IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 27, 2021, accepted January 11, 2022, date of publication January 14, 2022, date of current version January 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3143818

Rapid Quality Evaluation of Camellia Oleifera Seed Kernel Using a Developed Portable NIR With Optimal Wavelength Selection

AMORNDEJ PUTTIPIPATKAJORN[1] **AND AMORNRIT PUTTIPIPATKAJORN**[2]
[1]Department of Food Engineering, Kasetsart University, Nakhon Pathom 73140, Thailand
[2]Department of Computer Engineering, Kasetsart University, Nakhon Pathom 73140, Thailand

Corresponding author: Amornrit Puttipipatkajorn (fengarp@ku.ac.th)

This work was supported by the Tea Oil Plant and Other Oil Crops Research and Development Center, The Chaipattana Foundation.

ABSTRACT Moisture content is one of the factors measured to evaluate the quality of Camellia oleifera seeds. High quality *C. oleifera* seeds used for trading must have a low moisture content, specifically not more than 15% on a dry basis (db). Moisture content analysis requires a prolonged laboratory investigation so that the development of fast and effective determination methods is helpful. The objective of this paper was to develop a low-cost portable NIR reflectance spectrometer collaborating with an android application for the rapid prediction of the moisture content in *C. oleifera* seeds. To calibrate the prediction model, an effective chemometric algorithm, based on partial least squares regression was established, and models based on wavelength selection algorithms such as backward interval partial least squares (*bi*PLS) and partial least squares coupled with variable importance projection (VIP-PLS) were implemented as an improved version of PLS. Both algorithms (*bi*PLS and VIP-PLS) improved the predictive performance and accuracy of the model. The experimental results showed that the *bi*PLS model with the 1st derivative transformation provided the best prediction for measuring the moisture content of *C. oleifera* seeds with a coefficient of determination (R^2) value of 0.927, standard error of prediction (SEP) of 0.848%db, bias of −0.067%db, function slope of 1.005, and ratio of performance deviation (RPD) of 3.696. Finally, the device was tested according to the ISO 12099:2017(E) standard and confirmed the reliability of the device for in-field use.

INDEX TERMS *C. oleifera* seed, moisture content, portable spectrometer, backward interval partial least squares, variable importance projection.

I. INTRODUCTION

The *C. oleifera* tree grows in tropical and subtropical regions of Asia, especially in southwest China [1]. It has been extensively cultivated at mountainous elevations of 500–1,300 m in northern Thailand. *C. oleifera* is an edible tree oil like olive oil, palm oil, and coconut oil. It is also known as the oriental olive oil because of its composition being similar to olive oil. As a major important economic oilseed crop, *C. oleifera* seed has 40–60% oil content [2] and is rich in unsaturated fatty acids (>90%), especially oleic acid (74–87%) [3] that plays an important role in reducing cholesterol and triglycerides in the blood [4]. *C. oleifera* oil is also excellent for healthy cooking and contains a low amount of saturated fat. With its high quality and healthy properties, *C. oleifera* seed oil has become one of the most popular and expensive edible vegetable oils in global markets.

The moisture content is one of the essential components in determining the quality of *C. oleifera* seeds. The accurate measurement of the moisture content is the main factor in evaluating the quality and trading price. In terms of production technique in oilseed crop industries, the maximum permissible moisture content of *C. oleifera* seeds is not more than 13%wb, because greater values result in higher production costs. Technically, the moisture content of samples should be analyzed using the traditional laboratory method. Some traders have visually evaluated the moisture content based on with uncertified measurement standards with the grading according to the color, size, and observed completeness. This may lead to unfair trading. Although the traditional laboratory methods can provide a correct result, they require time-consuming steps. Because of these

The associate editor coordinating the review of this manuscript and approving it for publication was Norbert Herencsar.

problems, the current research aimed to find a method that could predict the moisture content of *C. oleifera* seeds rapidly and precisely.

Near-infrared spectroscopy (NIRS) is known as a powerful, convenient, and rapid analytical tool and can measure the quality and compositional attributes of many substances [5]. It is performed in the wavelength range 750–2500 nm which is the region between visible and infrared light [6]. NIRS has proven its effectiveness for both qualitative and quantitative analyses [7]. It quantifies organic compounds by the absorption of near-infrared light by the chemical bonds. Due to the strong combination of absorption bands for water at around 1450 nm (O-H), NIRS has been used for the estimation of the moisture content in grains and seeds [8], [9]. In recent years, studies have shown that NIRS is a suitable method for quantifying trace amounts of the moisture content in Camellia gauchowensis Chang seed kernels [10]. The results shown that NIRS predicted accurately with a high coefficient of determination ($R^2 = 0.92$) and a good standard error of prediction ($SEP = 0.29$). Another investigation demonstrated the potential of NIRS to predict the moisture content of groundnut seeds [11] with a high coefficient of determination ($R^2 = 0.99$) and good standard error of prediction ($SEP = 0.55$). Due to its low processing time and simplicity of sample preparation, NIRS has become an extremely important adjunct to the grain and food industries. Traditional laboratory methods such as loss on drying require time-consuming steps including oven heating and drying time in a desiccator to determine the loss of weight. Even newer automated moisture analysis equipment still requires a few minutes of drying time to determine the drying curve. All these techniques require sample preparation. Therefore, the objective of this paper was to develop a portable NIR reflectance spectrometer covering the range 900–1700 nm that could communicate to an application running on a smartphone based on an android operating system. In addition, the study proposed predictive models based on partial least squares regression (PLSR) with different wavelength selection algorithms to evaluate the moisture content of *C. oleifera* seeds and the accuracy of the models was tested according to the ISO 12099:2017(E) standard to confirm the reliability of the developed instrument for in-field use.

II. DESIGN OF PORTABLE NIR REFLECTANCE SPECTROMETER AND MOBILE APPLICATION

In the past few years, some portable NIRs have been developed but have not seen widespread adoption due to their high cost. Therefore, the goal was to develop a low-cost portable NIR spectrometer and to make it small and easy-to-use. The portable NIRS was developed based on the Texas Instruments engine (digital light projector or DLP). It acts as a wavelength selector incorporating a digital mirror device (DMD) and a single point detector using indium-gallium-arsenide (InGaAs). The DMD acts as a scanner. It selects a wavelength at a time and reflects it to the InGaAs detector. The dark scan achieves by covering all mirrors. A reference spectrum was scanned from the white reflectance standard (Spectralon) and stored in device memory for calculating the absorbance value. The device was designed for use in reflectance mode covering the range 900–1700 nm with a resolution of approximately 10 nm and was powered by a battery (Li-Polymer 3.7 V 1800 mAh) with overall dimensions of 10.0 cm × 15.0 cm × 9.0 cm and a weight of approximately 500 g, including the battery. The sample window was equipped for sample placement at the top of the spectrometer and there were four command buttons (a power button, a reset button, a wake button, and a scan button) that were installed on the base of the instrument, as depicted as in Fig. 1.

