Ann. For. Res. 68(1): 125-136, 2025
https://doi.org/10.15287/afr.2025.4012

ANNALS OF FOREST RESEARCH
www.afrjournal.org

Development of acorn discrimination model for warm-temperature evergreen oaks using hyperspectral analysis*

Gye-Hong Cho[1], Ye-Ji Kim[1], Koeun Jeon[1] and Kyu-Suk Kang[1,2]✉

Cho G.-H., Kim Y.-J., Jeon K., Kang K.-S., 2025. Development of acorn discrimination model for warm-temperature evergreen oaks using hyperspectral analysis. Ann. For. Res. 68(1): 125-136.

Abstract We used hyperspectral analysis to distinguish between acorns of Japanese red oak (*Quercus acuta* Thunb.) and ring-cup oak (*Quercus glauca* Thunb.), two closely related species of the evergreen oaks. To accomplish this, 631 Japanese red oak acorns and 505 ring-cupped oak acorns were collected from the seed orchard in Jeju Island, Korea, and hyperspectral imaging was performed. Two types of hyperspectral devices, Corning and Korea Spectral Products (KSP), were used to calibrate images and extract regions of interest. Average spectra were obtained from the extracted regions of interest, and morphological variables were added to the Corning data to form a dataset. Partial least square (PLS) was used as the learning model, Standard normal variate, Multiplicative scatter correction, and Savitzky-Golay filtering were applied as preprocessing techniques, and competitive adaptive reweighted sampling and successive projection algorithm were applied as variable selection techniques; and the combination of preprocessing method, the number of PLS components, and the number of selected variables were optimized. The lightweight model was generated from the selected variables, and the performance was improved by combining the morphological variables. As a result, the lightweight model based on Corning dataset showed 45~85% accuracy, and the lightweight model based on the KSP dataset showed 75~90% accuracy. The model utilizing morphological variables in the Corning-based lightweight model showed a high accuracy of 98-100%, so we were able to discriminate the acorns of evergreen oaks between *Q. acuta* and *Q. glauca*. The results of this study are expected to serve as a basis for future model development for seed classification of hybrid oak acorns.

Keywords: hyperspectral analysis; seed identification; variable selection; evergreen oak; species classification.

Addresses: [1]Department of Agriculture, Forestry and Bioresources, College of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea | [2]Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

✉ **Corresponding Author:** Kyu-Suk Kang (kangks84@snu.ac.kr).

Manuscript: received January 22, 2025; revised June 15, 2025; accepted June 27, 2025.

Introduction

Hybrids are common in *Quercus* spp. because species in the genus tend to hybridize well with each other (Rushton 1993). This trend is also observed in the evergreen oaks found in southern Korea. Hybridization occurs frequently in oak species such as Japanese red oak (*Quercus acuta* Thunb.) and ring-cup oaks (*Quercus glauca* Thunb.), despite the altitudinal differences in their reproductive zones (Lee et al. 2014). This hybridization can be used to induce hybridization among species in the genus to produce wood with

*This article was presented in the framework of the IUFRO Seed Orchards Conference, Brasov (Romania), 20-24 May 2024

high survivability and high quality in harsh environments caused by climate change (Pelegrín 2017).

It is very important to know whether the result of hybridization is true hybrid seeds or seeds from pollen contamination (Nie et al. 2019, Michelon et al. 2023). Especially in forestry, it is important to know if the hybridization was performed correctly, as pollination and fertilization can be improperly controlled in open field orchards (Ribeiro-Oliveira & Ranal 2014). There are several ways to determine breeding success after hybridization, including examining the plant phenotype, using genetic markers, or employing biotechnological methods. However, using plant phenotypes in practice is challenging due to the tendency for morphological overlap among species within the genus *Quercus* spp. (Valencia 2021). Methods using genetic markers (Matsumoto et al. 2009, Valencia 2021). Methods using genetic markers (Matsumoto et al. 2009) or molecular chemistry can be accurate, but they are costly, labor-intensive, and destructive to seeds (Shrestha & Hardeberg 2015, Boelt et al. 2018).

High-throughput phenotyping using hyperspectral analysis is a fast and non-destructive method for seed quality classification and is also being used for varietal classification at the seed stage (Feng et al. 2019). Varietal classification has been studied for seeds of various tree species, including hybrid seeds between European and Japanese larch (Farhadi et al. 2016), birch species (Tigabu et al. 2018), cultivated species in the genus *Medicago* (Jia et al. 2022), and the imaging classification of eucalyptus seeds (Michelon et al. 2024).

Prior to the hyperspectral classification of oak hybrids, there is a lack of research on the possibility of classifying hybrids between species within the genus using spectroscopic analysis at the seed level. Therefore, this study aims to first assess the classification of proximate evergreen oak species which are potential hybrid breeding materials within the *Quercus* genus in Korea, and to develop a foundational model for spectroscopic techniques to classify hybrid acorns in the future.

Q. acuta and *Q. glauca* are species of evergreen oaks distributed in the southern part of the Korean Peninsula. Initially, their distribution range was limited to Jeju Island and the coastal areas of the island (Lee & Choi 2010). However, due to climate change, their distribution range is expected to expand into parts of the inland Korean Peninsula within the next 50 to 100 years (Yun et al. 2014, Kim et al. 2023). Additionally, *Q. acuta* is expected to have a high carbon uptake capacity, as its BEF does not decrease with increasing age (Kim & Lee 2017). Breeding for cold tolerance is thus required for future inland planting due to unstable winter temperatures and sudden cold waves caused by climate change (Kretschmer et al. 2018).

The aim of this study was to characterize the acorns of *Q. acuta* and *Q. glauca* using hyperspectral analysis. To this end, the following objectives were set: 1) to develop a model that can discriminate between acorns of different species within the evergreen oaks, 2) to identify the key variables influencing species discrimination, and 3) to develop a lightweight model using these main variables and explore ways to improve the performance of the model.

