Beyond Image-Level Evaluation: Specimen-Aware Multi-View Feature Fusion for Wood Charcoal Classification

Dieu-Donné Fangnon[1], Diane Lingrand[1], Marco Corneli[2,3], Isabelle Théry-Parisot[2], and Antoine Pasqualini[2]

[1] I3S, CNRS, Univ. Côte d'Azur, INRIA - Maasai team, France
[2] LJAD, CNRS, Univ. Côte d'Azur, INRIA - Maasai team, France
[3] CEPAM, CNRS, Univ. Côte d'AzurNice, France

Abstract. Automated identification of wood species from microscopic images of charcoal is a key challenge in anthracology, with applications in archaeology, ecology, and forest heritage studies. Existing deep learning approaches for this task suffer from inflated accuracy estimates caused by data leakage, images from the same physical specimen appearing in both training and test sets, making it difficult to assess true generalization. We address this by proposing a specimen-level, leakage-free evaluation protocol based on stratified group 5-fold cross-validation, and systematically study multi-view feature aggregation strategies over three anatomical sections using frozen convolutional backbones. On a curated dataset of four gymnosperm species (81 trees, 897 images), our learned MLP-fusion with EfficientNet-V2L achieves 91.32±0.08% specimen-level accuracy on 500×500 crops, outperforming both statistical pooling baselines and self-supervised Vision Transformers. Preliminary experiments on an ongoing 21-species dataset (520 trees) yield 81.73% specimen-level and 94.42% family-level accuracy, suggesting that the proposed pipeline generalizes to more realistic, large-scale settings.

Keywords: Computer Vision · Deep Learning · Classification · Anthracology · Features fusion· Leakage-free.

1 Introduction

Anthracology, the study of archaeological and sedimentary charcoal, provides irreplaceable evidence for reconstructing past vegetation, human land use, and climate change over millennia. A central task in this discipline is the taxonomic identification of wood species from microscopic images of charcoal fragments, typically examined across three anatomical sections: radial, tangential, and transversal. Today, this identification is performed manually by trained experts, a process that is time-consuming, requires years of specialist training, and is difficult to scale many of fragments typically recovered from a single excavation site. Different specimens from the same species cannot be recognized by the

anthracologists. Automating this process would not only accelerate archaeological analysis but also improve reproducibility and open the door to large-scale ecological and paleoenvironmental studies.

Unlike fresh wood samples, charcoal fragments present specific visual challenges: the carbonization process degrades fine anatomical structures, alters surface texture, and introduces artifacts such as cracks and shrinkage deformations. This makes charcoal classification a genuinely difficult fine-grained recognition problem, distinct from the better-studied problem of fresh wood identification, and one where visual representations learned on natural images may not transfer straightforwardly.

In recent years, deep learning has significantly advanced wood and charcoal species recognition. Early work relied on handcrafted features and classical machine learning pipelines [1, 6], which have since been surpassed by Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) across a range of tasks [2, 12, 13, 21]. A key insight in this literature is that exploiting the multi-view nature of the data, combining features from all three anatomical sections of a specimen consistently improves accuracy over single-section approaches [14, 15, 22]. However, a critical and underappreciated flaw affects most reported results in this field: the train/test split is typically performed at the image level rather than at the specimen level. As a consequence, images from the same physical tree may appear in both training and test sets [7–9], artificially inflating accuracy estimates and preventing any reliable assessment of generalization to unseen specimens [16]. We refer to this phenomenon as data leakage, and we argue it is the primary reason why many published results cannot be meaningfully compared or reproduced.

As part of the AIWOOD project, this work focuses on a new, large and structured database of European modern wood charcoals, still under construction. While the ultimate goal of this research is the identification of archaeological charcoal specimens, these are often scarce, highly degraded by centuries of environmental constraints, and not yet fully digitized, also without labels. Therefore, we first establish a robust framework using modern charcoal samples where we have access to the ground truth, to develop a generalizable model, providing a foundation for future domain adaptation to archaeological data. In collaboration with anthracologists, we were provided with a selection of four softwood species that remain particularly challenging to distinguish using traditional identification methods. Consequently, our study focuses on these type-species, selecting those represented by at least 16 individual trees to ensure statistical robustness (Table 1). This application has the specificity of a varying number of images per section and per specimen: a fusion mechanism is thus needed and must be invariant to permutations of the cropping order (images) and sectional views (specimens). We address both the evaluation problem and the fusion problem in a unified experimental study. Working with a curated dataset of four gymnosperm species spanning 81 trees and 897 high-resolution micrographs, we systematically compare multi-view feature aggregation strategies built on top of frozen ImageNet-pretrained convolutional backbones, under a strict leakage-free

evaluation protocol. We also report preliminary results on an ongoing 21-species extension of the dataset, as first evidence of scalability. This work makes the following contributions:

