
Available online 20 April 2026

It would be highly beneficial to include information on stem characteristics related to wood quality already during forest inventory surveys on standing trees [12]. Although there is currently no harmonization regarding which characteristics should be assessed [13], a few highly descriptive traits can be identified, among which branchiness and stem form are particularly relevant [14]. Visual grading of standing trees can be based on these stem properties; however, their assessment is challenging due to time and resource constraints. Manual evaluation is often impractical for large-scale surveys, which is why stem quality assessment in standing trees is not commonly performed.

In this regard, technology could provide valuable support [15], starting with laser scanning and photogrammetry techniques, which are widely applied for the determination of quantitative forest parameters such as tree diameter, height, and volume [16–19]. The application of these technologies for assessing the visually detectable characteristics of standing trees related to wood quality was reviewed by [14], who highlighted their effectiveness during field surveys, but also emphasized the need for further improvements, particularly in data processing, algorithm optimization, and software development.

To address the limitations of the existing instrumental methods, a new approach exploiting the potential of deep neural networks for wood quality prediction through visual inspection appears highly promising. This approach relies on recent advances in image-based techniques and artificial intelligence. Developments in hardware and software technologies have opened new possibilities, improving both the speed and accuracy of surveys and enabling more detailed data acquisition. In particular, computer vision and deep learning techniques show strong potential, as they allow real-time evaluations and enable the automatic extraction of stem attributes such as diameter, branch dimension, and branch angle from stereo images with sufficient accuracy for size class definition [20,21].

Image-based solutions have gained increasing attention due to their affordability, portability, and the availability of easy-to-use devices for data acquisition, typically a camera. In this context, the use of smartphone applications for tree measurements is growing, making forest surveys progressively more cost-effective [22,23]. Among them, Trestima® is a mobile application used for forest inventory that estimates forest attributes by capturing images with the smartphone camera [24,25]. During field surveys, the operator walks through the forest compartment, acquires images, and uploads them for cloud-based processing. The estimated variables, including tree species, basal area, median diameter, and height, are then provided to the user in near real time. The resulting report describes the forest compartment in terms of quantitative metrics. However, such applications are currently used primarily for determining tree dimensions, but the same images employed for metric estimation could also be exploited to assess stem quality, thereby enabling the integration of qualitative descriptors in forest inventory outputs [26].

Unlike traditional machine learning approaches that rely on manual feature extraction, Deep Neural Networks, and in particular Transformer architectures, provide enhanced capabilities for processing complex visual data and detecting subtle patterns such as those related to wood quality. Originally developed for Natural Language Processing tasks, Transformer architectures rely on self-attention mechanisms that were subsequently found to be effective for image classification as well [27]. Following this insight, Vision Transformers (ViTs) have been investigated across a wide range of application domains. Maurício et al. [28] reviewed studies comparing ViTs with the previously dominant Convolutional Neural Networks (CNNs) for image classification, reporting that ViTs generally outperformed CNN-based models. One of the main characteristics of ViTs is their ability to capture global contextual information through self-attention, which can improve performance in the presence of noisy visual data. However, while the self-attention mechanism enables efficient modeling of long-range dependencies, ViTs may exhibit limited generalization when trained on small datasets without adequate regularization or pre-training. Both ViTs and CNNs are well suited for transfer learning and have demonstrated strong performance when adapted from large pre-trained models to downstream tasks.

In forest monitoring, ViTs have been investigated mainly in remote sensing applications such as forest health assessment [29], deforestation monitoring [30], forest fire detection [31], and canopy-height estimation [32]. Most existing studies operate on large-scale aerial imagery and rely on datasets that are not organized at the forest-plot (compartment) level. Despite the strong performance of modern deep models on generic computer-vision benchmarks, only a limited number of approaches have been tailored to proximal, ground-level images of forest inventory plots. Due to the scarcity of curated plot-level datasets, models pre-trained on natural images often exhibit limited performance in this specific domain, thus requiring adaptation.

Low-Rank Adaptation (LoRA) [33] is a parameter-efficient fine-tuning technique designed to adapt large pre-trained models without updating all their weights. Instead of updating the entire network, LoRA injects trainable low-rank decomposition matrices into the self-attention layers of the frozen pre-trained ViT backbone. This allows the model to adapt its attention patterns to task-specific visual features while maintaining the general visual representations learned during large-scale pre-training, making it particularly suitable for scenarios with limited data or computational resources. It has been adopted across a wide range of applications, including computer vision. Yang et al. [34] provided a comprehensive review of LoRA techniques across multiple domains, ranging from natural language processing and code generation to speech processing and computer vision, highlighting its ability to preserve model performance while substantially reducing computational requirements. Applications of LoRA-based adaptation have also been reported in remote sensing imagery analysis [35,36]. However, to the best of our knowledge, no prior study has investigated the use of LoRA for the analysis of forest proximity images acquired at ground level.

The present work addresses this gap by adapting a ViT-based model to a dataset of terrestrial forest images acquired by a smartphone camera and subsequently labeled for stem quality with both compartment- and image-level granularity. A stratified k-Fold Cross-Validation strategy was adopted to ensure robust evaluation and fair comparison. Standard full fine-tuning of the ViT backbone was compared with parameter-efficient fine-tuning using LoRA, also evaluating model robustness under domain shifts.

The overall objective of this study was to apply a Vision Transformer model to analyse terrestrial forest images and classify forest compartments according to tree-stem quality, with the aim of supporting forest inventory and management applications. To achieve this aim, the following specific objectives were pursued: (1) to develop and evaluate a ViT-based framework for stem quality classification using smartphone-acquired terrestrial forest images, considering both image-level and compartment-level prediction scales; (2) to compare full and parameter-efficient model adaptation strategies, assessing their impact on classification performance and robustness under data-scarce and domain-shift conditions; (3) to assess the generalization capability of the proposed approach through spatially independent, compartment-level cross-validation aligned with operational forest inventory requirements. The resulting output, a qualitative classification of forest compartments, could be integrated as an additional stand-level attribute within forest inventory systems, thereby enhancing decision-making processes along the forest-wood value chain.

The remainder of this paper is organized as follows: Section 2 details the dataset acquisition, the quality assessment protocol, and the dataset splitting strategy; Section 3 describes the proposed methodology for developing the ViT model; Section 4 presents the experimental results; Section 5 discusses the main findings; Section 6 addresses the limitations of the study and suggests possible future developments; finally, Section 7 summarizes the concluding remarks.

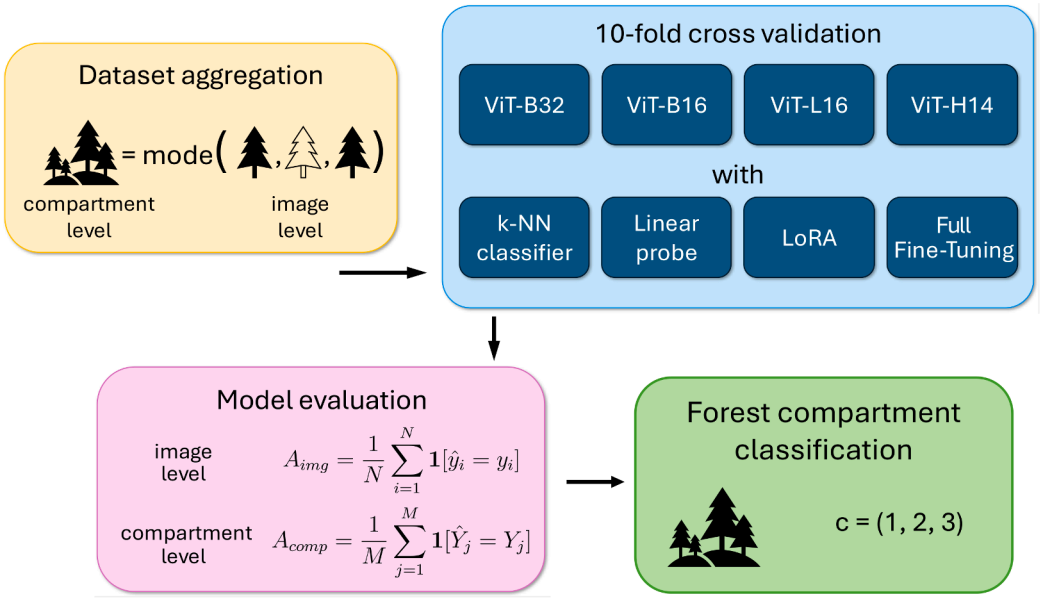


Fig. 1. Overview of the experimental pipeline. *Top-left*: Dataset aggregation where compartment-level classes are determined via majority voting (mode) of image-level predictions. *Top-right*: The 10-fold cross-validation framework benchmarking four Vision Transformer backbones (ViT-B32, B16, L16, H14) using adaptation strategies ranging from shallow classifiers (k-NN, Linear probe) to Low-Rank Adaptation (LoRA) and full Fine-tuning. *Bottom-left*: Model evaluation metrics for accuracy at both compartment (A_{comp}) and image (A_{img}) levels. *Bottom-right*: classification task.

Table 1
Characteristics of the forest stands surveyed for image data acquisition.

Variable	Mean	Min	Max
Area (ha)	1.9	0.2	7.3
Basal area (m^2/ha)	65.8	36.2	103.9
Mean DBH (cm)	43.8	27.6	71.7
Mean Height (m)	35.1	27.4	43.4
N trees / ha	455	190	996
Volume (m^3/ha)	954	504	1571
Age (years)	70	40	97

Fig. 2. Trestima® picture example.

2. Dataset

The pipeline of the experiment is outlined in Fig. 1 and detailed in the following paragraphs.

