

Rapid Quantification of Patchouli Alcohol in Patchouli Oil (Pogostemon Cablin) using Near Infrared Spectroscopy with Multivariate and Machine Learning Approaches

Nurul Ainina

Padjadjaran University

Sarifah Nurjanah

Padjadjaran University

Uma Rani Sinniah

Universiti Putra Malaysia

Rahmat Hidayat

Bandung Institute of Technology

Euis Juliaeha

Padjadjaran University

Muhammad Achirul Nanda

Padjadjaran University

Mimin Muhaemin

m.muhaemin@unpad.ac.id

Padjadjaran University

Article

Keywords: near-infrared spectroscopy, patchouli oil, patchouli alcohol, partial least square regression (PLS), support vector regression (SVR)

Posted Date: May 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9409718/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Patchouli oil (*Pogostemon Cablin*) is an essential oil primarily used as a fixative ingredient in the perfumery and cosmetic industries. The major component, patchouli alcohol (PA), is recognized as a key quality parameter, with higher contents being desirable. However, the conventional method for quantifying PA, namely gas chromatography-mass spectrometry (GC-MS), is relatively expensive, time-consuming, and destructive. Therefore, this study aimed to explore near-infrared (NIR) spectroscopy combined with multivariate and machine learning algorithms, including partial least squares regression (PLSR), support vector regression (SVR), and a hybrid method, partial least squares-support vector regression (PLS-SVR). The optimum results were obtained using the PLS-SVR approach in combination with Savitzky-Golay 2nd derivative (SG 2nd) preprocessing, achieving $R^2_C = 0.9182$, RMSEC = 2.1344, $R^2_{CV} = 0.9124$, RMSECV = 2.2085, $R^2_P = 0.9317$, RMSEP = 2.2095, and RPD = 3.849. In conclusion, NIR spectroscopy, in combination with the hybrid PLS-SVR method, could be used as a rapid, simple, low-cost alternative method for estimating PA content in patchouli oil.

1. Introduction

Patchouli oil (*Pogostemon Cablin*) is widely recognized as a fixative ingredient in many products, especially in perfume and skincare products. The blending of patchouli oil with other essential oil provides a strong base and long-lasting fragrance due to the slow evaporation rate, leading to high scent tenacity [1]. Patchouli oil is also popular due to the therapeutic properties, including antibacterial, antidepressant, anti-inflammatory, anti-allergy, and antinociceptive [2, 3, 4, 5]. The major constituent is patchouli alcohol (PA), which is responsible for the unique aroma [6]. In addition, the quality of patchouli oil is commonly determined by PA concentration. The higher the concentration of PA, the greater the quality of patchouli oil. PA content depends on several factors, including harvested time, cultivated location, and production method [7].

The most frequently used methods for chemical composition analysis of patchouli oil are gas chromatography (GC) coupled with mass spectrometry (MS). However, GC-MS is often considered to be time-consuming, relatively expensive, destructive, and requires skilled operators. Consequently, near-infrared spectroscopy (NIR) has the potential to be an alternative analytical method for obtaining chemical and physical properties of a sample [8]. This method is quick, non-destructive, low-cost, easy to operate, and requires minimal or no sample preparation [9, 10]. NIR absorption bands are in the range between 780–2500 nm and originate from overtones and combinations of hydrogen functional groups, including C – H, N – H, and O – H [11]. PA ($C_{15}H_{26}O$) is a part of the sesquiterpene group containing hydrogen bonds. Therefore, NIR spectroscopy application is probably suitable for the analysis.

Based on previous studies, NIR was found useful for quantitative and qualitative analysis of essential oil. In quantitative analysis, this method demonstrated significant potential in predicting the yield of essential oil directly from the sources [12, 13] and quantifying the major compounds [14, 15, 16, 17, 18]. For qualitative purposes, NIR is capable of differentiating between pure and adulterated essential oil [19, 20, 21] and for chemotype determination [14, 22]. Moreover, NIR spectrum is associated with chemical constituents in organic materials as the molecules are rich in hydrogen-bearing groups. This information is foundational for quantitative and qualitative analysis using NIR spectroscopy [11].

NIR information is difficult to interpret due to the large amount of data and the presence of many overlapping peaks, causing broad bands [23]. To address this problem, chemometrics have been developed to extract relevant information between NIR spectral information and the chemical composition of samples [24]. By using chemometrics, a calibration model was established, and the sample analysis could be executed. Multivariate data analysis, as a part of chemometrics, is used to interpret NIR spectral information, both for quantitative and qualitative analysis. It includes principal component analysis (PCA), soft independent modelling of class analogy (SIMCA), partial least squares regression (PLSR), multiple linear regression (MLR), principal component regression (PCR), and linear discriminant analysis (LDA) [25]. Currently, machine learning algorithms are also used to extract NIR information, such as support vector machine/regression (SVM/SVR), and multilayer perceptron (MLP) [22, 26].

In the context of quantitative analysis, regression methods such as PCR, MLR, and PLSR are commonly applied to build a calibration model on NIR data. Among these methods, PLSR is the most frequently used due to simplicity, speed, and strong performance. However, this method only deals with linear relationships. Consequently, machine learning algorithms that work with nonlinear relationships, such as support vector regression (SVR), have been implemented in many cases of NIR spectroscopy analysis. Studies on the prediction of volatile oil components in *Mentha haplocalyx* powder showed that SVR outperformed PLSR, achieving higher coefficients of determination ($R^2 > 0.9$) and lower prediction errors [27]. Detection of cyclic adenosine monophosphate content from red jujube was successfully predicted by SVR, providing better results than PLSR [28]. In another study, SVR showed more favorable results in NIR prediction of ascorbic acid and titratable acidity of acerola fruit [29].

Based on the discussion above, this study aims to develop a rapid and non-destructive method for PA determination in patchouli oil using NIR spectroscopy. Although several works have explored the quantification of PA using NIR [30, 31, 32], none have performed a nonlinear regression method for model development. Therefore, in this study, both linear and nonlinear regression approaches—represented by PLS and SVR, respectively, were used to build NIR prediction model of PA. A hybrid method of PLS-SVR was also investigated, in which dimensionality reduction was performed first using PLS, followed by regression using SVR. The combination of NIR spectroscopy with linear and nonlinear regression models was developed to achieve robust models, thereby benefiting the evaluation of patchouli oil quality.

2. Material and Method

2.1 Patchouli oil samples and reference analysis

A total of 75 samples of patchouli oil were obtained from different suppliers and locations in Indonesia. The majority were sourced from the biggest producers of patchouli oil; Sumatera, Java, and Sulawesi, then the rest were from Nusa Tenggara, Maluku, and Papua. To expand the range of concentration of PA, twenty-five samples were prepared by spiking patchouli oil with a lower PA concentration with patchouli oil containing a higher PA content. All samples were stored in amber glass bottles to protect from the light.

As the reference values of PA quantification, gas chromatography coupled with mass spectrometer (GC-MS) analysis was carried out using 7890 GC with auto sampler and 5975 mass selective detector (Agilent Technologies). The experimental conditions included the following oven temperature program: initial temperature at 60°C and held for 0 minute, increased at 2°C/min to 150°C and held for 1 minute, then raised to 20°C/min to 210°C, where it was held for 10 minutes. The injector temperature (split ratio 250) was set to 250°C, and helium was used as the carrier gas at a flow rate of 0.6 mL/min.

2.2 NIR spectroscopy measurements

The spectra of each sample of patchouli oil were recorded from 1000–2500 nm using FT-NIR N-500 Fiber Optic Solid with Transreflectance Adapter (Buchi, Switzerland). About 5 millilitres of patchouli oil were placed in the beaker, and the transreflectance head was put into it to collect NIR spectrum. The transreflectance mode was applied as the sampling method. Three times scanning was carried out for each sample, hence, 300 spectra were acquired (100 samples x 3 scans). The spectra of each sample were not averaged to determine the reproducibility of NIR spectroscopy method [21].

2.3 Data analysis

2.3.1 Spectral preprocessing

Before the spectral preprocessing, patchouli oil samples were divided into a calibration set (70% of samples) and prediction set (30% samples). As each sample was scanned three times, it is essential to ensure that all replicates of a single sample remain on the same data partition (either calibration or prediction) to avoid data leakage. Each of spectra dataset then were preprocessed, as raw spectra typically contain noise, baseline drift, background effects, and stray light interference [24] that could lead to inaccurate predictions. Various preprocessing methods, such as Savitzky-Golay (SG) polynomial derivatives (1st and 2nd derivative), Autoscale, Standard Normal Variate (SNV), and Multiplicative Scatter Correction (MSC) were performed

before regression modelling. SNV and MSC is used to minimize the physical variability between samples due to scattering effects caused by dispersion, particle size, and the difference in signal intensity [33, 25]. The 1st derivative SG (SG 1st) is capable of removing baseline from spectra, while 2nd derivative SG (SG 2nd) can remove baseline and linear trend [33]. In this study, both SG 1st and SG 2nd incorporated a smoothing process in the calculation and used 15 points of window length with a polynomial order of 2.