Materials and Methods

Materials

In October and December 2022, 631 acorns of *Q. acuta* and 505 acorns of *Q. glauca* were collected respectively from the seed orchards of the National Forest Seed Variety Center (NFSV) in Jeju Island, Korea. A total of 42 families of *Q. acuta* and 33 families of *Q. glauca* provided acorns, with 10 to 15 seeds per family and 5 seeds per individual tree. The collected seeds were stored in a refrigerator at 5°C for approximately 1-2 months and dried

in an indoor environment at 25°C and 50-70% humidity for approximately 1-2 hours to control moisture and temperature conditions just prior to hyperspectral imaging.

Hyperspectral imaging

Measurements were conducted using two hyperspectral imaging instruments. The sensors used were Corning (400-1000 nm) and Korea Spectral Products (KSP, 400-1700 nm). For the Corning sensor, 150 spectral bands with 4 nm intervals were measured in the visible and near-infrared regions. For KSP, 640 bands with 2 nm intervals were measured in the visible and part of the short wavelength infrared (SWIR) region. During the measurements, a darkroom environment was created by installing a box coated with absorbent paint to prevent external light from entering.

Hyperspectral images were calibrated using white reference (WR) and dark reference (DR) images (Equation 1). To segment the region of interest where the acorns are located in the image, a thresholding strategy and a contouring strategy were used. For thresholding, the image data from the Corning sensor was converted to the difference between the two wavelength variables used in generalized difference vegetation index (GDVI) (Sripada et al., 2006). Pixels with $R_{800-550}$ intensities above 0.9 were initially selected, and the unfiltered areas were removed using the fill hole technique.

$$R_N = \frac{I_N - W}{W - D} \quad (1)$$

(R_N: reflectance image calibrated for the Nth sample, I_N: hyperspectral image taken for the Nth sample, W: white reference, D: dark reference).

$$R_{800-550} = R_{800} - R_{550} \quad (2)$$

(R_{800}: spectral image at 800 nm, R_{550}: spectral image at 550 nm)

For contouring, we used a masking method based on Orth's thresholding for 932 nm grayscale images. The mean spectra, which averages the spectral values of the pixels in the seed region of interest, were extracted to form the Corning and KSP datasets along with the species data. For the region of interest on the Corning image, we extracted 24 additional morphological data from the red, green, and blue (480 nm, 540 nm, and 630 nm) images (Fig. S1). For the KSP dataset, we removed the region around 400-600 nm, where hot pixels and dead pixels occur due to inherent instrumental errors. Hyperspectral images were calibrated and extracted using python and the open source PlantCV (Gehan et al. 2017) packages.

Data analysis

For the extracted data of 1136 samples, we divided the training and validation sets in a 3:1 ratio, with an equal number of *Q. acuta* and *Q. glauca* acorns in each set. The samples in the training set were analyzed using the following process.

First, statistical preprocessing techniques were applied to remove potential errors, such as baseline and scatter effects, that may occur during hyperspectral imaging. We used Standard normal variate (SNV), Multiplicate scatter correction (MSC), and Savitzky-Golay filtering (SGF).

Multivariate analyses were conducted using principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA). PLS has been widely used in fields like chemometrics where there are many variables and covariance problems, such as hyperspectral analysis. There are several model methods for PLS-DA, and in this study, we used PLS2 - Hard PLS-DA technique. The modeling was performed using the open-source PyChemAuth (Mahynski 2023) package.

Finally, the number of variables required in the model was reduced to identify the most important spectral variables and reduce multicollinearity. For variable selection, we used the Successive projection algorithm (SPA) and Competitive adaptive reweighted

sampling (CARS). SPA is a technique used to remove multicollinearity among multiple variables (Zhang et al. 2008). CARS selects the k most important variables with the highest PLS coefficients (Li et al. 2009). The selected significant variables were used to build a lightweight model and test the classification performance. To enhance the performance of the lightweight model for the Corning dataset, an improved model was constructed by combining the important variables with morphological variables to create and use a new dataset. The combined variables were scaled to mitigate the effects of varying variable scales.

For model validation, we used nested K-fold validation. To optimize the combination of PLS principal components and variable selection, which are the hyperparameters of the model in this study, a 5-fold cross validation was conducted to record the accuracy of each combination, and the combination with the highest average accuracy was selected. To evaluate the performance of the optimized model, accuracy was evaluated using 3-fold cross validation (Fig. 1). Finally, the model's generalization performance was validated with the test set.

Results

Mean spectra and PCA plot

Significant differences occurred between the spectral variables of the acorns of *Q. acuta* and *Q. glauca* (Fig. 2). For the Corning spectra, the main differences appeared between 700 and 900 nm. In the case of KSP spectra, significant differences emerged in the near-infrared region, starting from 700 nm to 1500 nm, and the differences between the mean values of the acornsdecreased after 1500 nm (Fig. 2).

Dimensionality reduction using PCA showed that the clustering of the spectra of *Q. acuta* and *Q. glauca* tended to differ. For the Corning dataset, the difference between the clusters of the two species tended to occur in the PC2 direction. The PC2 loading on the Corning PCA plot was mainly influenced by spectral variables in the visible region (400-700 nm) and near-infrared range (900-1000 nm). For the KSP dataset, the clustering also showed differences in the PC2 direction (Fig. 3).

Accuracy according to preprocessing and spectral range

Models based on the Corning dataset distinguished between the two species an accuracy ranging from approximately 88~97.5%. When all spectral variables in the 400-1000 nm range were preprocessed, with MSC preprocessing, a model achieved up to 97.5% accuracy (Table 1).

