Vol. 14, No. 2 (2025): 515 - 526 http://dx.doi.org/10.23960/jtep-l.v14i2.515-526

JURNAL TEKNIK PERTANIAN LAMPUNG

ISSN 2302-559X (print) / 2549-0818 (online)
Journal homepage : https://jurnal.fp.unila.ac.id/index.php/JTP

Portable Near-Infrared Spectroscopy and Support Vector Regression for Fast Quality Evaluation of Vanilla (*Vanilla planifolia*)

Widyaningrum[1,5], Y. Aris Purwanto[2,✉], Slamet Widodo[2], Supijatno[3], Evi Savitri Iriani[4]

[1] Agricultural Engineering Science Study Program, Department of Mechanical and Biosystems Engineering, IPB University, Bogor, INDONESIA.
[2] Department of Mechanical and Biosystems Engineering, IPB University, Bogor, INDONESIA.
[3] Department of Agronomy and Horticulture, IPB University, Bogor, INDONESIA.
[4] Standardization Agency for Agricultural Instruments – Refreshing and Industrial Crops, Ministry of Agriculture, Sukabumi, INDONESIA.
[5] Agricultural Development Polytechnic of Manokwari, Ministry of Agriculture, Manokwari, INDONESIA.

Article History:

Received : 29 November 2024
Revised : 18 December 2024
Accepted : 02 January 2025

Keywords:

Moisture content,
Portable NIR spectroscopy,
Support vector regression,
Vanilla planifolia,
Vanillin content.

Corresponding Author:
✉ arispurwanto@apps.ipb.ac.id
(Y. Aris Purwanto)

ABSTRACT

Vanilla (Vanilla planifolia) *is a high-value agricultural product, with its quality influenced by essential factors such as moisture and vanillin content. Conventional techniques for evaluating these characteristics are inefficient, require sample destruction, and are impractical for swift assessments. This research explores the feasibility of using portable Near-Infrared (NIR) spectroscopy combined with Support Vector Regression (SVR) to enable quick and noninvasive property prediction. Spectral information was obtained from vanilla samples using two portable NIR instruments, SCiO (740–1070 nm) and Neospectra (1350–2550 nm). Preprocessing techniques such as normalization, SNV, MSC, first derivative, first derivative-SNV, and first derivative-MSC were applied. For moisture content prediction, SCiO achieved an R^2 of 0.768, an RMSE of 4.720%, an RPD of 2.075 and an RER 10.197 using Min-Max normalization, while Neospectra yielded an R^2 of 0.758, an RMSE of 5.161%, an RPD of 2.033 and an RER 9.325 with MSC preprocessing. In contrast, predicting vanillin concentration proved more challenging, with SCiO achieving moderate accuracy with an R^2 0.406, an RMSE 0.379%, an RPD 1.297, an RER 5.039, and Neospectra demonstrating limited performance with an R^2 0.172, an RMSE 0.576%, an RPD 1.098 and an RER 3.315. These findings highlight the potential of portable NIR spectroscopy as a practical tool for assessing vanilla quality, particularly for moisture content, in industrial and field applications.*

1. INTRODUCTION

Vanilla (*Vanilla planifolia*) is one of the high-value agricultural products due to its complex and lengthy production process. Additionally, vanilla is renowned for its distinct flavor and aroma, which come from the primary aromatic component, vanillin (Ranadive, 2019). The global demand for high-quality vanilla continues to rise, driven by its widespread use in various industries, including food, beverages, cosmetics, and pharmaceuticals (Baqueiro-Peña & Guerrero-Beltrán, 2017). Several important factors, including vanillin and moisture content, influence the quality of vanilla. Moisture content significantly impacts the texture, aroma release, and shelf life of vanilla beans. At the same time, vanillin concentration is a key factor in flavor intensity and consumer acceptance, which ultimately affects the grade and price of the vanilla (Ranadive, 2019).

Traditionally, the assessment of moisture content and vanillin concentration has relied on conventional laboratory methods, such as oven-drying (Havkin-Frenkel & Frenkel, 2008), UV spectrophotometry, and high-performance liquid chromatography (HPLC) (Ranadive, 2019). Although these techniques yield accurate results, they are typically time-

consuming, costly, and involve destructive sampling. Moreover, the use of chemicals in these analyses makes them environmentally unfriendly. Additionally, the need for specialized equipment and skilled personnel makes these methods impractical for field applications (Cozzolino, 2016; Bittner *et al*., 2013). This underscores the demand for fast, affordable, and noninvasive analytical tools that can be used directly in the field or industrial environments.

NIR spectroscopy has emerged as a powerful method for assessing the quality characteristics of agricultural commodities (Pandiselvam *et al*., 2022). This method operates by analyzing how near-infrared light interacts with molecular structures in organic compounds, generating spectral data that can be correlated with specific chemical and physical properties (Schwanninger *et al*., 2011; Workman & Weyer, 2007; Zhang *et al*., 2022). However, despite its advantages, NIR has the drawback of having relatively large instruments, making it difficult to perform direct field analysis, as samples still need to be taken to the laboratory for analysis. To address this challenge, scientists have developed portable versions of NIR. Portable NIR devices offer the advantage of on-site measurements, enabling analysis without damaging the sample due to their compact and small size (Beć *et al*., 2021; Huck, 2020; Pu *et al*., 2021). Recently, there has been significant research into the use of portable NIR spectroscopy for effectively assessing the moisture content in apples (Malvandi *et al*., 2022) and mango (Wokadala *et al*., 2020), protein content in cereals (Chadalavada *et al*., 2022), and other key attributes in crops such as cocoa bean (Anyidoho *et al*., 2021) coffee (Correia *et al*., 2018), and other agricultural materials.

