Received: 3 July 2022 | Revised: 14 March 2023 | Accepted: 26 March 2023

DOI: 10.1002/cem.3479

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Evaluation of spectral collection strategies for identification of *Dalbergia* spp. using handheld laser-induced breakdown spectroscopy

Caelin P. Celani[1] | **Rachel A. McCormick**[1] | **Amelia M. Speed**[1] | **William Johnston**[1] | **James A. Jordan**[2] | **Tyler B. Coplen**[2] | **Karl S. Booksh**[1]

[1]Department of Chemistry and Biochemistry, University of Delaware, Newark, Delaware, USA

[2]U.S. Geological Survey, Reston, Virginia, USA

Correspondence
Karl S. Booksh, Department of Chemistry and Biochemistry, University of Delaware, The Green, Newark, DE 19716, USA.
Email: kbooksh@udel.edu

Funding information
National Science Foundation, Grant/Award Number: CHE2003839

Abstract

The illegal timber trade has significant impact on the survival of endangered tropical hardwood species like *Dalbergia* spp. (rosewood), a world-wide protected genus from the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Due to increased threat to *Dalbergia* spp., and lack of action to reduce threats, port of entry analysis methods are required to identify *Dalbergia* spp. Handheld laser-induced breakdown spectroscopy (LIBS) has been shown to be capable of identifying species and establishing provenance of *Dalbergia* spp. and other tropical hardwoods, but analysis methods for this work have yet to be investigated in detail. The present work investigates five well-known algorithms—partial least squares discriminant analysis (PLS-DA), classification and regression trees (CART), *k*-nearest neighbor (*k*-NN), random forest (RF), and support vector machine (SVM)—two training/test set sampling regimes, and data collection at two signal-to-noise (S/N) ratios to assess the potential for handheld LIBS analyses. Additionally, imbalanced classes are addressed. For this application, SVM and RF yield near identical results (though RF takes nearly 100 longer to compute), while the S/N ratio has a significant effect on model success assuming all else is equal. It was found that forming a training set with replicate low S/N analyses can perform as well as higher precision training sets for true prediction, even if the predicted samples have low signal to noise! This work confirms handheld LIBS analyzers can provide a viable method for classification of hardwood species, even within the same genus.

KEYWORDS

chemometrics, classification, forensics, forestry, laser-induced breakdown spectroscopy

1 | INTRODUCTION

Handheld spectroscopic instrumentation has been rapidly increasing in analytical power and breadth of application over the last 5 to 10 years.[1] Many spectroscopic instruments are commercially available in handheld or portable form, including Raman, infrared (IR), X-ray fluorescence (XRF), and laser-induced breakdown spectroscopy (LIBS). The

ability to readily collect large quantities of data in the field, with minimal sample transport, preparation, and storage, makes handheld spectrometers an intriguing tool in metallurgy,[2,3] botany,[4] archeology,[5,6] mining,[7] geology,[8,9] recycling,[10] and forensics,[11] among other fields. While many traditional laboratory analyses can now be performed in the field with handheld analogs to benchtop instruments, handheld spectrometers generally have poorer resolution and signal-to-noise (S/N) performance. Consequently, chemometric and machine learning tools are commonly employed to extract the maximal quality of information from collections of spectra acquired with handheld instrumentation.

One potential use for handheld spectroscopic instrumentation is to help combat the illegal timber trade by rapidly and reliably identifying timber from endangered species in the field or at ports of entry. The illegal timber trade is valued by Global Financial Integrity between $52 to $157 billion USD annually.[12] In addition to face value costs, illegally logged timber is estimated by the World Bank to result in losses of $5 billion USD in tax revenue and depreciation of global timber prices by 7% to 16%.[13] A major contributor to the total mass of illegally logged timber is rosewood, or *Dalbergia* spp. The genus *Dalbergia* comprises 250 species and contains all "true rosewoods," which are prized for their deep red hue, pleasant, sweet scent, high density, and cultural significance. These traits make rosewood a highly desirable timber to be manufactured into furniture, cabinetry, flooring, and string instruments, despite its protected status.

It was estimated that *Dalbergia* spp. accounts for nearly 35% of all seized illegal flora and fauna with approximately 85% listing China as the destination.[14] In late 2017, the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) uplisted all species of *Dalbergia* to Appendix II, "permitting international trade only with an export permit," and uplisting *Dalbergia nigra* further to Appendix I, "prohibiting international trade outside of exceptional circumstances."[15] Despite these new regulations, the illegal harvesting and trade of rosewood species have increased by 129% and 120%, respectively, from 2016 to 2018.[16]

Because it is challenging to interdict the illegal exportation of rosewood at the source, an alternative route to policing this trade is via field detection with handheld instrumentation by customs agencies at the port of entry. By applying chemometric and machine learning classification models, the species and provenance may be rapidly discerned.[17] Current state of the art analyses rely on time consuming laboratory-based methods, making point-of-entry detection difficult.[18] Chemometric analysis applied to handheld instrumentation can enable rapid field detection, thereby requiring laboratory analyses only when field measurements suggest the shipment materials and listed manifest materials do not match.

Multiple spectroscopic techniques have been applied to classification of woods in general with a few studies looking specifically at *Dalbergia*, which will be discussed here. IR spectroscopy has shown classification accuracies across multiple species ranging from 50% to 100% depending on preprocessing, modeling algorithms, sample species, and so on.[19–21] A recent study compared partial least squares discriminant analysis (PLS-DA) to soft independent modeling by class analogy (SIMCA) showing ∼90% efficiency for PLS-DA and ∼50% efficiency for SIMCA to classify seven species of *Dalbergia*.[22] This study also reported efficiency rates of 90% for PLS1-DA and 85% for SIMCA to separate *D. nigra* (CITES Appendix I) from all other measured rosewood species (CITES Appendix II) in a 1 versus the world classification.[22] Though XRF, Raman, and other techniques are available options, there are few XRF studies for wood classification, none of which include classification of *Dalbergia* species,[23] and Raman has limited instances of wood classification and only a single analysis of *Dalbergia*.[24] Other methods have recently been published to address *Dalbergia* characterization.[25]

LIBS offers an intriguing option for classification of *Dalbergia* species because there is a strong correlation between *Dalbergia* species and geographic origin. Consequently, the metal signature of the soil, as transferred into the wood, may form a pathway to determine *Dalbergia* species of rosewood lumber and products. In recent years, handheld LIBS has been used in a wide variety of areas where elemental signature characterization is important, including mining,[2,3] astrochemistry,[26] wood analysis,[27] and other applications.[28] In one recent study, a homemade portable LIBS system in tandem with a near IR spectrometer was used to discriminate varying wood genera and showed a 90% classification accuracy.[29] In a similar vein, LIBS was coupled with an electronic nose sensor to classify four Brazilian woods of differing genus with perfect classification success.[27] Alternatively, a benchtop LIBS system was used to differentiate seven *Dalbergia* species with algorithm dependent classification rates from 52% to 100%.[30] It should also be noted here that LIBS frequently requires multivariate analysis and chemometrics regardless of application due to the highly overlapping and information rich signals.[17,26,27,29–32]

There are multiple classification algorithms that have been applied to problems analogous to the classification of *Dalbergia* species via LIBS.[17,26,27,29–32] One facet of this study considers how six common supervised classification methods perform modeling large amounts low S/N data compared with smaller amounts of high S/N data.