2.1. Dataset acquisition and stem quality evaluation

Data acquisition was conducted in 31 forest compartments of Douglas fir (*Pseudotsuga menziesii* (Mirb.) Franco) monoculture. Sampling was carried out across different properties aimed at the cultivation of this species as an even-aged plantation for timber production. Douglas fir was therefore the dominant species, although other tree species, such as silver fir, beech, black pine, and chestnut, were occasionally and locally present, depending on site-specific factors including soil type, exposure, and altitude. The characteristics of the stands are reported in Table 1. The average elevation was 920 m asl, ranging from 350 m asl for the lowest forest compartment to 1340 m.

Pictures were taken using the Trestima® (https://www.trestima.com) mobile application during a survey campaign carried out from June to the following October. The survey was conducted in daylight, under sunny or cloudy conditions (no rain or snow), while walking across the stand, without maintaining fixed distances from the trees. Three different mobile phone models were used: Samsung Galaxy A34, Motorola Moto G72, and Google Pixel 7a, with apertures of f/1.8, f/2.2, and f/1.89, and sensor sizes of 1/2.0", 1/1.67", and 1/1.73", respectively. Exposure settings were automatic, and HDR was enabled on all devices. The image resolution was set by the Trestima® application to 1600×900. The forest compartments were walked systematically, and an average of 15 pictures per stand was taken so to cover the entire stand area. The pictures are in landscape format; an example is depicted in Fig. 2.

The qualification of standing trees was performed by visual assessment. Since no grading rule is available for standing tree trunks, general guidance was taken from the European Standard "Quality grading of round timber of conifers. Part 3: Larch and Douglas fir" [37]. Three classes were developed to describe stem quality, primarily based on branchiness, with class 1 the best and class 3 the poorest quality (Table 2).

Fig. 3. Example of quality classification: class 1 on the left; class 2 in the center; class 3 on the right.

Table 2
Description of stem quality classes: branch parameters and limitation.

Quality class	Branch diameter	Number of branches
1	< 1 cm	< 2 / m
2	< 3 cm	2 / m < N < 5 / m
3	> 3 cm	> 2 / m

Table 3
Distribution (in percentages) of image classes within each forest compartment quality.

Forest	N. Images	Image class (%)		
Compartment class		Class 1	Class 2	Class 3
1	102	87.3	9.8	2.9
2	281	9.9	77.6	12.5
3	77	0.0	19.5	80.5

Quality 1: Generally, no branches are present in the first sawn log (the first 4-5 m of the stem); very small branches (less than 1 cm in diameter) may be present, but in limited numbers.

Quality 2: Branches may be present in the first sawn log; either numerous very small branches or a few medium-sized branches (less than 3 cm in diameter) are allowed.

Quality 3: Branches are present in the first sawn log; either numerous medium-sized branches or larger branches are allowed.

The forest compartments were classified according to the three qualities: to each image a quality class was assigned, determined by field observation and subsequent desk verification of the images (Fig. 3); the class assignment was guided by the 3-4 trees in the foreground of the picture. To each forest compartment was assigned the class most frequently observed among the images taken within that compartment.
The dataset consisted of 460 landscape pictures: 117 images were classified as class 1, 243 as class 2 and 100 as class 3. Accordingly, 6 forest compartments were classified as class 1, 18 as class 2 and 7 as class 3. Table 3 shows the distribution of the image classification for each forest compartment quality. The images and their classification as well as the classification by forest compartments are published on Zenodo [38]. As shown in Table 3, the dataset exhibits a mild class imbalance, particularly between class 2 and the other classes. This likely occurs because the middle class represents the average quality of the studied forests, and was therefore more frequent than the low and high quality classes. Details on how this issue was addressed during model development are provided in the following section.

2.2. Problem formulation and definition of evaluation metrics

The quality assessment task was formalized at two levels of granularity: the individual image level and the forest compartment level. Let the dataset be denoted as $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{N}$, consisting of N total images. Here, x_i represented the ith input image, $y_i \in \mathcal{L} = \{1, 2, 3\}$ denoted the ground truth quality label, and $c_i \in \{1, \ldots, M\}$ identified the unique forest compartment to which the image belonged, with M being the total number of compartments. The model $f(\cdot)$ mapped an input x_i to a predicted label $\hat{y}_i = f(x_i)$.

Two categories of evaluation metrics were defined, namely image-level and compartment-level metrics. Image-Level Accuracy (A_{img}) was used to quantify fine-grained classification performance and was defined as:

$$A_{img} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i], \tag{1}$$

where $\mathbb{1}[\cdot]$ denoted the indicator function.

In addition to accuracy, macro recall (equivalently, macro sensitivity) and macro F1-score were also considered in order to provide a more comprehensive assessment of the multiclass classification performance. Since the problem involved three classes, recall and F1-score were computed in a one-vs-rest fashion for each class and then macro-averaged across classes. For each class $\ell \in \mathcal{L}$, the image-level recall/sensitivity was defined as

$$R_{img}^{(\ell)} = \frac{TP_{img}^{(\ell)}}{TP_{img}^{(\ell)} + FN_{img}^{(\ell)}}, \tag{2}$$

where $TP_{img}^{(\ell)}$ and $FN_{img}^{(\ell)}$ denote the true positives and false negatives for class ℓ, respectively. The corresponding macro-averaged image-level recall was then computed as

$$mR_{img} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} R_{img}^{(\ell)}. \tag{3}$$

Similarly, the image-level F1-score for class ℓ was defined as

$$F1_{img}^{(\ell)} = \frac{2TP_{img}^{(\ell)}}{2TP_{img}^{(\ell)} + FP_{img}^{(\ell)} + FN_{img}^{(\ell)}}, \tag{4}$$

where $FP_{img}^{(\ell)}$ denotes the false positives for class ℓ. The macro-averaged image-level F1-score was then given by

$$mF1_{img} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} F1_{img}^{(\ell)}. \tag{5}$$

To align the evaluation with the operational requirements of the forest inventory application, in which aggregate quality information describes the quality of an entire forest compartment, compartment-level metrics were also introduced. These metrics were computed employing a majority voting aggregation strategy.

For a given compartment j, $S_j = \{i \mid c_i = j\}$ represented the set of indices corresponding to all images associated with that compartment. The aggregated compartment-level prediction, $\hat{Y}_j$, was defined as the mode of the image-level predictions:

$$\hat{Y}_j = \text{mode}(\{\hat{y}_i \mid i \in S_j\}). \tag{6}$$

Given the compartment ground truth Y_j, the compartment-level accuracy was computed as

$$A_{comp} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}[\hat{Y}_j = Y_j]. \tag{7}$$

Using the same majority-vote compartment predictions $\hat{Y}_j$, macro recall/sensitivity and macro F1-score were also computed at the compartment level. In particular,

$$R_{comp}^{(\ell)} = \frac{TP_{comp}^{(\ell)}}{TP_{comp}^{(\ell)} + FN_{comp}^{(\ell)}}, \quad mR_{comp} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} R_{comp}^{(\ell)}. \tag{8}$$

$$F1_{comp}^{(\ell)} = \frac{2TP_{comp}^{(\ell)}}{2TP_{comp}^{(\ell)} + FP_{comp}^{(\ell)} + FN_{comp}^{(\ell)}}, \quad mF1_{comp} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} F1_{comp}^{(\ell)}. \tag{9}$$

Here, $TP_{comp}^{(\ell)}$, $FP_{comp}^{(\ell)}$, and $FN_{comp}^{(\ell)}$ were computed by comparing the aggregated compartment predictions $\hat{Y}_j$ with the corresponding ground-truth compartment labels Y_j.

These metrics were used to assess the robustness of the proposed method against individual noisy viewpoints by exploiting spatial consistency within compartments.

2.3. Cross-validation strategy

To mitigate overfitting and evaluate generalization to unseen forest compartments, a stratified k-Fold Cross-Validation strategy ($k = 10$) was employed. Crucially, train-test data splitting was performed at the compartment level rather than the image level, ensuring that all images from a given compartment appeared exclusively in either the training or the test set of each fold. This approach explicitly evaluated the model's capacity for domain generalization, promoting learning of robust wood quality features rather than overfitting to site-specific environmental characteristics present in the training locations.

Within each fold, the test set comprised a balanced selection of compartments, specifically two compartments from each quality class, to ensure uniform evaluation. The remaining compartments formed the training set. While compartment selection was balanced, the number of images per fold varied due to the intrinsic distribution of images (see Table 3). Fig. 4 visually illustrates the resulting composition of the training and testing splits across all ten folds. This rigorous procedure allowed an evaluation of the model's ability to generalize to entirely new geographical locations. Finally, for reproducibility, the compartment IDs assigned to the test set for each fold are detailed in Table 4.

Table 4
Composition of the test sets for the 10-fold cross-validation. The table lists the IDs of the forest compartments (FC) [38] included in the test split for each fold.

Fold	Class 1		Class 2		Class 3	
	FC1	FC2	FC1	FC2	FC1	FC2
1	2881501	2881593	2880634	2881811	2880631	2938215
2	2881593	2882024	2881522	2881822	2880542	2938215
3	2881501	2881567	2881523	2881528	2880495	2938215
4	2881567	2881593	2880486	2880634	2759870	2880495
5	2881526	2882024	2880673	2881555	2881575	2938215
6	2881501	2882024	2880673	2881522	2880542	2881575
7	2881502	2881526	2771528	2881710	2880491	2880542
8	2881526	2881567	2880485	2880486	2880495	2880542
9	2881501	2881526	2880485	2880490	2880491	2881575
10	2881502	2882024	2881487	2881610	2759870	2938215

Table 5
Architectural hyperparameters of the vision transformer variants employed.

Model	Layers	Hidden Size D	MLP Size	Heads	Params
ViT-B/32	12	768	3072	12	86M
ViT-B/16	12	768	3072	12	86M
ViT-L/16	24	1024	4096	16	307M
ViT-H/14	32	1280	5120	16	632M

3. Method

Wood quality assessment was formulated as a supervised classification task. A ViT architecture was employed as the feature extractor. Two strategies were evaluated for adapting the pre-trained weights to the target domain: Full Fine-Tuning and LoRA.