Outlier detection was performed using PCA in combination with Hotelling T^2 ellipse after each preprocessing only in calibration set. Preprocessed spectra of patchouli oil were projected into PCA map, then Hotelling T^2 ellipse (confidence level at 95%) was adopted onto this PCA projection [34]. Samples found outside the ellipse confidence limit were identified as the outliers and were removed before calibration to increase model robustness.

2.3.3 Regression tools

PLSR, SVR, and PLS-SVR algorithms were applied to investigate the correlation between NIR spectra and PA concentration. The number of latent variables (LVs) in PLSR model was optimized using GridSearchCV procedure from Scikit-learn library (Python), with the range of 1 to 5 LVs. The same procedure was used to determine the optimal hyperparameters for SVR, including C as the cost of constraints violation (0.1, 1, 5, 10, 20, 25, 50), epsilon (ϵ) representing the margin of tolerance (0.1, 0.25, 0.50, 0.75, 0.9), and gamma (γ) parameter (0.0001, 0.001, 0.01, 0.1, 'scale'). A radial basis function (RBF) kernel was selected for SVR, as it typically provides better performance than other kernels (linear, polynomial, and sigmoid) [27] and can handle non-linear cases. Ten-fold cross-validation was applied to all models, facilitating the assessment of predictive ability using an internal validation procedure. This step was performed by dividing the calibration set into 10 subsets and consecutively excluding each group, fitting a model, and assessing model predictions of the excluded samples.

In the case of the PLS-SVR method, dimensionality reduction was first conducted using PLS to transform the high-dimensional spectra into a reduced set of latent variables (LVs), with the number of LVs adjusted from 1 to 5, as in PLSR. The resulting latent variables were then used as inputs for the SVR model, following the same hyperparameter optimization strategy as applied to SVR alone. Summary of NIR data analysis is presented in Fig. 1.

2.3.4 Model evaluation

The performance of models was evaluated using statistical terms, such as correlation coefficient of calibration (R_C^2), root mean squared error of calibration (RMSEC), correlation coefficient of prediction (R_P^2), root mean squared error of prediction (RMSEP), correlation coefficient of cross-validation (R_{CV}^2), root mean squared error of cross-validation (RMSECV), and residual predictive deviation (RPD) of prediction set. The R^2 values for each dataset must be close to 1 with low RMSE values to obtain a robust model. RPD is defined as the ratio between the standard deviation of PA concentration on the prediction set and the RMSECV. A model with RPD value higher than 2.5 indicates strong predictive ability, while a value ranging from 2–2.5 and 1.5–2 represents good and moderate predictive performance, respectively. RPD value below 1.5 demonstrates poor accuracy and is insufficient for practical application [35].

2.3.5 Software for NIR data analysis

NIR data analysis was conducted in Python 3.12.4 coding language by Jupyter Notebook in an Anaconda environment. The codes use a library from Scikit-learn and numerical SciPy, NumPy, and Pandas.

3. Results and discussion

3.1 Reference data analysis

The concentration of PA on patchouli oil samples determined by GC-MS analysis had values in the range of 5.22–58.20%. Table 1 shows that the differences in mean and standard deviation between the full dataset, calibration, and prediction set on PA content were not statistically significant. These results show that the NIR data splitting into calibration and prediction sets was reasonable and reliable for further analysis [36].

Table 1
Descriptive statistics for patchouli alcohol prediction on patchouli oil

Dataset groups	Sample number	Maximum (%)	Minimum (%)	Mean (%)	Standard deviation
Full dataset	100	58.2	5.22	26.56%	8.53%
Calibration set	70	58.2	5.22	26.85%	8.54%
Prediction set	30	39.41	6.13	25.88%	8.50%

3.2 NIR spectral features

The NIR spectra of patchouli oil raw data are shown in Fig. 2. The absorption peaks of the spectrum are observed in several regions, with the most intense peaks found in the range of 2300–2500 nm. Other peaks were also observed in 1195, 1415, 1637, 1708–1715, 1752, 2045–2170 nm. The band at around 1195 nm corresponds to C-H second overtone[37] while the peak at around 1415 could be associated with the O-H stretch first overtone [38, 20]. The bands between 1610–1752 nm were probably due to C-H first overtone [39] [37]. The region of 2045–2170 and 2300–2500 might be related to O-H and C-H combination bands, respectively [39, 40, 38]. Due to the many overlapping bands found in these spectra, quantitative analysis is complex to investigate by direct interpretation. Therefore, multivariate analysis was conducted to predict PA content from NIR spectral characteristics.

3.3 Spectral preprocessing result

Five preprocessing techniques were applied, followed by outlier detection using Hotelling T^2 (significance level $\alpha = 0.05$). The visualization of NIR spectra after preprocessing –autoscaling, SG 1st, SG 2nd, SNV and MSC– and outlier removal are shown in Fig. 3. The original data shows notable amplitude variation in the region above 2200 nm. The SG 1st and SG 2nd treatments flattened the baseline to zero, with the SG 2nd revealing sharper absorption peaks. SNV and MSC preserved the spectral shape, with SNV rescaled the spectra absorbance unit into 5, while MSC maintained in the original scale. Compared to the original spectra, SNV and MSC brought the spectra visually close together, particularly above 2200 nm. In contrast, autoscaling exhibited chaotic spectral features.

The outlier classification results for each method were also slightly different. A summary of outlier removal process is presented in Table 2. Notably, the outlier detection was exclusively performed for calibration set to enhance model robustness.

Table 2
The summary of outlier detection on calibration set for each preprocessing method

Preprocessing	Initial spectra (3 times scan/samples)	Outlier detected	Final spectra
None	300	25	275
Autoscale	300	25	275
SG 1st	300	27	273
SG 2nd	300	27	273
SNV	300	19	281
MSC	300	19	281

As shown in Table 2, the SG 1st and SG 2nd derivative technique detected a higher number of outliers compared to the other preprocessing methods, whereas MSC and SNV resulted in fewer outliers. However, the quantity of outliers cannot be used to judge whether a method is suitable. These outlier variations can be attributed to the distinct mathematical operations of each preprocessing techniques.

3.4 Model evaluation using PLSR, SVR, and PLS-SVR

This section presents the results of NIR spectroscopy modeling using three different regression methods, namely PLSR, SVR, and PLS-SVR, under various preprocessing methods. The models were evaluated based on the performance metrics (R^2 , RMSE, and RPD) on calibration, cross-validation, and prediction datasets, to determine the most accurate method for predicting PA content on patchouli oil. Several pre-processing data (SNV, SG 1st, SG 2nd, MC, and the combination) were applied into NIR spectral to remove undesired information, commonly through noise and scattering effects removal, baseline correction, and other physical phenomena during measurement before building calibration models. Due to the independent outlier removal applied to each preprocessed dataset, the final number of spectra used for calibration (N) slightly varied between models.

3.4.1 PLSR Model Results

In this study, the number of latent variables (LVs) for all PLSR models was set in the range of 1–5, as LVs are a critical parameter in PLS that directly affect model accuracy. An insufficient number of LVs can eliminate the useful information from the original data, leading to reduced model predictive ability, while excessive LVs may incorporate irrelevant noises, resulting in an overfitting model [41]. Therefore, finding the optimum number of LVs is a critical step in PLSR. The GridSearchCV procedure implemented in Python was used to calculate the optimal number of LVs by minimizing the RMSE in cross-validation.

Table 3 shows the performance of PLSR models. Compared to the original data, preprocessing of NIR data by SG 1st and SG 2nd methods slightly enhanced the model performances of PLSR, with almost all R^2 exceeding 0.85 (except R^2_{cv} for SG 1st) and for RMSE less than less than 3.2. across all datasets. The best model of PLSR was obtained by applying SG 2nd, which produced the lowest values of RMSEP, as well as the highest values of R^2_p and RPD. The R^2_c and R^2_p values were close to 0.9, and the value of RPD was higher than 3, indicating that the model has excellent predictive accuracy and robustness [24]. Although SG 1st model showed a higher R^2_c value, the SG 2nd model provided better prediction performance, as indicated by higher R^2_p and lower RMSEP. Second derivative preprocessing method eliminates baseline and linear trend, while also smoothing the NIR spectrum to minimize reduction of signal-to-noise ratio [33]. Furthermore, PLSR with autoscaling exhibited identical performance to model without preprocessing. This is expected, as the PLSRegression library in Python performs autoscaling by default during its operation. Scattering plots between predicted and measured PA using the best PLSR models are shown in Fig. 4.