Principal component analysis (PCA)[33] is perhaps the most common "classification" strategy, likely due to the method's simplicity and model interpretability. However, PCA is an unsupervised method and lacks the capacity to directly employ known class memberships in the training set to achieve a more optimal model. PLS-DA[34,35] is a supervised parametric method that employs partial least squares regression (PLSR) to perform a binary split between two groups of classes. Like PCA, PLS-DA works on a lower dimensional subspace of the data to reduce the impact of embedded noise. The PLSR space can be optimized for one binary split (PLS1-DA) or multiple simultaneous binary splits (PLS2-DA). Similarly, a multiclass classification problem can be approached by multiple PLS1-DA models where each classification model is not constrained to share a common subspace or by PLS2-DA where all classification models are constrained to share a common subspace. In this study, PLS2-DA is employed.

k-Nearest neighbor (*k*-NN)[36] is a supervised, nonparametric method that classifies new samples based on their Euclidean (or other) distances from samples of known classification. *k*-NN has the advantage of making very few assumptions about the data or noise structures. The plurality of the classes of the k training set samples nearest to the unknown sample predicts the unknown sample's class membership. Note that $k = 1$ gives an upper-bound estimate of the Bayesian error rate of a classification problem,[37] as *k*-NN1 converges to an error rate of at most twice the Bayesian error rate.

Classification methods more associated with machine learning have also become increasingly popular. Support vector machine (SVM)[38–40] is a supervised parametric method that employs linear or nonlinear transformations of the observed variables to increase the dimensionality of the dataspace. Classification and regression trees (CART)[41,42] is a supervised nonparametric method where classification is approached by performing binary splits of the dataset. Information theory is employed to find the "optimal" rule for every split in succession. CART is considered a "greedy" algorithm because the optimal split at an early step may not lead to the final optimal model. Lastly, random forest (RF)[43,44] is a supervised, nonparametric method that seeks to overcome the "greedy algorithm" limitation of CART by averaging over many CART models. RF is effectively a bootstrap of CART by randomly seeding multiple CART models with different starting variables.

In the context of employing handheld instrumentation, this manuscript proposes the hypothesis that rapid replicate analyses can compensate for lower signal-to-nose instrumental responses. This hypothesis is tested with respect to classification method choice, balance of class sizes within a training set, and average class size within a training set. The chosen data testbed of *Dalbergia* ssp. identification by handheld LIBS presents classification challenges due to seven or more highly overlapped classes—a real-world application that pushes the current limits of LIBS capabilities.

2 | EXPERIMENTAL

2.1 | Samples

The complete set of hardwood samples consisted of 235 labeled ~0.5 × 2 × 3 inch wood chips across 20 classes (species) containing both *Dalbergia* (11 classes) and non-*Dalbergia* (nine classes) genera purchased from commercial sources. From this sample pool, an exploratory collection of 159 wood chips across 13 classes were analyzed (Table 1). Class assignment was based on the vendor-reported species, verified by Coplen et al.,[45] and each wood chip reported to be from a distinct tree. All sample assignments were validated with mass spectrometry and wood anatomy analysis.[46] Additionally, the number of wood chips per class varied between 6 and 17 (Table 1). Samples of a given species were randomly chosen over a series of days for spectral analysis via handheld LIBS with one of two sampling parameter options, outlined below. Samples were not modified and were analyzed as they arrived (i.e. no sample preparation was used).

2.2 | Spectra collection

LIBS spectra were collected using a handheld SciAps Z-300 LIBS Analyzer (SciAps, Boston, MA) with Argon purge (1064 nm laser source, 6 mJ pulse energy, 50 Hz repetition rate) over the full 190–950 nm spectral range. No instrument optimization was performed prior to wood chip analysis with calibration and wavelength selection being completed automatically with an internal standard of 316 stainless steel. Sampling parameters were varied, creating a lower signal-to-noise dataset containing 1779 spectra across 12 classes and a higher signal-to-noise dataset containing 282 spectra across eight classes, which will be referred to as Set 1 and Set 2 (S1 and S2), respectively (Table 1). It should

TABLE 1 Class information of analyzed samples.[45]

Class	Sets	Number of wood chips	Location	Genus sp.	Common name
1[a]	1	12	Honduras	*Dalbergia stevensonii*	Rosewood
2[a]	1	11	Madagascar	*Dalbergia* spp.	Rosewood
3[b]	1, 2	6	Para, Brazil	*Dalbergia spruceana*	Rosewood
4[a]	1, 2	11	Surinam	*Dicorynia paraensis*	Angelique
5[a]	1	16	Brazil	*Dalbergia cearensis*	Kingwood
6[a]	1, 2	13	Malaysia	*Dalbergia* spp.	Indonesian rosewood
7[a]	1	15	Brazil	*Dalbergia frutescens*	Tulipwood
9[b]	1, 2	6	Madagascar	*Dalbergia maritima*	Rosewood
10[a]	1	17	Tanzania	*Dalbergia melanoxylon*	African blackwood
11[a]	1, 2	8	Mexico	*Caesalpinia platyloba*	Chakte Viga
12[a]	1, 2	13	India	*Dalbergia latifolia*	Indian rosewood
14[a]	1, 2	17	Brazil	*Dalbergia nigra*	Brazilian rosewood
18[a]	2	14	Brazil	*Phoebe porosa*	Imbuia

[a]Eisenbrand Inc. Torrance, CA.
[b]Gilmer Wood Co. Portland, OR.

be noted that minor deviations from 12 spectra occurred in Set 1 leading to the discrepancy between the expected 1740 spectra and 1779 spectra seen in Set 1. The acquisition parameters for Set 1 are as follows: 12 sampling locations/wood chip, 1 × 1 raster grid (single point measurement), one "cleaning shot" (unrecorded laser blast to remove surface debris), and four collection shots (averaged). The acquisition parameters for Set 2 are as follows: three sampling locations, 1 × 1 raster grid, four "cleaning shots," and 12 collection shots (averaged). This sampling scheme thus created a theoretical increase of S/N by a factor of 1.7 for Set 2 over Set 1 for any single spectrum. Sets 1 and 2 also differed in the total number of analyzed species. The full set of 20 classes could not be completed before the borrowed LIBS was returned to SciAps.