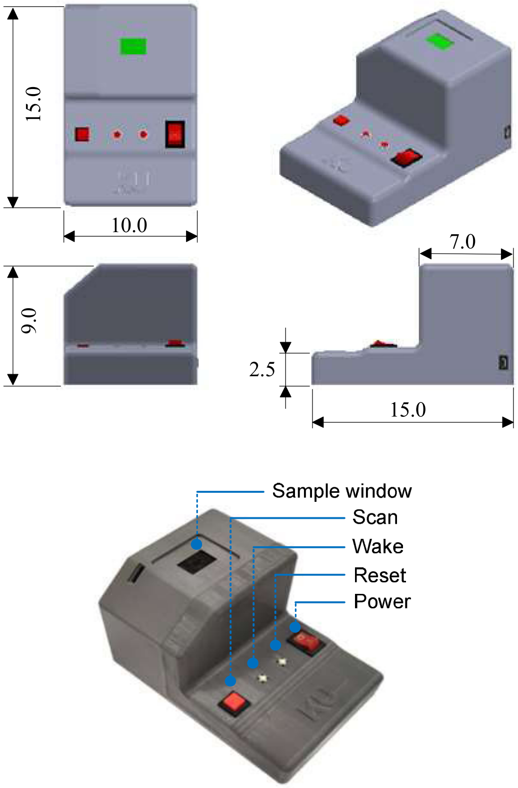


FIGURE 1. Details and dimensions of designed portable NIR spectrometer.

The application on the smartphone was designed for the android operating system using android studio 2.3.3. The development of the application utilized the spectrum library and SDK from KS Technologies, USA. The application was designed to be easy to use with two activity windows. The main window appears automatically when the application is launched and then Bluetooth paring is ready for connection. At first, it makes sure that Bluetooth is turned on in both a smartphone and a spectrometer. After Bluetooth pairing succeeds, the second window appears immediately, and the unit is ready to measure the moisture content of the sample.

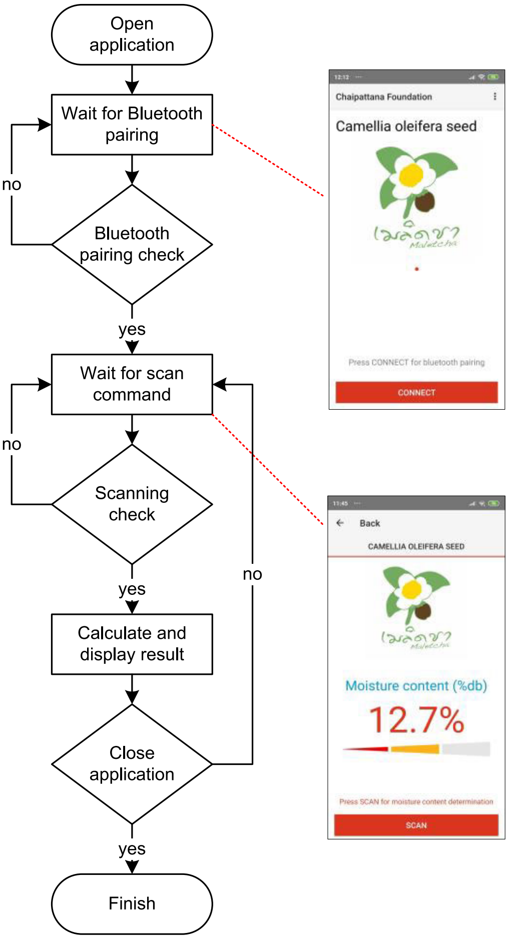


FIGURE 2. Flowchart of android application for moisture content determination in *C. oleifera* seed, with screen display shown at right.

When the SCAN button is pressed, the light source is directed over the sample and the spectrum is recorded. The spectrum is pre-processed, and the measurement result is produced based on the developed predictive model. Subsequently, the result is displayed in the application window. All events can be represented by the flowchart in Fig. 2. However, the predictive model and its parameters do not allow editing after deploying the application on a mobile device.

III. UNITS SAMPLING AND SPECTRAL ACQUISITION

During the harvest season (2020), all *C. oleifera* seeds were provided by the Tea Oil Plant and Other Oil Crops Research and Development Center, The Chaipattana Foundation, Mae Sai District, Chiang Rai province, Thailand. The 184 samples were randomly divided into two independent groups using the ratio 75:25, with 138 seeds for the calibration set used in the model development phase to obtain calibration models and the remainder were used for model testing. Each *C. oleifera* seed was cut in half after peeling and placed on the sample window. The sample spectrum was acquired for a single point, based on the average of 10 readings at an ambient temperature varying from 28 °C to 35 °C. This research did not take account of the effect of temperature difference on the models as the spectrum region of 970 nm and 1400 nm (related to water absorption band) had an insignificant influence on the prediction when the samples were scanned between 30 °C and 35 °C [44]. Then, the spectrum was gathered and saved to the computer. The samples were finally analyzed for their moisture content using the laboratory method according to the AOAC standard [12]. The moisture content on a dry basis was calculated as follows: $M\ (\%db) = \left((W_1 - W)_2 / W_2\right) \times 100$, where W_1 is the weight of the sample before drying and W_2 is the weight of the sample after drying.

IV. CALIBRATION MODEL DEVELOPMENT

PLSR uses a set of algorithms to address problems in econometric path modeling. It has been adopted to solve regression problems in chemometric and spectrometric modeling [13]. The number of PLSR applications has steadily increased in research fields such as bioinformatics, machine learning, and chemometrics. PLSR can cope with datasets containing large amounts of noise, collinear variables, and missing values. PLSR is exploited to linearly correlate an analyte (Y) with its spectral signature (X) by projecting high dimensional data into a few latent variables (LVs) that capture most of the variance in the response variable and maximizing their covariance via a simple and robust method for dealing with multivariate variables [14]. The smallest possible set of latent variables (LVs) must be assigned as a parameter to construct the PLSR model. If too many LVs are used, the model can capture more noise in the model that leads to an over-fitting problem, whereas too few LVs may result in low model performance [15]. Therefore, optimal LVs must be identified to obtain the best model performance. The cross-validation technique is a popular way to implement this task, and the root means-squared error of cross-validation (RMSECV) is used as the criterion for choosing the optimal LVs [16] that produce the lowest RMSECV based on the 5-fold cross-validation method.

Basically, in spectroscopy, each compound in a sample absorbs the light spectrum in different wavelengths. Therefore, certain wavelengths of the spectrum may not strongly relate to a specific compound of interest and consequently, a subset of wavelengths relevant to the target compound may lead to better predictive performance compared to using the complete wavelength. In the current study, the important wavelengths of the spectrum relating to measuring analytes were highlighted using the variable importance projection (VIP) method [17]. The VIP score measures the explicative power of predictor variables regarding the response variable based on the PLS model [18]. The VIP score of the j^{th} variable (the j^{th} bands) is calculated using Equation (1) [19]:

$$v_j = \sqrt{p\sum_{a=1}^{A}\left[SS_a\left(w_{aj}\Big/\|w_a\|^2\right)\right]\Bigg/\sum_{a=1}^{A}SS_a} \tag{1}$$

where SS_a is the sum of squares explained by the a^{th} component and the weight is a measure of the contribution of each variable according to the variance explained by each PLS component and represents the importance of the j^{th} variable. Thus, if v_j is less than some user-defined threshold, v_j can be eliminated. A threshold between 0.83 and 1.21 can yield more relevant variables [18]. The threshold of the VIP score for wavelength selection is generally set to 1 because the mean of the squared scores equals 1 [20].