A key factor in utilizing NIR spectroscopy is the development of robust predictive models that can convert spectral data into accurate quality predictions (Zareef *et al*., 2020). Machine learning algorithms have shown great potential, particularly Support Vector Regression (SVR). SVR is a machine learning technique designed for regression tasks, particularly effective in handling complex and high-dimensional datasets (Zhang & O'Donnell, 2020), such as those generated by NIR spectroscopy. Unlike traditional regression methods, SVR uses a kernel-based approach to capture nonlinear relationships, making it well-suited to handle complex variability (Wani *et al*., 2024).

This study aims to create a quick and noninvasive method for estimating the moisture and vanillin content of vanilla beans by combining the capabilities of portable NIR spectroscopy with SVR. By integrating spectral preprocessing techniques such as min-max normalization, multiplicative scatter correction (MSC), Standard Normal Variate (SNV), first derivative, and combinations of the first derivative with SNV or MSC, it is expected to improve the quality of the spectral data and enhance model performance. The methodology includes using the Kennard-Stone algorithm for data splitting, ensuring that spectral variability is well-represented in both training and testing datasets while optimizing the SVR parameters through grid search for better prediction accuracy.

2. MATERIALS AND METHODS

2.1. Materials and Instruments

Forty-nine dried vanilla beans (*Vanilla planifolia*) samples were collected from various Indonesian vanilla processing industries. The samples represented a range of quality grades, with vanillin content varying between 2.09% and 0.18% and moisture content ranging from 56.04% to 7.91%. The sample set was categorized into nine samples of Grade II, 16 samples of Grade III, and 24 samples of Grade IV (cuts). The grading system was based on vanilla's Indonesian National Standard (SNI) 01-0010-2002.

This study utilized two portable NIR spectrometers with distinct wavelength ranges. The first, the SCiO spectrometer (Consumer Physics, SF, CA, USA), operates within the 740-1070 nm range and is equipped with an LED lamp, bandpass filters, and a 12-element silicon photodiode array. It provides a spectral resolution of 1 nm and produces 331 data points. The second instrument, the Neospectra spectrometer (Si-Ware, Menlo Park, CA, USA), functions in the 1350-2550 nm range, featuring a tungsten halogen lamp, a MEMS Michelson interferometer, and a single-element InGaAs detector. This device offers a spectral resolution of 9 nm and generates 257 data points.

2.2. Spectra Acquisition

To guarantee consistency, 60 g of vanilla samples were weighed and knotted at both ends. Before measurements, the SCiO and Neospectra instruments were calibrated. Calibration and measurements were conducted through

applications connected via Bluetooth between the instruments and a smartphone. While the Neospectra device was managed by the "Neospectra Collect" application, which is compatible with iOS, the SCiO instrument was controlled by the "The Lab" application, which is compatible with Android and iOS. SCiO calibration was performed by placing the device inside its cover with the optical sensor facing the cover. The SCiO function button was pressed, or 'Calibrate' was selected in The Lab App to initiate the calibration process. Meanwhile, Neospectra calibration was performed by placing the lid on top of the optical window. Then, the "BG/Calibrate" button was pressed in the Neospectra Collect app to begin the calibration process. The instrument could be used after the calibration process is completed. The time required for spectral acquisition with the SCiO ranged from 2 to 5 seconds, while Neospectra requires 4 to 5 seconds. Measurements were conducted at multiple points along each sample to maintain consistency and account for spatial variability. Each sample was measured in three distinct points: the stem end, the center, and the blossom end. A total of 147 spectral data points were collected from each spectrometer. The spectral data from the three points were averaged to enhance accuracy. Measurements were performed at an ambient temperature of approximately 25°C. The spectral data were stored in the cloud and downloaded in .csv format. After obtaining the data in reflectance form, the log (1/R) was applied to convert it to absorbance.

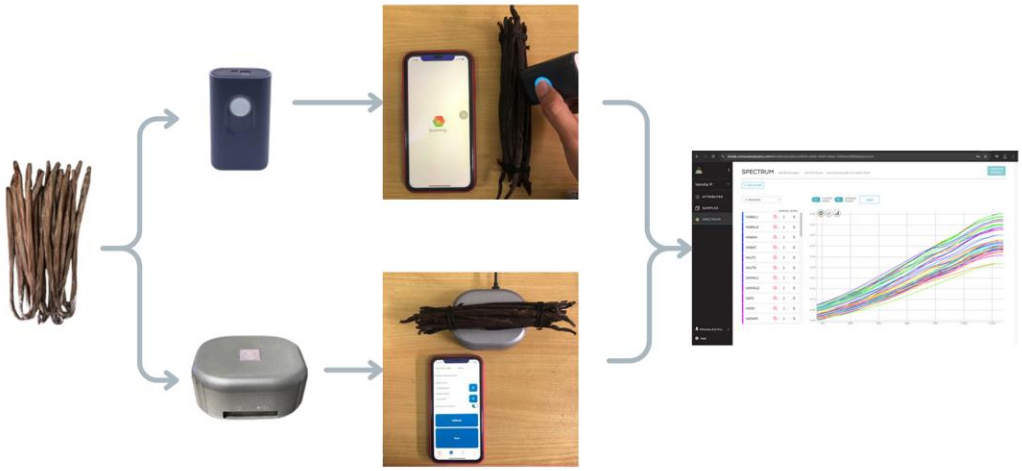

Figure 1. Spectral data acquisition

2.3. Determination of Moisture Content and Vanillin Content

Vanillin and moisture content are measured in compliance with ISO 948:1980. The distillation method determined moisture content, while vanillin content was assessed using UV spectrophotometry. The moisture content of vanilla was calculated using Equation 1, and vanillin content was determined using Equation 2.

$$\text{WC } (\%) = \frac{w}{v} \times 100\% \quad (1)$$

$$\text{VC } (\%) = \frac{C \times 5 \times 100}{M \times 100H} \quad (2)$$

where WC refers to moisture content (%), w represents the weight of the sample (g), v denotes the water volume (mL), VC indicates vanillin content (%), and C signifies the sample solution concentration (expressed in µg/mL or ppm). Using solution C as the blank, the concentration (C) was calculated from the standard curve based on absorbance measurements at a wavelength of 348 nm.