2.3 | Normalization and preprocessing

All data analysis was done in RStudio[47] version 1.2.5042 running R version 3.6.3 base R[48] and using the packages "doParallel,"[49] "R.matlab,"[50] "signal,"[51] "caret,"[52] "chemometrics,"[53] "rpart.plot,"[54] "dplyr,"[55] "stats,"[48] and "car."[56] Prior to application of any of the classification algorithms, spectra were normalized to the mean intensity of each spectrum to account for matrix effects, shot-to-shot variance, and other environmental factors by taking the ratio of each variable in a given spectrum to the mean of all variables in that spectrum (Figure 1).[57]

After normalization, 10 training/test set pairs were created using two pseudorandom sampling strategies to assign 25% of the data to the test set. The two methods chosen were "Stratified by Spectrum," which treats each spectrum as independent variables, and "Stratified by Sample," which treats each wood chip as an independent variable. These methods differ in that the former includes all wood chips in the training and test sets, thus yielding more optimistic results, while the latter simulates a model that has new samples it has never seen before introduced for prediction. For the "Stratified by Spectrum" strategy, each spectrum within every class was randomly assigned to the test set with a probability of 0.25. This resulted in the optimistic case where spectra from the same wood sample, but from different locations on that sample, would be in both the training and test sets. The "Stratified by Spectra" strategy yielded training/test splits of average size 1339/440 for Set 1 and 193/89 for Set 2. For the "Stratified by Sample" strategy, all the spectra for each of the wood samples within every class were assigned to the test set with a probability of 0.25. This is a more realistic split where the spectra from all locations within a single wood sample are assigned to either the training or test set. The "Stratified by Sample" strategy yielded training sets ranging from 1386 to 1445 spectra and test sets ranging from 393 to 334 spectra for Set 1; training sets ranged from 217 to 228 spectra for Set 2 with test sets between 65 and 54 spectra. To see train/test sizes for each of the 10 resamplings, see Table SI1. For both sampling techniques, the pseudorandom sampling seeds used for the 10 resampled sets were 1, 579, 609,430, 7, 934, 42, 2153, 570, 347, and 12,340.

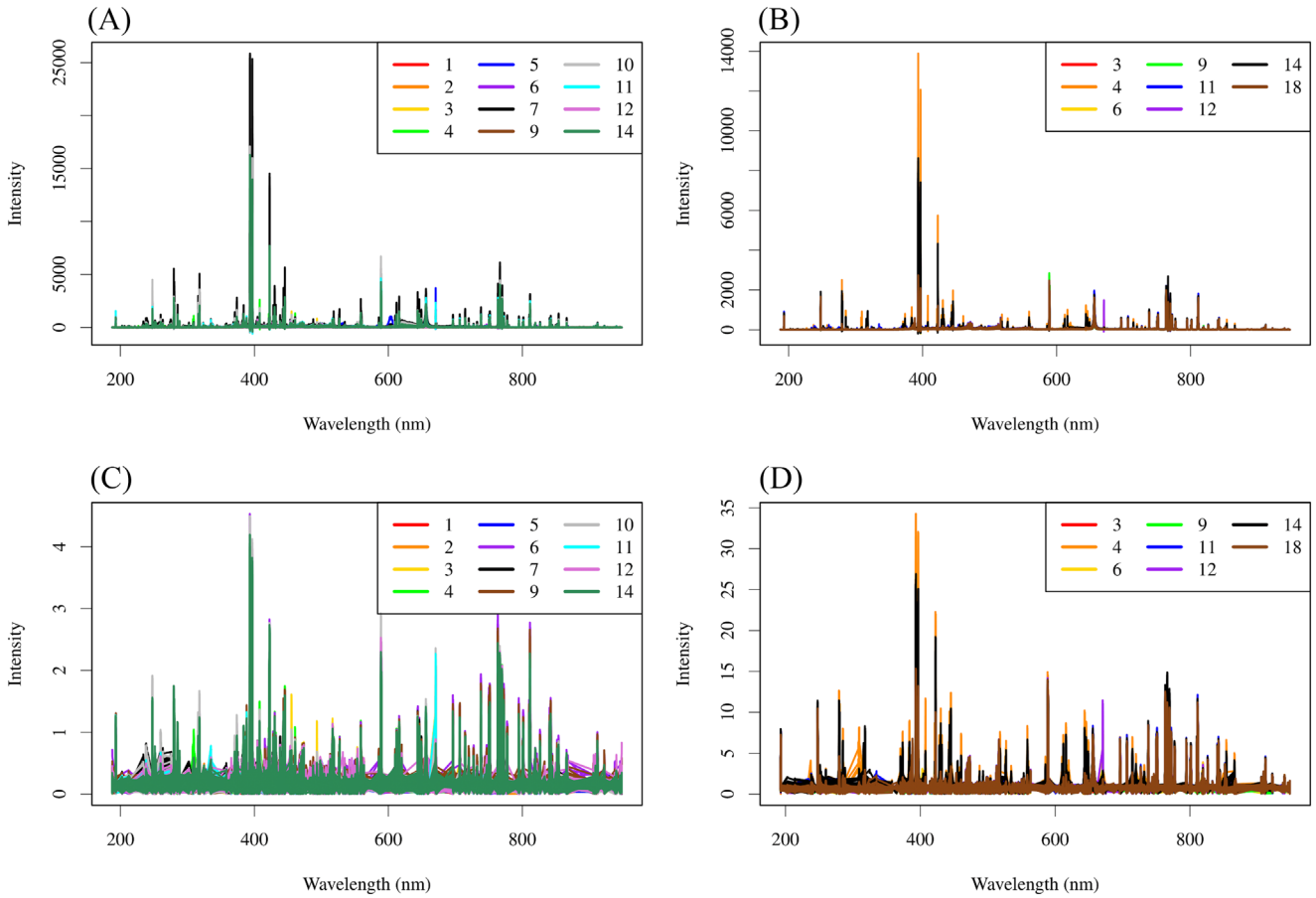

FIGURE 1 Raw spectra from the full dataset for (A) Set 1 and (B) Set 2 and example preprocessed spectra prior to autoscaling from the first training set of the full dataset for (C) Set 1 and (D) Set 2.

Prior to analyses, each spectrum was preprocessed in a similar but modified manner as the previous investigation of *Dalbergia* classification with LIBS, which will also be detailed here.[17] First, each spectrum was baseline corrected and smoothed by Savitzky–Golay filtering using a first derivative, second-order polynomial, and 15-point window ($p = 2$, $n = 15$, $m = 1$). In order to minimize spectral distortion and minimize residual noise, a 15-point window was chosen such that it was approximately one quarter of the peak width at half max. As the Savitzky–Golay window approaches one half peak width at half max, spectral distortion begins to become noticeable. A second-order polynomial was chosen based on best practices; higher order polynomials leave more residual noise. Lastly, a first derivative was chosen to remove local baseline offsets. Higher order derivatives remove sloping or curved baselines; these derivatives also introduce more residual noise. Because a nonnormal distribution of values at each variable was observed, the absolute value of each datum was calculated and transformed to its square root. A more traditional transformation by a log function was investigated, but a log transformation, compared with a square root transformation, amplified the baseline noise relative to the LIBS peaks. To partially reduce noise in the data, baseline variables were identified and removed by a naïve variable selection method. For each training set, any variable with a mean value less than the global mean of the training set was designated as "baseline" and removed from both the training set and test set. It is understood that a small amount of useful variance may be discarded when using this variable selection method. The variable selected data were then mean centered and variance scaled (autoscaled) as the final step of preprocessing.

