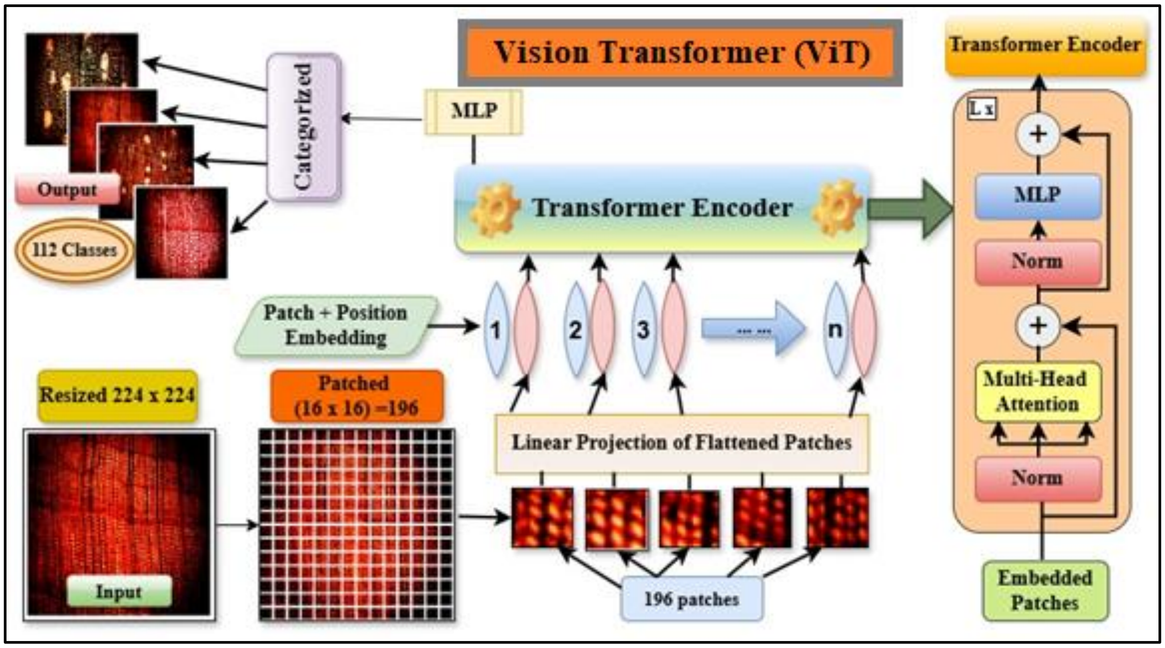
Categorization of Microscopic Wood Images with Transfer Learning Approach on Pretrained Vision Transformer Models

Kenan Kılıç *

* Corresponding author: kenan.kilic@bozok.edu.tr

DOI: 10.15376/biores.20.3.6394-6405

GRAPHICAL ABSTRACT



Categorization of Microscopic Wood Images with Transfer Learning Approach on Pretrained Vision Transformer Models

Kenan Kılıç *

Four Vision Transformer (ViT)-based models were optimized to classify microscopic wood images. The models were DeiT, Google ViT, BeiT, and Microsoft Swin Transformer. Training was performed on a set enriched with data augmentation techniques. The generalization ability of the model was strengthened by increasing the number of images for each class with data augmentation. The dataset used in the study consisted of 112 different species belonging to 30 families, 37 of which were coniferous and 75 were angiosperms. The samples had been softened, cut into thin sections, colored with the triple staining method, and imaged with fixed magnification. The Google ViT model was the most successful, with 99.40% accuracy. The DeiT model, which stood out with its data efficiency, ranked second with 98.51% accuracy, while the BEiT and Microsoft Swin Transformer models reached 96.43% and 98.21% accuracy, respectively. The Microsoft Swin Transformer model required the least training time. Data augmentation techniques improved the performance of all models by 3% to 5%, thus increasing the resistance of the models to overfitting and providing more robust predictions. It was found that ViT-based models gave superior performance in microscopic wood image classification tasks and that data augmentation significantly improved model performance.

DOI: 10.15376/biores.20.3.6394-6405

Keywords: Wood products industrial engineering; Wood classification; Vision transformer; Deep learning; Computer vision

*Contact information: Department of Design, Yozgat Vocational School, Yozgat Bozok University, 66200 Yozgat, Türkiye; *Corresponding author: kenan.kilic@bozok.edu.tr*

INTRODUCTION

Identification and classification of tree species is a critical and important step in understanding their biodiversity, role in the ecosystem, economic value, and cultural importance. It is of great importance to correctly identify the species to use wood correctly and efficiently (Wheeler and Baas 1998). Microscopic wood anatomy is a widely used basic method for the classification of coniferous and broad-leaved trees and has become more objective and scalable with modern image processing techniques (Filho *et al.* 2014). Species are usually easy to identify when organs, such as flowers, leaves, or seeds, are present. However, once the tree is processed, its identification can become quite challenging. In this case, identification is based solely on the macroscopic and microscopic properties of the wood (Khalid *et al.* 2008). Traditionally, professional woodworkers have classified wood based on macroscopic and microscopic features to identify wood species. However, these methods depend on the knowledge and experience of experts and can be time-consuming, impractical, costly, and inadequate for classification processes requiring high accuracy (Mohan *et al.* 2014; Rajagopal *et al.* 2019). This poses a major problem, especially for industries where large quantities of wood species need to be identified in a short period of time (Kırbaş and Çifci 2022). In this context, machine learning-based, deep learning-based computer vision approaches can offer an important solution in the development of faster and more accurate wood species identification methods.

