EVALUATION OF TEST SPECIMEN SURFACE PREPARATION ON MACROSCOPIC COMPUTER VISION WOOD IDENTIFICATION

Prabu Ravindran
Scientist III
Department of Botany
University of Wisconsin
Madison, WI
and
Center for Wood Anatomy Research
USDA Forest Service, Forest Products Laboratory
Madison, WI
E-mail: pravindran@wisc.edu

Frank C. Owens†*
Associate Professor
Department of Sustainable Bioproducts
Mississippi State University
Starkville, MS
E-mail: fco7@msstate.edu

Adriana Costa†
Assistant Professor
Department of Sustainable Bioproducts
Mississippi State University
Starkville, MS
E-mail: adc751@msstate.edu

Brunela Pollastrelli Rodrigues†
Assistant Professor
Department of Forestry and Environmental Conservation
Clemson University
Clemson, SC
E-mail: brunelr@clemson.edu

Manuel Chavesta
Profesor Principal
Department of Wood Industry
Universidad Nacional Agraria La Molina
Lima, Peru
E-mail: mchavesta@lamolina.edu.pe

Rolando Montenegro
Profesor Asociado
Department of Wood Industry
Universidad Nacional Agraria La Molina
Lima, Peru
E-mail: rmontenegro@lamolina.edu.pe

* Corresponding author
† SWST member

Wood and Fiber Science, 55(2), 2023, pp. 176-202
https://doi.org/10.22382/wfs-2023-15

Rubin Shmulsky†
Professor and Head
Department of Sustainable Bioproducts
Mississippi State University
Starkville, MS
E-mail: rs26@msstate.edu

Alex C. Wiedenhoeft
Research Botanist and Team Leader
Center for Wood Anatomy Research
USDA Forest Service, Forest Products Laboratory
Madison, WI
and
Department of Botany
University of Wisconsin
Madison, WI
and
Department of Sustainable Bioproducts
Mississippi State University
Starkville, MS
and
Department of Forestry and Natural Resources
Purdue University
West Lafayette, IN
and
Departamento de Ciências Biológicas (Botânica)
Universidade Estadual Paulista—Botucatu
São Paulo, Brasil
E-mail: acwieden@wisc.edu

(Received June 2023)

Abstract. Previous studies on computer vision wood identification (CVWID) have assumed or implied that the quality of sanding or knifing preparation of the transverse surface of wood specimens could influence model performance, but its impact is unknown and largely unexplored. This study investigates how variations in surface preparation quality of test specimens could affect the predictive accuracy of a previously published 24-class XyloTron CVWID model for Peruvian timbers. The model was trained on images of Peruvian wood specimens prepared at 1500 sanding grit and tested on images of independent specimens (not used in training) prepared across a series of progressively coarser sanding grits (1500, 800, 600, 400, 240, 180, and 80) and high-quality knife cuts. The results show that while there was a drop in performance at the lowest sanding grit of 80, most of the higher grits and knife cuts did not exhibit statistically significant differences in predictive accuracy. These results lay the groundwork for a future larger-scale investigation into how the quality of surface preparation in *both* training *and* testing data will impact CVWID model accuracy.

Keywords: XyloTron, computer vision wood identification, machine learning, deep learning, surface preparation.

INTRODUCTION

The implementation and enforcement of sustainable, legal, and monetizable forest products value chains require access to wood identification expertise (Gasson 2011; Johnson and Laestadius 2011; Lowe et al 2016; UNODC 2016). This expertise, using traditional laboratory methods, is currently restricted to a handful of wood identification experts and centers and is difficult to scale up to meet global needs (Wiedenhoeft et al 2019). Multiple technologies for automated wood identification have been researched to help address this dearth of wood identification experts (Schmitz et al 2020). Powered by advances in machine

learning, democratized access to macroscopic imaging tools such as the XyloTron and XyloPhone (Ravindran et al 2020; Wiedenhoeft 2020), and cloud-based processing platforms, computer vision wood identification (CVWID, Khalid et al 2008; Hwang and Sugiyama 2021) has been demonstrated to be an effective and affordable technology for automated wood identification at country and continental scales for field screening purposes (Ravindran et al 2018, 2019, 2020, 2022a, 2022b; de Geus et al 2020; Arévalo et al 2021).

In traditional macroscopic wood identification (Panshin and de Zeeuw 1980; Hoadley 1990; Wheeler and Baas 1998; Ruffinatto et al 2015), experts with extensive training view anatomical features on the transverse, and often radial and/or tangential surface(s), and make taxonomical determinations based on the size, frequency, combinations, and/or patterns of features they observe (Miller et al 2002; Wiedenhoeft 2011; Arévalo et al 2020; Arévalo and Wiedenhoeft 2022). In CVWID, images with relevant wood anatomical features are processed by a model to make an identification (Khalid et al 2008; Esteban et al 2009; Filho et al 2014; Kwon et al 2017, 2019; Rosa da Silva et al 2017; Barmpoutis et al 2018; Figueroa-Mata et al 2018; Tang et al 2018; Damayanti et al 2019; de Andrade et al 2020; He et al 2020; Lens et al 2020; Souza et al 2020; Fabijańska et al 2021; Wu et al 2021). While a taxonomic determination can be based on microscopic or macroscopic features in the wood specimen, the focus of this study is macroscopic wood identification, and CVWID will refer to image-based macroscopic wood identification.

In both methods, since it is presumed that disparate taxa can be identified based on visual differences in anatomy, surface preparation is commonly employed to make anatomical features visible (or more visible). For traditional macroscopic wood identification, a specimen's transverse surface is typically cut cleanly with a hand-held utility knife (Hoadley 1990; Wiedenhoeft 2011), whereas both knife cuts and sanding have been used for specimen surface preparation in prior CVWID works (Cerre 2016; Barbosa et al 2021; Hwang and Sugiyama 2021).

While previous CVWID studies have assumed or implied that the quality of surface preparation could influence model performance, its impact is unknown and largely unexplored (Wang et al 2013; Tang et al 2018; Damayanti et al 2019; de Andrade et al 2020). Hwang and Sugiyama (2021) suggest that effective models can be developed only by using images that show species-specific anatomical features and posit that surface preparation with a knife or sandpaper is necessary to clearly reveal those characteristics. Ravindran and Wiedenhoeft (2022) also argue that surface preparation must be sufficient to reveal anatomical characteristics relevant for identification. To the authors' knowledge, no study to date has attempted to evaluate the impact of surface preparation on CVWID model performance, especially when there is a difference in sample preparation between images of the training and testing specimens.

Ravindran et al (2021) presented the first national scale, anatomically informed 24-class CVWID model for Peruvian woods. They evaluated model predictive accuracies using a novel surrogate field testing methodology whereby mutually exclusive training and testing specimens were procured from separate xylaria—a practice that has been shown to be critical for evaluating model generalizability (Ravindran and Wiedenhoeft 2022). The specimens in both the training dataset and the testing dataset were polished to a sanding grit of 1500. As this Peruvian CVWID field model is ready for deployment, there is an urgent need to determine the approximate minimum surface quality needed for test specimens to maintain high predictive accuracy in real-life implementation.

This study evaluates the predictive accuracy of Ravindran et al's (2021) 24-class Peruvian CVWID model, trained on images of specimens prepared at 1500 sanding grit, with new testing images of the prior test specimens prepared across a series of progressively coarser sanding grits (1500, 800, 600, 400, 240, 180, and 80) and high-quality knife cuts to determine if the reduction in surface quality leads to any reduction in model

performance. Results of this initial study are expected to inform future investigations into the impact of surface preparation quality (of training *and* testing data) on CVWID models. It should be noted that the Peruvian CVWID model is evaluated as previously published, with no subsequent training or fine-tuning of the model weight parameters.

MATERIALS AND METHODS

Datasets

Image datasets for model testing were captured using 167 specimens from the David A. Kribs Wood Collection (PACw) and teaching collections at Mississippi State University. One hundred fifteen specimens used in field model testing in Ravindran et al (2021) were included and supplemented with 52 correctly identified and validated specimens from the teaching wood collection. Four specimens used for model testing in Ravindran et al (2021) were not included in this study due to size and geometry constraints regarding the effects of repeated sanding on specimen longevity.

The transverse surfaces of the 167 specimens were sanded by hand on a benchtop disc sander at descending grits in the following order: 1500, 800, 600, 400, 240, 180, and 80. The sander was equipped with an adjustable worktable to ensure that the transverse surface remained perpendicular to the longitudinal axis of each specimen during preparation. After sanding, each sample was inspected for burn marks and atypical sanding artifacts. If any were found, they were gently resanded at the same grit to remove them. Up to five images of the prepared surfaces were collected using the Xylo-Tron from each specimen after every grit setting.

A subset (75/167) of the sanded specimens were cut cleanly with a hand-held utility knife (Wiedenhoeft 2011) on their transverse surface. After cutting, each sample was inspected with a hand lens to ensure that the surface condition left by the 80-grit preparation had been removed. The specimens were imaged to yield up to five images per specimen.

For the knife-cut dataset, it was not possible to prepare specimens in PACw from two of the classes (Poulsenia and Schizolobium) due to reduced specimen size from repeated sanding. Ten specimens of each of these missing classes were prepared and imaged from the reference collection housed at the Universidad Nacional Agraria La Molina (UNALM) in Peru. This resulted in seven multigrit, 24-class image datasets using 167 specimens and one knife-cut image dataset from 95 specimens.

Dataset details are summarized in Table 1. As in Ravindran et al (2021), when referring to a

Table 1. Specimen counts for the test image datasets.

Class label	Sanded (specimen counts)	Knife-cut[a] (specimen counts)
Amburana	2	1
Aniba	2	1
Aspidosperma	5	2
BrosimumA	9	4
BrosimumU	2	2
Calycophyllum	7	2
Cariniana	8	4
Cedrela	20	7
Cedrelinga	6	6
Chorisia	5	4
Copaifera	3	3
Dipteryx	5	4
Eucalyptus	28	11
Guazuma	5	1
Hura	4	3
Maquira	2	1
Myroxylon	6	3
Ormosia	6	1
Pinus	2	1
Poulsenia	1	–
Pouteria	4	3
Schizolobium	3	–
Swietenia	15	8
Virola	17	3
Total	167	75

[a]For the knife-cut dataset, images from the Poulsenia and Schizolobium classes were not possible to source from the PACw and Mississippi State University teaching collections due to reduced size from repeated sanding. Ten additional images for each of those missing classes were sourced from the UNALM collection; however, those 20 images were not used in the main comparison with the data from the sanding grits due to statistical constraints, though, for completeness, they are mentioned in Appendix A.

CVWID class we use the class name without italics, but when referring to the same woods as botanical entities, we use italicization (eg the class Schizolobium vs the genus *Schizolobium*). Exemplar images from one specimen of *Cariniana* sp. were compiled to show the differences in visual surface quality at each grit (Fig 1).