We enforce stratified group 5-fold cross-validation at the specimen level, ensuring that no images from the same tree appear across training and test folds. This yields realistic generalization estimates and exposes the degree to which prior results may be optimistic. A systematic study of multi-view feature aggregation. We compare three families of aggregation strategies over the three anatomical sections: a majority-voting baseline, non-learnable statistical pooling (mean, max, standard deviation), and a learnable permutation-invariant MLP-fusion. All strategies are evaluated under identical conditions, providing a controlled and reproducible comparison. Under our evaluation protocol, learned MLP-fusion with EfficientNet-V2L achieves 91.32±0.08% specimen-level accuracy on 500×500 crops, substantially outperforming the voting baseline and statistical pooling, and exceeding self-supervised Vision Transformer features. Preliminary results on 21 species suggest the pipeline generalizes to a more realistic and challenging setting.

2 Materials and Methods

2.1 Dataset

2.1.1 Data Collection and Composition: This work uses images from an ongoing structured database of European wood charcoals, built in collaboration with expert anthracologists under the AIWOOD project. For our primary experiments, we used the four gymnosperm species provided by anthracologists, with the number of available specimens, applying a minimum inclusion threshold of 16 individuals per species to ensure sufficient statistical power for cross-validation. The resulting dataset comprises 81 trees across four species. For each specimen, microscopic images were acquired across three anatomical sections, radial (R), tangential (T), and transversal (Tr) yielding 897 high-resolution micrographs in total with different original sizes (2880×2048, 2559×1917, 1023×724, 1021×745, 949×744). The full breakdown by species and section is provided in Table 1. Each specimen contributes a variable number of images per section, reflecting the practical constraints of microscopic acquisition: not every fragment yields images of equal quality or quantity across all three planes. This variability motivates the need for aggregation strategies that are invariant in permutation of images and adapted to a variable number of inputs.

2.1.2 Preprocessing and Crop Generation: Raw micrographs vary substantially in resolution and aspect ratio. Rather than resizing entire images to a fixed resolution which would discard fine anatomical texture information critical for species discrimination, we tile each image into non-overlapping square crops. We evaluate four tiling configurations: the full resized original image, and crops of 224×224, 300×300, and 500×500 pixels, all converted to grayscale. These sizes

correspond to standard input resolutions of the pretrained backbones used in our experiments, allowing us to feed crops directly without upsampling. The full resized image is included as an additional baseline condition to assess the effect of resolution loss. In our results tables, each crop size represents an independent experimental condition, and comparisons across crop sizes should be interpreted accordingly rather than as a direct head-to-head evaluation.

2.1.3 Evaluation Protocol: All experiments are evaluated using stratified group 5-fold cross-validation at the specimen level. Specifically, the folds are constructed such that all images from a given tree are assigned exclusively to either the training set or the test set, never both. Stratification ensures that the class distribution is approximately balanced across folds. This protocol prevents data leakage, a well-documented issue in specimen-based classification tasks [16] where image-level splits allow the model to exploit low-level appearance correlations between images of the same individual. We report the mean accuracy and standard deviation across the five folds.

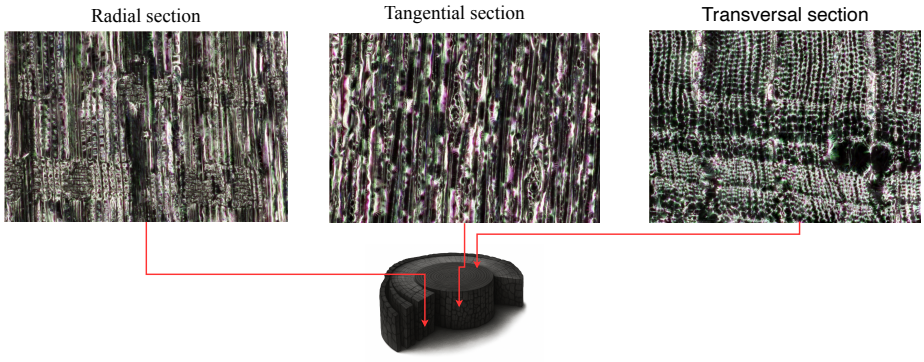


Fig. 1: Three anatomical sections of a wood charcoal sample from the *Picea abies* species.

2.2 Methods

Our pipeline operates at the specimen level: given a set of microscopic images from the three anatomical sections of a single tree, it produces a single species prediction. Figure 2 provides an overview of the full pipeline. In the followings subsections, we describe its components in order: feature extraction, the voting baseline, non-learnable statistical aggregation, learnable MLP-fusion, and classification part.

Table 1: Preliminary experiment on 4 species: Number of specimens (trees) per species and number of images for each anatomical section. **R:** Radial, **T:** Tangential, **Tr:** Transversal.