Computer-aided machine vision-based systems based on visual and textural features are gaining increasing interest for automatic wood species identification (Herrera-Poyatos *et al.* 2024). Most machine vision-based identification systems have been developed for use in laboratory environments (Tou *et al.* 2007; Khalid *et al.* 2008; Hermanson *et al.* 2013). In Hermanson *et al.* (2013), XyloTron, a wood species identification system with field application, was developed by the Forest Products Laboratory of the United States Department of Agriculture (USDA). In recent years, researchers have adapted deep learning approaches for feature extraction and classification of tree images at different scales. Hafemann *et al.* (2014) constructed convolutional neural network (CNN) models for classifying and identifying macroscopic (41 classes) and microscopic (112 species) images of wood.

Tang *et al.* (2017) proposed automatic wood species identification methods with macroscopic images of 60 tropical timber species. Kwon *et al.* (2017) developed an automatic identification system to identify five different Korean softwood species. In these studies, macroscopic images taken from the cross-section of the wood and recorded with a digital camera or a smartphone camera were used. Ravindran *et al.* (2018) used transfer learning method with CNN models to identify 10 neotropical species belonging to the Meliaceae family. Machine learning and deep learning methods have been often used for the classification of wood species. However, according to the current literature, there has been no study on wood categorization with vision transformers (ViT), which is the subject of this research.

Transformers are the name given to models that use a self-attention mechanism that independently evaluates the importance of each component of the input data (Maurício *et al.* 2023). Transformers are models developed for the analysis of sequential data and have achieved great success, especially due to their self-attention mechanism (Vaswani *et al.* 2017). These models optimize information transfer by considering the relationship between each component of the input data. Transformers, which have made breakthroughs in fields, such as natural language processing (NLP) and computer vision, are also widely used in tasks, such as image classification and object detection, as an alternative to convolutional neural networks (CNN) (Dosovitskiy *et al.* 2016). Multi-layer deep learning architectures, such as CNN, have high GPU utilization (Kılıç *et al.* 2025). A similar situation also exists in ViT approaches.

In recent years, the success of deep learning models in image classification has enabled the development of new approaches for the identification and classification of wood species. Transformer-based models have demonstrated high performance with their attention mechanisms on visual data and broken new ground in the field of image processing. In this study, automatic classification of wood species is considered using the Vision Transformer (ViT) model. ViT offers more effective classification performance compared to traditional CNN due to its attention mechanisms that analyze images in parts and capture the global context. The study aims to demonstrate the adaptation of the ViT model to the task of wood species classification and the advantages it provides in this context.

EXPERIMENTAL

Data Set

In this study, a database containing 112 different forest species, catalogued by the Laboratory of Wood Anatomy of the Federal University of Parana, was used. Obtaining the images had included the steps of softening the wood by boiling, cutting thin slices with a microtome, coloring with the triple staining technique, dehydration with an alcohol series, and recording the images with an Olympus Cx40 microscope at 100× magnification. A total of 2,240 microscopic images were obtained in uncompressed PNG format and at a resolution of 1024 x 768 pixels. The database contains species belonging to a total of 85 genera and 30 families, 37 of which are coniferous (23 genera, 8 families) and 75 leafy (62 genera, 22 families) (Filho *et al.* 2014). Table 1 shows softwood species (gymnosperms), Table 2 shows hardwood species (angiosperms).

Examples of microscopic wood images used in the research are given in Fig. 1.

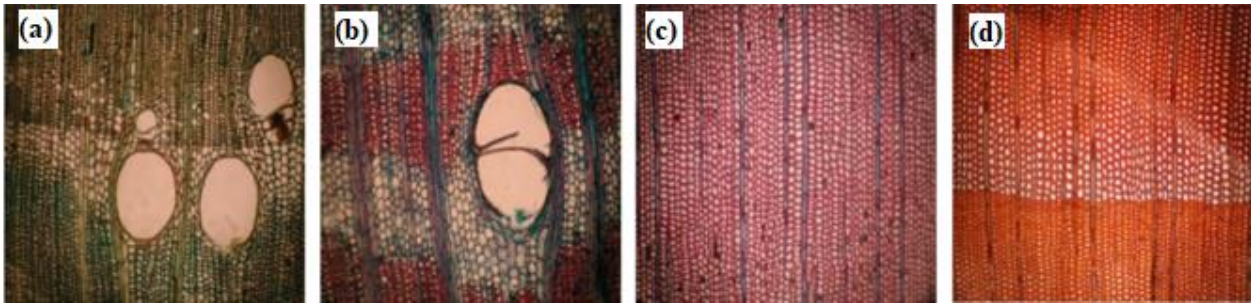


Fig. 1. Some examples of microscopic wood images used in the study: (a) *Cedrela fissilis*, (b) *Ficus gomelleira*, (c) *Tetraclinis articulata*, and (d) *Cedrus libani*