Interval partial least squares regression (*i*PLS) has been applied to discover the optimal bands for improvement of model prediction [21]. The main idea behind the *i*PLS algorithm is to find only a single interval that provides better prediction instead of using the full spectrum, where the interval width is the major tuning parameter of the *i*PLS model. For efficient variable selection, an improved version of *i*PLS (synergy interval partial least square regression or *si*PLS) has been investigated to allow the combination of intervals to be selected for model building [22]. The tuning parameters for *si*PLS include the interval width, the number of intervals, and LVs to be tuned. Initially, the algorithm splits the whole spectrum into equal-width and non-overlapping sub-intervals. The interval can be either a single wavelength or a window of adjacent wavelengths, where adjacent wavelengths are related to each other. To find the best combination of intervals relevant to the information of the response variable, the algorithm starts by creating local PLS models with identical dimensionality for each predefined wavelength interval. Cross-validation is performed for each of these models and the interval which provides the lowest RMSECV is selected. Additional cycles can be performed by combining the first selected interval with each of the other remaining intervals, one at a time. In the same way, the new local PLS model is created for each combined interval, and the model with the combined interval providing the lowest RMSECV is retained. These processes are repeated until all intervals have been selected. Among all the combined intervals, the one with the lowest RMSECV is considered as the optimized interval for the model [23]. Backward mode (*bi*PLS) can be performed in the opposite manner to the above forward mode procedure [24], [25]. First, all intervals are included in the model. The algorithm discards a single interval which, when excluded, produces a model with the lowest RMSECV. Subsequent cycles discard the next worst interval based on an improved RMSECV value after which that interval is discarded. Finally, the best combination of intervals with the best RMSECV is identified as the optimized interval for the model.

V. EVALUATION OF CALIBRATION MODEL

In this study, the calibration models for the prediction of the moisture content of *C. oleifera* seeds were externally validated using an independent prediction set. The predictive performance [26], [27] of the model was interpreted based on the R^2 bias, standard error of calibration (SEC) and standard error of prediction (SEP), as well as on the ratio of performance deviation (RPD). The RPD is defined as the ratio of the standard deviation of the reference value (SD) to the root mean square error of prediction (RMSEP) [28]. In general, an RPD value less than 2.0 does not provide sufficient prediction, while an RPD value between 2.0 and 3.0 is useful for screening purpose, and an RPD value greater than 3.0 indicates that the model gives good prediction [29], [30]. The SEC and SEP can be calculated using Equation (2) and (3):

$$SEC = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p-1}} \tag{2}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}} \tag{3}$$

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{4}$$

where e_i, $\bar{e}$, p, and n are the residual of the i^{th} sample (difference between reference (y_i) and predicted ($\hat{y}_i$) value), bias (Equation (4)), the number of variables or factors in the model, and the number of samples, respectively.

Furthermore, the accuracy of the developed models was evaluated according to the International Organization for Standardization (ISO) 12099:2017(E) [31] using statistical parameters such as a bias, SEP, and slope. Under this testing, unexplained error confidence limits (T_{UE}) of SEP are determined by performing an F-test (comparison of two variances) using Equation (5). Such calculation relates the SEP with the SEC of the calibration model. If the SEP value is less than the T_{UE}, the SEP is acceptable.

$$T_{UE} = SEC\sqrt{F(\alpha, \vartheta, M)} \tag{5}$$

where $\vartheta = n - 1$ is the degrees of freedom of the numerator associated with the SEP of the prediction set, n is the number of samples in the prediction set, $M = n_c - p - 1$ is the degrees of freedom of the denominator associated with the SEC, n_c is the number of samples in the calibration set, and p is the number of dimensions in the PLS model.

There are many sources of predictive error associated with NIR-based measurement such as variation of sample characteristics, noise, and NIR instrument drift which may explain any bias. Thus, the statistics are applied to verify whether or not the measured bias value is significantly different from 0. The significance of the bias is determined based on a t-test to determine the confidence limits of the bias-value (T_b) (Equation (6)). If the bias value is less than T_b, the bias is not significantly different from 0.

$$T_b = \pm\frac{t_{(1-\alpha/2)}SEP}{\sqrt{n}} \tag{6}$$

where α is the probability of making a type I error, t is the appropriate Student's t-value for a two-tailed test with the degrees of freedom associated with SEP and the selected probability of a type I error, and n is the number of samples.

TABLE 1. Reference measurement of moisture content of *C. oleifera* seeds.

Sample set	Number of samples	Maximum (%db)	Minimum (%db)	Mean±SD
Calibration	138	14.185	1.576	4.481±3.366
Prediction	46	11.927	1.905	5.088±3.134

The slope of a simple regression of the predicted versus the reference value was calculated and compared to the ideal slope (of 1). The t-test was applied to test the hypothesis that the slope value is not significantly different from 1. The observed t-value (Equation (7)) was calculated and compared to the t-value obtained from a table of the t-distribution ($t_{(1-\alpha/2)}$) for a probability of $\alpha = 0.05$. The slope is considered as significantly different from 1 when $t_{obs} \geq t_{(1-\alpha/2)}$.

$$t_{obs} = |b-1|\sqrt{\frac{s^2_{\hat{y}}\,(n-1)}{s^2_{res}}} \tag{7}$$

where b is the slope of the simple regression line, $s^2_{\hat{y}}$ is the variance of the predicted values, n is the number of samples, and s_{res} is the residual standard deviation (Equation (8)), respectively.

$$S_{res} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - a - b\hat{y}_i\right)^2}{n-2}} \tag{8}$$

where a is the interception of the simple regression line, y_i is the i^{th} reference value, $\hat{y}_i$ is the i^{th} predicted value obtained from the model, and n is the number of samples in the prediction set, respectively.

In this study, all data analysis (spectral pretreatments, model developments, evaluations, and data visualization) were implemented using the python 3.7 software which is a very popular and powerful programming language for data analytics. Using this open-source environment, there are many useful libraries that can be applied in the spectroscopic area such as for regression and statistical analysis.

VI. RESULTS AND DISCUSSION

Before building the calibration models, the moisture content of the sample sets (calibration and validation set) was measured using a laboratory method according to the AOAC standard to obtain the reference value. The minimum, maximum, mean, and standard deviation of the samples are shown in Table 1. The statistics for a non-biased prediction set should be within the range of the calibration set with not much difference in the sample variance.

The characteristics of the sample spectrum at different moisture contents are presented in Fig. 4. Since the key chemical compounds of *C. oleifera* seeds consist mainly of water and oil, a large variation in the absorbance value was visible at approximately 930 nm [32], [33] that is related to $-CH_2$ functional groups in the 3rd overtone, at approximately 1210 nm [34], [35] due to the second overtones of C-H and CH = CH- stretching vibrations, at approximately 1652 nm [32], [33] connected to the $-CH_3$ functional group in the first overtone relevant to absorption of oil, and at approximately 960 nm [30]–[36] and 1455 nm [34] relevant to water absorption due to the vibration of O-H bonds in the second and first overtones, respectively. The spectra data had high values of absorbance at 960 nm and 1455 nm, corresponding to the high moisture content in the sample and at approximately 930 nm, 1210 nm and 1652 nm indicating high amounts of oil.

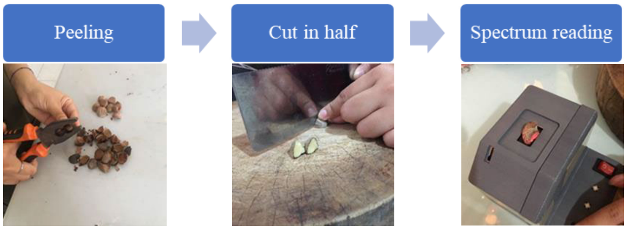


FIGURE 3. Spectral acquisition process using the designed portable spectrometer.

