https://doi.org/10.48048/tis.2025.9546

Support Vector Machine Coupled with Modular Visible-Near Infrared Spectroscopy for Chili (*Capsicum annuum*) Seed Viability Discrimination

Hanim Zuhrotul Amanah[1], Yumna Fauzia Rahmannisa[1], Reza Adhitama Putra Hernanda[2], Hoonsoo Lee[2], Nadya Hafidzatun Nisa[1], Rizki Maftukhah[1] and Rudiati Evi Masithoh[1,*]

[1]Department of Agricultural and Biosystems Engineering, Faculty of Agricultural Technology, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[2]Department of Biosystems Engineering, College of Agriculture, Life, and Environment Sciences, Chungbuk National University, Cheongju 28644, Republic of Korea

([*]Corresponding author's e-mail: evi@ugm.ac.id)

Received: 17 December 2024, Revised: 26 December 2024, Accepted: 26 January 2025, Published: 25 March 2025

Abstract

This work investigated the performance of spectroscopic data in conjunction with a Gaussian support vector machine (SVM) for non-destructive discrimination of chili pepper (CP) seed viability. This present study also involved 2 wavelength selection methods, namely variable important in projection (VIP) and backward PLS (bPLS), which were realized and examined. The spectra data of CP seeds were collected at 2 regions: Visible-near infrared (Vis/NIR, 400 - 1000 nm) and shortwave near-infrared (SWNIR, 1000 - 1700 nm). The individual and mixed (generalized) discrimination models were then examined. This study demonstrated that a generalized model with effective Vis/NIR wavelengths through VIP achieved the optimum discrimination prediction accuracy (97.22 %). Thus, our findings successfully provided a general calibration model in a non-destructive way to discriminate CP seed viability. Our findings hold promises for practical implementation in the seed industry.

Keywords: *Capsicum annuum*, Gaussian SVM, Viability, Spectroscopy, Variable importance in projection

Introduction

CPs are a horticultural crop used as spices originally from South America and belong to the genus Capsicum of the Solanaceae family. CPs are consistently more well-known than other food materials due to their pungency flavor and strong aroma. Their spiciness is caused by the presence of capsaicin found in the crop. Moreover, several studies have revealed the benefits of CP to human health, such as its antioxidants, anti-inflammatory, anti-obesity, and other properties [1,2].

CP plants are easily grown in tropical countries, such as Indonesia, due to their warmth and humidity [1]. Not only due to proper climatic conditions, but the CP productivity is also affected by the quality of CP seed; specifically, seed viability determines the success of growing CP plants. Seed viability refers to the ability of seeds to germinate in a seed lot [3]. Various factors affect seed viability, including postharvest handling practices, such as drying and storage [4,5], and the inherent characteristics of the seeds. Therefore, conducting viability tests is essential for ensuring sustainable CP production.

Yet, it is nearly impossible for farmers to assess the CP viability through their naked eyes. Several measurements are used to test seed viability. For instance, the viability of *Jatropha curcas* L. was confirmed by sowing the seed in the sand and performing a tetrazolium chloride (TTZ) test [6]. Likewise, this method was also carried out to check the viability of barley [7] and papaya seeds [8]. Nevertheless, as the current germination test method is time-consuming and inefficient, the seed could die

during measurement before obtaining the results [3]. Moreover, this method produces chemical waste and requires skilled operators. Thus, an alternative method, i.e., non-destructive measurement, is necessary to replace the common-conventional method. Therefore, this study will evaluate a non-destructive method, implementing a spectroscopic technique. The spectroscopy method is used to overcome the limitations of the conventional methods since it offers rapid, less sample preparation, and non-destructive measurement.

Since the spectroscopy technique is measurable in a wide range of electromagnetic spectrum, a visible-near infrared (Vis/NIR) and shortwave near-infrared (SWNIR) will be used in our study. Various studies have demonstrated the feasibility of Vis/NIR and SWNIR spectroscopies in agricultural practices. Evidence showed that organic compounds found in agriculture samples are noticeable in these regions [9-12]. Spectroscopy describes the positive interaction between electromagnetic waves and chemical bonds, such as N-H, O-H, C-H, etc., which may help identify any specific chemical compound of the objects [10], which is the seed. In recent years, spectroscopy has been used for food safety and quality inspections. For instance, water content in porang powder was measured using Vis/NIR spectroscopy combined with partial least squares regression (PLSR), yielding 0.96 of R^2 with 0.56 % wb of the standard error of prediction (SEP) [13]. One study achieved 86 % accuracy in discriminating against the cocoa ripeness through Vis/NIR spectroscopy combined with principal component analysis (PCA) and linear discriminant analysis (LDA) [14]. The ability to predict soybean seeds' protein and lipid content using Raman spectroscopy in the PLSR has been investigated with R^2 and SEP of 0.872 and 0.759 %, respectively [15]. Aside from these applications, spectroscopy techniques are also applicable for seed viability discrimination. For example, a Fourier transform-near infrared (FT-NIR) spectroscopy equipped with partial least squares discriminant analysis (PLS-DA) generated > 99.2 % accuracy in detecting soybean seeds [16] and super sweet corn viabilities with 98.7 % accuracy [17]. Meanwhile, another work demonstrated the feasibility of FT-NIR in detecting corn seed viability by PLS-DA with accuracy ranging from 62.2 to 100 % [4]. Therefore, these published works reported the high accuracy of spectroscopy in discriminating seed viability.

Support Vector Machine (SVM) is a well-known nonlinear traditional machine learning technique used for decades in various fields, including spectroscopy. The SVM technique commonly implements kernel equations, such as linear, Gaussian, and polynomial. Among them, the Gaussian kernel has been used frequently due to its ability to handle highly non-linear data, elsewhere cited [18]. The equation of the Gaussian kernel is the formulae in Eq. (3) of this manuscript in the SVM subsection. In-depth investigations of SVM have been done for years to facilitate spectroscopic analysis in various fields, such as pharmacy, environment, and agriculture. However, the following description will only highlight the application of SVM in tandem with spectral data in agriculture. Previously, scholars employed SVM with NIR spectral data to classify the tea quality, attaining only a 5 % classification error [19]. Compared with a few classifiers, namely LDA and PLS-DA, an on-line NIR hyperspectral imager combined with linear kernel SVM yielded up to 100 % accuracy in discriminating corn viability [20]. SVM also generated a similar result in detecting the adulterant in Tartary buckwheat using a benchtop NIR spectrometer, while the maximum accuracy of PLS-DA was only 94.22 % [21]. Moreover, combined with mid-infrared (MIR) spectroscopy, SVM achieved 100 % accuracy in discriminating coconut from palm kernel oil seeds [22]. Likewise, Windarsih *et al*. [23] accurately predicted lard in tuna oil, indicating an R^2 of 0.993. SVM was powerful for sample classification and regression based on vibrational spectroscopy techniques in these examples. Aside from the flexibility offered by SVM, its inherent characteristic of maximizing the margin between classes and samples enables those researchers to achieve good classification and regression performance [23]. Furthermore, SVM can deal with a limited sample, whereas other nonlinear models, such as neural networks, require thousands of sample variations within the calibration dataset [12,24].