Table 1. Softwood Species (Gymnosperms)

ID	Family	Genus	Species	ID	Family	Genus	Species
1	Ginkgoaceae	*Ginkgo*	*biloba*	20	Pinaceae	*Cedrus*	*atlantica*
2	Araucariaceae	*Agathis*	*beccarii*	21	Pinaceae	*Cedrus*	*libani*
3	Araucariaceae	*Araucaria*	*angustifolia*	22	Pinaceae	*Cedrus*	*sp*
4	Cephalotaxaceae	*Cephalotaxus*	*drupacea*	23	Pinaceae	*Keteleeria*	*fortunei*
5	Cephalotaxaceae	*Cephalotaxus*	*harringtonia*	24	Pinaceae	*Picea*	*abies*
6	Cephalotaxaceae	*Torreya*	*nucifera*	25	Pinaceae	*Pinus*	*arizonica*
7	Cupressaceae	*Calocedrus*	*decurrens*	26	Pinaceae	*Pinus*	*caribaea*
8	Cupressaceae	*Chamaecyparis*	*formosensis*	27	Pinaceae	*Pinus*	*elliottii*
9	Cupressaceae	*Chamaecyparis*	*pisifera*	28	Pinaceae	*Pinus*	*greggii*
10	Cupressaceae	*Cupressus*	*arizonica*	29	Pinaceae	*Pinus*	*maximinoi*
11	Cupressaceae	*Cupressus*	*lindleyi*	30	Pinaceae	*Pinus*	*taeda*
12	Cupressaceae	*Fitzroya*	*cupressoides*	31	Pinaceae	*Pseudotsuga*	*macrolepsis*
13	Pinaceae	*Larix*	*laricina*	32	Pinaceae	*Tsuga*	*canadensis*
14	Pinaceae	*Larix*	*leptolepis*	33	Pinaceae	*Tsuga*	*sp*
15	Pinaceae	*Larix*	*sp*	34	Podocarpaceae	*Podocarpus*	*lambertii*
16	Cupressaceae	*Tetraclinis*	*articulata*	35	Taxaceae	*Taxus*	*baccata*
17	Cupressaceae	*Widdringtonia*	*cupressoides*	36	Taxodiaceae	*Sequoia*	*sempervirens*
18	Pinaceae	*Abies*	*religiosa*	37	Taxodiaceae	*Taxodium*	*distichum*
19	Pinaceae	*Abies*	*vejarii*				

Table 2. Hardwood Species (Angiosperms)