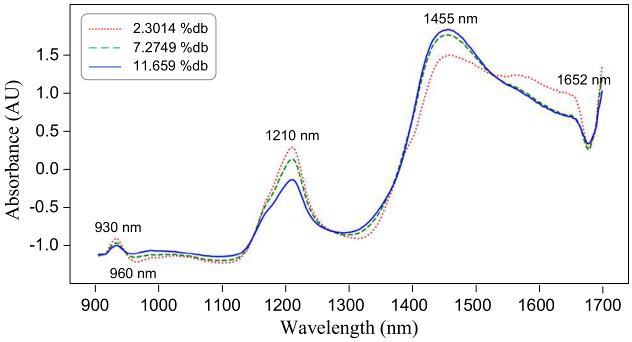
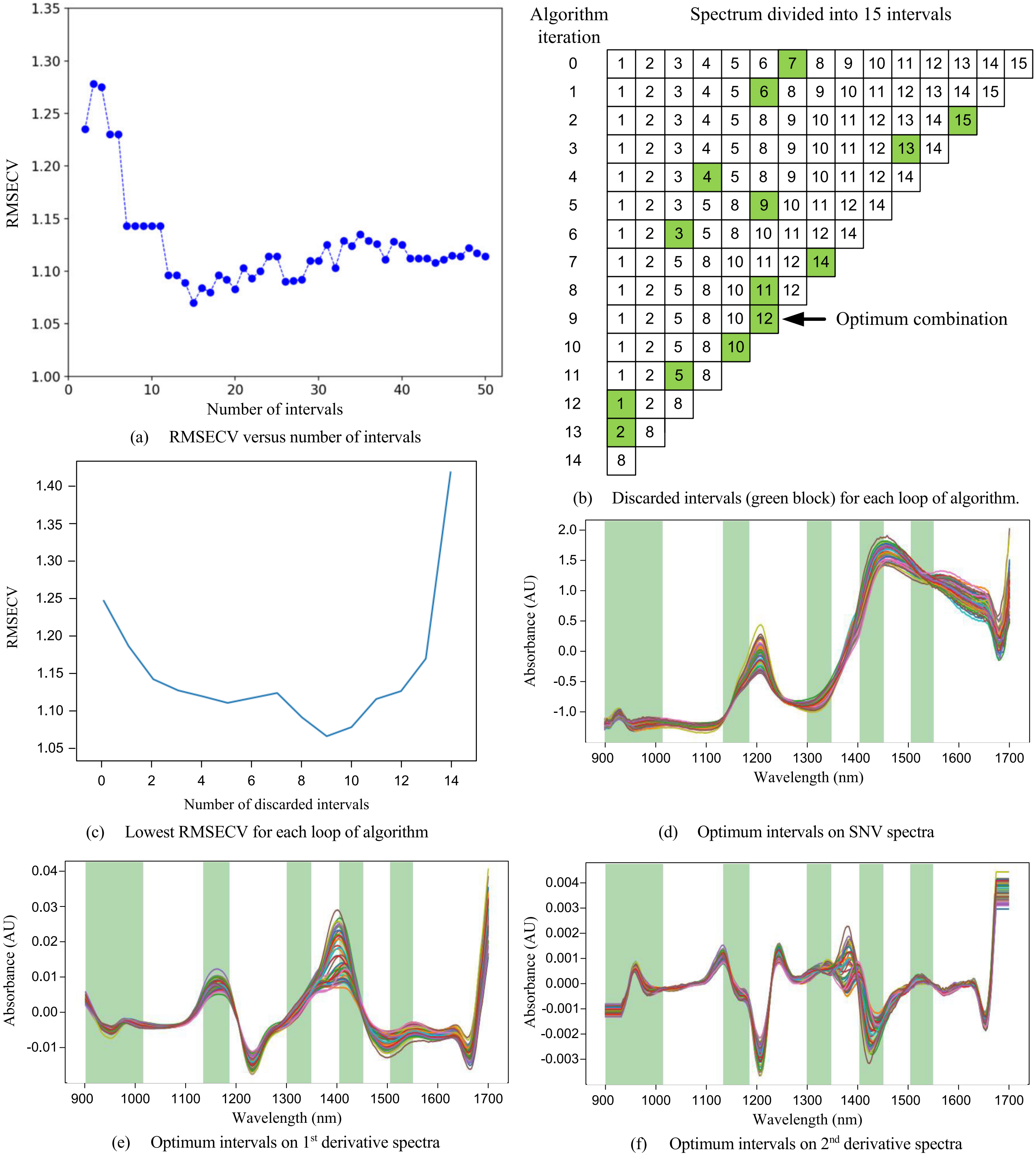


FIGURE 4. Spectra of *C. oleifera* seeds at different moisture contents.

To study the possibility of using spectra to predict the moisture content of *C. oleifera* seeds, a multivariate PLS model was established. As the cutting surface of *C. oleifera* seeds has different degrees of smoothness in each sample, their spectra are transformed to minimize errors from baseline shifts caused by the light scattering phenomenon of the measuring surface. Furthermore, other errors such as additive and multiplicative effects and other nonlinearities frequently associated with the raw spectra should be reduced and this can improve the performance of the models to be created [13]. SNV [37] and MSC [38] are the most frequently used spectral pre-processing techniques for scattering correction. Both methods provide the same results, but SNV is less complicated as regression analysis of a common reference signal is not required. In addition, Savtizky-Golay derivatives [39] capable of removing both additive and multiplicative effects are probably the most popular



(a) RMSECV versus number of intervals

Algorithm iteration	Spectrum divided into 15 intervals														
0	1	2	3	4	5	6	**7**	8	9	10	11	12	13	14	15
1	1	2	3	4	5	**6**	8	9	10	11	12	13	14	15	
2	1	2	3	4	5	8	9	10	11	12	13	14	**15**		
3	1	2	3	4	5	8	9	10	11	12	**13**	14			
4	1	2	3	**4**	5	8	9	10	11	12	14				
5	1	2	3	5	8	**9**	10	11	12	14					
6	1	2	**3**	5	8	10	11	12	14						
7	1	2	5	8	10	11	12	**14**							
8	1	2	5	8	10	**11**	12								
9	1	2	5	8	10	**12**	← Optimum combination								
10	1	2	5	8	**10**										
11	1	2	**5**	8											
12	**1**	2	8												
13	**2**	8													
14	8														

(b) Discarded intervals (green block) for each loop of algorithm.

(c) Lowest RMSECV for each loop of algorithm

(d) Optimum intervals on SNV spectra

(e) Optimum intervals on 1st derivative spectra

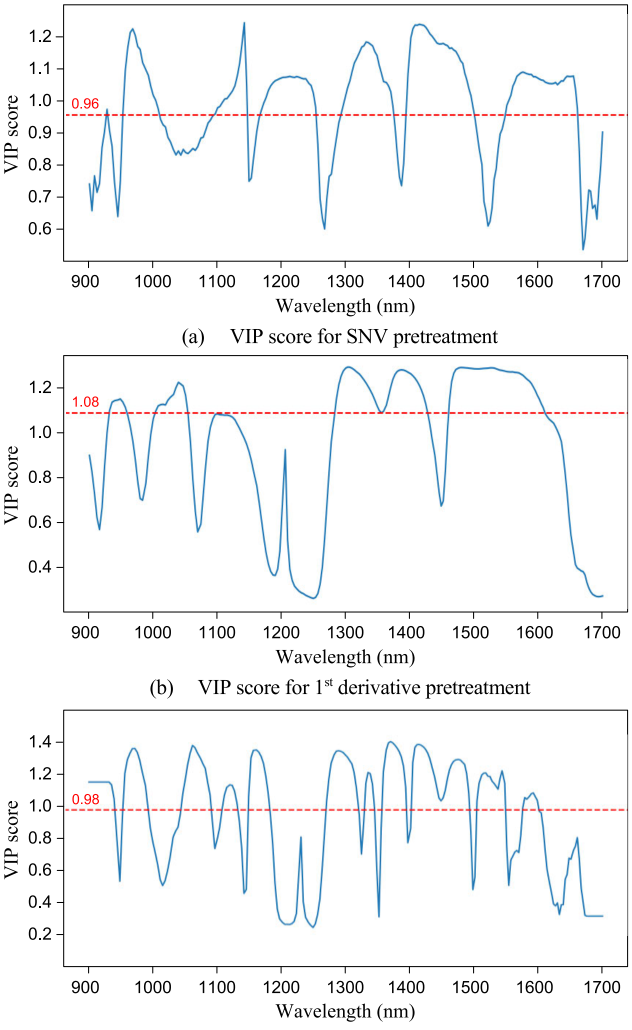
(f) Optimum intervals on 2nd derivative spectra

FIGURE 5. Visualizations of intervals discarded.

method for numerical derivation of a spectrum that includes a smoothing step, the 1st and 2nd derivatives [40]–[42]. Therefore, in this study, three pre-processing methods (SNV, the 1st derivative and the 2nd derivative) were implemented to transform raw spectra before inputting them into the PLS algorithm. The influence of the pre-processing techniques was investigated based on the results of the model testing.

