Accuracy Evaluation of Visible-Near Infrared Spectroscopy for Detecting Insect Damage in Acorns of *Quercus acuta*

Gye-Hong Cho[1], Ye-Ji Kim[1], Koeun Jeon[1], Hye-Jun Joo[2] and Kyu-Suk Kang[1,3,*]

[1] Department of Agriculture, Forestry and Bioresources, Seoul National University, Seoul 08826, Korea
[2] Department of Seed and Seedling Management, National Forest Seed Variety Center, Chungju 27495, Korea
[3] Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Korea

* Corresponding author: Kyu-Suk Kang, E-mail: kangks84@snu.ac.kr

Abstract

This study aimed to test near-infrared spectroscopy to assess insect damage to oak acorns collected from a seed orchard of *Quercus acuta* in Jeju Island, Korea. A total of 550 acorns were sorted into 362 sound and 178 unsound (insect-damaged) acorns, followed by near-infrared spectroscopy. To minimize spectral data errors, preprocessing techniques such as first derivative, multiplicative scatter correction, standard normal variate, and Savitzky-Golay filter were applied, along with multivariate analysis methods like partial least squares. Then the model performance, including accuracy and precision, was evaluated using the Variable Importance in Projection. The near-infrared wavelength of the acorns showed strong absorption peaks at 660~720nm and a slight downward trend at 900~1000nm. The most effective model for distinguishing unsound acorns was Savitzky-Golay filtering treatment applied in the 400~1000nm range and used partial least squares, showing prediction accuracy of 86 % ($p<0.05$). The performance was significantly influenced by absorption points at 660~720nm and 960~1000nm, with the latter range believed to be affected by changes in moisture and carbohydrates due to insect damage. The former range showed lower classification capability due to chlorophyll and color variation but affected the model performance when used with near-infrared wavelength range. These findings can narrow down the scope of investigation for future research using wider wavelength ranges or multispectral analysis.

Keywords: *Quercus acuta; insect damage; near-infrared spectroscopy; multivariate analysis; nondestructive testing*

Introduction

Quercus acuta Thunb is recognized as a promising carbon sink species to deal with climate change due to its high carbon absorption capacity and the shifts in growth distribution areas resulting from climate change. Studies on carbon absorption capacity have shown that *Q. acuta*. is superior compared to the dominant tree species in Korea, such as pine trees (*Pinus densiflora*) and other tree species including evergreen broad-leaved tree species (Son et al. 2014, Hwang et al. 2015, Kim and Lee 2017). Currently, *Q. acuta*. is mainly distributed in piedmont and valley regions at elevations between 170~500 meters in southern part of Korea, including Jeju Island and Wando (Lee and Choi 2010). However, its natural growth areas are expected to expand substantially across the Korean Peninsula with the progression of climate change (Kim et al. 2023). As a result, *Q. acuta*. is one of the subtropical evergreen broad-leaved tree species that is expected to in forestation demand according to the Korean government's "2050 carbon neutrality scenarios" (Government of Korea 2021). To meet the future demand for afforestation, "The development and management plan for 2nd generation seed orchards" proposes to expand the area of existing orchards of evergreen *Quercus* species by 2.5 times (NFSV 2021).

In the mass production of acorns, the quality of the acorns is important, directly influencing the cost-effectiveness of establishing nurseries, the quality of seedlings, and the success rate of afforestation (Novikov and Novikova 2018). High-quality acorns are also beneficial for the rapid growth and good quality of trees under the changing conditions brought by climate change. Factors such as germination rate, viability, and insect-damage are crucial components of acorn quality (FAO 2010).

DOI:10.2478/sg-2024-0010
edited by the Thünen Institute of Forest Genetics

A significant issue of *Q. acuta* trees is the high rate of insect damage to acorns. The biennial acorn masting characteristic of *Q. acuta* trees makes them particularly vulnerable to insect damage during years of lower acorn yield (Hirayama et al. 2012). In the southern regions of the Korean peninsula, over half of the acorns matured between September and October are harmed by pests, highlighting the need for a thorough differentiation between healthy and insect-damaged acorns (Kim and Park 2021).

Traditionally, the floating method has been used to sort acorns based on insect damage, which involves floating the acorns on water and discarding those that float due to their lower density, just before sowing (Gribko and Jones 1995). However, this method alone is insufficient to eliminate all insect-damaged acorns. Furthermore, insect damaging can occur during the acorn storage period (Connor 2004). Other acorn testing methods such as the tetrazolium test, winding method, and cutting inspection are although effective, but are challenging to implement in mechanized production processes (Novikov et al. 2020). Conversely, acorn testing using near-infrared spectroscopy has been in the limelight as a cost-effective, non-destructive method (Novikov 2019). It is particularly useful for simultaneously testing multiple traits (Shenk and Westerhaus 1993).

Non-destructive testing with near-infrared (NIR) spectroscopy has been used to screen seeds for a broader range of qualities. The presence of insect damage affecting seed and fruit quality or vigor has also been studied for nearly two decades across a wide range of tree species: *Cordia africana* fruits (Tigabu and Odén 2002), *Juniperus procera* seeds (Tigabu et al. 2007), jujubes (Wang J et al. 2010), chestnuts (Moscetti et al. 2014), Chinese fir seeds (Tigabu et al. 2019), pomegranate (Jamshidi 2020), apples (Zhang et al. 2022). However, few studies have been conducted on the seeds of evergreen broadleaf tree species.

The main objective of this study is to investigate the feasibility of using near-infrared spectroscopic analysis to assess insect damage in *Quercus acuta* acorns without depreciating acorn quality. We try to develop an accurate classification model based on this approach. The specific objectives are: 1) to verify the possibility of insect-damage testing and establishment of classification models, using spectral wavelength of 400~1000nm and acorn quality classification data, 2) to compare models to determine the classification accuracy based on the combination of data preprocessing and multivariate analysis techniques, and 3) to identify the wavelength variables in the near-infrared range that are significant for classification of insect-damaged acorns.

Materials and methods

Material

For the experiment, materials were obtained from 38 families in the seed orchard of *Q. acuta* (San 1-5, Sanghyo-dong, Seogwipo-si, Jeju-do, Republic of Korea), managed by the National Forest Seed Variety Center (NFSV). Fruiting individuals were chosen from each family, and a total of 550 acorns were collected in October 2021. When collecting acorns, we evenly harvested from four directions of the trees, considering. However, as the yield of the *Q. acuta* seed orchard was poor in the year of 2021, it was challenging to collect acorns uniformly from each family. Consequently, the maximum number of acorns collected from family was 18, the minimum was 1, and the average number of acorns per family was 14.47.