Species	Nb. Trees	Images/Section			Nb. Img
		R	T	Tr	
Picea abies	20	95	48	50	193
Juniperus communis	19	143	81	86	310
Larix decidua	19	62	41	40	143
Pinus sylvestris	23	121	65	65	251
Total	**81**				**897**

2.2.1 Feature Extraction Each crop image is passed through a frozen convolutional backbone pretrained on ImageNet, followed by Global Average Pooling (GAP). GAP computes the spatial average of each feature map produced by the final convolutional layer, yielding a fixed-dimensional vector regardless of the spatial dimensions of the input. Formally, for a feature map of size $H \times W \times D$, GAP produces a vector in $\mathbb{R}^D$ by averaging over the $H \times W$ spatial positions. This produces a compact, translation-invariant representation that summarizes the global texture and structural content of a crop, which is well-suited to the fine-grained micro-texture patterns present in charcoal micrographs. We evaluate four backbone architectures: ResNet-50 [3], EfficientNet-B7 [19], EfficientNet-V2L [20], and ConvNeXt-XLarge [5]. These were selected based on their strong performance in fine-grained visual recognition and transfer learning benchmarks, and their coverage of a range of architectural scales and inductive biases. All backbones are kept frozen throughout training: only the aggregation and classification components are optimized. This choice is motivated by the small size of our dataset, where full fine-tuning risks overfitting. We additionally evaluate three self-supervised Vision Transformer (ViT) backbones from the DINOv2 [10] and DINOv3 [17] families (table 4), using their CLS token embedding as the crop-level feature vector. These are included to assess whether large-scale self-supervised pretraining provides competitive representations for this domain.

2.2.2 Voting Baseline As a first baseline, we train a classification head consisting of two fully connected layers (256 and 4 neurons respectively) on top of the frozen GAP features of individual crops. At inference time, each crop produces an independent class prediction, and the specimen-level decision is obtained by majority voting across all crops from all sections of that tree. This approach treats each crop independently and ignores the multi-view structure of the data entirely. As shown in section 3.1, this yields poor specimen-level accuracy, motivating the aggregation strategies described below.

2.2.3 Fusion methods As explained in previous section 2.1, we need a aggregation method that can take variable number of inputs and that is permutation invariant. The first method is simple and do not need any learning while the

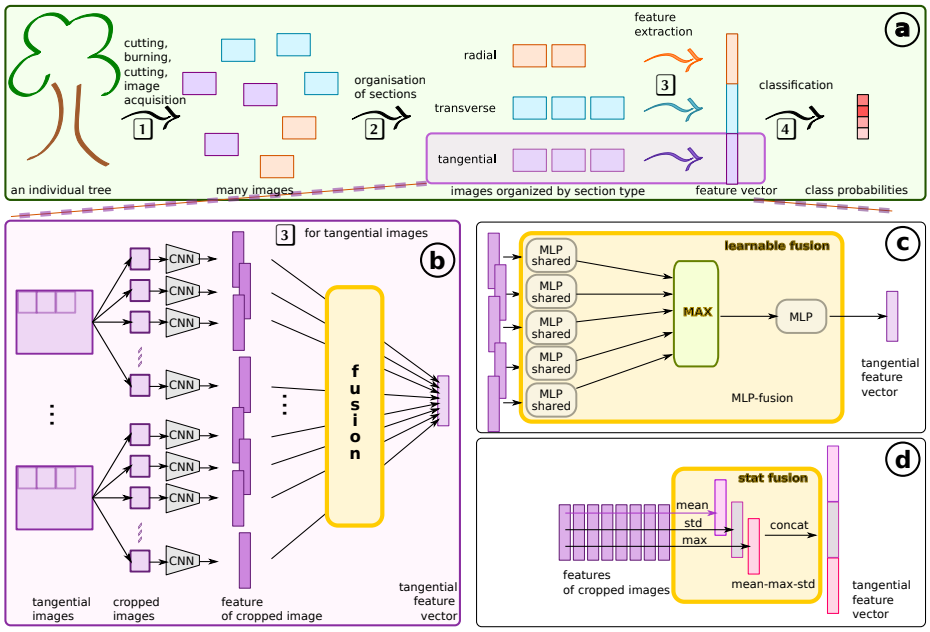


Fig. 2: Global pipeline: a) general process, b) features extraction and fusion process from a single section, c) MLP-fusion process, d) Stat fusion process.

second has been inspired by PointNet [11] making fusion of variable number of 3D points in order to classify 3D point clouds. The fusion algorithm of PointNet needed some adaptation for our problem.