3.1. Network architecture

The Vision Transformer [27] processes images as sequences of patch embeddings. An input image $x \in \mathbb{R}^{H \times W \times C}$—where H, W, and C represent the height, width, and number of channels, respectively—is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (P, P) denotes the patch resolution and $N = HW/P^2$ is the sequence length. The patches are mapped to a latent vector of size D via a trainable linear projection. A learnable position embedding is added to preserve spatial structure, and a learnable classification token (`[CLS]`) is prepended to the sequence. The resulting vectors are processed by an encoder consisting of L layers of Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. The final representation y is obtained from the state of the `[CLS]` token at the output of the Lth layer.

To analyze the influence of model capacity and architectural design on performance, four Vision Transformer variants were considered in this study: ViT-B/32, ViT-B/16, ViT-L/16, and ViT-H/14. All models utilized weights pre-trained on the ImageNet-21k dataset to ensure robust feature initialization. The architectural details for these models, including network depth, embedding dimensions, and parameter counts, are summarized in Table 5.

3.2. Model adaptation

3.2.1. Full fine-tuning

In this configuration, the pre-trained projection head was replaced with a randomly initialized linear classifier producing $C = 3$ class logits. During training, all encoder and classifier parameters were updated. Differential learning rates were adopted, with a lower learning rate for the encoder and a higher one for the classifier.

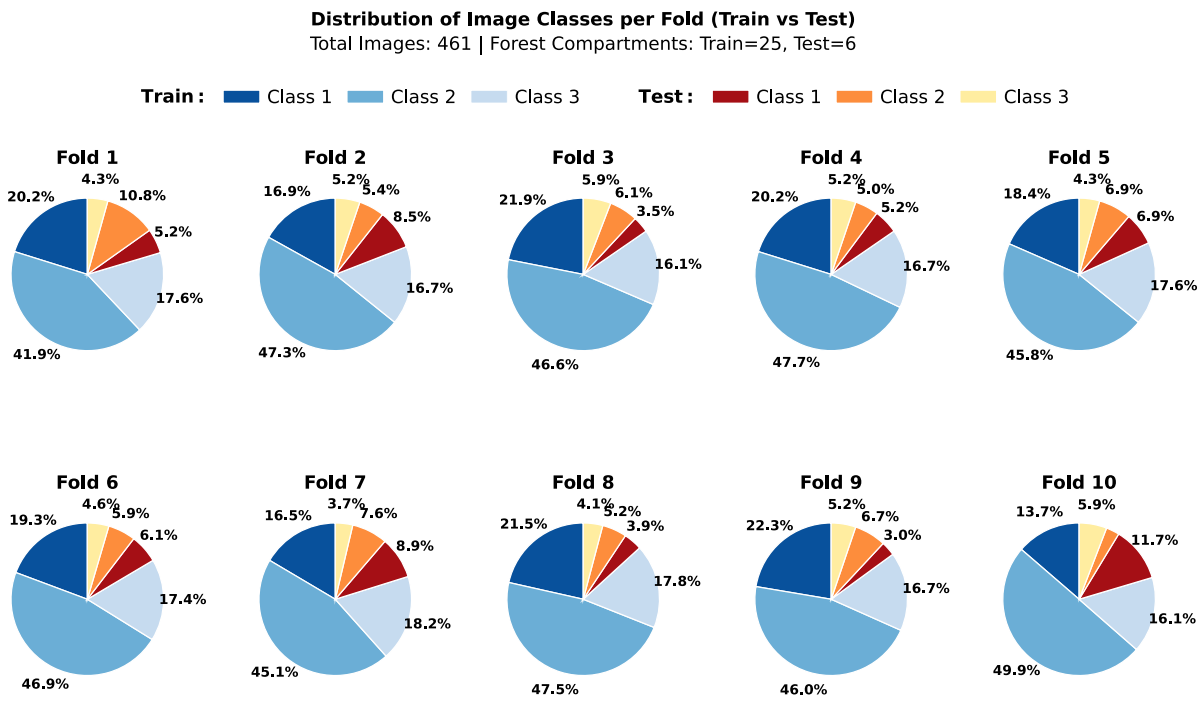


Fig. 4. 10-fold splits of dataset.

3.2.2. Low-rank adaptation (LoRA)

To enable efficient adaptation in a data-scarce regime, Low-Rank Adaptation (LoRA) [33] was employed. LoRA freezes the pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$ and models the weight updates ΔW via low-rank decomposition. The forward pass for a linear layer is modified as:

$$h = W_0 x + \Delta W x = W_0 x + \frac{\alpha}{r} BAx \tag{10}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$, and α is a scaling factor. The weight W_0 remains frozen during training.

LoRA modules were applied to the query (W_Q) and value (W_V) projection matrices within the self-attention mechanism, restricting optimization to a small subset of parameters.

3.3. Objective function

The model was optimized using the cross-entropy loss. Let $z_i \in \mathbb{R}^C$ denote the vector of raw logits for the ith sample in a batch of size B, and y_i denote the corresponding ground truth class index. The loss was defined as:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp(z_{i,y_i})}{\sum_{j=1}^{C}\exp(z_{i,j})}\right) \tag{11}$$

3.4. Compared methods

To assess the efficacy of the proposed adaptation strategies, their performance was compared against three baseline methods with frozen backbones:

- **k-Nearest Neighbors (k-NN):** the pre-trained encoder was used as a fixed feature extractor, and classification was performed on the `[CLS]` embeddings with a k-nearest neighbors classifier ($k = 5$).
- **Linear Probing:** the encoder was kept frozen and a single linear classifier was trained on the extracted `[CLS]` embeddings.
- **MLP Head:** the encoder was kept frozen and the linear classifier was replaced with a two-layer MLP with ReLU activations.
- **Training from Scratch:** the Vision Transformer backbone and the final classification layer were randomly initialized and trained directly on the target dataset, without using pre-trained weights. This setting was included to assess the contribution of transfer learning with respect to learning the model entirely from the available training data.

Friedman test and Nemenyi post-hoc test [39] were used to compare the models. These comparison methods provide reference points for evaluating the effect of frozen-feature baselines, random initialization, and encoder adaptation through Full Fine-Tuning and LoRA.

4. Experimental results

This section presents a comprehensive experimental evaluation of the proposed methodology. All experiments were conducted using an NVIDIA GeForce RTX 4090 GPU. The results are first reported in terms of model performance, with a comparison between the proposed adaptation strategies and the baseline methods. A sensitivity analysis was also performed on the best-performing model. Subsequently, qualitative analyses are provided by examining classification errors through confusion matrices. Finally, to validate the obtained results and to gain further insight into the model's behavior, an explainability analysis of the best-performing model is presented.

4.1. Implementation details

The LoRA rank was set to $r = 32$ with a scaling factor $\alpha = 32$. Training was performed using the AdamW optimizer, incorporating a weight decay of 0.0005 for regularization, and a strictly controlled learning rate schedule. A cosine annealing strategy was utilized, initializing the learning rate at 0.0001 and decaying it to zero over 100 epochs. No warmup strategy was adopted. Because the dataset is mildly imbalanced, balanced batch sampling was adopted during training so that each mini-batch contained a balanced representation of the three classes; therefore, no class-weighted loss was used. The batch size was set to 32 to accommodate the GPU memory constraints during experimentation. Rather than performing configuration-specific hyperparameter

Table 6
Performance comparison of different fine-tuning strategies across four Vision Transformer architectures at the image level. Image-Level Accuracy (A_{img}), macro recall (mR_{img}), and macro F1-score ($mF1_{img}$) are reported. The values represent the mean ± standard deviation over the 10-fold cross-validation. The best value in each column is reported in bold.

Method	A_{img}	mR_{img}	$mF1_{img}$
(a) ViT-B/32			
k-NN classifier	0.48 ± 0.06	0.45 ± 0.05	0.45 ± 0.06
Linear Probing	0.59 ± 0.11	0.58 ± 0.10	0.56 ± 0.11
MLP head	0.55 ± 0.10	0.54 ± 0.10	0.54 ± 0.09
From Scratch	0.65 ± 0.09	0.65 ± 0.09	0.64 ± 0.09
Full Fine-Tuning	0.36 ± 0.09	0.33 ± 0.12	0.34 ± 0.10
LoRA	**0.69 ± 0.08**	**0.68 ± 0.07**	**0.67 ± 0.08**
(b) ViT-B/16			
k-NN classifier	0.54 ± 0.08	0.53 ± 0.08	0.54 ± 0.07
Linear Probing	0.65 ± 0.08	0.63 ± 0.10	0.61 ± 0.09
MLP head	0.64 ± 0.07	0.64 ± 0.08	0.63 ± 0.09
From Scratch	0.64 ± 0.09	0.63 ± 0.10	0.63 ± 0.10
Full Fine-Tuning	0.49 ± 0.09	0.48 ± 0.10	0.45 ± 0.11
LoRA	**0.68 ± 0.08**	**0.66 ± 0.10**	**0.65 ± 0.09**
(c) ViT-L/16			
k-NN classifier	0.52 ± 0.10	0.50 ± 0.10	0.49 ± 0.09
Linear Probing	0.62 ± 0.06	0.59 ± 0.05	0.58 ± 0.07
MLP head	0.63 ± 0.09	0.61 ± 0.10	0.60 ± 0.10
From Scratch	0.65 ± 0.09	0.65 ± 0.10	0.64 ± 0.10
Full Fine-Tuning	0.38 ± 0.13	0.35 ± 0.14	0.37 ± 0.14
LoRA	**0.69 ± 0.08**	**0.67 ± 0.07**	**0.67 ± 0.08**
(d) ViT-H/14			
k-NN classifier	0.52 ± 0.06	0.52 ± 0.05	0.50 ± 0.04
Linear Probing	0.61 ± 0.06	0.60 ± 0.05	0.59 ± 0.06
MLP head	0.67 ± 0.04	0.65 ± 0.05	0.64 ± 0.04
From Scratch	0.63 ± 0.09	0.63 ± 0.10	0.62 ± 0.09
Full Fine-Tuning	0.35 ± 0.08	0.34 ± 0.07	0.33 ± 0.09
LoRA	**0.68 ± 0.08**	**0.67 ± 0.07**	**0.65 ± 0.08**

Table 7
Performance comparison of different fine-tuning strategies across four Vision Transformer architectures at the compartment level. Aggregated Compartment-Level Accuracy (A_{comp}), macro recall (mR_{comp}), and macro F1-score ($mF1_{comp}$) are reported. The values represent the mean ± standard deviation over the 10-fold cross-validation. Since each test fold contains two compartments for each of the three classes, mR_{comp} numerically coincides with A_{comp}. The best value in each column is reported in bold.