Sorting and pretreatment

We stored the collected acorns in a refrigerated environment at 5°C for about 4 months to decompose chlorophyll (Bonner 1987), before undergoing spectroscopic analysis. Afterwards, insect-damaged acorns were sorted based on visual investigation. For insect damage, we referred to prior research on damage caused by acorn predators in the classification (Dorsey et al. 1962, Gribko 1995). If the presence of acorn weevil larvae was confirmed (Figure 1-a), or if oviposition holes (Figure 1-b) and escape holes (Figure 1-c) were observed, the acorns were classified as insect-damaged. As a result of the visual inspection, a total of 368 acorns were classified as sound, and 182 acorns were classified as insect damaged.

Pretreatment was conducted to eliminate factors interfering near-infrared-spectral analysis, such as moisture and temperature. Regarding the temperature, the acorns, wrapped in plastic bags, were immersed in a water bath at 25°C for approximately 10 minutes, in order to normalize the temperature of the acorns to a level similar to room temperature (Takahashi et al. 2011). The immersion time was determined to minimize the effect of temperature controlling on the acorns, which are recalcitrant seed prone to desiccation and water percentage change (Xia et al. 2012). Subsequently, the acorns were dried with a tissue to remove any potential moisture on their surface. This temperature treatment time was determined based on the consideration that the deterioration rate of starch inside the acorn is affected by temperature changes for at least 24 hours to several days, so 10 minutes of room temperature treatment was judged not to affect the germination rate and acorn quality (Khan and Habib 2015).

Near-infrared wavelength measurement

The near-infrared spectra were measured using the spectrometer, Flame S-Vis-Nir (400-1000nm, Oceanview). An integrating sphere for reflectance (ISP-REF, Oceanoptics) was used to provide the light source. The resolution, calculated using Ocean Insight's resolution calculator, was approximately 1.39 cm-1 cm. A Spectralon reflectance material (WS-1SL, Oceanoptics) was used as a white standard material.



Figure 1
Caterpillars of *Curculio dentipes* (a), acorn with a small hole (b) and acorn with an exit hole (c)

Reflectance was measured through the spectrometer and the integrating sphere (Equation 1). The reflectance (Rn) at the nth wavelength was calculated as the ratio of the reflectance of the sample (Rs) minus the background reflection measured without the light source (Rback), to the reflectance of the standard material (Rref) minus the Rback. Since reflectance is a relative figure, it was converted into absorbance (An) spectrum (Equation 2).

$$R_n = \frac{R_s - R_{back}}{R_{ref} - R_{back}} \times 100(\%) \quad (1)$$

$$A_n = \log_e(1/R_n) \quad (2)$$

The digital data collected through the spectrophotometer were processed with Oceanview spectroscopy software 2.0.8 (Ocean Insight) to calculate the reflectance. Then, the original spectral variables were interpolated into 601 wavelength variables at 1nm intervals within the 400~1000nm range. The samples were positioned directly above an entrance hole of 1cm diameter on the integrating sphere, allowing for the measurement of a single aspect of the acorn (Figure 2).

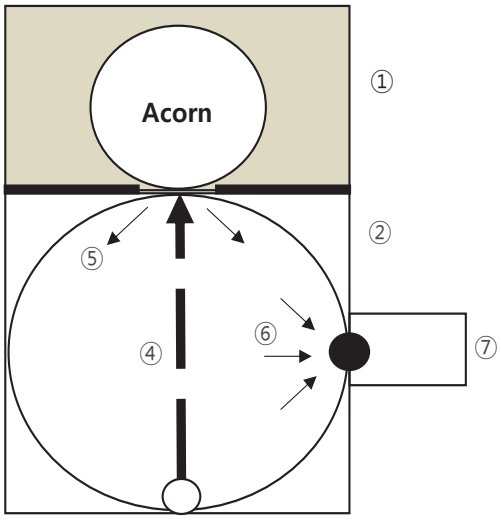

Figure 2
Acorn spectroscopy using integrating sphere. ① Lip, ② Integrating sphere, ③ Light source, ④ Light from the source, ⑤ Sample reflected light, ⑥ Reflected light processed by scattered light exception within the sphere, and ⑦ Optical fiber moves the reflected light from the sphere to a detector

Aa black box-shaped barrier (Figure 2, Lid) was made and used to prevent interference from ambient lights during the measurement of near-infrared spectra (Daneshvar et al. 2015). The box was painted with light absorbing colors to enable the interference of ambient light. Also, considering the possibility of instrumental drift (i.e., a measurement error occurring after several measurements), the light and darkness references were remeasured every five samples (Farhadi et al. 2016). The integrating sphere was operated for more than 30 minutes before starting the measurements to account for errors caused by light saturation in the detector (Moscetti et al. 2014). Lastly, the reflectance was captured 10 times at every 90 degrees of an acorn, totaling 40 measurements, and the averaged value of these measurements was used as the spectral data.

Data preprocessing and analysis

Prior to data analysis, the samples were divided into a calibration set and a validation set for cross-validation. The sizes of each set were designated as 366 acorns and 184 acorns respectively, and the distribution of samples within each set was randomized using built-in functions in R. Ultimately, the calibration set consisted of 246 sound acorns and 120 insect-damaged acorns, while the validation set consisted of 123 sound acorns and 61 insect-damaged acorns.

For preprocessing methods, mean-centering was applied across all samples, followed by statistical preprocessing and wavelength range selection. Several statistical preprocessing techniques were used to eliminate errors that occur when measuring the near-infrared spectra. Standard normal variate (SNV) is a normalizing technique which make mean and standard deviation set at 0, 1, respectively. Multiplicative scatter correction (MSC) is a technique that eliminate the additive effect of interactions between wavelength variables. Both of preprocessing techniques help to decrease light scattering effect (Næs et al. 2002). First derivative (FD) is differential preprocessing which removes base line and linear trend error of spectra. Savitzky-Golay filter (SGF) is a smoothing technique that decreases the noise within the spectra (Næs et al. 2002; Rinnan et al, 2009). For the selection of wavelength range, the range was divided into the full (400~1000nm), the visible light (400~780nm), and the near-infrared (780~1000nm).