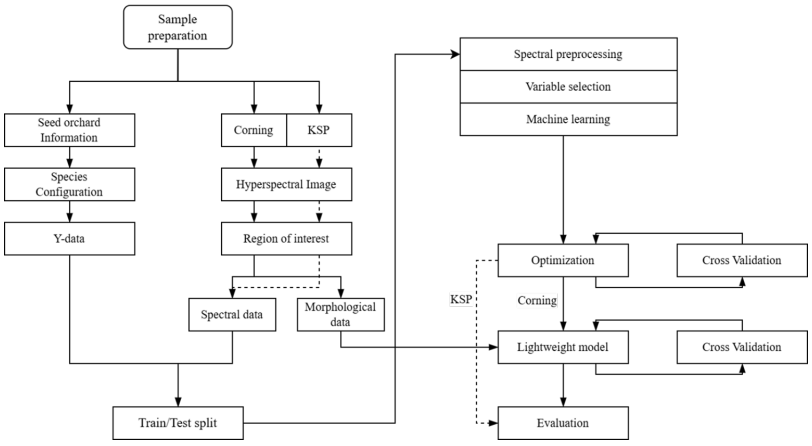


Figure 1 Workflow of spectrometry analysis used in this study.

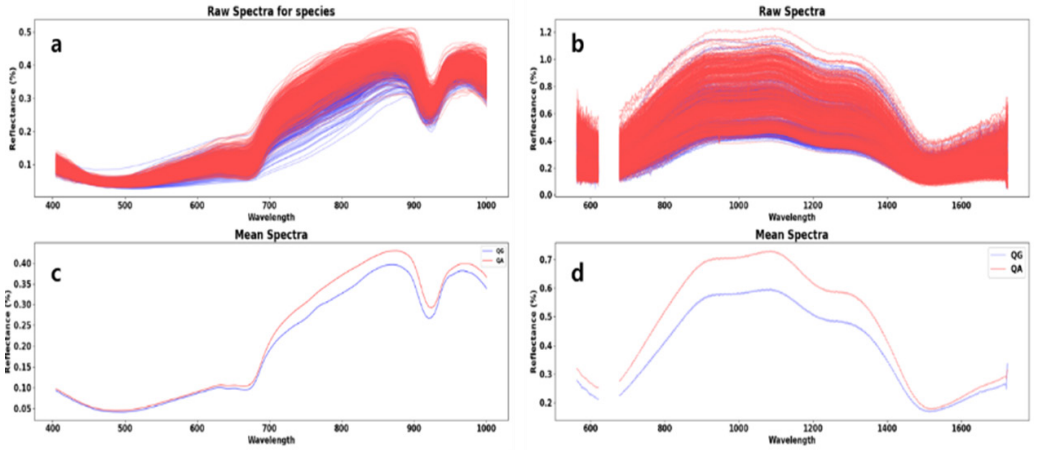


Figure 2 The whole and mean reflectance spectra of *Quercus acuta* and *Quercus glauca* acorns: 400~560, 623~675nm within the KSP dataset were excluded due to their high variances. Blue: Q. glauca; Red: Q. acuta; Corning dataset: a) whole spectra, c) mean spectra and KSP dataset: b) whole spectra, d) mean spectra.

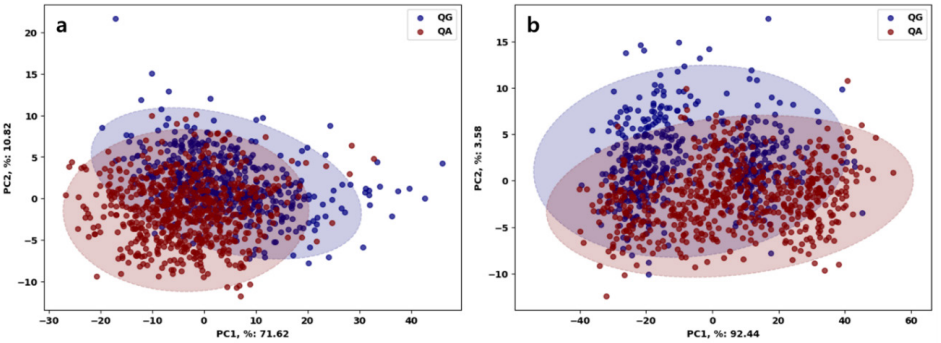


Figure 3 Principal component analysis (2D score plots) for component 1 and 2. (a): Corning dataset, (b) KSP dataset.

Table 1 Species classification accuracy for spectral devices.

Model	Device	Corning dataset								
	Range (4nm)	400-1000			400-780			780-1000		
	Metrics	Accuracy	Acc(QA)	Acc(QG)	Accuracy	Acc(QA)	Acc(QG)	Accuracy	Acc(QA)	Acc(QG)
PLS-DA	Raw	0.94718	0.92405	0.97619	0.95070	0.93040	0.97622	0.91197	0.89241	0.93651
	SNV	0.97183	0.95570	0.99206	0.96831	0.96835	0.96825	0.90490	0.93040	0.87302
	MSC	0.97535	0.96203	0.99206	0.96479	0.95570	0.97619	0.91197	0.93671	0.88095
	SG	0.93310	0.89873	0.97619	0.94014	0.90506	0.98413	0.88732	0.87975	0.89683
	Device	KSP dataset								
	Range (2nm)	400-1700			400-780			780-1700		
	Metrics	Accuracy	Acc(QA)	Acc(QG)	Accuracy	Acc(QA)	Acc(QG)	Accuracy	Acc(QA)	Acc(QG)
PLS-DA	Raw	0.95423	0.96203	0.94444	0.71831	0.72785	0.70635	0.94366	0.93671	0.95238
	SNV	0.96479	0.97468	0.95238	0.65845	0.77848	0.50794	0.96831	0.96835	0.96825
	MSC	0.95423	0.95570	0.95238	0.66197	0.79747	0.49206	0.94718	0.94937	0.94444
	SG	0.96479	0.97468	0.95238	0.73239	0.71520	0.75397	0.96127	0.96203	0.96032

Note: PLS-DA: partial least squared discriminant analysis; Acc: accuracy; QA: *Quercus acuta*; QG: *Quercus glauca*; SNV: standard normal variate; MSC: multiplicative scatter correction; SGF: Savitzuky-Golay Filtering.