Table 3
Statistical evaluation of the regression models for PA content determination using PLSR

Preprocessing	N	LVs	Calibration		Cross-validation		Prediction		RPD
			R^2_c	RMSEC(%)	R^2_{cv}	RMSECV(%)	R^2_p	RMSEP(%)	
None	275	5	0.8260	3.0546	0.7945	3.319	0.8521	3.2513	2.616
Autoscale	275	5	0.8260	3.0546	0.7945	3.319	0.8521	3.2513	2.616
SG 1st	273	4	0.9113	2.1775	0.8394	2.9294	0.8628	3.1318	2.715
SG 2nd	273	3	0.8995	2.3652	0.8571	2.8203	0.8965	2.7206	3.126
SNV	281	3	0.7425	3.9268	0.7025	4.2204	0.659	4.9374	1.722
MSC	281	3	0.7431	3.9214	0.7031	4.2161	0.6563	4.9571	1.716

N = number of spectra in calibration dataset after outlier removal

Interestingly, the application of standard normal variate (SNV) and multiplicative scatter correction (MSC), resulted in less satisfactory performance for the NIR spectra of patchouli oil. The R^2 and RMSE values for calibration and prediction of both

models were approximately 0.66–0.74 and 3.93–4.94, respectively, with an RPD less than 2, which is insufficient for model applications. Moreover, these results were also inferior compared to that of non-preprocessed spectra.

This implies that some valuable information related to the properties of interest may have been removed during preprocessing using SNV and MSC. SNV and MSC belong to the group of scatter-correction methods, which are developed to minimize the physical variability between samples due to light scattering [33]. Typically, light scattering occurs in powder materials, while in pure liquids, such as patchouli oil, it is minimized. Consequently, SNV and MSC failed to improve the PA prediction models. The result is consistent with previous studies in NIR analysis, showing that NIR preprocessing method could sometimes lead to poorer model performance compared to the original spectra [42, 43]. In some cases, the scattering information may include crucial details and the removal can adversely affect model accuracy. Furthermore, a multiplicative effect may also be present, making modelling more challenging [43]. Another study have reported that MSC may eliminate information correlated to the target quantification (y) from the spectral signal (X), thereby leading to innacurate model prediction [44].

In the case of patchouli oil, previous studies have also applied NIR spectroscopy and PLSR to quantify PA content, albeit with a smaller sample sizes and narrower concentration ranges. For instance, [30] reported that PA models had R^2 of 0.74, RMSEC of 4.162, RMSEP of 4.666, and an RPD of 1.8 using PLSR. A better result was obtained by [31], which reported R^2 , RMSEC, and RPD values of 0.8852, 3.847, and 3.02, respectively. Compared to these studies, the current work presents a notable improvement in PA prediction accuracy.

Furthermore, a recent investigation[32] reported that PA yielding r values between 0.910 and 0.980, with SEC, SEP, and RPD values ranging from 0.658% to 0.908%, 0.759% to 0.920%, and 2.43 to 5.04, respectively. While that study reported excellent statistical parameters, their model was developed using a narrower concentration (24–34%) from only 8 provinces in Indonesia. In contrast, the present study XXXtilizes broader range of PA concentration (5–58%) with samples obtained from more diverse regions across Indonesia. Despite the complexity of the dataset, this work demonstrates high model robustness, achieving an R^2 for prediction of 0.8965 with RMSEC and RMSEP values less than 3 and RPD of 3.126. These results suggest that the current model may be more suited for industrial applications.

3.4.2 SVR model results

Support vector regression (SVR) is the extended version of the machine learning application support vector machine (SVM), usually applied for quantitative analysis in various subjects. A comprehensive explanation of SVM and SVR has been published in previous studies [45, 46]. In SVM, the core concept is the construction of an optimal hyperplane that separates data in a high-dimensional feature space using kernel functions. The optimal hyperplane is defined by maximizing the margin, that is, the distance between the hyperplane and the nearest data points (support vectors) [47]. Numerous studies have used NIR spectroscopy with SVR to perform multivariate calibration due to the ability to handle non-linear problems. The accuracy of SVR model is based on the selection of hyperparameters, cost (C), epsilon (ϵ), and kernel parameter (gamma (γ)) [48]. The C parameter acts as penalty control for any cases that are outside of the regression tolerance margin (set based on ϵ). A high value of C indicates cases outside of the tolerance margin are severely penalized, decreasing the training bias, making it potentially overfit. Meanwhile, low value may lead to a greater training bias, producing an underfitted model [26, 48]. ϵ is tolerance margin where no penalty given when the training cases in the regression do not fall more from the hyperplane and within ϵ value. A higher value of ϵ will probably produce high error and an underfitted model, while an extremely low value may result in an overfitted model [26].

In contrast to PLSR, SVR models used full spectrum without dimensionality reduction. As presented in Table 4, the SG 1st preprocessing with SVR produced the highest prediction performance, with an R^2_p of 0.8518, RPD values of 2.612, and the lowest RMSEP of 3.2554, indicating acceptable model performance. Furthermore, calibration and cross-validation results supported this performance, with $R^2_c = 0.8763$, RMSEC = 2.571, $R^2_{cv} = 0.8243$, RMSECV = 3.064, suggesting good generalization ability and a minimal risk of overfitting, shown by a small gap between R^2_c and R^2_p .

The optimal combination of SVR hyperparameters (C , γ , and ϵ) was determined using GridSearchCV. As shown in Table 3, all data preprocessed were best fitted with γ value of 0.001, while the optimal values of C and ϵ values varied depending on the

preprocessing method. Scatter plots between predicted and actual reference using the best models of SVR with SG 1st preprocessing method are presented in Fig. 5.

Table 4
Statistical evaluation of the regression models for PA content determination using SVR

Preprocessing	N	Hyperparameter			Calibration		Cross-validation		Prediction		RPD
		γ	C	ϵ	R^2_C	RMSEC	R^2_{CV}	RMSECV	R^2_P	RMSEP	
None	275	0.1	50	0.75	0.7704	3.5084	0.7438	3.7061	0.8106	3.6800	2.311
Autoscale	275	0.0001	20	0.75	0.8459	2.8746	0.8003	3.2721	0.8435	3.3452	2.542
SG 1st	273	0.0001	10	0.1	0.8763	2.571	0.8243	3.064	0.8518	3.2554	2.612
SG 2nd	273	0.0001	25	0.1	0.9453	1.7453	0.8939	2.4305	0.8137	3.6491	2.331
SNV	281	0.0001	50	0.1	0.8775	2.7087	0.8152	3.3259	0.7765	3.9969	2.128
MSC	281	0.0001	50	0.1	0.8773	2.71	0.8158	3.3211	0.7829	3.9399	2.159

N = number of spectra in calibration dataset after outlier removal

Similar to the PLSR method, SNV and MSC preprocessing also produced less satisfactory results using SVR approach, as indicated by the lowest R^2_P values (< 0.8) and RPD values (2.128). Furthermore, the SNV and MSC-based model exhibited overfitting, evidenced by a notable gap between R^2_C and R^2_P (> 0.1). These results strongly support the earlier observation in the PLSR section that SNV and MSC may eliminate substantial chemical information relevant to the prediction of PA in patchouli oil, even when dimensionality reduction is not applied and a non-linear method, such as SVR, is used.

Regression of NIR spectra using SVR tends to produce overfitted model for SG 2nd, SNV, and MSC preprocessing methods, as reflected in the noticeable gaps between their R^2_P and R^2_C values. The most remarkable gap was found on NIR spectra preprocessed with SG 2nd which had a gap above 0.1. Generally, R^2_C has higher values than R^2_P , but when the gap between R^2_C and R^2_P is too large, it indicates that the model fits the calibration dataset well, but fails to generalize to the new data (prediction dataset) [49]. Overfitting the model in SVR may be attributed to improper selection of the hyperparameters (C, γ , and ϵ). In this case, the selected SVR hyperparameters may be unsuitable for NIR spectra preprocessed using the SG 2nd, SNV, and MSC in PA prediction.

