

Wood Physics/Mechanical Properties

Jördis Sieburg-Rockel*, Stephanie Helmling, Lars Nieradzik, Stephanie Wrage, Thomas Weibel, Tim Lehne, Volker Haag, Immo Heinz, Jonas Heddier, Henrike Stephani and Andrea Olbrich

Wood identification in fiber materials: a comparative blind test study of artificial intelligence and human experts

<https://doi.org/10.1515/hf-2025-0099>

Received August 18, 2025; accepted December 22, 2025;

published online February 5, 2026

Abstract: The conversion of processes controlled by human expertise and know-how into automated operations is the goal of a wide range of modern research. This study reports on a blind test in which twelve unknown samples were analyzed by human experts and an AI system developed to check the genera of hardwoods used in paper production. Trained to detect hardwood cells in microscopic images of maceration slides, it assigns them to the nine most commonly used hardwood species. Softwoods were also added to the samples, to make the test as realistic as possible. Human experts achieved an accuracy of 0.96 for recognizing a genus contained in a sample. The two-stage machine recognition system achieved an accuracy of 0.79 at the defined threshold values of 0.7 for object recognition and 0.75 for genus probability. Human experts failed to recognize some genera contained in the samples, whereas the AI models did not miss any genus at these thresholds. A comparison of the results revealed the strengths and weaknesses

of the independent methods. This study is an important step towards optimizing the automated image recognition system developed to support the implementation of the European Union Deforestation Regulation (EUDR) with regard to deforestation-free products/supply chains.

Keywords: wood identification; macerate; artificial intelligence (AI); vessel elements; EUDR; machine learning

1 Introduction

In 2023, over 400 million tons of paper and paperboard were produced globally (FAO 2025). As wood is the raw material for these products, it is important to consider the sources. The new European Regulation 2023/1115 on deforestation-free supply chains (European Union 2023) aims to protect forests and biodiversity worldwide. However, regulations and laws can only be effective if they are evaluated (Dormontt et al. 2015). In the global timber trade, there are various methods (e.g. genetic analysis, stable isotope analysis, mass spectrometry, near-infrared spectroscopy and wood anatomy) for identifying timber and tracing it back to its source in order to combat illegal logging and protect forest habitats, biodiversity and local communities (Beeckman et al. 2020; Dormontt et al. 2015; Low et al. 2022). And every method has its benefits and drawbacks. In recent years, artificial intelligence (AI), particularly machine learning and deep learning, has become a major focus of research in wood species identification, as it enables fast, accurate and scalable recognition of wood (He et al. 2024; Hwang and Sugiyama 2021; Silva et al. 2022a). This reduces reliance on experienced experts and supports law enforcement. In addition, data can be efficiently linked and used for automated identification (Lens et al. 2020). Wood in paper and pulp can be identified by chemical analysis of the extractives, but so far only to a very limited extent because references are still lacking (Flaig et al. 2024). The established method for identifying wood in this material, which is so structureless and has lost most of its characteristics compared to solid wood, is wood anatomy (Beeckman et al. 2019). It

*Corresponding author: **Jördis Sieburg-Rockel**, Johann Heinrich von Thünen Institute (TI), Institute of Wood Research, Leuschnerstr. 91, 21031 Hamburg-Bergedorf, Hamburg, Germany, E-mail: joerdis.sieburg-rockel@thuenen.de. <https://orcid.org/0009-0001-7547-269X>

Stephanie Helmling, Stephanie Wrage, Tim Lehne, Volker Haag, Immo Heinz and Andrea Olbrich, Johann Heinrich von Thünen Institute (TI), Institute of Wood Research, Leuschnerstr. 91, 21031 Hamburg-Bergedorf, Hamburg, Germany. <https://orcid.org/0009-0009-6611-3140> (S. Helmling). <https://orcid.org/0009-0000-0112-6113> (S. Wrage). <https://orcid.org/0000-0002-2376-3436> (T. Lehne). <https://orcid.org/0000-0002-5913-6485> (V. Haag). <https://orcid.org/0000-0002-5840-9066> (I. Heinz). <https://orcid.org/0009-0007-2249-2797> (A. Olbrich)

Lars Nieradzik, Thomas Weibel and Henrike Stephani, Image Processing Department, Fraunhofer ITWM, Fraunhofer Platz 1, 67663, Kaiserslautern, Germany. <https://orcid.org/0000-0002-7523-5694> (L. Nieradzik). <https://orcid.org/0009-0009-5848-5024> (T. Weibel). <https://orcid.org/0000-0002-9821-1636> (H. Stephani)

Jonas Heddier, Institute of Wood Science, University of Hamburg, Leuschnerstr. 91, 21031 Hamburg-Bergedorf, Hamburg, Germany

requires only a few remaining structural features of individual cells for identification (Helmling et al. 2018; Ilvessalo-Pfäffli 1995) and, of course, only accurate at the generic level (Gasson 2011). However, the analysis of processed wood species in fibrous materials is time-consuming and covers only a very small part of the global production. It must be carried out by specialized wood anatomists who examine several hundreds of cells for each sample under a light microscope (Beeckman et al. 2020). Therefore, a project was established with the aim of automatically identifying wood species in microscopic images of fibrous materials (pulp and paper products) using AI. At least one previous study has reported the segmentation and characterization of macerated fibers and vessel elements of a single genus using machine learning (Qamar et al. 2024). However, the recent work by Nieradzic et al. 2024a represents the first implementation that extends this approach to multiple wood species and applies it to the detection of diagnostically relevant cells for subsequent taxonomic classification.

An AI system for the automated identification of wood species in fibrous material was trained with an image data set of microscopic images of hardwood (angiosperm) samples at five focal planes as described in detail by Nieradzic et al. 2024a–c). In this data set, the cells important for identifying hardwoods, the vessel elements, were initially annotated manually and can now be easily detected and classified by the system. An AI model for the detection and classification of softwoods (gymnosperms) is not yet available. However, paper is sometimes made from a combination of hardwood and softwood, and it is common for these two types of wood to be processed together to create a single product. That is why mixed samples of hardwoods and softwoods were used, even though no fully trained model for softwood recognition was available. Mixed samples represent real samples much better than pure hardwood samples. The genera used are *Acacia*, *Betula*, *Eucalyptus*, *Fagus*, *Hevea*, *Liquidambar*, *Populus*, *Salix*, *Schima*, *Abies*, *Cunninghamia*, *Picea*, *Pinus* and *Pseudotsuga*. Some of them are cultivated in plantations worldwide and are particularly well adapted for pulp and paper production due to their rapid growth, good fiber properties and high yields (Kanninen 2010). The blind test, which compares the results between man and machine, is intended to show how well the existing system has worked so far and also demonstrates the knowledge gained, the limitations, and the potential for improvement.