2.4. Data Preprocessing

Improving and enhancing the quality of spectral data is essential before developing predictive models (Hayati *et al*., 2020). Although many studies have investigated different data preprocessing approaches, the most effective method is typically determined through experimentation (Torniainen *et al*., 2020). This study employed several preprocessing

approaches, including min-max normalization, SNV, MSC, first derivative transformation, and combinations of first derivative with SNV or MSC. Each preprocessing method, along with its equation, is explained in Table 1.

Table 1. The preprocessing method and its equation

Pre-processing method	Definition	Equation
Min-Max Normalisation	A data standardization method that scales the lowest and maximum values of each feature to 0 and adjusts other values within the range of 0 to 1 (Raju *et al*., 2020)	$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$ (3) where; X is original value; max(X) is the maximum value in the dataset; min(X) is the minimum value in the dataset
Standard Normal Variate (SNV)	A method used to reduce scattering variability, particularly in backscatter measurements, by centering and scaling each spectrum to correct for light scatter and particle size.	$Z = (X - \mu)/\sigma$ (4) where; X is the original data point; μ is the mean of the spectrum; σ is the SD of the spectrum
Multiplicative Scatter Correction (MSC)	Corrects for scattering effects by adjusting spectra to the same scatter level as the ideal sample, improving consistency and quality.	$X_i msc = \frac{X_i - a_i}{b_i}$ (5) where; X_i is the original value, a_i is the estimates mean spectrum, b_i is the SD of the spectrum.
Derivative	A technique used to address peak overlap and correct baseline drifts in spectral data, improving analytical precision and clarity.	$X = dA/d\lambda$ (6) where; A is the absorbance and λ is the wavelength.

2.5. Support Vector Regression

The Support Vector Machine (SVM) technique was modified for regression tasks and is known as Support Vector Regression (SVR), which makes it possible to predict numerical responses (Rodríguez-Pérez & Bajorath, 2022). SVR is a form of supervised learning frequently applied in regression analysis (Drucker *et al*., 1996). This method is useful for analyzing the connection between a target variable and one or more independent variables. For regression problems, Support Vector Regression (SVR) works well because it maximizes the trade-off between prediction accuracy and model complexity while demonstrating strong performance in handling high-dimensional data (Zhang & O'Donnell, 2020). In contrast to SVM classification, which generates binary outputs (i.e., class labels), Support Vector Regression (SVR) is tailored for regression tasks, facilitating the estimation of real-valued functions (Zhang & O'Donnell, 2020).

Several studies have successfully applied the SVR model for prediction using NIR instruments. Among them is the use of NIR combined with SVR to predict caMP content in red jujube. The results showed that SVR outperformed conventional chemometric models such as PLS, achieving an R^2p of 0.93 and RMSEP of 13.07, while PLS yielded an R^2p of 0.83 and RMSEP of 29.04 (Chen *et al*., 2019). Another research, which assessed the moisture level of black tea during processing, proved the advantage of SVR in predictive model development. The results showed that the SVR model achieved an R^2p of 0.98 and RMSEP of 0.03, while the model built with PLS yielded an R^2p of 0.94 and RMSEP of 0.07 (Zou *et al*., 2022).

2.6. Data Analysis

The data were imported and processed using Python on a Google Colab notebook. The Kennard-Stone approach was used to divide the dataset into 30% testing and 70% training subsets. Table 2 shows the descriptive statistics of the actual measured vanilla data, including vanillin and moisture content. Several preprocessing techniques were applied, including no preprocessing, min-max normalization, SNV, MSC, first derivative, and combinations of first derivative with SNV or MSC. Grid search optimization was then applied to identify the best parameters for the Support Vector Regression (SVR) model. In this study, the hyperparameter settings and their tuning results are shown in Table 3.

Table 2. Reference measurement of vanillin and water content of vanilla samples in datasets

Dataset	Number of samples	Range (%)	Mean(%)	Std Dev(%)
Summary of vanillin content				
Data training	34	0.22 – 2.07	1.04	0.49
Data testing	15	0.18 – 2.09	0.99	0.56
Summary of water content				
Data training	34	7.91 – 48.59	24.11	9.47
Data testing	15	9.65 – 56.04	24.45	14.34

Table 3 Grid search optimization for SVR

Parameter setting	Parameter optimum
'C': {0.1, 1, 10, 100}, 'epsilon': {0.01, 0.1, 0.5}, 'kernel': {'linear', 'rbf'}	{'C': 100, 'epsilon': 0.5, 'kernel': 'linear'}

Furthermore, ten-fold cross-validation (CV) was used to assess the robustness of the model and minimize overfitting. Model performance was evaluated using R^2 (coefficient of determination), RMSEP (root mean square error of prediction), RPD (ratio of prediction to deviation) and RER (range error ratio) metrics. In general, the lower the RMSE and the closer the R^2 is to 1, the more accurate the model predictions will be (Pereira *et al*., 2019). Furthermore, RPD values below 1.0 are considered very poor, while values ranging from 1.0 to 1.4 imply poor predictions. An RPD range of 1.4 to 1.8 indicates reasonable performance, whereas values between 1.8 to 2.0 imply accurate predictions. RPD values between 2.0 to 2.5 indicate very accurate predictions, with RPD values above 2.5 considered excellent. RER values greater than 20 indicate an excellent prediction model. RER values between 15 to 20 mean the model is considered successful. RER values between 10 to 15 classify the model as moderately successful. Meanwhile, RER values between 8 to 10 indicate that the model is still moderately useful (Williams & Norris, 1987). Additionally, for a model to be considered effective, it should ideally exhibit high R^2 and RPD values, alongside a small RMSEP, which together reflect a model's accuracy and reliability in making predictions (Douglas *et al*., 2018). Equations 7, 8, 9 and 10 compute and express the values of R^2, RMSE, RPD and RER.