Method	A_{comp}	mR_{comp}	$mF1_{comp}$
(a) ViT-B/32			
k-NN classifier	0.43 ± 0.11	0.43 ± 0.11	0.40 ± 0.12
Linear Probing	0.67 ± 0.21	0.67 ± 0.21	0.64 ± 0.20
MLP head	0.63 ± 0.15	0.63 ± 0.15	0.61 ± 0.16
From Scratch	0.72 ± 0.17	0.72 ± 0.17	0.71 ± 0.16
Full Fine-Tuning	0.38 ± 0.15	0.38 ± 0.15	0.37 ± 0.15
LoRA	**0.78 ± 0.15**	**0.78 ± 0.15**	**0.78 ± 0.15**
(b) ViT-B/16			
k-NN classifier	0.58 ± 0.13	0.58 ± 0.13	0.57 ± 0.12
Linear Probing	0.73 ± 0.13	0.73 ± 0.13	0.71 ± 0.20
MLP head	0.77 ± 0.13	0.77 ± 0.13	0.74 ± 0.13
From Scratch	0.70 ± 0.18	0.70 ± 0.18	0.70 ± 0.18
Full Fine-Tuning	0.57 ± 0.13	0.57 ± 0.13	0.56 ± 0.13
LoRA	**0.78 ± 0.15**	**0.78 ± 0.15**	**0.76 ± 0.14**
(c) ViT-L/16			
k-NN classifier	0.52 ± 0.17	0.52 ± 0.17	0.51 ± 0.18
Linear Probing	0.70 ± 0.16	0.70 ± 0.16	0.70 ± 0.15
MLP head	0.60 ± 0.21	0.60 ± 0.21	0.59 ± 0.19
From Scratch	0.70 ± 0.18	0.70 ± 0.18	0.70 ± 0.18
Full Fine-Tuning	0.50 ± 0.15	0.50 ± 0.15	0.48 ± 0.15
LoRA	**0.78 ± 0.15**	**0.78 ± 0.15**	**0.77 ± 0.14**
(d) ViT-H/14			
k-NN classifier	0.65 ± 0.16	0.65 ± 0.16	0.62 ± 0.14
Linear Probing	0.63 ± 0.18	0.63 ± 0.18	0.60 ± 0.18
MLP head	**0.78 ± 0.08**	**0.78 ± 0.08**	0.75 ± 0.07
From Scratch	0.72 ± 0.17	0.72 ± 0.17	0.71 ± 0.16
Full Fine-Tuning	0.43 ± 0.11	0.43 ± 0.11	0.43 ± 0.10
LoRA	**0.78 ± 0.15**	**0.78 ± 0.15**	**0.77 ± 0.15**

searches, we maintained these standardized settings across all experiments to guarantee a fair, controlled comparison among the various ViT architectures and adaptation methods. To ensure robustness, all images were resized to 224 × 224 pixels and normalized using the standard statistics (mean and standard deviation) of the ImageNet dataset. Data augmentation techniques, including Random Crop, Horizontal Flip, and RandAugment, were applied during training to enhance the model's invariance to a variety of visual variations.

4.2. Model capacity performance

Quantitative results of the experimental evaluation are summarized in Tables 6 and 7. Performance was analysed along two dimensions: the effectiveness of the adaptation strategy and the impact of model architecture size, considering both image-level and compartment-level metrics as defined in Section 2.2.

A systematic comparison of the adaptation methods revealed distinct performance patterns across all backbone architectures. These trends were consistent not only in terms of accuracy, but also in macro recall and macro F1-score, indicating that the relative behavior of the compared strategies remained stable when class-wise performance was taken into account. As for the three baseline tested, the Linear Probing and MLP Head demonstrated similar performance, with the zero-shot *k-NN* classifier showed slightly lower metric values.

The training-from-scratch setting yielded competitive results across all backbone architectures. In most cases, it outperformed the frozen-feature baselines and consistently exceeded Full Fine-Tuning, showing that the model can learn task-relevant features directly from the target dataset. However, it remained below LoRA across all architectures and metrics, indicating that random initialization is viable but suboptimal with respect to adaptation from pretrained weights.

A notable observation was the degradation in performance associated with Full Fine-Tuning. As shown in Tables 6 and 7, updating all network parameters resulted in a sharp decline in accuracy (e.g., 0.36 A_{img} for ViT-B/32) relative to Linear Probing (0.59).

The proposed LoRA strategy yielded the highest performance consistently across all metrics and architectures. In particular, LoRA achieved the best, or tied-best, macro recall and macro F1-score in nearly all settings, showing that its advantage was not limited to overall accuracy but also reflected a more balanced recognition of the three classes. As a confirmation, statistical comparisons through Friedman test revealed highly significant differences across models (e.g., Chi-squared = 29.8, df = 4, $p < 0.001$ for image-level accuracy and Chi-squared = 21.8, df = 4, $p < 0.001$ for compartment-level accuracy, with the ViT-B/32 backbone). According to Nemenyi post-hoc test LoRA differed significantly from k-NN classifier ($p < 0.05$), and from Full Fine-Tuning ($p < 0.001$), but not from Linear Probing or MLP.

Analysis of the results across the four ViT variants highlighted a clear performance improvement when transitioning from ViT-B/32 to ViT-B/16. While both models share the same backbone capacity, the reduction in patch size from 32 × 32 to 16 × 16 resulted in a more detailed spatial decomposition of the input images, enabling more detailed representation of stem-related visual features.

Table 8
Sensitivity analysis and ablation study for the LoRA rank (r) and scaling factor (α) over 10-fold cross-validation. Image-level accuracy (A_{img}), macro recall (mR_{img}), and macro F1-score ($mF1_{img}$), as well as compartment-level accuracy (A_{comp}), macro recall (mR_{comp}), and macro F1-score ($mF1_{comp}$), are reported. The values represent the mean ± standard deviation over the 10-fold cross-validation.

α	r	Image Level			Compartment Level		
		A_{img}	mR_{img}	$mF1_{img}$	A_{comp}	mR_{comp}	$mF1_{comp}$
16	8	0.63 ± 0.11	0.62 ± 0.10	0.60 ± 0.13	0.73 ± 0.14	0.73 ± 0.14	0.71 ± 0.13
	16	0.66 ± 0.09	0.65 ± 0.11	0.63 ± 0.14	0.73 ± 0.11	0.73 ± 0.11	0.72 ± 0.11
	32	0.64 ± 0.08	0.63 ± 0.07	0.62 ± 0.12	0.70 ± 0.15	0.70 ± 0.15	0.68 ± 0.14
32	16	0.67 ± 0.09	0.66 ± 0.09	0.64 ± 0.09	0.73 ± 0.13	0.73 ± 0.13	0.71 ± 0.14
	32	0.68 ± 0.08	0.66 ± 0.07	0.65 ± 0.08	0.78 ± 0.15	0.78 ± 0.15	0.78 ± 0.15
	64	0.65 ± 0.08	0.63 ± 0.07	0.62 ± 0.07	0.75 ± 0.16	0.75 ± 0.16	0.73 ± 0.16
64	32	0.67 ± 0.08	0.66 ± 0.08	0.65 ± 0.12	0.70 ± 0.17	0.70 ± 0.17	0.69 ± 0.15
	64	0.68 ± 0.10	0.67 ± 0.07	0.66 ± 0.13	0.75 ± 0.15	0.75 ± 0.15	0.72 ± 0.16
	128	0.65 ± 0.09	0.64 ± 0.10	0.63 ± 0.11	0.73 ± 0.17	0.73 ± 0.17	0.73 ± 0.17

Under the LoRA adaptation strategy, the best results were already obtained with the ViT-B/32 backbone.

A targeted sensitivity analysis was then conducted on the LoRA hyperparameters, namely the rank r and the scaling factor α with ViT-B/32 backbone. The results are reported in Table 8. The best overall performance was obtained for ($\alpha = 32, r = 32$), which achieved the strongest compartment-level results and competitive image-level performance. More generally, the results suggested that intermediate values of r and α provide the most effective trade-off, while smaller or larger configurations do not yield further improvements.

4.3. Qualitative analysis and explainability

To complement the quantitative evaluation, a detailed qualitative analysis was conducted focusing on the ViT-B/32 architecture. This specific backbone was selected as a representative case to examine the model's behavior, error distribution, and learned representations under different adaptation strategies.

4.3.1. Error analysis via confusion matrices

The confusion matrices for image-level classification are presented in Fig. 5, while those for compartment-level classification are shown in Fig. 6. The values reported in these matrices correspond to averaged raw counts, rather than percentages. For each cross-validation fold, a confusion matrix was computed on the corresponding test set, and the final matrix was obtained by averaging the 10 fold-wise matrices. Therefore, each cell represents the mean number of samples for a given ground-truth/predicted class combination. Given the very similar performance observed for the Linear Probing and MLP-head baselines, only the results for Linear Probing are reported.