Model Evaluation at Each Surface Quality

The XyloTron field model presented in Ravindran et al (2021) was trained with transfer learning (Pan and Yang 2010) using transverse surface images from 1300 specimens (housed in the USDA Forest Products Laboratory's MADw and SJRw collections) and prepared at exactly one

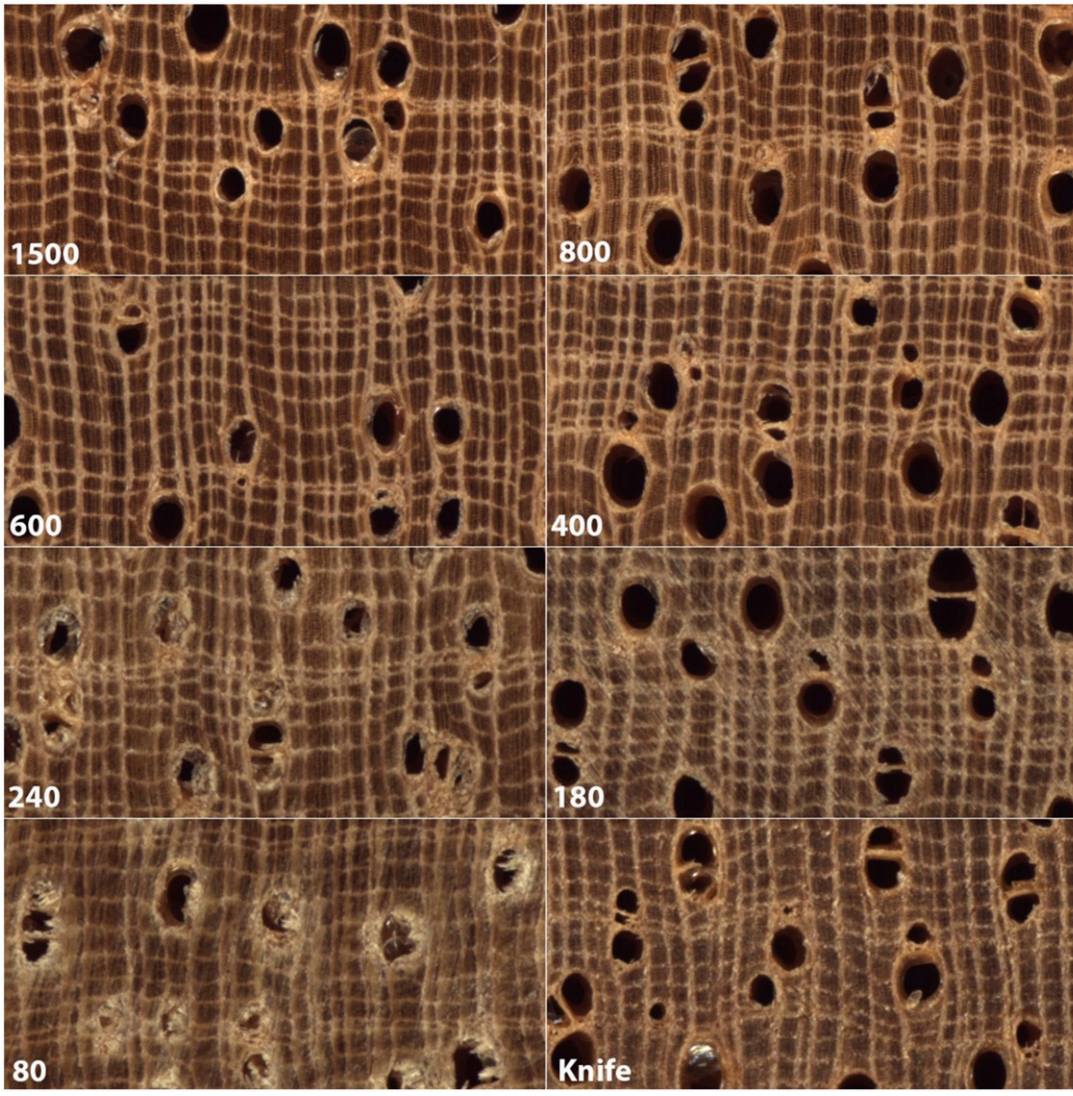

Figure 1. Exemplar images from the same specimen of *Cariniana* sp. showing the gradual decline in visual quality by descending grit number. With coarser grits (smaller number), a gradual decline in feature visibility/clarity and gradual increase in sandpaper artifacts can be observed. Each image shows 3.174 mm × 1.587 mm of tissue.

sanding grit: 1500. The published weights for the 24-class convolutional neural network (CNN, LeCun et al 1989; Goodfellow et al 2016), using the ResNet50 (He et al 2015, 2016) backbone, were used **as-is** with no further model training or finetuning to generate specimen-level class predictions for the eight image datasets. The model prediction pipeline was implemented using PyTorch (Paszke et al 2019) and Scikit-learn (Pedregosa et al 2011).

For each sample preparation condition, the top-1 and top-2 specimen-level prediction accuracies were generated using image predictions from up to five images contributed by the specimen. The majority vote, with equal weighting, of the top-k image-level predictions was taken as the corresponding top-k specimen-level prediction for the specimen contributing the images. A specimen is correctly classified if its true label is in the k predicted labels.

Statistical Analyses

Differences in field model predictive accuracy values (percent of specimens correctly identified) among the eight sanding/knife preparation groups were evaluated for statistical significance by Cochran's Q using IBM SPSS Statistics 28. Multiple comparisons were made by the method proposed in Dunn (1964). p-Values were adjusted using a Bonferroni correction to ensure an experiment-wise error rate of 0.05. p-Values are summarized in the main text. All test statistics and p-values for individual multiple comparisons are provided in Appendix A. Additionally, since the statistical tests require that all data come from the same specimens (N=75), the data from the 20 UNALM specimens were not used for comparison with the sanding data, but, for the sake of completeness, have been included as a separate analysis (where N=95) in Appendix A.

RESULTS AND DISCUSSION

The exemplar images of *Cariniana* sp. show a gradual decline in surface quality and visibility of anatomical features as the grit number decreases and the coarseness of the surface increases (Fig 1). While the decline is barely perceptible in the higher grits, a noticeable decrease in feature clarity and an increase in sanding artifacts can be observed at around 240 grit. The knife-cut image is perhaps most comparable to the sanding quality at 400 grit.

For the specimens listed by class in Table 1, confusion matrices for the top-1 specimen-level predictions of the ResNet50-based model are provided to indicate visually where and how frequently inaccurate predictions occurred (Fig 2). The more inaccurate the performance, the more off-diagonal cells appear darkened. A quick visual comparison of the matrices reveals that the 80-grit matrix in the lower left exhibits the most off-diagonal cells. The individual confusion matrices in Fig 2, along with cell values and annotations, are presented in Appendix A.

As the sample size for the knife-cut group (N = 75) was smaller than those of the sanding grit groups (N = 167), the results are presented in two ways. When comparing the predictive accuracies (percentages of correctly identified specimens) by sanding grit group only, where N = 167, there was no statistically significant difference ($p > 0.05$) among the percentages of correctly identified specimens for 1500 (89.2%), 800 (95.2%), 600 (94.6%), 400 (93.4%), and 240 (88.6%) (Fig 3). Grit 180 (84.4%) was significantly different ($p < 0.05$) from 800 (95.2%), 600 (94.6%), and 80 (62.9%) only. Grit 80 (62.9%) showed a statistically significant difference from all other grits.

When comparing the predictive accuracies by sanding grit and knife-cut groups, where N = 75, there was no statistically significant difference ($p > 0.05$) among the percentages of correctly identified specimens for grits 1500 (93.3%), 800 (94.7%), 600 (96.0%), 400 (92.0%), 240 (92.0%), 180 (85.3%), and knife cuts (90.7%) (Fig 4). Grit 80 (69.3%) showed a statistically significant difference between all other grit and knife-cut groups ($p < 0.05$).

The results of this investigation suggest that, for these datasets, the CVWID model's predictive accuracy is robust to differences in surface

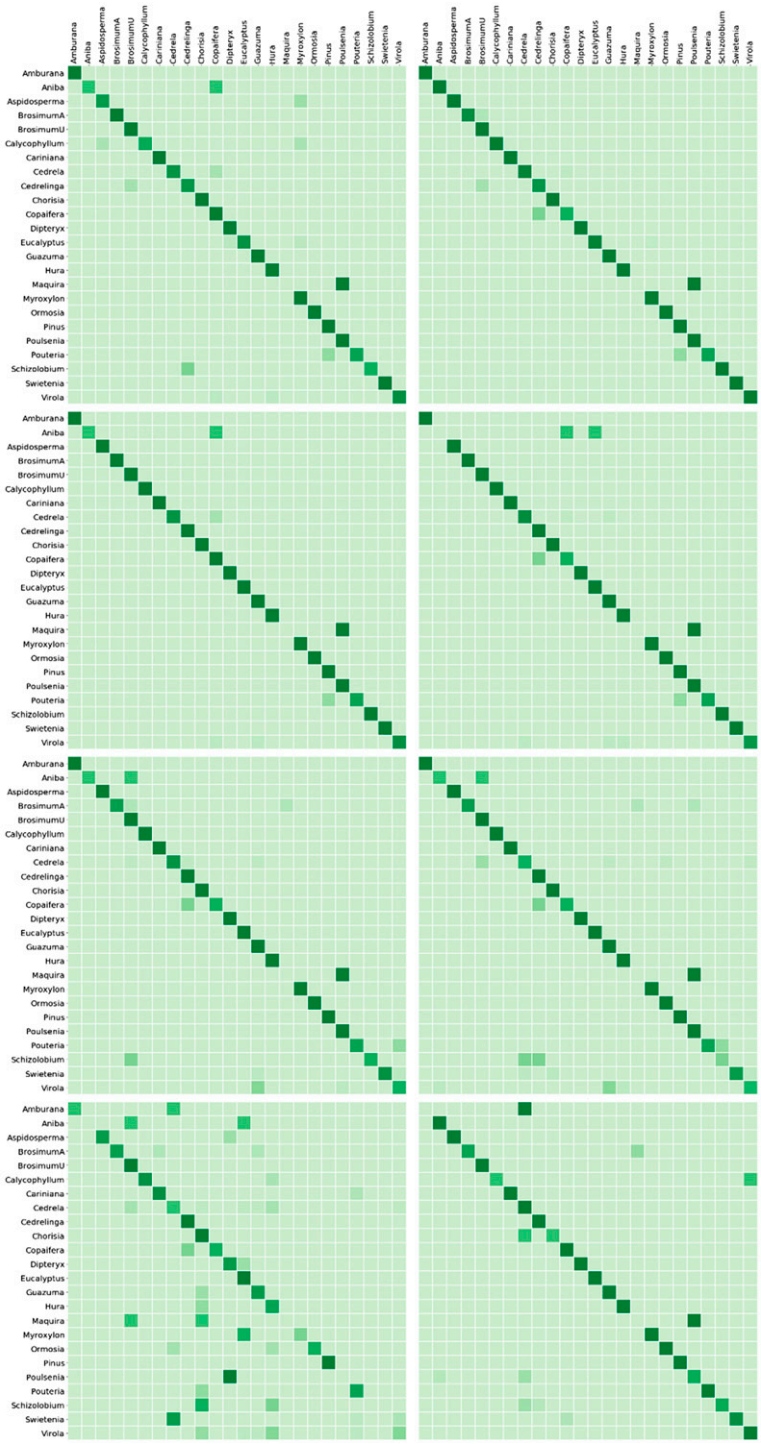

Figure 2. Confusion matrices for top-1 specimen-level predictions of the ResNet50-based model. The sample preparation settings, in raster order, are sanding at grits 1500, 800, 600, 400, 240, 180, 80, and knife cuts. The individual confusion matrices in this figure, along with annotations, are presented in Appendix A.