Models based on the KSP dataset cshowed an accuracy ranging from approximately 65~96.8%, and when the variables in the 780-1700 nm range were processed with SNV preprocessing, a model achieved up to 96.8% accuracy in classifying the species (Table 1).

When the visible and near-infrared variables were analyzed separately, the Corning dataset had similar or increased accuracy in the visible but showed decreased accuracy in the near-infrared range. The KSP dataset had a significant decrease in accuracy in the visible region, with accuracy below 0.75, and similar or reduced accuracy in the near-infrared range, with results between 94% and 96%. For species-specific classification performance, the Corning dataset tended to discriminate *Q. glauca* better than *Q. acuta*, while the KSP dataset showed the opposite trend (Table 1).

Model performance by variable selection method

Variable selection using SPA showed poor discriminative performance compared to CARS. Reducing variables using SPA yielded low classification accuracy of approximately 45-82% on the Corning dataset and 79-93% on the KSP dataset (Table 2). On the other hand, CARS showed 94-96.8% classification accuracy on the Corning dataset and 89-96% classification accuracy on the KSP dataset. Among preprocessing techniques, SG achieved the highest discrimination accuracy, except when combined with CARS for the Corning dataset (Table 2).

For the model with the highest accuracy, the Corning dataset showed the optimal discrimination with CARS and MSC preprocessing, achieving 96.83% accuracy (Fig. 4). For this model, variables between 550 and 780 nm were selected. On the KSP dataset, the model with CARS and SGF preprocessing showed 96.13% accuracy, with selected variables in the ranges of 820-920 nm, 1050-1300 nm, 1421 nm, and 1675 nm (Fig. 4).

Trends in model performance and morphological data combined the model

Models constructed using the variables selected via SPA did not show a trend of performance fluctuations with changes in the number of selected variables or principal components, except when SGF preprocessing was applied to reduce noise (Fig. 5).

With CARS, models built using variables selected by this algorithm showed similar performance improvements in the Corning and KSP datasets. In particular, except when SNV or MSC preprocessing was applied to the KSP dataset, model performance tended to increase in proportion to the number of variables selected and the number of principal components (Fig. 5).

When analyzing the Corning dataset, combining variables selected by SPA or CARS with morphological variables improved classification accuracy. The SPA-

Table 2 The classification metrics of variable selection methods for SPA and CARS.

		Corning dataset				KSP dataset			
		Raw	SNV	MSC	SG	Raw	SNV	MSC	SG
SPA	components	2	2	5	3	3	8	7	9
	selected	9	5	5	9	13	12	16	18
	train accuracy	0.7236	0.7410	0.7934	0.7570	0.8955	0.8826	0.8720	0.9307
	test accuracy	0.4507	0.4824	0.7254	0.7570	0.8556	0.7993	0.8345	0.9014
	recall	0.4747	0.5127	0.7089	0.7215	0.8354	0.8418	0.8608	0.8861
	precision	0.5068	0.5364	0.7778	0.8201	0.8980	0.8061	0.8447	0.9333
CARS	components	13	13	14	13	12	9	10	12
	selected	16	19	14	19	16	18	19	16
	train accuracy	0.9823	0.9859	0.9835	0.9788	0.9424	0.9284	0.9295	0.9683
	test accuracy	0.9542	0.9613	0.9683	0.9472	0.9296	0.8944	0.9085	0.9613
	recall	0.9367	0.9430	0.9557	0.9114	0.9430	0.9241	0.9367	0.9557
	precision	0.9801	0.9868	0.9869	0.9931	0.9313	0.8902	0.9024	0.9742

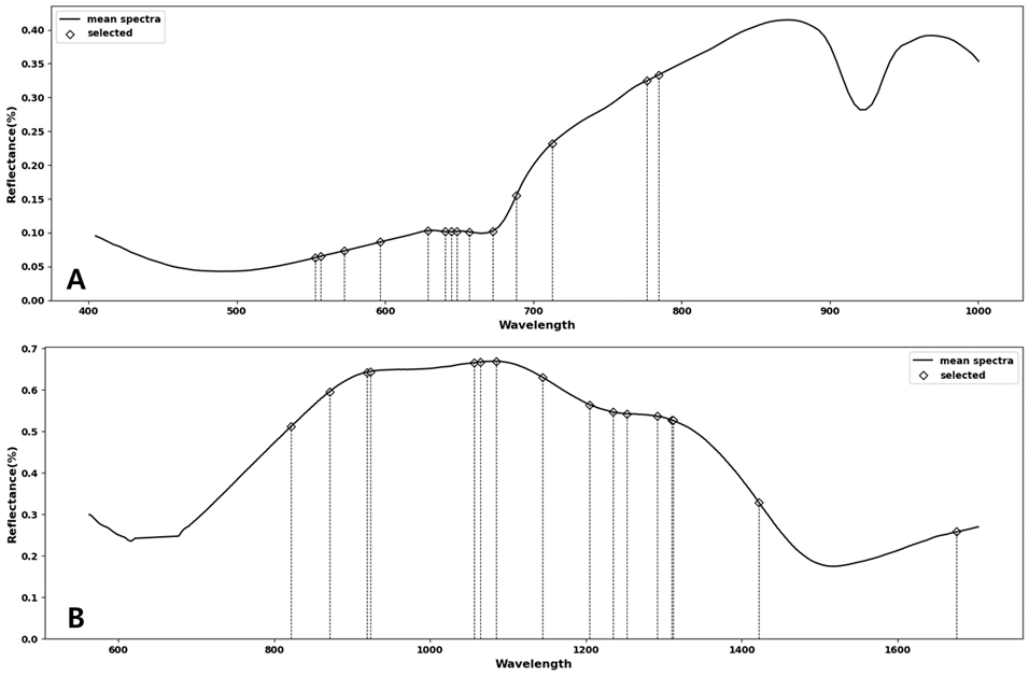


Figure 4 Variables selected. (A) Corning variables selected by CARS-MSC: 526.9, 556.68, 572.67, 596.65, 628.63, 640.62, 644.61, 648.61, 656.6, 672.59, 688.58, 712.56, 776.51, 784.5 nm. (B) KSP variables selected by CARS-SGF: 822, 871.5, 919, 923.4, 1056.2, 1065, 1084.8, 1144.6, 1205, 1234.4, 1252.6, 1291.4, 1309.8, 1312.1, 1421.9, 1675 nm.

