JST: Engineering and Technology for Sustainable Development
Volume 33, Issue 1, March 2023, 001-008

Rapid Identification of the Geographical Origin of Black Pepper in Vietnam Using Near-Infrared Spectroscopy and Chemometrics

Le Tuan Phuc[1*], Tran Thi Thanh Hoa[1], Cung Thi To Quynh[1], Nguyen Hoang Dzung[2], Pham Ngoc Hung[1]

[1]*School of Biotechnology and Food Technology, Ha Noi University of Science and Technology, Ha Noi, Vietnam*
[2]*Faculty of Chemical Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam*
[*]*Corresponding author email: phuc.lt211194m@sis.hust.edu.vn*

Abstract

The aim of this study was to develop a rapid method for determining geographical origin of black pepper based on combination of near-infrared spectroscopy and chemometric algorithms. Near-infrared spectroscopy was used to determine the geographical origin of 111 black pepper samples from three regions in Vietnam: North Central Region, Central Highlands and Southeast Region. A portable NIR spectrometer was used to collect the spectra of intact black pepper. Support vector machine (SVM), linear discriminant analysis (LDA) and K-nearest neighbors (KNN) were established and the identification effect of these models with different pre-processing methods were optimized and compared. By using support vector machine (SVM) classification and standard normal variate (SNV) spectral pre-processing, classification accuracy was 98.7% on the calibration and 100% on the validation sets for the determination of origin of black pepper. This study has demonstrated that portable spectrometers can be effective tools to use in the classification of black pepper samples according to their geographical origin.

Keywords: NIR, black pepper, geographical origin, chemometrics, SVM, KNN, LDA.

1. Introduction

Black Pepper (*Piper nigrum L.*) is one of the most widely consumed spices in the world [1]. Pepper is rich in piperine, essential oil, and is a source of numerous biological activities that are beneficial for human health [2]. Piperine is the major alkaloid that constitutes 98% of the total alkaloids in pepper, the presence of piperine brings a characteristic pungent taste to pepper [3]. Piperine is associated with antioxidant, antimicrobial; anti-carcinogenic, anti-inflammatory, and anti-ulcer [4]. It also exhibits diuretic properties and breast cancer [1].

The quality of pepper from different regions can vary significantly, so the origin serves as the primary criterion for the purchasing department when making purchasing decisions. In recent years, regulations have been put in place to verify the authenticity of the geographical origin in order to fight against adulteration and misbranding of foods [5]. Various methods available for the identification of geographic of black pepper have been used such as DNA Fingerprinting, Gas Chromatography-Mass Spectrometry (GC-MS) and Volatile Organic Compounds (VOCs). These methods have high accuracy but can be time-consuming and expensive. Consequently, to make the pepper trading process more convenient and quicker, it is necessary to develop a method that is simple, cost-effective and rapid for identifying the geographical origin of pepper.

Near-infrared spectroscopy, combined with chemometrics tools, is an effective and widely used method for non-destructive testing of agricultural products, particularly in the identification of geographical origin [6]. This technique is not only rapid but also does not require complex chemical pre-treatment, which has received more and more attention in recent years [7]. Chemometrics models extract useful information from multivariate data obtained from spectroscopy by using mathematical and statistical techniques [8]. The application of NIR spectra and chemometrics has been shown to achieve a high degree of accuracy in determining geographical origin. For example, Saffron can be identified with over 93% accuracy [9]; Coffee can be identified with more than 98% accuracy by continent and 100% accuracy by country [10]; Curcumae Radix can be identified with 100% accuracy. Besides, quantitative models based on NIR and chemometrics have also been developed and shown excellent results, such as the determination of melamine in milk with R^2 and the root mean square error (RMSE) values of 0.98 and 0.56%, respectively [12]; the determination of lipid and protein content in green coffee with $R^2 > 0.982$ and RMSEP < 0.106 [13]; and the determination of soluble solids content and titratable acidity of mango with R^2 of 0.91 and 0.98, respectively [14].