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)^2}{\Sigma(\hat{y}_i - \bar{y}_i)^2} \quad (7)$$

$$RMSE = \sqrt{\frac{\Sigma(\hat{y}_i - y_i)^2}{n}} \quad (8)$$

$$RPD = \frac{SD}{RMSE} \quad (9)$$

$$RER = \frac{\text{Range of the reference data}}{RMSE} \quad (10)$$

where $\hat{y}_i$ is sample i's real value; y_i is sample i's predicted value; y is mean of the real values; n is quantity of samples; SD is standard deviation.

3. RESULTS AND DISCUSSION

3.1. Spectral Analysis

Vanillin is the major aromatic compound found in vanilla, with the chemical formula $C_8H_8O_3$ (Anand *et al*., 2019). Figure 2 presents the absorbance spectra of vanilla beans obtained through a portable NIR spectrometer. Figure 3(a) illustrates the absorption peaks in the average vanilla spectrum within the wavelength range of 740-1070 nm, obtained using the SCiO instrument. The image illustrates how the vanilla spectra in this range display almost no peaks. On the other hand, Figure 3 (b) shows the average vanilla spectrum produced in the 1350–2550 nm wavelength region using the portable NIR Neospectra devices. It is evident from the graphic that there are many absorption peaks in this range.

Compared to the reference in Practical Near-Infrared Spectroscopy (Osborne *et al*., 1993), the spectral analysis results show an absorption peak at 970 nm due to O–H stretching bond vibration at the second overtone, indicating the presence of H_2O. At 1450 nm, an absorption peak is seen related to the O–H stretching bond vibration at the first overtone, indicating the presence of starch and water. Absorption peaks at 1725 and 1780 nm show C–H stretching bond vibrations in the first overtone region, suggesting the presence of CH_2 compounds and cellulose. The absorption peak at 1940 nm was induced by O–H stretching and deformation bond vibrations, signifying the presence of H_2O.

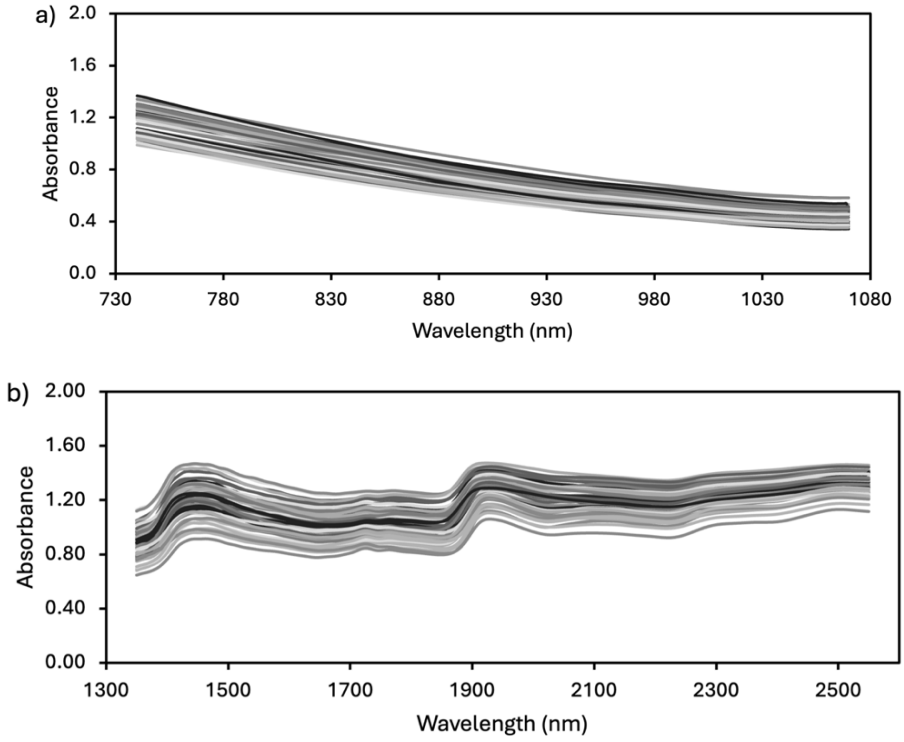

Figure 2. Absorbance spectra of vanilla at wavelengths of 740-1070 nm (a) and 1350-2050 nm (b).

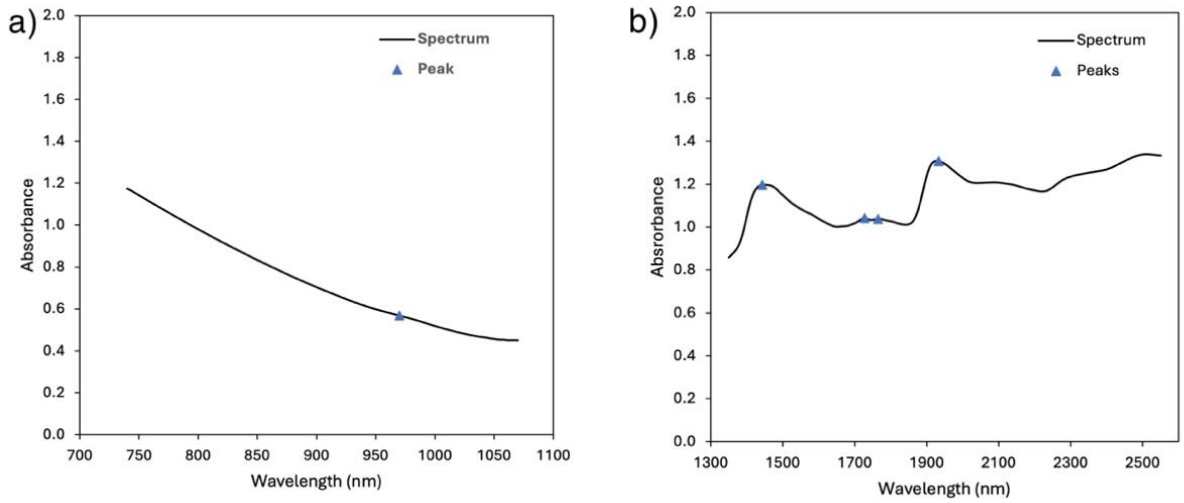

Figure 3. Average spectral of vanilla with identified absorption peaks wavelengths of 740-1070 nm (a) and 1350-2050 nm (b).