To remove the overlap between absorption peaks, multivariate analysis techniques were used to eliminate severe

correlations between wavelength variables (Chung and Kim 2000). The multivariate analysis techniques utilized were principal component analysis (PCA), partial least squares (PLS), and orthogonal partial least squares (OPLS). The optimal number of principal components for use in the multivariate analysis was determined using K-fold cross-validation. The K-value was set to 7, and the mean squared error (MSE) was used as the evaluation criterion. Subsequently, regression equations were formulated from the principal components generated by the multivariate analysis techniques to derive predicted values (y-predicted). A threshold of 0.5 was used to discriminate the soundness of the acorns (discriminant analysis, DA). The analysis procedures were conducted using Simca-P 14.1 (Umetrics, Sweden).

Model performance evaluation

The performance analysis of the models was conducted using classification evaluation index. The basic metrics for classification included TP (true positive), TN (true negative), FP (false positive), and FN (false negative). The TP refers to the number of cases where observations classified as 1 or true by the model are indeed true. the TN stands for the number of cases where the model correctly classifies the actual negative observations as negative. The FP indicates the number of cases where the model classifies the observations as positive, but they are actually negative. Conversely, the FN denotes the number of cases where the model classifies the observations as negative, but they are actually positive. Based on these four metrics, various classification indexes were developed to evaluate the performance of the classification model (Table 1). Subsequently, the variable importance in projections (VIP) was used to determine which wavelength variables had the most impact on each model (Equation 3).

$$VIP_k = \sqrt{\frac{\sum_{f=1}^{F} w_{kf}^2 SSY_f}{SSY_{total} * F}} * K \qquad (3)$$

VIP_k: VIP of the kth variable X, w_{kf}: PLS weight of the k-th variable and f-th principal compenent,
SSY_k: explainable variance of Y within the fth principal component, SSY_{total}: explainable variance of Y within all principal component, F: the number of principal components.

Table 1
Evaluation of accuracy and precision

Metrics	Equation
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Recall	(TP)/(TP+FN)
Precision	(TP)/(TP+FP)
F1 Score	(2*Precision*Recall)/(Precision+Recall)
G mean	$\sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}}$
TP	Actual positive-pred postive
FP	Actual negative-pred positive
FN	Actual positive-pred negative
TN	Actual negative-pred negative

Results

Spectral figures of the samples

The mean of absorbance was obtained by averaging the 40 measurements (550 spectra) from each acorn in all four directions (Figure 3). High absorption rate was observed in the visible light range of 400~780nm, while a lower absorption rate was noted in the near-infrared range of 780~1000nm. A distinct absorption peak was observed around 680nm, and a looser absorption peak was noted at 938nm. Regarding the average spectra for each class, distinction between sound and insect-damaged acorns began to rise after 600nm, with clear difference between average values occurring in the infrared range after 780nm (Figure 3, below).

Raw data was analyzed to determine the impact of preprocessing on the classification. Principal component analysis with a 95 % confidence interval was applied to the calibration set for the wavelength ranges of 400~1000nm, 400~780nm, and 780~1000nm to understand the distribution of the raw data. As a result, clustering between sound and insect-damaged acorn classes was observed in the score plot (Figure 4). The clusters among sound acorn groups and loose clustering of insect-damaged acorn groups were observed in the score plot composed of the first principal component (96.7 %) and the second principal component (3.94 %) for the wavelength range of 780~1000nm (Figure 4). The direction of cluster made differences, where the sound clustered towards the first component and the unsound clustered towards the second component.

Unprocessed spectral variables were analyzed to compare the performance of multivariate analysis models according to the wavelength range and preprocessing (Table 2). All ranges showed a recall rate of over 95 % for the classification of sound acorns, with particularly high recall rates of 98.37 % and 99.19 % for PLS-DA and OPLS-DA, respectively, in the 400~780nm range. On the other hand, the classification of insect-damaged acorns presented a different trend. Within the 400~1000nm range, PLS-DA and OPLS-DA showed accuracies of 85.87 % and 77.17 %, respectively (Table 2). In the near-infrared range of 780~1000nm, both PLS and OPLS-DA showed the same accuracy of 85.87 %. However, in the visible light range of 400~780nm, the accuracy was 67-69 %, indicating lower model performance compared to the other two ranges (Table 2). This difference was more pronounced in terms of the G mean, with PLS-DA and OPLS-DA showing G means of 33.60 % and 22.09 % respectively in the 400~780nm range, which was significantly lower than the 65-80 % in the other ranges.