Compared to the best model in PLSR (SG 2nd), the SVR model (SG 1st) produced a lower model performance, characterized by lower R^2 values across all datasets, as well as lower RPD, and higher RMSE values. Despite this, this model also did not result overfit models because the difference values between R^2_C and R^2_P were relatively close. A similar trend was observed in the rapid quantification of total volatile basic nitrogen (TVB-N) content in beef and pork; there, NIR combined with PLSR provided superior performance ($R^2_C = 0.9858$, RMSEC = 1.41, $R^2_P=0.9366$, RMSEP = 3.15) compared to full-spectrum SVR ($R^2_C = 0.9846$, RMSEC = 1.58, $R^2_P=0.8314$, RMSEP = 4.61) [50]. Another study investigating the prediction of cellulose, hemicellulose and lignin content in crop straws similarly found that the full-spectrum was inferior to PLSR for all three components, as indicated by its lower R^2_P values and RPD, and higher RMSEP [51]. Additionally, the full spectrum of SVR models also exhibited relatively worse performances compared to PLSR in predicting soluble solid content in citrus fruit, although the differences between the two methods were not statistically significant [52].

3.4.3 PLS-SVR model results

PLS-SVR combines PLS for dimensionality reduction with SVR as a nonlinear regressor. Dimensionality reduction aims to improve model accuracy by removing redundant features, irrelevant information, and noise associated with the target variable [53]. Additionally, this step also enables faster computational processing. In this method, PLS replaces the original with a few

latent variables (LVs) [54]. PLS dimensionality reduction result was used as the input for SVR. The LVs number applied in SVR was similar to those used in PLSR. The results of PLS-SVR method for PA prediction using NIR spectroscopy are shown in Table 5. Additionally, the graphics depicting the relationship between predicted and actual reference of PA content using the best models NIR by PLS-SVR with SG 2nd preprocessing methods are shown in Fig. 6.

Table 5
Statistical evaluation of the regression models for PA content determination using PLS-SVR

Preprocessing	N	Hyperparameter			Calibration		Cross-validation		Prediction		RPD
		γ	C	ϵ	R^2_C	RMSEC	R^2_{cv}	RMSECV	R^2_P	RMSEP	
None	275	0.0001	10	0.75	0.8007	3.2685	0.7779	4.3508	0.7841	3.9287	2.165
Autoscale	275	0.0001	10	0.75	0.8007	3.2685	0.7779	4.3508	0.7841	3.9287	2.165
SG 1st	273	0.0001	50	0.5	0.9327	1.8959	0.9179	2.0949	0.8779	2.9542	2.879
SG 2nd	273	0.0001	50	0.25	0.9182	2.1344	0.9124	2.2085	0.9317	2.2095	3.849
SNV	281	0.0001	5	0.75	0.7223	4.0775	0.7046	4.2058	0.6419	5.0597	1.681
MSC	281	0.0001	5	0.9	0.7242	4.0632	0.7055	4.1989	0.6387	5.0825	1.152

N = number of spectra in calibration dataset after outlier removal

The best PLS-SVR (SG 2nd) model resulted in better predictive performance, as evidenced by higher R^2_P (> 0.9) and RPD (> 3.5) values, as well as lower RMSEP compared to the best PLSR and SVR models. Moreover, using this preprocessing technique, no model overfitting was observed, as indicated by minimal discrepancy between R^2_C and R^2_P . This demonstrates that preprocessing NIR spectra with SG 2nd, followed by PLS dimensionality reduction, effectively removed irrelevant features for PA prediction. Furthermore, SVR model's using RBF kernel successfully captured the non-linear relationships from the PLS-reduced data.

Similar to PLSR, the most accurate prediction of PLS-SVR model was obtained using NIR data preprocessed by SG 2nd, but with addition of SVR tuning parameters of $C = 50$, $\epsilon = 0.25$, and $\gamma = 0.0001$. Moreover, the effect of autoscaling did not result in any difference compared to the original spectra, similar with PLSR result. This suggests that the default autoscaling in the Python PLSRegression library is performed internally by PLS algorithm itself. Consequently, the model performances results for both the original spectra and the autoscaled were identical.

The optimized PLS-SVR (SG 2nd) has higher prediction accuracy over both the best PLSR (SG 2nd) and the full-spectrum SVR model (SG 1st). A significant improvement was observed in the RPD values, where PLS-SVR achieved a higher RPD (3.849) compared to PLSR (3.126) and SVR (2.612). This indicates that combining PLS-based dimensionality reduction with nonlinear SVR enhances the model performance at handling spectral data for PA prediction.

These results are in agreement with [27], who reported that volatile compounds in *Mentha haplocalyx* leaves could be predicted significantly better using PLS-SVR ($C = 31.6228$, $\gamma = 0.0031623$, $\epsilon = 0.1$) than PLSR based on NIR spectroscopy. The PLS-SVR model achieved R^2 values greater than 0.9 for both calibration and prediction sets, along with low RMSEC and RMSEP values (< 0.1). In contrast, the PLSR model produced R^2 values between 0.8 and 0.9, with RMSEC and RMSEP values generally exceeding 0.1. Quantification of essential oil content in spearmint (*Mentha spicata*) leaves using NIR spectroscopy coupled with PLS-SVR also demonstrated better predictive ability compared to PLSR and SVR [26].

The improved performance of the PLS-SVR models in predicting PA content after PLS-based dimensionality reduction in patchouli oil may be attributed to several factors, including better handling of non-linearity relationship and/or the characteristics of sample constituents [55]. In addition, the selection of RBF as a non-linear kernel plays an important role in improving model performance. The choice of kernel is highly impactful on the execution of SVR. Several studies in NIR

application coupled with SVR used RBF kernel due to the ability in reducing computing intricacy of the calibration set and the excellent performance compared to others [27, 56]. In the context of essential oil authenticity, SVR was also used to estimate adulterant levels. Using an RBF kernel, the detection of castor oil levels as an adulterant in sandalwood oil could be predicted accurately ($R^2_p = 0.9998$ and RMSEP 0.00804) with C , γ , and ϵ values of 31.6228, 0.001, and 0.01, respectively [22].

4. Conclusion

In conclusion, PA content from patchouli oil can be predicted using NIR spectroscopy coupled with multivariate and machine learning methods. Among the various methods evaluated, including PLSR (linear regression), SVR (non-linear regression), and PLS-SVR, the PLS-SVR model using SG 2nd derivative demonstrated the best predictive ability as evidenced by higher R^2 values and lower error performance, particularly in the prediction set ($R^2_p = 0.9317$, RMSEP = 2.2095). RPD obtained from this method was also satisfactory, as the values were close to four, indicating excellent model prediction. Furthermore, the sample size and range of patchouli oil concentration used in this study were quite large, suggesting strong model generalization and has potential for practical application.

Declarations

Author Contribution

N. Ainina: Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization; S. Nurjanah: Writing – Review & Editing, Validation, Supervision, Funding Acquisition, Conceptualization; U. R. Sinniah : Methodology, Supervision, Validation; R. Hidayat : Methodology, Supervision, Validation; E. Julaeaha: Methodology, Supervision, Validation; M. A. Nanda : Methodology, Supervision, Validation; M. Muhaemin : Writing – Review & Editing, Validation, Conceptualization, Supervision, Methodology, Software, Data Curation.

Data availability statement

The data that supports the results of this study are included in the supplementary material of this paper.

Declaration of competing interest

The authors declare that they have no conflict of interest

Funding

This work was supported with the financial support of the Universitas Padjadjaran through the Padjadjaran Doctoral Program Scholarship (2203/UN6.3.1/PT.00/2022) with Sarifah Nurjanah as the promotor.

Acknowledgement

This publication charge is funded by Unpad through the Indonesian Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology and managed under the EQUITY Program (Contract No. 4303/B3/DT.03.08/2025 and 3927/UN6.RKT/HK.07.00/2025).