2 Materials and methods

2.1 Material and datasets

The reference material for producing the macerates according to Franklin 1945 was provided from the scientific

Thünen wood collection (RBHw) (Thünen-Institut 2025) and other documented sources determined at least to the generic level. The starting materials consisted exclusively of self-produced macerates to prevent contamination with other types of wood (Helmling et al. 2016, 2018; Nieradzic et al. 2024a). The training dataset for the detection of vessel elements in the overview images (number of images: 2,236) consisted of 106 hardwood and softwood samples representing 57 species across 21 genera, specifically: *Abies* (3 samples, 2 species: *A. alba*, *A. grandis*), *Acacia* (10 samples, 4 species: *A. aulacocarpa*, *A. confusa*, *A. mangium*, *A. mearnsii*), *Acer* (5 samples, 4 species: *A. platanoides*, *A. pseudoplatanus*, *A. rubrum*, *A. saccharum*), *Alnus* (3 samples, 3 species: *A. glutinosa*, *A. japonica*, *A. rubra*), *Betula* (8 samples, 3 species: *Betula alba*, *B. pendula*, *B. utahensis*), *Casuarina* (2 samples, 1 species: *C. equisetifolia*), *Cunninghamia* (3 samples, 2 species: *C. lanceolata*, *C. sinensis*), *Diospyros* (2 samples, 2 species: *D. kaki*, *D. sakalavarum*), *Eucalyptus* (16 samples, 6 species: *E. globulus*, *E. grandis*, *E. saligna*, *E. smithii*, *E. urograndis*, *E. viminalis*), *Fagus* (5 samples, 1 species: *F. sylvatica*), *Hevea* (5 samples, 2 species: *H. brasiliensis*, *H. guianensis*), *Liquidambar* (5 samples, 3 species: *L. excelsa*, *L. formosana*, *L. styraciflua*), *Melaleuca* (1 sample, 1 species: *M. leucadendra*), *Picea* (3 samples, 3 species: *P. abies*, *P. jezoensis*, *P. rubens*), *Pinus* (14 samples, 6 species: *P. canariensis*, *P. caribea*, *P. maximinoi*, *P. patula*, *P. radiata*, *P. sylvestris*), *Populus* (8 samples, 6 species: *P. canadensis*, *P. deltoides*, *P. grandidentata*, *P. nigra*, *P. tremula*, *P. tremuloides*), *Pseudotsuga* (3 samples, 1 species: *P. menziesii*), *Quercus* (1 sample, 1 species: *Q. robur*), *Robinia* (2 samples, 1 species: *R. pseudoacacia*), *Salix* (3 samples, 3 species: *S. acutifolia*, *S. alba*, *S. fragilis*), and *Schima* (4 samples, 2 species: *S. superba*, *S. wallichii*). All images of the above-mentioned samples of the genera *Acacia*, *Betula*, *Eucalyptus*, *Fagus*, *Hevea*, *Liquidambar*, *Populus*, *Salix* and *Schima* were used for classification training - including the images of the macerates used for mixing the blind test samples. As a matter of fact, always new cells from a macerate are used for each preparation (each slide) – i.e., always other cells that neither the experts nor the models have ever seen before, therefore leakage is highly unlikely. The blind test samples contain nine hardwood species *Acacia confusa*, *Betula alba*, *Eucalyptus grandis*, *Fagus sylvatica*, *Hevea brasiliensis*, *Liquidambar formosana*, *Populus tremuloides*, *Salix fragilis* and *Schima wallichii* and six softwood species *Abies alba*, *Cunninghamia lanceolata*, *Picea abies*, *Pinus sylvestris*, *Pinus taeda* and *Pseudotsuga menziesii*. The species are listed here to enhance the transparency of the data set. In the following sections, only generic names are given as differentiation at the specific level is usually impossible.

2.2 Preparation of the sample mixtures

12 blind test samples (Table 1) were blended from the individual macerated samples. The mixing ratios and composition of genera were previously randomly generated under the following conditions: All genera should be included at least once in a blind sample. The proportion of each macerate should be roughly equal. Except in the case of *Hevea*, where experience has shown that the proportion of vessel elements in the macerate is very low, a double amount was added to ensure that it is well represented in the sample and sufficient vessel elements on each of the slides can be examined. Up to a maximum of four different genera could be contained in one sample. The “confusion partners” (genera with very similar anatomical features of the vessel elements) *Populus-Salix* and *Liquidambar-Schima* had to be included at least once together and at least once alone. In total, 12 blind samples were compiled combining six hardwood samples, four hardwood-softwood samples, and two softwood samples (Table 1).

Three slides stained with Alexander-Herzberg solution (AH) and three stained with Nigrosin (Ni) were prepared per sample, as described in Helmling et al. 2018 resulting in a total of six slides per blind test sample. These 72 slides were then divided into three sets of 24 slides (12 × one AH and one Ni slide each) for the experts, but were summarized in a single evaluation at the end (Tables 4 and 5). Thus, the evaluation of the results of the AI models was as comparable as possible, with all six slides (3 × AH and 3 × Ni) being evaluated together. The composition of the blind sample mixture was kept secret until the end of the evaluation by all

experts. The samples contained hardwoods and softwoods in order to match an industrial paper as realistic as possible.

2.3 Imaging

A Zeiss Axioscan 7 microscope (Zeiss, Germany) was used for the overview scans of the entire slide surface. The customized scan profile was already used to create the training data. An N-Achroplan 5×/0.15 objective was used to scan an area of approximately 8 cm² with a scale per pixel of 0.69 × 0.69 × 16.33 μm³ in five focal planes per slide (software: ZEN slidescan 3.5, Zeiss, Germany) (Nieradzki et al. 2024a).

2.4 Human experts' evaluation

The experts' evaluation for the composition of the samples is based on the routine method for the examination of sample material, as used in daily practices. The fiber material is usually spread out on two microscope slides. The cells are then stained and examined under a light microscope. The scientists' decision which genera are contained in a sample is based on their ability to clearly identify cells and classify these cells according to their characteristic morphological features. Usually, individual vessel elements are the most useful cells for identification, as they can be distinguished by their characteristic features such as the type of perforation plates, the presence or absence of helical thickenings, and the type and arrangement of vessel-ray pits (Flaig et al. 2024; Helmling et al. 2018; IAWA Committee 1989; Ilvessalo-Pfäffli 1995; Richter and Dallwitz 2000; Wheeler 2004). Those cells are assigned to a genus with the greatest probability.

Each of the six participating scientists is a wood anatomist with extensive experience in fiber analysis. All received one AH slide and one Ni slide to examine. The blind test included 12 samples × 14 genera = 168 decisions per expert and 1,008 decisions all together. Each of the scientists could select multiple genera per sample.