To date, the application of spectroscopic methods for seed viability discrimination, as aforementioned, remains limited. One possibility is that internal quality improvements, such as those achieved through genetic engineering, have been implemented to increase viability rates, as cited elsewhere [25]. Notably,

although genetic engineering could improve the crop's productivity and increase the germination rate [26], it remains controversial in several countries and among certain belief systems [27]. Furthermore, despite the high accuracy of FT-NIR and Raman spectroscopy [28], they are relatively more costly in practice than Vis/NIR or shortwave near-infrared (SWNIR) spectroscopy. Thus, this study used modular Vis/NIR spectroscopy, comprising two spectra in a single measurement, i.e., Vis/NIR (400 - 1000 nm) and SWNIR (1000 - 1700 nm), with the purpose as follows: 1) Investigate the feasibility of Vis/NIR and SWNIR spectroscopy to discriminate CP seed based on its viability and 2) Develop a non-destructive model of CP viability test using Vis/NIR and SWNIR spectroscopy by the SVM method through the whole and selected spectrum. This study used 3 common CP seed varieties, namely *Capsicum annum* var. Grossum, *Capsicum annuum* L. var. Longum, and *Capsicum frutescens* L. will be involved. Furthermore, our main goal is to develop one general model such that the spectra from these varieties will be mixed and examined.

Materials and methods

Sample preparation and germination test

Three CP seed varieties, *Capsicum annum* var. Grossum, *Capsicum annuum* L. var. Longum, and *Capsicum frutescens* L., were used in this study and purchased from a CP seed supplier (Infarm, Surabaya, East Java, Indonesia). The inviable CP seed group was created by artificially inactivating the CP seeds by placing the CP seeds in aluminum foil for an hour in a 95 °C water bath (Memmert type WNB10, Memmert GmbH + Co.KG, Germany), following modified procedures from Wen [29]. The destructive method was done after scanning each CP seed using Vis/NIR and SWNIR spectrometers. The viability test of CP seeds was conducted by transferring them between the rolled straw paper, which was previously humidified by spraying it with distilled water. This arrangement was then put vertically inside the chamber with a black high-density polyethylene (HDPE) net around it.

Vis/NIR spectroscopy: Instrumentation and data collection

A Vis/NIR spectrometer (Flame-TVIS-NIR Ocean Optics, Dunedin, FL, USA; 350 - 1000 nm) with a spectral interval of 0.22 nm and a SWNIR spectrometer (Flame-NIR Ocean Optics, Orlando, FL, USA) were used to capture the CP's Vis/NIR and SWNIR fingerprints in our study. During the spectra measurement, the scanning integration time was set at 130 ms with an average of 100 simultaneous scans for the Vis/NIR. Meanwhile, the SWNIR instrument was set at 400 ms for integration time with an average of 350 simultaneous scans. A boxcar width of 1 was employed to ensure robust spectra quality. For the remaining instruments and instrument calibration, the protocol from our previous study was used [13]. In addition, the reflectance (%) mode was employed as the instrument's default setting. Details of the spectrometer are presented in **Figure 1**.

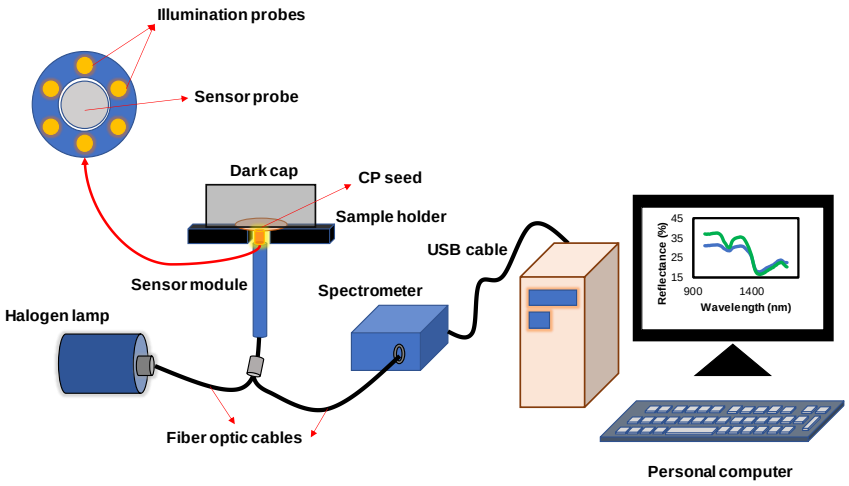

Figure 1 A modular Vis/NIR spectrometer configuration system used in this study.

The spectra measurement was performed by placing each CP single kernel perpendicular to the spectrometer probe on a sample holder. In addition, a dark cap was used to cover the measurement area from ambient light interference. It is important to note that the seeds were carefully wiped using dried tissues to avoid measurement errors due to moisture and dirt. In addition, two spectra profiles, Vis/NIR and SWNIR, were obtained simultaneously for every measurement. We extracted 800×3182 and 800×120 matrices from Vis/NIR and SWNIR (sample×spectra). Furthermore, the spectrum and germination results were compiled and stored in a computer using the .xlsx format before analysis. Noteworthy, due to the low signal-to-noise ratio, the noisy part was removed, and the result was 400 - 1000 nm and 1000 - 1700 nm for Vis/NIR and SWNIR, respectively.

Data preparation: Spectra labeling and preprocessing

Most machine learning models require numeric input for the reference matrix. To facilitate this requirement, our spectra data was numerically labeled as provided below.

$$Y_{class} = \begin{cases} 1 = viable\ group \\ 2 = inviable\ group \end{cases}$$

Second, our original matrices were divided into calibration and prediction data sets for 70% and 30 % of the total data. It is worth noting that the calibration dataset refers to the data for the SVM training, while the prediction dataset is considered the external validation for the calibrated SVM and was not included during the calibration stage. Unlike routine procedures, we do not report any preprocessing techniques since no significant accuracy was found with the raw spectra (data not shown).

Principal component analysis (PCA)

PCA is one of the unsupervised methods known for its ease of use during the computation process. It belongs to dimensional reduction that returns the original spectra matrix into the new variable. In the field of spectroscopy, PCA is applicable for data exploration mainly to visualize whether each sample has different characteristics of spectra information, in other words, is clustering. This study emphasizes the application of PCA to demonstrate its effectiveness in distinguishing the viable and inviable groups based on the Vis/NIR and SWNIR spectra. The linear relationship of the reflectance value (X), scores (T), and loading (P) is indicated in Eq. (1).

$$X = TP^T + error \quad (1)$$

Support vector machine (SVM)

The SVM is essential in optimizing the minimum distance of the hyperplane to the closest training sample [30]. This current study applied nonlinear SVM, the extended aim of the linear SVM. The Gaussian or radial basis function (RBF) kernel with the SVM method was implemented to address this issue. Furthermore, the RBF kernel enables the maximum-margin-separation hyperplane among linear and polynomial kernels to distinguish viable and inviable classes (**Figure 2**) [31]. This kernel function maps the original spectra dataset into high-dimensional space to easily classify the CP seed viability. The mathematical computation of SVM is expressed in Eqs. (2) and (3).