Non-Learnable Statistical Aggregation Our first family of aggregation strategies summarizes the set of GAP feature vectors from all crops of a given specimen and section using fixed, permutation-invariant statistics. Let $f_{ijk}^{(n)} \in \mathbb{R}^D$ denote the GAP feature from backbone n, specimen j, anatomical section $i \in \{1, 2, 3\}$, and crop index k. We aggregate the set $\{f_{ijk}^{(n)}\}_k$ into a single section descriptor using three statistics:

$$\mu_{ij}^{(n)} = \text{mean}_k\{f_{ijk}^{(n)}\}, \quad \delta_{ij}^{(n)} = \max_k\{f_{ijk}^{(n)}\}, \quad \sigma_{ij}^{(n)} = \text{std}_k\{f_{ijk}^{(n)}\} \tag{1}$$

Two aggregation variants are evaluated. In **single-statistic aggregation**, one statistic (mean, max, or std) is selected and the three section descriptors are concatenated to form the specimen representation $z_j^{(n)} = [\mu_{1j}^{(n)} \parallel \mu_{2j}^{(n)} \parallel \mu_{3j}^{(n)}] \in \mathbb{R}^{3D}$. In **three-statistic aggregation** (denoted mean-max-std), all three statistics are concatenated within each section to form a richer descriptor $u_{ij}^{(n)} = [\mu_{ij}^{(n)} \parallel \delta_{ij}^{(n)} \parallel \sigma_{ij}^{(n)}] \in \mathbb{R}^{3D}$, which are then concatenated across sections: $z_j^{(n)} \in \mathbb{R}^{9D}$. Both variants are permutation-invariant with respect to the ordering of crops within a section and accept a variable number of crops per specimen.

Learnable MLP-Fusion Our second aggregation strategy learns a permutation-invariant encoder that maps a variable-size set of crop embeddings to a single section descriptor (fig. 2-c). For a given anatomical section, the input is a set $X = \{x_i\}_{i=1}^{N} \subset \mathbb{R}^D$ of N, GAP feature vectors, where N varies across specimens. The encoder proceeds in three steps: (i) a shared residual MLP is applied independently to each crop embedding; (ii) the transformed embeddings are aggregated with a max pooling operator across crops; (iii) a second residual MLP refines the pooled summary into the final section descriptor. Each MLP consists of two linear layers (dimensions $2D$ and D respectively), ReLU activations, and a residual connection from input to output. The same encoder is applied independently to each of the three anatomical sections, producing section descriptors $Z_{\text{rad}}, Z_{\text{tan}}, Z_{\text{tra}} \in \mathbb{R}^D$, which are then concatenated into a single specimen representation $Z_j \in \mathbb{R}^{3D}$. This design is permutation-invariant with respect to both the ordering of crops within a section.

2.2.4 Classification Head and Ensemble The specimen representation produced by either aggregation strategy is passed to a classification head (fig. 3). During training of the learnable fusion, a shallow softmax head is used to provide supervision. At evaluation time, we replace this head with an ensemble of four classifiers on the fused Z features: a linear SVM, PCA followed by a linear SVM, Logistic Regression, and a Random Forest, whose predictions are combined by majority voting. This replacement is motivated by an empirical finding: the learned specimen representations are highly linearly separable, and the ensemble of simple linear and tree-based classifiers consistently outperforms the neural softmax head at test time. This is consistent with observations in the transfer learning literature that deep features from strong pretrained backbones tend to be linearly separable even for out-of-domain tasks. Only ensemble classifier results are therefore reported in tables tables 3 and 4.

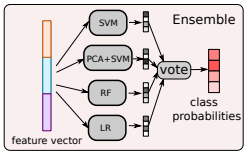


Fig. 3: Ensemble classification.

3 Results and Discussions

We organize our results in four parts: baseline voting results (Section 3.1), multi-view aggregation results on the four-species dataset (Section 3.2), a comparison with self-supervised Vision Transformer features (Section 3.3), and preliminary results on the ongoing 21-species extension (Section 3.4). All results are reported as mean accuracy and standard deviation across the five folds of the stratified group cross-validation protocol described in Section 2.1.

3.1 Baseline Results

Table 2 reports specimen-level accuracy for the voting baseline across all backbone and crop size combinations. Accuracy ranges from approximately 45%

to 61%, with the best result obtained by EfficientNet-V2L on 500×500 crops (60.69±0.01%). Several observations are worth noting. First, performance is strongly coupled to the backbone-crop-size pairing. Larger crops consistently produce higher accuracy across most backbones, which is consistent with the hypothesis that fine anatomical texture cues, the primary discriminative signal in charcoal micrographs are preserved better at higher resolution and lost when images are aggressively downsampled. This motivates our choice to evaluate all subsequent methods at all four crop sizes rather than selecting a single configuration. Second, the overall accuracy level hovering near chance for a four-class problem at smaller crop sizes demonstrates that treating each crop independently and aggregating predictions by voting is fundamentally insufficient for this task. This motivates the feature-level aggregation strategies described in sections 2.2 and 2.2. Third, we note that different backbones achieve their best performance at different crop sizes. This reflects differences in the inductive biases and receptive field sizes of the architectures rather than an unfair comparison: each backbone-crop-size pair represents an independent experimental condition, and we report all combinations to provide a complete picture.

Table 2: Mean accuracy (%) and standard deviations (±) across the 5 folds of the baseline method. The best results on each cropping dimension are highlighted in bold. **Original** = resized original image.