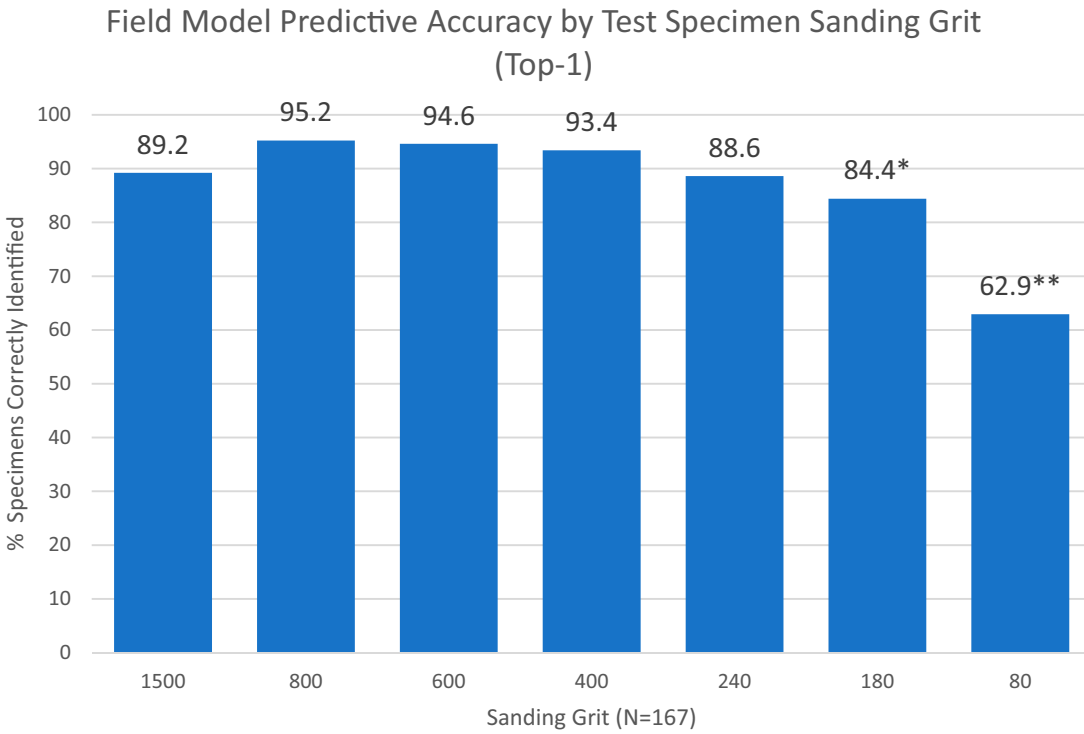

Figure 3. Comparison of Peru field model predictive accuracy (top-1) by sanding grit only ($N = 167$). The percentages of specimens correctly identified among the sanding grits were compared using Cochran's Q with post hoc multiple comparisons per Dunn (1964) with an experiment-wise error of 0.05. Multiple comparisons showed that there were no significant differences ($p > 0.05$) among the groups with the following exceptions. * There was a statistically significant difference ($p < 0.05$) between 180 grit and 800, 600, and 80, respectively. ** There was a statistically significant difference ($p < 0.05$) between 80 grit and all other grits, respectively.

preparation among the knife cut and sanding grits from 240 upward. At 80 grit, a sharp reduction in predictive accuracy can be observed, which might be due to the decline in feature visibility/clarity (pore boundaries, rays, etc.) and an increase in sandpaper artifacts of the kind exemplified by Fig 1. This suggests that there is a point at which the obscuring of anatomical detail and/or the increase in surface preparation artifacts results in reduced model performance, though this point could vary by species due to differences in specific gravity, hardness, and/or other properties. As variation in surface quality is a practical concern for field deployment, the impact of obscured anatomical features and sanding/knifing artifacts merits further investigation to ensure that predictive accuracy is not unnecessarily reduced by insufficient specimen preparation.

The deployment of CVWID systems can be impacted by covariate shifts (between training and deployment data), which can occur due to wood anatomy variations (eg natural forest grown vs plantation grown, changes due to environmental stresses) or operating condition variations (eg sample preparation methods, imaging parameter settings, operator skill). The focus of this study was to evaluate the predictive accuracy of a CVWID model with respect to one potential source of operator-induced covariate shift—explicit sample preparation differences between training and testing stages. Our results show that a model trained on a 1500-grit dataset has predictive accuracies that are acceptable for field screening, on images from surfaces sanded at grits coarser than 1500. Even though the predictive accuracy drops sharply for the 80-grit dataset, a

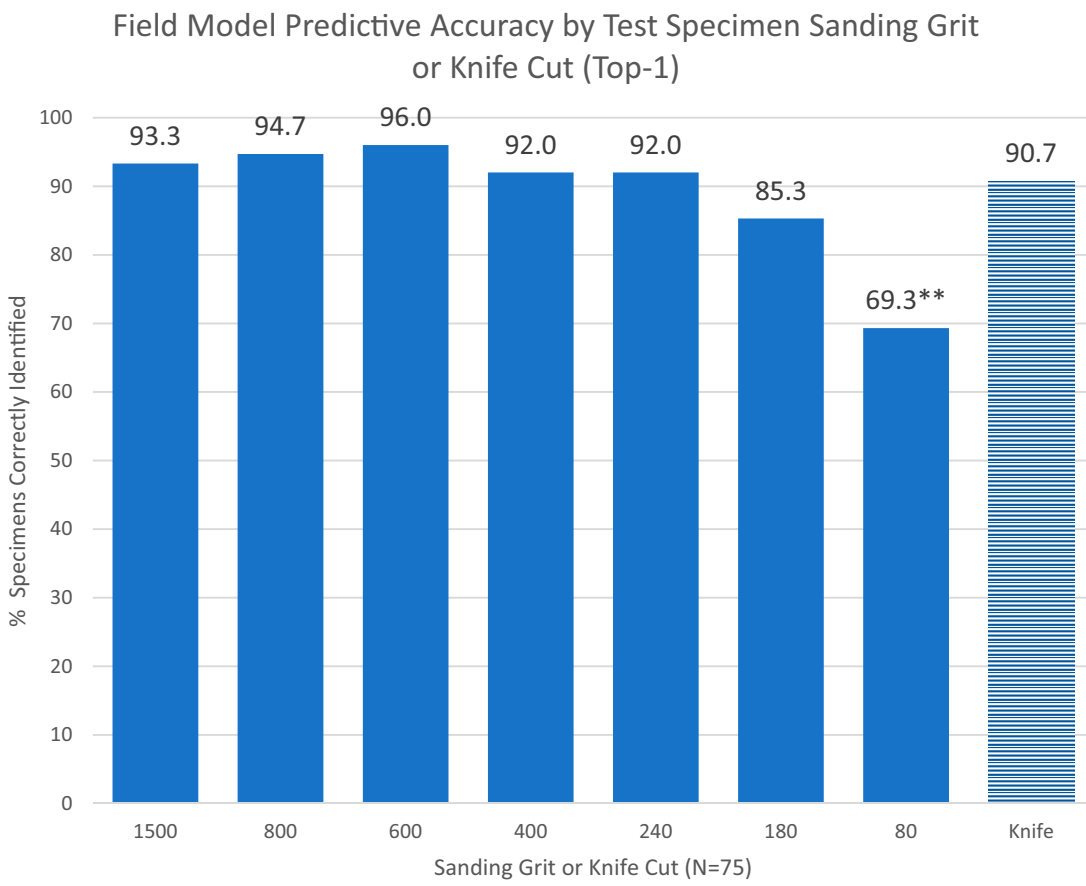

Figure 4. Comparison of Peru field model predictive accuracy (top-1) by sanding grit (solid bars) and knife-cut (patterned bar) groups ($N = 75$). The percentages of specimens correctly identified among all the groups were compared using Cochran's Q with post hoc multiple comparisons per Dunn (1964) with an experiment-wise error of 0.05. Multiple comparisons showed that there were no significant differences ($p > 0.05$) among the groups with the following exceptions. ** There was a statistically significant difference ($p < 0.05$) between 80 grit and all other grit and knife cut groups, respectively.

surface coarseness where the complete range of anatomical features may not be reliably discerned, the obtained accuracy of 62.9% is still considerably better than random guessing (ie for a model with 24 classes, an accuracy of ~4% would be expected). This suggests that the model generalizes acceptably in the presence of sanding grit differences between training and deployment data when the surrogate field-testing approach is employed for datasets from specimens from multiple xylaria.

It has been shown that there are biases inherent in surrogate field testing (Ravindran and Wiedenhoeft 2022). It should be noted that the question of optimal sample preparation during the training and testing phases is still unknown and unexplored. The results of this study suggest that for field implementation of a model trained on high-grit images with a human-in-the-loop modality (eg in the *xyloinf* software for the XyloTron platform, as distributed in Ravindran et al 2020), the predictions are highly reliable at grits as low as 240. Field polishing of a specimen from 80 to 240 grit with a cordless drill and an appropriate sanding pad requires only about 20 s (Wiedenhoeft, personal observation). Field polishing requires significantly less skill and training than making high-quality utility knife cuts but has a higher hardware burden (cordless drill, batteries, chargers,

access to power to recharge, the need to carry abrasives, etc.), and so may well not be practical in some field contexts.

A comprehensive investigation of the effect surface preparation has on the predictive accuracy of CVWID models would necessitate examining the effects each surface quality has on *both* training *and* testing. As specimen preparation for the type of imaging performed in this study is both labor and time intensive, the authors elected to examine first how varying the surface quality of the test specimens might affect predictive accuracy when the model was trained on images prepared at the highest quality (1500 grit) as that directly impacts the timely release of the Peruvian CVWID field model. Results showing both statistically significant and practical differences in performance among various surface preparations in testing justify a much greater investment of time and resources introducing a second variable of training specimen preparation quality. As the training specimens from the MADw and SJRw collections outnumber the testing specimens in the PACw collection by approximately eight times, repeating this study with two variables would require many more person-hours.

Prior works in CVWID have used sanding or knife cuts for sample preparation before imaging the prepared surfaces. Both these methods have their pros and cons. The knife-cut method is "quick" and field deployment friendly but making consistent and safe knife cuts to produce a clean surface on which the anatomical features are readily observable is a skill that is practice-intensive. In addition, manipulating the knife, hand, and wrist to cleanly cut the ends of flush and tightly packed bundled lumber *in situ* without first extracting a small subsample may prove difficult or impossible. On the other hand, sanding produces consistent surfaces and is well-suited for collecting large image datasets, thus a method of choice heretofore. One way to leverage the positives of the two approaches would be to employ sanding for training image datasets and using good, clean, and skillful utility knife cuts in the field, but, as noted above, this requires operator training and skill and may not be logistically feasible in some field contexts. Another advantage of constructing training datasets from finely sanded surfaces is that they can be used for the creation of digital archives of reference images which can be used for educational purposes. It is also easier to apply software-based image degradation approaches for data augmentation during the training procedure to improve model generalizability if one begins with high-quality images showing anatomical detail—there is no known method to enhance a low-quality dataset to include anatomical features not already observable. While our results suggest that the approach of using data from sanded surfaces for training and using utility knife cuts during deployment may be feasible for the creation of operational, field-deployable CVWID systems, the question of the optimal sanding grit for the training dataset is still open.

In principle, a path to robust models would be to train models by pooling datasets from different studies thus capturing a wide range of botanically and operationally induced covariate shifts. Until efforts for sharing and pooling multiple CVWID datasets gain traction, and workers publish the spatial scale of the images they collect (which is often not done), the method of using finely sanded specimens for training models and employing a quicker method of sample preparation (clean knife cuts or coarser grit sanding) in the field may be a reasonable pathway to operational, field-deployable CVWID systems.