RAW model (raw variables selected by SPA without preprocessing) and the CARS-SNV model (variables selected by CARS after SNV preprocessing) achieved 100% accuracy (Table 3). For the CARS-SNV model, 572.67 nm showed an absolute coefficient ≥ 0.3, and 480.74 nm showed an absolute coefficient ≥ 0.2. For the SPA-RAW model, tno variables had coefficients exceeding 0.2, but the coefficient differences among variables were not significant (Fig. 6).

Discussion

The models used in this study effectively classified *Q. acuta* and *Q. glauca* acorns from two hyperspectral instrument datasets by combining preprocessing techniques, multivariate analysis, variable selection, and morphological data. In addition, we extracted the variables with significant influence from the model to create a lightweight model and explored the utilization of acorn morphological data to design an effective method for future classification studies of oak hybrids.

Table 3 Comparison of accuracy between models built with selected variables from corning dataset and morphological variables.

		Selected variables and morphological variables			
		Raw	SNV	MSC	SGF
SPA	Selected	9	5	5	9
	Whole	33	29	29	33
	Components	13	13	13	13
	Accuracy	1.000	0.993	0.98	0.993
CARS	Selected	16	19	14	19
	Whole	40	43	38	43
	Components	13	10	12	10
	Accuracy	0.993	1.000	0.99	0.997

Note: SNV: standard normal variate, MSC: multiplicate scatter correction and SGF: Savitzky-Golay filtering

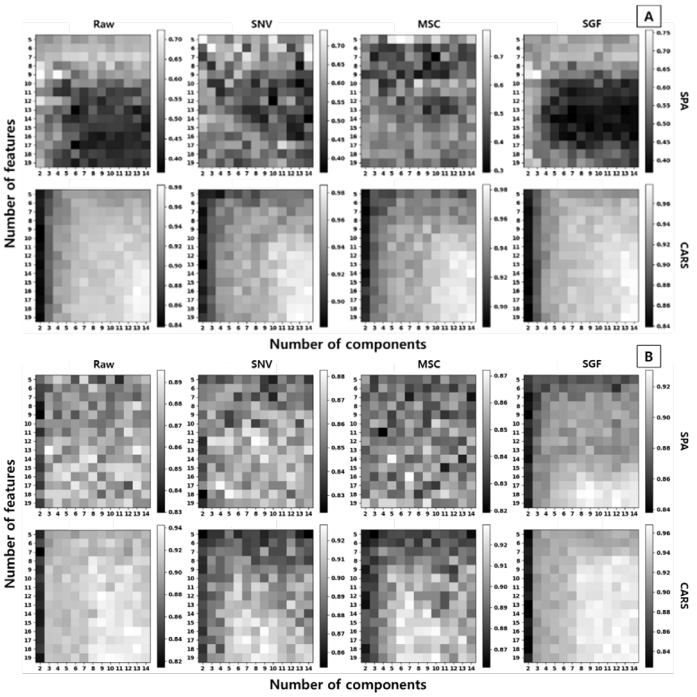
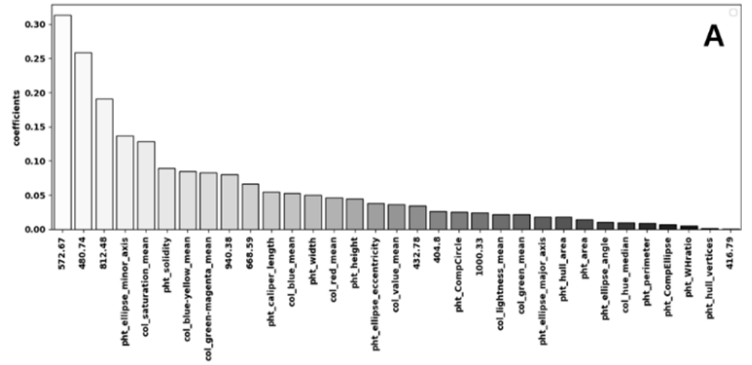


Figure 5 Heatmaps of classification accuracies for different variable selection methods are presented for (A) the Corning dataset and (B) the KSP dataset. Spectral preprocessing) Raw: preprocessing was not performed, SNV: standard normal variate, MSC: multiplicative scatter correction, SGF: Savitzuky Golay filtering; Variable selection model) SPA: successive projections algorithm, CARS: competitive adaptive reweighted sampling; Hyperparameters) Number of components: principal components used in the partial least square discriminant analysis model; Number of features: features selected in the variable selection algorithm.



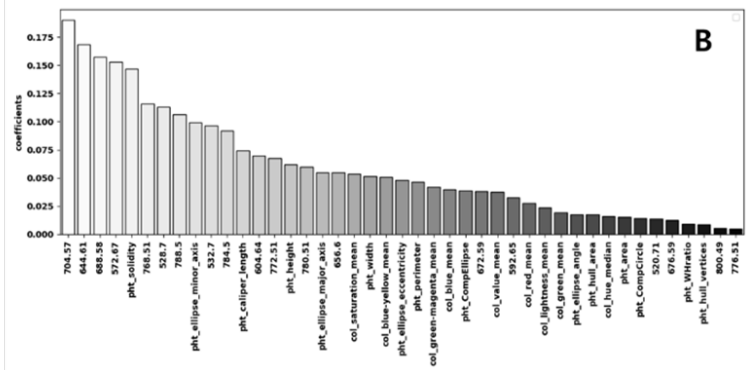


Figure 6 Absolute coefficients of the classification models from selected variables and morphological variables. A) SPA-RAW selected variables; B) CARS-SNV selected variables.