In this study, established not only ordinary PLS models based on the full spectrum, but also models based on a wavelength selection algorithm, namely the *bi*PLS, and the VIP-PLS were built as an improved version of PLS. For each algorithm, three models were constructed using spectral data derived from the three different pre-processing techniques. In the *bi*PLS algorithm, the number of intervals must be selected. The experiment varied the number of intervals (2–50) in single steps. In each round, the whole spectrum was divided into a given interval, then the algorithm with backward mode was performed. In each inner loop of the algorithm, a discarded interval was searched and was removed when it caused the remainder to produce the lowest RMSECV in the cross-validation process. The algorithm proceeded until only one interval remained; the combination of intervals that resulted in the lowest RMSECV was regarded as the optimal combination. The results showed that the RMSECV significantly decreased when the number of intervals was less than 10 and slowly increased thereafter (Fig. 5(a)). The optimal number of intervals was 15, and the intervals consisting of the 1st, 2nd, 5th, 8th, 10th, and 12th intervals were considered as informative wavelengths. The discarded intervals and the RMSECV during the algorithm procedure are depicted in Fig. 5(b) and 5(c). Hence, the bands 900–950 nm, 1030–1090 nm, 1180–1310 nm, 1390–1430 nm, 1510–1630 nm, and 1660–1700 nm were maintained for model construction, and visualized on the SNV, the 1st and 2nd derivative spectra, as shown in Fig. 5(d)–(f), respectively. Among these bands, the bands at approximately 960 nm and 1400–1450 nm [34] corresponding to the absorption of water were preserved, but bands strongly relevant to the oil content such as at approximately 1200 nm [43] were discarded from spectra so that the models built with these bands should improve the predictive performance of the moisture content prediction for *C. oleifera* seeds. The results showed that *bi*PLS models improved the SEP from 0.916%db to 0.848%db and the RPD from 3.422 to 3.696 for the 1st derivative spectral pretreatment.

The wavelength selection based on the VIP algorithm was performed for each spectral pre-processing method, and their VIP scores are shown in Fig. 6. To obtain a suitable threshold for the VIP score, VIP scores of more than 0.8 were tested as a starting score producing a high impact on the model [19]. Thus, for a given score, the wavelength with a score greater than for that score was selected as the wavelength for the model. The wavelengths selected from any given scores based on the minimum RMSECV in the cross-validation procedure was considered as providing the optimal score for that model. As a result, the optimum scores for the models where the spectra were preprocessed using SNV, the 1st derivative, and the 2nd derivative were 0.96, 1.08 and 0.98, respectively, which we close to 1 and consistent with the literature [17], [18], [20]. It was noticed that wavelengths at around 1200 nm that were mainly related to the oil content in *C. oleifera* seeds [42] did not participate in the model for predicting moisture content due to the score being less than 1.08 and 0.98 for the 1st and 2nd derivative spectra, respectively, but they were present in the SNV spectra. Consequently, the models associated with the 1st and 2nd derivative spectra produced better results than for the SNV spectra where the bands around 1200 nm were included. The results showed that the predictive performance of the models with the 1st and 2nd derivative spectra were clearly superior to the SNV spectra based on all criteria, namely R^2 (0.876, 0.919, 0.917), SEP (1.104%db, 0.848%db, 0.991%db), bias (−0.269%db, −0.049%db, −0.045%db), and RPD (2.838, 3.524, 3.461) for SNV, the 1st and 2nd derivative spectra, respectively. Furthermore, the model with the 1st derivative spectra had the best performance, because its selected wavelengths (Fig. 6(b)) contained less bands such as around 930 nm and 1120–1160 nm that were not relevant to water absorption than for the 2nd derivative spectra, as shown in Fig. 6(c). These bands were preserved in the 2nd derivative spectra which have degraded the moisture content prediction.



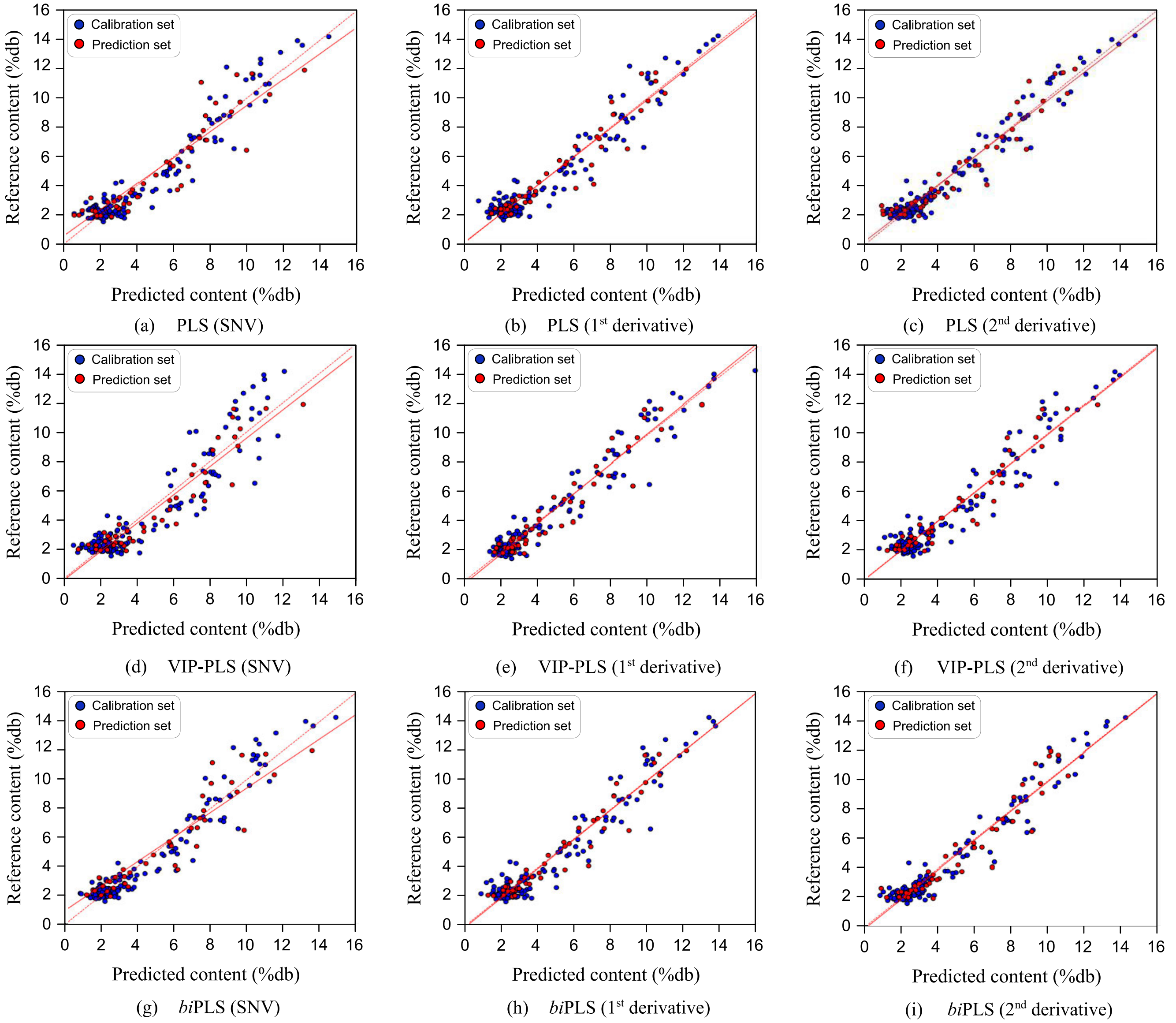
(a) VIP score for SNV pretreatment

(b) VIP score for 1st derivative pretreatment

(c) VIP score for 2nd derivative pretreatment

FIGURE 6. VIP scores of PLS models for different kinds of pretreatment methods.