To assess the agreement of the scientists, Krippendorff's alpha (α_k) was used: A statistical method for evaluating agreement without taking truth into account, that is often used when more than two expert opinions are available on the same question. The calculated intraclass correlation coefficient gives values between −1 and 1. The following evaluation of these values is recommended: systematic disagreement (below 0), insufficient agreement (below 0.67), acceptable agreement (between 0.67 and 0.80), strong agreement (above 0.80), perfect agreement (1) (Krippendorff 2004). α_k was calculated using this code (Castro 2017).

Table 1: Blind test sample mixtures. Genera contained in samples 1 to 12 for the comparative blind test.

Sample	Hardwood genera	Softwood genera
1	<i>Liquidambar</i> , <i>Populus</i> , <i>Schima</i>	
2	<i>Fagus</i> , <i>Liquidambar</i> , <i>Populus</i>	
3	<i>Acacia</i> , <i>Salix</i>	<i>Abies</i> , <i>Cunninghamia</i>
4	<i>Hevea</i> , <i>Schima</i>	<i>Cunninghamia</i>
5	<i>Fagus</i> , <i>Hevea</i>	<i>Picea</i> , <i>Pseudotsuga</i>
6	<i>Betula</i> , <i>Schima</i>	<i>Pinus</i>
7		<i>Abies</i> , <i>Picea</i> , <i>Pseudotsuga</i>
8	<i>Acacia</i> , <i>Eucalyptus</i> , <i>Populus</i> , <i>Salix</i>	
9	<i>Acacia</i> , <i>Hevea</i>	
10		<i>Abies</i> , <i>Cunninghamia</i> , <i>Pinus</i> , <i>Pseudotsuga</i>
11	<i>Betula</i> , <i>Liquidambar</i> , <i>Populus</i>	
12	<i>Eucalyptus</i> , <i>Schima</i>	

2.5 Used networks

Two machine learning based models were used for the automated analysis of the microscopic overview images. For the analysis of the samples by the models, all six slides per sample were evaluated as one unit.

Similar to the experts' analysis procedure, the relevant cells were detected first by WoodYOLO (Nieradzic et al. 2024c). The detected cells are localized with x- and y-coordinates in bounding boxes (BB) within the image. Although many object detection models are available, WoodYOLO offers a key advantage as it directly optimizes for recall. This is in line with the objective of identifying all vessel elements, even at the cost of a slightly higher false positive rate. False positives can be controlled later in the classification by adjusting the confidence threshold, but ensuring high recall is crucial. This is especially important for genera like *Hevea*, as it naturally has fewer vessel elements. Furthermore, because WoodYOLO is specifically optimized for microscopic images of macerated wood cells, it provides more accurate bounding boxes tailored to this domain.

In the second step, the image sections within the bounding boxes that contained supposedly relevant cells were classified using ConvNeXt-tiny. This architecture was selected based on both empirical performance and recent findings from the literature that question the traditional reliance on ImageNet accuracy as a proxy for real-world effectiveness. ConvNeXt has consistently demonstrated strong performance across a wide range of tasks and domains, often outperforming older convolutional architectures such as ResNet and DenseNet (Nieradzic et al. 2024a).

3 Results and discussion

The task for both expert and machine was to analyze the 12 blind samples consisting of different mixtures of 9 hardwoods and 5 softwoods. Similar to the assessment of the experts' wood anatomical examination, AI also pays attention to the anatomical characteristics of the relevant cells (Nieradzic et al. 2024b) in terms of pattern recognition (Richter and Dallwitz 2000-onwards; Wheeler 2004-onwards). But convolutional neural networks learn patterns based on a data set, experts achieve it through experience in a similar way. But experts learn at a much faster rate (with fewer images as examples) and can extrapolate better, while the machine is limited to the knowledge of the data set. And the risk is that color, brightness or other repetitive, process-related patterns are also learned, while

human experts can focus better on defined features and exclude less important features (Lens et al. 2020; Silva et al. 2022b). These are exemplary reasons why it is important to distinguish between the experts' evaluation procedures and those based on AI models. The experts analyzed two slides of each blind sample. Each vessel element on these two slides was analyzed, but the decision whether a hardwood genus was present in the sample or not was taken only on clearly identifiable vessel elements. In contrast, the decision of the models as to whether a genus was present in a sample was based on all detected and classified vessel elements in all six images/slides of a sample, regardless of the clarity/confidence. While the experts in many instances refrain from classifying a vessel element due to its unclear morphology, the AI algorithms assign one class (the one with highest likelihood) regardless. In addition, the wood anatomists know softwood, while the models are only trained on recognition and classification of nine hardwoods so far. Cell fragments of the softwoods contained in the blind samples could therefore be falsely attributed, but subsequently not correctly classified in any case. For that reason, the results are presented separately for experts and automated evaluation and then the potentials and limitations are discussed together.

The accuracy, i.e. how correct the detection is with respect to the known ground truth, is given for the results. The precision indicates the proportion of correctly recognized genera (= true positive (TP)) of the total number of recognized genera. This means false positive (FP) hits are included in the calculation. Recall, on the other hand, indicates the proportion of correctly recognized genera that were recognized. In other words, the incorrectly unrecognized (= false negatives (FN)) genera are included, so that the recall becomes worse when genera were overlooked. The genera not falsely added to a sample are counted as true negative (TN).

3.1 Human experts' results

As described before, the results of the wood anatomists who examined the samples are evaluated on the basis of recognized and unrecognized genera in a sample. Decisions are made here on the basis of a number of clearly assignable individual vessel elements in a sample, but not on the basis of each visible individual vessel element on a slide. Every participant had to make 168 decisions whether a genus was contained in the sample. The participants had knowledge which genera were used for the entire project of the blind test. All participants who are not directly involved in the project have experience with the analysis of paper and fiber

Table 2: Human experts' results of hardwoods and softwoods.

	True	False
Positive	204	24
Negative	756	24
Accuracy	0.95	
Precision	0.89	
Recall	0.89	

Table 3: Human experts' results of hardwoods without softwoods.

	True	False
Positive	140	13
Negative	485	10
Accuracy	0.95	
Precision	0.92	
Recall	0.93	

material from their work as experts at the Thünen Centre of Competence on the Origin of Timber. The results of the experts' analysis are summarized in Table 2. For completeness, the evaluation was presented once with and, for better comparability with AI results, once without softwood genera (Table 3). The results differ slightly, being only a bit better without the softwoods.

