BRILL

IAWA Journal 46 (1), 2025: 131-146

iawa

Robustness of a macroscopic computer-vision wood identification model to digital perturbations of test images

Frank C. Owens [1,*], Prabu Ravindran [2,3], Adriana Costa [1], Manuel Chavesta [4], Rolando Montenegro [4], Rubin Shmulsky [1] and Alex C. Wiedenhoeft [1,2,3,5,6]

[1] *Department of Sustainable Bioproducts, Mississippi State University, Starkville, MS, USA*
[2] *Department of Botany, University of Wisconsin, Madison, WI, USA*
[3] *Center for Wood Anatomy Research, USDA Forest Service, Forest Products Laboratory, Madison, WI, USA*
[4] *Department of Wood Industry, Universidad Nacional Agraria La Molina, Lima, Peru*
[5] *Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN, USA*
[6] *Departamento de Ciências Biológicas (Botânica), Universidade Estadual Paulista — Botucatu, São Paulo, Brasil*
**Corresponding author; email: fco7@msstate.edu*
ORCID iDs: Owens: 0000-0002-5421-3269; Ravindran: 0000-0001-7240-7713; Costa: 0000-0002-8755-4609; Chavesta: 0000-0002-5774-6159; Montenegro: 0000-0002-7300-856X; Wiedenhoeft: 0000-0002-7053-8565

Accepted for publication: 20 June 2024; published online: 11 July 2024

Summary – Distribution shift, a phenomenon in machine learning characterized by a change in input data distribution between training and testing, can reduce the predictive accuracy of deep learning models. As operator and hardware conditions at the time of training are not always consistent with those after deployment, computer vision wood identification (CVWID) models are potentially susceptible to the negative impacts of distribution shift in the field. To maximize the robustness of CVWID models, it is critical to evaluate the influence of distribution shifts on model performance. In this study, a previously published 24-class CVWID model for Peruvian timbers was evaluated on images of test specimens digitally perturbed to simulate four kinds of image variations an operator might encounter in the field including (1) red and blue color shifts to simulate sensor drift or the effects of disparate sensors; (2) resizing to simulate different magnifications that could result from using different or improperly calibrated hardware; (3) digital scratches to simulate artifacts of specimen preparation; and (4) a range of blurring effects to simulate out-of-focus images. The model was most robust to digital scratches, moderately robust to red shift and smaller areas of medium-to-severe blur, and was least robust to resizing, blue shift, and large areas of medium-to-severe blur. These findings emphasize the importance of formulating and consistently applying best practices to reduce the occurrence of distribution shift in practice and standardizing imaging hardware and protocols to ensure dataset compatibility across CVWID platforms.

Keywords – computer vision, deep learning, distribution shift, illegal logging, machine learning, Peruvian wood identification, XyloTron.

Introduction

Wood identification is an invaluable tool for the deterrence of illegal logging and wood trade. Advances in computer vision, deep learning, and macroscopic imaging have led to the development of field-deployable, rapid, accurate, and economical wood identification screening tools such as the XyloTron, the XyloPhone and others (Khalid *et al.* 2008; Ravindran *et al.* 2018, 2019, 2020, 2022a,b; de Geus *et al.* 2020; Wiedenhoeft 2020; Arévalo *et al.* 2021; Hwang & Sugiyama 2021).

DOI 10.1163/22941932-bja10167

Distribution shift, a phenomenon in machine learning characterized by a change in input data distribution between training and testing, can reduce the predictive accuracy of deep learning models. As operator and hardware conditions at the time of training are not always consistent with those after deployment, computer vision wood identification (CVWID) models are potentially susceptible to the negative impacts of distribution shift in the field.

Potential sources of distribution shift in CVWID are various. Some can be found in the selection of the training specimens. For example, wood specimens might be sampled in a way that does not adequately represent the distribution of anatomical features in the greater population encountered in the field. In other cases, even if the sampling of training specimens adequately represents a local population, the distribution of anatomical features might differ in a population originating elsewhere. In the context of field-deployed CVWID systems, other sources of distribution shift could be testing on freshly felled wood when trained on dry xylarium specimens or testing on commercial timbers wherein the anatomical features might vary relative to individuals chosen for scientific collection.

Another source of distribution shift is operator-induced. This might include sanding/cutting artifacts or general inconsistencies in the quality of wood specimen surface preparation to expose the wood's anatomical features (Lopes *et al.* 2020, as discussed in Ravindran & Wiedenhoeft 2022). Ravindran *et al.* (2023) examined the influence of test specimen surface preparation on the accuracy of a CVWID model for the XyloTron platform. Results showed that while differences among the highest-quality surface preparations had little effect on accuracy, the poorest-quality preparation had a significant negative impact on model performance.

A third source of distribution shift relates to hardware specifications and settings. For example, some camera and lens systems are unequipped to control important external factors such as variations in ambient light. Other systems employ fundamentally different sensors such as those found among different smartphone models or between laptop- and smartphone-based systems such as the XyloTron (Ravindran *et al.* 2020) and XyloPhone (Wiedenhoeft 2020), which if not studied and understood can lead to cross-platform-incompatibility of image-based identification systems. In addition, many systems require regular calibration of focus, color balance, resizing, and lens clarity/distortions. These adjustments may not be readily controlled by an operator, especially if they are trained only as a field user of the device.

To maximize the robustness of deployed CVWID models, it is critical to evaluate the influence of distribution shifts on model performance. In this study, we induce distribution shift in CVWID test data by digitally perturbing macroscopic test specimen images of Peruvian woods to simulate four kinds of confounding image variations an operator might encounter or induce in the field including (1) red and blue channel shifts to simulate sensor drift or the effects of disparate sensors; (2) resizing to simulate different magnifications that could result from using different or improperly calibrated hardware; (3) digital scratches to simulate artifacts of specimen preparation; and (4) a range of blurring effects to simulate out-of-focus images that could be the result of operator error, dirty lenses, or uneven specimen surfaces. The effect these perturbations have on model accuracy will be evaluated, and suggestions for incorporating these results in future CVWID model development will be proposed. Establishing the effects of such distribution shifts will inform the standardization of imaging hardware and protocols and help determine how best to ensure the deployment of robust CVWID systems.

Materials and methods

CVWID model

The convolutional neural network (CNN; LeCun *et al.* 1989, Goodfellow *et al.* 2016) CVWID model for Peruvian timbers employed in this investigation was previously used in two studies (Ravindran *et al.* 2021, 2023). The 24-class model was trained by transfer learning on 5715 macroscopic images of 1300 specimens from xylaria housed at the U.S. Forest Service, Forest Products Laboratory (MADw, SJRw), and other institutions (BCTw, BOFw, Tw, FORIGw). The

Table 1. Class labels and specimen counts from the testing dataset.