To compare the predictive performance of the models based on each pre-processing technique used, the statistical parameters, R^2, SEC/SEP, and RPD were calculated [27]. The results in Table 2 indicated that for all algorithms, the model with the 1st derivative spectra provided the best accuracy as they had the lowest SEP for the prediction set with values of 0.916%db, 0.889%db, and 0.848%db for PLS,



(a) PLS (SNV) (b) PLS (1st derivative) (c) PLS (2nd derivative)

(d) VIP-PLS (SNV) (e) VIP-PLS (1st derivative) (f) VIP-PLS (2nd derivative)

(g) *bi*PLS (SNV) (h) *bi*PLS (1st derivative) (i) *bi*PLS (2nd derivative)

FIGURE 7. Scatter plots of reference and predicted moisture contents of PLS (a–c), VIP-PLS (d–f), and *bi*PLS (g–i) models for SNV, 1st derivative and 2nd derivative spectral pretreatments, respectively, where dashed and solid lines represent ideal and predicted slopes, respectively, fitted with the prediction set.

VIP-PLS, and *bi*PLS, respectively. Due to non-smoothness of the measuring surface of the sample, sample spectra were possibly influenced by multiplicative effects and other nonlinearities, which the 1^{st} and 2^{nd} derivative transformations should have reduced from the raw spectra [39]. Thus, the predictive performance of models with the 1^{st} and 2^{nd} derivative spectra was clearly superior to the SNV spectra for all algorithms. Furthermore, the results showed that the models with wavelength selection algorithms (VIP-PLS, *bi*PLS) improved predictive performance (R^2, SEP, RPD) over the ordinary PLS, and the predictive performance of VIP-PLS was better than for *bi*PLS for both SNV and the 2^{nd} derivative pretreatment, but less than for the 1^{st} derivative pretreatment. The slope of the regression line between the reference and predicted values, as shown in Fig. 7, was also enhanced when *bi*PLS and VIP-PLS were applied, especially for the 1^{st} and 2^{nd} derivative pretreatments which had slope values very close to 1 (ideal slope). This made the model more reliable for prediction.

As a result, the *bi*PLS model with the 1^{st} derivative pretreatment (R^2 0.927; SEP 0.848%db; RPD 3.696) was regarded as the best model for moisture content prediction compared to the other proposed models. In addition, the predictive quality of the model was interpreted based on RPD and its value of 3.696 (more than 3) indicated it provided a good estimation of moisture content for use in the classification of *C. oleifera* seeds during the commercial procedure.

TABLE 2. Comparison of predictive performance of ordinary PLS and PLS based on wavelength selection.

Algorithm	Pretreatment	Factor	R^2_C	SEC	R^2_P	SEP	Bias	Slope	RPD
PLS	SNV	11	0.929	0.937	0.827	1.305	0.021	0.893	2.402
	1st	**9**	**0.929**	**0.930**	**0.915**	**0.916**	**-0.037**	**0.985**	**3.422**
	2nd	12	0.955	0.744	0.910	0.939	0.019	0.961	3.337
*bi*PLS	SNV	12	0.932	0.921	0.820	1.331	0.190	0.846	2.335
	1st	**9**	**0.927**	**0.940**	**0.927**	**0.848**	**-0.067**	**1.005**	**3.696**
	2nd	11	0.945	0.825	0.900	0.991	-0.090	1.005	3.162
VIP-PLS	SNV	3	0.857	1.287	0.876	1.104	-0.269	0.968	2.838
	1st	**15**	**0.945**	**0.833**	**0.919**	**0.889**	**-0.049**	**1.027**	**3.524**
	2nd	5	0.922	0.956	0.917	0.906	-0.045	0.995	3.461

TABLE 3. Bias, SEP, and slope testing according to ISO 12099:2017(E).

Algorithm	Pretreatment	T_b	Bias<T_b	T_{UE}	SEP<T_{UE}	t_{obs}	$t(\alpha-\alpha/2)$	$t_{obs}<t(\alpha-\alpha/2)$
PLS	SNV	±0.387	PASS	1.136	FAIL	1.818	2.014	PASS
	1st	±0.272	PASS	1.127	PASS	0.331	2.014	PASS
	2nd	±0.279	PASS	0.902	FAIL	0.856	2.014	PASS
*bi*PLS	SNV	±0.395	PASS	1.117	FAIL	2.893	2.014	FAIL
	1st	±0.252	PASS	1.139	PASS	0.127	2.014	PASS
	2nd	±0.294	PASS	1.000	PASS	0.105	2.014	PASS
VIP-PLS	SNV	±0.328	PASS	1.557	PASS	0.606	2.014	PASS
	1st	±0.264	PASS	1.011	PASS	0.595	2.014	PASS
	2nd	±0.269	PASS	1.158	PASS	0.107	2.014	PASS

The predictive performance of the models was evaluated based on bias, SEP, and slope according to the ISO 12099:2017(E) standard so that the results could be accepted for routine use. The confidence limits of the bias value (T_b), the unexplained error confidence limits (T_{UE}), and the observed t-value (t_{obs}) are shown in Table 3. Since the T_b values or bias limits of all tested models were higher than their bias values, their bias was considered not different from 0 and passed the bias test. The T_{UE} value is the SEP limit and allows for evaluation of the validation of the models. Among the PLS models, only one (1st derivative) had a T_{UE} value higher than its SEP and thus passed the SEP test. For VIP-PLS, all models had a T_{UE} value greater than SEP and thus passed the test. Furthermore, the SNV of *bi*PLS was the only model not passing the SEP test. Finally, only one model (SNV of *bi*PLS) that had an observed t-value greater than the critical t-value; thus, it was evaluated as having a slope different from 1 and so did not pass the test. It should be noticed that there was only one model (1st derivative) for the PLS which completely satisfied the ISO 12099:2017(E) standard, and two models (1st and 2nd derivative) for the *bi*PLS, and all models for VIP-PLS passed the test as well. These results showed that there was a significant improvement in model accuracy when *bi*PLS, VIP-PLS had been applied. In particular, VIP-PLS had the greatest influence.

In the next harvest season (2021), the designed instrument coded with the biPLS algorithm was evaluated using a new batch of 60 samples in the range of 2.290-11.251%db with a standard deviation of 2.142. The predictive performance (R^2 0.839, SEP 0.861%db, and RPD 2.490) slightly decreased compared to the current season (R^2 0.927, SEP 0.848%db, and RPD 3.696), but its performance still complied with the ISO 12099:2017(E) standard, as shown in Fig. 8.