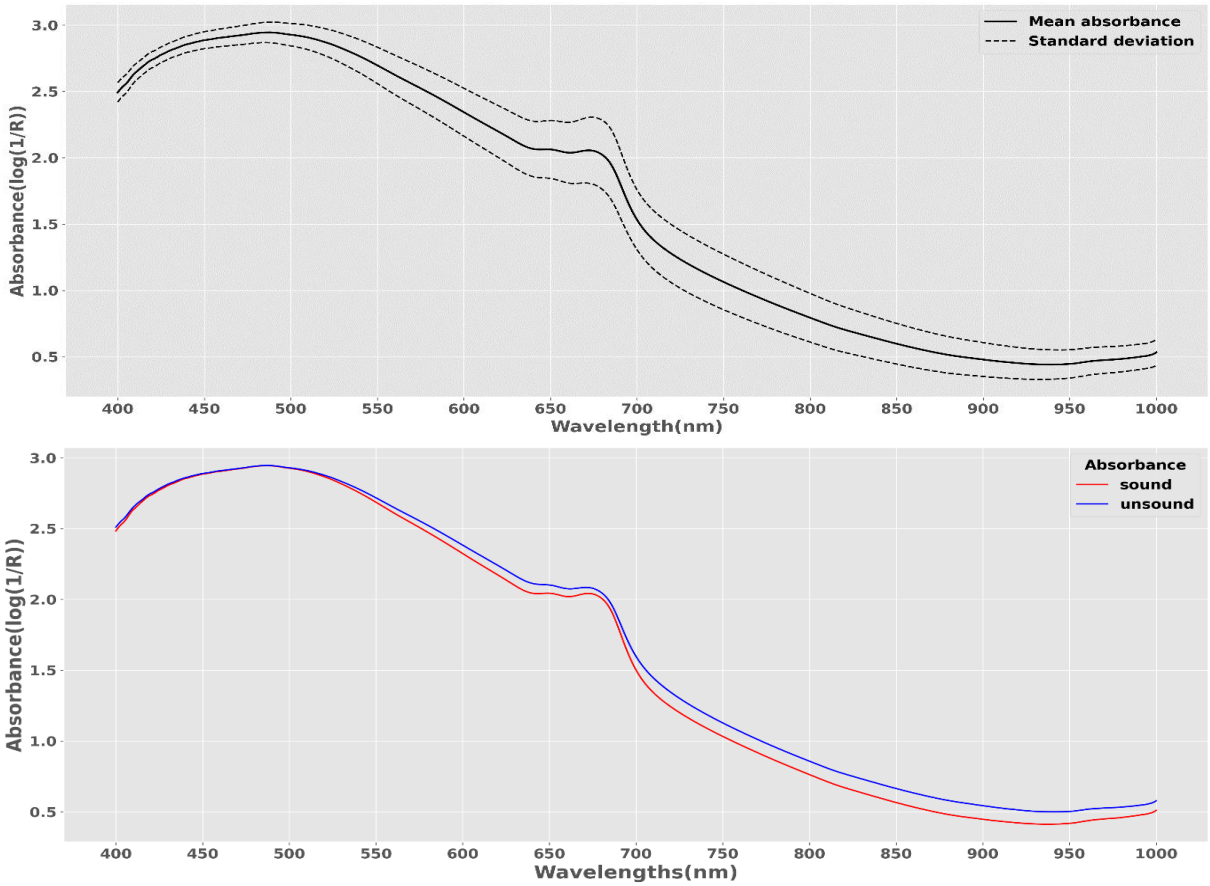

Figure 3
Whole Vis-NIR raw spectra from 400 to 1000nm for acorns of *Q. acuta* (above), and averages of wavelength between sound and insect-damaged acorns from 400 to 1000nm (below)

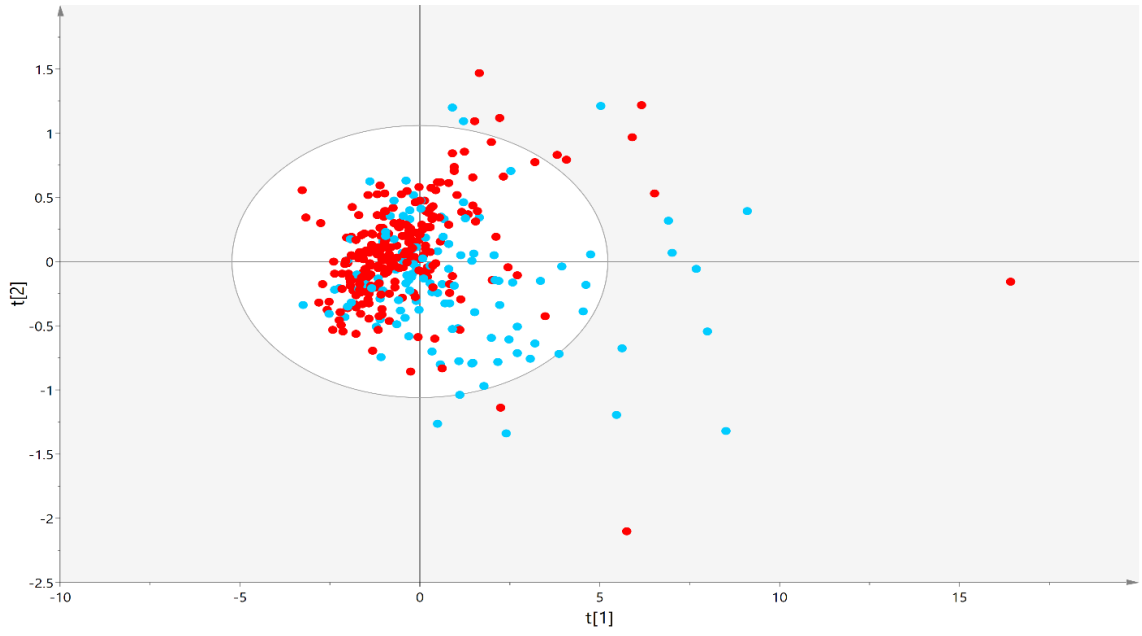

Figure 4
Principal component analysis for sound and damaged acorns in the wave range from 780 to 1000nm. Orange and blue dots indicate sound and unsound acorns, and a circle is Hotelling's 95 % confidence intervals.

Table 2. Evaluation index for raw spectra

Preprocessed	Raw spectra					
Wavelength	400~1000nm		400~780nm		780~1000nm	
Model name	NWP	NWO	NVP	NVO	NIP	NIO
Analysis	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA	PLS-DA	OPLS-DA
Components	12	1+3	2	1+0	7	1+5
Accuracy	0.8587	0.7717	0.6957	0.6793	0.8587	0.8587
Recall	0.9512	0.9350	0.9837	0.9919	0.9593	0.9512
Precision	0.8540	0.7718	0.6914	0.6778	0.8489	0.8540
F1_Score	0.9000	0.8456	0.8121	0.8053	0.9008	0.9000
G_Mean	0.7996	0.6433	0.3360	0.2209	0.7931	0.7996

Note: N: no preprocessing, W: whole, V: visual, P: partial least squares (PLS), O: orthogonal partial least squares (OPLS), I: infrared, DA: discriminant analysis

For data preprocessed with the multiplicative scatter correction (MSC) and the standard normal variate (SNV), a similar pattern was observed. While it was difficult to find absorbance peaks that differentiated the classes in the full wavelength and near-infrared wavelength ranges, the differences between classes were observed in the 780~1000nm range. With the first derivative (FD) preprocessing, absorbance peaks were observed around 680nm, 690nm, and 980nm for the range of 400~1000nm. However, it was difficult to find a tendency of separation between the classes. When the preprocessing was only applied to the 780~1000nm range, a strong absorbance peak was observed around 960nm, and the differences between sound and insect-damaged acorns were observed after 850nm.

With the Savitzky-Golay filter (SGF) preprocessing on the 400~1000nm range, absorbance peaks similar to the raw spectra were observed at 680nm and in the 900~1000nm range, and differences between classes were observed across the entire wavelength range. When preprocessing was applied only to the near-infrared range of 780~1000nm, the differences between classes widened based on the absorbance points in the 900~1000nm range.

For multivariate analysis, the partial least squares discriminant analysis (PLS-DA) discriminated the seeds with an accuracy of 66.85-86.41 %, while the orthogonal partial least squares discriminant analysis (OPLS-DA) showed an accuracy of 66.85-86.41 %. Except when the FD preprocessing was applied to the 400-1000 nm range, PLS-DA showed higher or similar discriminatory power to OPLS-DA (Table S1).