References

1. M. K. Swamy and U. R. Sinniah, "A Comprehensive Review on the Phytochemical Constituents and Pharmacological Activities of Pogostemon cablin Benth.: An Aromatic Medicinal Plant of Industrial Importance," *Molecules*, vol. 20, 2015, <https://doi.org/10.3390/molecules20058521>
2. P. Astuti, Khairan, Marthoenis and K. Hasballah, "Antidepressant-like Activity of Patchouli Oil var. Tapak Tuan (Pogostemon cablin Benth) via Elevated Dopamine Level: A Study Using Rat Model," *Pharmaceuticals*, vol. 15, 2022, <https://doi.org/10.3390/ph15050608>

3. J. He, H.M.Chen, C.W.Li, D.W.Wu, X.L.Wu, S.J.Shi, Y.C.Li, J.N.Chen, Z.R.Su and X.P.Lai, "Experimental study on antinociceptive and anti-allergy effects of patchouli oil," *Journal of Essential Oil Research*, vol. 25, no. 6, pp. 488-496, 2013, <https://doi.org/10.1080/10412905.2013.809319>
4. Y. Shin, S. Lee, J. Lee, P. Kang and G. H. Seol, "Effects of Short-Term Inhalation of Patchouli Oil on Professional Quality of Life and Stress Levels in Emergency Nurses : A Randomized Controlled Trial," *The Journal of Alternative and Complementary Medicine*, vol. 26, 2020, <https://doi.org/10.1089/acm.2020.0206>
5. A. Karimi, "Characterization and antimicrobial activity of patchouli essential oil extracted from pogostemon cablin [Blanco] Benth. [lamiaceae]," *Advances in Environmental Biology*, 2014.
6. S. K. Pandey, S. Bhandari, N. Sarma, T. Begum, S. Munda, J. Baruah, R. Gogol, S. Haldar and M. Lal, "Essential oil compositions, pharmacological importance and agro technological practices of Patchouli (Pogostemon cablin Benth.): A review," *Journal of Essential Oil Bearing Plants*, vol. 24, 2021, <https://doi.org/10.1080/0972060X.2021.1995511>
7. H. S. Lee, J. Lee, D. Smolensky and S. H. Lee, "Potential benefits of patchouli alcohol in prevention of human diseases: A mechanistic review," *International Immunopharmacology*, vol. 89, 2020, <https://doi.org/10.1016/j.intimp.2020.107056>
8. A. M. Elfiky, E. Shawky, A. R. Khatib and R. S. Ibrahim, "Integration of NIR spectroscopy and chemometrics for authentication and quantitation of adulteration in sweet marjoram (*Origanum majorana* L.)," *Microchemical Journal*, vol. 183, 2022, <https://doi.org/10.1016/j.microc.2022.108125>
9. X. Li, L. Zhang, Y. Zhang, D. Wang, X. Wang, L. Yu, W. Zhang and P.Li, "Review of NIR spectroscopy methods for nondestructive quality analysis of oilseeds and edible oils," *Trends in Food Science & Technology*, vol. 101, 2020, <https://doi.org/10.1016/j.tifs.2020.05.002>
10. F. K. G. S. N. Kollman, P. Mishra, L. Zhang and A. J. v. d. Goot, "Near-infrared spectroscopy-based quantification of sunflower oil and pea protein isolate in dense mixtures for novel plant-based products," *Journal of Food Composition and Analysis*, vol. 121, 2023, <https://doi.org/10.1016/j.jfca.2023.105414>
11. Y. Yang, R. Mao, L. Yang, J. Liu, S. Wu, M. Wu, X. Zhang, H. Tong and X. Ji, "Rapid and comprehensive quality assessment of Bupleuri Radix through near-infrared spectroscopy combined with chemometrics," *Infrared Physics and Technology*, vol. 121, 2022, <https://doi.org/10.1016/j.infrared.2022.104051>
12. N. Celikbas, A. Kaya and M. Turkmen, "Application of Near Infrared Spectroscopy Combined with Multivariate Analysis for Screening Foliar Main Essential Oil Components in Bay Laurel," *Engenharia Agricola*, vol. 42, 2022, <https://doi.org/10.1590/1809-4430-Eng.Agric.v42n2e20200040/2022>
13. B. Wedding, R. White, S. Grauf, B. Tilse and P. Gadek, "Near infrared spectroscopy as a rapid method for sandalwood oil determination," *Sandalwood Research Newsletter*, vol. 23, pp. 5-7, 2008.
14. S. Lafhal, P. Vanloot, I. Bombarda, J. Kister and N. Dupuy, "Chemometric analysis of French lavender and lavandin essential oils by near infrared spectroscopy," *Industrial Crops and Products*, vol. 80, 2016, <https://doi.org/10.1016/j.indcrop.2015.11.017>
15. H. Andrianoelisoa, P. Danthu, M. Tomazello, C. Menut and G. Chaix, "Near Infrared (NIR) spectroscopy for determination of essential oil chemotypes from *Ravensara aromatica*," in *16th International Conference on Near Infrared Spectroscopy*, la Grande-Motte, 2013.
16. N. Amusant, J. Beauchene, A. Digeon and G. Chaix, "Essential oil yield in rosewood (*Aniba rosaedora* Ducke): initial application of rapid prediction by near infrared spectroscopy based on wood spectra," *Journal of Near Infrared Spectroscopy*, vol. 24, 2016, <https://doi.org/10.1255/jnirs.1241>
17. Y. W. Wu, S. Q. Sun, Q. Zhou and H. W. Leung, "Fourier transform mid-infrared (MIR) and near-infrared (NIR) spectroscopy for rapid quality assessment of Chinese medicine preparation Honghua Oil," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 46, 2008, <https://doi.org/10.1016/j.jpba.2007.11.021>
18. M. Sandasi, G. P. P. Kamatou, C. Gavaghan, M. Baranska and A. M. Viljoen, "A quality control method for geranium oil based on vibrational spectroscopy and chemometric data analysis," *Vibrational Spectroscopy*, vol. 57, 2011, <https://doi.org/10.1016/j.vibspec.2011.08.002>

19. D. A. Selim, R. S. Darwish and E. Shawky, "Authentication and detection of common adulterants in clove buds (*Syzygium aromaticum* L.) powders and oils using near IR combined to multivariate analysis," *Microchemical Journal*, vol. 191, 2023, <https://doi.org/10.1016/j.microc.2023.108890>
20. A. C. O. Moreira, A. H. L. Machado, F. V. Almeida and J. W. B. Braga, "Rapid Purity Determination of Copaiba Oils by a Portable NIR Spectrometer and PLSR," *Food Analytical Methods*, vol. 11, 2018, <https://doi.org/10.1007/s12161-017-1079-8>
21. J. B. Johnson, P. R. Thani and M. Naiker, "Through-container detection of tea tree oil adulteration using near-infrared spectroscopy (NIRS)," *Chemical Papers*, vol. 77, 2022, <https://doi.org/10.1007/s11696-022-02603-4>
22. S. Kuriakose and H. Joe, "Qualitative and quantitative analysis in sandalwood oils using near infrared Is using near infrared," *Food Chemistry*, vol. 135, 2012, <https://doi.org/10.1016/j.foodchem.2012.04.073>
23. E. Shawky, L. Nahar, S. M. Nassief, S. D. Sarker and R. S. Ibrahim, "Spice authentication by near-infrared spectroscopy: Current advances, limitations, and future perspectives," *Trends in Food Science & Technology*, vol. 148, 2024, <https://doi.org/10.1016/j.tifs.2024.104522>
24. H. Zhang, X. Hu, L. Liu, J. Wei and X. Bian, "Near infrared spectroscopy combined with chemometrics for quantitative analysis of corn oil in edible blend oil," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 270, 2022, <https://doi.org/10.1016/j.saa.2021.120841>
25. H. N. Moghaddam, Z. Tamiji, M. A. Lakeh, M. R. Khoshayand and M. H. Mahmoodi, "Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics," *Journal of Food Composition and Analysis*, vol. 107, 2022, <https://doi.org/10.1016/j.jfca.2021.104343>
26. S. Haute, A. Nikkhah, D. Malavi and S. Kiani, "Prediction of essential oil content in spearmint (*Mentha spicata*) via near infrared hyperspectral imaging and chemometrics," *Scientific Report*, vol. 13, 2023, <https://doi.org/10.1038/s41598-023-31517-8>
27. H. Yan, C. Guo, Y. Shao and Z. Ouyang, "Rapid Detection of Volatile Oil in *Mentha haplocalyx* by NearInfrared Spectroscopy and Chemometrics," *Pharmacognosy Magazine*, vol. 13, 2016. <https://doi.org/10.4103/0973-1296.211026>
28. C. Chen, H. Li, X. Lv, J. Tang, C. Chen and X. Zheng, "Application of near infrared spectroscopy combined with SVR algorithm in rapid detection of cAMP content in red jujube," *Optik*, vol. 194, 2019. <https://doi.org/10.1016/j.ijleo.2019.163063>
29. C. Malegori, E. J. N. Marques, S. T. Freitas, M. F. Pimentel, C. Pasquini and E. Casiraghi, "Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms," *Talanta*, vol. 165, 2017. <https://doi.org/10.1016/j.talanta.2016.12.035>
30. V. Windawati, Y. A. Purwanto, I. W. Budiastara, S. Widodo and S. Kuroki, "Prediction of chemical content in patchouli oil (*Pogostemon cablin* Benth) by portable NIR spectroscopy," in *IOP Conference Series: Materials Science and Engineering*, Bogor, 2019. <https://doi.org/10.1088/1757-899X/557/1/012087>
31. Zulfahrizal, S. Muhammad and A. A. Munawar, "Rapid determination of patchouli oil quality as a tropical plant using infrared technology," in *IOP Conference Series: Earth and Environmental Science*, Banda Aceh, 2021. <https://doi.org/10.1088/1755-1315/922/1/012027>
32. D. M. Cano-Reinoso, Y. A. Purwanto, I. W. Budiastara, S. Kuroki, S. Sutrisno and S. Widodo, "Evaluation of the patchouli essential oil (*Pogostemon cablin* Benth.) aromatic characteristic by near-infrared spectroscopy," *Indonesian Journal of Biotechnology*, vol. 28, 2023. <https://doi.org/10.22146/ijbiotech.69073>
33. A. Rinnan, F. v. d. Berg and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, 2009. <https://doi.org/10.1016/j.trac.2009.07.007>
34. A. Agussabti, R. Rahmaddiansyah, P. Satriyo and A. A. Munawar, "Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in Aceh Province, Indonesia," *Data in brief*, vol. 29, no. 105251, 2020. <https://doi.org/10.1016/j.dib.2020.105251>
35. G. Tsegay, Y. Ammare and S. Mesfin, "Development of non-destructive NIRS models to predict oil and major fatty acid contents of Ehiopian sesame," *Journal of Food Composition and Analysis*, vol. 115, 2023.