2.4 | Cross-validation (CV), training, and prediction

The caret[52] package in R was used for CV, model training, and model prediction. CV parameters were specified in the trainControl function and 5-fold repeated CV with 10 repeats was selected. These CV parameters were then passed to

the train function and used to tune the training models for each of the 10 training sets for all sampling methods and datasets. After a training model was established, the test set was predicted through the predict function. Confusion matrices, Cohen's kappa coefficient, and the no information rate for CV and prediction were compiled after each analysis as well as averaged across all 10 analyses. Model assessment was primarily based on Cohen's kappa coefficient,[58] or κ, which is an accuracy metric that accounts for random chance correct guesses and is mathematically expressed as

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_0 is accuracy and p_e is the no information rate. Other metrics, such as Matthew's correlation coefficient (MCC), do not account for differences in chance of correct guesses when comparing model efficacy across unbalanced datasets. That is to say that MCC can be fooled into presenting an optimistic assessment of accuracy for the largest classes in an unbalanced data collection and for datasets that are most unbalanced. Error of models was assessed using the standard error of the mean (SEM, $\sigma_{\bar{x}}$), which is defined as $\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$, where s is the standard deviation and N is the total number of samples (or in this case, total number of averaged κ coefficients).

It is important to note that for this study κ was calculated based on classification of individual spectra. Thus, although the training set for Dataset 1 has the opportunity to average the classification model over many replicate spectra, each spectrum in the prediction set is classified by itself without averaging replicate spectra.

2.5 | Analysis—Comparison of algorithms

The full preprocessed datasets (training/test sets) present in Set 1 and Set 2 for both sampling methods were analyzed independently in the caret[52] package by six classification methods: PCA, PLS2-DA, CART, RF, *k*-NN, and SVM. The primary comparison to be made between these analyses was to investigate overall algorithm success with multiple parameters varied between datasets to mimic the level of variability that would be associated with field measurements in a real scenario. These datasets will be referred to as "Full Set 1" and "Full Set 2."

2.6 | Analysis—Comparison of S/N ratio

To directly compare the effect of S/N across all classification methods, subsets of Set 1 and Set 2 were selected to be constituted of the same number of spectra per sample from only the samples held in common for both sets. Consequently, only Classes 3, 4, 6, 9, 11, 12, and 14 were employed for this study and three of the 12 spectra from each wood sample in Set 1 were randomly selected for retention. The subsets of data from Set 1 and Set 2 each had 229 spectra across 76 samples from seven classes. These samples were then used for both the "Stratified by Spectrum" and "Stratified by Sample" assignment to the training and test sets as outlined in Section 2.3. These datasets will be referred to as "Mutual Set 1" and "Mutual Set 2."

2.7 | Analysis—Comparison of balanced/imbalanced classes

It is well documented that when classes are imbalanced, machine learning classifiers tend to favor majority classes, decreasing the sensitivity for minority classes.[59,60] Some classes within the reported datasets, namely, Classes 3, 9, and 11, contain significantly fewer spectra than other classes due to a smaller number of wood chips of those species in the provided collection. Thus, to directly compare the effect of balanced versus imbalanced classes on sensitivity and selectivity, the datasets were reduced in size such that each class contained the same number of samples as the smallest class in that dataset. Samples were selected once again through random number generation of a subset of wood chips in each class, which yielded classes of 72 spectra per class (six wood chips, 12 spectra/chip) for Set 1 and 18 spectra per class (six wood chips, three spectra/chip) for Set 2. After the datasets were modified, the sets were analyzed identically to the previous sections. Because classes were balanced, all train/test splits were identical for both sampling regimes, which gave train/test split sizes of 54/18 for Set 1 and 12/6 for Set 2. These datasets will be referred to as "Balanced Set 1" and "Balanced Set 2."

3 | RESULTS AND DISCUSSION

In total, 60 different analyses (600 if each train/test split is counted as one analysis) were performed across the two datasets (Set 1 and Set 2) that differed by their S/N level. Each dataset was split into training and test sets based on two opposed sampling strategies: Stratified by Spectrum and Stratified by Sample. These random, without replacement, assignments were repeated 10 times using bootstrap statistics on model and method performance. The placements into training and test sets were determined prior to analyses and held immutable to facilitate better comparisons of outcomes. The datasets were parsed in three manners to isolate features during comparisons: the full data where the two sets had different classes and numbers of exemplars in each class, the mutual data subsets where only classes held mutually between both full datasets are modeled, and the balanced data subsets where the classes in each full dataset are constrained to have identical numbers of exemplar samples and spectra per exemplar sample. All the data subsets were analyzed by PLS2-DA, CART, *k*-NN, RF, and SVM.

3.1 | Full dataset

The general trend in the analyses of the full datasets is that SVM and RF return greater average κ than the other classification models for both CV and true prediction (Figure 2), albeit at the expense of interpretability. CART and *k*-NN performed moderately well but did not reliably reach the average κ levels of SVM and RF. These relatively poorer performances can be ascribed to the highly overlapped data clouds of each class in the full ~4100-dimensional space. The preponderance of noisy variables and lack of a clear subset of discriminating variables makes it easy for a greedy method like CART to become trapped in local minima. While the extensive class overlap in many variables confounds distance-based methods such as *k*-NN. Presumably, the performance metrics for CART and *k*-NN would each improve with a thorough application of variable selection. The flat PLS2-DA model was clearly the worst performing of the five methods tested. This is also a probable consequence of the overlapping, noisy classes. In previous analyses of similar data, PLS2-DA did not perform well for classifying nine exotic hardwood classes; a hierarchical decision tree of PLS1-DA models was needed to achieve an acceptable model.[17] It is noteworthy that PLS2-DA performed significantly better with the less-noisy Dataset 2 than with the more-noisy Dataset 1, further confirming the assertion regarding noisy data clouds. Additionally, it is worth noting that the average κ for CV is comparable with the average κ for true prediction; this provides confidence that the models are not significantly overfitting the data.

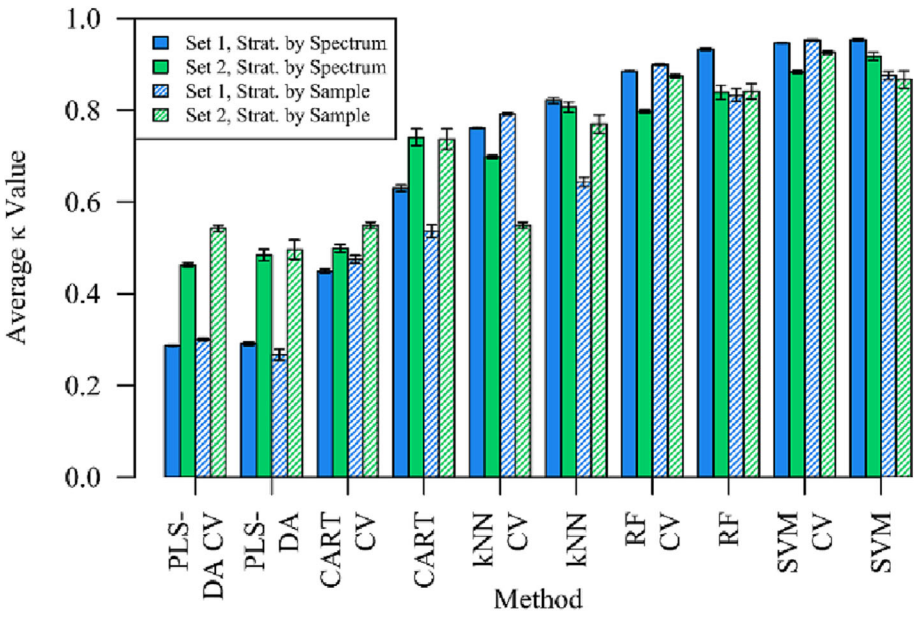

FIGURE 2 Bar plot showing the average $\kappa \pm 1\sigma_{\bar{x}}$ coefficient across each classification algorithm for each dataset and sampling strategy of the full (as-is) dataset.