ID	Family	Genus	Species	ID	Family	Genus	Species
38	Ephedraceae	*Ephedra*	*californica*	76	Lauraceae	*Nectandra*	*sp*
39	Lecythidaceae	*Cariniana*	*estrellensis*	77	Lauraceae	*Ocotea*	*porosa*
40	Lecythidaceae	*Couratari*	*sp*	78	Lauraceae	*Persea*	*racemosa*
41	Lecythidaceae	*Eschweilera*	*matamata*	79	Annonaceae	*Porcelia*	*macrocarpa*
42	Lecythidaceae	*Eschweilera*	*chartacea*	80	Magnoliaceae	*Magnolia*	*grandiflora*
43	Sapotaceae	*Chrysophyllum*	*sp*	81	Magnoliaceae	*Talauma*	*ovata*
44	Sapotaceae	*Micropholis*	*guyanensis*	82	Melastomataceae	*Tibouchina*	*sellowiana*
45	Sapotaceae	*Pouteria*	*pachycarpa*	83	Myristicaceae	*Virola*	*oleifera*
46	Fabaceae-Cae.	*Copaifera*	*trapezifolia*	84	Myrtaceae	*Campomanesia*	*xanthocarpa*
47	Fabaceae-Cae.	*Eperua*	*falcata*	85	Myrtaceae	*Eucalyptus*	*globulus*
48	Fabaceae-Cae.	*Hymenaea*	*courbaril*	86	Myrtaceae	*Eucalyptus*	*grandis*
49	Fabaceae-Cae.	*Hymenaea*	*sp*	87	Myrtaceae	*Eucalyptus*	*saligna*
50	Fabaceae-Cae.	*Schizolobium*	*parahyba*	88	Myrtaceae	*Myrcia*	*racemulosa*
51	Fabaceae-Fab.	*Pterocarpus*	*violaceus*	89	Vochysiaceae	*Erisma*	*uncinatum*
52	Fabaceae-Mim.	*Acacia*	*tucunamensis*	90	Vochysiaceae	*Qualea*	*sp*
53	Fabaceae-Mim.	*Anadenanthera*	*colubrina*	91	Vochysiaceae	*Vochysia*	*laurifolia*
54	Fabaceae-Mim.	*Anadenanthera*	*peregrina*	92	Proteaceae	*Grevillea*	*robusta*
55	Fabaceae-Fab.	*Dalbergia*	*jacaranda*	93	Proteaceae	*Grevillea*	*sp*
56	Fabaceae-Fab.	*Dalbergia*	*spruceana*	94	Proteaceae	*Roupala*	*sp*
57	Fabaceae-Fab.	*Dalbergia*	*variabilis*	95	Moraceae	*Bagassa*	*guianensis*
58	Fabaceae-Mim.	*Dinizia*	*excelsa*	96	Moraceae	*Brosimum*	*alicastrum*
59	Fabaceae-Mim.	*Enterolobium*	*schomburgkii*	97	Moraceae	*Ficus*	*gomelleira*
60	Fabaceae-Mim.	*Inga*	*sessilis*	98	Rhamnaceae	*Hovenia*	*dulcis*
61	Fabaceae-Mim.	*Leucaena*	*leucocephala*	99	Rhamnaceae	*Rhamnus*	*frangula*
62	Fabaceae-Fab.	*Lonchocarpus*	*subglaucescens*	100	Rosaceae	*Prunus*	*sellowii*
63	Fabaceae-Mim.	*Mimosa*	*bimucronata*	101	Rosaceae	*Prunus*	*serotina*
64	Fabaceae-Mim.	*Mimosa*	*scabrella*	102	Rubiaceae	*Faramea*	*occidentalis*
65	Fabaceae-Fab.	*Ormosia*	*excelsa*	103	Meliaceae	*Cabralea*	*canjerana*
66	Fabaceae-Mim.	*Parapiptadenia*	*rigida*	104	Meliaceae	*Carapa*	*guianensis*
67	Fabaceae-Mim.	*Parkia*	*multijuga*	105	Meliaceae	*Cedrela*	*fissilis*
68	Fabaceae-Mim.	*Piptadenia*	*excelsa*	106	Meliaceae	*Khaya*	*ivorensis*
69	Fabaceae-Mim.	*Pithecellobium*	*jupunba*	107	Meliaceae	*Melia*	*azedarach*
70	Rubiaceae	*Psychotria*	*carthagenensis*	108	Meliaceae	*Swietenia*	*macrophylla*
71	Rubiaceae	*Psychotria*	*longipes*	109	Rutaceae	*Balfourodendron*	*riedelianum*
72	Bignoniaceae	*Tabebuia*	*roseoalba*	110	Rutaceae	*Citrus*	*aurantium*
73	Bignoniaceae	*Tabebuia*	*sp*	111	Rutaceae	*Fagara*	*rhoifolia*
74	Oleaceae	*Ligustrum*	*lucidum*	112	Simaroubaceae	*Simarouba*	*amara*
75	Lauraceae	*Nectandra*	*rigida*				

Preprocessing

Resize

In the first step, all input images were rescaled to 224 x 224 pixels with transforms. Resize((224, 224)). This is necessary to adapt to the input sizes of models, such as the ViT, and to achieve consistency in training. It also optimizes GPU memory usage and ensures stability in data loading.

In the second step, the image was transformed into a PyTorch tensor with transforms. ToTensor(). In this process, the color channel layout was converted from the PIL format (height x width x channel) layout to the PyTorch format (channel x height x width) layout, and the pixel values are normalized from the range [0, 255] to the range [0, 1].

In the last step, each pixel value was normalized with transforms. Normalize(mean, std). Here, mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225], representing the mean and standard deviation of RGB channels for the ImageNet dataset. This process allows the model to learn faster and more stably and improves performance when performing transfer learning with models trained on ImageNet.

Augmentation

This research, which was carried out to categorize microscopic wood images using ViT, aimed at creating more diversity in the training set of the model and to increase its generalization ability through data augmentation processes. The images were sized as 224 x 224. Then, a horizontally symmetric version of each image was created with a 50% probability by random horizontal flipping, allowing the model to recognize objects at different orientations. With random rotation, each image was randomly rotated between -20 and +20°, allowing the model to gain the ability to recognize objects at different angles. In addition, color variation was added, and the features of each image such as brightness, contrast, saturation and hue were randomly changed. In this way, the model can classify correctly in different lighting conditions and color changes. Finally, each image was normalized, allowing the model to learn faster and more accurately.