Crop Dim.	Backbone			
	ResNet-50	EffNet-B7	EffNet-V2L	ConvNeXt-XL
Original	**52.67±0.06**	48.57±0.09	51.02±0.01	51.95±0.03
224×224	45.23±0.04	54.78±0.10	**55.55±0.11**	50.34±0.08
300×300	50.67±0.07	**59.35±0.08**	58.09±0.10	55.75±0.09
500×500	51.45±0.08	59.75±0.10	**60.69±0.01**	52.14±0.09

3.2 Multi-View Feature Aggregation Results

Table 3 reports specimen-level accuracy for all aggregation strategies, single-statistic (mean, max, std), three-statistic (mean-max-std), and learnable MLP-fusion across all backbones and crop sizes. Both families of aggregation strategy substantially outperform the voting baseline across all conditions, confirming that feature-level aggregation over the three anatomical sections is the primary driver of performance in this task.

At 500×500 crops with EfficientNet-V2L, MLP-fusion achieves 91.32±0.08%, closely followed by mean-max-std at 90.15±0.09%. The gap between learnable and non-learnable aggregation narrows at smaller crop sizes, suggesting that the benefit of learned fusion is most pronounced when the individual crop embeddings are richer consistent with larger crops preserving more discriminative texture information. Across all crop sizes, the mean-max-std strategy consistently

outperforms single-statistic variants, confirming that capturing the distribution of crop features (through the combination of mean, max, and standard deviation) is more informative than any single statistic alone.

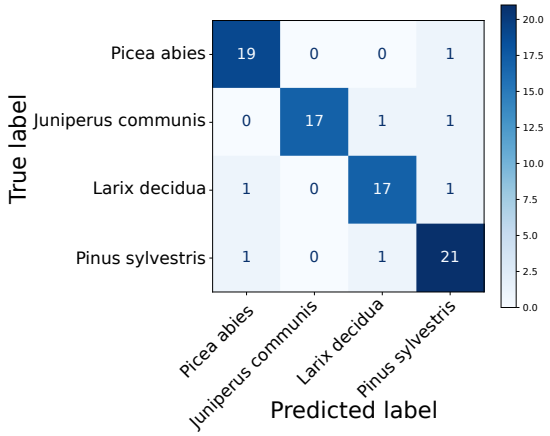


Fig. 4: Confusion matrix for the best MLP-fusion result across the 5 test folds for the 4 species. No systematic confusion between any pair of species is observed (even the 2 difficult one).

EfficientNet-V2L is the strongest backbone across most conditions, followed by ConvNeXt-XLarge and EfficientNet-B7, with ResNet-50 consistently lagging behind. This ranking is broadly consistent with ImageNet benchmark performance, suggesting that backbone capacity and training recipe matter even when the backbone is frozen and the downstream domain differs substantially from natural images. The 500×500 configuration yields the best results for both MLP-fusion and mean-max-std, across most backbones.

Figure 4 shows the confusion matrix for the best MLP-fusion configuration across the five test folds. Errors are rare and largely isolated: no systematic confusion between any pair of species is observed, and predictions are strongly concentrated on the diagonal. The few misclassified specimens are distributed across all species without a clear pattern, suggesting that residual errors correspond to genuinely ambiguous specimens, for instance, fragments where carbonization has severely degraded discriminative anatomical features rather than systematic model failures.

Sun et al. [18], Zielinski et al. [22] and Huang et al. [4] report competitive results on wood identification benchmarks, but on datasets that differ in species composition, imaging conditions, and evaluation protocol, making direct comparison unreliable. We therefore present our results as a reproducible baseline under a rigorous evaluation protocol rather than as a definitive state-of-the-art

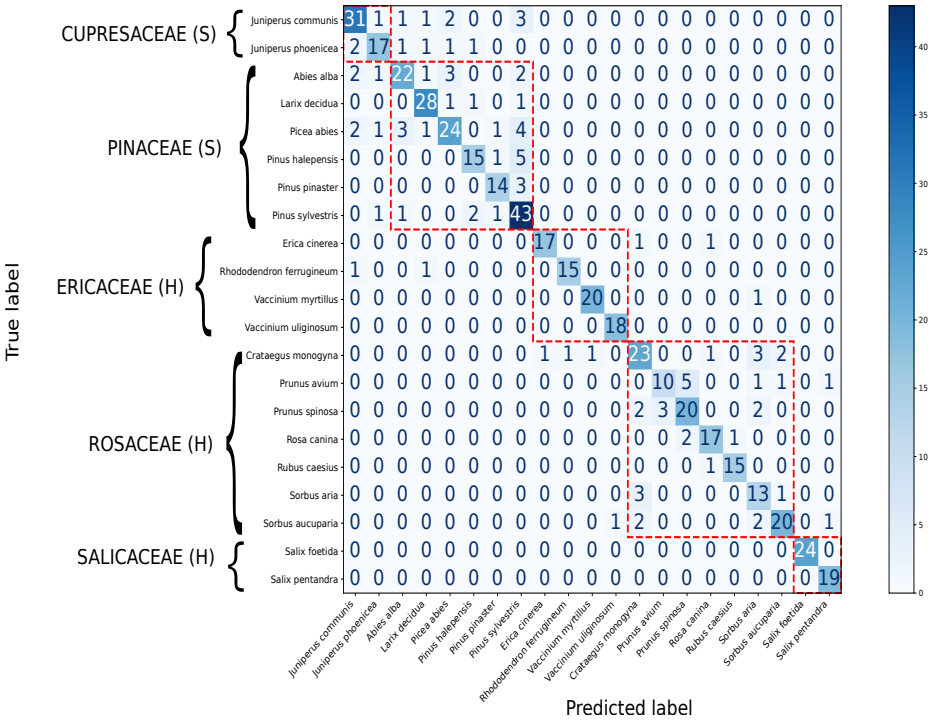
claim, and encourage future work to adopt the same evaluation standard to enable meaningful cross-paper comparison.