<https://doi.org/10.1016/j.jfca.2022.104908t>

36. Y. Zhao, C. Zhu and H. Jiang, "Quantitative detection of zearalenone in wheat using intervals selection coupled to near-infrared spectroscopy," *Infrared Physics and Technology*, vol. 136, 2024. <https://doi.org/10.1016/j.infrared.2023.105004>
37. J. S. Shenk, J. J. W. Jr. and M. O. Westerhaus, "Application of NIR Spectroscopy to Agricultural Products," in *Handbook of Near-Infrared Analysis (Third Edition)*, Boca Raton, CRC Press, 2008, pp. 347 - 386.
38. L. G. Weyer and S.-C. Lo, "Spectra– Structure Correlations in the Near-Infrared," in *Handbook of Vibrational Spectroscopy*, Chichester, John Wiley & Sons, Ltd, 2006, pp. 1817 - 1836.
39. B. G. Osborne, "Near-Infrared Spectroscopy in Food Analysis," *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, pp. 1-14, 2006.
40. M. Manley, G. Downey and V. Baeten, "Spectrometric Technique: Near-infrared (NIR) Spectroscopy," in *Modern Techniques for Food Authentication*, Dublin, Elsevier, 2008, pp. 65-115.
41. X. Bian, Y. Liu, R. Zhag, H. Sun, P. Liu and X. Tan, "Rapid quantification of grapeseed oil multiple adulterations using near-infrared spectroscopy coupled with a novel double ensemble modelling method," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 311, 2024. <https://doi.org/10.1016/j.saa.2024.124016>
42. H. B. Qu, D. L. Ou and Y. Y. Cheng, "Background correction in near-infrared spectra of plant extracts by orthogonal signal correction," *Plant & Animal Sciences and Biotechnology*, vol. 6, 2005. <https://doi.org/10.1631/jzus.2005.B0838>
43. P. Mishra, D. N. Rutledge, J. M. Roger, K. Wali and H. A. Khan, "Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction," *Talanta*, vol. 229, 2021. <https://doi.org/10.1016/j.talanta.2021.122303>
44. M. Zeaiter, J. M. Roger and V. B. Maurel, "Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods," *Trends in Analytical Chemistry*, vol. 24, pp. 437 - 445, 2005. <https://doi.org/10.1016/j.trac.2004.11.023>
45. A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," in *Statistic and Computing*, Kluwer Academic Publishers, 2004, pp. 199-222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
46. O. A. M. Lopez, A. M. Lopez and J. Crossa, "Support Vector Machines and Support Vector Regression," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer, 2022, pp. 337 - 378. https://doi.org/10.1007/978-3-030-89010-0_9
47. S. Chanda, A. K. Hazarika, N. Choudhury, S. A. Islam, R. Manna, S. Sabhapondit, B. Tudu and R. Bandyopadhyay, "Support vector machine regression on selected wavelength regions for quantitative analysis of caffeine in tea leaves by near infrared spectroscopy," *Journal of Chemometrics*, vol. 33, 2019. <https://doi.org/10.1002/cem.3172>
48. S. Kuriakose, X. Thankappan, H. Joe and V. Venkataraman, "Detection and quantification of adulteration in sandalwood oil through near infrared spectroscopy," *Analyst*, vol. 135, 2010. <https://doi.org/10.1039/C0AN00261E>
49. C. Alferis and G. Simon, "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*, Cham, Springer, 2024, pp. 477-524. https://doi.org/10.1007/978-3-031-39355-6_10
50. T. Leng, F. Li, Y. Chen, L. Tang, J. Xie and Q. Yu, "Fast quantification of total volatile basic nitrogen (TVB-N) content in beef and pork by near-infrared spectroscopy: Comparison of SVR and PLS model," *Meat Science*, vol. 180, 2021. <https://doi.org/10.1016/j.meatsci.2021.108559>
51. Y. Zhao, Y. Zhu, Y. Ren, Y. Lu, C. Yu, G. Chen, Y. Hong and Q. Liu, "Predictive Modeling of Lignocellulosic Content in Crop Straws Using NIR Spectroscopy," *Plants*, vol. 14, 2025. <https://doi.org/10.3390/plants14101430>
52. Y. Xiao, C. Li, C. Jin, J. Luo, H. Qi and C. Zhang, "Detection of soluble solid content in citrus fruit using near-infrared spectroscopy with machine learning regression: An exploration of the influence of sampling positions," *Journal of Food Composition and Analysis*, vol. 142, 2025. <https://doi.org/10.1016/j.jfca.2025.107554>
53. S. Velliangiri, S. .Alagumuthukrishnan and S. I. T. Joseph, "A Review of Dimensionality Reduction Techniques for Efficient Computation," *Procedia Computer Science*, vol. 165, 2019. <https://doi.org/10.1016/j.procs.2020.01.079>

54. B. C. Deng, Y. H. Yun, Y. Z. Liang, D. S. Cao, Q. S. Xu, L. Z. Yi and X. Huang, "A new strategy to prevent over-fitting in partial least squares models based on model population analysis," *Analytica Chimica Acta*, vol. 880,, 2015.
<https://doi.org/10.1016/j.aca.2015.04.045>
55. F. C. B. Bedin, M. V. Faust, G. A. Guarneri, T. S. Assmann, C. B. B. Lafay, L. F. Soares, P. A. V. Oliveira and L. M. Santos-Tonial, "NIR associated to PLS and SVM for fast and non-destructive determination of C, N, P, and K contents in poultry litter," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 245, 2021.
<https://doi.org/10.1016/j.saa.2020.118834>
56. S. Sarkar, J. K. Basak, B. E. Moon and H. T. Kim, "A Comparative Study of PLSR and SVM-R with Various Preprocessing Techniques for the Quantitative Determination of Soluble Solids Content of Hardy Kiwi Fruit by a Portable Vis/NIR Spectrometer," *Foods*, vol. 9, 2020. <https://doi.org/10.3390/foods9081078>