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x^*, x_i) + b\right) \quad (2)$$

$$K(x^*, x_i) = e^{\left(-\frac{\|x^* - x_i\|^2}{2\gamma^2}\right)} \quad (3)$$

where $\hat{y}$ is the SVM-predicted class, α_i refers to Lagrange multipliers calculated on the training step, y_i corresponds with the target class for ith sample, $K(x^*, x_i)$ is the RBF kernel function that calculates similarities between the new sample x^* with the trained dataset x_i, γ is the RBF kernel parameter, and the sign operation converts any result generated from the SVM model into binary, such as 1 and 2. The illustration of SVM is depicted in **Figure 2**.

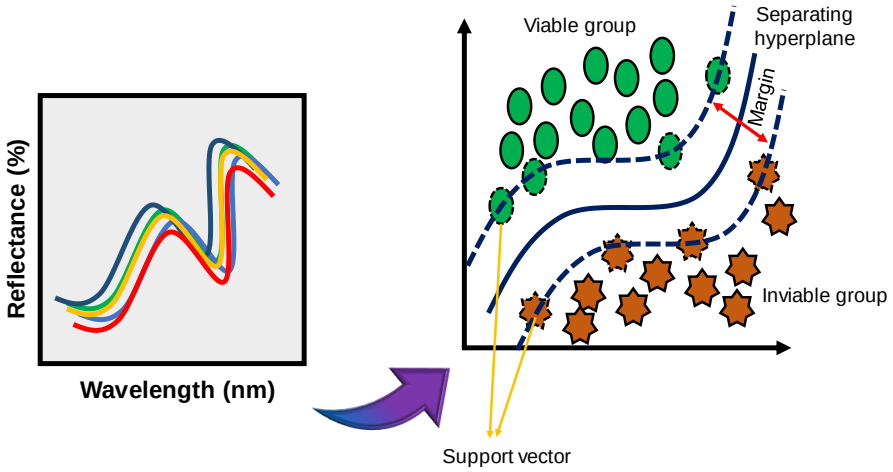

Figure 2 Schematic diagram of SVM brief concept for seed viability discrimination.

Noteworthy, RBF-SVM involves two tuning parameters, the critical point of model quality: 1) The optimization of minimum distances with the hyperplane depends on the cost parameters C and γ in kernel calculation. Therefore, the best combination of these hyperparameters should be matched. In our case, a grid search technique was performed with tenfold cross-validation to determine the best combination of hyperparameters.

Variable selection

More than 3000 wavelengths ranging from 400 to 1000 nm could be extracted for Vis/NIR and 120 wavelengths from SWNIR (1000 - 1700 nm) from the spectrometer instruments. All these wavelengths are considered when detecting CP seed viability. Despite being a robust technique for classification tasks in spectroscopy, as reported by Shrestha et al. [32], SVM takes a relatively more prolonged training step due to massive amounts of wavelengths. Furthermore, only a few wavelengths contribute significantly to discriminating CP seed viability. To that end, the wavelengths with less contribution can be efficiently executed. To achieve this purpose, two types of feature selection algorithms, namely, variable importance in projection (VIP) and backward partial least squares (bPLS), were proposed by Fernández *et al*. [33] and were used. The VIP formula is expressed in Eq. (4) [13,34].

$$VIP_j = \sqrt{\frac{\sum_{i=1}^{I} W_{ji}{}^2 SSY_i J}{SSY_{total} I}} \quad (4)$$

where, W_{ji} denotes the weight value in any element at i and j. SSY_i refers to the sum square of the measured variables in elements i, whereas SSY_{total} describes the total sum of squares measured for the response variable. Finally, J and I correspond with several variables and elements.

The main idea behind the bPLS is to eliminate the unneeded wavelength. The strategy of bPLS was setting a variable's initial and subjected to a partial least squares (PLS) algorithm with full cross-validation to avoid overfitting. The selection of essential wavebands while considering the value of root means squared error (RMSE) during the iteration. This process will continue until the RMSE yields significantly unchanged from the variables. Moreover, the reduction percentage was calculated by the ratio of differences between initial and selected initial wavelengths.

Evaluation metrics

The performance of the SVM classifier for its ability to detect viable and inviable CP seeds was evaluated by calculating the percentage of accuracy, sensitivity, and specificity [34]. These evaluation metrics were computed for both results in the calibration and prediction groups. Mathematically, the evaluation metrics can be formulated in Eq. (5) - (7), respectively.

$$Acc = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \times 100 \quad (5)$$

$$Sn = \left(\frac{TP}{TP + FN}\right) \times 100 \quad (6)$$

$$Sp = \left(\frac{TN}{TN + FP}\right) \times 100 \quad (7)$$

where Acc, Sn, and Sp refer to accuracy, sensitivity, and specificity; TP and TN indicate true positive and true negative; and of FP and FN denote false positives and false negatives. These values are from a confusion matrix, as illustrated in **Figure 3**.

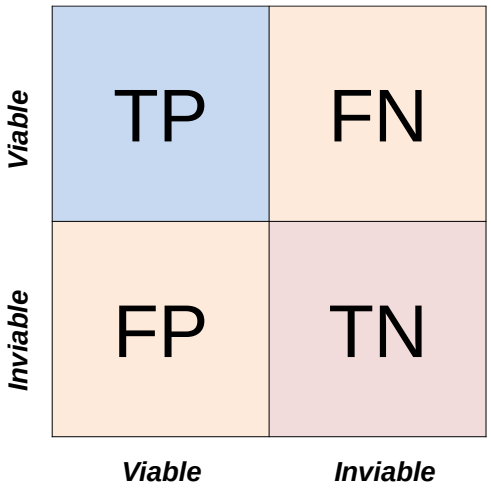

Figure 3 Scenario of confusion matrix. x- and y-axis represent the predicted and actual class.

Software

All calculations in this current study were performed and customized in Python version 3.9 run in Visual Studio Code (version 1.88.1, Microsoft Corporation, Redmond, WA, USA). Microsoft Excel 365 (Microsoft Corporation, Redmond, WA, USA) was also used to organize all data obtained during the experiment. Additionally, this study employed the "Scikit-learn" package to develop the SVM and PCA model [35] and "NumPy" for numerical computations. Meanwhile, the "Pandas" package was used to read and save data. All computational processes were executed in an Acer Aspire 5 (11th Gen Intel(R) Core (TM) i3-1115G4 @ 3.00 GHz 3.00 GHz, RAM 8.00 GB) notebook. Overall, the entire process of this study is illustrated in **Figure 4**.