3.3 Comparison with Self-Supervised Vision Transformer Features

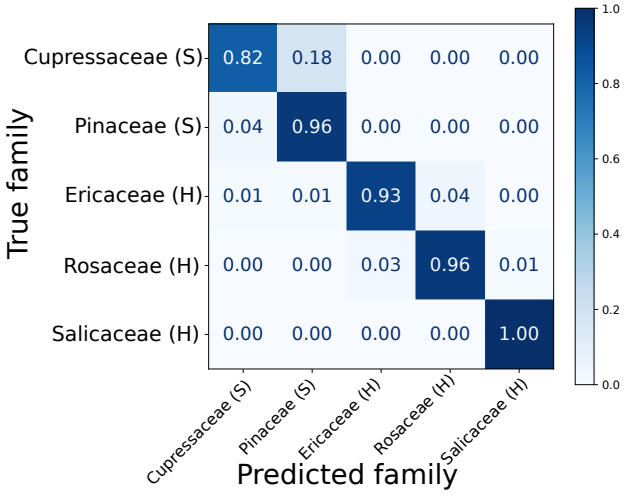
Table 4 reports results for three self-supervised ViT variants from the DINOv2 and DINOv3 families, evaluated with MLP-fusion at 224×224 and 518×518 crop sizes. The best ViT result is obtained by DINOv3-ViTb16(lvd) at 224×224 crops (83.97±0.10%), which remains below the best CNN-based result of 91.32±0.08%. This performance gap warrants careful interpretation: self-supervised objectives may produce representations that are less specialized for fine-grained texture discrimination; the CLS token embedding used by DINO models may capture less local texture information than the GAP features of CNNs; or the domain gap between natural images and charcoal micrographs may affect the two architectures differently. In the meantime, our results suggest that, for the specific task of fine-grained texture classification in charcoal micrographs, frozen convolutional features from large supervised backbones provide a stronger starting point than frozen self-supervised transformer features.

3.4 Preliminary Results on the 21-Species Extension

To probe the generalization of our best-performing pipeline to a more realistic and challenging setting, we apply it without modification to a preliminary version of an ongoing 21-species dataset comprising 520 trees (4,677 original images) across 14 genera and 5 families, spanning both softwood and hardwood categories. We emphasize that this dataset is still under active construction: species counts and specimen numbers will grow in future releases, and the results reported here should be interpreted as preliminary evidence of scalability rather than definitive performance figures. Under the same stratified group 5-fold cross-validation protocol, MLP-fusion with EfficientNet-V2L on 500×500 crops achieves 81.73±0.02% specimen-level accuracy and 94.42% family-level accuracy. The species-level confusion matrix (Figure 5a) reveals that the large majority of errors are confined within family blocks: species from the same botanical family are occasionally confused with one another, while cross-family confusions are rare. This is ecologically meaningful, species within the same family share more anatomical features than those from different families and is further confirmed by the family-level confusion matrix (Figure 5b), where all five families achieve above 82% recall with Salicaceae reaching perfect classification. These preliminary results suggest two encouraging conclusions. First, the pipeline scales gracefully from 4 to 21 species without any architectural modification or hyperparameter re-tuning, indicating that the learned aggregation strategy captures genuinely discriminative specimen-level features rather than overfitting to the small four-species setting. Second, the high family-level accuracy suggests that the model has implicitly learned a representation that respects the taxonomic hierarchy of the data, which may be exploited in future work through hierarchical classification objectives or taxonomically-informed loss functions.



(a) Species-level confusion matrix. Each red box indicates species grouped by their specific family. S/H = Soft/Hardwood.



(b) Family-level confusion matrix. We observe that the confusion probabilities between families are rare and very negligible. S/H = Soft/Hardwood.

Table 3: Mean accuracy (%) and standard deviations (±) across folds of the ensemble results using each fusion method on $(R + Tan + Tr)$ GAP features, each dataset, and backbone. The best results on each cropping dimension and fusion method are highlighted in bold. **Original** = resized original image.