Krippendorff's alpha was calculated to provide a measure of agreement for the genera recognized by experts. The results of the six participants vary for the different genera (Tables 4 and 5). There is perfect agreement with *ak* for the genera *Betula*, *Populus* and *Pseudotsuga*. There is a strong agreement with *ak* above 0.8 for the genera *Pinus*, *Salix*, *Fagus* and *Hevea* and acceptable agreement between *ak* 0.67 and 0.80 for *Schima*, *Acacia* and *Eucalyptus*. For the genera *Liquidambar*, *Cunninghamia*, *Picea* and *Abies* only insufficient agreement was achieved. It should be noted that this is an analysis of the agreement between the participants' results; the truth is not taken into account in this evaluation, nor the presence of a confusion partner. Therefore, the fact that the participants differ in their assessments is no reason to fundamentally question them. However, the differentiation of some genera, where insufficient agreement has occurred, should be considered critically. Discussion of these results will be continued later. In summary, some genera were recognized equally well by all participants, while the recognition of other genera was less homogeneous. In the following, the results will be considered in terms of their accuracy, precision, and recall, taking into account the truth. At least one participant recognized 100 % correctly all genera without false positives (not shown in the table).

Whether a genus was contained in one of the 12 samples was decided by the 6 experts, resulting in 72 decisions per genus (Tables 4 and 5). Three genera (*Betula*, *Populus* and *Pseudotsuga*) were recognized with absolute certainty by all participants in all samples (accuracy = 1). This is not surprising, as the anatomical characteristics of at least two of these genera (very fine scalariform perforation plates and minute intervessel pits (*Betula*) and softwood tracheids with helical (wall) thickenings (*Pseudotsuga*)) differ significantly from the characteristics of the other genera included. *Populus* and *Salix*, on the other hand, are actually a confusion pair with very similar anatomical features. Both genera show very distinct vessel-ray pits. But in most cases this confusion occurs in one direction. Many vessel elements of *Salix* look like *Populus* because the distinguishing feature (upright marginal cells for *Salix*) is missing. The good result for *Populus* was achieved because anatomists considered only the significant cells for their decisions. *Salix* was also never overlooked in any sample (recall = 1) and only two of the 72 decisions for the genus were false positives.

The genera *Acacia* and *Pinus* were never falsely positive assigned (precision = 1). The participants achieved the worst precision with the genus *Picea*. It was falsely detected six times (precision 0.63). This result is surprising, as one would expect that the very small cross-field pits would be more easily overlooked than imagined. The worst recall was achieved for the genus *Abies* (recall 0.67). This was overlooked a total of six times. The results for the genus *Cunninghamia* are similarly poor. This means that three softwood genera (*Abies*, *Cunninghamia* and *Picea*) were the least well identified. The two other softwood genera *Pinus* and *Pseudotsuga* were identified very reliably. The genus *Pinus* forms distinctive window-like or pinoid cross-field pits, which allow reliable differentiation from the other softwood genera. *Pinus* was overlooked only once. The genus *Pseudotsuga* was neither overlooked nor misidentified. *Pseudotsuga* forms unmistakable helical thickenings in tracheid cell walls in combination with piceoid cross-field pits that make identification easier. There were three pairs of confusion partners in the blind test, two hardwood pairs (*Liquidambar-Schima* and *Populus-Salix*) and one softwood pair (*Abies-Cunninghamia*). All participants were very sure about the combination of *Populus* and *Salix*. Contrary to the assumption that these closely related genera with very similar structural characteristics are difficult to differentiate, the anatomists made almost no mistakes. Only twice the genus *Salix* was noted as false positive. The number of errors was significantly higher for the pair of *Liquidambar* and *Schima*. *Liquidambar* was falsely identified four times and overlooked twice. *Schima* was falsely identified three times and overlooked once. The reason for this is that the

Table 4: Human experts' performance metrics for hardwoods. Six experts decided, whether a genus was contained in one of the 12 samples, resulting in 72 decisions per genus.

Genus	TP	TN	FP	FN	Accuracy	Precision	Recall	ak
<i>Acacia</i>	15	54	0	3	0.96	1	0.83	0.7841
<i>Betula</i>	12	60	0	0	1	1	1	1
<i>Eucalyptus</i>	11	58	2	1	0.96	0.85	0.92	0.7223
<i>Fagus</i>	11	59	1	1	0.97	0.92	0.92	0.8028
<i>Hevea</i>	16	53	1	2	0.96	0.94	0.89	0.8026
<i>Liquidambar</i>	16	50	4	2	0.92	0.80	0.89	0.6450
<i>Populus</i>	24	48	0	0	1	1	1	1
<i>Salix</i>	12	58	2	0	0.97	0.86	1	0.8251
<i>Schima</i>	23	45	3	1	0.94	0.88	0.96	0.7863
Mean	140	485	13	10	0.96	0.92	0.93	–

Table 5: Human experts' performance metrics for softwoods.

Genus	TP	TN	FP	FN	Accuracy	Precision	Recall	ak
<i>Abies</i>	12	51	3	6	0.88	0.80	0.67	0.4851
<i>Cunninghamia</i>	13	52	2	5	0.90	0.87	0.72	0.5848
<i>Picea</i>	10	54	6	2	0.89	0.63	0.83	0.4929
<i>Pinus</i>	11	60	0	1	0.99	1	0.92	0.8942
<i>Pseudotsuga</i>	18	54	0	0	1	1	1	1
Mean	64	271	11	14	0.93	0.85	0.82	–

distinguishing feature, the number and thickness of the bars of the scalariform perforation plates, tends to differ, but there are overlaps in the expression of the feature (Helmling et al. 2018). The identification of some softwoods was not easy and therefore the differentiation of the softwood pair worked least well.

3.2 AI results

The relevant cells were detected in the large overview images by the WoodYOLO model (Nieradzic et al. 2024c). In a future application, not necessarily all detected cells have to be analyzed, but analogous to the experts' analysis only the clearly recognizable and also classifiable cells need to be evaluated. The predicted object probability (detection confidence) was used as a threshold for further evaluation. Therefore, threshold values of over 0.9, over 0.7 and over 0.3 were assumed for the object probability. For example, vessel elements overlain by other cells in the preparation and whose characteristics are therefore obscured are removed from the analysis. Such a masked cell (Figure 1I) would inevitably be misclassified in the next step and result in unnecessary false positives.

The classification step was performed by the ConvNeXt-tiny model, which has already shown strong performance (Nieradzic et al. 2024a). A certain genus probability (classification confidence) was also found as a threshold value for the analysis. Thus, in a future application, the results displayed are reduced to the cells that have been classified with a high degree of certainty. The threshold value for the analysis was not defined in advance, so the results for several threshold values were determined. Threshold values of over 0.95, over 0.75 and over 0.5 were assumed for the generic accuracy.