With the full datasets, SVM and RF perform slightly better on the dataset with more samples (Set 1) than the dataset with greater S/N (Set 2). By way of contrast, CART and PLS2-DA perform significantly better on the greater S/N dataset. This highlights the reliance on high-quality training sets for successful application of CART and PLS2-DA models; these methods did not fully leverage the benefits of the larger training sets, when the data is noisy, compared with SVM and RF.

As expected for applications with replicate spectra, Stratification by Sample returns slightly lower average κ than does stratification by spectrum for true prediction. The average κ for true prediction is lower for Stratification by Sample six out of 10 times and statistically comparable the remaining four or 10 times. This effect is due, in part, to Stratification by Sample having a smaller basis set of unique samples in the training set. Also, replicate spectra residing in both the training and test sets for Stratification by Spectrum providing the best-case scenario for model performance. The average κ for CV is comparable in magnitude to the average κ for Stratification by Spectrum because the CV routines in caret are limited to Stratification by Spectrum i.e. sample-level data partitioning. However, the average κ for true prediction with Stratification by Sample provides a reasonable prediction of future model performance.

With respect to using handheld instrumentation as an alternative to higher S/N laboratory methods, the data show that exploiting handheld instrumentation's capability to rapidly collect replicate spectra can compensate for the relative deficiencies in S/N. For the best performing models (SVM and RF) using proper experimental design (Stratification by Sample), there is no difference in average κ between the low S/N yet many spectra Set 1 versus the high S/N yet few samples Set 2. However, some of the inferior models (PLS-DA and CART) do require high S/N data for construction of adequate calibration models and are, perhaps, not as attuned for use with handheld instrumentation.

3.2 | Mutual dataset

In order to more directly compare the effect of different S/N ratios on classification performance, the datasets were parsed to contain only classes mutually held and the number of spectra per sample in Set 1 was reduced to 4 in order to match the number of spectra in Set 2. This removed the possibility of the machine learning methods averaging over more, albeit noisier, spectra in Set 1 than in Set 2. Looking at the average κ values, the same broad trends hold in the mutual dataset (Figure 3) as hold in the full dataset (Figure 2). SVM and RF are the two best performing methods. CART and *k*-NN perform moderately well but do not consistently return κ values as large as SVM and *k*-NN. The flat PLS2-DA model was the worst performing of the five tested, returning average κ less than 0.5. The two stratification

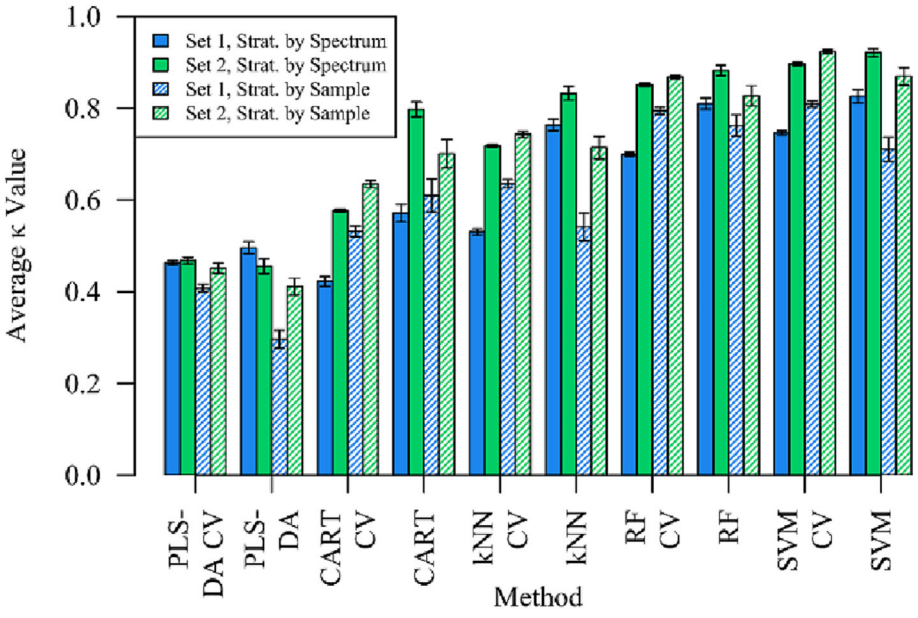

FIGURE 3 Bar plot showing the average $\kappa \pm 1\sigma_{\bar{x}}$ coefficient across each classification algorithm for each dataset and sampling strategy of the S/N dataset.

strategies perform comparably as with the full dataset. Stratification by Sample returns lower average κ in nine of the 10 true prediction instances and Stratification by Spectrum performs equivalently with CV and true prediction.

Isolating the effects of S/N ratio from other factors, such as number of spectra in each class, illustrates how handheld instrumentation can compensate for poorer S/N by exploiting the capability to easily perform additional samplings. When the number of spectra per wood sample is held constant, the greater S/N Set 2 returns greater average κ than does the lesser S/N Set 1. Recall that individual spectra in Set 2 have a S/N ratio that is a factor of 1.7 greater than individual spectra from Set 1, so this effect is not surprising. The relative improvement of average κ is especially prominent in the better performing SVM and RF models. By comparison, in the full datasets where Set 1 includes 12 low S/N spectra to only three high S/N spectra in Set 2, the average κ for SVM and RF in Set 1 is greater than the average κ for Set 2. In other words, the collection of more spectra for the classification models more than balances the effect of poor S/N. This effect was also observed in multivariate curve resolution applied to hyperspectral Raman imaging when the total spectral collection time was equal for high sample density /low S/N images and low sample density/high S/N images.[61]

With respect to hand held instrumentation, isolating the effect of S/N in each dataset reinforces the importance of multiple replicates per sample. When constraining each dataset to have the same number of samples, the high S/N dataset (Set 2) outperforms the low S/N dataset (Set 1) for true prediction as assessed by average κ.