CONCLUSIONS

This evaluation of a previously published 24-class CVWID model, trained on images of Peruvian specimens prepared at 1500 sanding grit, with new testing images of the prior test specimens prepared across a series of progressively coarser sanding grits (1500, 800, 600, 400, 240, 180, and 80) showed a robustness of predictive accuracy at grits 240 and higher and a reduction in performance at 180 grit and 80 grit. Moreover, it indicated that the predictive accuracy on the specimens prepared with utility knife cuts was not statistically different from that of the images

prepared at 180 grit and above. These results suggest that there is a point at which the obscuring of anatomical detail and/or the increase in surface preparation artifacts on test specimens result in reduced model performance, but minor differences at higher levels of surface preparation quality have little to no practical effect. As variation in surface preparation quality is a pragmatic concern for field deployment, the impact it has on predictive accuracy merits further investigation. These results lay the groundwork for a future larger-scale investigation into how the quality of surface preparation in *both* training *and* testing data will impact CVWID model accuracy. Models will be trained at progressively coarser grits and tested, for each grit, against the images of each testing grit used in this study.

ACKNOWLEDGMENT

The authors wish to gratefully acknowledge the specimen preparation and imaging efforts of Nicholas Bargren, Karl Kleinschmidt, Caitlin Gilly, Richard Soares, Adriana Costa, Fernanda Sofia Silva Gerbi, Jessica Francesca Figueroa Mendoza, Flavio Ruffinatto, Adam Wade, Chris Luckett, and Brunela Rodrigues.

Code Availability: 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.

Data Availability Statement: A minimal data set can be obtained by contacting the corresponding author, but the full data set used in the study is protected for up to 5 yr by a CRADA among FPL, UW-Madison, and FSC.

Author Contributions: F.C.O. and R.S. provided access to and supervised data acquisition from the PACw test specimens. B.P.R. prepared and imaged the PACw specimens. M.C. and R.M. prepared and imaged the UNALM specimens. P.R. implemented the machine learning pipelines for the study. P.R. and A.C.W. conducted data analysis and synthesis. F.C.O. provided statistical analysis. P.R., A.C.W., A.C., and F.C.O. wrote the paper. All authors contributed actionable feedback that improved the presentation of the paper.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: This work was supported in part by a grant from the U.S. Department of State via Interagency Agreement number 19318814Y0010 to A.C.W. and in part by research funding from the Forest Stewardship Council to A.C.W. P.R. 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 BR and MSU graduate students. Any opinions, findings, conclusion, 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.

REFERENCES

Arévalo R, Ebanyenle E, Ebeheakey AA, Bonsu KA, Lambog O, Soares R, Wiedenhoeft AC (2020) Field identification manual for Ghanaian timbers. General technical report FPL-GTR-190. U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, Madison, WI. 130 pp.

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. Colomb For 24(1):5-16.

Arévalo R, Wiedenhoeft AC (2022) Identification of Central American, Mexican, and Caribbean woods. General technical report FPL-GTR-190. U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, Madison, WI. 376 pp.

Barbosa ACF, Gerolamo CS, Lima AC, Angyalossy V, Pace MR (2021) Polishing entire stems and roots using sandpaper under water: An alternative method for

macroscopic analyses. Appl Plant Sci 9(5):10.1002/aps3.11421.

Barmpoutis P, Dimitropoulos K, Barboutis I, Nikos G, Lefakis P (2018) Wood species recognition through multidimensional texture analysis. Comput Electron Agric 144:241-248.

Cerre JC (2016) Incident light microphotography at high depth of focus. IAWA J 37(3):506-510.

Damayanti R, Prakasa E, Krisdianto Dewi LM, Wardoyo R, Sugiarto B, Pardede HF, Riyanto Y, Astutiputri VF, Panjaitan GR, Hadiwidjaja ML, Maulana YH, Mutaqin IN (2019) LignoIndo: Image database of Indonesian commercial timber. IOP Conf Ser: Earth Environ 374: 012057.

de Andrade BG, Basso VM, Latorraca JVF (2020) Machine vision for field-level wood identification. IAWA J 41(4):681-698.

de Geus A, da Silva SF, Gontijo AB, Silva FO, Batista MA, Souza JR (2020) An analysis of timber sections and deep learning for wood species classification. Multimedia Tools Appl 79:34513-34529.

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

Esteban LG, Fernández FG, de Palacios P, Romero RM, Cano NN (2009) Artificial neural networks in wood identification: The case of two *Juniperus* species from the Canary Islands. IAWA J 30(1):87-94.

Fabijańska A, Danek M, Barniak J (2021) Wood species automatic identification from wood core images with a residual convolutional neural network. Comput Electron Agric 181:105941.

Figueroa-Mata G, Mata-Montero E, Valverde-Otárola JC, Arias-Aguilar D (2018) Automated image-based identification of forest species: Challenges and opportunities for 21st century Xylotheques. *in* Proceedings, 2018 IEEE International Work Conference on Bioinspired Intelligence, July 18-20, 2018, San Carlos, Costa Rica, 1–8.

Filho PLP, Oliveira LS, Nisgoski S, Britto AS Jr. (2014) Forest species recognition using macroscopic images. Mach Vis Appl 25:1019-1031.

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(2):137-154.

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge, MA. 800 pp.

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *in* Proceedings, IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile, 1-11.

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *in* Proceedings, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, 1-12.

He T, Lu Y, Jiao L, Zhang Y, Jiang X, Yin Y (2020) Developing deep learning models to automate rosewood tree species identification for CITES designation and implementation. Holzforschung 74(12):1123-1133.

Hoadley RB (1990) Identifying wood: Accurate results with simple tools. Taunton Press, Newtown, CT. 223 pp.

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, 1-21.

Johnson A, Laestadius L (2011) New laws, new needs: The role of wood science in global policy efforts to reduce illegal logging and associated trade. IAWA J 32(2): 125-136.

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(3):9-19.

Kwon O, Lee H, Yang S-Y, Kim H, Park S-Y, Choi I-G, Yeo H (2019) Performance enhancement of automatic wood classification of Korean softwood by ensembles of convolutional neural networks. J Korean Wood Sci Technol 47:265-276.

Kwon O, Lee HG, Lee M-R, Jang S, Yang S-Y, Park S-Y, Choi I-G, Yeo H (2017) Automatic wood species identification of Korean softwood based on convolutional neural networks. J Korean Wood Sci Technol 45: 797-808.

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.

Lens F, Liang C, Guo Y, Tang X, Jahanbanifard M, da Silva FSC, Ceccantini G, Verbeek FJ (2020) Computer-assisted timber identification based on features extracted from microscopic wood sections. IAWA J 41(4):660-680.

Lowe AJ, Dormontt EE, Bowie MJ, Degen B, Gardner S, Thomas D, Clarke C, Rimbawanto A, Wiedenhoeft A, Yin Y, Sasaki N (2016) Opportunities for improved transparency in the timber trade through scientific verification. Bioscience 66(11):990-998.

Miller R, Wiedenhoeft AC, Ribeyron M-J (2002) CITES identification guide—Tropical woods. Environment Canada, Canada.

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345-1359.

Panshin AJ, de Zeeuw C (1980) Textbook of wood technology: Structure, identification, properties, and uses of the commercial woods of the United States and Canada. McGraw-Hill, New York. 736 pp.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. Pages 8026-8037 *in* Proceedings, 2019 NeurIPS, December 3, 2019, Vancouver, Canada. https://arxiv.org/abs/1912.01703

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12(85): 2825-2830.

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, 1-10.

Ravindran P, Ebanyenle E, Ebeheakey AA, Abban KB, Lambog O, Soares R, Costa A, Wiedenhoeft AC (2019) Image based identification of Ghanaian timbers using the XyloTron: Opportunities, risks, and challenges. *in* Proceedings, 33rd NeurIPS, December 2019, Vancouver, Canada, 1-10. https://arxiv.org/abs/1912.00296

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.

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.

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, 1-8.

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.

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.

Rosa da Silva N, De Ridder M, Baetens JM, den Bulcke JVB (2017) Automated classification of wood transverse cross-section micro-imagery from 77 commercial Central-African timber species. Ann Sci 74:30, 1-14.

Ruffinatto F, Crivellaro A, Wiedenhoeft AC (2015) Review of macroscopic features for hardwood and softwood identification and a proposal for a new character list. IAWA J 36(2):208-241.

Schmitz N, Beeckman H, Blanc-Jolivet C, Boeschoten L, Braga JWB, Cabezas JA, Chaix G, Crameri S, Deklerck V, Degen B, Dormontt E, Espinoza E, Gasson P, Haag V, Helmling S, Horacek M, Koch G, Lancaster C, Lens F, Lowe A, Martínez-Jarquín S, Nowakowska JA, Olbrich A, Paredes-Villanueva K, Pastore TCM, Ramananantoandro T, Razafimahatratra AR, Ravindran P, Rees G, Soares LF, Tysklind N, Vlam M, Watkinson C, Wheeler E, Winkler R, Widenhoeft AC, Zemke VT, Zuidema P (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.

Souza DV, Santos JX, Vieira HC, Naide TL, Nisgoski S, Oliveira LES (2020) An automatic recognition system of Brazilian flora species based on textural features of macroscopic images of wood. Wood Sci Technol 54: 1065-1090.

Tang XJ, Tay YH, Siam NA, Lim SC (2018) MyWood-ID: Automated macroscopic wood identification system using smartphone and macro-lens. Pages 37–43 *in* Proceedings, 2018 International Conference on Computational Intelligence and Intelligent Systems (CIIS 2018). Association for Computing Machinery, New York, NY.

UNODC (2016) Best practice guide for forensic timber identification. United Nations Office on Drugs and Crime, New York, NY.

Wang HJ, Qi HN, Wang XF (2013) A new Gabor based approach for wood recognition. Neurocomputing 116: 192-200.

Wheeler EA, Baas P (1998) Wood identification—A review. IAWA J 19(3):241-264.

Wiedenhoeft AC (2011) Identification of Central American woods (Identificacion de las Especies Maderables de Centroamerica). Forest Products Society, Madison, WI. Publication #7215-11.

Wiedenhoeft AC (2020) The XyloPhone: Toward democratizing access to high-quality macroscopic imaging for wood and other substrates. IAWA J 41(4):699-719.

Wiedenhoeft AC, Simeone J, Smith A, Parker-Forney M, Soares R, Fishman A (2019) Fraud and misrepresentation in retail forest products exceeds U.S. forensic wood science capacity. PLoS One 14(7):e0219917.

Wu F, Gazo R, Haviarova E, Benes B (2021) Wood identification based on longitudinal section images by using deep learning. Wood Sci Technol 55:553-563.

APPENDIX A

Table A1. Cochran's Q test for Fig 3.

Total *N*	167
Test statistic	152.437
Degree of freedom	6
Asymptotic sig. (2-sided test)	0.000

Table A2. Pairwise comparisons for Fig 3.