The confusion matrices highlighted the impact of the different training strategies on class-specific performance, in line with the trends observed in the accuracy metrics. The Full Fine-Tuning approach (Fig. 5c) exhibited the highest level of confusion, particularly for Class 2, the most represented class in the dataset. In this case, correct predictions accounted for only approximately one third of the samples, with substantial misclassifications into Class 1 (on average 11.9 images) and Class 3 (9.8 images). This behavior indicated that, although Full Fine-Tuning modified the network weights extensively, it struggled to generalize across the morphological features associated with stem quality.

Lower levels of confusion were observed for the k-NN (Fig. 5a) and Linear Probing (Fig. 5b) baselines, with correct classification rates of approximately 50% across all classes. The k-NN baseline performed slightly worse, particularly for Classes 2 and 3, where a substantial number of samples were incorrectly assigned to Class 1. On average, nearly one third of the images belonging to Class 3 in the test set were over-classified as Class 1.

In contrast, the LoRA-adapted model (Fig. 5d) exhibited a strong concentration along the diagonal, corresponding to the highest true positive rates across all classes. Notably, the model also demonstrated robust performance on the less represented classes (Classes 1 and 3), with limited confusion between distant quality categories (i.e., Class 1 vs. Class 3).

The advantages of the compartment-level aggregation strategy were further illustrated in Fig. 6. Compared to image-level predictions, the aggregated compartment-level results exhibited a clear "smoothing" effect. The relative performance ranking of the different adaptation strategies was consistently emphasized. Full Fine-Tuning again showed high confusion, particularly for Class 2, as did k-NN, whose performance was especially poor for Class 1 and 3, and Linear Probing.

For the LoRA-based approach, diagonal dominance was further strengthened at the compartment level (Fig. 6d). Class 1 achieved a notably high accuracy, with minimal confusion, while Class 3 exhibited substantially improved separability and an absence of extreme misclassifications. Overall, strong performance was observed across all classes, despite the inherent class imbalance of the dataset.

4.3.2. Explainability with score-CAM

To validate the morphological features driving the model's predictions, Score-CAM [40] attention maps were generated for the LoRA-adapted model using the ViT-B/32 backbone (Fig. 7).

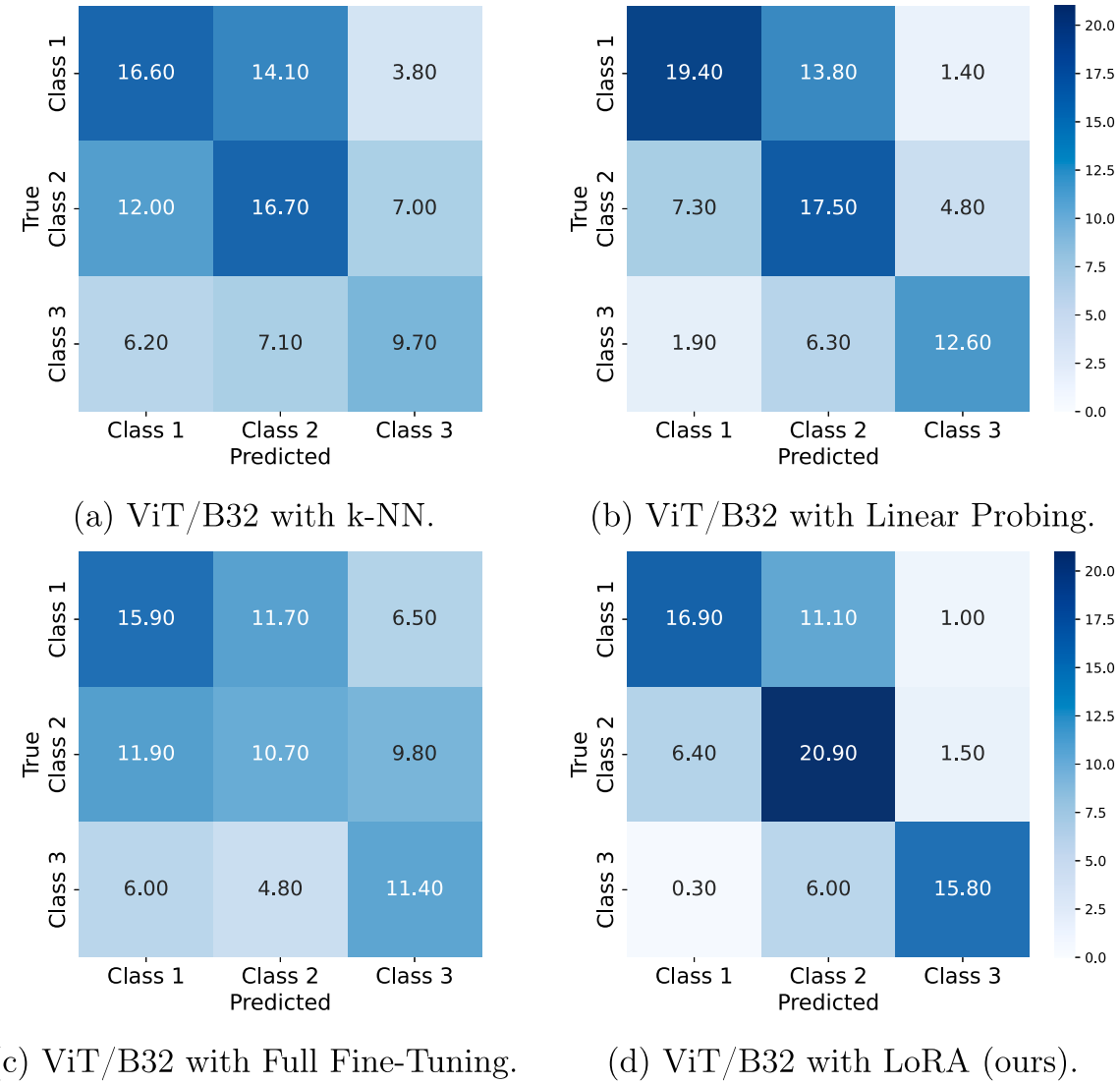
In correctly classified images (Fig. 7, left), the model consistently directed its attention towards the foreground stem and branching patterns, the primary determinants of quality classes. This focus is consistent with the class definition, which is based on the presence, size and distribution of branches. In contrast, misclassified samples (Fig. 7, middle and right) exhibited attention shifts toward background elements or environmental clutter.

4.3.3. Feature space visualization

To further analyze the learned feature space, t-SNE [41] projections were computed (Fig. 8).

In this study, t-SNE was applied to the feature embeddings to assess the separability of the quality classes in the learned feature space. The progression of cluster separability across models highlighted the effectiveness of the adaptation strategies. The k-NN baseline (Fig. 8a) exhibited a largely overlapping distribution, indicating that the pre-trained features alone were insufficient to separate the quality classes. This lack of clear clustering suggests limited discriminative capability in the absence of domain-specific adaptation.

Linear Probing (Fig. 8b) introduced partial separation, particularly for Class 1, indicating some alignment between the pre-trained features and the target task. In contrast, the LoRA-adapted model (Fig. 8c) produced the most compact and well-separated clusters. This clear separation demonstrates that parameter-efficient adaptation successfully



(a) ViT/B32 with k-NN.

(b) ViT/B32 with Linear Probing.

(c) ViT/B32 with Full Fine-Tuning.

(d) ViT/B32 with LoRA (ours).

Fig. 5. Confusion matrices comparing the performance of the ViT-B/32 backbone using different adaptation strategies on the image-level classification. Adaptation method compared: (a) k-NN, (b) Linear Probing, (c) Full Fine-Tuning, and (d) proposed LoRA method. The reported values represent the average over the 10 fold repetitions of the set test classification of each category.
Darker colors indicate a higher density of samples.

learned discriminative features specific to stem quality while preserving the semantic structure of the pre-trained ViT backbone.

5. Discussions

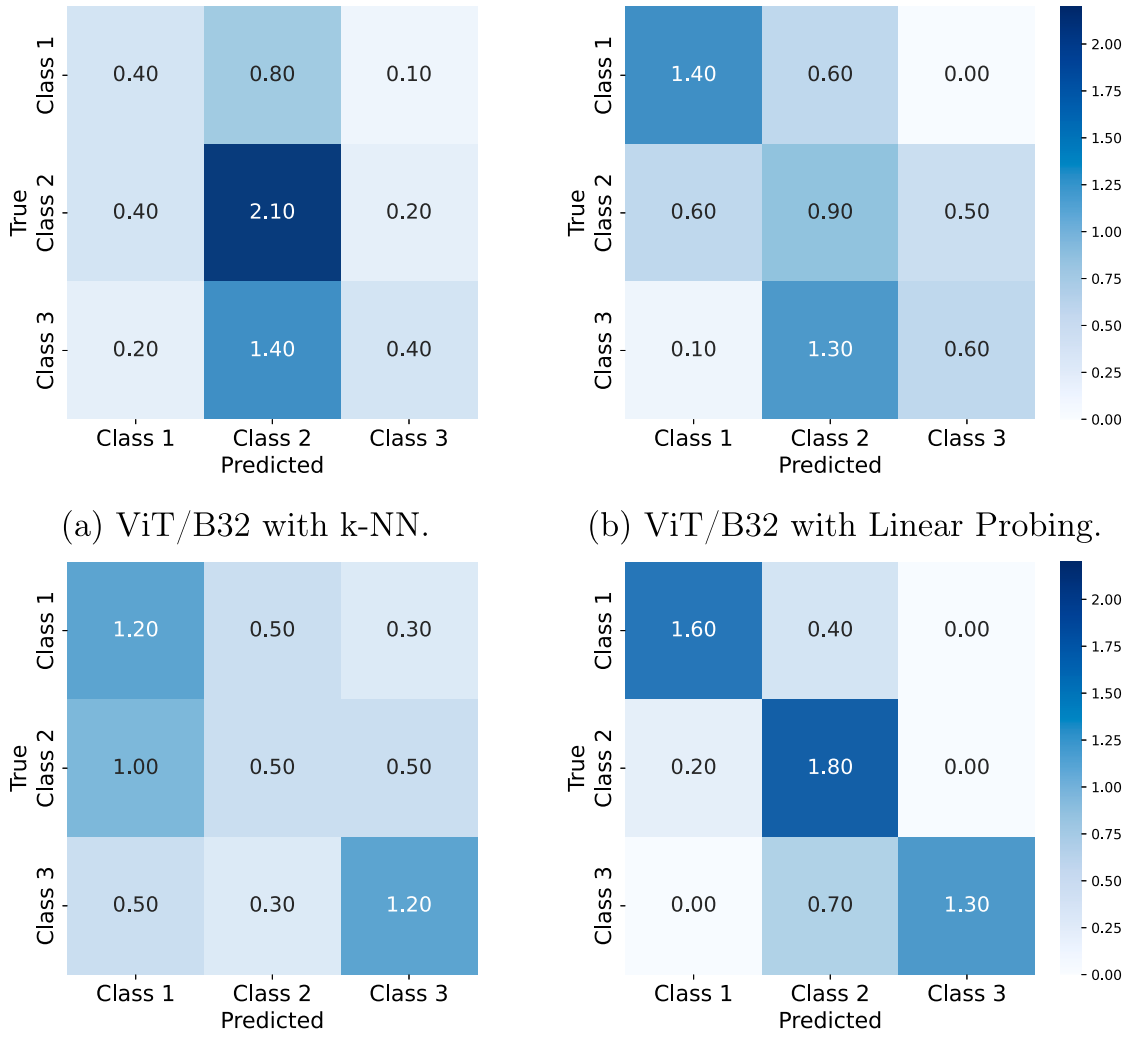
5.1. Application of vision transformers for qualitative forest classification