ISSN 2734-9381
https://doi.org/10.51316/jst.164.etsd.2023.33.1.1
Received: December 7, 2022; accepted: February 2, 2023

The aim of this study is to develop a highly accurate, non-destructive analytical method for determining the geographical origin of Black Pepper in Vietnam through the use of Near-infrared (NIR) spectroscopy and chemometrics. Despite the numerous cases of success using portable NIR spectrometers in food analysis, there is currently a lack of studies applying NIR spectroscopy to the discrimination of geographical origins in Vietnamese Pepper. The spectral data were processed through Principal Component Analysis (PCA) and then subjected to classification using Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA) algorithms. The optimization of the parameters of the classification model was accomplished through the use of the grid search algorithm, thereby improving its predictiveness and accuracy.

2. Material and Method

2.1. Materials

2.1.1. Black pepper samples

The set of pepper samples used in this study included a total of 111 samples: 19 samples from the North Central Region (Quang Tri province); 55 samples from the Central Highlands, including Dak Lak (21), Dak Nong (18), and Gia Lai (16); 37 samples from the South East Region from Ba Ria - Vung Tau (19) and Dong Nai (18) from December 2021 to March 2022. Pepper samples were preliminarily processed, which were accepted for inclusion in the study only if their moisture content was below 15%. Subsequently, the samples were preserved in PE zip-lock bags at room temperature for later study.

2.1.2. Experimental instruments

The NIR spectra of samples were measured by using a Digital Light Processing (DLP) NIRscan™ Nano portable spectrometer (Texas Instrument, USA) in the wavelength range of 900-1700 nm. This spectrometer is a compact battery-operated evaluation module (EVM) that uses a DLP micromirror array, a single InGaAs detector, and two tungsten halogen lamps.

The software package DLPR350 (Texas Instrument, USA) can be installed on a computer to help manipulating measurements and storing them directly on a personal computer via a micro-USB cable. In this experiment, the spectra were acquired in reflective mode, but the software interface converted and saved the spectra in an absorbance unit.

Fig. 1. DLP NIRscan Nano EVM spectrometer

2.2. Methods

2.2.1. Spectra collection

Spectra of the collected samples were recorded with a NIR spectrometer fixed in a closed box to prevent external light and vibration during the experiment. The settings of scan mode, scan range, and resolution were adjusted to the reflectance mode, 900 - 1700 nm and 2 nm, respectively. The sample, which weighed about 30 grams, was poured into a white porcelain cup. Then, the glass rod was slid over the mouth of the cup to ensure that the measuring surface was as flat as possible. The sample was analyzed ten times by returning it to the PE zip-lock bags and shaking the bag between each measurement. The average of 6 iterations under the spectrometer's default settings was taken for each scan.

2.2.2. Data analysis methods

In this study, the process of data mining was divided into three stages: (1) data pre-processing; (2) the establishment and training of the classification model; (3) the validation of the classification model. The flow chart for data treatment was depicted in Fig. 2. The spectra were pre-processed using the following methods: standard normal variate (SNV), first-order polynomial (SG1), second-order polynomial (SG2), combined standard normal variate and first-order polynomial (SNV + SG1), combined standard normal variate and second-order polynomial (SNV + SG2).