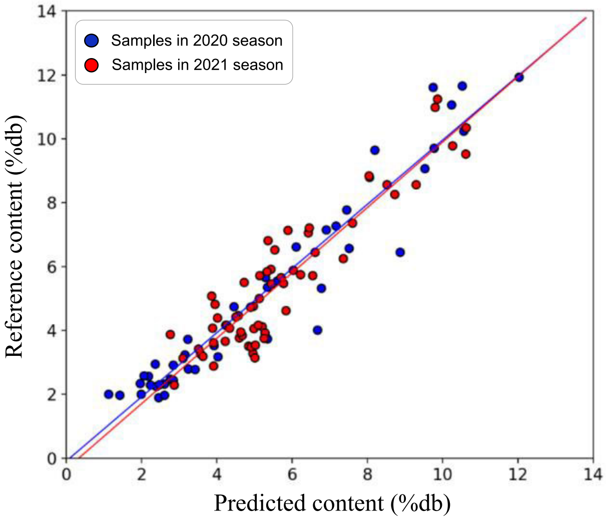


FIGURE 8. Scatter plot of reference and predicted moisture contents for current (2020) and new batch (2021).

In addition to the ISO standard test, the repeatability of the portable NIR was tested. This statistical measure was the proportion of the variation between measurements due to consistent differences between the individuals measured and can be calculated using analysis of variance (ANOVA) using the terminology of Fowler & Cohen in Equation (9):

$$r = \frac{(B - W)}{B + (N - 1)\,W} \tag{9}$$

where N is the number of measurements for an individual and B and W represent the between individual variance and the within individual variance, respectively. A set of 60 samples was used for the repeatability test by with N = 5. The ANOVA generated the value of B = 4.2221 and W = 0.0247. From Equation (9), the repeatability (r) was 0.9174, which was considered very high [45].

VII. CONCLUSION

The experiments indicated that the NIR spectroscopy technique had high potential for the rapid estimation of the moisture content of *C. oleifera* seeds across the full range of anticipated moisture contents because the sample spectrum correlated with the water absorption bands at around 960 nm and 1455 nm. In terms of accuracy, the PLS algorithm was still a good option regarding spectroscopy, because it could predict the moisture content of the *C. oleifera* seeds with accepted results against a recognized standard. In addition, the proposed wavelength selection algorithms (*bi*PLS and VIP-PLS) were able to enhance the predictive performance of the model, especially the *bi*PLS algorithm that generated the best model ($R^2$0.927, SEP 0.848%db, and RPD 3.696). In particular, the 1st derivative spectral pretreatment was the most useful method for improving the predictive performance of all algorithms. The results indicated that the PLS models with wavelength selection algorithms could improve both the predictive performance and accuracy of the model according to the ISO 12099:2017(E) standard. This provided assurance for utilizing the technique for in-field rapid determination of the moisture content of *C. oleifera* seeds. However, the available samples used for training the model had a maximum moisture content of 14.185% dry basis. Thus, the predictive performance may deviate from expectation when the moisture content is higher than 14.185%.

In the next harvest season, The Chaipattana Foundation in Thailand plans to use this instrument in commercial trading of *C. oleifera* seeds in the mountainous Chiang Rai province, Thailand. The developed instrument is compact, lightweight, and has a built-in battery and consequently it will be useful in such areas and will provide acceptable stability, repeatability, and accuracy.

REFERENCES

[1] X. Cheng, T. Yang, Y. Wang, B. Zhou, L. Yan, L. Teng, F. Wang, L. Chen, Y. He, K. Guo, and D. Zhang, "New method for effective identification of adulterated *Camellia* oil basing on *Camellia oleifera*-specific DNA," *Arabian J. Chem.*, vol. 11, no. 6, pp. 815–826, Sep. 2018.

[2] M. H. Su, M. C. Shih, and K. H. Lin, "Chemical composition of seed oils in native Taiwanese *Camellia* species," *Food Chem.*, vol. 156, pp. 369–373, Mar. 2014.

[3] X. Zhou, S. Jiang, D. Zhao, J. Zhang, S. Gu, Z. Pan, and Y. Ding, "Changes in physicochemical properties and protein structure of surimi enhanced with camellia tea oil," *Lebensmittel-Wissenschaft Technol. Food Sci. Technol.*, vol. 84, pp. 562–571, Dec. 2017.

[4] I. Sramala, W. Pinket, P. Pongwan, S. Jarussophon, and K. Kasemwong, "Development of an *in vitro* system to simulate the adsorption of self-emulsifying tea (*Camellia oleifera*) seed oil," *Molecules*, vol. 21, no. 5, pp. 479–487, May 2016, doi: 10.3390/molecules21050479.

[5] J. S. Shenk, I. Landa, M. R. Hoover, and M. O. Westerhaus, "Description and evaluation of a near infrared reflectance spectro-computer for forage and grain analysis," *Crop Sci.*, vol. 21, no. 3, pp. 355–358, May 1981.

[6] C. Pasquini, "Near infrared spectroscopy: A mature analytical technique with new perspectives—A review," *Anal. Chim. Acta*, vol. 1026, pp. 8–36, Oct. 2018.

[7] K. H. Norris, "Design and development of a new moisture meter," *J. Agricult. Eng.*, vol. 45, no. 7, pp. 370–372, 1964.

[8] T. M. Baye, T. C. Pearson, and A. M. Settles, "Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy," *J. Cereal Sci.*, vol. 43, no. 2, pp. 236–243, Mar. 2006.

[9] A. Fassio and D. Cozzolino, "Non-destructive prediction of chemical composition in sunflower seeds by near infrared spectroscopy," *Ind. Crops Prod.*, vol. 20, pp. 321–329, Mar. 2004.

[10] Y. Zhang, L. Zhang, J. Wang, X. Tang, H. Wu, and M. Wang, "Rapid determination of the oil and moisture contents in *Camellia gauchowensis* chang and *Camellia semiserrata* chi seeds kernels by near-infrared reflectance spectroscopy," *Molecules*, vol. 23, no. 9, p. 2332, 2018, doi: 10.3390/molecules23092332.

[11] D. Deshmukh, P. Kona, A. Teggi, and M. Variath, "Rapid measurement of moisture content in groundnut kernels using non-destructive near infrared reflectance spectroscopy," *Int. J. Chem. Stud.*, vol. 6, no. 4, pp. 1686–1689, 2018.

[12] *AOAC-Official Methods of Analysis*, Association of Official Analytical Chemists, Arlington, VA, USA, 1990.

[13] A. Stevens, W. Van, H. Bartholomeus, D. Rosillon, B. Tychon, and E. Ben-Dor, "Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils," *Geoderma*, vol. 144, pp. 395–404, May 2008.

[14] P. Lin, Y. Chen, and Y. He, "Identification of geographical origin of olive oil using visible and near-infrared spectroscopy technique combined with chemometrics," *Food Bioprocess Technol.*, vol. 5, pp. 235–242, Mar. 2012.

[15] X. Yu, H. Lu, and D. Wu, "Development of deep learning method for predicting firmness and soluble solid content of postharvest Korla fragrant pear using Vis/NIR hyperspectral reflectance imaging," *Postharvest Biol. Technol.*, vol. 141, pp. 39–49, Jul. 2018.

[16] S. J. Mazivila, F. B. de Santana, H. Mitsutake, L. C. Gontijo, D. Q. Santos, and W. B. Neto, ''Discrimination of the type of biodiesel/diesel blend (B5) using mid-infrared spectroscopy and PLS-DA,'' *Fuel*, vol. 142, pp. 222–226, Feb. 2015.

[17] R. A. V. Rossel and T. Behrens, ''Using data mining to model and interpret soil diffuse reflectance spectra,'' *Geoderma*, vol. 158, nos. 1–2, pp. 46–54, 2010.