In the visible range, the blue and red-light regions were thought to have a significant effect on species discrimination. Discrimination using only the 400-780 nm range of the Corning dataset showed an accuracy above 95%, while the discrimination using 400-780 nm range of the KSP dataset with the red and blue light regions removed was below 70% (Table 1). Also, the key variables selected by CARS were clustered around 550-650 nm in the Corning dataset, which corresponds to the red-light region (Fig. 4). This aligns with studies in birch, where the absorption points of 400-750 nm played a major role in distinguishing between species (Tigabu et al. 2018), and in maize, where differences between varieties occurred in the 500-700nm range (Huang et al. 2016), and in alfalfa, where variables in the 400-500 nm range affected discrimination between species (Yang et al. 2020). In the red-light region, color variations due to pigments such as tannins and anthocyanins are known to influence species discrimination (Mortensen et al. 2021, Michelon et al. 2024).

This suggests that the acorn shell color, especially in the red and blue regions, plays a key role in distinguishing the two species. This is because the color of the seed coat can be used as a phenotypic marker for a tree species, as it carries the genetic information of the mother trees (Bacherikov et al. 2022). However, since chlorophyll levels decrease during the maturation process of acorns (Bonner & Vozzo 1987), it should be considered that color changes caused from other substances in the acorn shell besides chlorophyll may have influenced the species classification.

In the near-infrared range, the key influencing variables seem to be around 820-925 nm, 1050-1312 nm, 1422 nm, and 1675 nm. The 820~925nm is the range where starch is expected to have a major impact on varietal classification (Yang et al. 2015). The spectral variables at 1205, 1309.8, and 1312.1 nm are considered to be valid for variety classification due to the second overtone of the C-H stretching vibration (Feng et al. 2017). 919 and 923.4 nm are influenced by the third overtone of the C-H2 stretching vibration. 1144, 1205, and 1421.9 nm are influenced by the second overtone of the C-H2 stretching vibration, and 1675 nm is influenced by the first overtone of the C-H2 vibration (Metrohm NIRSystems, 2013). Since starch is a major component in acorns containing C-H and C-H2 groups, it is believed that starch is responsible for the distinction between *Q. acuta* and *Q. glauca* acorns. The quantity, composition, and structural characteristics of starch have been shown to play a key role in variety classification (Jeong et al. 2010, Valková et al. 2019). With advances in spectral analysis of starch (Yang et al. 2015, Seo et al. 2020, Wang et al. 2023).

In terms of variable selection methods, this study employed CARS and SPA, as these algorithms are commonly used in hyperspectral analyses for classification tasks such as crop-variety discrimination or seed-quality assessment. For instance, Zhang et al. (2018) applied CARS and SPA to select optimal wavelengths for distinguishing okra varieties, and Pang et al. (2021) used both algorithms to reduce the feature set to 25 bands while preserving high classification performance for Quercus seed viability. In our results, CARS achieved a minimum classification accuracy of 89%, confirming its efficiency for inter-species discrimination using only spectral variables. To further optimize CARS, various spectral preprocessing techniques were evaluated; within a certain range, increasing the number of PLS components in CARS improved the final model's performance, thereby enhancing overall algorithm efficiency (Sun et al. 2021, Dilillo et al. 2025).

Notably, when SNV and MSC preprocessing were applied, there was a range in which accuracy plateaued and then declined as the number of PLS components increased (Figure 5 Dataset B with CARS-SNV). This suggests that SNV and MSC effectively removed spectral noise, simplified the dataset's complexity, and appropriately narrowed the hyperparameter

search space (Gariso et al. 2025). By contrast, SPA-based models exhibited lower accuracy relative to CARS and, except for the KSP dataset with SGF preprocessing, did not show a trend of improved performance with increasing hyperparameters. This indicates that, for SPA, optimizing solely based on the number of selected variables—as in prior studies—may be more appropriate.

Combining variable selection with morphological data seems to be very effective for acorn classification. In particular, the accuracy of the variable selection model extracted using SPA was 70%, but the classification performance improved to 100% when morphological data was added. This suggests that there was an interaction between the variables representing the color of oak acorns and the morphological variables. This result is similar to other studies where adding morphological data increased model performance when using only multispectral variables alone yielded low classification accuracy. But the difference was that morphological data was ranked higher in the importance ranking of the variables in this result (Jia et al. 2022, Fu et al. 2024).

The reason for the difference in the importance ranking of morphological data is that there was a significant difference in the mean for morphological variables with high importance (Fig. S1). Therefore, further research on the importance of morphological data should be conducted to determine whether the same pattern occurs when models are built for more complex problems, such as the identification of oak hybrids. In addition, we believe that the model, important variables, and methods developed in this study for discriminating between acorns of oak relatives using morphological data, can serve as a foundation for more complex and sophisticated models for future successful breeding programs.

Conclusions

This study developed a hyperspectral-based discrimination model for the acorns of two closely related oak species, *Quercus acuta* and *Quercus glauca*, and subsequently improved a lightweight model by incorporating important spectral and morphological variables. Specifically, CARS and SPA were optimized to determine the optimal combination of preprocessing techniques, the number of PLS components, and the number of selected variables. Using these methods, we demonstrated that wavelengths in the blue and red regions of the visible spectrum, as well as features responsive to the C–H and C–H_2 bonds of starch—the primary constituent of acorns—play a significant role in species discrimination.