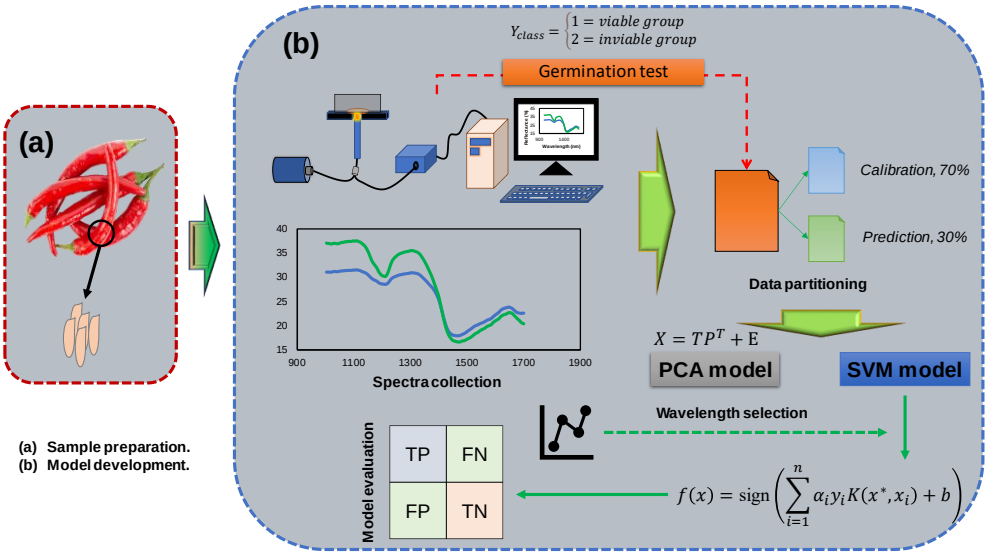

Figure 4 Schematic diagram of nondestructive CP seed viability discrimination using Vis/NIR spectroscopy and multivariate modelling.

Results and discussion

Spectral analysis

During our measurement, hundreds of spectra were obtained from both Vis/NIR and SWNIR regions, as detailed in the subsection **Vis/NIR spectroscopy: Instrumentation and data collection**. Similar patterns can be observed for both spectra regions' viable and inviable groups. For a straightforward interpretation, an averaged spectrum of each group is presented in **Figure 5**, showing spectral differences between viable and inviable CP seeds. In addition, our study found that viable CP seeds were less likely to reflect the light from the illumination source than those of inviable seeds. The influence of heat treatment during the seed inactivation was remarkably responsible for these findings, making our results contradictory to the conclusions from Yasmin *et al*. [36]. Nevertheless, our findings on spectra patterns were consistent with those of previous studies [20,37,38].

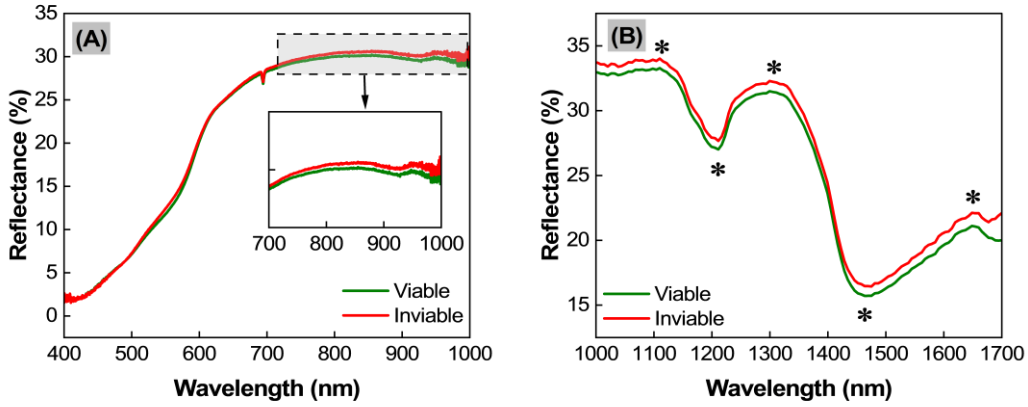

Figure 5 Averaged CP spectrum at (a) Vis/NIR and (b) SWNIR regions. The inset in **Figure 1(a)** visualizes the zoom in the spectrum at 700 to 1000 nm, and the asterisk in **Figure 1(b)** highlights the significant peaks occurring in the SWNIR region positioned at 1109, 1210, 1305, 1463, and 1654 nm.

Overlapped spectra of viable and inviable CP seeds at approximately 400 - 550 nm were recognized and started to separate at the NIR region (~700 - 1000 nm). As these wavelengths (400 - 550 nm) were responsible for the visual appearance (color) of viable and inviable seeds, no peculiar color was observed between the feasible and inviable CP seeds (data not shown). The higher reflectance of inviable seed at 500 - 600 nm was perhaps due to chlorophyll *a* and other plant pigment degradation available in CP seeds [39]. Moreover, significant variations in the SWNIR area were distinguishable due to complex chemical responses in the CP seed matrix, such as X-H molecules.

Proximate content changes of inviable CP seeds, such as water, protein, fat, and carbohydrates, were responsible for the lower absorption. These macromolecules were the main reason for the occurrence of peaks and valleys at 1109, 1210, 1305, 1463, and 1654 nm. Several environmental factors, such as temperature, moisture, storage time, fungal infection, and many more, hold important factors for the seed quality [40]. X-ray imaging clearly distinguished that internal damage was found in the inviable watermelon seed [39] (**Figure 8**). Further, chemical alterations might occur in inviable seeds elsewhere cited [41], which is reasonable for the lower absorption at the wavelengths above—in our case. These findings suggest that improper storage conditions lead to this poor seed quality. Unfortunately, no chemical data was available to support this information related to the proximate analysis of viable and inviable CP seeds. Thus, future studies should provide chemical information to investigate the spectral data in depth. Notably, due to various findings from previous studies about the spectral patterns between the viable and inviable seeds—as we have mentioned earlier, it is difficult for us to suggest or decide whether the CP seed is viable or inviable. Thereby, hereinafter, we will demonstrate the ability of several machine learnings to provide accurate decisions on CP seed viability discrimination.

PCA result

Unsupervised machine learning, namely PCA, was used for qualitative study. The PCA decreases and transforms the high-dimensional spectra matrix into new variables known as principal components (PC) [20]. Previous studies used PCA to cluster the NIR spectra information from fresh-cut vegetables and foreign materials, wherein PCA could explain up to 96.51 % by the 1st 3 PCs [34]. PCA was also applicable for Arabica coffee authentication based on Vis/NIR and SWNIR, with > 95 % explaining variables using the first 2 PCs [42]. Wakholi *et al.* [20] found a clear separation between the treated and non-treated white corn seeds based on SWIR hyperspectral imaging with 99.70 % of the total variables explained by PC1, PC2, and PC3. In conjunction with LDA, Vis/NIR spectroscopy could attain up to 86 % accuracy in discriminating against the maturity of cocoa fruits [14] and contaminated rice with paraffin with 90.10 % accuracy [43].

.