3.3 | Balanced dataset

A balanced training set that eliminates model bias towards fitting the largest classes was constructed by constraining each class to have the same number of samples and spectra as the smallest class within each individual dataset. This reduced the size of both Set 1 and Set 2 relative to their original sizes but kept Set 2 significantly smaller than Set 1. From a practical perspective, the balanced dataset is most similar to the full dataset, except that the full dataset has a majority of classes that are well spanned by unique samples (i.e., 13–17 wood chips) and only a few classes with a small number of samples (i.e., six to eight wood chips) while all the classes are constrained to contain only six unique wood chips in the balanced dataset. Consequently, the balanced dataset probes the application of handheld instrumentation where the number of unique samples are low, but the number of replicates for each sample may be great.

Overall, the same trend holds that SVM and RF provide the largest average κ for both CV and true prediction (Figure 4). CART and *k*-NN perform worse than SVM and RF, with *k*-NN slightly outperforming CART. PLS2-DA performs poorly, especially on the low S/N Set 1. The average κ of true prediction, across all methods except the poorly

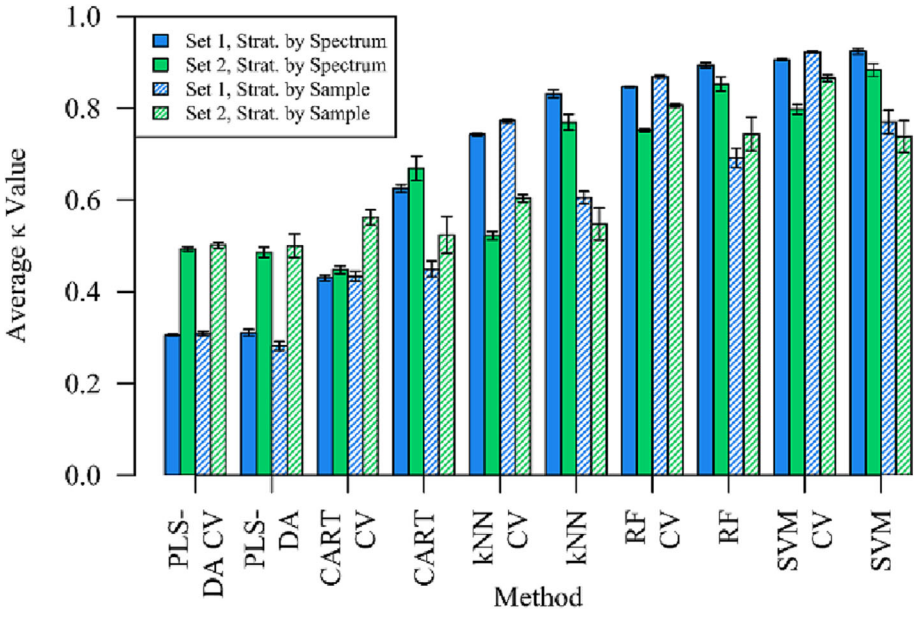

FIGURE 4 Bar plot showing the average $\kappa \pm 1\sigma_{\bar{x}}$ coefficient across each classification algorithm for each dataset and sampling strategy of the balanced dataset.

performing PLS2-DA, shows a significant drop from Stratification by Spectrum compared with Stratification by Sample. This is justifiable because Stratification by Sample pulls all instances of a unique wood chip from the training set, further reducing the span of the training set, while, on the other hand, Stratification by Spectrum has a 20% broader basis for the training set (five vs. six unique wood chips), and the spectra in the test set are replicates of samples in the training set.

From the perspective of deploying handheld instrumentation, the balanced dataset indicates that collection of replicate spectra cannot overcome a limited number of unique samples in the experimental design. That the average κ for true prediction with SVM and RF drops from 0.9 in the full dataset where most classes have 11+ unique wood chips to 0.7 in the reduced dataset with only six unique wood chips per class indicates that six unique samples are at, or below, the minimum tolerance for this application. There is no practical difference in outcome when collecting large numbers of noisy spectra (Set 1) versus lower numbers of precise spectra (Set 2).

3.4 | Statistical comparison of the classification models

To assess statistically significant differences between methods, one-way ANOVA was applied to the κ coefficients for each dataset. The Levene test for equal variance was applied, and no differences in variances were found at 99% confidence except for Balanced Set 2 when stratified by spectra ($p = 0.006$). Additionally, the Shapiro–Wilk test for normality was applied, and no deviations from normality were identified at 99% confidence except for Mutual Set 1 when stratified by spectra ($p = 0.009$). Once these conditions were checked, ANOVA was performed. It was found that there was a statistically significant difference between methods in each case, with all tests resulting in a p value of $<2 \times 10^{-16}$. Because all ANOVA tests showed statistically significant difference, multiple pairwise comparison was performed with a Tukey honest significant difference test. The results are included in Tables SI5.1–S5.3 and qualitatively summarized in Figure 5.

3.5 | Analyses in context

This study provides insight into optimizing data acquisition schemes for the determination of exotic hardwood species in the field by LIBS. Considering the practical collection of spectra with a handheld LIBS, it is advantageous to rapidly collect many (noisy) spectra than it is to slowly collect high S/N spectra. The handheld LIBS employed here ablates the sample and collects atomic emission spectra in a sequence of rapid laser pulses. A typical measurement would entail a small number of "cleaning shots" to remove a contaminated surface (e.g., 2) followed by a series of "data shots" that return optical emission spectra (e.g., 4–12). Such a scheme is often repeated over one to 12 locations, separated by 24 μm, in a serpentine pattern. All the collected spectra are averaged to return the final LIBS signal. Any movement of the sample during acquisition misaligns the "cleaning" and "data" locations. The typical duty cycle for laser pulses is 5–10 Hz; faster tends to overheat the instrument during long sampling campaigns. Holding the instrument steady with

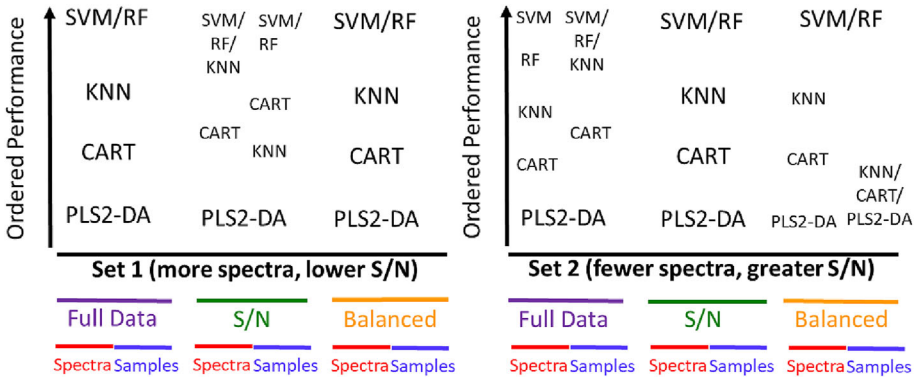

FIGURE 5 Qualitative ordering of predictive performance for the five methods tested across all data treatments. Method performance is based on Cohen's kappa value averaged over 10 resamplings of the training/test set assignment.