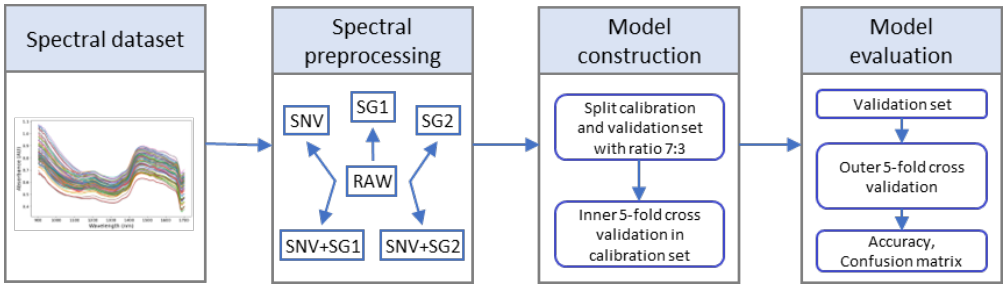

Fig. 2. The flowchart for data treatment

- **Spectral data pre-processing**

NIR spectra are usually pre-processed before performing analyses in order to remove possible light-scattering effects [15]. The standard normal variate transformations (SNV) and the multiplicative scatter correction (MSC) were employed to mitigate multiplicative scatter effect [15]. Besides, the Savitzky-Golay algorithm with a first-order polynomial (SG1) or second-order polynomial (SG2) was performed to remove noise and smoothing the spectra [16]. In this study, the above pre-processing methods were applied individually and combined in series to optimize the predictive power of the classification models. The pre-processing method and classification model with the highest accuracy were selected to establish the final models. All the spectral pre-processing were performed in Python with pyspectra package.

- **Model construction**

In this study, analysis was performed using both supervised and unsupervised learning algorithms. With unsupervised algorithm, the Principal Component Analysis (PCA) was applied to visualize the present data. PCA is a popular and efficient tool for extracting features of spectral data. This method not only reduces the dimension of the complex and highly dimensional data but also extracts the information with the most significant difference and diagnostic significance for classification. The supervised classification methods that were used in this study include Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA). SVM algorithm generates an optimal hyperplane to classify different sample classes in a dataset, while LDA algorithm seeks to maximize the ratio between between-class variance and within-class variance by linear combination. KNN is one of the simplest supervised-learning algorithms, used for prediction based on the nearest data point in the training set.

The original data set was divided randomly into two datasets using a stratified train-test split. Since the dataset does not have a balance for each class label, it is desirable to split the dataset into train and test sets in a way that preserves the same proportions of examples in each class as observed in the original dataset. The two datasets, the calibration set and the validation set, were used for training and validation purposes, with the calibration set accounts for 70% for training model and the validation set 30%.

- **Model evaluation**

Cross-validation was performed to avoid overfitting and optimize model parameters. The optimization of the model parameters was carried out through the inner cross-validation, which was conducted only on the calibration set, consisting of 70% of the original data set. Meanwhile, the outer cross-validation was performed on both calibration and validation sets, consisting of 30% of the original data set, to evaluate the performance of the models. Both cross-validation were done with 5-fold cross-validation. The results on each dataset were the mean of these cross-validations and the standard deviation between the estimates. Models were evaluated by accuracy (*Acc*) according to (1)

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

where *TP* is true positive, *FN* is false negative, *TN* is true negative, and the *FP* is false positive.

- **Software**

All the spectra pre-processing method, chemometrics model to classification, tools for visualizing data and algorithm for optimizing the classification model were done using Python in version 3.9.0 with libraries such as Numpy, Pandas, Scikit-learn, Matplotlib, and Scipy.

3. Results and Discussion

3.1. NIR Spectroscopy of Black Pepper

The NIR spectrum collected during the study is characterized by absorption bands related to different oscillations of organic bonds present in the samples. In pepper samples grown from different regions, the observed absorption bands were related to the characteristic organic compounds of pepper, mainly piperine, volatile essential oil and content ash. The NIR spectra of the samples are shown in Fig. 3a, it can be seen that there were two distinct absorption peaks at 1200 nm and 1470 nm, which were related to chemical variations in the samples of different geographical origin.