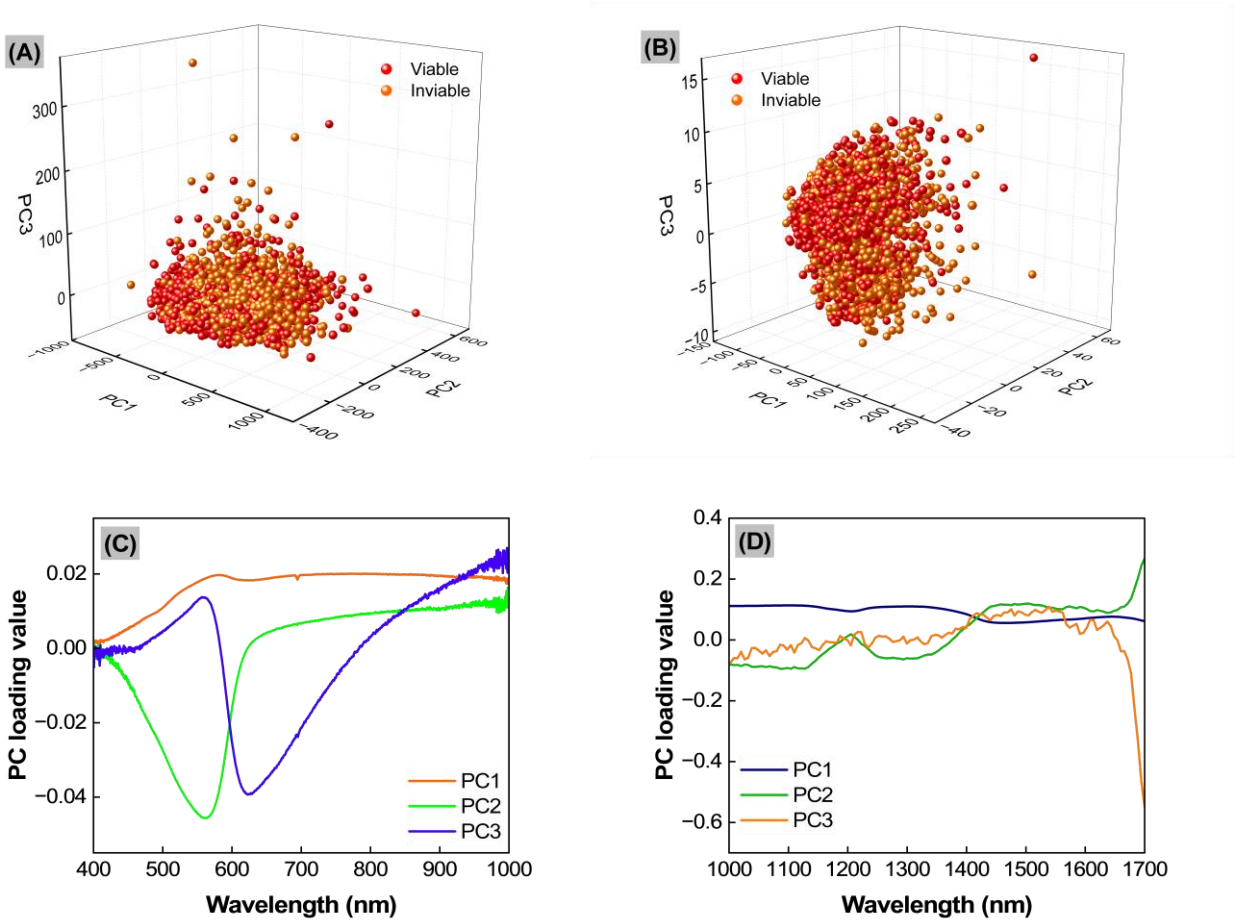

Figure 6 PCA results of the viable and inviable CP seed groups in different regions: (a) and (c) for the PCA score and PCA loading plot generated from Vis/NIR bands, whereas (b) and (d) show the PCA score and PCA loading plots based on SWNIR bands.

Figure 6 illustrates the PCA results in this current study: Score and loading plot. Considering previous studies, PCA was hypothesized to find a distinct cluster between the viable and inviable CP seed based on Vis/NIR and SWNIR spectra. After subjecting our spectra matrices, the 1st three components, which achieved beyond 99 % of the total explained variables by Vis/NIR and SWNIR spectra, were obtained. Interestingly, the PCA failed to cluster the seed's viability using Vis/NIR or SWNIR spectra. Nonetheless, the results provided clear information about the insignificance spectra characteristics between the two groups [20], disabling the PCA in creating two distinct clusters.

PCA loadings along the Vis/NIR and SWNIR wavelengths showed little information in specific bands. Two major wavelengths were observed in PC loading, namely, 562 and 624 nm, where these wavelengths contributed to chlorophyll *a* and the C-O vibration [39]. Conversely, PC loading variations were noticed in the SWNIR region, such as 1128, 1317, 1443, and 1639 nm. These SWNIR wavelengths were associated with complex chemical compositions found in CP seeds, particularly the X-H bonds [13].

SVM result using full wavelength

The advancement of SVM is the ability to convert data into high dimensionality and more flexibility to differentiate the class [20]. In this case, it is CP seed viability. Similar to the PLS algorithm and other machine learning techniques, SVM requires at least two hyperparameters (i.e., cost (C) and kernel parameters (γ)), as mentioned in previous sections of this manuscript. A grid search was performed with a tenfold cross-validation for model calibration efficiency to fulfill the tuning parameters. In contrast, our previous study assembled full cross-validation to determine the optimum latent variables in PLS [13]. We initially determined the C ranging from 10^{-2} to 10^{6}, whereas 10^{-9} to 10^{3} for the γ parameter. The optimal hyperparameters during the calibration were then used to validate the trained SVM model and evaluated with accuracy, sensitivity, and specificity. The results of the SVM model in detecting CP seed viability are listed in **Table 1**, respectively.

Table 2 SVM classifier performance using the full wavelengths for CP seed viability discrimination.

Spectra	Variety	Calibration			Prediction		
		Acc	Sn	Sp	Acc	Sn	Sp
Vis/NIR	*Capsicum annuum var*. Grossum	98.93	100	97.91	98.33	100	96.75
	Capsicum annuum L. *var*. Longum	96.07	99.62	92.95	97.92	99.11	96.88
	Capsicum frutescens L.	99.82	100	99.65	97.08	98.31	95.9
	Mixed	97.98	99.51	96.55	97.78	98.56	97.05
SWNIR	*Capsicum annuum var*. Grossum	98.57	100	97.21	98.75	100	97.56
	Capsicum annuum L. *var*. Longum	96.25	100	92.95	97.92	100	96.09
	Capsicum frutescens L.	98.04	98.19	97.89	94.58	92.37	96.72
	Mixed	97.38	98.4	96.43	95.97	97.12	94.91

Note: Acc, Sn, and Sp stand for accuracy, sensitivity, and specificity, respectively. All metrics are reported in percent.

In this study, the SVM model was developed for 3 different CP varieties for a single model. Furthermore, we also aimed to provide a general model by mixing those 3 CP varieties. Each model was evaluated by calculating accuracy, sensitivity, and specificity. Accuracy quantifies the effectiveness of SVM in correctly determining viable and inviable CP seeds. Simultaneously, sensitivity and specificity describe the proportion of the predicted true positive (viable) and true negative (inviable) correctly classified by SVM. Overall, SVM (Vis/NIR and SWNIR) showed an outperformance in discriminating the CP seeds according to viability, indicated by > 96 % accuracy, sensitivity, and specificity.