Sample 1-Sample 2	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.[a]
Grit80-Grit180	0.216	0.032	6.815	<0.001	0.000
Grit80-Grit240	0.257	0.032	8.140	<0.001	0.000
Grit80-Grit1500	0.263	0.032	8.329	0.000	0.000
Grit80-Grit400	0.305	0.032	9.655	0.000	0.000
Grit80-Grit600	0.317	0.032	10.033	0.000	0.000
Grit80-Grit800	0.323	0.032	10.222	0.000	0.000
Grit180-Grit240	0.042	0.032	1.325	0.185	1.000
Grit180-Grit1500	0.048	0.032	1.514	0.130	1.000
Grit180-Grit400	0.090	0.032	2.840	0.005	0.095
Grit180-Grit600	0.102	0.032	3.218	0.001	0.027
Grit180-Grit800	0.108	0.032	3.407	<0.001	0.014
Grit240-Grit1500	0.006	0.032	0.189	0.850	1.000
Grit240-Grit400	0.048	0.032	1.514	0.130	1.000
Grit240-Grit600	0.060	0.032	1.893	0.058	1.000
Grit240-Grit800	0.066	0.032	2.082	0.037	0.784
Grit1500-Grit400	−0.042	0.032	−1.325	0.185	1.000
Grit1500-Grit600	−0.054	0.032	−1.704	0.088	1.000
Grit1500-Grit800	−0.060	0.032	−1.893	0.058	1.000
Grit400-Grit600	0.012	0.032	0.379	0.705	1.000
Grit400-Grit800	0.018	0.032	0.568	0.570	1.000
Grit600-Grit800	0.006	0.032	0.189	0.850	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

[a]Significance values have been adjusted by the Bonferroni correction for multiple tests.

Table A3. Cochran's Q test for Fig 4.

Total *N*	75
Test statistic	56.751
Degree of freedom	7
Asymptotic sig. (2-sided test)	<0.001

Table A4. Pairwise comparisons for Fig 4.

Sample 1-Sample 2	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.[a]
Grit80-Grit180	0.160	0.043	3.735	<0.001	0.005
Grit80-Knife	−0.213	0.043	−4.980	<0.001	0.000
Grit80-Grit400	0.227	0.043	5.292	<0.001	0.000
Grit80-Grit240	0.227	0.043	5.292	<0.001	0.000
Grit80-Grit1500	0.240	0.043	5.603	<0.001	0.000
Grit80-Grit800	0.253	0.043	5.914	<0.001	0.000
Grit80-Grit600	0.267	0.043	6.225	<0.001	0.000
Grit180-Knife	−0.053	0.043	−1.245	0.213	1.000
Grit180-Grit400	0.067	0.043	1.556	0.120	1.000
Grit180-Grit240	0.067	0.043	1.556	0.120	1.000
Grit180-Grit1500	0.080	0.043	1.868	0.062	1.000
Grit180-Grit800	0.093	0.043	2.179	0.029	0.822
Grit180-Grit600	0.107	0.043	2.490	0.013	0.358
Knife-Grit400	0.013	0.043	0.311	0.756	1.000
Knife-Grit240	0.013	0.043	0.311	0.756	1.000
Knife-Grit1500	0.027	0.043	0.623	0.534	1.000
Knife-Grit800	0.040	0.043	0.934	0.350	1.000
Knife-Grit600	0.053	0.043	1.245	0.213	1.000
Grit400-Grit1500	0.013	0.043	0.311	0.756	1.000
Grit240-Grit1500	0.013	0.043	0.311	0.756	1.000
Grit400-Grit800	0.027	0.043	0.623	0.534	1.000
Grit240-Grit800	0.027	0.043	0.623	0.534	1.000
Grit400-Grit600	0.040	0.043	0.934	0.350	1.000
Grit240-Grit600	0.040	0.043	0.934	0.350	1.000
Grit400-Grit240	0.000	0.043	0.000	1.000	1.000
Grit1500-Grit800	−0.013	0.043	−0.311	0.756	1.000
Grit1500-Grit600	−0.027	0.043	−0.623	0.534	1.000
Grit800-Grit600	−0.013	0.043	−0.311	0.756	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

[a]Significance values have been adjusted by the Bonferroni correction for multiple tests.

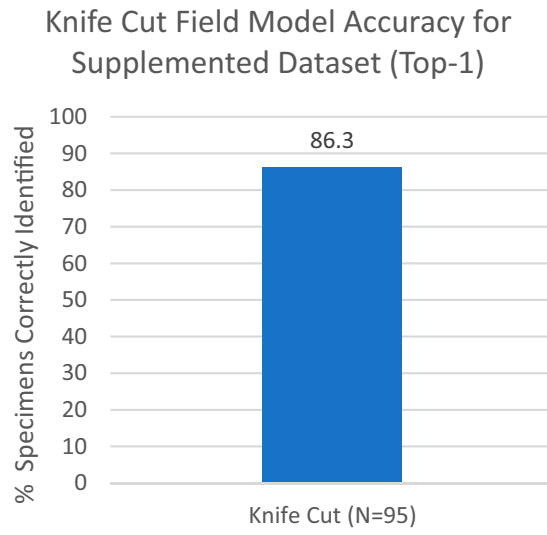

Figure A1. Predictive accuracy for knife-cut dataset including 10 additional specimens of Poulsenia and 10 additional specimens of Schizolobium, all from the UNALM collection ($N = 95$).

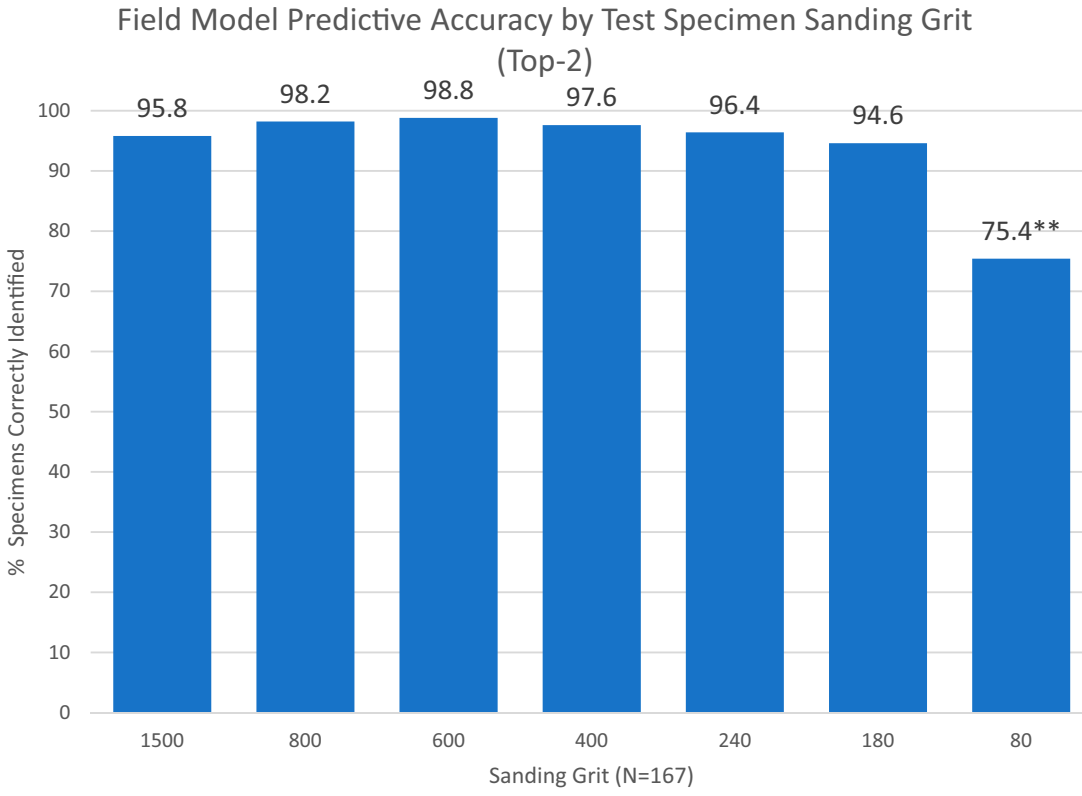

Figure A2. Comparison of Peru field model predictive accuracy (top-2) by sanding grit only ($N = 167$). The percentages of specimens correctly identified among all the groups were compared using Cochran's Q with post hoc multiple comparisons per Dunn (1964) with an experiment-wise error of 0.05. Multiple comparisons showed that there were no significant differences ($p > 0.05$) among the groups with the following exceptions. ** There was a statistically significant difference ($p < 0.05$) between 80 grit and all other grits, respectively.

Table A5. Cochran's Q test for Fig A2.

Total *N*	167
Test statistic	141.089
Degree of freedom	6
Asymptotic sig. (2-sided test)	0.000

Table A6. Pairwise comparisons for Fig A2.

Sample 1-Sample 2	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.[a]
T2_Grit80-T2_Grit180	0.192	0.024	7.976	<0.001	0.000
T2_Grit80-T2_Grit1500	0.204	0.024	8.475	0.000	0.000
T2_Grit80-T2_Grit240	0.210	0.024	8.724	0.000	0.000
T2_Grit80-T2_Grit400	0.222	0.024	9.223	0.000	0.000
T2_Grit80-T2_Grit800	0.228	0.024	9.472	0.000	0.000
T2_Grit80-T2_Grit600	0.234	0.024	9.721	0.000	0.000
T2_Grit180-T2_Grit1500	0.012	0.024	0.499	0.618	1.000
T2_Grit180-T2_Grit240	0.018	0.024	0.748	0.455	1.000
T2_Grit180-T2_Grit400	0.030	0.024	1.246	0.213	1.000
T2_Grit180-T2_Grit800	0.036	0.024	1.496	0.135	1.000
T2_Grit180-T2_Grit600	0.042	0.024	1.745	0.081	1.000
T2_Grit1500-T2_Grit240	−0.006	0.024	−0.249	0.803	1.000
T2_Grit1500-T2_Grit400	−0.018	0.024	−0.748	0.455	1.000
T2_Grit1500-T2_Grit800	−0.024	0.024	−0.997	0.319	1.000
T2_Grit1500-T2_Grit600	−0.030	0.024	−1.246	0.213	1.000
T2_Grit240-T2_Grit400	0.012	0.024	0.499	0.618	1.000
T2_Grit240-T2_Grit800	0.018	0.024	0.748	0.455	1.000
T2_Grit240-T2_Grit600	0.024	0.024	0.997	0.319	1.000
T2_Grit400-T2_Grit800	0.006	0.024	0.249	0.803	1.000
T2_Grit400-T2_Grit600	0.012	0.024	0.499	0.618	1.000
T2_Grit800-T2_Grit600	−0.006	0.024	−0.249	0.803	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

[a]Significance values have been adjusted by the Bonferroni correction for multiple tests.