Amburana	2	Cedrelinga	6	Myroxylon	6
Aniba	2	Chorisia	5	Ormosia	6
Aspidosperma	5	Copaifera	3	Pinus	2
BrosimumA	9	Dipteryx	5	Poulsenia	1
BrosimumU	2	Eucalyptus	28	Pouteria	4
Calycophyllum	7	Guazuma	5	Schizolobium	3
Cariniana	8	Hura	4	Swietenia	15
Cedrela	20	Maquira	2	Virola	17

Total specimens 167, 23 genera, 13 families.

training images were captured with the XyloTron computer vision wood identification system from the transverse surfaces of specimens progressively sanded to 1500 grit. Detailed specifications for the XyloTron's camera, lens, and housing can be found in Ravindran *et al.* 2020. For disparate hardware configurations or systems that rely on proprietary imaging information (e.g. most smartphones for use with systems like AIKO-KLHK (Arifin *et al.* 2020) or the XyloPhone (Wiedenhoeft 2020)), the imaging details may be unknown and/or unknowable. To be able to test controlled perturbations, we had to start with images from a known and knowable system. The model prediction pipeline was implemented using PyTorch (Paszke *et al.* 2019) and Scikit-learn (Pedregosa *et al.* 2011).

Original test dataset

The original test dataset comprised XyloTron images from 167 specimens progressively sanded to 1500 grit. The specimens were obtained from the David A. Kribs (PACw) and teaching collections at Mississippi State University (MSU). Specimens from the PACw and MSU teaching collections did not contribute images to the training dataset. The specimen counts for each class in the dataset are provided in Table 1.

As in prior publications (Ravindran *et al.* 2020, 2021, 2022a,b, 2023; Ravindran & Wiedenhoeft 2022), the term "class" refers to an image classification category and not to the rank between "division" and "order" used in taxonomy. Images that share the same class label are presumed to be of the same kind of wood — mutually indistinguishable from each other macroscopically yet distinguishable from woods in other classes. Classes often correspond to genera but not necessarily so. For example, in Table 1, the genus *Brosimum* is divided into two classes ("BrosimumA" for those species macroscopically indistinguishable from *B. alicastrum* and "BrosimumU" for those indistinguishable from *B. utile*) because those classes can be distinguished from each other at a subgeneric level. While some CVWID models have been trained to discriminate woods at the species level, we chose this class composition based on the macroscopic separability of these classes as defined herein. In this and previous publications, scientific names are italicized when referring to a taxon and unitalicized when referring to classes (as in Table 1) to establish a convention to facilitate clarity when discussing the CVWID classes in contrast to botanical taxa.

The specimens were obtained from the PACw and teaching collections at MSU. The names of the genera in the table are class labels (as opposed to botanical taxa) and hence are not italicized. The botanical name for all species can be found in Table A1 in the Appendix at 10.6084/m9.figshare.26232311.

Image perturbations and transformed test datasets

Every image in the original test dataset was subjected to the perturbations described below. After each perturbation was applied to every image in the original test dataset, the resulting perturbed images comprised a new test dataset that was used to evaluate the published Peruvian CVWID model, without any further training or fine-tuning.

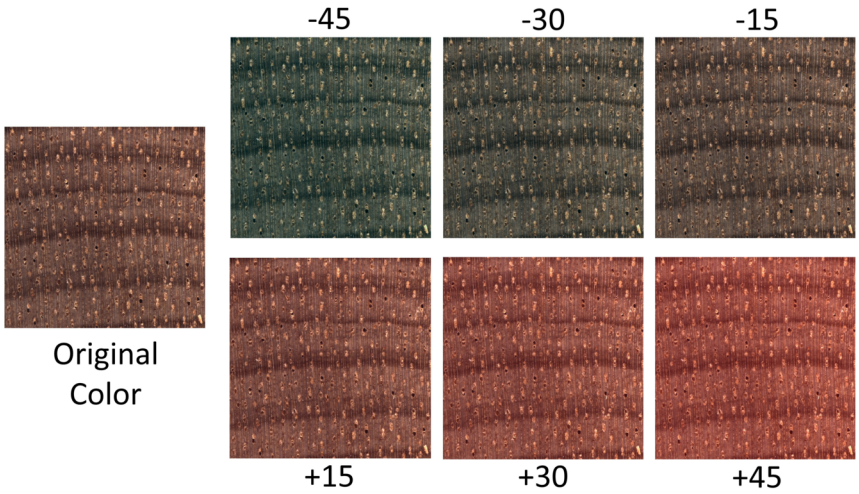

Fig. 1. The effects of red channel shift on an image (of 6.35 mm on a side) of class Virola. The six transformed images differ only in the red component value of their RGB pixel intensities. The red intensity value of each pixel (ranging from 0 to 255) has been adjusted from the original by the positive or negative value above/below each image. As the red intensity value increases, the image becomes redder. As it decreases, the image becomes less red.

Red and blue channel shifts

Each pixel in an 8-bit RGB digital image has a red channel, a green channel, and a blue channel, each with an intensity value in the range 0 to 255. Shifting the red channel by δr transforms a pixel with values (r, g, b) to a pixel with values $(r + \delta r, g, b)$, with a minimum shifted value of 0 and a maximum shifted value in the channel of 255, and likewise for the blue channel. Transformed images were generated for shifts of −45, −30, −15, 15, 30 and 45 for both δr and δb each forming six new datasets for a total of 12 datasets. Note that as the value of δr increases, the images become redder, and as the value of δr decreases, the images become less red (Fig. 1). As the value of δb increases, the images become bluer. As the value of δb decreases, the images become less blue (Fig. 2). The camera used in the XyloTron (Ravindran *et al.* 2020), the FLIR Flea3, allows for calibration of the red and blue channels but not the green channel; therefore, we did not perturb the green channel.

Resizing

Each original test image was resized by six resizing factors and center-cropped to capture a smaller tissue area while keeping the final image dimensions fixed at 2048 × 2048 pixels (Fig. 3). Resizing factors less than one, corresponding to more tissue area being captured in an image, will require image padding (say, with zeros) in order to keep the image dimensions at 2048 × 2048 pixels. For this reason, only resizing factors greater than one were studied. The resulting digitally resized images contain bilinearly interpolated pixels and do not provide finer optical resolution. Six digital resizing (zoom) factors of 1.17, 1.33, 1.50, 1.67, 1.83 and 2.0× were used, each forming its own dataset for a total of 6 sets.

Digitally simulated scratches

Five parallel lines were added to each image as simplified representations of scratches (Fig. 4). The length of the lines was chosen so that the entirety of each line was visible at all orientations. The line color was set to be the 95th percentile of image pixel intensities with a Gaussian profile opacity across the line thickness. The orientation of the scratches with respect to vertical varied over the values 0, 18, 36, 54, 72 and 90 degrees, each forming its own dataset for a total of 6 sets.