Data augmentation techniques were applied to prevent over-learning on the training data and to increase the generalization ability of the model. In this context, horizontal flip (RandomHorizontalFlip) with a 50% probability, random rotation between -20° and +20° (RandomRotation) and ColorJitter (change in brightness, contrast, saturation and hue values within the range of 0.2) were applied to each image during training. These transformations were applied to each training example in a random order and together in each epoch, thus increasing the model's ability to learn against different variations.

These augmentation operations produced various alternatives for each image. For example, horizontal flipping offers 2 alternatives for each image (original and flipped), while rotation and color swapping can create many more alternatives for each image. In this way, the training set, which initially started with 14 images, grew significantly, thanks to the augmentation, thereby increasing the generalization capacity of the model and reducing the risk of overfitting. According to the parameters used, 3 augmented images were created for each original image, resulting in a total of 42 augmented images. Training was performed with 56 images in each class, including the original images.

ViT-based Image Classification Methods

In 2017, the Google team proposed the Transformer structure based solely on the Attention mechanism, abandoning the traditional CNN and RNN structures to solve machine translation tasks. This innovative approach has become widely used in deep learning. In 2020, the Google team proposed the ViT model by adapting the Transformer structure to image classification tasks. The ViT reached a milestone in the application of transformers in computer vision (CV) by offering strong scalability with its "simple" and efficient design, and inspired subsequent research (Huo *et al.* 2023).

In this study, four different ViT-based models were used. Details, features, and explanations of mathematical structures of each model are given below.

DeiT (Data-efficient ımage transformer) Base Patch16 224

The DeiT model is a model proposed by Touvron *et al.* (2021) and based on ViT. DeiT is specifically focused on improving data efficiency and aims to achieve successful performance with smaller data sets. The model is trained with the help of a "teacher model" using distillation techniques.

Mathematically, the DeiT model is based on the standard ViT structure. The input image is in the form of $x \in R^{H \times W \times C}$ where, H, W, C represent the image height, width, and number of channels, respectively. This image is divided into patches of a fixed size, 16 × 16 pixel patches. Each patch is vectorized as follows,

$$x_p \in R^{N \times (P^2 \cdot C)} \quad (1)$$

where P is the patch size and $N = \frac{H \cdot W}{P^2}$ is the total number of patches. The following operations are performed on the patches,

$$z_0 = \left[x_{p1}E; x_{p2}E; \dots; x_{pN}E\right] + E_{\text{pos}} \quad (2)$$

where E is a learnable matrix encoding patch features, and E_{pos} the positional feature matrix. Then, the processing is done with the multi-head attention mechanism and feed-forward network.

Google ViT Base Patch16 224

The original Vision Transformer model proposed by Google (Dosovitskiy *et al.* 2021) is a model that adapts the pure transformer structure to image classification tasks. The original Vision Transformer model proposed by Google divides the input image into patches and feeds these patches to an Encoder.

Patch transformation:

$$x_p = Patchify(x), \quad x_p \in \mathbb{R}^{\left(N \times (P^2 \cdot C)\right)} \quad (3)$$

Here, P denotes the patch size and $N = \frac{HW}{P^2}$ denotes the total number of patches.

Attention Mechanism:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (4)$$

Here, Q, K, V represent query, key, and value matrices, respectively d_k, represents dimensionality reduction.

Finally, the classification process is performed using the [CLS] token.

BEiT (Bidirectional encoder representations from ımages) Base Patch16

The BEiT model (Bao *et al.* 2021) was developed with a pretraining method based on masked image modeling. This method is an image-adapted version of masked language modeling in language models.

Masked Patch Modeling:

$$\hat{y} = \arg\max_{y} P\left(y \mid x_{\text{masked}}\right) \quad (5)$$

Here, x_{masked} denotes the masked patch input. BEiT learns image features by estimating these masked patches.

Microsoft/swin-tiny-patch4-window7-224

Swin Transformer (Liu *et al.* 2021) performs the attention process within local windows with the shifted window mechanism. It gathers broader context information by scrolling between windows.

Attention Calculation:

$$Attention(Q,K,V) = Softmax\left(\frac{QW_Q \cdot KW_K^T}{\sqrt{d_k}}\right)VW_V \quad (6)$$

Here, Q, K, V represent Query, Key, and Value matrices, respectively, W_Q, W_K, W_V are the learnable weight matrices, d_k is the size of the key vectors, and $Softmax$ is the process that normalizes the distribution of attention.

Experimental Setup

In this experiment, different ViT models and data augmentation techniques were used to solve the problem of categorizing wood images into 112 microscopic classes. In these experiments performed on the NVIDIA Tesla T4 GPU on the Kaggle platform, three different model architectures, DeiT, Google ViT, and BEiT, and Microsoft Swin Transformer models, were optimized and used. All four models were loaded with pre-trained versions of Hugging Face and adapted to visual classification tasks such as microscopic wood categorization.