Among the models combining each preprocessing and multivariate techniques, the model applied with SGF and PLS-DA on the 400~1000nm wavelength range showed the highest accuracy of 86.41 %, with precision, F1 score, and G mean of 86.03 %, 90.35 %, and 80.93 % respectively. This demonstrated the best classification performance for sound and insect-damaged acorns. However, models without preprocessing showed high classification performance with precision, F1 score, and G mean of 85.4 %, 90 %, and 79.96 % which were almost similar to or higher than those with SGF preprocessing and most other preprocessing methods.

Performance of the model was visually assessed using a score scatter plot. Although it was difficult to clearly observe a distinct separation trend solely based on the 1st principal component (explaining 77.7 % of the variance) and the 2nd principal component (explaining 13 % of the variance), clustering of the classes could still be identified (Figure 5).

For the partial least squares (PLS) model, the actual sample classification indicator, y-predictability, is set around a baseline (threshold) of 0.5, which divides the observations into 0 and 1. In the case of the SGWP model, 111 out of the actual 116 sound acorns were correctly identified, and 41 out of 58 infested acorns were correctly distinguished.

When using all variables, computational issues can arise during the actual spectral analysis classification. Therefore, to determine which wavelength variables are preferable to use in the actual model and which wavelengths should be selected in the future, important variable selection was conducted through VIP. There were 161 variables with a VIP exceeding 1, mainly in the visible light range between 650~750nm, around 400nm, and variables above 950nm (Figure 6). Similarly, regression coefficients also showed strong peaks in the 650~750nm range, followed by other peaks around 400nm and 950nm.

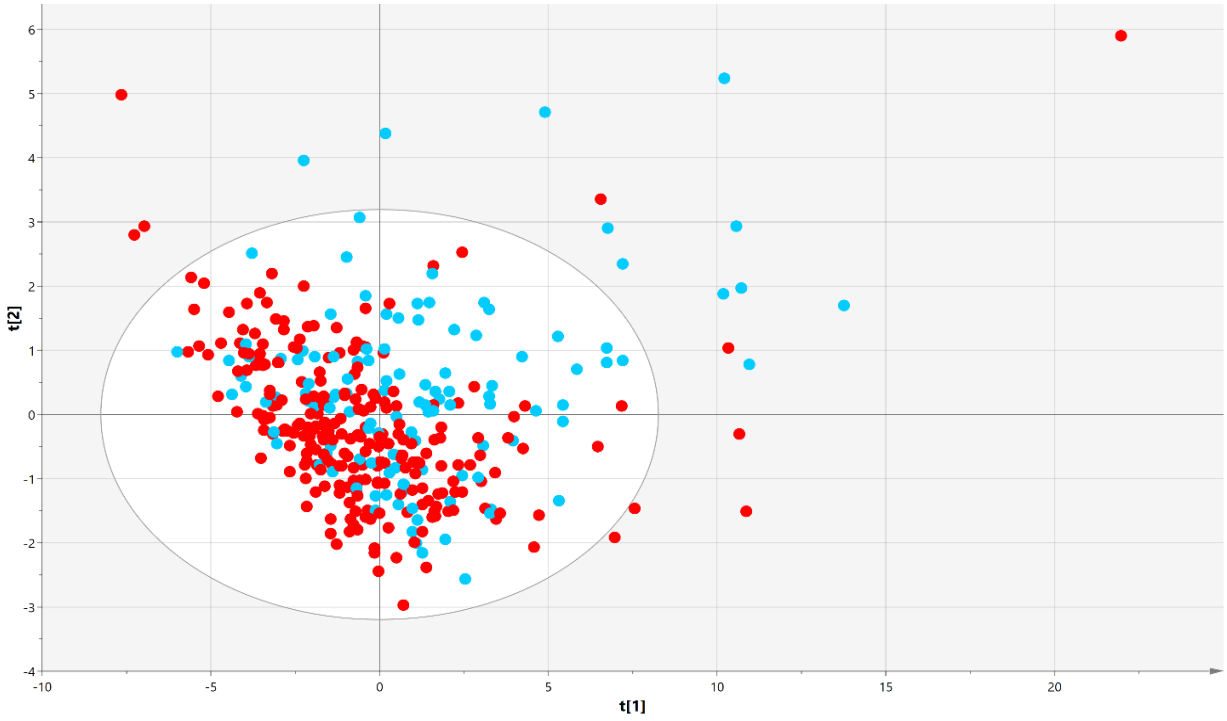

Figure 5
Principal component (PC) score plot for 400~1000nm Savitzky-Golay filter processed model. t[1]: partial least square PC that take the largest part of the explainable variation (77 %), and t[2]: the 2nd largest PC (13 %)

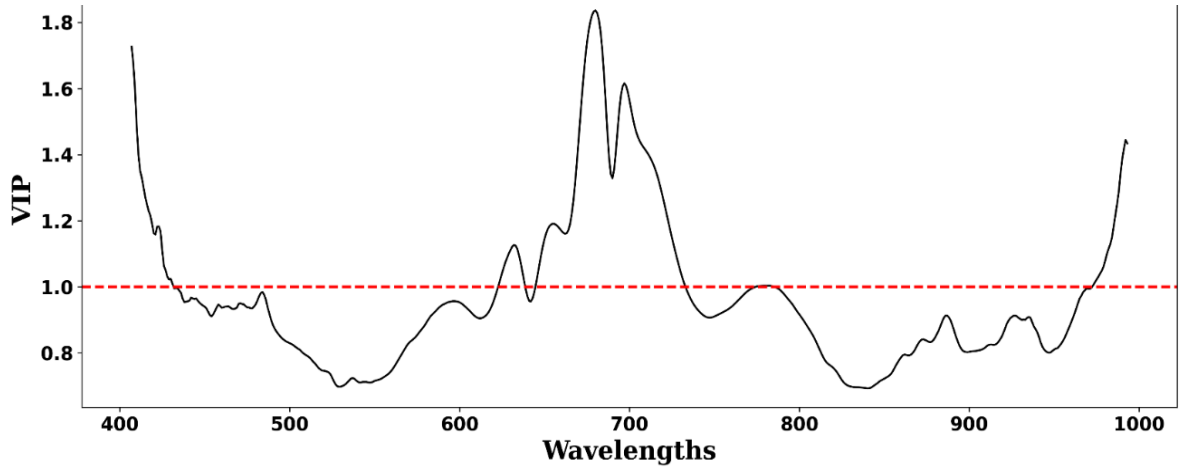

Figure 6
Variable importance in projections (VIP). A normalized index that explains which spectra in the model were responsible for y-predicted-value decision.