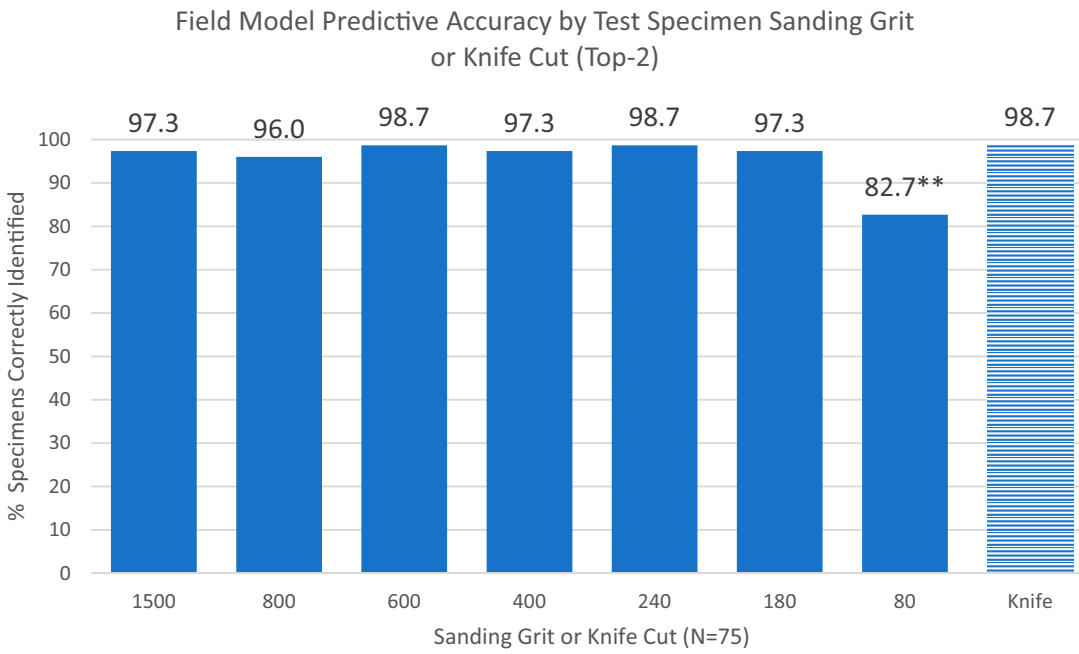

Figure A3. Comparison of Peru field model predictive accuracy (top-2) by sanding grit (solid bars) and knife-cut (patterned bar) groups ($N = 75$). The percentages of specimens correctly identified among all the groups were compared using Cochran's Q with post hoc multiple comparisons per Dunn (1964) with an experiment-wise error of 0.05. Multiple comparisons showed that there were no significant differences ($p > 0.05$) among the groups with the following exceptions. ** There was a statistically significant difference ($p < 0.05$) between 80 grit and all other grits, respectively.

Table A7. Cochran's Q test for Fig A3.

Total *N*	75
Test statistic	47.652
Degree of freedom	7
Asymptotic sig. (2-sided test)	<0.001

Table A8. Pairwise comparisons for Fig A3.

Sample 1-Sample 2	Test statistic	Std. error	Std. test statistic	Sig.	Adj. sig.[a]
T2_Grit80-T2_Grit800	0.133	0.029	4.554	<0.001	0.000
T2_Grit80-T2_Grit1500	0.147	0.029	5.010	<0.001	0.000
T2_Grit80-T2_Grit400	0.147	0.029	5.010	<0.001	0.000
T2_Grit80-T2_Grit180	0.147	0.029	5.010	<0.001	0.000
T2_Grit80-T2_Grit600	0.160	0.029	5.465	<0.001	0.000
T2_Grit80-T2_Grit240	0.160	0.029	5.465	<0.001	0.000
T2_Grit80-T2_ Knife	−0.160	0.029	−5.465	<0.001	0.000
T2_Grit800-T2_Grit1500	0.013	0.029	0.455	0.649	1.000
T2_Grit800-T2_Grit400	−0.013	0.029	−0.455	0.649	1.000
T2_Grit800-T2_Grit180	−0.013	0.029	−0.455	0.649	1.000
T2_Grit800-T2_Grit600	−0.027	0.029	−0.911	0.362	1.000
T2_Grit800-T2_Grit240	−0.027	0.029	−0.911	0.362	1.000
T2_Grit800-T2_Knife	−0.027	0.029	−0.911	0.362	1.000
T2_Grit1500-T2_Grit400	0.000	0.029	0.000	1.000	1.000
T2_Grit1500-T2_Grit180	0.000	0.029	0.000	1.000	1.000
T2_Grit1500-T2_Grit600	−0.013	0.029	−0.455	0.649	1.000
T2_Grit1500-T2_Grit240	−0.013	0.029	−0.455	0.649	1.000
T2_Grit1500-T2_Knife	−0.013	0.029	−0.455	0.649	1.000
T2_Grit400-T2_Grit600	0.013	0.029	0.455	0.649	1.000
T2_Grit180-T2_Grit600	0.013	0.029	0.455	0.649	1.000
T2_Grit400-T2_Grit180	0.000	0.029	0.000	1.000	1.000
T2_Grit400-T2_Grit240	−0.013	0.029	−0.455	0.649	1.000
T2_Grit400-T2_ Knife	−0.013	0.029	−0.455	0.649	1.000
T2_Grit180-T2_Grit240	0.013	0.029	0.455	0.649	1.000
T2_Grit180-T2_ Knife	−0.013	0.029	−0.455	0.649	1.000
T2_Grit600-T2_Grit240	0.000	0.029	0.000	1.000	1.000
T2_Grit600-T2_ Knife	0.000	0.029	0.000	1.000	1.000
T2_Grit240-T2_ Knife	0.000	0.029	0.000	1.000	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

[a]Significance values have been adjusted by the Bonferroni correction for multiple tests.

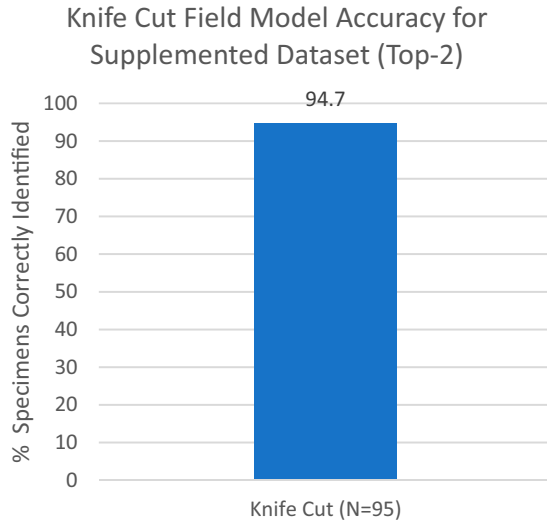

Figure A4. Predictive accuracy for knife-cut dataset including 10 additional specimens of Poulsenia and 10 additional specimens of Schizolobium, all from the UNALM collection ($N = 95$).

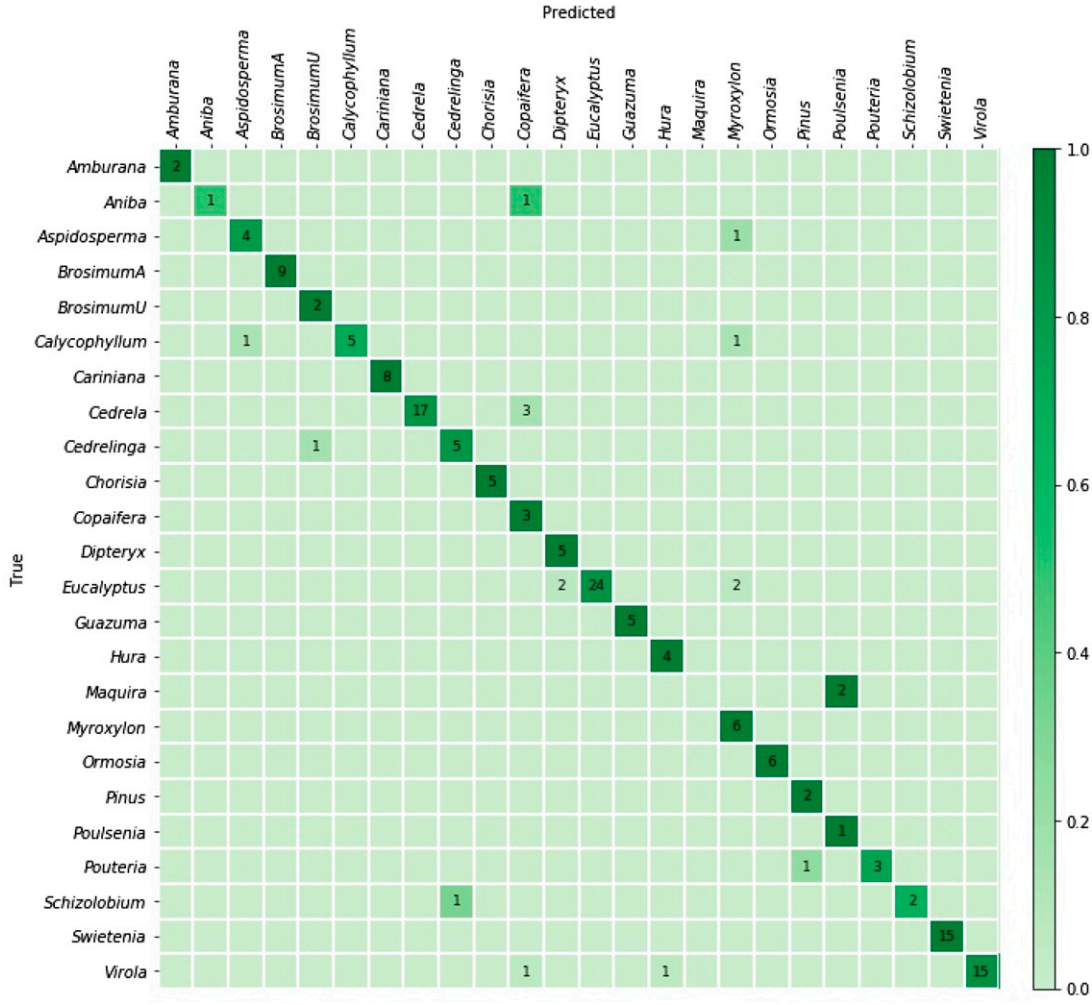

Figure A5. Confusion matrix for ResNet50, 1500 sanding grit.

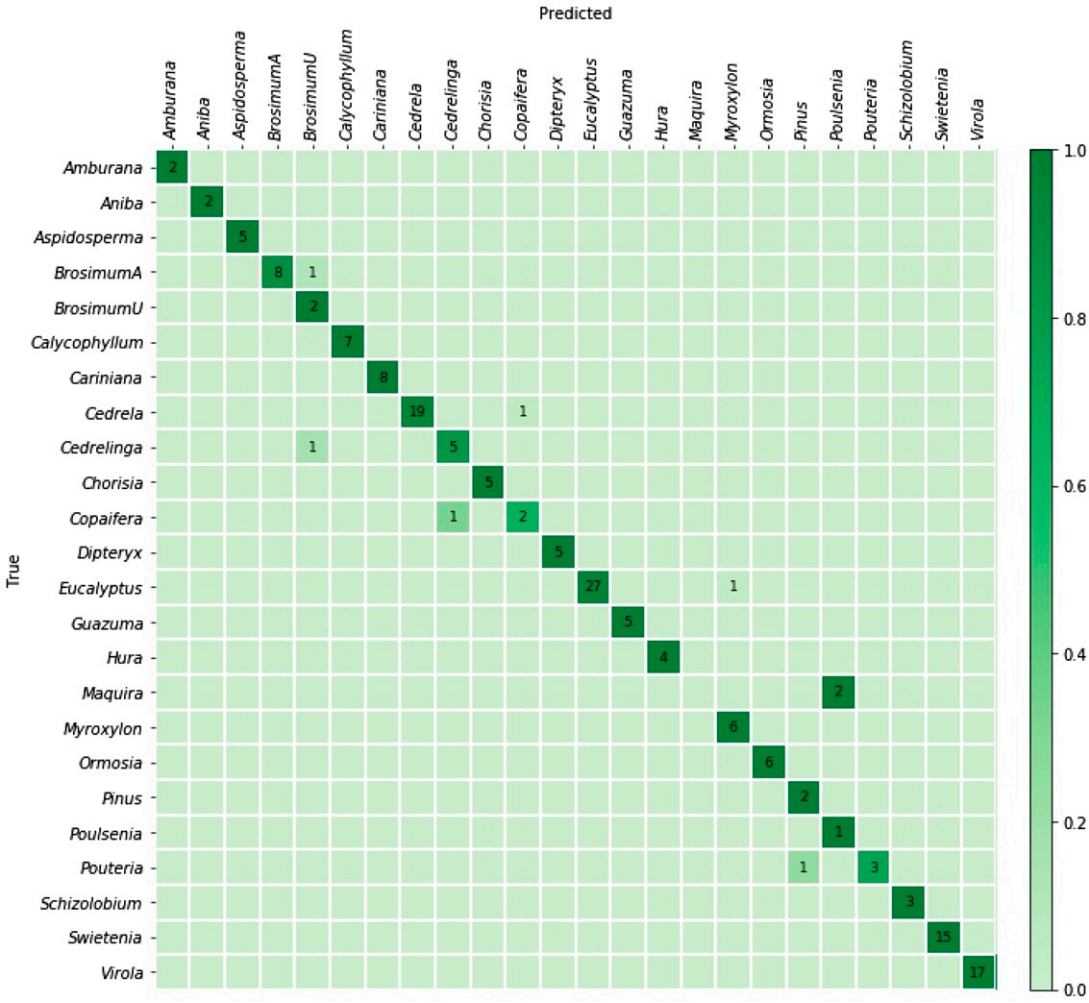

Figure A6. Confusion matrix for ResNet 50, 800 sanding grit.