DeiT is an architecture specifically designed to provide data efficiency. This model has the ability to perform better with less data. Google ViT has a unique architecture that performs well on larger dataset sizes and is particularly successful on large datasets. BEiT, on the other hand, is a model for learning contextual representations from visual data and attracts attention with its transformer-based structures. Microsoft Swin Transformer uses a window-based approach to rendering visual data. Swin Transformer is a hybrid model designed to effectively learn local features and produce more efficient results. Dataset and data separation, a dataset consisting of microscopic images with 112 classes was used. The dataset is appropriately divided for training (70%), validation (15%), and testing (15%); 14 images for each class were divided into a training set, 3 images validation set and 3 images testing set. When data augmentation is performed, the training set consists of 56 images.

Data augmentation: Various augmentations were applied to the training data to increase the generalization ability of the model and prevent over-learning. In this way, the size and diversity of the training set was increased, and the model was able to make more accurate predictions under different conditions. In the validation and test sets, the actual performance of the model was evaluated using only normal transformations.

The AdamW optimization algorithm was used in training the models. The learning rate was initially set to 0.0001 and the weight decay value was 0.01. CrossEntropyLoss was preferred as the loss function. The ReduceLROnPlateau strategy was applied to automatically reduce the learning rate when the validation loss did not improve during training. This mechanism reduces the learning rate by a factor of 0.1 when the validation loss did not decrease for a certain period of time, allowing the model to learn more stably. In order to prevent overfitting of the model, the early stopping mechanism was triggered three times during the training period. In this way, the training process was stopped when the validation loss did not show improvement for a certain period of time, thus preventing unnecessary long training times and overlearning. A maximum of 10 epochs and batch size = 8 were used in the training process. All experiments were conducted with the same parameters, with the goal of ensuring equality.

The hardware and software environment accelerated the training process of the model using the NVIDIA Tesla T4 GPU in the Kaggle environment. PyTorch framework and Hugging Face Transformers libraries were used in the training process. Training, validation and testing processes have been successfully completed with Python and related libraries.

This experimental setup aims to compare different transformer-based models and investigate the impact of data augmentation techniques on model performance.

Evaluation metrics

In this research paper, commonly used metrics, namely F1-score, accuracy, precision, and recall, were used to evaluate the performance of machine learning classification algorithms. The formulas used to calculate these metrics are presented in Eqs. (7 to 10), respectively.

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (7)$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative values, respectively.

RESULTS AND DISCUSSION

There have been no studies conducted with ViT in the existing literature regarding wood categorization. There are different machine learning and deep learning approaches to categorize wood images. ViT technologies are new computer vision technologies that have been a research area in the last few years under the topic of deep learning.

The performances of ViT models were tested with different parameters. Table 3 presents the success of these models in categorizing images, while Table 4 shows the categorization performance of the same models on augmented images.

Table 3. Performance of ViT Models in Categorizing Images

ViT Model	Duration	Precision	Recall	F1-Score	Accuracy
DeiT	12.32 min	0.9538	0.9315	0.9208	0.9315
Google ViT	12.30 min	0.9610	0.9464	0.9401	0.9464
BEiT	13.55 min	0.9536	0.9345	0.9206	0.9345
Microsoft Swin Transformer	4.53 min	0.9603	0.9435	0.9330	0.9435

Table 3 shows that the DeiT model was quite successful, with an accuracy of 93.15% in approximately 12.32 min. The model, which exhibited high performance with a Precision value of 95.38% and a Recall value of 93.15%, also reached a value of 92.08% in terms of F1-Score. Although Google ViT was just behind this model, it attracted attention with its 94.64% accuracy and 96.10% precision values. Although BEiT lagged behind other models, it delivered impressive results with 93.45% accuracy and 95.36% precision. Microsoft Swin Transformer was the model that showed the best performance, especially in terms of speed, with 94.35% accuracy and 96.03% precision, achieving high accuracy in just 4.53 min.

Table 4. Performance of ViT Models in Categorizing Augmented Images

ViT Model	Duration	Precision	Recall	F1-Score	Accuracy
DeiT	19.44 min	0.9911	0.9851	0.9820	0.9851
Google ViT	19.24 min	0.9955	0.9940	0.9939	0.9940
BEiT	21.04 min	0.9779	0.9643	0.9607	0.9643
Microsoft Swin Transformer	10.06 min	0.9913	0.9821	0.9790	0.9821

Table 4 demonstrates the performance of similar models in tests performed on augmented images. Google ViT showed the highest success with 99.55% precision, 99.40% recall, and 99.39% F1-Score, and was the most successful result in this category with 99.40% accuracy. Although other models were also successful, DeiT maintained its high performance with an F1-Score of 98.51% and ranked second with an accuracy of 98.51%. BEiT and Microsoft Swin Transformer ranked third and fourth with an accuracy of 96.43% and 98.21%, respectively. It is observed that the accuracy and F1-Score values of all models increased significantly, especially in the augmented images, which shows how data augmentation improves the model performance.