Figures

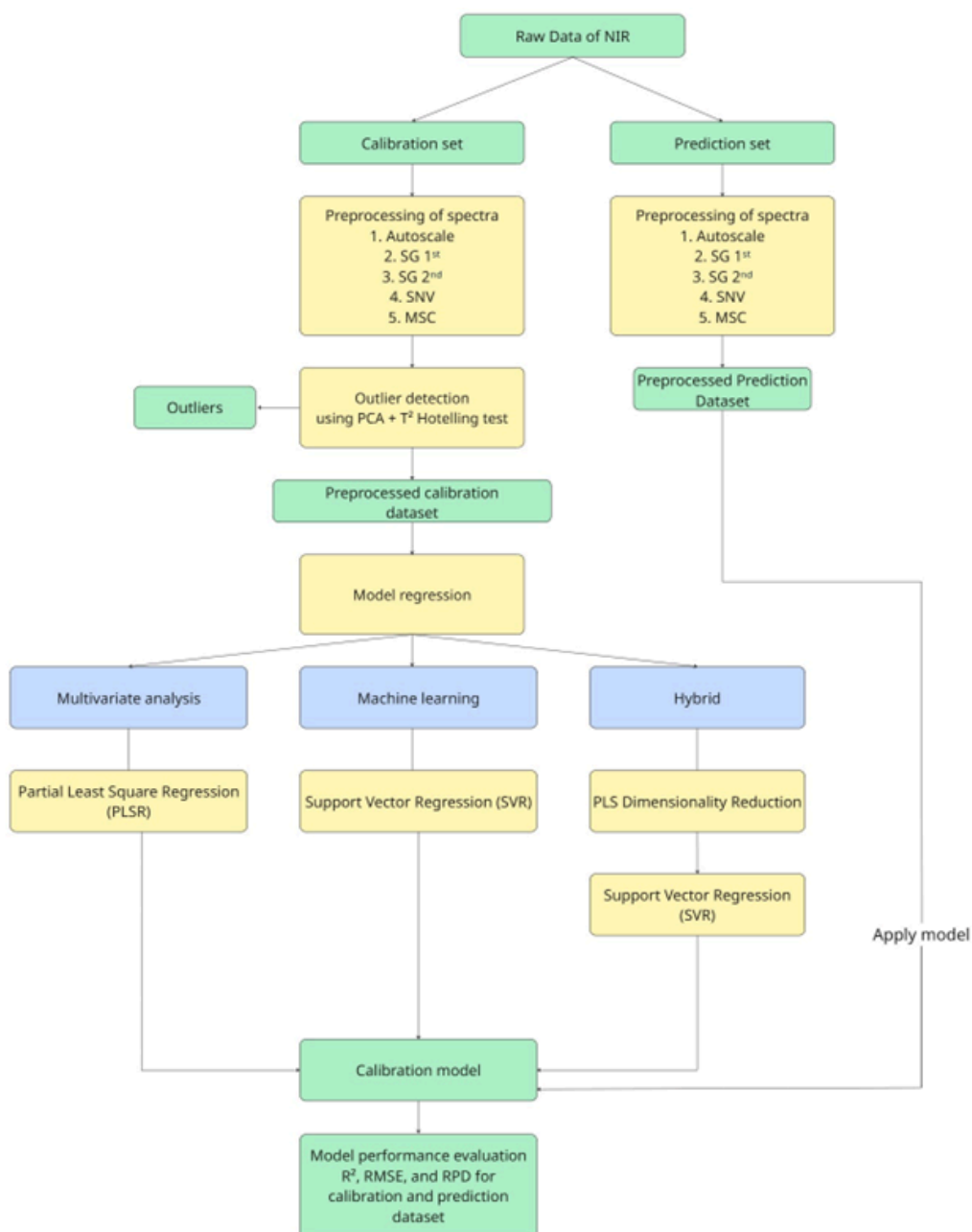


Figure 1

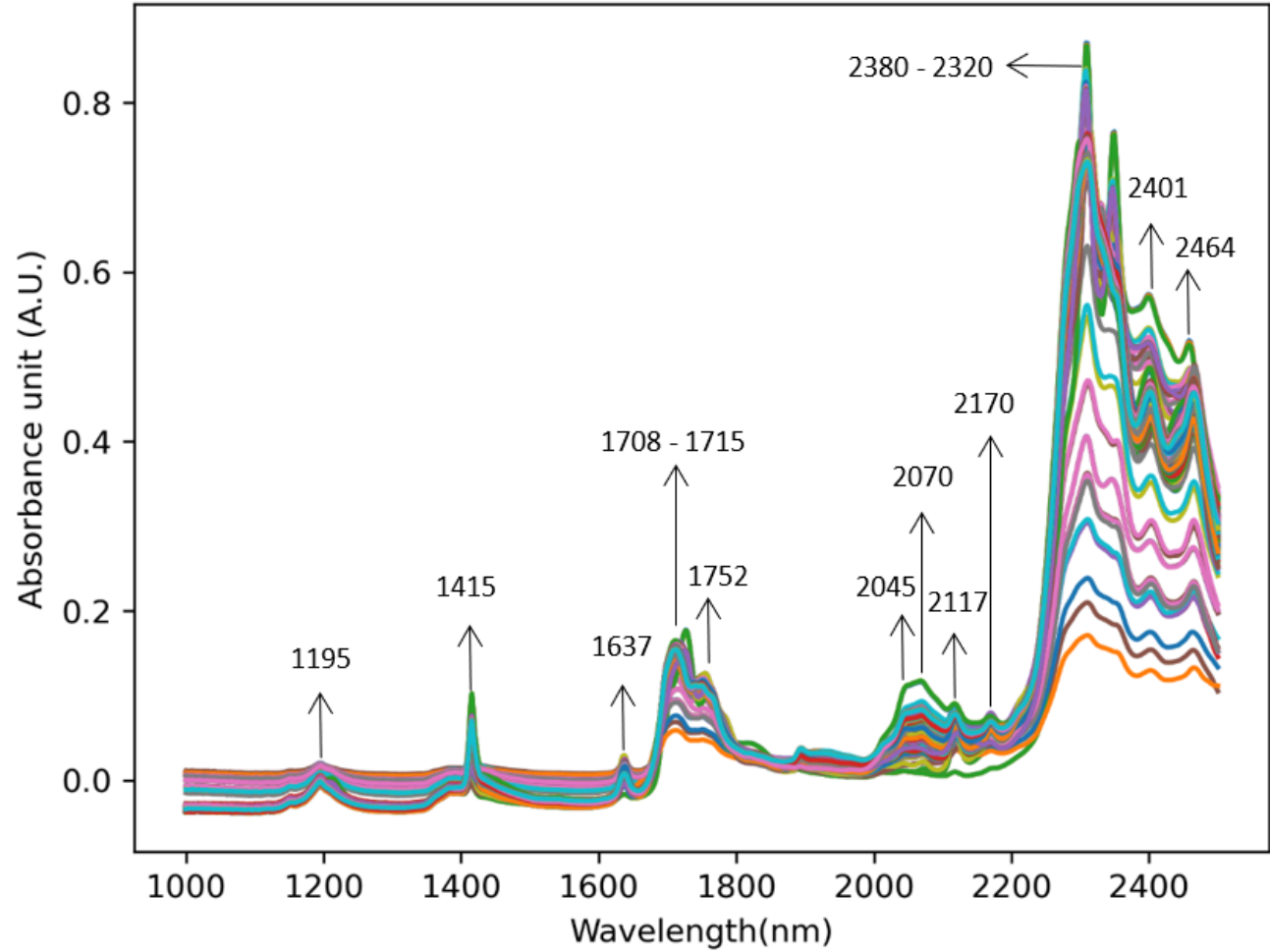


Figure 2
NIR spectra of 100 patchouli oil without preprocessing

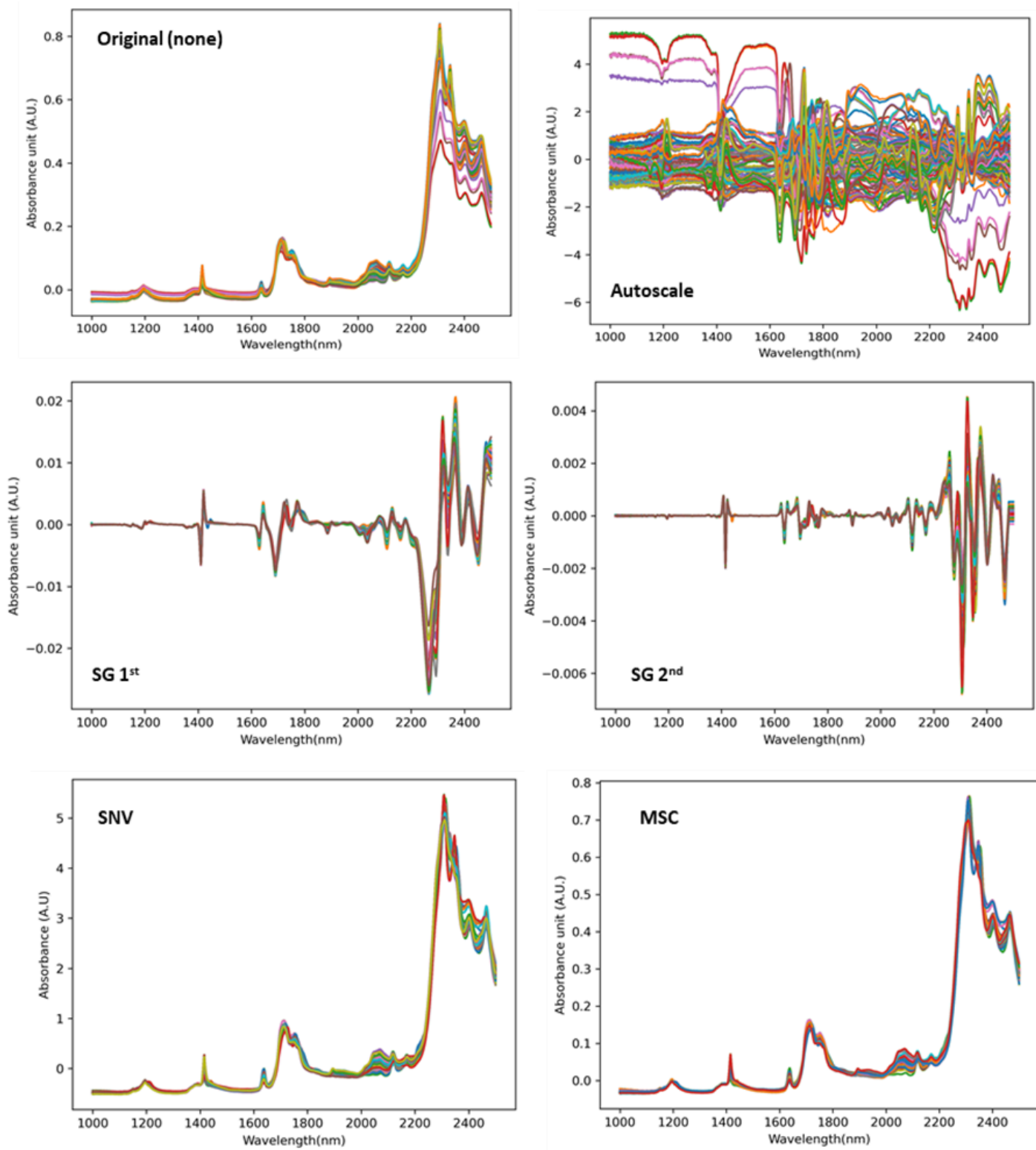


Figure 3

Original and pretreated of NIR spectra (after preprocessing and outlier removal)

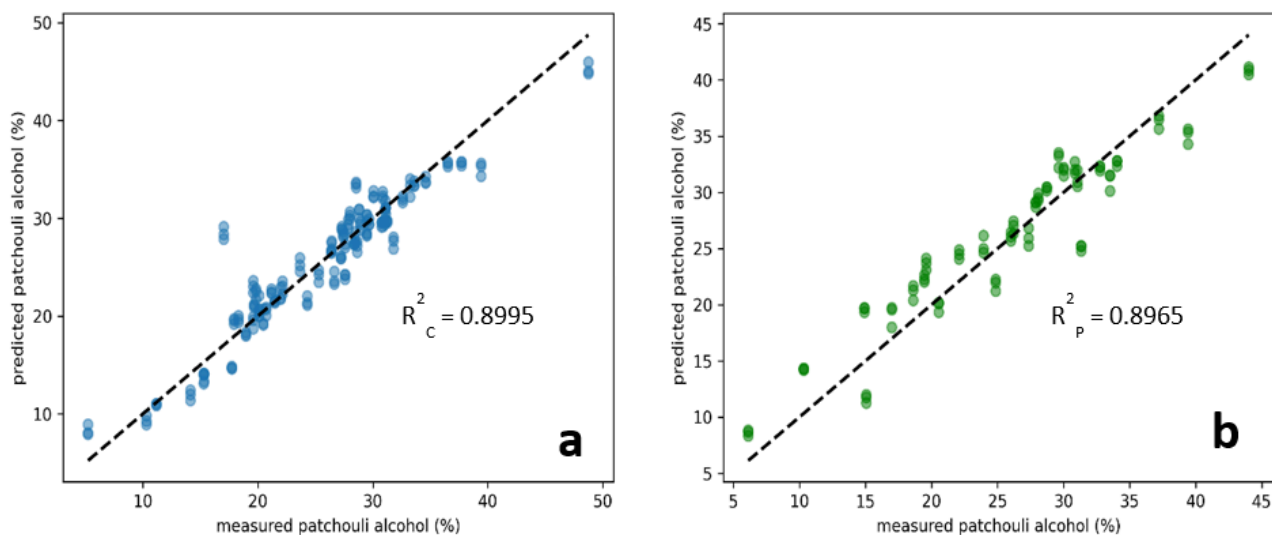


Figure 4

Graphics of the best models NIR for PA Prediction using SG 2nd preprocessing coupled with PLSR. (a) Calibration set, and (b) Prediction set