In general, using Vis/NIR and SWNIR spectra, the SVM model for *Capsicum annum* var. Grossum exhibited higher accuracy than those of *Capsicum annuum* L. var. Longum, *Capsicum frutescens* L., and mixed varieties, owing to 98.33 % of accuracy (Vis/NIR) and 98.75 % of accuracy by SWNIR spectra. However, the performance of the mixed model was close to that of other varieties. Compared to the two

spectra regions, the mixed model by Vis/NIR spectra yielded better performance than SWNIR spectra, indicated by 97.78 % accuracy, 98.56 % sensitivity, and 97.05 % specificity. This study result was contrary to our previous study, in which SWNIR likely performs better than Vis/NIR [13]. Otherwise, our finding was consistent with the results from Zhang *et al*. [39], where PLSR with Vis/NIR strongly predicted the germination percentage, germination energy, and simple vigor index in wheat seeds. This event was caused by the stronger Vis/NIR relationship with seed characteristics than in SWNIR. One logical reason is the plant pigment content, such as chlorophyll, relatable to the viability of the seed, where it can be detected in the Vis/NIR region [44].

The mixed model revealed consistent outcomes in discriminating CP seeds according to viability from these investigations. Thus, the SVM model with mixed varieties could be promoted as a good candidate for the CP seed viability model. A previous study found similar results, where a mixed model yielded $R^2 = 0.92$ and SEP = 0.15 mg/g to other groups in predicting anthocyanin in black rice using near-infrared hyperspectral imaging [45]. Another study reported that SVM could achieve up to 100 % accuracy in detecting corn seed viability using an ANN-line near-infrared hyperspectral imaging system [20]. Furthermore, SVM, combined with a near-infrared sensor, can effectively detect the wrong powder in simultaneous food production with an accuracy of 91.68 to 99.52 % [46]. Detecting soybean seed viability using FT-NIR spectroscopy was highly accurate, up to 100 %, using the PLS-DA model [16]. Our proposed model was comparable to the previous works with an accuracy of 97.78 %.

SVM result using selected wavelengths

In fact, despite the benefits of using SVM, calibrating the model is relatively costly. One might reason that SVM requires at least two hyperparameters necessary for modelers to find the perfect combination of these parameters. Furthermore, the massive size of variables (wavelengths) is the main issue in the spectroscopy field. For instance, this study extracted 3182 wavelengths for Vis/NIR and 120 SWNIR spectra. Our analysis revealed that controlling the variable size is easier than using wavelength selection algorithms for SVM parameters. The results of the SVM using effective wavebands are summarized in **Table 2**, respectively.

Table 2 SVM classifier performance using the selected wavelengths for CP seed viability discrimination.

Spectra	Variety	Accuracy (%)		
		Full	**VIP**	**bPLS**
Vis/NIR	*Capsicum annum* var. Grossum	98.33	97.92	97.92
	Capsicum annuum L. var. Longum	97.92	98.33	92.5
	Capsicum frustescens L.	97.08	97.5	88.75
	Mixed**	97.78	97.22	95.14
SWNIR	*Capsicum annum* var. Grossum	98.75	98.33	98.33
	Capsicum annuum L. var. Longum	97.92	97.5	94.58
	Capsicum frustescens L.	94.58	93.33	97.08
	Mixed	95.97	94.44	91.94

Note: Double asterisks indicate the optimum model for CP seed viability discrimination.

First, the VIP score was calculated for each wavelength to consider the selection in Eq. (4). Then, to extract the effective variables, wavelengths with a VIP score of < 1.0 were excluded and regarded as a variable with a low contribution [13-47]. The strategy of choosing the threshold followed that of the previous

study [47]. Our study found that the VIP method reduced the number of wavelengths in Vis/NIR (69.70 to 73.19 % reduction) and SWNIR (59.17 to 66.67 % reduction). Second, decision parameters for selecting effective wavebands using the bPLS method were based on the RMSE for every trial and error from the combined wavelengths. In bPLS, this algorithm could reduce the number of Vis/NIR variables by 98.65 %. Meanwhile, the effectiveness of bPLS and VIP slightly differed, producing a 66.67 % reduction when using SWNIR spectra.

Regarding model performance, SVM with effective wavebands demonstrated a slightly weaker accuracy in discriminating against the viability of our CP seeds (**Table 2**), though it is still acceptable. Poor accuracy was somewhat reasonable as the variables were significantly reduced. In **Table 2**, prediction accuracies varied; nonetheless, the performance of SVM with Vis/NIR spectra remained constant, which was higher than using SWNIR spectra. Compared with those variable selection methods, VIP achieved greater accuracy than bPLS. Moreover, in VIP-SVM using Vis/NIR, the accuracy of mixed varieties was close to the individual variety model, which was 97.22 %. These variable selection methods differed in the procedure during the search process. Fernández *et al.* [33] classified variable selection techniques into filter and wrapper. The filter method, i.e., VIP, indirectly selected the wavelengths; if the VIP score was lower than the threshold, the unimportant wavelengths were eliminated. Conversely, bPLS involves a learning algorithm, such as PLS, to evaluate the subset of variables. Our study showed significant results between VIP and bPLS: 1) the number of selected wavelengths and 2) the prediction accuracy [33]. Considering the number of important wavelengths with the accuracy, the mixed Vis/NIR-VIP-SVM model is eagerly suggested, considering the highest accuracy and simplified model, to confirm the CP seed viability. The representative of selected wavelengths from the optimum model—mixed varieties using Vis/NIR— is presented in **Figure 7**.

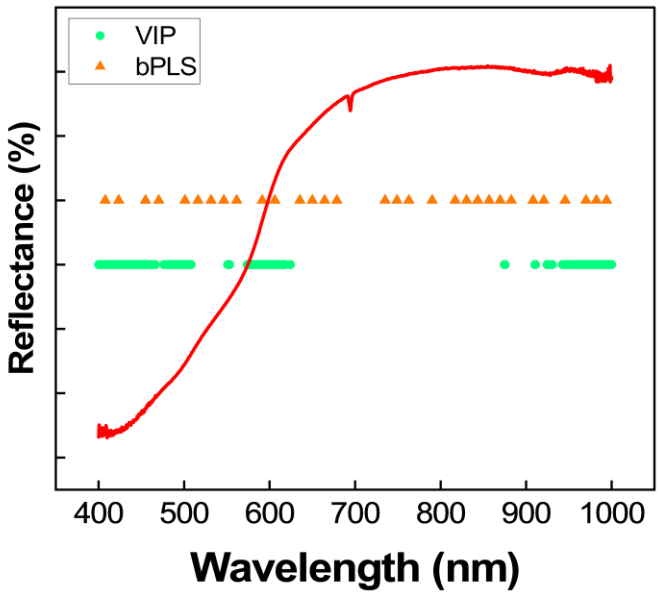

Figure 7 Representation of full Vis/NIR spectrum (red line) and selected wavelengths using VIP and bPLS algorithms (shown in circle and triangle plots).