3.2. Moisture Content of Vanilla

The moisture content of a product is an important quality indicator in various commodities, including spices, teas, fruits, vegetables, and herbal remedies. It plays a vital role in maintaining product stability over time and directly impacts shelf life by influencing how long the product remains safe and functional (Beć *et al*., 2022). Due to its impact on both product stability and longevity, moisture content requires careful and consistent monitoring to maintain quality standards and prevent spoilage or degradation.

Vanilla is an essential primary ingredient in the extraction industry, with market requirements varying by region. In the United States, the demand focuses on vanilla with low moisture content (20–25%), tailored for industrial processing. In contrast, the European market, which mainly serves household consumption, focuses on whole vanilla beans that are visually attractive, have a high vanillin concentration, a strong fragrance, and a moisture content ranging from 30–35% (Wahyuningsih *et al*., 2022). Meanwhile, in Indonesia, the maximum moisture content for Grade 1 vanilla, according to the Indonesian National Standard (SNI), ranges between 30–38% (Badan Standardisasi Nasional). To address the critical role of moisture content in vanilla quality, this research used Support Vector Regression (SVR) and preprocessing approaches to estimate the moisture content of vanilla beans. The moisture content prediction results for vanilla beans using the SVR algorithm are shown in Table 4. According to the SVR algorithm investigation on vanilla moisture content prediction (Table 4), the SCiO instrument, which operates in the 740–1070 nm wavelength region, produces somewhat better predictions than the Neospectra instrument, which operates in the 1350–2550 nm wavelength range.

The SCiO instrument achieved an R^2 value of 0.768 with Min-Max Normalization preprocessing, while the Neospectra instrument achieved an R^2 value of 0.758 with MSC preprocessing. This indicates that both models exhibit good performance in explaining data variation. The RMSE values obtained were 4.720 and 5.161, respectively. Better model performance is indicated by a lower RMSE value that is nearer 0 (Chicco *et al*., 2021). Meanwhile, the RPD values produced by both instruments were also comparable, with the SCiO-based model achieving an RPD of 2.075 and the Neospectra-based model achieving an RPD of 2.033. The RER value of 10.197 indicates that the RF model with min-max normalization preprocessing is moderately successful and can be used for quality screening or initial screening purposes. These outcomes are comparable to the PLS model, where PLS combined with first derivative-SNV preprocessing yielded an R^2 of 0.78, an RPD of 3.05, and an RMSE of 3.61 (Widyaningrum *et al*., 2024). Figure 4 shows the results of the plot between actual moisture content and predicted moisture content from each instrument.

This performance indicates that both instruments are suitable for practical applications in rapid moisture content analysis of vanilla. However, further validation with a more extensive and more diverse dataset is recommended to ensure the robustness of the models across different conditions. The robustness of a model largely depends on the quality and quantity of data utilized during its training and testing stages. Employing extensive and diverse datasets enables models to generalize more effectively, enhancing their resilience to data variability and noise (Zhou *et al*., 2021). The slight difference in performance between the SCiO and Neospectra instruments could be attributed to their respective spectral sensitivity ranges and interaction with moisture-related features.

Table 4. Prediction results of vanilla bean moisture content using the SVR algorithm.

Wavelength	Preprocessing	R^2	RMSE	RPD	RER
740-1070 nm	No preprocessing	0.412	7.504	1.304	6.413
	Min-max normalization	0.768	4.720	2.075	10.197
	SNV	0.753	4.862	2.013	9.899
	MSC	0.434	7.370	1.328	6.530
	First Derivative	0.698	5.382	1.819	8.942
	First Derivative-SNV	0.731	5.074	1.930	9.485
	First Derivative-MSC	0.740	4.990	1.962	9.645
1350-2550 nm	No preprocessing	0.709	5.660	1.854	8.503
	Min-max normalization	0.746	5.280	1.987	9.115
	SNV	0.717	5.583	1.879	8.620
	MSC	0.758	5.161	2.033	9.325
	First Derivative	0.687	5.867	1.788	8.203
	First Derivative-SNV	0.738	5.364	1.956	8.972
	First Derivative-MSC	0.718	5.571	1.883	8.639