A lightweight model built with the selected spectral variables achieved moderate classification accuracy, whereas an enhanced model that combined spectral and morphological data classified nearly all acorns with 98–100% accuracy. This improvement likely reduced multicollinearity among hyperspectral variables and enhanced generalization by leveraging variables with complementary properties. However, as seed samples were exclusively collected from Jeju Island, further validation is required to assess the model's generalizability to acorns from other provenances or collection periods. Overall, these findings provide a valuable foundation for designing more complex models aimed at identifying oak hybrid acorns in future studies.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This study was carried out with the support of ′R&D Program for Forest Science Technology (Project No. RS-2024-00404133)‘ provided by Korea Forest Service (Korea Forestry Promotion Institute).

References

Bacherikov I.V., Raupova D.E., Durova A.S., Bragin V.D., Petrishchev E.P., Novikov A.I., Danilov D.A. and

Zhigunov A.V. 2022. Coat colour grading of the scots pine seeds collected from faraway provenances reveals a different germination effect. Seeds 1: 49-73. https://doi.org/10.3390/seeds1010006

Boelt B., Shrestha S., Salimi Z., Jørgensen J.R., Nicolaisen M. and Carstensen J.M. 2018. Multispectral imaging – a new tool in seed quality assessment? Seed Science Research 28: 222-228. https://doi.org/10.1017/S0960258518000235

Bonner F.T. and Vozzo J.A. 1987. Seed biology and technology of *Quercus*. U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station.

Dilillo N., Sanna A., Belcore E., Smith K., Piras M., Montrucchio B., Ferrero R., 2025. Enhancing lettuce classification: Optimizing spectral wavelength selection via CCARS and PLS-DA. Smart Agricultural Technology, 11, 100962. https://doi.org/10.1016/j.atech.2025.100962

Farhadi M., Tigabu M., Stener L.-G. and Odén P.C. 2016. Feasibility of visible + near infrared spectroscopy for non-destructive verification of European × Japanese larch hybrid seeds. New Forests 47: 271-285. https://doi.org/10.1007/s11056-015-9514-4

Feng X., Peng C., Chen Y., Liu X., Feng X. and He Y. 2017. Discrimination of CRISPR/Cas9-induced mutants of rice seeds using near-infrared hyperspectral imaging. Scientific Reports 7: 15934-10. https://doi.org/10.1038/s41598-017-16254-z

Feng L., Zhu S., Liu F., He Y., Bao Y. and Zhang C. 2019. Hyperspectral imaging for seed quality and safety inspection: A review. Plant Methods 15: 91. https://doi.org/10.1186/s13007-019-0476-y

Fu X., Bai M., Xu Y., Wang T., Hui Z., Hu X., 2023. Cultivars identification of oat (*Avena sativa* L.) seed via multispectral imaging analysis. Frontiers in Plant Science, 14, 1113535–1113535. https://doi.org/10.3389/fpls.2023.1113535

Gariso R., Coutinho J. P. L., Rato T. J., Reis M. S., 2025. A comparative analysis of deep learning and chemometric approaches for spectral data modeling. Analytica Chimica Acta, 1347, 343766. https://doi.org/10.1016/j.aca.2025.343766

Gehan M.A., Fahlgren N., Abbasi A., Berry J.C., Callen S.T., Chavez L., Doust A.N., Feldman M.J., Gilbert K.B., Hodge J.G., Hoyer J.S., Lin A., Liu S., Lizárraga C., Lorence A., Miller M., Platon E., Tessman M. and Sax T. 2017. PlantCV v2: Image analysis software for high-throughput plant phenotyping. PeerJ 2017: e4088-e4088. https://doi.org/10.7717/peerj.4088

Gil-Pelegrín E., Peguero-Pina J., Sancho-Knapik D. (eds) 2017. Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L. Tree Physiology, vol 7. Springer, Cham. https://doi.org/10.1007/978-3-319-69099-5_5

Huang M., He C., Zhu Q. and Qin J. 2016. Maize seed variety classification using the integration of spectral and image features combined with feature transformation based on hyperspectral imaging. Applied Sciences 6: 183. https://doi.org/10.3390/app6060183

Jeong W. H., Harada, K., Yamada T., Abe J., Kitamura K., 2010. Establishment of new method for analysis of starch contents and varietal differences in soybean seeds. Breeding Science, 60(2), 160–163. https://doi.org/10.1270/jsbbs.60.160.

Jia Z., Sun M., Ou C., Sun S., Mao C., Hong L., Wang J., Li M., Jia S. and Mao P. 2022. Single seed identification in three medicago species via multispectral imaging combined with stacking ensemble learning. Sensors 22: 7521. https://doi.org/10.3390/s22197521

Jinnuo Z., Xuping F., Xiaodan L., Yong H., 2018. Identification of hybrid okra seeds based on near-infrared hyperspectral imaging technology. Applied Sciences, 8(10), 1793. https://doi.org/10.3390/app8101793

Kim C.Y., Kim W.M., Song W.K., Cho J.P. and Choi J.Y. 2023. Prediction of native seed habitat distribution according to SSP scenario and seed transfer zones: A focus on *Acer pictum* subsp. *Mono* and *Quercus acuta*. Forests 14: 87. https://doi.org/10.3390/f14010087

Kim H.J. and Lee S.H. 2017. Estimating carbon storage and CO_2 absorption by developing allometric equations for *Quercus acuta* in South Korea. Forest Science and Technology 13: 55-60. https://doi.org/10.1080/21580103.2017.1308888

Kretschmer M., Coumou D., Agel L., Barlow M., Tziperman E. and Cohen J. 2018. More-persistent weak stratospheric polar vortex states linked to cold extremes. Bulletin of the American Meteorological Society 99: 49-60. https://doi.org/10.1175/bams-d-16-0259.1

Lee J.H. and Choi B.H. 2010. Distribution and northernmost limit on the Korean peninsula of three evergreen trees. Korean Journal of Plant Taxonomy 40: 267-273. https://doi.org/10.11110/kjpt.2010.40.4.267