Crop Dim.	Fusion method	Backbone			
		ResNet-50	EffNet-B7	EffNet-V2L	ConvNeXt-XL
Original	Mean	75.37±0.08	83.90±0.06	79.12±0.05	86.40±0.07
	Max	79.04±0.08	86.40±0.04	80.29±0.09	86.40±0.07
	Std	75.37±0.08	85.22±0.07	79.12±0.09	76.54±0.08
	mean-max-std	77.79±0.09	**86.40±0.04**	79.12±0.09	85.15±0.08
	MLP-fusion	77.87±0.08	82.72±0.07	85.15±0.05	76.54±0.09
224×224	Mean	80.22±0.04	83.97±0.11	84.04±0.11	80.22±0.07
	Max	85.07±0.10	81.54±0.13	85.07±0.10	80.15±0.10
	Std	86.32±0.12	83.97±0.09	83.90±0.10	80.22±0.08
	mean-max-std	86.32±0.12	**86.40±0.12**	85.15±0.11	81.40±0.11
	MLP-fusion	81.47±0.08	79.56±0.11	81.54±0.08	76.54±0.09
300×300	Mean	80.22±0.08	85.15±0.14	**86.47±0.09**	81.54±0.10
	Max	86.32±0.12	81.54±0.13	85.15±0.11	77.79±0.08
	Std	85.15±0.11	83.97±0.11	82.72±0.09	81.47±0.08
	mean-max-std	85.15±0.11	85.22±0.12	83.97±0.11	82.65±0.09
	MLP-fusion	80.22±0.07	75.20±0.04	83.82±0.10	75.37±0.11
500×500	Mean	81.54±0.07	86.32±0.12	85.29±0.12	82.72±0.09
	Max	85.15±0.09	83.82±0.10	87.65±0.10	83.90±0.10
	Std	81.54±0.08	85.07±0.10	86.40±0.08	77.87±0.09
	mean-max-std	85.15±0.09	85.07±0.10	90.15±0.09	83.90±0.13
	MLP-fusion	80.22±0.07	77.79±0.06	**91.32±0.08**	75.29±0.06

Table 4: Mean accuracy (%) and standard deviations (±) across folds of ensemble using DINO's features and **MLP-fusion**. DINOv2-ViTb14, DINOv3-ViTb16(lvd) and DINOv3-ViTl16(sat) use a 'Base (b)' and 'Large (l)' architecture with 14 and 16-pixel patches respectively. Both Dinov3 are pretrained on the lvd (Large Video Dataset) and (sat) satellite imagery.

Crop Dim.	Backbone		
	DINOv2-ViTb14	DINOv3-ViTb16(lvd)	DINOv3-ViTl16(sat)
224×224	80.15±0.10	**83.97±0.10**	73.97±0.07
518×518	**81.32±0.13**	80.15±0.10	78.90±0.06

3.5 How much does data leakage inflate accuracy in this field?

Table 5 provides a direct and striking quantification of the data leakage effect described in section 2.1. When the same EfficientNet-V2L backbone and 500×500 crops are evaluated using a standard image-level 5-fold cross-validation on the same four species dataset, the protocol used in the majority of prior work in this field, accuracy rises dramatically across all fusion strategies, ranging from 94.82% (max pooling) to 98.98% (MLP-fusion), compared to 87.65% and 91.32% respectively under our specimen-level protocol. This gap of approximately 7 percentage points represents a systematic and reproducible overestimation of generalization performance: under image-level splits, crops from the same physical tree appear in both training and test folds, allowing the model to recognize individual specimen characteristics rather than genuinely discriminative species-level features, thereby artificially inflating accuracy to near-ceiling values. Crucially, this inflation is not specific to any particular fusion strategy, it affects all five methods evaluated here by a consistent and substantial margin, confirming that the effect is structural and inherent to the evaluation protocol rather than a property of any individual model. These results provide concrete empirical support for our central methodological argument: published accuracy figures obtained under image-level evaluation in this domain should be interpreted with caution, and specimen-level stratified cross-validation is a necessary condition for any meaningful assessment of generalization in wood and charcoal species recognition.

Table 5: Leakage mean accuracy (%) and standard deviations (±) across standard 5-fold cross-validation with the EffNet-V2L features on the same four species dataset.