Studies on wood categorization in the existing literature are presented in Table 5.

Table 5. Studies on Wood Categorization

Study	Method	Number of Wood Types	Accuracy Rate
Hafemann *et al.* (2014)	CNN	41 (macroscopic), 112 (microscopic)	Macroscopic: 95.77%, Microscopic: 97.32%
Tang *et al.* (2017)	CNN (Macroscopic Images)	60 tropical timber species	96.00%
Kwon *et al.* (2017)	Automatic Identification System	5 (softwood types)	99.30%
Ravindran *et al.* (2018)	CNN and Transfer Learning	10 (family Meliaceae)	97.50%
He *et al.* (2021)	CNN	41 macroscopic	98.81%
Wood Type Categorization with ViT (This research)	ViT tabanlı modeller	112 classes of microscopic images	99.40%

In the study by Hafemann *et al.* (2014), 95.77% accuracy was achieved with macroscopic images and 97.32% accuracy was achieved with microscopic images. These are higher accuracy rates than achieved using traditional CNN. ViT-based models, especially models, such as Google ViT, have outperformed these accuracy rates with 99.40% test accuracy rates. This suggests that ViT-based models may perform better on more complex and larger datasets.

Tang *et al.* (2017) classified 60 tropical timber species with 96% accuracy using macroscopic images. In this research, the ViT models used reached much higher accuracy rates and 99.40% accuracy was reached with Google ViT. This difference is evidence that ViT can perform better, especially in visual recognition tasks.

He *et al.* (2021) proposed an ensemble structure combining three deep CNN models using magnified macroscopic wood images and achieved wood species identification with up to 98.81% accuracy on two different datasets. CNN architectures perform quite well in these subjects. However, studies using ViT approach among deep learning approaches are gaining importance nowadays. Ensemble learning is difficult and time consuming. In this proposed research, it is seen that only optimization yields highly successful results.

Kwon *et al.* (2017) achieved 99.30% accuracy in identifying five different Korean softwood species. In this study, the 99.40% accuracy obtained with ViT models shows that ViT-based models offer superior success.

In Ravindran *et al.* (2018), 97.50% accuracy was achieved using CNN and transfer learning methods. ViT models have achieved high success rates, especially with Google ViT and DeiT, such as 99.40% accuracy and 98.51% accuracy, which once again confirms the success of ViT-based models in transfer learning and large datasets.

The high accuracy, speed, and generalization capabilities provided by ViT-based models, especially in visual recognition tasks, are quite promising for their applications in the field of categorizing wood species. It is revealed that ViT-based models achieve much higher accuracies in wood type classification compared to traditional methods and offer advantages in terms of speed. This is a significant improvement over previous studies in the literature and suggests that ViT-based models will become more common in future applications.

CONCLUSIONS

1. The accuracy performances of ViT-based models achieved in this work were quite high. While Google ViT exhibited the best performance with 99.40% accuracy, DeiT (98.51%) and Microsoft Swin Transformer (98.21%) also attracted attention with their high accuracy rates. These results indicate that ViT models offer overall strong performance in visual recognition tasks such as wood type classification.
2. In tests with augmented images, a significant increase in accuracy of ViT-based models of 3 to 5% was observed. This shows that data augmentation techniques improve the generalization ability of the model and provide better results.
3. ViT-based models can flexibly achieve high accuracy on different image types and classification tasks, making them suitable for various visual recognition applications.
4. It is envisaged that the method can be used and developed in areas such as wood categorization and wood defect detection. It has shown higher speed and accuracy performance compared to existing methods. It is anticipated that it can be used in many areas in the wood industry.
5. The high accuracy, speed, and generalization capabilities provided by ViT-based models offer great potential in field-applicable tasks such as wood species identification and classification. The barriers to the use of such deep learning-based systems in industry are gradually decreasing and it is expected to become more widespread.
6. Microsoft Swin Transformer model exhibited the fastest training time of 10.06 minutes, while the other models have approximately 19.44 minutes for DeiT, 19.24 minutes for Google ViT, and 21.04 minutes for BEiT. The Swin Transformer completed the normal classification task in approximately 50-66% shorter time compared to other models. This time advantage was especially evident in the classification with augmented images, indicating that the model can be used more efficiently in practical applications.