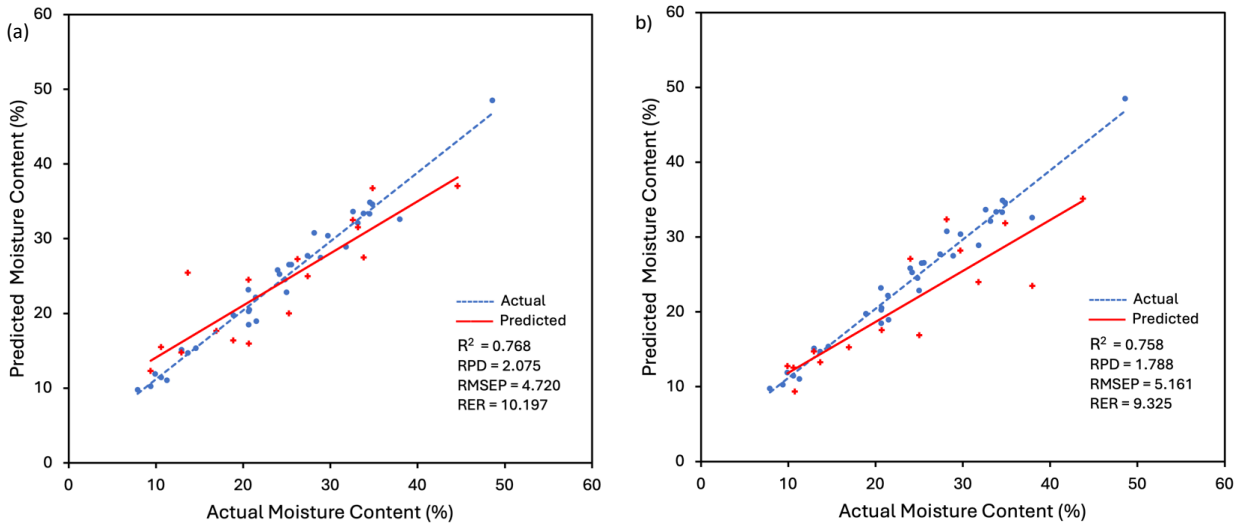

Figure 4. Plot comparing the actual and predicted moisture content of vanilla: (a) wavelength 740-1070 nm; and (b) wavelength 1350-2550 nm.

3.3. The Vanillin Concentration in Vanilla

The distinctive flavor and aromatic strength of vanilla are primarily attributed to vanillin, its most significant component. Approximately one-third of a vanilla product's flavor intensity is derived from its vanillin concentration. Consequently, vanillin content serves as a key factor in determining the market value of vanilla beans, with higher vanillin levels commanding premium prices (Ranadive, 2019). The vanillin content varies significantly across different vanilla species. Under optimal conditions, *Vanilla planifolia* beans can produce between 2% – 2.5% vanillin (Ranadive, 2019).

To predict the vanillin content in vanilla beans, this study applied the SVR algorithm, utilizing various preprocessing methods. The prediction results, presented in Table 5, illustrate how different preprocessing techniques influence the accuracy of vanillin content estimation. The prediction results obtained using the SCiO instrument with the SVR algorithm and MSC preprocessing showed moderate performance, with an R^2 value of 0.406, RMSE of 0.379, and RPD of 1.297. These values suggest that the model could explain a portion of the variability in the data, but there is still significant room for improvement. In contrast, the Neospectra instrument with the SVR algorithm without any preprocessing achieved lower performance, with an R^2 value of 0.172, RMSE of 0.576, and RPD of 1.098. This indicates a weaker predictive capability, with the model explaining only a small portion of the variance in the data. Overall, the results highlight the varying performance of different instruments and underscore the need for further optimization of both the modeling approach and the spectral data quality for more accurate predictions of vanilla quality parameters.

Several factors, such as the model employed, the quality of the spectral data, and the inherent characteristics of the instruments, may contribute to the observed difference in R^2 values between the Neospectra and SCiO instruments when employing the SVR method. The lower R^2 value (0.172) with Neospectra at wavelengths 1350-2550 nm suggests that the spectral data might be less representative or noisier, making it more challenging for the SVR model to capture meaningful patterns. In contrast, SCiO's moderate R^2 value (0.406) at wavelengths 740-1070 nm with the SVR algorithm could indicate that the instrument provides more consistent and relevant spectral information, enabling the SVR model to better capture the underlying trends in the data. However, the moderate performance also highlights to the complexity of predicting vanilla parameters, where spectral features alone may not fully explain variations in moisture or vanillin content, requiring further optimization of the model or data. The efficacy of machine learning models depends on the specific task and dataset (Sachindra & Kanae, 2019).

Table 5 Prediction results of vanillin content of vanilla bean using the SVR algorithm.

Wavelength	Preprocessing	R^2	RMSE	RPD	RER
740-1070 nm	No preprocessing	0.346	0.398	1.237	4.798
	Min-max normalization	0.118	0.462	1.065	4.134
	SNV	0.265	0.422	1.166	4.526
	MSC	0.406	0.379	1.297	5.039
	First Derivative	0.334	0.402	1.225	4.751
	First Derivative-SNV	0.326	0.404	1.218	4.727
	First Derivative-MSC	0.374	0.389	1.264	4.910
1350-2550 nm	No preprocessing	0.172	0.576	1.098	3.315
	Min-max normalization	0.170	0.577	1.097	3.310
	SNV	0.055	0.615	1.029	3.105
	MSC	0.141	0.587	1.788	3.253
	First Derivative	0.051	0.617	1.026	3.095
	First Derivative-SNV	0.012	0.629	1.006	3.036
	First Derivative-MSC	0.074	0.609	1.039	3.136

In comparison to the SVR results, the PLS model with SCiO showed a higher R^2 value (0.72) using the first derivative-SNV preprocessing method (Widyaningrum *et al*., 2024), suggesting that the PLS algorithm might better handle the spectral data in this context. This highlights the significant role that the choice of machine learning model plays in determining predictive accuracy (Wu *et al*., 2020). While preprocessing can enhance the data (Alasadi & Bhaya, 2017; Fan *et al*., 2021), the model's ability to effectively interpret and predict the target variables is critical (Murdoch *et al*., 2019), with PLS demonstrating better performance than SVR in this case.