The results were determined once for all 25,366 detected and classified bounding boxes (BBs) in all samples (Table 6). For the evaluation, the results for which the classification gave the highest confidence for a genus that was actually contained in the individual blind sample were considered as potentially correct and thus evaluated as a true positive. Conversely, the incorrectly classified cells in the bounding boxes were evaluated as false positives. However, this method did not allow numerical values to be determined for false negative (overlooked) and true negative cells. Therefore, no accuracy can be calculated for this evaluation. Only precision can be achieved. Without any thresholds, 11,225 BBs were potentially correctly recognized and classified at

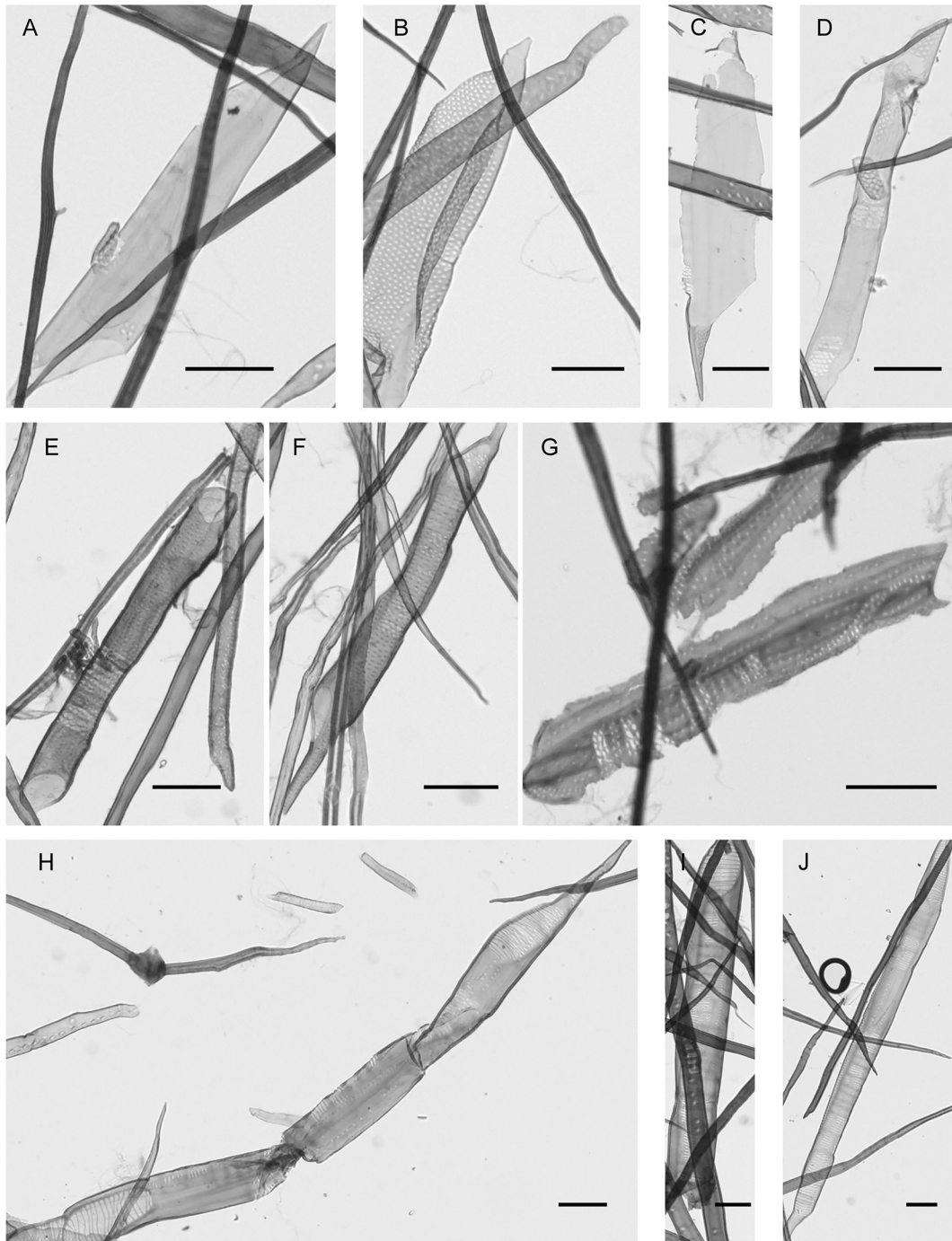


Figure 1: False positive decisions by the AI in original bounding box dimensions. Vessel elements of *Salix*, classified as *Populus* (A–D). *Fagus* vessel elements, classified as *Eucalyptus* (E–G). Vessel elements from *Schima*, classified as *Liquidambar* (H–J). All scale bars = 100 μm .

this “bounding box level” (BB level). 14,141 boxes were categorized with the highest confidence to a genus that was not part of the respective sample. These were either not correctly recognized as vessel elements and thus inevitably misclassified or a real vessel element was incorrectly classified.

A further way of displaying the results is whether a genus was correctly recognized at least once in a sample or not (Table 7). This “genus level” (G level) is important in order to recognize whether a genus has been incorrectly categorized as contained or overlooked. Nine hardwoods were included in the 12 blind samples. This results in

Table 6: AI results on the bounding box level: The results were determined for all 25,366 detected and classified bounding boxes. The object probability and the genus probability serve as threshold values here in order to select the detected and classified cells in the bounding boxes with a high probability.

Object prob.	Genus prob.	TP	FP	Precision
0.9	0.95	203	0	1
	0.75	231	4	0.98
	0.5	260	23	0.92
	0	264	30	0.90
0.7	0.95	3,479	54	0.98
	0.75	4,176	246	0.94
	0.5	5,134	1,101	0.82
0.3	0	5,299	1,380	0.79
	0.95	4,263	217	0.95
	0.75	5,333	994	0.84
0	0.5	6,985	4,623	0.60
	0	7,390	5,735	0.56
	0.95	4,983	527	0.90
	0.75	6,745	2,433	0.73
	0.5	9,880	10,647	0.48
	0	11,225	14,141	0.44

12 × 9 = 108 decisions as to whether a genus was included in a sample or not. In fact, hardwood was included 25 times in the 12 blind samples. Thus a total of 25 true positive and false negative decisions and 83 true negative and false positive decisions could be made. If the recall is 1, then no genus was overlooked. The analysis comes close to the experts' analysis and is therefore most suitable for comparison.

With a threshold value of 0.9 for the object probability, good to very good results were achieved for the precision at the “BB level” (e.g. 100 % at 0.9 and 0.95), but at the “G level” 7 genera in the 12 samples are simply overlooked (7 false negatives). The fact that genera are overlooked is not suitable for downstream use. Furthermore, only 203 BBs of the original 11,225 potentially correctly detected BBs remain for evaluation, which is not even two percent. That represents too much loss of useful data. With a threshold value of 0.7 for the object probability and 0.95 for the genus probability, 3,533 BBs are still classified. This corresponds to a good 31 % of all potentially correctly detected BBs. The precision here is 0.98. Therefore, a threshold value of 0.7 for the object probability is the highest threshold value that should be assumed for the evaluation.