This study presented the application of Vision Transformers (ViTs) for the qualitative classification of terrestrial forest images, with the ultimate goal of integrating a qualitative index into the description of forest compartments, alongside conventional dimensional and quantitative inventory metrics. The qualitative assessment focused on one of the key characteristics, branchiness, that broadly determines wood quality and consequently the value of the timber obtainable from the forest harvesting and the downstream processing. The images of the dataset were classified according to the number and size of branches present in the tree trunk and the resulting quality class was the only input for the training of the model.

The results demonstrated that the use of ViTs is feasible for effectively classifying the forest compartments according to the quality of the tree stem present within them, even when operating on a relatively small and class-imbalanced dataset dominated by the most prevalent quality class in the investigated forests. Comparison with similar works in the literature is not possible directly, due to the absence of studies addressing a comparable domain. Nevertheless, studies on high resolution images collected by drone flights for weed and crop classification demonstrated that Vision Transformers outperformed CNN even with small labeled dataset [42]. In the present work, however, the model adaptation strategy proved to be a critical factor.

5.2. Effectiveness of the adaptation strategy and impact of the model architectural size

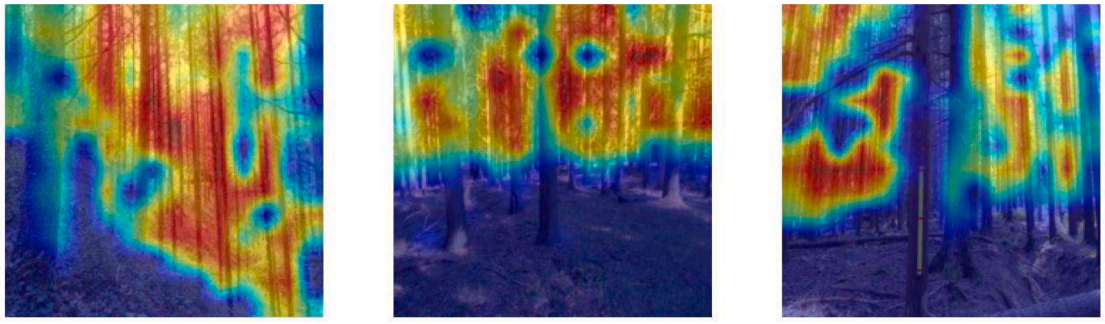
Comparing the adaptation methods, the adoption of LoRA enabled high classification performance, while performance dropped with Full Fine Tuning. This behavior is characteristic of catastrophic forgetting and overfitting, which are prevalent when over-parameterized models such as Vision Transformers are fine-tuned on small-scale datasets (460 images in the present study) without sufficient regularization. In this scenario, the model likely compromised the robust pre-trained fea-



(a) ViT/B32 with k-NN. (b) ViT/B32 with Linear Probing.

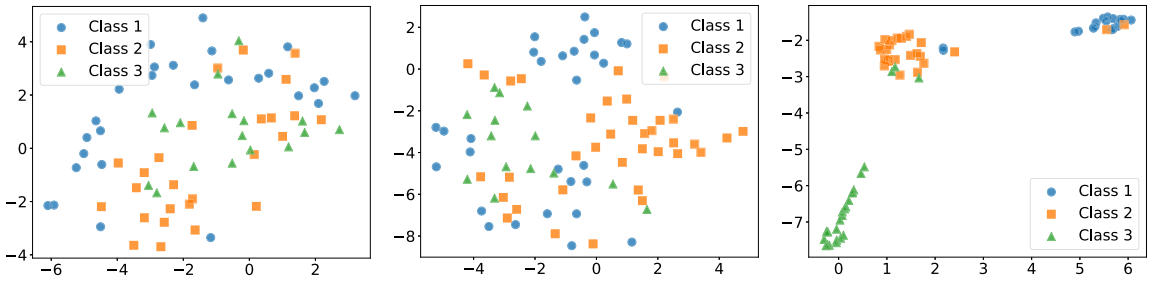
(c) ViT/B32 with Full Fine-Tuning. (d) ViT/B32 with LoRA (ours).

Fig. 6. Confusion matrices comparing the performance of the ViT-B/32 backbone using different adaptation strategies on the compartment-level classification. Adaptation methods compared: (a) k-NN, (b) Linear Probing, (c) Full Fine-Tuning, and (d) proposed LoRA method. The reported values represent the average over the 10 fold repetitions of the test set classification of each category.
Darker colors indicate a higher density of samples.



(a) Pred: 3, GT: 3 (b) Pred: 3, GT: 2 (c) Pred: 2, GT: 3

Fig. 7. Score-CAM visualizations for the best-performing model (ViT-B/32 with LoRA). Warmer colors indicate regions contributing most to the prediction. The left panel shows a correct classification, while the middle and right panels illustrate misclassifications. Correct predictions focus on foreground tree structures, whereas failure cases emphasize background regions, leading to erroneous decisions.



(a) ViT/B32 with k-NN. (b) ViT/B32 with Linear Probing. (c) ViT/B32 with LoRA (ours).

Fig. 8. t-SNE visualization of the feature embeddings learned by the ViT-B/32 backbone under different adaptation strategies. We compare the latent space representations for: (a) k-NN, (b) Linear Probing, and (c) our proposed LoRA method. Colors correspond to the ground-truth classes. While the baseline methods (a) and (b) exhibit significant overlap between classes, indicating entangled feature representations, LoRA (c) produces compact and well-separated clusters. This demonstrates that LoRA effectively enhances the discriminative power of the learned features.

tures in an attempt to fit the limited training samples. On the other hand, by optimizing a low-rank subset of parameters, LoRA successfully adapted the model to the specific domain of forest landscape images while preserving the generalization capability of the pre-trained backbone.

Similarly, LoRA adaptation strategy exhibited superior efficiency compared to Full Fine-Tuning in previous studies on medical imagery classification, despite the significant lower number of trainable parameters [43]. LoRA proved to be effective also in plant health monitoring, when applied to a ViT backbone pretrained on RGB remote sensing images and used to adapt the model to near infrared images [35].

The results obtained by training from scratch further support this interpretation. Although random initialization produced competitive performance, often exceeding the frozen-feature baselines and consistently outperforming Full Fine-Tuning, it remained inferior to LoRA. This suggests that, in the present small-data regime, a strong prior provided by pretrained weights is beneficial: it supports adaptation to stem-quality cues more effectively than learning entirely from the available target data, while avoiding the instability associated with updating all parameters.

Across the four ViT backbones analysed, performance improvements were primarily driven by finer input granularity rather than increased model capacity. Although the variants differ in both size, from Base (B) to Large (L) and Huge (H), and patch resolution (from 32×32 to 16×16), baseline methods benefited only from the reduction in patch size (Tables 6 and 7). In contrast, scaling up to larger architectures (ViT-L/16 and ViT-H/14) did not yield further gains.

Conversely, LoRA performance plateaued at roughly 0.69 at image-level and 0.78 at compartment-level, indicating that accuracy is constrained by the limited dataset rather than by representational capacity. Previous studies have shown that ViT accuracy tends to improve as the base-model size increases [43]. In the present case, however, once a certain capacity threshold is reached, further gains in quality classification are more likely to result from additional training data rather than from architectural scaling.

From a computational standpoint, backbone choice introduces substantial trade-offs. Model size ranges from 86M parameters for ViT-B variants to 632M for ViT-H/14 (Table 5), yet the larger models offer no measurable advantage in this task, making their additional computational load unjustified.

Selecting ViT-B/32 over ViT-B/16 further reflects this balance between accuracy and efficiency. Although both share the same number of parameters, ViT-B/16 generates four times more tokens, and its self-attention complexity scales quadratically with sequence length, greatly increasing FLOPs and inference latency. Since experiments under LoRA showed no accuracy benefit from this higher resolution, ViT-B/32 emerged as the most efficient architecture, offering the best accuracy-to-cost ratio. This makes it particularly suitable for smartphone-based forest inventory applications, where low latency and reduced power consumption are essential for real-time integration into operational workflows.