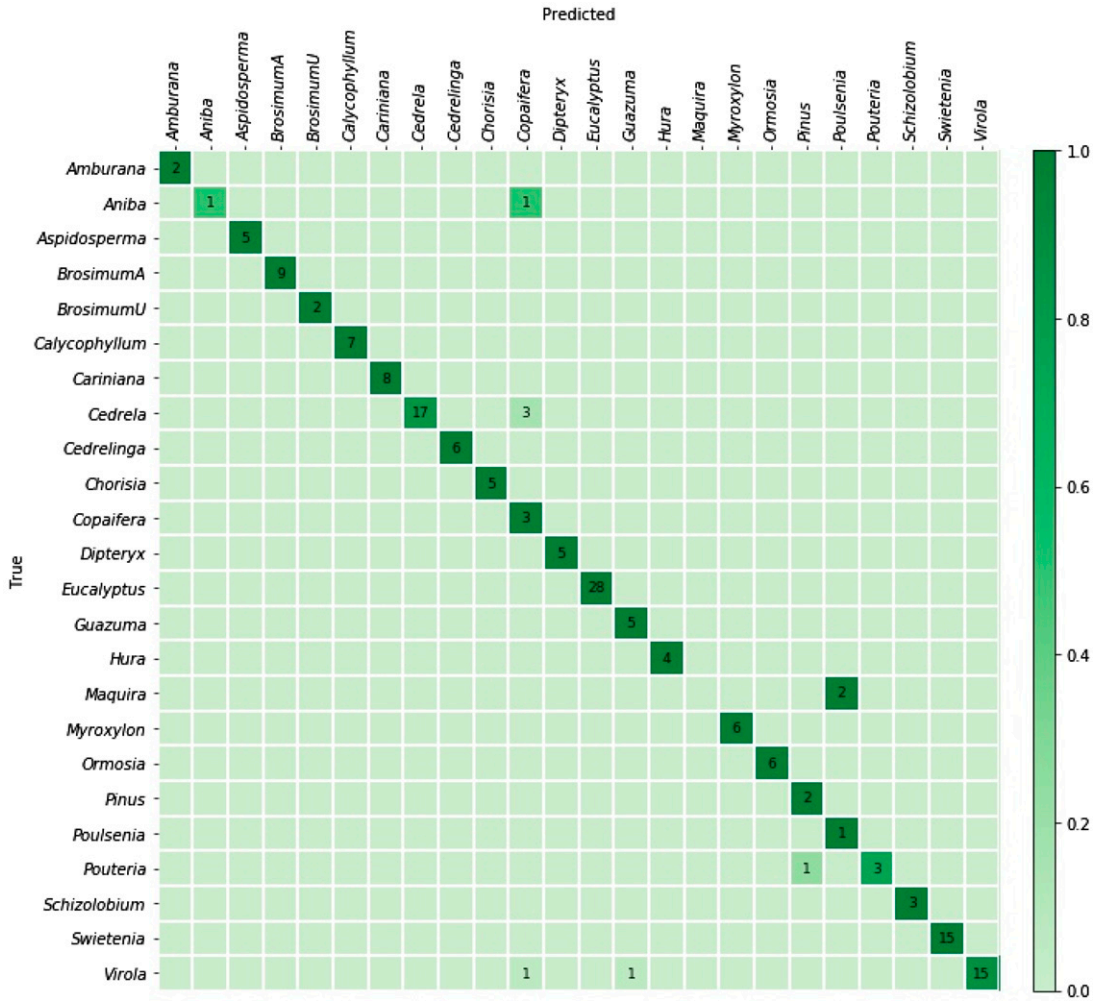

Figure A7. Confusion matrix for ResNet 50, 600 sanding grit.

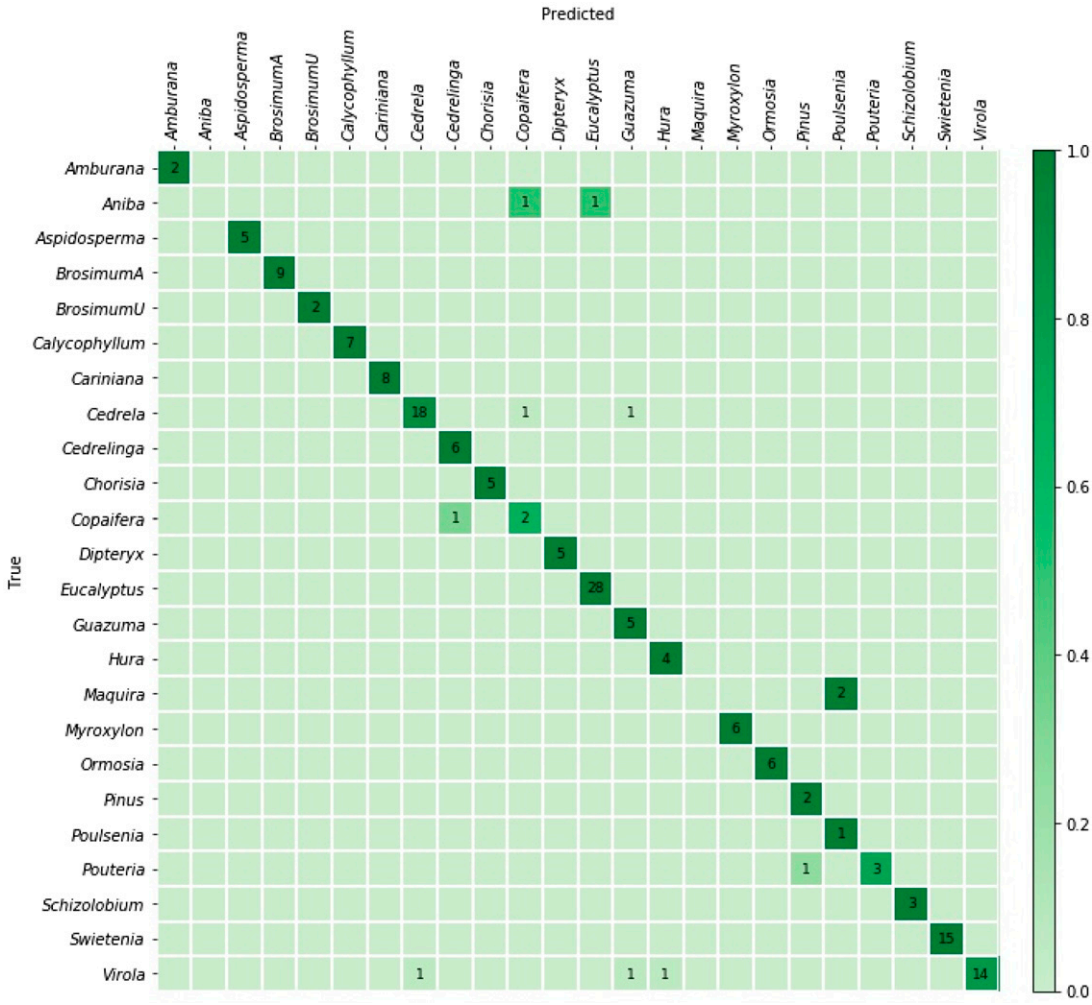

Figure A8. Confusion matrix for ResNet 50, 400 sanding grit.

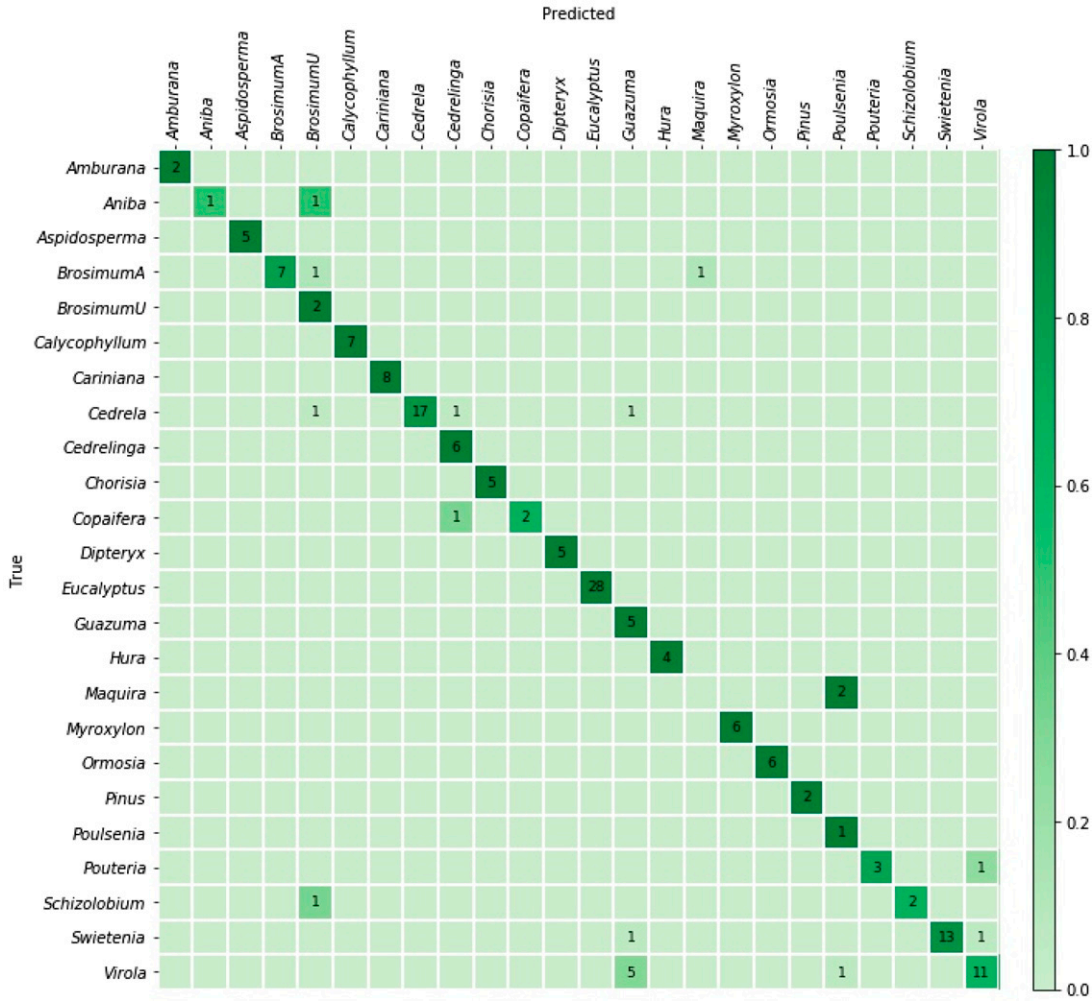

Figure A9. Confusion matrix for ResNet 50, 240 sanding grit.