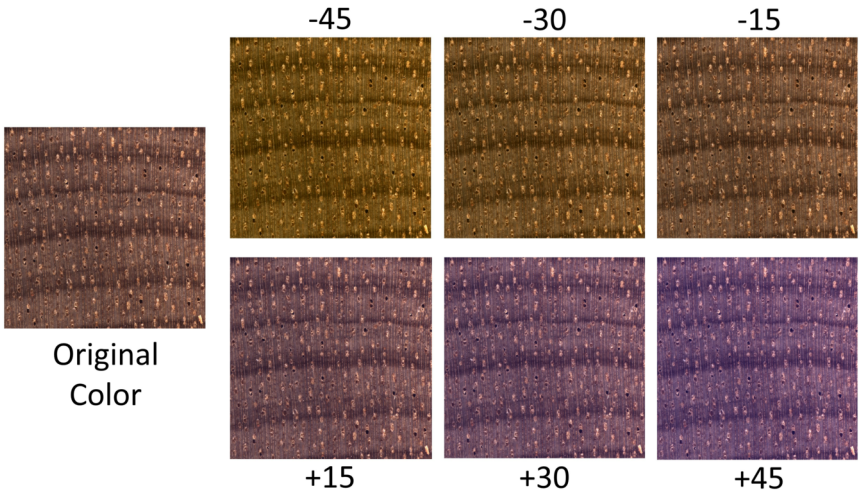

Fig. 2. The effects of blue channel shift on an image (of 6.35 mm on a side) of class Virola. The six transformed images differ only in the blue component value of their RGB pixel intensities. The blue intensity value of each pixel (ranging from 0 to 255) has been adjusted from the original by the positive or negative value above/below each image. As the blue intensity value increases, the image becomes bluer, and as it decreases, the image becomes less blue.

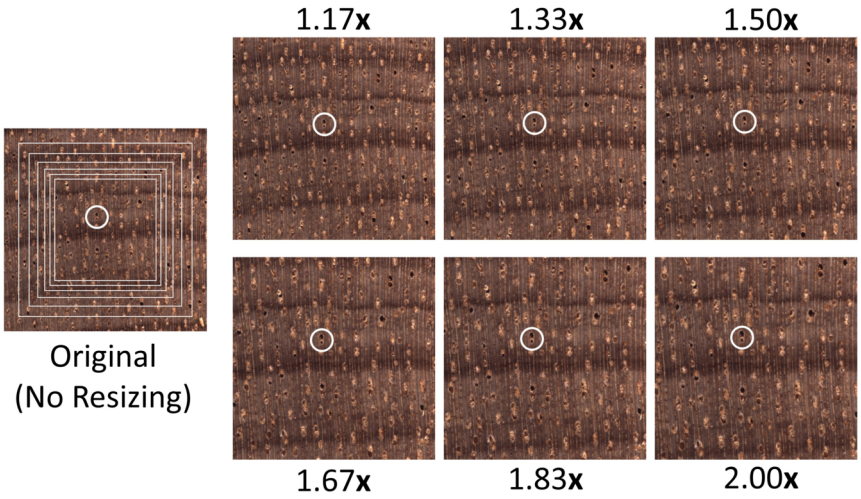

Fig. 3. The effects of resizing an image (of 6.35 mm on a side) of class Virola on anatomical feature size and position. The exemplar images show the kind of effect each resizing factor has on the size and position of anatomical features. The transformed images are center-cropped to equalize their dimensions thereby reducing the area of wood tissue captured. As the resizing factor increases, the features become larger and migrate toward the image edges. A fixed-size white circle has been added to the same tissue in each image to facilitate size and location comparisons.

Gaussian blur

Each image in the original test dataset was blurred. The blurring was varied both by magnitude and by the proportion of the image's area. The Gaussian Blur class in the ImageFilter module of the open source PIL package (Clark 2015) was employed for image blurring with the following 6 values for the radius argument: 2, 4, 6, 8, 10 and 12 (Fig. 5). The radius is the standard deviation in pixels of the Gaussian blurring kernel. The larger the value of the radius argument, the more out-of-focus the image becomes. Additionally, every image was partitioned into a 3 × 2

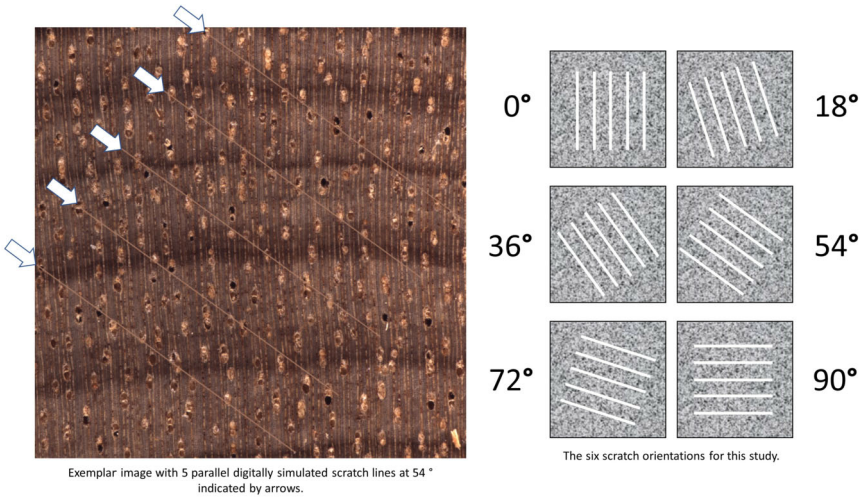

Fig. 4. Digital simulation of scratches in an image (of 6.35 mm on a side) of class Virola (left) with digital scratches angled at 54 degrees. Five parallel lines were drawn in each image as indicated by the arrows. Orientation of the scratch lines with respect to the vertical was 0, 18, 36, 54, 72 and 90 degrees shown on a grey background, right.

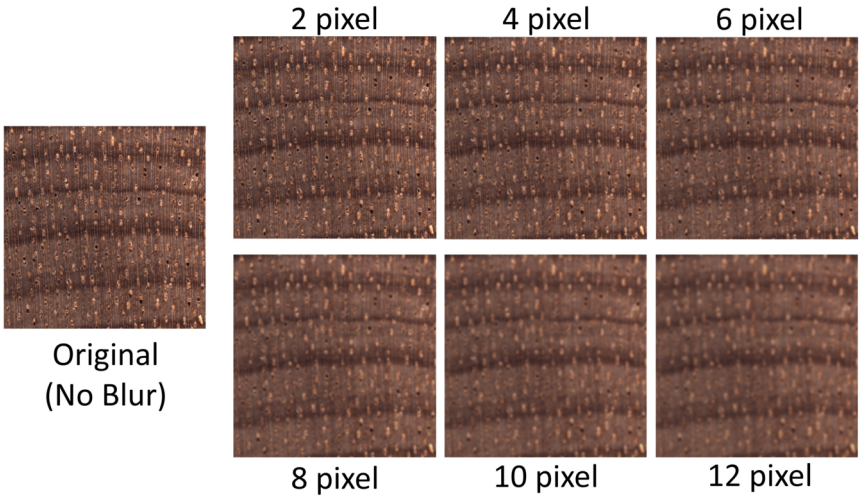

Fig. 5. Varying the magnitude of blur in an image (of 6.35 mm on a side) of class Virola. The 2-pixel image exhibits the least and the 12-pixel image the most blurring. All images are blurred in their entirety (i.e. patch value 6).

grid of patches, and the number of patches blurred (at each of the 6 standard deviation settings for the smoothing kernel) was varied from 1 (top-left patch) to 6 (full image) in raster order (Fig. 6). Table 2 and Fig. 6 show the numbers of patches as percentages of the total pixels blurred: 16.7. 33.3, 50.0, 66.7, 83.3 and 100.0%. A dataset was generated for each of 6 blur patches at each of 6 blur magnitudes for a total of 36 transformed test datasets.