In practical field scenarios, inference can be executed either locally on the mobile device or through a hybrid architecture in which images are sent to the cloud for processing (the option adopted by Trestima® mobile application). If network connectivity is not available images can be temporarily stored and processed once connectivity is retrieved. The moderate computational requirements of the selected model architecture make on-device inference feasible on modern smartphones equipped with mobile neural processing units (NPUs) or GPUs, enabling near-real-time predictions during data acquisition, but also the option cloud-based can advantage from the lower computational requirements offering almost real-time outputs.

5.3. *Image-level versus compartment-level classification*

Across all successful experiments, Compartment-Level Accuracy consistently exceeded Image-Level Accuracy. The performance gap between these metrics indicates that individual images within a compartment often present challenging visual conditions. However, the collective evidence from multiple viewpoints enabled reliable compartment-level classification, which is better aligned with the operational requirements of forest inventory and management applications.

Looking at the confusion matrices (Figs. 5 and 6), LoRA-adapted model exhibited the highest classification performance. This result indicated a superior ability to discriminate between stem quality grades. Although image-level classification exhibited a non-negligible number of errors, misclassifications between the most distant quality classes were rare. More importantly, compartment-level classification, representing the intended operational application, proved to be substantially more robust. By aggregating predictions across multiple images, the compartment-level approach mitigated the impact of individual erroneous predictions and provided a reliable qualitative description of forest stands.

Overall, strong performance was observed across all classes, despite the inherent class imbalance of the dataset. These results confirmed that majority voting at the compartment level effectively mitigated the impact of noisy image-level predictions by exploiting spatial consistency within forest stands, thereby improving robustness and operational reliability.

5.4. Explicability

To gain insight of the features driving models' prediction, Class Activation Mapping (CAM) techniques [40] and t-SNE [41] were applied. CAM is commonly used to interpret model predictions by highlighting image regions that contribute most strongly to a given classification decision. Although originally developed for CNN, these techniques can be adapted to ViTs.

t-SNE is a non-linear dimensionality reduction technique commonly used to visualize high-dimensional data in two or three dimensions. By preserving local neighborhood relationships through pairwise similarity modeling, it is particularly effective for analyzing the clustering behavior of learned representations in deep neural networks.

Together, Score-CAM and t-SNE provided complementary local and global insights into the model's behavior. The explainability analysis supported the quantitative results, highlighting the superior performance and robustness of the proposed LoRA-based adaptation strategy.

In particular, these techniques revealed how correct classifications were driven by the foreground items (stems and branches), while misclassifications occurred when the attention shifted on the background objects. These patterns suggest that, while the LoRA adaptation enhanced feature discrimination, occlusions and background complexity remained the primary sources of erroneous inference.

5.5. Methodological novelty and annotation considerations

To the authors' knowledge, no directly comparable studies are currently available, preventing direct performance comparisons with existing approaches. It is noteworthy that the reported results were achieved using supervised learning based solely on image-level class labels, without explicit annotation of the specific visual features used for classification (e.g., branch presence and size). While such detailed annotations might have further improved performance, they would have required significantly greater manual effort during dataset preparation.

On the other hand, while manual classification reduces annotation burden compared to object-level labeling, it still requires expert knowledge. Semi-supervised, self-supervised, or weakly supervised learning paradigms could leverage large amounts of unlabeled forest imagery to reduce annotation effort while maintaining or improving performance. Contrastive pretraining or pseudo-labeling strategies may be particularly effective in forestry applications, where labeled datasets are often limited.

6. Limitations and future research directions

Despite the encouraging results, several limitations should be acknowledged when interpreting the findings of this study.

6.1. Dataset size and representativeness

The main limitations of the present study are related to the dataset. The experimental evaluation was conducted on a relatively small dataset. Although the use of transfer learning and parameter-efficient fine-tuning (LoRA) enabled stable convergence and strong performance, limited sample size intrinsically constrains the variability of visual conditions represented during training. This may reduce the model's robustness when deployed under substantially different illumination conditions, stand densities, management regimes, or site characteristics.
In forest imagery, dense shadows caused by strong sunlight can alter the visibility of branches or knots on tree trunks and degrade image quality. In our dataset, approximately half of the analyzed images exhibited pronounced shadows due to intense illumination; however, no specific patterns of performance degradation attributable to this factor were observed. Nevertheless, a dedicated analysis focusing on illumination effects could provide valuable insight.

In this context, future research should systematically investigate the minimum training dataset size required to achieve reliable and stable performance, with the aim of supporting scalable deployment in diverse forest conditions.

6.2. Class imbalance

The dataset exhibited class imbalance, reflecting the natural distribution of qualitative classes within the studied forest stands. While this improves ecological realism, it may bias the learned decision boundaries toward the dominant class. The approach adopted during model development, which maintained class balance in the test set, helped achieve similar accuracy across classes, although slightly lower performance was observed for those less represented. Misclassifications between highly dissimilar classes were limited; however, performance for minority classes may still be sensitive to distribution shifts. Future work should evaluate alternative imbalance mitigation strategies, such as focal loss, class-balanced sampling, or synthetic data augmentation, to assess their impact on robustness and calibration.

6.3. Compartment-level aggregation strategy

In the present study, compartment-level predictions were obtained through simple majority voting over the image-level predicted labels. While this choice provides a straightforward and interpretable aggregation scheme, it does not account for the confidence associated with each image-level prediction. As a result, highly confident predictions and uncertain predictions contribute equally to the final compartment-level decision. Future work could investigate confidence-weighted aggregation strategies, in which images associated with higher prediction confidence exert a stronger influence on the final compartment label. Such an approach may further improve robustness, particularly in the presence of ambiguous views or heterogeneous visual conditions within the same forest compartment.

6.4. Species-specific focus

In the absence of a standardized and widely accepted protocol for qualitative classification of standing trees, the adopted quality scheme was specifically designed for the Douglas fir stands considered in this study. Stem morphology, branching patterns, bark texture, and defect expression vary considerably across species and forest types. As a result, the trained model may not directly generalize to mixed stands, uneven-aged forests, or species with markedly different architectural traits. Its applicability to other species and forest types therefore requires further validation and potential adaptation.

Extending the dataset to include multiple species, diverse stand structures, and geographically distinct sites would allow evaluation of cross-domain generalization and domain adaptation strategies. Such expansion would also support the development of more universally applicable qualitative indices.

Moreover, although the visual classification was based on criteria designed to be as objective as possible, a certain degree of subjectivity cannot be entirely excluded. The potential impact of this subjectivity on model training and performance represents an additional aspect worthy of further investigation.

7. Conclusions

This study investigated the feasibility of applying Vision Transformers (ViTs) to the qualitative classification of terrestrial forest images, with the aim of integrating a stem quality indicator into conventional forest inventory descriptions. Using a dataset of smartphone-acquired images collected in Douglas fir forest compartments, the proposed approach classified stands according to the degree of branchiness, a key factor influencing timber quality and economic value.

The results demonstrated that ViT-based models can effectively perform qualitative forest classification even in the presence of limited and class-imbalanced training data. Among the evaluated adaptation strategies, parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) consistently outperformed full fine-tuning, highlighting the importance of preserving the generalization capability of the pre-trained backbone when operating in small-data regimes. The analysis of different ViT architectures further indicated that improvements in performance were driven more by input representation granularity than by increased model capacity, with larger models providing limited benefits relative to their computational cost.

From an operational perspective, aggregating predictions at the forest-compartment level significantly improved classification robustness compared to image-level predictions. This finding confirms the suitability of the proposed approach for practical forest inventory applications, where decisions are typically made at the stand level rather than for individual observations.

Although further validation across larger and more diverse datasets is required, the proposed method demonstrates that qualitative stem assessment can be derived from low-cost terrestrial imagery. This approach facilitates the integration of qualitative evaluations into conventional forest inventory surveys, thereby enhancing knowledge of stand conditions and providing valuable information to support forest management and timber utilization strategies based not only on quantity but also on wood quality.

CRediT authorship contribution statement

Niccolò Biondi: Writing – review & editing, Supervision, Formal analysis; **Simone Ricci:** Writing – review & editing, Writing – original draft, Formal analysis; **Niccolò Arati:** Writing – original draft, Software, Formal analysis; **Michela Nocetti:** Writing – review & editing, Writing – original draft, Project administration, Data curation, Conceptualization; **Giovanni Aminti:** Investigation, Data curation; **Pietro Pala:** Writing – review & editing, Supervision, Funding acquisition; **Michele Brunetti:** Writing – review & editing, Funding acquisition, Data curation, Conceptualization.

Data availability

The dataset is published on Zenodo, https://doi.org/10.5281/zenodo.18484448.

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.

Acknowledgment

The authors want to thank the Carabinieri Biodiversity Group of Vallombrosa and Davide Pozzi of Tenuta Podernovo for their collaboration during the field work.

Funding

This work was supported by the DIGIMEDFOR project, Digital tools and technology systems for the sustainable management of Mediterranean forest resources, European Union's HORIZON-CL62022-CIRCBIO-02-two-stage (Project No. 101081928).

References

[1] F. Negro, C. Cremonini, R. Zanuttini, CE marking of structural timber: the European standardization framework and its effects on Italian manufacturers, Drvna Industrija 64 (1) (2013) 55–62. https://doi.org/10.5552/drind.2013.1214

[2] M. Brunetti, P. Burato, C. Cremonini, F. Negro, M. Nocetti, R. Zanuttini, Visual and machine grading of larch (*Larix decidua* Mill.) Structural timber from the Italian Alps, Mater. Struct. 49 (7) (2016) 2681–2688. https://doi.org/10.1617/s11527-015-0676-5

[3] E. Macdonald, J. Hubert, A review of the effects of silviculture on timber quality of Sitka spruce, Forestry 75 (2) (2002) 107–138. https://doi.org/10.1093/forestry/75.2.107

[4] G. Murphy, D. Cown, J. Moore, Economics of Segregation Based on Wood Properties, Technical Report GCFF TN-03, Scion, 2015.