The threshold values were set to 0.7 and 0.75 for the results of the AI models so that nothing was overlooked and a sufficient number of vessel elements in the BB could be evaluated. The image sections of the false-positive BBs with these threshold values were examined more closely in order to identify any remaining weaknesses in the detection or classification. These were exactly 246 BBs. 17 BBs contained image sections of the two confusion partners (4 × *Liquidambar-Schima* and 13 × *Populus-Salix*). These genera are extremely difficult to distinguish, also for experts when it comes to evaluating every single vessel element. *Populus* and *Salix* are botanically very closely related, both belong to the Salicaceae family. The only difference between the vessel elements of these two genera is that *Salix* shows upright marginal cells in the rays whose presence can be

Table 7: AI results on the genus level. The table shows, whether a genus was correctly recognized at least once in a sample or not. Nine hardwoods were included in the 12 blind samples, this results in 12 × 9 = 108 decisions. Recall = 1 indicates that no genus was overlooked.

Object prob.	Genus prob.	TP	TN	FP	FN	Accuracy	Precision	Recall
0.9	0.95	18	83	0	7	0.94	1	0.72
	0.75	21	79	4	4	0.93	0.84	0.84
	0.5	22	71	12	3	0.86	0.68	0.88
	0	22	69	14	3	0.84	0.61	0.88
0.7	0.95	25	67	16	0	0.85	0.61	1
	0.75	25	60	23	0	0.79	0.52	1
	0.5	25	48	35	0	0.68	0.42	1
	0	25	41	42	0	0.61	0.37	1
0.3	0.95	25	64	19	0	0.82	0.57	1
	0.75	25	49	34	0	0.69	0.42	1
	0.5	25	36	47	0	0.57	0.35	1
	0	25	27	56	0	0.48	0.31	1
0	0.95	25	58	25	0	0.77	0.5	1
	0.75	25	35	48	0	0.56	0.34	1
	0.5	25	16	67	0	0.38	0.27	1
	0	25	7	76	0	0.30	0.25	1

inferred from the corresponding pit fields in the vessel walls (Figure 1A+D). But many vessel elements of *Salix* look like *Populus* because the upright marginal cells are not visible (Figure 1B+C).

Although the genera *Liquidambar* and *Schima* do not even belong to the same botanical family, the anatomical features of the vessel elements differ only sometimes in the number of bars of the scalariform perforation plates at the vessel element ends. It is therefore understandable that confusion arises here four times. *Eucalyptus* was most frequently incorrectly detected and classified (102 times). In most cases (62 times), cell parts of softwood could be seen in the BBs. In 27 cases, the genus *Fagus* was confused with *Eucalyptus* (Figure 1E–G). The genus *Schima* was the second most frequently misclassified (63 times). 62 of these BBs contained softwood cells, only once *Fagus* was confused with *Schima*. *Liquidambar* was the third most frequently misclassified genus (58 times). 53 BBs actually contained softwood cells and four times *Schima* was mistaken for *Liquidambar* (Figure 1H–J).

In total, 178 of 246 BBs were misclassified because they contained softwood cells. A class that is unknown for both models, but occurs repeatedly in real-world paper mixtures. Considering how many cells in the pure softwood samples did not lead to any confusion with vessel elements, the detection of the vessel elements is basically very good. The targeted exclusion of softwood tracheids by the detection is a task for future improvements. Despite of the relevant number of false vessel element detections, a precision of 0.94 was achieved for these threshold values at the BB level and an accuracy of 0.79 at the G level.

Table 8 shows the results for the individual hardwood genera at the threshold values 0.7 and 0.75. They show whether a genus was detected at least once in a sample (G level) or not. This makes the results comparable with the experts' results. As explained above, there are no false negatives with these threshold values. Overlooking a genus would not be expedient with regard to the subsequent application. Furthermore, the genera *Acacia*, *Hevea* and *Salix* were never falsely predicted in a sample in which they did not occur (no false positives). *Fagus* was only falsely identified in one sample, *Betula* in two and *Populus* in three of the samples. The analysis of the individual images of the falsely positive genera showed that the supposed *Fagus* vessel elements were in fact *Liquidambar* vessel elements. Although these two genera are not typical confusion partners, they have a certain similarity. Both have or can have scalariform perforation plates and scalariform intervessel pits. Each image contained at least one of these features, but an expert would not have used them for serious identification purposes in case of doubt. A very similar situation exists

Table 8: AI model performance for hardwood genera of all blind samples at the threshold values 0.7 (object prob.) and 0.75 (genus prob.).

Genus	TP	TN	FP	FN	Accuracy	Precision	Recall
<i>Acacia</i>	3	9	0	0	1	1	1
<i>Betula</i>	2	8	2	0	0.83	0.50	1
<i>Eucalyptus</i>	2	3	7	0	0.42	0.22	1
<i>Fagus</i>	2	9	1	0	0.92	0.67	1
<i>Hevea</i>	3	9	0	0	1	1	1
<i>Liquidambar</i>	3	3	6	0	0.50	0.33	1
<i>Populus</i>	4	5	3	0	0.75	0.57	1
<i>Salix</i>	2	10	0	0	1	1	1
<i>Schima</i>	4	4	4	0	0.67	0.50	1
Mean	25	60	23	0	0.79	0.52	1

with the false positive *Betula* vessel elements. In these cases, they were in fact *Schima* vessel elements. Here, too, the similarity between the genera lies in the scalariform perforation plates, which could explain the false classification. However, since the pits have a completely different shape and size, confusion by experts would be unlikely. Most of the vessel elements classified as *Populus* were in fact *Salix* vessel elements. Here the typical confusion partner was incorrectly assigned in the classification.

The results permit a closer look at the samples that do not contain softwood (Table 9). In the six pure hardwood samples, 7,300 potentially true positive and 2,277 false positive BBs were detected and classified. Thus, an accuracy of 0.76 was achieved even without threshold values in the BB level. For the threshold values 0.7 and 0.75, 3,183 potentially true positive BBs and only 28 false positive BBs were found and classified. A precision of 0.99 was achieved at BB level. At G level, all genera were correctly recognized at least once in each sample and none were overlooked. Five times a genus was misidentified in the samples (5 false positives). This results in an accuracy of 0.91, a precision of 0.77 and a recall of 1.

Table 9: AI model performance for hardwood genera of blind samples containing exclusively hardwoods.