Discussion

Consideration for wavelength ranges

The selection of wavelength ranges considers various factors. For instance, the model with SGF preprocessing and PLS-DA in the 780~1000nm range showed the highest accuracy, F1 score, and G mean. When it comes to specific metrics, however, accuracies for the 400~1000nm and 780~1000nm ranges were between 0.84 and 0.86, which showed a slight difference of about 1~2 %. This suggests that the difference in model performance between the wavelength ranges used is minimal.

Using only the visible light wavelength range reduced the accuracies to 66~68 %. While sensitivity was very high at 98~99 %, and G mean was 22~33 %. This indicates that using only the visible light range could result in the misclassification of insect-damaged acorns as sound acorns. This result is consistent with the previous study (Tigabu et al. 2007) that showed about 90 % accuracy within a 2 % error margin regardless of whether the visible light range was included. But the results using only visible range data showed an high accuracy of about 86 %.

Several reasons can be speculated for these results. The visible light range contains numerous interfering factors. For instance, the 680nm region represents a specific absorption point, which is associated with red and chlorophyll light absorption (Wati et al. 2021). Factors such as undecomposed chloroplasts and acorn lipid content influence color changes in the visible light range (Bonner 1987). Additionally, the acorns with severe insect infestation have their acorn coat darkened. These factors likely made classification based solely on analysis in the visible light range challenging.

However, it appears that using the visible range along with the near-infrared range does affect the performance of classification models. The VIP technique showed a high variable importance at the 680nm absorption peak (Figure 6), and the regression coefficient also confirmed a sharp peak within the same range, which is close to red edge range utilized in various plants' disease detection studies (Lowe et al. 2017). The first principal component of PLS included visible and near-infrared wavelength variables over 600nm, suggesting the inclusion of the visible light range in subsequent acorn insect-damage studies. In future study, increasing the number of samples would be the best solution to offset the random variation within the visible range and increase the discrimination power. However, this may be inefficient due to the low seed yield of *Q. acuta*. Therefore, more advanced observation techniques such as hyperspectral imaging should be used to utilize tunnels and blackening areas that are common in insect-damaged part of the seeds through segmentation (Munera et al. 2021).

In the near-infrared range, most wavelength variables with VIP > 1 were located between 960nm and 1000nm with significant absorption peaks of high regression coefficient in this range. The absorption peaks between 960nm and 1000nm are known to be influenced by the 3rd and 4th overtones of molecular O-H and C-H bending motions (Cen and He 2007), particularly reacting with O-H groups in water and carbohydrates in fruits (Guo et al. 2016).

Acorns primarily consist of water, carbohydrates (especially starch), and other components including tannins (Chung et al. 1975), and it seems that the internal effects of insect infestation on the water and carbohydrates led to the current results (Wang et al. 2010, Jamshidi et al. 2019, Jamshidi 2020). Still, factors other than insect infestation, such as fungi infection, could be the confounding one by changing the structure of starch granules (He et al. 2021), which should be strictly managed in the successive study. Meanwhile, the absorption point of tannin components at 938nm was noted in the regression coefficient (Takahashi et al. 2011) but did not significantly impact the determination of the insect-damaged acorns.

In this study, though the visible and near-infrared regions are different ranges, it seemed that the information of various chlorophylls in the visible range and the material information available in the near-infrared region, as well as the interaction between them, improved the performance of the models.

Effects of preprocessing methods and multivariate techniques

The preprocessing methods showed differences in G mean scores, which seemed to be due to specificity included in the G mean calculation. In the 780~1000nm range, the specificity (capability) to correctly classify insect-damaged acorns was between 62~67 %, showing no significant variance. However, in the 400~1000nm range, when using SNV and MSC, and both PLS and OPLS-DA, the specificity displayed only 1.64 %, almost unable to distinguish insect-damaged acorns.

The results of applying SNVs to the near-infrared range were similar to previous studies. Moscetti et al. detected insect-damaged chestnuts with an accuracy of about 92 % using the near-infrared wavelength variables processed with MSC (Moscetti et al. 2014), and Rahman et al. showed that both MSC and SNV were more than 80 % specificity for macadamia seed classification (Rahman et al. 2021). On the other hand, Jamshidi et al found a decrease in specificity when SNVs were processed for the 400-1000 nm wavelength band, but an increase in classification performance (Jamshidi et al. 2019). This discrepancy might be resulted from the characteristics of the SNV and MSC preprocessing methods.

Although SNV and MSC use different variables in their formulas, both remove unnecessary noise caused by scatter effects in the wavelength variables. Hence they have the similar statistical preprocessing outcome of removing scatter effects when used in actual analysis (Rinnan et al. 2009).

If the scatter effect is not significant and the compositional variance is greater, these preprocessing methods can remove both of them.(Næs et al. 2002). Statistics used in SNV and MSC, such as mean, standard deviation, additive effect, and multiplicative effect, include not only color variations in the visible light range but also compositional variations within the near-infrared range. This leads to the conclusion that the data intended for classification were also removed along with scatter effects.

These scatter effects can also be caused by the presence of heterogeneous structures within plant tissues (Wang et al. 2020). Since the acorn weevil can create such structures during

its attack on the cotyledon of acorns, the erased scatter effect may contain important information needed for the discrimination of insect-damaged seeds (Csóka and Hirka 2006).

On the other hand, other preprocessing methods such as FD and SGF, which remove baseline effects and amplify absorption peaks without removing a wide range of variance including scatter effects, seemed to lead to higher specificity in both the visible and near-infrared ranges. Thus, statistical preprocessing using SNV and MSC thought to be unsuitable for visual-near-infrared analysis of acorn insect-damage detection.

PLS-DA is a statistical technique that compresses dimensionality by maximizing the covariance between the target class and predictors while minimizing other variances (Barker and Rayens 2003). For example, seed family is a factor that can act as a random variation in this study. Ideally, we would collect many seed samples from each family to compare models across families. However, due to the low production of *Quercus acuta* seeds, increasing the number of samples was not an available option. Therefore, it was assumed that PLS-DA would minimize the variance arising from these families and maximize the covariance between sound and unsound seeds. In future studies, the uniformity in the number of samples between families need to be maintained.