[5] G.E. Murphy, J.R. Moore, SEGMOD–a techno-economic model for evaluating the impact of segregation based on internal wood properties, Ann. For. Sci. 75 (3) (2018) 73. https://doi.org/10.1007/s13595-018-0755-1

[6] X. Wang, Recent advances in nondestructive evaluation of wood: in-forest wood quality assessments, Forests 12 (7) (2021) 949. https://doi.org/10.3390/f12070949

[7] A. Rais, H. Pretzsch, J.-W.G. Van De Kuilen, Roundwood pre-grading with longitudinal acoustic waves for production of structural boards, Eur. J. Wood Wood Prod. 72 (1) (2014) 87–98. https://doi.org/10.1007/s00107-013-0757-5

[8] J. Malinen, M. Haring, H. Kilpeläinen, E. Verkasalo, Comparison of alternative roundwood pricing systems - a simulation approach, Silva Fennica 49 (3) (2015) 1293. https://doi.org/10.14214/sf.1293

[9] J. Marenče, B. Šega, D. Gornik Bučar, Monitoring the quality and quantity of beech-wood from tree to Sawmill product, Croatian J. For. Eng. 41 (1) (2020) 119–128. https://doi.org/10.5552/crojfe.2020.613

[10] M. Balasso, M. Hunt, A. Jacobs, J. O'Reilly-Wapstra, Development of a segregation method to sort fast-grown Eucalyptus nitens (H. Deane & Maiden) Maiden plantation trees and logs for higher quality structural timber products, Ann. For. Sci. 79 (1) (2022) 9. https://doi.org/10.1186/s13595-022-01122-2

[11] M. Nocetti, G. Aminti, M. Vicario, M. Brunetti, Assessment of oak roundwood quality using photogrammetry and acoustic surveys, Forests 16 (3) (2025) 421. https://doi.org/10.3390/f16030421

[12] A. Ruano, I. Alberdi, P. Adame, D. Moreno-Fernández, A.C. Amiano, J. Fernández-Golfín, E. Hermoso, L. Hernández, E. Merlo, V. Sandoval, I. Cañellas, Improving stem quality assessment based on national forest inventory data: an approach applied to Spanish forests, Ann. For. Sci. 80 (1) (2023) 20. https://doi.org/10.1186/s13595-023-01187-7

[13] M. Bosela, J. Redmond, M. Kučera, G. Marin, R. Adolt, T. Gschwantner, R. Petráš, K. Korhonen, A. Kuliešis, G. Kulbokas, C. Fischer, A. Lanz, Stem quality assessment in European national forest inventories: an opportunity for harmonised reporting?, Ann. For. Sci. 73 (3) (2016) 635–648. https://doi.org/10.1007/s13595-015-0503-8

[14] M. Nocetti, M. Brunetti, Advancements in wood quality assessment: standing tree visual evaluation–a review, Forests 15 (6) (2024) 943. https://doi.org/10.3390/f15060943

[15] M. Rudnicki, X. Wang, R.J. Ross, R.B. Allison, K. Perzynski, Measuring Wood Quality in Standing Trees–A Review, Technical Report, U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, (2017).

[16] J. Iglhaut, C. Cabo, S. Puliti, L. Piermattei, J. O'Connor, J. Rosette, Structure from motion photogrammetry in forestry: a review, Current For. Reports 5 (3) (2019) 155–168. https://doi.org/10.1007/s40725-019-00094-3

[17] K. Kędra, I. Barbeito, M. Dassot, P. Vallet, A. Gazda, Single-image photogrammetry for deriving tree architectural traits in mature forest stands: a comparison with terrestrial laser scanning, Ann. For. Sci. 76 (1) (2019) 5. https://doi.org/10.1007/s13595-018-0783-x

[18] G. Krok, B. Kraszewski, K. Stereńczak, Application of terrestrial laser scanning in forest inventory - an overview of selected issues, For. Res. Pap. 81 (4) (2020) 175–194. https://doi.org/10.48538/FRP-2020-0021

[19] D. Xu, H. Wang, W. Xu, Z. Luan, X. Xu, LiDAR applications to estimate forest biomass at individual tree scale: opportunities, challenges and future perspectives, Forests 12 (5) (2021) 550. https://doi.org/10.3390/f12050550

[20] N. Niknejad, R. Bidese-Puhl, Y. Bao, K.G. Payn, J. Zheng, Phenotyping of architecture traits of loblolly pine trees using stereo machine vision and deep learning: stem diameter, branch angle, and branch diameter, Comput. Electron. Agric. 211 (2023) 107999. https://doi.org/10.1016/j.compag.2023.107999

[21] L.A. Wells, W. Chung, Real-time computer vision for tree stem detection and tracking, Forests 14 (2) (2023) 267. https://doi.org/10.3390/f14020267

[22] A. Sandim, M. Amaro, M.E. Silva, J. Cunha, S. Morais, A. Marques, A. Ferreira, J.L. Lousada, T. Fonseca, New technologies for expedited forest inventory using smartphone applications, Forests 14 (8) (2023) 1553. https://doi.org/10.3390/f14081553

[23] R. Magnuson, Y. Erfanifard, M. Kulicki, T.A. Gasica, E. Tangwa, M. Mielcarek, K. Stereńczak, Mobile devices in forest mensuration: a review of technologies and methods in single tree measurements, Remote Sens. 16 (19) (2024) 3570. https://doi.org/10.3390/rs16193570

[24] M. Vastaranta, E. Latorre, V. Luoma, N. Saarinen, M. Holopainen, J. Hyyppä, Evaluation of a smartphone app for forest sample plot measurements, Forests 6 (4) (2015) 1179–1194. https://doi.org/10.3390/f6041179

[25] V. Vähä-Konka, L. Korhonen, K. Kärhä, M. Maltamo, Estimating the accuracy of smartphone app-based removal estimates against actual wood-harvesting data from clear cuttings, iForest - Biogeosci. For. 17 (3) (2024) 140–147. https://doi.org/10.3832/ifor4377-017

[26] M. Nocetti, G. Aminti, M. Brunetti, G. Fontani, S. Kivimaki, T. Rouvinen, L. Saulino, Early wood quality assessment for a better use of the forest resource. The DigiMedFor project, in: Proceedings of the 26th International Wood Machining Seminar, 14-15 April 2025, Firenze, 2025, pp. 290–292.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations (ICLR), 2021. arXiv:2010.11929.
[28] J. Maurício, I. Domingues, J. Bernardino, Comparing vision transformers and convolutional neural networks for image classification: a literature review, Appl. Sci. 13 (9) (2023) 5521. https://doi.org/10.3390/app13095521
[29] D. Joshi, C. Witharana, Vision transformer-based unhealthy tree crown detection in mixed Northeastern US forests and evaluation of annotation uncertainty, Remote Sens. 17 (6) (2025). https://doi.org/10.3390/rs17061066
[30] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, A. Doulamis, A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring, IEEE Trans. Neural Netw. Learn. Syst. 34 (7) (2023) 3299–3307. https://doi.org/10.1109/TNNLS.2022.3144791
[31] R. Ghali, M.A. Akhloufi, M. Jmal, W. Souidene Mseddi, R. Attia, Wildfire segmentation using deep vision transformers, Remote Sens. 13 (17) (2021) 3527. https://doi.org/10.3390/rs13173527
[32] T. Chang, K. Ndegwa, A. Gros, V.A. Landau, L.J. Zachmann, B. State, M.A. Gritts, C.W. Miller, N.E. Rutenbeck, S. Conway, G. Bayes, VibrantVS: a high-resolution vision transformer for forest canopy height estimation, Remote Sens. 17 (6) (2025). https://doi.org/10.3390/rs17061017
[33] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, W. Wang, Y. Chen, LoRA: low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2022).
[34] M. Yang, J. Chen, J. Tao, Y. Zhang, J. Liu, J. Zhang, Q. Ma, H. Verma, R. Zhang, M. Zhou, I. King, R. Ying, Low-rank adaptation for foundation models: a comprehensive review, 2025, arXiv:2501.00365v2
[35] I. Ulku, O. Ozgur Tanriover, E. Akagündüz, LoRA-NIR: low-rank adaptation of vision transformers for remote sensing with near-infrared imagery, IEEE Geosci. Remote Sens. Lett. 21 (2024) 1–5. https://doi.org/10.1109/LGRS.2024.3449372
[36] B. Xue, H. Cheng, Q. Yang, Y. Wang, X. He, Adapting segment anything model to aerial land cover classification with low-rank adaptation, IEEE Geosci. Remote Sens. Lett. 21 (2024) 1–5. https://doi.org/10.1109/LGRS.2024.3357777
[37] EN , 1927-3 Qualitative classification of softwood round timber. Part 3: Larches and Douglas fir, European standard, CEN, Bruxelles, 2008.
[38] M. Nocetti, G. Aminti, G. Fontani, B. Kumar, M. Brunetti, Landscape images of Douglas fir forests classified for stem quality, Zenodo (2026). https://doi.org/10.5281/zenodo.18484448
[39] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30. http://jmlr.org/papers/v7/demsar06a.html.
[40] H. Wang, Z. Chen, B. Yang, Score-CAM: score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 24–25.
[41] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[42] R. Reedha, E. Dericquebourg, R. Canals, A. Hafiane, Transformer neural network for weed and crop classification of high resolution UAV images, Remote Sens. 14 (3) (2022) 592. https://doi.org/10.3390/rs14030592
[43] Y. Zhu, Z. Shen, Z. Zhao, S. Wang, X. Wang, X. Zhao, D. Shen, Q. Wang, MeLo: low-rank adaptation is better than fine-tuning for medical image diagnosis, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 2024, pp. 1–5. https://doi.org/10.1109/ISBI56570.2024.10635615