[18] Y. Y. Pu and D. W. Sun, ''Prediction of moisture content uniformity of microwave-vacuum dried mangoes as affected by different shapes using NIR hyperspectral imaging,'' *Innov. Food Sci. Eng. Technol.*, vol. 33, pp. 348–356, Dec. 2016.

[19] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, ''A review of variable selection methods in partial least squares regression,'' *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 62–69, Aug. 2012.

[20] I.-G. Chong and C.-H. Jun, ''Performance of some variable selection methods when multicollinearity is present,'' *Chemometrics Intell. Lab. Syst.*, vol. 78, pp. 103–112, Jul. 2005.

[21] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, ''Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy,'' *Appl. Spectrosc.*, vol. 54, no. 3, pp. 413–419, 2000.

[22] Z. Yang, H. Xiao, L. Zhang, D. Feng, F. Zhang, M. Jiang, Q. Sui, and L. Jia, ''Fast determination of oxides content in cement raw meal using NIR spectroscopy combined with synergy interval partial least square and different preprocessing methods,'' *Measurement*, vol. 149, Jan. 2020, Art. no. 106990, doi: 10.1016/j.measurement.2019.106990.

[23] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin, ''Variable's selection methods in near-infrared spectroscopy,'' *Anal. Chim. Acta*, vol. 667, pp. 14–32, May 2010.

[24] M. Farrokhnia and S. Karimi, ''Variable selection in multivariate calibration based on clustering of variable concept,'' *Anal. Chim. Acta*, vol. 902, pp. 70–81, Dec. 2016.

[25] R. Leardi and L. Nørgaard, ''Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions,'' *J. Chemometrics*, vol. 18, no. 11, pp. 486–497, Nov. 2004.

[26] C. Lin, X. Chen, L. Jian, C. Shi, X. Jin, and G. Zhang, ''Determination of grain protein content by near-infrared spectrometry and multivariate calibration in barley,'' *Food Chem.*, vol. 162, pp. 10–15, 2014.

[27] R. Sheng, W. Cheng, H. Li, S. Ali, A. A. Agyekum, and Q. Chen, ''Model development for soluble solids and lycopene contents of cherry tomato at different temperatures using near-infrared spectroscopy,'' *Postharvest Bio. Technol.*, vol. 156, pp. 110952–110963, Oct. 2019, doi: 10.1016/j.postharvbio.2019.110952.

[28] O. Escuredo, A. Seijo-Rodriguez, M. I. Gonzalez-Martin, M. S. Rodriguez-Flores, and M. C. Seijo, ''Potential of near infrared spectroscopy for predicting the physicochemical properties on potato flesh,'' *Microchem. J.*, vol. 141, pp. 451–457, Dec. 2018.

[29] J. U. Porep, D. R. Kammerer, and R. Carle, ''On-line application of near infrared (NIR) spectroscopy in food production,'' *Trends Food Sci. Technol.*, vol. 46, no. 2, pp. 211–230, Dec. 2015.

[30] M. Ye, Z. Gao, Z. Li, Y. Yuan, and T. Yue, ''Rapid detection of volatile compounds in apple wines using FT-NIR spectroscopy,'' *Food Chem.*, vol. 190, pp. 701–708, Jul. 2016.

[31] *Statistics for Performance Measurement ISO 12099:2017(E): Animal Feeding Stuffs, Cereals and Milled Cereal Products, Guidelines for Application of Near Infrared Spectrometer*, International Organization for Standardization, Cham, Switzerland, 2017, pp. 5–12.

[32] P. Hourant, V. Baeten, M. T. Morales, M. Meurens, and R. Aparicio, ''Oil and fat classification by selected bands of near-infrared spectroscopy,'' *Appl. Spectrosc.*, vol. 54, pp. 1168–1174, Mar. 2000.

[33] A. Riaublanc, D. Bertrand, and E. Dufour, ''Chapitre 6: La spectroscopie infrarouge et ses applications analytiques,'' in *Collection Sciences & Techniques, Agro-Alimentaires*. Paris, France: Lavoisier, 2006, pp. 141–174.

[34] J. A. Cayuela and J. F. García, ''Sorting olive oil based on alpha-tocopherol and total tocopherol content using near-infra-red spectroscopy (NIRS) analysis,'' *J. Food Eng.*, vol. 202, pp. 79–88, Jun. 2017.

[35] J. A. Cayuela and J. F. García, ''Nondestructive measurement of squalene in olive oil by near infrared spectroscopy,'' *LWT*, vol. 88, pp. 103–108, Feb. 2018.

[36] L. Kou, D. Labrie, and P. Chylek, ''Refractive indices of water and ice in the 0.65-to 2.5-μm spectral range,'' *App. Opt.*, vol. 32, pp. 3531–3540, Mar. 1993.

[37] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, ''Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra,'' *Appl. Spectrosc.*, vol. 43, no. 5, pp. 772–777, Jul. 1989.

[38] P. Geladi, D. MacDougall, and H. Martens, ''Linearization and scatter-correction for near-infrared reflectance spectra of meat,'' *Appl. Spectrosc.*, vol. 39, pp. 491–500, Mar. 1985.

[39] A. Savitzky and M. J. E. Golay, ''Smoothing and differentiation of data by simplified least squares procedures,'' *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[40] S. Farres, L. Srata, F. Fethi, and A. Kadaoui, ''Argan oil authentication using visible/near-infrared spectroscopy combined to chemometrics tools,'' *Vib. Spectrosc.*, vol. 102, pp. 79–84, 2019.

[41] R. A. V. Rossel, D. J. J. Walvoort, A. B. McBratney, L. J. Janik, and J. O. Skjemstad, ''Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties,'' *Geoderma*, vol. 131, nos. 1–2, pp. 59–75, Mar. 2006.

[42] S. R. Delwiche and J. B. Reeves, ''A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: Example with Savitzky—Golay filters and partial least squares regression,'' *Appl. Spectrosc.*, vol. 64, no. 1, pp. 73–82, Jan. 2010.

[43] P. A. Picouet, P. Gou, R. Hyypiö, and M. Castellari, ''Implementation of NIR technology for at-line rapid detection of sunflower oil adulterated with mineral oil,'' *J. Food Eng.*, vol. 230, pp. 18–27, Aug. 2018.

[44] D. Cozzolino, L. Liu, W. U. Cynkar, R. G. Dambergs, L. Janik, C. B. Colby, and M. Gishen, ''Effect of temperature variation on the visible and near infrared spectra of wine and the consequences on the partial least square calibrations developed to measure chemical composition,'' *Anal. Chim. Acta*, vol. 588, pp. 224–230, Dec. 2007.

[45] D. G. C. Harper, ''Some comments on the repeatability of measurements,'' *Ringing Migration*, vol. 15, pp. 84–90, Mar. 1994, doi: 10.1080/03078698.1994.9674078.

AMORNDEJ PUTTIPIPATKAJORN received the B.S. degree in food engineering from Kasetsart University at Kamphaeng Saen, Nakhon Pathom, Thailand, in 1997, and the M.S. degree in chemical engineering from Kasetsart University, Bangkok, Thailand, in 2001. He is currently an Assistant Professor with Kasetsart University. His research interests include the spectroscopy and spectral imaging analysis, and its applications in food and agriculture.

AMORNRIT PUTTIPIPATKAJORN was born in Bo Thong, Chonburi, Thailand, in 1975. He received the B.S. degree in electrical engineering from Kasetsart University, Bangkok, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the University of Montpellier II, Montpellier, France, in 2000 and 2005, respectively. He is currently an Assistant Professor with Kasetsart University. His research interests include machine learning and non-destructive techniques in agriculture.