4. CONCLUSION

This study underscores the critical importance of moisture and vanillin concentration in affecting the quality of vanilla beans. The study demonstrated the efficacy of Support Vector Regression (SVR) in predicting moisture content and vanillin concentration using spectral data from two instruments, SCiO and Neospectra, with different preprocessing methods. Both SCiO and Neospectra exhibited strong predictive capabilities for moisture content, suggesting suitability for practical applications. Notably, the SCiO instrument, operating in the 740–1070 nm wavelength range, showed slightly better performance than the Neospectra instrument, which operates in the 1350–2550 nm range. This difference is likely due to the SCiO's shorter wavelength range being more sensitive to moisture-related features in vanilla beans. The spectral sensitivity of the instruments directly influenced their ability to capture relevant patterns, emphasizing the importance of wavelength selection in optimizing predictive performance. However, further validation with more diverse datasets is recommended to enhance model robustness. In contrast, the prediction of vanillin content proved more challenging, with SCiO achieving moderate performance and Neospectra demonstrating limited predictive accuracy. The lower predictive performance of the Neospectra instrument may stem from its wavelength range being less effective at capturing vanillin-related spectral features, leading to noisier data and less representative models. These results highlight the inherent complexity of predicting vanillin concentration and suggest that additional optimization in spectral data quality, preprocessing methods, and modeling techniques is necessary. Future research should focus on expanding the dataset to include a more diverse range of vanilla samples, ensuring the robustness of models across varying conditions. Additionally, efforts should be directed toward improving spectral preprocessing techniques and exploring alternative machine learning models, such as hybrid or ensemble approaches, to enhance predictive accuracy for complex parameters like vanillin concentration. Investigating novel spectral regions or combining complementary instruments may provide deeper insights and improve prediction performance for vanilla quality attributes. Ultimately, these innovations will pave the way for dependable, efficient, and noninvasive tools for assessing vanilla quality.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Ministry of Education, Culture, Research, and Technology of Indonesia for funding this research through the 2022 Doctoral Dissertation Program under grant number 3828/IT3.L1/PT.01.03/P/B/2022. The authors also thank the Indonesian Agency of Agricultural Extension and Human Resources Development (IAAEHRD), Ministry of Agriculture, Indonesia, for their invaluable support.

REFERENCES

Alasadi, S.A., & Bhaya, W.S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, ***12***(16), 4102-4107.

Anand, A., Khurana, R., Wahal, N., Mahajan, S., Mehta, M., Satija, S., Sharma, N., Vyas, M., & Khurana, N. (2019). Vanillin: A comprehensive review of pharmacological activities. *Plant Archives*, ***19***(2), 1000-1004.

Anyidoho, E.K., Teye, E., Agbemafle, R., Amuah, C.L.Y., & Boadu, V.G. (2021). Application of portable near infrared spectroscopy for classifying and quantifying cocoa bean quality parameters. *Journal of Food Processing and Preservation*, ***45***(5), e15445. https://doi.org/10.1111/jfpp.15445

Badan Standardisasi Nasional. (n.d.). *SNI 01-0010-2002 Panili*.

Baqueiro-Peña, I., & Guerrero-Beltrán, J.Á. (2017). Vanilla (*Vanilla planifolia* Andr.), its residues and other industrial by-products for recovering high value flavor molecules: A review. *Journal of Applied Research on Medicinal and Aromatic Plants*, ***6***, 1–9. https://doi.org/10.1016/j.jarmap.2016.10.003

Beć, K.B., Grabska, J., & Huck, C.W. (2021). Principles and applications of miniaturized Near-Infrared (NIR) spectrometers. In *Chemistry - A European Journal*, ***27***(5), 1514-1532. https://doi.org/10.1002/chem.202002838

Beć, K.B., Grabska, J., & Huck, C.W. (2022). Miniaturized NIR spectroscopy in food analysis and quality control: Promises, challenges, and perspectives. *Foods*, ***11***(10), 1465. https://doi.org/10.3390/foods11101465

Bittner, L.K., Schönbichler, S.A., Bonn, G.K., & Huck, C.W. (2013). Near infrared spectroscopy (NIRS) as a tool to analyze phenolic compounds in plants. *Current Analytical Chemistry*, ***9***(3), 417–423. https://doi.org/10.2174/15734110113090300010

Chadalavada, K., Anbazhagan, K., Ndour, A., Choudhary, S., Palmer, W., Flynn, J.R., Mallayee, S., Pothu, S., Prasad, K.V.S.V., Varijakshapanikar, P., Jones, C.S., & Kholová, J. (2022). NIR instruments and prediction methods for rapid access to grain protein content in multiple cereals. *Sensors*, ***22***(10), 3710. https://doi.org/10.3390/s22103710

Chen, C., Li, H., Lv, X., Tang, J., Chen, C., & Zheng, X. (2019). Application of near infrared spectroscopy combined with SVR algorithm in rapid detection of cAMP content in red jujube. *Optik*, ***194***, 163063. https://doi.org/10.1016/j.ijleo.2019.163063

Chicco, D., Warrens, M.J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, ***7***, e623. https://doi.org/10.7717/PEERJ-CS.623

Correia, R.M., Tosato, F., Domingos, E., Rodrigues, R.R.T., Aquino, L.F.M., Filgueiras, P.R., Lacerda, V., & Romão, W. (2018). Portable near infrared spectroscopy applied to quality control of Brazilian coffee. *Talanta*, ***176***, 59-68. https://doi.org/10.1016/j.talanta.2017.08.009

Cozzolino, D. (2016). Near infrared spectroscopy and food authenticity. In *Advances in Food Traceability Techniques and Technologies: Improving Quality Throughout the Food Chain*, 119-136. https://doi.org/10.1016/B978-0-08-100310-7.00007-7

Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., & Coulon, F. (2018). Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Science of the Total Environment*, *616–617*. https://doi.org/10.1016/j.scitotenv.2017.10.323

Drucker, H., Surges, C.J.C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems,* ***9***, 155-161.

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, ***9***, Article 652801. https://doi.org/10.3389/fenrg.2021.652801

Havkin-Frenkel, D., & Frenkel, C. (2008). Postharvest handling and storage of cured vanilla beans. *Stewart Postharvest Review*, ***2***(4), 1-9. https://doi.org/10.2212/spr.2006.4.6

Hayati, R., Munawar, A.A., & Fachruddin, F. (2020). Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango. *Data in Brief*, ***30***, 105571. https://doi.org/10.1016/j.dib.2020.105571

Huck, C.W. (2020). New trend in instrumentation of NIR spectroscopy-miniaturization. In *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications,* 193-210. https://doi.org/10.1007/978-981-15-8648-4_8

Malvandi, A., Feng, H., & Kamruzzaman, M. (2022). Application of NIR spectroscopy and multivariate analysis for Non-destructive evaluation of apple moisture content during ultrasonic drying. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, ***269***, 120733. https://doi.org/10.1016/j.saa.2021.120733

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, ***116***(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116

Osborne, B.G., Fearn, T., & Hindle, P.H. (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. Longman Scientific and Technical, Harlow, UK: 227 pp.