~10 μm tolerance is easy for the 0.5 s measurements in Set 1 and reasonable for the 2 s measurements in Set 2. However, it becomes increasingly difficult for rastered measurements if truly high S/N spectra are needed. This study shows that with the properly constructed training set, prediction of lower S/N can be just as accurate, on average, as prediction of higher S/N data. Consequently, demonstrating the ability to build successful classification models for rapidly collected data is of great benefit for future applications.

The limitation of this study is the relatively small scope of the training set. This study was limited to 10 classes per dataset. Current work is to scale the models up to 23 classes of exotic hardwoods in possession. The ultimate goal is to develop methodology for identifying exotic hardwood species in the field against a list of 250+ *Dalbergia* spp. and *Dalbergia* spp. lookalikes.

4 | CONCLUSIONS

This study shows that the relatively low S/N ratios often encountered in handheld instrumentation can be easily overcome by collecting more samples in the field and applying advanced machine learning methods such as SVM and RF. Such a strategy requires a philosophical shift from focusing on a few, high-quality measurements to embracing the capabilities of handheld instrumentation to perform near continuous sampling. This study also shows that collection of multiple replicates with handheld instrumentation cannot overcome a training set with limited numbers of unique exemplar samples. Specific to the *Dalbergia* application, this study demonstrates that rapidly collected, low S/N spectra are sufficient for classification provided that 10 or more unique exemplar samples are in each class for the training set.

ACKNOWLEDGMENTS

The authors gratefully acknowledge National Science Foundation (CHE2003839) for further support of this project. The authors also thank SciAps for the loan of a Z-300 LIBS Analyzer and Edgard Espinoza for the loan of wood samples. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US government.

DATA AVAILABILITY STATEMENT

Data are available in the Supporting Information.

ORCID

Caelin P. Celani https://orcid.org/0000-0002-9963-8340
Karl S. Booksh https://orcid.org/0000-0003-1803-376X

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3479.

REFERENCES

1. Crocombe RA. Portable spectroscopy. *Appl Spectrosc*. 2018;72(12):1701-1751. doi:10.1177/0003702818809719
2. Bennett BN, Martin MZ, Leonard DN, Garlea E. Calibration curves for commercial copper and aluminum alloys using handheld laser-induced breakdown spectroscopy. *Appl Phys B Lasers Opt*. 2018;124(42):58-66.
3. Rao AP, Jenkins PR, Vu DM, Auxier JD, Patnaik AK, Shattan MB. Rapid quantitative analysis of trace elements in plutonium alloys using a handheld laser-induced breakdown spectroscopy (LIBS) device coupled with chemometrics and machine learning. *Anal Methods*. 2021;13(30):3368-3378. doi:10.1039/D1AY00826A
4. Navarrete Gutiérrez DM, Pollard AJ, van der Ent A, et al. *Blepharidium guatemalense*, an obligate nickel hyperaccumulator plant from non-ultramafic soils in Mexico. *Chemoecology*. 2021;31(3):169-187. doi:10.1007/s00049-021-00338-4
5. Huisman DJ, Bach A, Joosten I, Ngan-Tillard DJM, van den Eynde G. Ceramic-faience hybrids were used to recycle bronze in North-Western European Iron Age egg-shaped crucibles. *J Archaeol Sci Rep*. 2020;32:102421. doi:10.1016/J.JASREP.2020.102421
6. Crocco R, Huisman H, Sablerolles Y, Henderson J, van Os B, Nieuwhof A. Hunting colours: origin and reuse of glass tesserae from the Wierum terp. *Archaeol Anthropol Sci*. 2021;13(9):1, 155-22. doi:10.1007/s12520-021-01391-4
7. Bellie V, Gokulraju R, Rajasekar C, Vinoth S, Mohankumar V, Gunapriya B. Laser induced breakdown spectroscopy for new product development in mining industry. *Mater Today Proc*. 2021;45:8157-8161. doi:10.1016/J.MATPR.2021.02.459