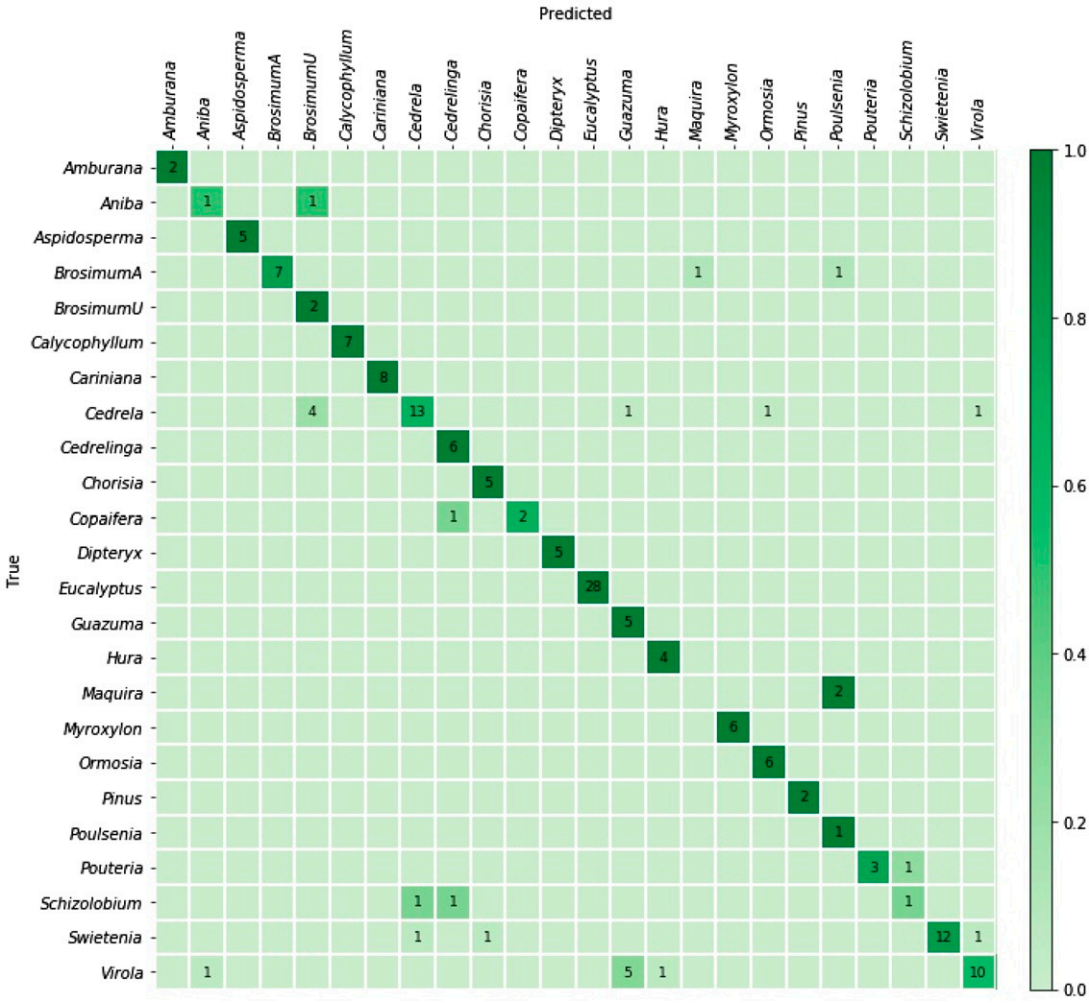

Figure A10. Confusion matrix for ResNet 50, 180 sanding grit.

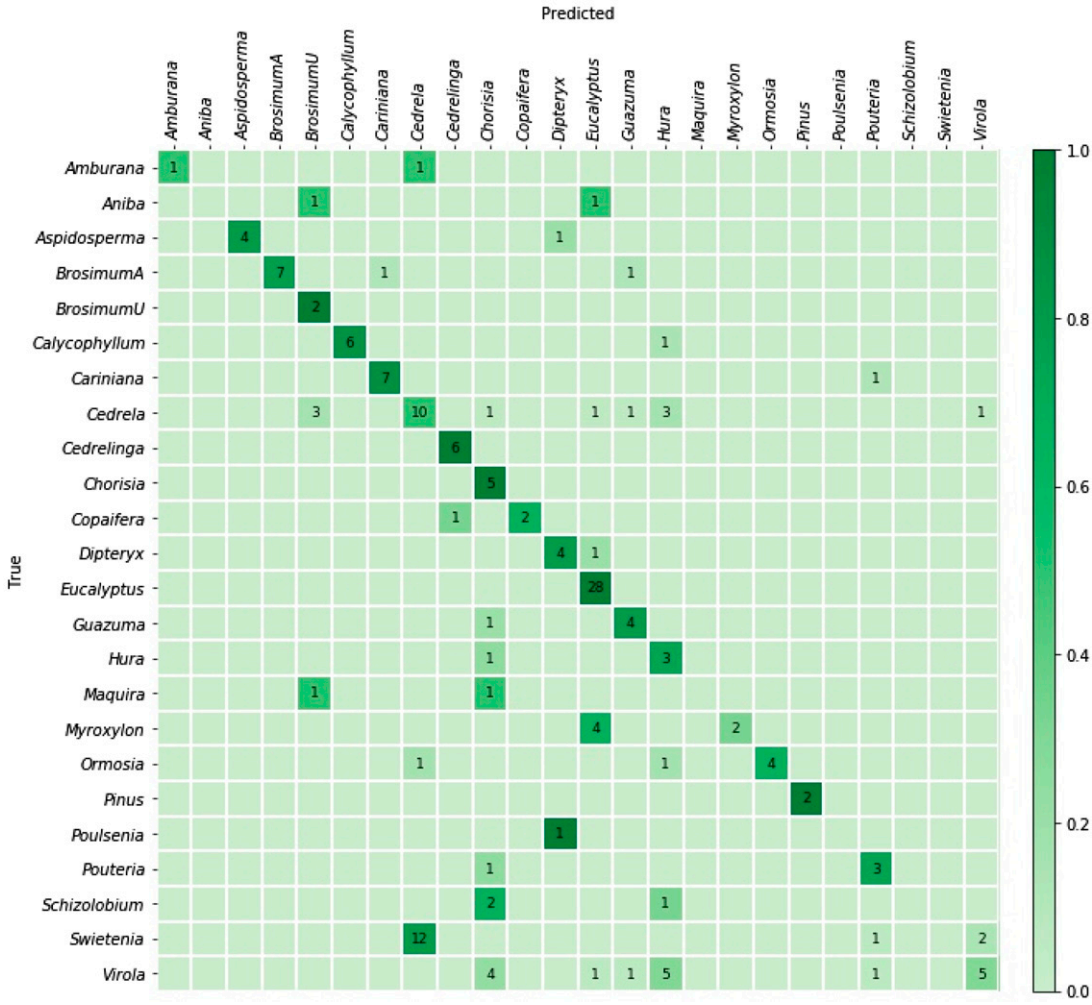

Figure A11. Confusion matrix for ResNet 50, 80 sanding grit.

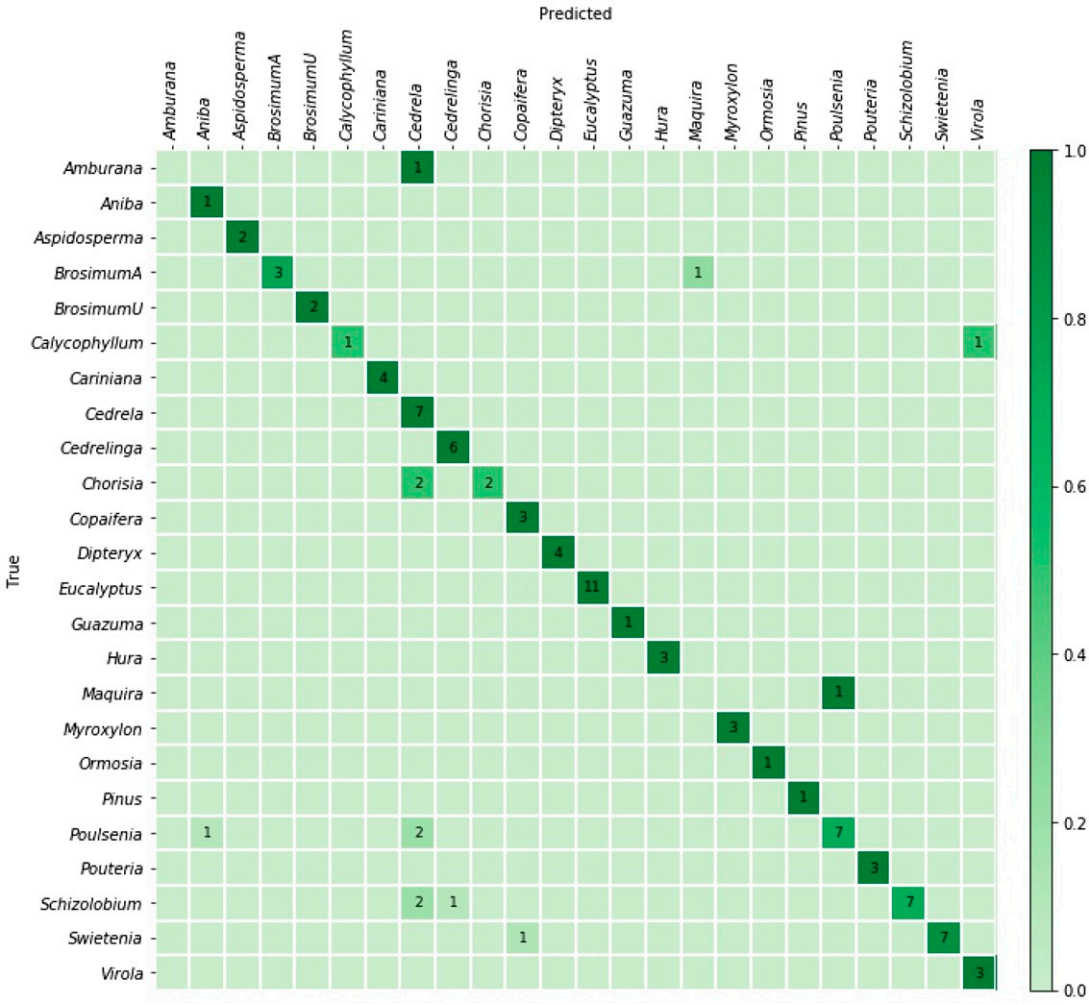

Figure A12. Confusion matrix for ResNet 50, knife cuts.