In total, the digital perturbations led to the generation of 60 transformed test datasets, the details of which are summarized in Table 3. The robustness of the trained model to the perturbations was evaluated by computing the model's predictive accuracy for each test dataset.

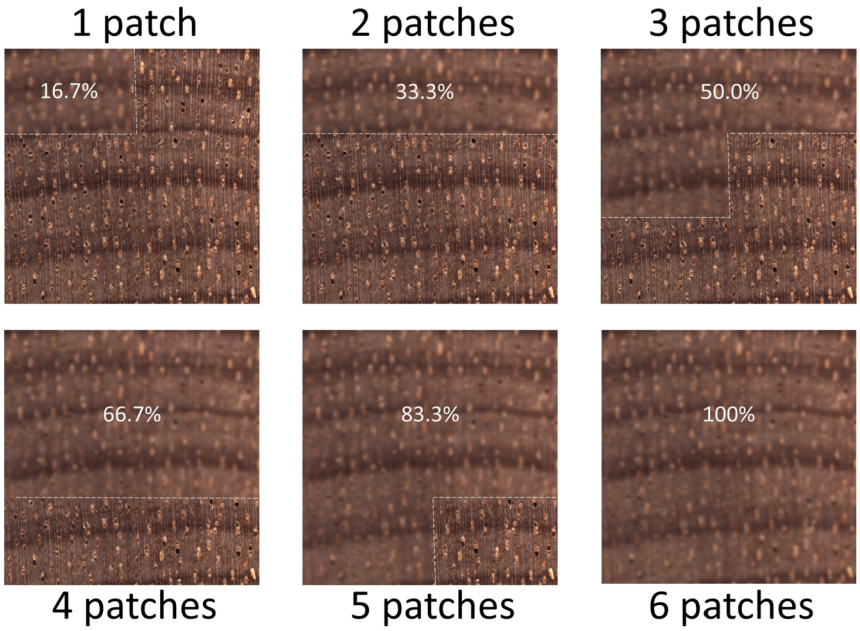

Fig. 6. Varying the number of blur patches (at the 12-pixel magnitude) in an image (of 6.35 mm on a side) of class Virola.

Table 2. Number of blurred patches and percentages of the total number of pixels blurred in the images.

Number of blur patches	Proportion of pixels blurred (%)
1	16.7
2	33.3
3	50.0
4	66.7
5	83.3
6	100.0

Table 3. Summary of perturbations and parameter settings used in this work.

Perturbation	Parameter varied	Parameter settings	No. of test datasets
Red channel shift	Shift magnitude	−45, −30, −15, 15, 30, 45 (intensity levels)	6
Blue channel shift	Shift magnitude	−45, −30, −15, 15, 30, 45 (intensity levels)	6
Resizing	Resizing factor	1.17x, 1.33x, 1.50x, 1.67x, 1.83x, 2.0x	6
Scratches	Orientation with vertical	0, 18, 36, 54, 72, 90 (degrees)	6
Gaussian blur	Kernel standard deviation and percent of image blurred	2, 4, 6, 8, 10, 12 (pixels) and 16.7%, 33.3%, 50%, 66.7%, 83.3%, 100%	6 × 6 = 36

CVWID model evaluation

The 24-class field model from Ravindran *et al.* (2021) was employed for this study "as-is" without retraining or fine-tuning to make predictions on the transformed test images. All predictions were obtained at the specimen level. The class label assigned to a specimen was obtained as the majority prediction from up to five images obtained from the specimen. Confusion matrices for the specimen-level predictions were generated for each parameter setting and are available in the Appendix at 10.6084/m9.figshare.26232311. For each type of perturbation (red-shift, blue-shift, resizing, scratches, and focus), the model's top-1 and top-2 predictive accuracies for each parameter setting were compared with the accuracy of the model when tested on the original (untransformed) dataset to evaluate change in performance. The top-1 accuracies are presented in the Results section, and the top-2 accuracies can be found in the Appendix at 10.6084/m9.figshare.26232311.

Statistical analyses

To assess any statistical differences among predictive accuracy percentages, a Cochran's Q test was employed using IBM SPSS Statistics 28 (IBM 2021). The method proposed by Dunn (1964) was used for multiple comparisons. A Bonferroni correction was used to ensure an experiment-wise error rate of 0.05.

Changes in predictive accuracy are expressed first in absolute terms along with statistical inferences and then in relative terms. The relative changes were classified into the *ad hoc* categories of "mild", "moderate" and "severe" to broadly characterize the effect of each perturbation on model performance. The baseline predictive accuracy of the model on the original, unperturbed dataset was 89.2%. If the predictive accuracy of the model on a transformed dataset differs by less than 90% of the baseline (i.e. when $98.1\% > x > 80.3\%$ in absolute terms), the change in performance was categorized as "mild". As most of the predictive accuracies greater than 80.3% and none of the accuracies greater than the baseline of 89.2% were statistically significant ($p < 0.05$) this classification seems reasonable. If the predictive accuracy of the model on a transformed dataset is between 90 and 75% that of the baseline (i.e. when $66.9\% < x < 80.3\%$ in absolute terms), the change in performance was categorized as "moderate". If the predictive accuracy of the model on a transformed dataset is less than 75% that of the baseline (i.e. when $x < 66.9\%$ in absolute terms), the change in performance was categorized as "severe". In Figs 7–10, "mild" changes in performance are indicated by grey bars, "moderate" changes in performance by yellow bars, and "severe" changes by red bars.

Results

Figures 7–10 compare the predictive accuracy of the Peruvian CVWID model when tested on the original unperturbed dataset (solid black bars) to the predictive accuracies of the model when tested on the transformed datasets for each perturbation parameter (the other bars) of red channel shift, blue channel shift, resizing, digital scratches, and focus. The striped bars are statistically different from the black bars ($p < 0.05$, per Dunn's multiple comparisons) while the solid bars are not ($p > 0.05$). Grey bars represent "mild", yellow bars represent "moderate", and red bars represent "severe" performance change.

Red and blue channel shifts

The left graph in Fig. 7 compares the predictive accuracy of the model at each red-shift value with that of the model tested on the original unperturbed image dataset. The predictive accuracies for the settings –45, –30, –15, +15, +30 and +45 differ in absolute terms by –11.4, –10.2, –4.2, +3.0, –1.8 and –4.2%, respectively. The differences for settings –45 and –30 are statistically lower ($p < 0.05$).