Limitations and future efforts

Our in-depth investigation has opened the possibility for the future of the CP seed viability discrimination based on SVM combined with Vis/NIR and SWNIR spectroscopy with selected wavelengths. Since our study is the first upon publication, several limitations occur. Our research used a k-fold cross-validation grid search method for SVM hyperparameter

tuning, which provides an exhaustive and easy approach [48]. However, computational efficiency was the main limitation, although the variables have been reduced using variable selection methods, as discussed in this paper. Other optimization methods, encompassing the Bayesian optimizer, particle-swarm optimizer, genetic algorithm, etc., are available to address the limitations of the grid-search method, as in Yu *et al*. [21]. The laboratory scale spectrometer was used to measure the CP spectra and develop the model. However, a portable Vis/NIR and SWNIR sensor could be explored more for practical applications, such as in the seed industry. Furthermore, the feasibility of online hyperspectral imaging should be investigated because of its use, but it is still limited in the case of CP seeds.

Conclusions

Finally, the performance of several machine learning techniques in conjunction with spectral data, implementing PCA and SVM for CP seed viability discrimination using individual and generalized calibration models, have been demonstrated. Our study revealed that the similar spectral characteristics between viable and inviable seeds, both in Vis/NIR and SWNIR regions, lead to the inability of PCA to establish a clear separation. The ability of SVM resulted in an optimum performance based on mixed CP seed varieties through Vis/NIR spectra with 97.78 % accuracy, 98.56 % sensitivity, and 97.05 % specificity. Furthermore, VIP and bPLS significantly decreased the wavelengths with the prediction accuracy of 97.22% and 95.14 %, respectively. Considering the simplified model while serving good accuracy, we suggest the utilization of generalized calibration Vis/NIR-VIP-SVM for effective and nondestructive CP seed viability discrimination. We also hope our proposed model can be used in real applications for chili seed industries, particularly in Indonesia. Furthermore, addressing the limitations listed in the paragraph above makes it possible to enhance the quality of the results and provide practical aspects for real-time use.

Acknowledgments

This work was supported by Universitas Gadjah Mada through the Academic Excellence B program with the grant number 6529/UN1.P1/PT.01.03/2024.

References

[1] A Azlan, S Sultana, CS Huei and MR Razman. Antioxidant, anti-obesity, nutritional and other beneficial effects of different chili pepper: A review. *Molecules* 2022; **27(3)**, 898.

[2] BK Saleh, A Omer and B Teweldemedhin. Medicinal uses and health benefits of chili pepper (*Capsicum* spp.): A review. *MOJ Food Processing and Technology* 2018; **6(4)**, 325-328.

[3] JT Sawma and CL Mohler. Evaluating seed viability by an unimbibed seed crush test in comparison with the tetrazolium test. *Weed Technology* 2002; **16(4)**, 781-786.

[4] A Ambrose, S Lohumi, WH Lee and BK Cho. Comparative nondestructive measurement of corn seed viability using Fourier transform near-infrared (FT-NIR) and Raman spectroscopy. *Sensors and Actuators B: Chemical* 2016; **224**, 500-506.

[5] F Corbineau. The effects of storage conditions on seed deterioration and ageing: How to improve seed longevity. *Seeds* 2024; **3(1)**, 56-75.

[6] LEPD Guzman, OB Zamora, TH Borromeo, PCS Cruz and TC Mendoza. Seed viability and vigor testing of *Jatropha curcas* L. *Philippine Journal of Crop Science* 2011; **36(3)**, 10-18.

[7] CRDS Grzybowski, ODC Ohlson, RCD Silva and M Panobianco. Viability of barley seeds by the tetrazolium test. *Revista Brasileira de Sementes* 2012; **34(1)**, 47-54.

[8] SMC Carvalho, SB Torres, EC Sousa, DMM Sousa, KTO Pereira, EPD Paiva, JR Matias and BRVD Santos. Viability of *Carica papaya* L. seeds by the tetrazolium test. *Journal of Agricultural Science* 2018; **10(2)**, 335-340.

[9] S Lohumi, S Lee, H Lee and BK Cho. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. *Trends in Food Science and Technology* 2015; **46(1)**, 85-98.

[10] RAP Hernanda, J Lee and H Lee. Spectroscopy imaging techniques as *in vivo* analytical tools to detect plant traits. *Applied Sciences* 2023; **13(18)**, 10420.

[11] J Ryu, S Wi and H Lee. Snapshot-based multispectral imaging for heat stress detection in

southern-type garlic. *Applied Sciences* 2023; **13(14)**, 8133.

[12] RAP Hernanda, J Kim, MA Faqeerzada, HZ Amanah, BK Cho, MS Kim, I Baek and H Lee. Rapid and noncontact identification of soybean flour in edible insect using NIR spectral imager: A case study in *Protaetia brevitarsis seulensis* powder. *Food Control* 2025; **169**, 111019.

[13] HZ Amanah, S Rahayoe, E Harmayani, RAP Hernanda, Khoirunnisaa, AS Rohmat and H Lee. Construction of a sustainable model to predict the moisture content of porang powder (*Amorphophallus oncophyllus*) based on pointed-scan visible near-infrared spectroscopy. *Open Agriculture* 2024; **9(1)**, 20220268.

[14] R Listanti, RE Masithoh, AD Saputro and HZ Amanah. Identification of maturity stage of cacao using visible near infrared (Vis-NIR) and shortwave near infrared (SW-NIR) reflectance spectroscopy. *In*: Proceedings of the 4th International Conference on Smart and Innovative Agriculture, Yogyakarta, Indonesia. 2023, p. 6003.

[15] H Lee, BK Cho, MS Kim, WH Lee, J Tewari, H Bae, SI Sohn and HY Chi. Prediction of crude protein and oil content of soybeans using Raman spectroscopy. *Sensors and Actuators B: Chemical* 2013; **185**, 694-700.

[16] D Kusumaningrum, H Lee, S Lohumi, C Mo, MS Kim and BK Cho. Non-destructive technique for determining the viability of soybean (Glycine max) seeds using FT-NIR spectroscopy. *Journal of the Science of Food and Agriculture* 2018; **98(5)**, 1734-1742.

[17] G Qiu, E Lü, H Lu, S Xu, F Zeng and Q Shui. Single-kernel FT-NIR spectroscopy for detecting supersweet corn (*Zea mays* L. saccharata sturt) seed viability with multivariate data analysis. *Sensors* 2018; **18(4)**, 1010.

[18] C Peng, L Zhong, L Gao, L Li, L Nie, A Wu, R Huang, W Tian, W Yin, H Wang, Q Miao, Y Zhang and H Zang. Implementation of near-infrared spectroscopy and convolutional neural networks for predicting particle size distribution in fluidized bed granulation. *International Journal of Pharmaceutics* 2024; **655**, 124001.

[19] Y Ding, Y Yan, J Li, X Chen and H Jiang. Classification of tea quality levels using near-infrared spectroscopy based on CLPSO-SVM. *Foods* 2022; **11(11)**, 1658.