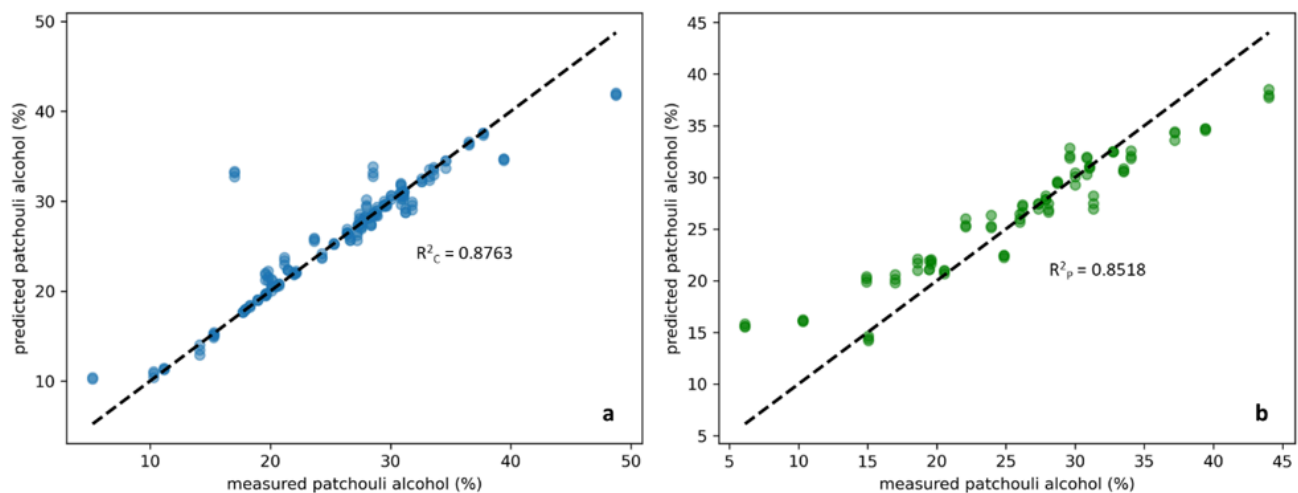


Figure 5

Graphics of the best models NIR for PA Prediction using SG 1st preprocessing coupled with SVR. (a) Calibration set, and (b) Prediction set

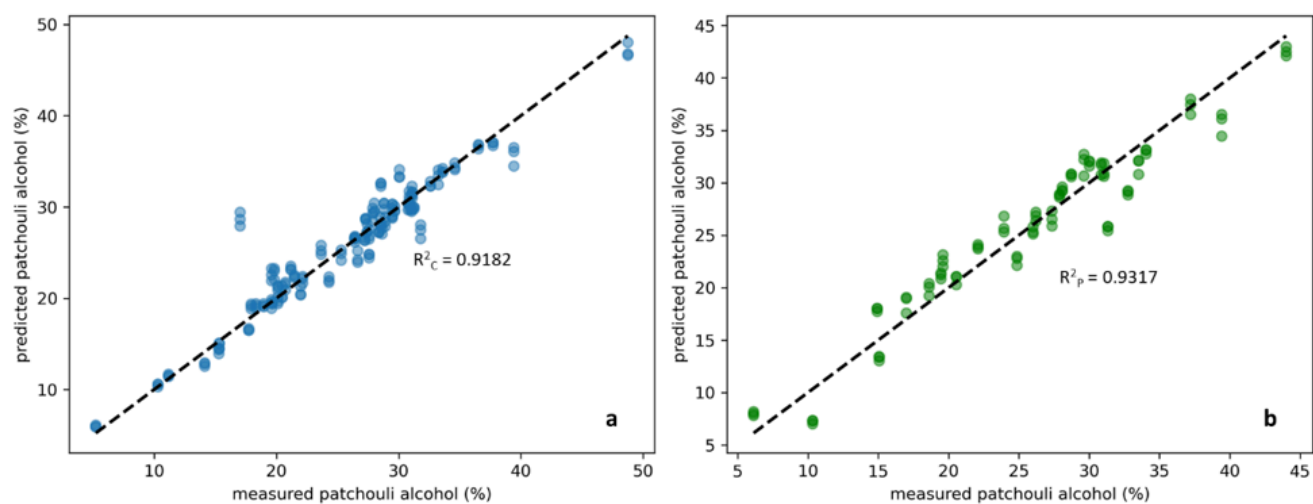


Figure 6

Graphics of the best models NIR for PA Prediction using SG 2nd + MC preprocessing coupled with PLS-SVR approach. (a) Calibration set, and (b) Prediction set

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NIRDataScientificReport.xlsx](#)