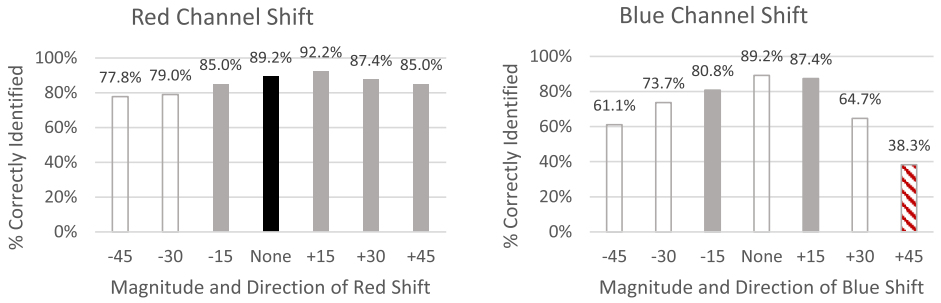

Fig. 7. Predictive accuracies of the model at various magnitudes of red and blue channel shift. The black column (None) shows the predictive accuracy of the model tested on the original untransformed images. The other bars represent the predictive accuracies when tested on the transformed datasets at each parameter setting. In relative terms, a grey bar indicates a predictive accuracy within 9.9% of the unperturbed dataset (98.1% < x < 80.3%), yellow bars indicate a predictive accuracy that is 75–90% that of the original unperturbed dataset (66.9% < x < 80.3%, where x = perturbed dataset accuracy values), and red bars indicates a predictive accuracy that is less than 75% that of the original unperturbed dataset (x < 66.9%). The values on the x-axes indicate the magnitude and direction of the pixel intensity shifts. The striped bars are statistically different from the black bars (p < 0.05, per Dunn's multiple comparisons) while the solid bars are not (p > 0.05).

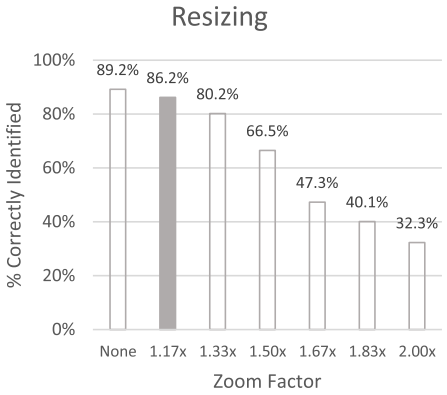

Fig. 8. Predictive accuracies of the model at various resizing factors. The black column (None) shows the predictive accuracy of the model tested on the original untransformed images. The other bars represent the predictive accuracies when tested on the transformed datasets at each parameter setting. In relative terms, a grey bar indicates a predictive accuracy within 9.9% of the unperturbed dataset (98.1% < x < 80.3%), yellow bars indicate a predictive accuracy that is 75–90% that of the original unperturbed dataset (66.9% < x < 80.3%, where x = perturbed dataset accuracy values), and red bars indicate a predictive accuracy that is less than 75% that of the original unperturbed dataset (x < 66.9%). The x-axis units, zoom factor, are resizing factors. The striped bars are statistically different from the black bars (p < 0.05) while the solid bars are not (p > 0.05, per Dunn's multiple comparisons).

The right graph in Fig. 7 compares the predictive accuracy of the model at each blue-shift value with that of the model tested on the original unperturbed image dataset. The predictive accuracies for the settings −45, −30, −15, +15, +30 and +45 differ in absolute terms by −28.1, −15.5, −8.4, −1.8, −24.5 and −50.9%, respectively. The differences for settings −45, −30, +30, and +45 are statistically lower (p < 0.05).

Resizing

Figure 8 compares the predictive accuracy of the model at each resizing factor with that of the model tested on the original unperturbed image dataset. The predictive accuracy for the factors 1.17, 1.33, 1.50, 1.67, 1.83 and 2.00× differed

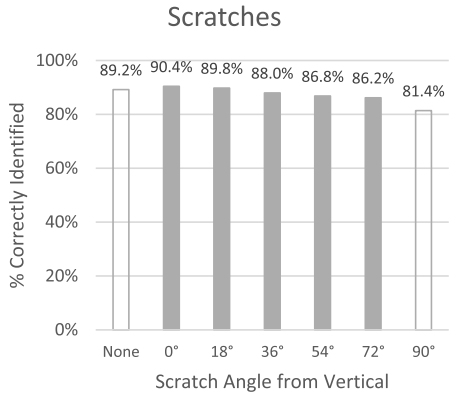

Fig. 9. Predictive accuracies of the model with five parallel scratches at various angles. The black column shows the predictive accuracy of the model tested on the original untransformed images. The other bars represent the predictive accuracies when tested on the transformed datasets at each parameter setting, and their grey color indicates a predictive accuracy within 9.9% of the unperturbed dataset (98.1% < x < 80.3%). The x-axis units for scratches are angles (from vertical). The striped bars are statistically different from the black bars (p < 0.05) while the solid bars are not (p > 0.05, per Dunn's multiple comparisons).

in absolute terms by –3.0, –9.0, –22.7, –41.9, –49.1 and –56.9%, respectively. The differences for factors 1.50, 1.67, 1.83 and 2.00× are statistically lower (p < 0.05).

Digital scratches

Figure 9 compares the predictive accuracy of the model at each scratch orientation (from vertical) with that of the model tested on the original unperturbed image dataset (no scratches). The predictive accuracy for angles 0, 18, 36, 54, 72 and 90° differed in absolute terms by +1.2, +0.6, –1.2, –2.4, –3.0 and –7.8%, respectively. Only the difference for the parameter 90° is statistically significant (p < 0.05).

Gaussian blur

Figure 10 compares the predictive accuracies of the model at each blur setting (percent of image blurred × blur magnitude) with that of the model tested on the original unperturbed image dataset (no blur). Table 4 summarizes the absolute decreases in accuracy for each parameter at each blur condition.

Discussion

The results show that many of the perturbation settings significantly impacted the predictive accuracy of the model. Additional comparison reveals that the influence of some perturbations was greater than that of others.

Color shift

Blue channel shifts resulted in lower predictive accuracies at every magnitude of perturbation than red channel shifts. The reductions in performance at the extremes are also "severe" in the case of blue shift while merely unilaterally (reduction of pixel values) "moderate" in the case of red shift. As such, the model appears to be moderately robust to red shift while less robust to blue shift. This may also suggest that for the woods in the Peruvian model, more information for wood classification is present in the blue than in the red channels in the images. Ongoing work by our team is exploring this.