Object prob.	Genus prob.	TP	FP/TN	Accuracy	Precision	Recall
BB level						
0.7	0.75	3,183	28	–	0.99	–
0	0	7,300	2,277	–	0.76	–
G level						
0.7	0.75	17	32	0.91	0.77	1

3.2.1 Differences and similarities

The differences in analysis between experts and machine have already been explained in detail. The experts' analysis is based on the visual inspection of wood samples and the experience of the analyst, while machine analysis is based on the processing of images and the application of algorithms. In a direct comparison the results of the experts and machine analysis show both similarities and differences. On average, the six scientists together achieved a better value for the accuracy of whether a genus was present in a sample (0.96) than the machine (0.79) at the defined threshold values of 0.7 for object recognition and 0.75 for genus certainty. However, the automated recognition worked perfectly for three genera, while human participants only recognized two of the hardwood genera without any errors. Surprisingly, the genus *Eucalyptus*, which has no expected confusion partners, was recognized below average by the machine. And even in the experts' results, a slight uncertainty can be seen for this genus when considering the only acceptable agreement according to *ak*. *Salix*, on the other hand, was recognized above average, although there is a high potential for confusion with *Populus*. Classification worked particularly well here because a detected vessel element could also be reliably assigned. The poor differentiation between *Schima* and *Liquidambar* can be explained by the below-average results for both experts' and machine analysis, due to the great anatomical similarity. The only acceptable agreement according to *ak* also confirms the previous approach in expert reports of declaring these two genera together. The same applies to the three softwood genera with insufficient agreement according to *ak*. But in the absence of a possibility of comparison, the results for the softwood identification from the experts' evaluation are not taken into account here. An important point is that automated analysis is faster and more efficient in analyzing large amounts of data, while the experts' analysis provides higher accuracy and reliability in recognizing wood species. The fact is that neither experts nor machine achieved a perfect result. Overall, the results show that both methods have strengths and weaknesses and that a combination of machine and subsequent experts' analysis or control in a planned application may currently be the best solution for wood genera recognition.

3.2.2 Potential and limitations

Some genera can be recognized and differentiated better than others. There are several reasons for this. On the one hand, not every vessel element detected has enough features to be clearly assigned to a genus. If a vessel element only has

a few features, experts may still be able to assign it to an already identified genus within the sample. But by now there is no algorithm programmed behind the AI model which sets the results of the individual BBs of a sample in relation to each other. How vessel elements that do not have sufficient characteristics and cannot be assigned to any other genus contained in the sample will actually be displayed in a later application has to be decided on a case-by-case basis by an expert. Currently, experts would have to review the BBs with low confidence and categorize them into (a) cells from a genus that has already been identified in the sample, (b) cells of an unknown genus for which no references are available, or (c) if only very few cells are observed, the respective taxon is below the detection limit. Another reason for a difficult differentiation is the similarity of the characteristics of some genera (Helmling et al. 2018; Ilvessalo-Pfäffli 1995; Nieradzik et al. 2024a) in pulp, paper or other fiber materials. It has been confirmed that genera that are difficult for experts to differentiate (confusion partners) are also difficult for models to distinguish, as other studies on images of solid wood structures have already shown at the macroscopic or microscopic level (Ravindran et al. 2020; Silva et al. 2022b). For a later planned application, this will mean that only groups of recognized genera can be output. Some genera are easier to classify than others, but the model must make a decision for the classification of each recognized vessel element or, in the worst case, another cell type or structure on the slide, regardless of how well or poorly features of the object are recognized. It is therefore important to be able to use the confidences as threshold values. In this way, uncertain decisions can be filtered out more easily. It is important to bear in mind that optimizing the system means it will be unable to recognize genera it has not been trained to identify. Softwoods show less clear characteristics than hardwoods. The size and shape of the cross-field pits, when no helical (wall) thickenings are present, are the diagnostic feature (IAWA Committee 2004). For reliable identification, these small structures must be aligned optimally on the slide.

4 Conclusions

The study aimed to compare the performance of AI models with the experts' examination of wood. Pattern recognition proved to be essential in both experts' and AI analysis. The AI models were not trained with softwoods, which is clearly reflected in the results. Nevertheless, the models are able to recognize relevant cells in large overview images and classify them based on their features. The application of object probability thresholds for further evaluation enabled the reduction of false positives and focused only on clearly

recognizable and classifiable cells, similar to experts' analysis practice. The results showed that a threshold of 0.7 for object probability and 0.75 for genus probability gave the best balance between accuracy and minimizing the loss of useful data. Higher thresholds led to an increased number of false negatives, while lower thresholds significantly reduced the number of evaluable bounding boxes. The investigation of the false positive bounding boxes revealed that softwood, when incorrectly detected in the classification, was mainly confused with the genera *Eucalyptus*, *Liquidambar* and *Schima*. Nevertheless, an overall accuracy of 0.94 was achieved at the bounding box level. The AI models performed exceptionally well on pure hardwood samples. They achieved an accuracy of 0.99 at the bounding box level. Further investigations on softwood detection and classification will probably improve the performance of the AI models. However as softwood has very different kinds of characteristics a completely different approach will have to be developed to that end. Overall, this study provides valuable insights into the potential and limitations of using AI in the anatomical study of wood. It highlights the importance of including real samples and training data to minimize errors and improve the accuracy of AI models.

Acknowledgments: The authors would like to sincerely thank Anne Wettich, Doris Helm, Claudia Piehl and Sergej Kaschuro for the great lab work and especially for the super accurate preparation of the slides. Many thanks to Christina Waitkus for her help with the images and Dr. Hans-Georg Richter as well as Dr. Gerald Koch for their attentive reading and constructive comments. All Thünen Institute of Wood Research.

Research ethics: Not applicable.

Informed consent: Not applicable.

Author contributions: Conceptualization, investigation: J.S.-R., S.H., L.N. S.W., T.W., H.S. and A.O.; formal analysis: J.S.-R., S.H., L.N. S.W., T.W., J.H., H.S. and A.O.; writing – original draft: J.S.-R.; writing – review and editing: J.S.-R., S.H., L.N., H.S., A.O.; microscopic sample analysis: J.S.-R., S.H., T.L., V.H., I.H. and A.O.; project administration: H.S. and A.O. All authors have read and agreed to the final version of the manuscript.

Use of Large Language Models, AI and Machine Learning Tools: DeepL translator and write was used for language proof reading.

Conflict of interest: The authors state no conflict of interest.

Research funding: This study was supported by the German Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) via the Fachagentur Nachwachsende Rohstoffe e. V. (FNR).

Data availability: The raw data can be obtained on request from the corresponding author.