[20] C Wakholi, LM Kandpal, H Lee, H Bae, E Park, MS Kim, C Mo, WH Lee and BK Cho. Rapid assessment of corn seed viability using short wave infrared line-scan hyperspectral imaging and chemometrics. *Sensors and Actuators B: Chemical* 2018; **255(1)**, 498-507.

[21] Y Yu, Y Chai, Y Yan, Z Li, Y Huang, L Chen and H Dong. Near-infrared spectroscopy combined with support vector machine for the identification of Tartary buckwheat (*Fagopyrum tataricum* (L.) Gaertn) adulteration using wavelength selection algorithms. *Food Chemistry* 2025; **463(4)**, 141548.

[22] CYE Tachie, D Obiri-Ananey, M Alfaro-Cordoba, NA Tawiah and ANA Aryee. Classification of oils and margarines by FTIR spectroscopy in tandem with machine learning. *Food Chemistry* 2024; **431**, 137077.

[23] A Windarsih, TH Jatmiko, AS Anggraeni and L Rahmawati. Machine learning-assisted FT-IR spectroscopy for identification of pork oil adulteration in tuna fish oil. *Vibrational Spectroscopy* 2024; **134**, 103715.

[24] TA Teklemariam. Raman and mid-infrared spectroscopy coupled with machine-deep learning for adulterant detection in ground turmeric. *Applied Spectroscopy Practica* 2024; **2(2)**, 1-19.

[25] KND Bhavanee, A Krishnamoorthi, HM Rathva, SC Mareguddikar, A Singh, BP Singh, Nageshwar and K Chittibomma. Advancements in genetic engineering for enhanced traits in horticulture crops: A comprehensive review. *Journal of Advances in Biology and Biotechnology* 2024; **27(2)**, 90-110.

[26] M Bhattacharjee, S Meshram, J Dayma, N Pandey, N Abdallah, A Hamwieh, N Fouad and S Acharjee. *Genetic engineering: A powerful tool for crop improvement*. *In*: Frontier technologies for crop improvement. Springer, Singapore, 2024, p. 223-258.

[27] Hernanda RAP, H Lee, J il Cho, G Kim, BK Cho and MS Kim. Current trends in the use of thermal imagery in assessing plant stresses: A review.

Computers and Electronics in Agriculture 2024; **224**, 109227.

[28] YW Seo, CK Ahn, H Lee, E Park, C Mo and BK Cho. Non-destructive sorting techniques for viable pepper (*Capsicum annuum* L.) seeds using Fourier transform near-infrared and Raman spectroscopy. *Journal of Biosystems Engineering* 2016; **41(1)**, 51-59.

[29] B Wen. Seed germination ecology of Alexandra palm (*Archontophoenix alexandrae*) and its implication on invasiveness. *Scientific Reports* 2019; **9(1)**, 4057.

[30] O Devos, C Ruckebusch, A Durand, L Duponchel and JP Huvenne. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems* 2009; **96(1)**, 27-33.

[31] J Sun, A Nirere, KD Dusabe, Z Yuhao and G Adrien. Rapid and nondestructive watermelon (*Citrullus lanatus*) seed viability detection based on visible near-infrared hyperspectral imaging technology and machine learning algorithms. *Journal of Food Science* 2024; **89(7)**, 4403-4418.

[32] S Shrestha, LC Deleuran and R Gislum. Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics. *Journal of Spectral Imaging* 2016; **5(1)**, a1.

[33] JAF Pierna, O Abbas, V Baeten and P Dardenne. A Backward Variable Selection method for PLS regression (BVSPLS). *Analytica Chimica Acta* 2009; **642(1-2)**, 89-93.

[34] SS Tunny, HZ Amanah, MA Faqeerzada, C Wakholi, MS Kim, I Baek and BK Cho. Multispectral wavebands selection for the detection of potential foreign materials in fresh-cut vegetables. *Sensors* 2022; **22(5)**, 1775.

[35] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot and É Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 2011; **12**, 2825-2830.

[36] J Yasmin, MR Ahmed, S Lohumi, C Wakholi, MS Kim and BK Cho. Classification method for viability screening of naturally aged watermelon seeds using FT-NIR spectroscopy. *Sensors* 2019; **19(5)**, 1190.

[37] S Yang, QB Zhu, M Huang and JW Qin. Hyperspectral image-based variety discrimination of maize seeds by using a multi-model strategy coupled with unsupervised joint skewness-based wavelength selection algorithm. *Food Analytical Methods* 2017; **10(2)**, 424-433.

[38] M Tigabu, A Daneshvar, R Jingjing, P Wu, X Ma and PC Odén. Multivariate discriminant analysis of single seed near infrared spectra for sorting dead-filled and viable seeds of three pine species: Does one model fit all species? *Forests* 2019; **10(6)**, 469.

[39] T Zhang, S Fan, Y Xiang, S Zhang, J Wang and Q Sun. Non-destructive analysis of germination percentage, germination energy and simple vigour index on wheat seeds during storage by Vis/NIR and SWIR hyperspectral imaging. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 2020; **239**, 118488.

[40] S Zhu, L Zhou, P Gao, Y Bao, Y He and L Feng. Near-infrared hyperspectral imaging combined with deep learning to identify cotton seed varieties. *Molecules* 2019; **24(18)**, 3268.

[41] SJ Hong, S Park, A Lee, SY Kim, E Kim, CH Lee and G Kim. Nondestructive prediction of pepper seed viability using single and fusion information of hyperspectral and X-ray images. *Sensors and Actuators A: Physical* 2023; **350**, 114151.

[42] A Dharmawan, RE Masithoh and HZ Amanah. Development of PCA-MLP model based on visible and shortwave near infrared spectroscopy for authenticating arabica coffee origins. *Foods* 2023; **12(11)**, 2112.

[43] MA Faqeerzada, T Akter, U Aline, MFR Pahlawan and BK Cho. Application of hyperspectral imaging for rapid and nondestructive detection of paraffine-contaminated rice. *In*: Proceedings of the 4th International Conference on Smart and Innovative Agriculture, Yogyakarta, Indonesia. 2023, p. 1001.

[44] D Ooms and MF Destain. Evaluation of chicory seeds maturity by chlorophyll fluorescence

imaging. *Biosystems Engineering* 2011; **110(2)**, 168-177.

[45] HZ Amanah, C Wakholi, M Perez, MA Faqeerzada, SS Tunny, RE Masithoh, MG Choung, KH Kim, WH Lee and BK Cho. Near-infrared hyperspectral imaging (NIR-HSI) for nondestructive prediction of anthocyanins content in black rice seeds. *Applied Sciences* 2021; **11(11)**, 4841.

[46] S Ozturk, A Bowler, A Rady and NJ Watson. Near-infrared spectroscopy and machine learning for classification of food powders during a continuous process. *Journal of Food Engineering* 2023; **341**, 111339.

[47] IG Chong and CH Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* 2005; **78(1)**, 103-112.

[48] Y Sun, S Ding, Z Zhang and W Jia. An improved grid search algorithm to optimize SVR for prediction. *Soft Computing* 2021; **25(7)**, 5633-5644.