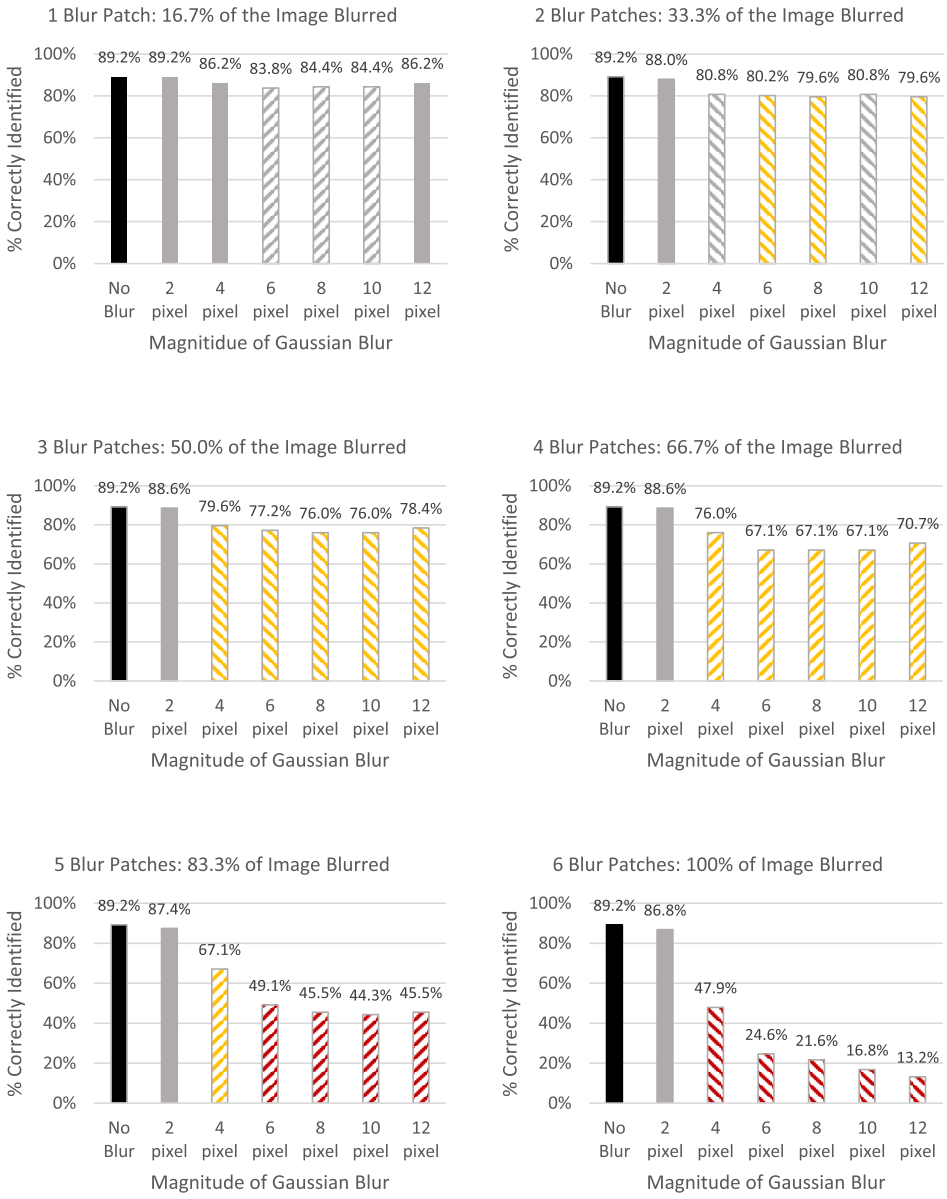

Fig. 10. Predictive accuracies under various blur settings. The black column (None) shows the predictive accuracy of the model tested on the original untransformed images. The other bars represent the predictive accuracies when tested on the transformed datasets at each parameter setting. In relative terms, a grey bar indicates a predictive accuracy within 9.9% of the unperturbed dataset ($98.1\% < x < 80.3\%$), yellow bars indicate a predictive accuracy that is 75–90% that of the original unperturbed dataset ($66.9\% < x < 80.3\%$, where x = perturbed dataset accuracy values), and red bars indicates a predictive accuracy that is less than 75% that of the original unperturbed dataset ($x < 66.9\%$). The x-axis values are standard deviations of the Gaussian blurring kernel in pixels, where greater pixel values yield greater blur. The striped bars are statistically different from the black bars ($p < 0.05$) while the solid bars are not ($p > 0.05$, per Dunn's multiple comparisons).

Table 4. Absolute differences in the model's predictive accuracy for each blur condition.

Blur magnitude	Proportion of the image blurred					
	16.7%	33.3%	50.0%	66.7%	83.3%	100%
2 pixels	0.0%	−1.2%	−0.6%	−0.6%	−1.8%	−2.4%
4 pixels	−3.0%	*−8.4%*	*−9.6%*	*−13.2%*	*−22.1%*	*−41.3%*
6 pixels	*−5.4%*	*−9.0%*	*−12.0%*	*−22.1%*	*−40.1%*	*−64.6%*
8 pixels	*−4.8%*	*−9.6%*	*−13.2%*	*−22.1%*	*−43.7%*	*−67.6%*
10 pixels	*−4.8%*	*−8.4%*	*−13.2%*	*−22.1%*	*−44.9%*	*−72.4%*
12 pixels	−3.0%	*−9.6%*	*−10.8%*	*−18.5%*	*−43.7%*	*−76.0%*

Percentages in italics indicate statistically significant decreases ($p < 0.05$, per Dunn's multiple comparisons). Blur magnitude is expressed by standard deviations of the Gaussian blurring kernel in pixels.

Resizing

The model showed progressive decreases in predictive accuracy as the resizing factor increased. Reductions in performance became "moderate" at 1.33× and progressively "severe" from 1.50×. As such, the model seems robust only to small magnitudes of digital resizing (1.17×). As noted in the materials and methods section, it is not possible to "zoom out" (that is, resize images with resizing factors less than 1.0) in digital images, as that would require capturing image data for adjacent tissue not present in the parent image.

Scratches

Digital scratches seemed to have the smallest impact on model performance. While only the 90-degree orientation resulted in a statistically significant reduction in predictive accuracy, the impact on performance was "mild". The XyloTron images used in this Peruvian wood model were oriented such that the rays were always vertical in the image. This may be important with regard to digital scratch angle, as digital scratches progressively less parallel to the rays in the image seem to have a greater impact on predictive accuracy. This is a hopeful result, as training human operators to make high-quality cuts on the transverse surface for field test specimens, whether the operators are professional trainees (e.g. customs agents or other government or law enforcement inspectors) or are university students in a laboratory setting, can be a limiting factor (Wiedenhoeft, Owens, Shmulsky and Costa, personal observations) for specimen imaging. This observation is also borne out in some published training datasets (Lopes *et al.* 2020, as detailed in Ravindran & Wiedenhoeft 2022). Our results suggest that a scientifically constructed model with training images that are essentially free of appreciable surfacing artifacts is comparatively robust to scratches, at least of the digital nature we induced in this study. The degree to which these digital scratches can serve as a surrogate for real-world scratches is still unknown.