While hyperspectral and multidimensional spectroscopy can be practical for automation of acorn sorting (Feng et al. 2019), the complexity of data analysis methods makes the identification of key variables significant. The results could contribute to testing for insect damage, a critical factor in determining acorn quality. The meaningful preprocessing combinations and wavelength ranges identified in this study could serve as reference materials for future research on subtropical evergreen broad-leaved tree acorns using the visible-shortwave infrared (400~2500nm) range. Furthermore, the significant absorption points and wavelength variables identified in this study could help narrow down the wavelength range in research methods using more advanced optical classification techniques like hyperspectral and multidimensional spectroscopy.

Conclusion

This study hypothesized that spectral analysis in the visual-near-infrared range could be used for insect infestation classification in acorns of *Q. acuta*. To test this hypothesis, spectral data from the visible to near-infrared range (VIS-NIR 400~1000nm) was utilized. The aim was to identify the most suitable model and wavelength variables for the classification of insect-damage acorns.

Models using all wavelength ranges and those using only the near-infrared range showed little difference in classification performances, with accuracies of 0.86~0.87 and G mean values around 0.80~0.82. In contrast, when only the visible light range was used, the power to test for insect-damaged acorns dropped close to 0, showing a low accuracy of 0.66. Ultimately, the model using the 400~1000nm range with Savitzky - Golay filter preprocessing followed by PLS-DA showed the best metrics, with accuracy of 0.864, precision of 0.860, F1 Score of 0.904, and G mean of 0.809.

The influence of wavelength variables within the model was examined using VIP score and regression, revealing that wavelength variables and absorption peaks between 660~720nm and 960~1000nm significantly impacted the classification model. The visible light range was interfered with by not only insect damage but also other factors, yet it still significantly influenced the model's testing power. The near-infrared range analysis suggested that differences in water and carbohydrate changes occurring during the larval infestation process impacted the model's discrimination capability. Therefore, it appears appropriate to use the entire 400~1000nm range. Moreover, focusing on narrowing down around the two mentioned absorption peaks was important in reducing the number of wavelength variables. But when using the entire range for the insect-damage detection of acorn, the SG or FD preprocessing seemed to perform well, while the SNV or MSC did not due to removal of scatter effects containing important information.

We believe the preprocessing and range combinations used in this study to help further research involving a broader wavelength range up to 2500nm. The significant wavelength ranges and absorption peaks identified would help narrow down the research scope when using multidimensional spectroscopy equipment.

Statements and Declarations

Funding

This work 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).

Conflict of interest

The authors declare no competing interests.

Acknowledgement

We wish to thank the National Forest Seed and Variety Center for the technical supports.

Data archiving statement

Datasets generated during the current study are available from the corresponding author upon reasonable request.

References

Barker M, Rayens W (2003) Partial least squares for discrimination. Journal of Chemometrics 17(3): 166-173. https://dx.doi.org/10.1002/cem.785

Bonner FT (1987) Seed biology and technology of *Quercus*. Louisiana, U.S.: US Department of Agriculture, Forest Service, Southern Forest Experiment Station, 5-6, 11-12 p

Cen H, He Y(2007) Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science and Technology 18(2): 72-83. https://dx.doi.org/10.1016/j.tifs.2006.09.003

Chung DH, Yu TJ, Choi BK (1975) Studies on the utilization of acorn starch -part I. properties of acorn starch-. Applied Biological Chemistry 18(2): 102-108

Chung HI, Kim HJ (2000) Near-infrared spectroscopy: principles. Analytical science & technology 13(1): 138-151

Connor K (2004) Storing acorns. Native Plants Journal 5(2): 160-166

Csóka G, Hirka AC (2006) Direct effects of carpophagous insects on the germination ability and early abscission of oak acorns. Acta Silvatica et Lignaria Hungarica 2: 57-67

Daneshvar A, Tigabu M, Karimidoost A, Oden PC (2015) Single seed near infrared spectroscopy discriminates viable and non-viable seeds of *Juniperus polycarpos*. Silva Fennica 49(5). https://dx.doi.org/10.14214/sf.1334

Dorsey CK, Tryon EH, Carvell KL (1962) Insect damage to acorns in West Virginia and control studies using granular systemic insecticides. Journal of Economic Entomology 55(6): 885-888. https://dx.doi.org/10.1093/jee/55.6.885

FAO (2010) Seeds in emergencies : a technical handbook. Rome: FAO, ISBN 9789251066768

Farhadi M, Tigabu M, Stener L-G, Odén PC (2016) Feasibility of visible+ near infrared spectroscopy for non-destructive verification of European× Japanese larch hybrid seeds. New Forests 47(2): 271-285

Feng L, Zhu S, Liu F, He Y, Bao Y, Zhang C (2019) Hyperspectral imaging for seed quality and safety inspection: a review. Plant Methods 15(1): 91-91. https://dx.doi.org/10.1186/s13007-019-0476-y

Government of Korea (2021) 2050 Carbon neutrality scenarios. Available from https://www.2050cnc.go.kr/eng/board/read?boardManagementNo=28&boardNo=608&searchCategory=&page=1&searchType=&searchWord=&menuLevel=2&menuNo=65

Gribko LS (1995) The effect of acorn insects on the establishment and vigor of northern red oak seedlings in north-central West Virginia. Proceedings of the 10th Central Hardwood Forest Conference, West Virginia, U. S.: 430-441

Gribko LS, Jones WE (1995) Test of the float method of assessing northern red oak acorn condition. Tree Planters' Notes 46(4): 143-147

Guo W, Zhao F, Dong J (2016) Nondestructive measurement of soluble solids content of kiwifruits using near-infrared hyperspectral imaging. Food Analytical Methods 9(1): 38-47

He X, Yan C, Jiang X, Shen F, You J, Fang Y (2021) Classification of aflatoxin B1 naturally contaminated peanut using visible and near-infrared hyperspectral imaging by integrating spectral and texture features. Infrared physics & technology 114: 103652. https://dx.doi.org/10.1016/j.infrared.2021.103652

Hirayama D, Fujii T, Nanami S, Itoh A, Yamakura T (2012) Two-year cycles of synchronous acorn and leaf production in biennial-fruiting evergreen oaks of subgenus *Cyclobalanopsis* (*Quercus, Fagaceae*). Ecological Research 27(6): 1059-1068. https://dx.doi.org/10.1007/s11284-012-0986-9

Hwang JS, Kang JT, Son YM, Jeon HS (2015) Prediction of the optimal growth site and estimation of carbon stocks for *Quercus acuta* in Wando Area. Journal of Climate Change Research 6(4): 319. https://dx.doi.org/10.15531/ksccr.2015.6.4.319