Crop Dim.	Fusion method				
	mean	max	std	mean-max-std	**MLP-fusion**
500×500	97.93±0.02	94.82±0.00	97.67±0.01	98.18±0.01	**98.98±0.02**

4 Conclusion

This work set out to address two interconnected problems in the automatic identification of wood species from microscopic charcoal images: the widespread use of evaluation protocols that overestimate generalization due to data leakage, and the lack of a systematic, controlled comparison of multi-view feature aggregation strategies under rigorous experimental conditions. On both fronts, our results are encouraging. Enforcing specimen-level cross-validation, ensuring that no images from the same tree appear across training and test folds, yields accuracy estimates that are substantially lower than those reported in most prior work, but considerably more meaningful as indicators of real-world generalization. Under

this protocol, learned MLP-fusion over GAP features from three anatomical sections with EfficientNet-V2L achieves 91.32% specimen-level accuracy on a four-species gymnosperm dataset, outperforming both statistical pooling baselines and self-supervised Vision Transformer features. The confusion matrix analysis suggests that residual errors are largely attributable to genuinely ambiguous specimens rather than systematic model weaknesses, which is a healthy sign for a pipeline intended for practical deployment. The preliminary 21-species results reinforce these findings. Without any architectural modification or tuning, the same pipeline achieves 81.73% specimen-level and 94.42% family-level accuracy on a considerably more challenging and taxonomically diverse dataset. The observation that errors are largely confined within botanical family boundaries is both ecologically interpretable and methodologically reassuring: the model has implicitly learned a representation that respects the hierarchical structure of the data. Future work will focus on extending this work on the new growing dataset by incorporate more classes and more data and explanation of the decision before an adaptation to ancient charcoal.

Acknowledgments. This work was carried out as part of the AIWOOD project, funded under grant ANR-23-CE38-0013. The authors are grateful to the OPAL infrastructure from University Côte d'Azur for providing resources and support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data Availability The dataset supporting the results reported in this paper is partially available upon reasonable request for research purposes. The complete dataset will be made publicly available upon completion of the project, in accordance with the data management plan of the funding institution by the end of 2027.

References

1. Filho, P., Soares de Oliveira, L., Nisgoski, S., Jr, A.: Forest species recognition using macroscopic images. MVA **25** (05 2014)
2. Hafemann, L.G.: Forest species recognition using deep convolutional neural networks. In: ICPR (09 2014)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
4. Huang, P., Zhao, F., Li, X., Wu, Z., Zhu, Z., Zhang, Y.: Variant transfer learning for wood recognition. In: 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics). pp. 743–748. IEEE (2020)
5. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s (2022)
6. Martins, J., Soares de Oliveira, L., Nisgoski, S.: A database for automatic classification of forest species. MVA **24** (04 2012)

7. Maruyama, T., Oliveira, L., Britto, A., Nisgoski, S.: Automatic classification of native wood charcoal. Ecological Informatics **46**, 1–7 (2018)
8. Menon, L.T., Laurensi, I.A., Penna, M.C., Oliveira, L.E.S., Britto, A.S.: Data augmentation and transfer learning applied to charcoal image classification. In: 2019 International Conference on Systems, Signals and Image Processing (IWSSIP). pp. 69–74 (2019)
9. Rodrigues de Oliveira Neto, R., Rodrigues Moreira, L., Mari, J., Naldi, M., Milagres, E., Vital, B., Carneiro, A., Binoti, D., Lopes, P., Leite, H.: Automatic identification of charcoal origin based on deep learning. Maderas: Ciencia y Tecnología **23** (09 2021)
10. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
11. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
12. Ravindran, P., Costa, A., Soares, R., Wiedenhoeft, A.: Classification of cites-listed and other neotropical meliaceae wood images using convolutional neural networks. Plant Methods **14** (03 2018)
13. Ravindran, P., Thompson, B.J., Soares, R.K., Wiedenhoeft, A.C.: The xylotron: flexible, open-source, image-based macroscopic field identification of wood products. Frontiers in plant science **11**, 1015 (2020)
14. Rosa, N., Deklerck, V., Baetens, J., Van den Bulcke, J., De Ridder, M., Rousseau, M., Beeckman, H., Van Acker, J., De Baets, B., Verwaeren, J.: Improved wood species identification based on multi-view imagery of the three anatomical planes. Plant Methods **18** (06 2022)
15. Seeland, M., Mader, P.: Multi-view classification with Convolutional Neural Networks. PLoS ONE **1**(16) (2021)
16. da Silva, N.R., *etal*, V.D.: Improved wood species identification based on multi-view imagery of the three anatomical planes. Plant Methods **18**(1), 79 (2022)
17. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
18. Sun, Y., Lin, Q., He, X., Zhao, Y., Dai, F., Qiu, J., Cao, Y.: Wood species recognition with small data: A deep learning approach. International Journal of Computational Intelligence Systems **14**(1), 1451–1460 (2021)
19. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks (2020), https://arxiv.org/abs/1905.11946
20. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training (2021), https://arxiv.org/abs/2104.00298
21. Wu, F., Gazo, R., Haviarova, E., Benes, B.: Wood identification based on longitudinal section images by using deep learning. Wood Science and Technology **55**, 1–11 (03 2021)
22. Zielinski, K.M., *etal*, L.S.: Advanced wood species identification based on multiple anatomical sections and using deep feature transfer and fusion. Computers and Electronics in Agriculture **231**, 109867 (2025)