8. Blinkenberg KH, Anderskouv K, Sheldon E, Bjerrum CJ, Stemmerik L. Stratigraphically controlled silicification in Danian chalk and its implications for reservoir properties, southern Danish Central Graben. *Mar Pet Geol*. 2020;115:104134. doi:10.1016/j.marpetgeo.2019.104134
9. Goff K, Schaetzl RJ, Chakraborty S, Weindorf DC, Kasmerchak C, Bettis EA. Impact of sample preparation methods for characterizing the geochemistry of soils and sediments by portable X-ray fluorescence. *Soil Sci Soc Am J*. 2020;84(1):131-143. doi:10.1002/SAJ2.20004
10. Robey NM, Solo-Gabriele HM, Jones AS, Marini J, Townsend TG. Metals content of recycled construction and demolition wood before and after implementation of best management practices. *Environ Pollut*. 2018;242(Pt B):1198-1205. doi:10.1016/J.ENVPOL.2018.07.134
11. de Araujo WR, Cardoso TMG, da Rocha RG, et al. Portable analytical platforms for forensic chemistry: a review. *Anal Chim Acta*. 2018;1034:1-21. doi:10.1016/J.ACA.2018.06.014
12. World Bank. *Illegal Logging, Fishing, and Wildlife Trade: The Costs and How to Combat It*. World Bank; 2019.
13. Illegal Logging. Global Forest Atlas. Accessed February 4, 2020. https://globalforestatlas.yale.edu/forest-use-logging/logging/illegal-logging
14. Zhu AL. China's rosewood boom: a cultural fix to capital overaccumulation. *Ann Am Assoc Geogr*. 2020;110(1):277-296. doi:10.1080/24694452.2019.1613955
15. Convention on International Trade in Endangered Species of Wild Fauna and Flora. *The CITES Appendices*. Convention on International Trade in Endangered Species of Wild Fauna and Flora. Accessed February 5, 2020. https://www.cites.org/eng/app/index.php
16. Dumenu WK. Assessing the impact of felling/export ban and CITES designation on exploitation of African rosewood (*Pterocarpus erinaceus*). *Biol Conserv*. 2019;236:124-133. doi:10.1016/J.BIOCON.2019.05.044
17. Celani CP, Lancaster CA, Jordan JA, Espinoza EO, Booksh KS. Assessing utility of handheld laser induced breakdown spectroscopy as a means of *Dalbergia* speciation. *Analyst*. 2019;144(17):5117-5126. doi:10.1039/C9AN00984A
18. International Consortium on Combating Wildlife Crime. Best practice guide for forensic timber identification. 2016. https://www.unodc.org/documents/Wildlife/Guide_Timber.pdf
19. Yang Z, Liu Y, Pang X, Li K. Preliminary investigation into the identification of wood species from different locations by near infrared spectroscopy. *BioResources*. 2015;10(4):8505-8517. doi:10.15376/biores.10.4.8505-8517
20. Chen H, Ferrari C, Angiuli M, Yao J, Raspi C, Bramanti E. Qualitative and quantitative analysis of wood samples by Fourier transform infrared spectroscopy and multivariate analysis. *Carbohydr Polym*. 2010;82(3):772-778. doi:10.1016/j.carbpol.2010.05.052
21. Duca D, Pizzi A, Rossini G, Mengarelli C, Foppa Pedretti E, Mancini M. Prediction of hardwood and softwood contents in blends of wood powders using mid-infrared spectroscopy. *Energy Fuel*. 2016;30(4):3038-3044. doi:10.1021/acs.energyfuels.5b02994
22. Snel FA, Braga JWB, da Silva D, et al. Potential field-deployable NIRS identification of seven *Dalbergia* species listed by CITES. *Wood Sci Technol*. 2018;52(5):1411-1427. doi:10.1007/s00226-018-1027-9
23. Fellin M, Negri M, Zanuttini R. Multi-elemental analysis of wood waste using energy dispersive X-ray fluorescence (ED-XRF) analyzer. *Eur J Wood Prod*. 2014;72(2):199-211. doi:10.1007/s00107-013-0766-4
24. Almeida MR, Fidelis CHV, Barata LES, Poppi RJ. Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation. *Talanta*. 2013;117:305-311. doi:10.1016/J.TALANTA.2013.09.025
25. Shang D, Brunswick P, Yan J, et al. Chemotyping and identification of protected *Dalbergia* timber using gas chromatography quadrupole time of flight mass spectrometry. *J Chromatogr A*. 2020;1615:460775. doi:10.1016/J.CHROMA.2019.460775
26. Senesi GS, Manzari P, Consiglio A, De Pascale O. Identification and classification of meteorites using a handheld LIBS instrument coupled with a fuzzy logic-based method. *J Anal At Spectrom*. 2018;33(10):1664-1675. doi:10.1039/C8JA00224J
27. Cordeiro JR, Martinez MIV, Li RWC, et al. Identification of four wood species by an electronic nose and by LIBS. *Int J Electrochem*. 2012;2012:1-5. doi:10.1155/2012/563939
28. Senesi GS, Harmon RS, Hark RR. Field-portable and handheld laser-induced breakdown spectroscopy: historical review, current status and future prospects. *Spectrochim Acta Part B At Spectrosc*. 2021;175:106013. doi:10.1016/j.sab.2020.106013
29. Leandro JGR, Gonzaga FB, de Latorraca J. Discrimination of wood species using laser-induced breakdown spectroscopy and near-infrared reflectance spectroscopy. *Wood Sci Technol*. 2019;53(5):1079-1091. doi:10.1007/s00226-019-01119-7
30. Cui X, Wang Q, Zhao Y, Qiao X, Teng G. Laser-induced breakdown spectroscopy (LIBS) for classification of wood species integrated with artificial neural network (ANN). *Appl Phys B Lasers Opt*. 2019;125(4):1-12. doi:10.1007/s00340-019-7166-3
31. Harmon RS, Hark RR, Throckmorton CS, et al. Geochemical fingerprinting by handheld laser-induced breakdown spectroscopy. *Geostand Geoanal Res*. 2017;41(4):563-584. doi:10.1111/GGR.12175
32. Harmon RS, Throckmorton CS, Hark RR, et al. Discriminating volcanic centers with handheld laser-induced breakdown spectroscopy (LIBS). *J Archaeol Sci*. 2018;98:112-127. doi:10.1016/J.JAS.2018.07.009
33. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag*. 1901;2:559-572.
34. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal Methods*. 2013;5(16):3790-3798. doi:10.1039/C3AY40582F
35. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. *J Chemometr*. 2014;28(4):213-225. doi:10.1002/CEM.2609
36. Peterson LE. K-nearest neighbor. *Scholarpedia*. 2009;4(2):1883. doi:10.4249/SCHOLARPEDIA.1883
37. Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13(1):21-27. doi:10.1109/TIT.1967.1053964
38. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov*. 1998;2(2):121-167. doi:10.1023/A:1009715923555

39. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett*. 1999;9(3):293-300. doi:10.1023/A:1018628609742
40. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Association for Computing Machinery; 1992:144-152. doi:10.1145/130385.130401
41. Loh WY. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2011;1(1):14-23. doi:10.1002/WIDM.8
42. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Taylor & Francis; 1984.
43. Breiman L. Random forests. *Random For*. 2001;5(32):5-32. doi:10.1201/9780429469275-8
44. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. IEEE; 1995:278-282. doi:10.1109/ICDAR.1995.598994
45. Coplen TB, Qi H, Jordan JA, et al. *Chemical and Isotopic Compositions of Tropical Wood Samples*. U.S. Geological Survey data release. doi:10.5066/P9SIYCHJ
46. Lancaster C, Espinoza E. Analysis of select *Dalbergia* and trade timber using direct analysis in real time and time-of-flight mass spectrometry for CITES enforcement. *Rapid Commun Mass Spectrom*. 2012;26(9):1147-1156. doi:10.1002/rcm.6215
47. RStudio Team. RStudio: integrated development environment for R. 2020. http://www.rstudio.com/
48. R Core Team. R: a language and environment for statistical computing. 2020. https://www.r-project.org/
49. Corporation M, Weston S. doParallel: foreach parallel adaptor for the "parallel" package. 2020. https://cran.r-project.org/package=doParallel
50. Bengtsson HR. Matlab: read and write MAT files and call MATLAB from within R. 2018. https://cran.r-project.org/package=R.matlab
51. Signal Developers. {signal}: signal processing. 2014. http://r-forge.r-project.org/projects/signal/
52. Kuhn M. Caret: classification and regression training. 2020. https://cran.r-project.org/package=caret
53. Filzmoser P, Varmuza K. Introduction to multivariate statistical analysis in chemometrics. 2017. https://cran.r-project.org/package=chemometrics
54. Milborrow S. Rpart.Plot: Plot "rpart" Models: an enhanced version of "Plot.Rpart". 2020. https://cran.r-project.org/package=rpart.plot
55. Wickham H, François R, Henry L, Müller K. Dplyr: a grammar of data manipulation. 2020. https://cran.r-project.org/package=dplyr
56. Fox J, Weisberg S. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks; 2019.
57. Fu X, Duan FJ, Huang TT, Ma L, Jiang JJ, Li YC. A fast variable selection method for quantitative analysis of soils using laser-induced breakdown spectroscopy. *J Anal At Spectrom*. 2017;32(6):1166-1176. doi:10.1039/C7JA00114B
58. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104
59. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. *J Big Data*. 2018;5(1):1-30. doi:10.1186/s40537-018-0151-6
60. Chen JJ, Tsai CA, Young JF, Kodell RL. Classification ensembles for unbalanced class sizes in predictive. *Toxicology*. 2006;16(6):517-529. doi:10.1080/10659360500468468
61. Smith JP, Smith FC, Ottaway J, et al. Raman microspectroscopic mapping with multivariate curve resolution–alternating least squares (MCR-ALS) applied to the high-pressure polymorph of titanium dioxide, TiO_2-II. *Appl Spectrosc*. 2017;71(8):1816-1833. doi:10.1177/0003702816687573

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Celani CP, McCormick RA, Speed AM, et al. Evaluation of spectral collection strategies for identification of *Dalbergia* spp. using handheld laser-induced breakdown spectroscopy. *Journal of Chemometrics*. 2024;38(5):e3479. doi:10.1002/cem.3479