Jamshidi B (2020) Ability of near-infrared spectroscopy for non-destructive detection of internal insect infestation in fruits: Meta-analysis of spectral ranges and optical measurement modes. Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy 225: 117479. https://dx.doi.org/10.1016/j.saa.2019.117479

Jamshidi B, Mohajerani E, Farazmand H, Mahmoudi A, Hemmati A (2019) Pattern recognition-based optical technique for non-destructive detection of *Ectomyelois ceratoniae* infestation in pomegranates during hidden activity of the larvae. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 206: 552-557

Khan S, Habib A (2015) Effect of pre-sowing treatments on seed germination in *Quercus glauca* Thunb., collected from different sampling sites of the Himalayan region of Pakistan. International Journal of Biosciences 6(11): 42-48

Kim C, Kim W, Song W, Cho J, Choi J (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(1): 87

Kim HJ, Lee SH (2017) Estimating carbon storage and CO2 absorption by developing allometric equations for *Quercus acuta* in South Korea. Forest science and technology 13(2): 55-60

Kim S, Park I-H (2021) Acorn production and characteristics of *Quercus acuta* thunb-Focused on Wando, Jindo and Haenam in Jeollanam-do, Korea. Korean Journal of Environment and Ecology 35(6): 621-631

Lee JH, Choi BH (2010) Distribution and northernmost limit on the Korean Peninsula of three evergreen trees. Korean Journal of Plant Taxonomy 40(4): 267

Lowe A, Harrison N, French AP (2017) Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. Plant Methods 13(1): 80-80. https://dx.doi.org/10.1186/s13007-017-0233-z

Moscetti R, Haff RP, Saranwong S, Monarca D, Cecchini M, Massantini R (2014) Nondestructive detection of insect infested chestnuts based on NIR spectroscopy. Postharvest biology and technology 87: 88-94. https://dx.doi.org/10.1016/j.postharvbio.2013.08.010

Munera S, Rodríguez-Ortega A, Aleixos N, Cubero S, Gómez-Sanchis J, Blasco J (2021) Detection of invisible damages in 'rojo brillante' persimmon fruit at different stages using hyperspectral imaging and chemometrics. Foods 10(9): 2170. https://dx.doi.org/10.3390/foods10092170

Næs T, Isaksson T, Fearn T, Davies T (2002) A user-friendly guide to multivariate calibration and classification. Chichester, UK: NIR Publications, ISBN 978-1-906715-25-0

NFSV (The National Forest Seed and Variety Center) (2021) The 2nd plan for the management and establishment of the seed orchard (2022~2026).

Novikov AI (2019) Visible wave spectrometric features of scots pine seeds: the basis for designing a rapid analyzer. In: 2020 IOP Conference Series: Earth and Environmental Science.: Earth Environmental Science. Bristol, UK: IOP Publishing, p 12064

Novikov AI, Ersson BT, Malyshev VV, Petrishchev EP, Ilunina AA (2020) Mechanization of coniferous seeds grading in Russia: a selected literature analysis. In: 2020 IOP Conference Series: Earth and Environmental Science.: Earth Environmental Science. Bristol: IOP Publishing, p 12060

Novikov AI, Novikova TP (2018) Non-destructive quality control of forest seeds in globalization: Problems and prospects of output innovative products. Globalization and Its Socio-Economic Consequences: 1260-1267

Rahman A, Wang S, Yan J, Xu H (2021) Intact macadamia nut quality assessment using near-infrared spectroscopy and multivariate analysis. Journal of food composition and analysis 102: 104033. https://dx.doi.org/10.1016/j.jfca.2021.104033

Rinnan Å, Van Den Berg F, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 28(10): 1201-1222

Shenk JS, Westerhaus MO (1993) Near infrared reflectance analysis with single and multiproduct calibrations. Crop Science 33(3): 582-584

Son YM, Kim RH, Kang JT, Lee KS, Kim SW (2014) A practical application and development of carbon emission factors for 4 major species of warm temperate forest in Korea. Journal of Korean Forest Society 103(4): 593-598. https://dx.doi.org/10.14578/jkfs.2014.103.4.593

Takahashi A, Shimada T, Kawano S (2011) Nondestructive determination of tannin content in intact individual acorns by near-infrared spectroscopy. Ecological Research 26(3): 679-685. https://dx.doi.org/10.1007/s11284-011-0823-6

Tigabu M, Daneshvar A, Wu R, Ma X, Christer Odén P (2019) Rapid and non-destructive evaluation of seed quality of Chinese fir by near infrared spectroscopy and multivariate discriminant analysis. New Forests 51(3): 395-408. https://dx.doi.org/10.1007/s11056-019-09735-8

Tigabu M, Fjellström J, Odén PC, Teketay D (2007) Germination of *Juniperus procera* seeds in response to stratification and smoke treatments, and detection of insect-damaged seeds with VIS + NIR spectroscopy. New Forests 33(2): 155-169. https://dx.doi.org/10.1007/s11056-006-9020-9

Tigabu M, Odén PC (2002) Multivariate classification of sound and insect-infested seeds of a tropical multipurpose tree, *Cordia africana*, with near infrared

reflectance spectroscopy. Journal of Near Infrared Spectroscopy 10(1): 45-51

Wang J, Nakano K, Ohashi S, Takizawa K, He J (2010) Comparison of different modes of visible and near-infrared spectroscopy for detecting internal insect infestation in jujubes. Journal of Food Engineering 101(1): 78-84

Wang Z, Künnemeyer R, McGlone A, Burdon J (2020) Potential of Vis-NIR spectroscopy for detection of chilling injury in kiwifruit. Postharvest biology and technology 164: 111160. https://dx.doi.org/10.1016/j.postharvbio.2020.111160

Wati RK, Pahlawan MFR, Masithoh RE (2021) Development of calibration model for pH content of intact tomatoes using a low-cost Vis/NIR spectroscopy. In: 2021 IOP Conference Series: Earth and Environmental Science.: Earth Environmental Science. Bristol: IOP Publishing, p 12049

Xia K, Daws MI, Stuppy W, Zhou Z-K, Pritchard HW (2012) Rates of water loss and uptake in recalcitrant fruits of *Quercus* species are determined by pericarp anatomy. PLoS One 7(10): e47368-e47368. https://dx.doi.org/10.1371/journal.pone.0047368

Zhang Z, Liu H, Chen D, Zhang J, Li H, Shen M, Pu Y, Zhang Z, Zhao J, Hu J (2022) SMOTE-based method for balanced spectral nondestructive detection of moldy apple core. Food control 141: 109100. https://dx.doi.org/10.1016/j.foodcont.2022.109100