Pandiselvam, R., Prithviraj, V., Manikantan, M.R., Kothakota, A., Rusu, A.V., Trif, M., & Mousavi Khaneghah, A. (2022). Recent advancements in NIR spectroscopy for assessing the quality and safety of horticultural products: A comprehensive review. In *Frontiers in Nutrition*, ***9***, Article 973457. https://doi.org/10.3389/fnut.2022.973457

Pereira, C.G., Leite, A.I.N., Andrade, J., Bell, M.J.V., & Anjos, V. (2019). Evaluation of butter oil adulteration with soybean oil by FT-MIR and FT-NIR spectroscopies and multivariate analyses. *LWT*, ***107***, 1-8. https://doi.org/10.1016/j.lwt.2019.02.072

Pu, Y., Pérez-Marín, D., O'shea, N., & Garrido-Varo, A. (2021). Recent advances in portable and handheld NIR spectrometers and applications in milk, cheese and dairy powders. *Foods*, ***10***(10), 2377. https://doi.org/10.3390/foods10102377

Raju, V.N.G., Lakshmi, K.P., Jain, V.M., Kalidindi, A., & Padma, V. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Ranadive, A.S. (2019). Quality control of vanilla beans and extracts. In Havkin-Frenkel, D., & Belanger, F.C. (Eds.), *Handbook of Vanilla Science and Technology* (2nd Ed., pp. 239–259). Wiley. https://doi.org/10.1002/9781119377320.ch15

Rodríguez-Pérez, R., & Bajorath, J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. *Journal of Computer-Aided Molecular Design*, ***36***(5), 355–362. https://doi.org/10.1007/s10822-022-00442-9

Sachindra, D.A., & Kanae, S. (2019). Machine learning for downscaling: the use of parallel multiple populations in genetic programming. *Stochastic Environmental Research and Risk Assessment*, ***33***(8–9), 1497–1533. https://doi.org/10.1007/s00477-019-01721-y

Schwanninger, M., Rodrigues, J.C., & Fackler, K. (2011). A review of band assignments in near infrared spectra of wood and wood components. In *Journal of Near Infrared Spectroscopy*, ***19***(5), 287–308. https://doi.org/10.1255/jnirs.955

Torniainen, J., Afara, I.O., Prakash, M., Sarin, J.K., Stenroth, L., & Töyräs, J. (2020). Open-source python module for automated preprocessing of near infrared spectroscopic data. *Analytica Chimica Acta*, ***1108***, 1-9. https://doi.org/10.1016/j.aca.2020.02.030

Wahyuningsih, R., Fitriasari, B., & Suwardji, S. (2022). Development of vanilla agribusiness and its export opportunities to support triple export program (Gratitude) on Lombok Island. *Path of Science*, ***8***(6), 5020–5024. https://doi.org/10.22178/pos.82-18

Wani, O.A., Mahdi, S.S., Yeasin, Md., Kumar, S.S., Gagnon, A.S., Danish, F., Al-Ansari, N., El-Hendawy, S., & Mattar, M.A. (2024). Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Scientific Reports*, ***14***, Article number 27876. https://doi.org/10.1038/s41598-024-77687-x

Widyaningrum, W., Purwanto, Y.A., Widodo, S., Supijatno, & Iriani, E.S. (2024). Rapid assessment of vanilla (*Vanilla planifolia*) quality parameters using portable near-infrared spectroscopy combined with random forest. *Journal of Food Composition and Analysis*, ***133***(March), 106346. https://doi.org/10.1016/j.jfca.2024.106346

Williams, P., & Norris, K. (Eds.) (1987). *Near-Infrared Technology in the Agricultural and Food Industries*. American Association of Cereal Chemists, St. Paul, Minnesota, USA: 330 p.

Wokadala, O.C., Human, C., Willemse, S., & Emmambux, N.M. (2020). Rapid non-destructive moisture content monitoring using a handheld portable Vis–NIR spectrophotometer during solar drying of mangoes (*Mangifera indica* L.). *Journal of Food Measurement and Characterization*, ***14***(2), 790–798. https://doi.org/10.1007/s11694-019-00327-w

Workman, J., & Weyer, L. (2007). *Practical Guide to Interpretive Near-Infrared Spectroscopy* (1st ed.). CRC Press. https://doi.org/10.1201/9781420018318

Wu, Y., Guo, J., Sun, R., & Min, J. (2020). Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *Npj Computational Materials*, ***6***(1), Article number 120. https://doi.org/10.1038/s41524-020-00388-2

Zareef, M., Chen, Q., Hassan, M. M., Arslan, M., Hashim, M. M., Ahmad, W., Kutsanedzie, F. Y. H., & Agyekum, A. A. (2020). An overview on the applications of typical non-linear algorithms coupled with NIR spectroscopy in food analysis. *Food Engineering Reviews*, ***12***(2), 173–190. https://doi.org/10.1007/s12393-020-09210-7

Zhang, F., & O'Donnell, L.J. (2020). Support vector regression. In Machine learning: Methods and applications to brain disorders (pp. 123–140). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00007-9

Zhang, W., Kasun, L.C., Wang, Q.J., Zheng, Y., & Lin, Z. (2022). A review of machine learning for near-infrared spectroscopy. *Sensors,* ***22***(24), 9764. https://doi.org/10.3390/s22249764

Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., & Madabhushi, A. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, ***109***(5), 820 - 838. https://doi.org/10.1109/JPROC.2021.3054390

Zou, H., Shen, S., Lan, T., Sheng, X., Zan, J., Jiang, Y., Du, Q., & Yuan, H. (2022). Prediction method of the moisture content of black tea during processing based on the miniaturized near-infrared spectrometer. *Horticulturae*, ***8***(12), 1170. https://doi.org/10.3390/horticulturae8121170