References

- Beeckman, H., Blanc-Jolivet, C., Boeschoten, L., Braga, J.W., Cabezas, J.A., Chaix, G., Crameri, S., Degen, B., Deklerck, V., Dormontt, E., et al. (2020). Schmitz, N. (Ed.). *Overview of current practices in data analysis for wood identification. A guide for the different timber tracking methods*. Global Timber Tracking Network, GTTN Secretariat, European Forest Institute and Thünen Institute.
- Beeckman, H., Cabezas Martínez, J.A., Cervera, M.T., Espinoza, E., Fernández-Golfín, J.I., Fernández-Golfín, J.I., Gasson, P., Hermanson, J.C., Arteaga, M.J., Koch, G., et al. (2019). Schmitz, N. (Ed.). *The timber tracking tool infogram. Overview of wood identification methods' capacity*. Global Timber Tracking Network, GTTN Secretariat, European Forest Institute and Thünen Institute.
- Castro, S. (2017). Fast Krippendorff: fast computation of Krippendorff's alpha agreement measure, Available at: <https://github.com/pln-fing-udelar/fast-krippendorff>.
- Dormontt, E.E., Boner, M., Braun, B., Breulmann, G., Degen, B., Espinoza, E., Gardner, S., Guillery, P., Hermanson, J.C., Koch, G., et al. (2015). Forensic timber identification: it's time to integrate disciplines to combat illegal logging. *Biol. Conserv.* 191: 790–798.
- European Union (2023). Regulation (EU) 2023/1115 of the European Parliament and of the Council of 31 May 2023 on the making available on the Union market and the export from the Union of certain commodities and products associated with deforestation and forest degradation and repealing Regulation (EU), Available at: <http://data.europa.eu/eli/reg/2023/1115/oj>.
- FAO (2025). *FAOSTAT*, <https://www.fao.org/faostat/en/#data/FO> (Accessed 07 Jul 2025).
- Flaig, M.L., Berger, J., Helmling, S., Olbrich, A., Schaffrath, H.J., Zahn, D., and Saake, B. (2024). Chemotaxonomic and anatomic wood species identification in bleached pulp: blind test and method validation. *Holzforschung* 78: 487–502.
- Franklin, G. (1945). Preparation of thin sections of synthetic resins and wood-resin composites, and a new macerating method for wood. *Nature* 155: 51.
- Gasson, P. (2011). How precise can wood identification be? Wood anatomy's role in support of the legal timber trade, especially CITES. *IAWA J.* 32: 137–154.
- He, X., Pelt, D.M., Gao, J., Gravendeel, B., Zhu, P., Chen, S., Qiu, J., and Lens, F. (2024). Machine learning-based wood anatomy identification: towards anatomical feature recognition. *IAWA J.* 45: 457–475.
- Helmling, S., Olbrich, A., Heinz, I., and Koch, G. (2018). Atlas of vessel elements: identification of Asian timbers. *IAWA J.* 39: 249–352.
- Helmling, S., Olbrich, A., Tepe, L., and Koch, G. (2016). Qualitative and quantitative characteristics of macerated vessels of 23 mixed tropical hardwood (MTH) species: a data collection for the identification of wood species in pulp and paper. *Holzforschung* 70: 839–844.
- Hwang, S.-W. and Sugiyama, J. (2021). Computer vision-based wood identification and its expansion and contribution potentials in wood science: a review. *Plant Methods* 17: 47.
- IAWA Committee (1989). IAWA list of microscopic features for hardwood identification. In: *IAWA Bulletin*, Vol. 10. IAWA, Leiden, pp. 219–332.
- IAWA Committee (2004). IAWA list of microscopic features for softwood identification. *IAWA J.* 25: 1–70.

- Ilvessalo-Pfäffli, M.-S. (1995). *Fiber atlas: identification of papermaking fibers*. Springer Science & Business Media, Springer Verlag, Berlin, Heidelberg.
- Kanninen, M. (2010). Plantation forests: global perspectives. In: *Ecosystem goods and services from plantation forests*. Routledge, London, pp. 1–15.
- Krippendorff, K. (2004). *Content analysis: an introduction to its methodology*. Sage Publications, Thousand Oaks.
- Lens, F., Liang, C., Guo, Y., Tang, X., Jahanbanifard, M., da Silva, F.S.C., Ceccantini, G., and Verbeek, F.J. (2020). Computer-assisted timber identification based on features extracted from microscopic wood sections. *IAWA J.* 41: 660–680.
- Low, M.C., Schmitz, N., Boeschoten, L.E., Cabezas, J.A., Cramm, M., Haag, V., Koch, G., Meyer-Sand, B.R., Paredes-Villanueva, K., Price, E., et al. (2022). Tracing the world's timber: the status of scientific verification technologies for species and origin identification. *IAWA J.* 44: 63–84.
- Nieradzik, L., Sieburg-Rockel, J., Helmling, S., Keuper, J., Weibel, T., Olbrich, A., and Stephani, H. (2024a). Automating wood species detection and classification in microscopic images of fibrous materials with deep learning. *Microsc. Microanal.* 30: 508–520.
- Nieradzik, L., Stephani, H., Sieburg-Rockel, J., Helmling, S., Olbrich, A., and Keuper, J. (2024b). Challenging the black box: a comprehensive evaluation of attribution maps of CNN applications in agriculture and forestry. *VISGRAPP 2024* 2: 483–492.
- Nieradzik, L., Stephani, H., Sieburg-Rockel, J., Helmling, S., Olbrich, A., Wrage, S., and Keuper, J. (2024c). WoodYOLO: a novel object detector for wood species detection in microscopic images. *Forests* 15: 1910.
- Qamar, S., Baba, A.I., Verger, S., and Andersson, M. (2024). Segmentation and characterization of macerated fibers and vessels using deep learning. *Plant Methods* 20: 126.
- Ravindran, P., Thompson, B.J., Soares, R.K., and Wiedenhoef, A.C. (2020). The XyloTron: flexible, open-source, image-based macroscopic field identification of wood products. *Front. Plant Sci.* 11: 1015.
- Richter, H.G. and Dallwitz (2000-onwards). Commercial timbers: descriptions, illustrations, identification, and information retrieval, <https://www.delta-intkey.com/wood/en/index.htm> (Accessed 08 July 2025).
- Silva, J.L., Bordalo, R., Pissarra, J., and de Palacios, P. (2022a). Computer vision-based wood identification: a review. *Forests* 13: 2041.
- Silva da, N.R., Deklerck, V., Baetens, J.M., Van den Bulcke, J., De Ridder, M., Rousseau, M., Bruno, O.M., Beeckman, H., Van Acker, J., De Baets, B., et al. (2022b). Improved wood species identification based on multi-view imagery of the three anatomical planes. *Plant Methods* 18: 79.
- Thünen-Institut (2025). *Scientific wood collection (Xylothek)*, <https://www.thuenen.de/en/thuenen-institute/infrastructure/the-thuenen-centre-of-competence-on-the-origin-of-timber/the-scientific-wood-collection-xylothek-1> (Accessed 08 July 2025).
- Wheeler, E.A. (2004-onwards). *InsideWood*, <https://insidewood.lib.ncsu.edu/search;jsessionid=QlhMElom39gUT2qFJDnuE6vLudO1GnIyCEP340Bx?0> (Accessed 08 July 2025).