REFERENCES CITED

Bao, H., Dong, L., Piao, S., and Wei, F. (2021). "Beit: Bert pre-training of image transformers," *arXiv Preprint*, article ID 2106.08254. DOI: 10.48550/arXiv.2106.08254

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.* (2021). "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Preprint*, article ID 2010.11929. DOI: 10.48550/arXiv.2010.11929

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., and Brox, T. (2016). "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(9), 1734-1747. DOI: 10.1109/TPAMI.2015.2496141

Filho, P. L. P., Oliveira, L. S., Nisgoski, S., and Britto, A. S. (2014). "Forest species recognition using macroscopic images," *Machine Vision and Applications* 25, 1019-1031. DOI: 10.1007/s00138-014-0592-7

Hafemann, L. G., Oliveira, L. S., and Cavalin, P. (2014). "Forest species recognition using deep convolutional neural networks," in: *2014 22nd International Conference on Pattern Recognition*, Stockholm, Sweden, pp. 1103-1107. DOI: 10.1109/ICPR.2014.199

He, T., Mu, S., Zhou, H., and Hu, J. (2021). "Wood species identification based on an ensemble of deep convolution neural networks," *Wood Res*. 66(1), 1-14.

Hermanson, J., Wiedenhoeft, A., and Gardner, S. (2013). "A machine vision system for automated field–level wood identification," Regional Workshop for Asia, Pacific and Oceania on identification of timber species and origins, Beijing, China.

Herrera-Poyatos, D., Poyatos, A. H., Soldado, R. M., De Palacios, P., Esteban, L. G., Iruela, A. G., and Herrera, F. (2024). "Deep Learning methodology for the identification of wood species using high-resolution macroscopic images," in: *2024 International Joint Conference on Neural Networks (IJCNN)*, Rome, Italy, pp. 1-8. DOI: 10.48550/arXiv.2406.11772

Huo, Y., Jin, K., Cai, J., Xiong, H., and Pang, J. (2023). "Vision transformer (Vit)-based applications in image classification," in: *2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, New York, NY, USA, pp. 135-140. DOI: 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033

Khalid, M., Lee, E. L. Y., Yusof, R., and Nadaraj, M. (2008). "Design of an intelligent wood species recognition system," *International Journal of Simulation System, Science and Technology* 9(3), 9-19.

Kılıç, K., Atacak, İ., and Doğru, İ. A. (2025). "FABLDroid: Malware detection based on hybrid analysis with factor analysis and broad learning methods for android applications," *Engineering Science and Technology, an International Journal* 62, article ID 101945. DOI: 10.1016/j.jestch.2024.101945

Kırbaş, İ., and Çifci, A. (2022). "An effective and fast solution for classification of wood species: A deep transfer learning approach," *Ecological Informatics* 69, article ID 101633. DOI: 10.1016/j.ecoinf.2022.101633

Kwon, O., Lee, H. G., Lee, M. R., Jang, S., Yang, S. Y., Park, S. Y., Choi, I. G., and Yeo, H. (2017). "Automatic wood species identification of Korean softwood based on convolutional neural networks," *Journal of the Korean Wood Science and Technology* 45(6), 797-808. DOI: 10.5658/WOOD.2017.45.6.797

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp. 10012-10022. DOI: 10.1109/ICCV48922.2021.00986

Maurício, J., Domingues, I., and Bernardino, J. (2023). "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences* 13(9), article 5521. DOI: 10.3390/app13095521

Mohan, S., Venkatachalapathy, K., and Sudhakar, P. (2014). "An intelligent recognition system for identification of wood species," *Journal of Computer Science* 10(7), article 1231. DOI: 10.3844/jcssp.2014.1231.1237

Rajagopal, H., Khairuddin, A. S. M., Mokhtar, N., Ahmad, A., and Yusof, R. (2019). "Application of image quality assessment module to motion-blurred wood images for wood species identification system," *Wood Science and Technology* 53, 967-981. DOI: 10.1007/s00226-019-01110-2

Ravindran, P., Costa, A., Soares, R., and Wiedenhoeft, A. C. (2018). "Classification of CITES-listed and other neotropical Meliaceae wood images using convolutional neural networks," *Plant Methods* 14, article 25. DOI: 10.1186/s13007-018-0292-9

Tang, X. J., Tay, Y. H., Siam, N. A., and Lim, S. C. (2017). "Rapid and robust automated macroscopic wood identification system using smartphone with macro-lens," *arXiv Preprint* 2017, article ID arXiv:1709.08154. DOI: 10.48550/arXiv.1709.08154

Tou, J. Y., Lau, P. Y., and Tay, Y. H. (2007). Computer vision-based wood recognition system. *In Proceedings of the International Workshop on Advanced Image Technology* (pp. 197–202). Lisboa, Instituto Superior Técnico, Portugal.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers and distillation through attention," *arXiv Preprint* 2021, article ID 12877. DOI: 10.48550/arXiv.2012.12877

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need," *Advances in Neural Information Processing Systems*. DOI: 10.48550/arXiv.1706.03762

Wheeler, E. A., and Baas, P. (1998). "Wood identification – A review," *IAWA Journal* 19(3), 241-264. DOI: 10.1163/22941932-90001528

Article submitted: April 24, 2025; Peer review completed: June 13, 2025; Revised version received and accepted: June 14, 2025; Published: June 20, 2025.
DOI: 10.15376/biores.20.3.6394-6405