Blur

Predictive accuracy tended to drop as the magnitude of the blur and the area of the blur increased. The model seemed robust to a 2-pixel perturbation even when the entire image was blurred. From a 4-pixel blur upward, the decrease in predictive accuracy tended to grow more "moderate" as the area of blur increased from 33.3 to 66.7%. When greater than 2/3 of the image was perturbed, reductions in performance became "severe" from 6 pixels in the case of 5 patches (83.3% blur) and from 4 pixels in the case of 6 patches (100% blur). As such, the model seems robust to large areas of slight blur and progressively less robust as the area of medium-to-severe blur increases.

The results of this study suggest that distribution shift induced by color, resizing, scratches, and blur can negatively impact the predictive accuracy of a CVWID model, especially when the magnitude of the various perturbations was

greater. To maximize model performance, it is important to prevent data drift from occurring and/or mitigate its effect by establishing and consistently applying best practices at the operator level, augmenting training data to increase model robustness, and standardizing CVWID imaging protocols and hardware specifications/settings to ensure cross-platform compatibility.

Due to time constraints and training limitations, distribution shift can likely never be eliminated at the operator level, but consistent application of best practices can reduce its impact on real-world performance. In sample preparation, it is imperative that the surface of test specimens be as flat as possible to reduce the occurrence of non-planar areas that are out of focus. It is also important to keep the surface as free as possible from scratches. Encouraging the use of portable polishing tools such as orbital sanders instead of knives could reduce the occurrence of both unevenness and cutting artifacts. In addition, operator training and field performance could be augmented by a machine learning model that classifies the degree of blur in an image before submitting the test image to a CVWID model. In a field deployment context, such a model could evaluate the blur in a test image and direct the operator to prepare a flatter surface and re-image if the blur is excessive. The focus settings of the hardware could also be evaluated with a small set of reference specimens with known degrees of non-planarity, to test if the system is showing drift. With respect to hardware, control and regular verification of hardware settings against a reference or standard such as color balancing, disabling digital resizing, and refocusing could help eliminate distribution shift in color, focus, and scale.

As eliminating all sources of operator-induced distribution shift is likely not possible, designing context-specific data-driven image augmentation (Yang *et al.* 2023) to training datasets could bolster model performance by boosting robustness to less-than-ideal test image inputs. Designing such augmentation strategies and incorporating them during training will slightly increase computation costs associated with developing the model, but with ever-more-accessible GPU resources such increases in computational costs should be a small fraction of the costs of developing CVWID models.

Implementing such a context-specific data-driven image augmentation protocol has the potential to increase model robustness and thus lessen the "deployment gap" identified in Ravindran *et al.* (2019) — the difference in predictive accuracy between test data and actual field performance — by attempting to quantify those parameters that might be expected to experience distribution shift and incorporating image augmentation to encompass expected distribution shift digitally into the training data (to be addressed in upcoming/ongoing work by the authors).

Our work using digitally perturbed test images for evaluating model robustness highlights the importance of careful attention to all aspects of CVWID development and deployment, such as sampling design, data augmentation for model training, stress testing in practical distribution shift regimes, and sensor calibration. The frequency of sensor calibration/maintenance may be dependent on the taxa in the model or on the operating conditions, and likely needs to be adapted at the individual device level (some imaging systems or their instantiations need more attention than others). Adopting and adapting this holistic methodology for development, deployment, and maintenance can lead to robust, operational CVWID systems. It should be noted that while richer data augmentation strategies can help address sensor-induced distribution shifts, the impact of anatomical distribution shifts needs to be addressed with expansive specimen sampling strategies.

Even if all sources of operator-induced distribution shift could be eliminated for one CVWID platform, distribution shift is still likely to occur when imaging with another platform as hardware specifications and settings commonly differ. Because the proliferation of CVWID models at a scale needed to address illegal logging can probably not occur unless multi-dataset compatibility can be ensured, standardizing imaging hardware, settings, and protocols could be a positive step in that direction. Such a suggestion is in keeping with other literature for other wood identification modalities, where protocol development and standardization have been called for (Beeckman *et al.* 2020; Gasson *et al.* 2021).

It may be of interest to note that CVWID is not the only wood identification modality that is impacted by distribution shift. Chemometric methods of wood identification are also likely to be influenced by distribution shift,

especially as it pertains to the difference between the age, history, and nature of training specimens (e.g. dry, often air-dried, xylarium specimens) as compared to field or trade specimens (e.g. recently felled wood, kiln-dried wood) (Kunze *et al.* 2021, Deklerck *et al.* 2022). One set of modalities that will not suffer from distribution shift between input and test data are DNA-based methods. Instead, DNA methods suffer from the degradation of specimen DNA as a function of wood processing (Yoshida *et al.* 2007; Jiao *et al.* 2014; Jiao *et al.* 2020) so rather than having a shift, some real-world specimens are more likely to lack DNA in usable condition, resulting in a failure of the method rather than a reduction in the accuracy.

Conclusions

In this study, we induced distribution shifts in CVWID model testing by digitally perturbing test image datasets to simulate the kind of data drift an operator might encounter in the field and evaluated the effects those transformed datasets had on the performance of a previously published CVWID model for Peruvian timbers. The results showed that many of the perturbation settings significantly impacted the predictive accuracy of the model. The model was most robust to digital scratches, moderately robust to red shift and smaller areas of medium-to-severe blur, and was least robust to resizing, blue shift, and large areas of medium-to-severe blur. These findings emphasize the importance of standardizing imaging hardware and protocols to reduce the occurrence of distribution shift. They also have the potential to inform training data augmentation strategies to increase the robustness of CVWID models to distribution shifts caused by operator inconsistencies, disparate hardware systems, and/or improper hardware settings.

Acknowledgements

The authors wish to gratefully acknowledge the specimen preparation and imaging efforts of Nicholas Bargren, Karl Kleinschmidt, Caitlin Gilly, Richard Soares, Adriana Costa, Sofia Silva Gerbi, Jessica Francesca Figueroa Mendoza and Flavio Ruffinatto. The software apps for image dataset collection and trained model deployment along with the weights of the trained model will be made available at https://github.com/fpl-xylotron. The unperturbed test dataset is available on reasonable request from PR. This work was supported in part by a grant from the US Department of State *via* Interagency Agreement number 19318814Y0010 to ACW and in part by research funding from the Forest Stewardship Council to ACW. PR was partially supported by a Wisconsin Idea Baldwin Grant. The authors wish to acknowledge the support of U.S. Department of Agriculture (USDA), Research, Education, and Economics (REE), Agriculture Research Service (ARS), Administrative and Financial Management (AFM), Financial Management and Accounting Division (FMAD) Grants and Agreements Management Branch (GAMB), under Agreement No. 58-0204-9-164, specifically for support of FO, RS. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture. This publication is a contribution of the Forest and Wildlife Research Center, Mississippi State University.