Lee J.H., Jin D.P. and Choi B.H. 2014. Genetic differentiation and introgression among korean evergreen *Quercus* (*Fagaceae*) are revealed by microsatellite markers. Annales Botanici Fennici 51: 39-48. https://doi.org/10.5735/085.051.0105

Li H., Liang Y., Xu Q. and Cao D. 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Analytica Chimica Acta 648: 77-84. https://doi.org/10.1016/j.aca.2009.06.046

Mahynski N.A. 2023. PyChemAuth (v0.0.0-beta3). Zenodo. https://doi.org/10.5281/zenodo.10037568

Matsumoto A., Kawahara T., Kanazashi A., Yoshimaru H., Takahashi M. and Tsumura Y. 2009. Differentiation of three closely related Japanese oak species and detection of interspecific hybrids using AFLP markers. Botany 87: 145-153. https://doi.org/10.1139/b08-121

Metrohm NIRSystems 2013. A guide to near-infrared spectroscopic analysis of industrial manufacturing processes**.**

Michelon T.B., Serra Negra Vieira E. and Panobianco M. 2023. Spectral imaging and chemometrics applied at phenotyping in seed science studies: a systematic review. Seed Science Research 33: 9-22. https://doi.org/10.1017/S0960258523000028.

Michelon T.B., Carstensen J.M., Serra Negra Vieira E.

and Panobianco M. 2024. Multispectral imaging for distinguishing hybrid forest seeds of *Corymbia spp.* and *Eucalyptus spp.* from their progenitors. Journal of Environmental Management 363: 121383-121383. https://doi.org/10.1016/j.jenvman.2024.121383

Mortensen A.K., Gislum R., Jørgensen J.R. and Boelt B. 2021. The use of multispectral imaging and single seed and bulk near-infrared spectroscopy to characterize seed covering structures: Methods and applications in seed testing and research. Agriculture 11: 301. https://doi.org/10.3390/agriculture11040301

Nie P., Zhang J., Feng X., Yu C. and He Y. 2019. Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. Sensors and Actuators. B, Chemical 296: 126630. https://doi.org/10.1016/j.snb.2019.126630

Pang L., Wang J., Men S., Yan L. and Xiao J. 2021. Hyperspectral imaging coupled with multivariate methods for seed vitality estimation and forecast for *Quercus* variabilis. Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy 245: 118888. https://doi.org/10.1016/j.saa.2020.118888

Ribeiro-Oliveira J.P. and Ranal M.A. 2014. Sementes florestais brasileiras: Inicio precário, presente inebriante e o futuro, promissor? Ciência Florestal 24: 771-784. https://doi.org/10.5902/1980509815738

Rushton B. 1993. Natural hybridization within the genus Quercus L. Annales des Sciences Forestières 50: 73s-90s

Seo, Y., Lee, A., Kim, B., Lim, J., 2020. Classification of rice and starch flours by using multiple hyperspectral imaging systems and chemometric methods. Applied Sciences, 10(19), 6724. https://doi.org/10.3390/app1019672

Shrestha R. and Hardeberg J.Y. 2015. An experimental study of fast multispectral imaging using LED illumination and an RGB camera. Color and Imaging Conference, Society for Imaging Science and Technology 23: 36-40. https://doi.org/10.2352/CIC.2015.23.1.art00008

Sripada R.P., Heiniger R.W., White J.G. and Meijer A.D. 2006. Aerial color infrared photography for determining early in-season nitrogen requirements in corn. Agronomy Journal 98: 968-977. https://doi.org/10.2134/agronj2005.0200

Tigabu M., Farhadi M., Stener L.G. and Odén P.C. 2018. Visible + Near Infrared Spectroscopy as taxonomic tool for identifying birch species. Silva Fennica 52: 1. https://doi.org/10.14214/sf.9996

Valencia S.A. 2021. Species delimitation in the genus *Quercus* (*Fagaceae*). Botanical Sciences 99: 1-12. https://doi.org/10.17129/BOTSCI.2658

Valkov V. 2019. The content and quality of starch in different wheat varieties growing in experimental conditions. Journal of Microbiology, Biotechnology and Food Sciences, 9(Special), 462–466. https://doi.org/10.15414/jmbfs.2019.9.special.462-466

Wang T., Xu L., Lan T., Deng Z., Yun Y. H., Zhai C., Qian C., 2024. Nondestructive identification and classification of starch types based on multispectral techniques coupled with chemometrics. Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 311, 123976. https://doi.org/10.1016/j.saa.2024.123976

Yang L., Zhang Z. and Hu X. 2020. Cultivar discrimination of single alfalfa (*Medicago sativa* L.) seed via multispectral imaging combined with multivariate analysis. Sensors 20: 1-14. https://doi.org/10.3390/s20226575

Yang X., Hong H., You Z. and Cheng F. 2015. Spectral and image integrated analysis of hyperspectral data for waxy corn seed variety classification. Sensors 15: 15578-15594. https://doi.org/10.3390/s150715578

Yun J.-H., Nakao K., Tsuyama I., Higa M., Matsui T., Park C.-H., Lee B.-Y. and Tanaka N. 2014. Does future climate change facilitate expansion of evergreen broad-leaved tree species in the human-disturbed landscape of the Korean Peninsula? Implication for monitoring design of the impact assessment. Journal of Forest Research 19: 174-183. https://doi.org/10.1007/s10310-013-0401-6

Zhang J., Rivard B. and Rogge D.M. 2008. The successive projection algorithm (SPA), an algorithm with a spatial constraint for the automatic search of endmembers in hyperspectral data. Sensors 8: 1321-1342. https://doi.org/10.3390/s8021321

Zongbao S., Junkui L., Jianfeng W., Xiaobo Z., Chi T.H., Liming L., Xiaojing Y., Xuan Z. 2021. Rapid qualitative and quantitative analysis of strong aroma base liquor based on SPME-MS combined with chemometrics. Food Science and Human Wellness, 10(3), 362–369. https://doi.org/10.1016/j.fshw.2021.02.031