References

Arévalo R, Pulido ENR, Solórzano JFG, Soares R, Ruffinatto F, Ravindran P, Wiedenhoeft AC. 2021. Image based identification of Colombian timbers using the XyloTron: a proof of concept international partnership. *Colombia For.* 24(1): 5–16. DOI: 10.14483/2256201X.16700.

Arifin MR, Sugiarto B, Wardoyo R, Rianto Y. 2020. Development of mobile-based application for practical wood identification. *IOP Conf. Series: Earth and Environ. Sci.* 572(1): 012040. DOI: 10.1088/1755-1315/572/1/012040.

Beeckman H, Blanc-Jolivet C, Boeschoten L, Braga JWB, *et al.* 2020. 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, Hamburg. DOI: 10.13140/RG.2.2.21518.79689.

Clark A. 2015. Pillow (PIL Fork) Documentation, readthedocs. Available online at: https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf (accessed: 21 June 2023).

de Geus AR, da Silva SF, Gontijo AB, Silva FO, Batista MA, Souza JR. 2020. An analysis of timber sections and deep learning for wood species classification. *Multimed Tools Appl.* 79: 34513–34529. DOI: 10.1007/s11042-020-09212-x.

Deklerck V, Fowble KL, Coon AM, Espinoza EO, Beeckman H, Musah RA. 2022. Opportunities in phytochemistry, ecophysiology and wood research via laser ablation direct analysis in real time imaging-mass spectrometry. *New Phytol.* 234(1): 319–331. DOI: 10.1111/nph.17893.

Dunn OJ. 1964. Multiple comparisons using rank sums. *Technometrics* 6(3): 241–252.

Gasson PE, Lancaster CA, Young R, Redstone S, Miles-Bunch IA, Rees G, Guillery RP, Parker-Forney M, Lebow ET. 2021. WorldForestID: Addressing the need for standardized wood reference collections to support authentication analysis technologies; a way forward for checking the origin and identity of traded timber. *Plants People Planet* 3(2): 130–141. DOI: 10.1002/ppp3.10164.

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. MIT Press, Cambridge, MA.

Hwang SW, Sugiyama J. 2021. Computer vision-based wood identification and its expansion and contribution potentials in wood science: a review. *Plant Methods* 17: 47. DOI: 10.1186/s13007-021-00746-1.

IBM. 2021. IBM SPSS Statistics for Windows, Version 28.0. IBM, Armonk, NY.

Jiao L, Lu Y, He T, Guo J, Yin Y. 2020. DNA barcoding for wood identification: global review of the last decade and future perspective. *IAWA J.* 41(4): 620–643. DOI: 10.1163/22941932-bja10041.

Jiao L, Yin Y, Cheng Y, Jiang X. 2014. DNA barcoding for identification of the endangered species *Aquilaria sinensis*: comparison of data from heated or aged wood samples. *Holzforschung* 68: 487–494. DOI: 10.1515/hf-2013-0129.

Khalid M, Lew E, Lee Y, Yusof R, Nadaraj M. 2008. Design of an intelligent wood species recognition system. *Int. J. Simul. Syst. Sci. Technol.* 9: 9–19. Available online at https://ijssst.info/Vol-09/No-3/paper2.pdf.

Kunze DC, Pastore TC, Rocha HS, Lopes PVDA, Vieira RD, Coradin VT, Braga JW. 2021. Correction of the moisture variation in wood NIR spectra for species identification using EPO and soft PLS2-DA. *Microchem. J.* 171: 106839. DOI: 10.1016/j.microc.2021.106839.

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1(4): 541–551. DOI: 10.1162/neco.1989.1.4.541.

Lopes DJV, Burgreen GW, Entsminger ED. 2020. North American hardwoods identification using machine-learning. *Forests* 11(3): 298. DOI: 10.3390/f11030298.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, *et al.* 2019. Pytorch: An imperative style, high-performance deep learning library. ArXiv: 8026–8037. DOI: 10.48550/arXiv.1912.01703.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, *et al.* 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12(85): 2825–2830.

Ravindran P, Wiedenhoeft AC. 2022. Caveat emptor: On the need for baseline quality standards in computer vision wood identification. *Forests* 13(4): 632, 1–19. DOI: 10.3390/f13040632.

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

Ravindran P, Ebanyenle E, Ebeheakey AA, Bonsu KA, Lambog O, Soares R, Costa A, Wiedenhoeft AC. 2019. Image based identification of Ghanaian timbers using the XyloTron: opportunities, risks and challenges. In: *Proceedings of NeurIPS 2019 workshop on machine learning for the developing world: challenges and risks of ML4D. Thirty-third Conference on neural information processing systems*, Vancouver, BC, Canada, 12/8/2019–12/14/2019, available online at https://arxiv.org/abs/1912.00296.

Ravindran P, Thompson BJ, Soares RK, Wiedenhoeft AC. 2020. The XyloTron: Flexible, Open-Source, Image-Based Macroscopic Field Identification of Wood Products. *Front Plant Sci.* 11: 1015. DOI: 10.3389/fpls.2020.01015.

Ravindran P, Owens FC, Wade AC, Vega P, Montenegro R, Shmulsky R, Wiedenhoeft AC. 2021. Field-deployable computer vision wood identification of Peruvian timbers. *Front Plant Sci.* 12: 647515. DOI: 10.3389/fpls.2021.647515.

Ravindran P, Owens FC, Wade AC, Shmulsky R, Wiedenhoeft AC. 2022a. Towards sustainable North American wood product value chains, part I: Computer vision identification of diffuse porous hardwoods. *Front Plant Sci.* 12: 758455. DOI: 10.3389/fpls.2021.758455.

Ravindran P, Wade AC, Owens FC, Shmulsky R, Wiedenhoeft AC. 2022b. Towards sustainable North American wood product value chains, part 2: Computer vision identification of ring-porous hardwoods. *Can. J. For. Res.* 52: 1014–1027. DOI: 10.1139/cjfr-2022-0077.

Ravindran P, Owens FC, Costa A, Rodrigues BP, Chavesta M, Shmulsky R, Wiedenhoeft AC. 2023. Evaluation of test specimen surface preparation on computer vision wood identification. *Wood Fiber Sci.* 55(2): 177–202. DOI: 10.22382/wfs-2023-15.

Wiedenhoeft AC. 2020. The XyloPhone: toward democratizing access to high-quality macroscopic imaging for wood and other substrates. *IAWA J.* 41(4): 699–719. DOI: 10.1163/22941932-bja10043.

Yang Z, Sinnott RO, Bailey B, Ke Q. 2023. A survey of automated data augmentation algorithms for deep learning-based image classification tasks. *Knowl. Inf. Syst.* 65(7): 2805–2861. DOI: 10.1007/s10115-023-01853-2.

Yoshida K, Kagawa A, Nishiguchi M. 2007. Extraction and detection of DNA from wood for species identification. In: *Proceedings of the international symposium on development of improved methods to identify Shorea species wood and its origin*. Forestry and Forest Products Research Institute, Ibaraki, Japan: 27–34.

Edited by Susan Anagnost