The application of SNV pre-treatment to eliminate light scattering in pepper samples is shown in Fig. 3b. A slight difference in the absorption spectrum of pepper from different growing regions is still seen. The absorption peak at 1432 nm, related to the O-H absorption of water, is also observed in the NIR spectrum of black pepper and is made clearly visible through the Savitzky-Golay derivative transformation as seen in Fig. 3c or Fig. 3d [17]. Especially, the absorption peaks at 1323 and 1390 nm, which are observed in the black pepper spectra using the second derivative (Fig. 3d), are related to the fluctuations of the main structural functional groups such as piperine {1- [5- (1,3-Benzodioxol -5-yl) -1-oxo-2,4-pentadienyl] piperidina}, piperonol {1,3-Benzodioxole-5-methanol} and caryophyllene {β-caryophyllene}[18]

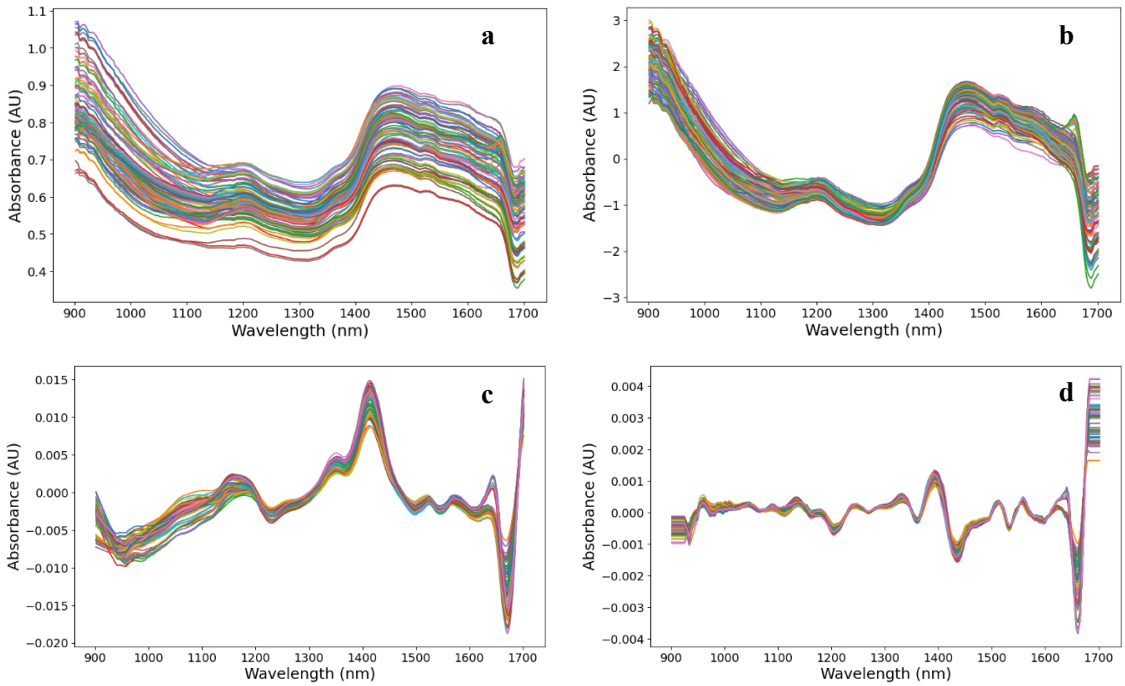

Fig. 3. Black pepper NIR spectra (a) Raw spectra (b) Spectra treated with SNV (c) Spectra treated with SG1 (d) Spectra treated with SG2

3.2. Principal Component Analysis

The application of PCA to all SNV pre-processed NIR datasets for the evaluation of separation based on the growing regions of pepper was carried out.

The plot for the first two principal components PC1 and PC2 (Fig. 4), explained 69.46%, 27.64% of the total variance, respectively and cumulatively explained 97.1% of the total variance shown in Fig. 5a. It was observed that samples from the 3 regions of North Central, Central Highlands and South East were almost separated (Fig. 4). The North Central samples, which have negative PC2 scores, are clearly separated from the positive PC2 scores of South East samples. This indicated that NIR combined with PCA can effectively determine the geographical origin of black pepper.

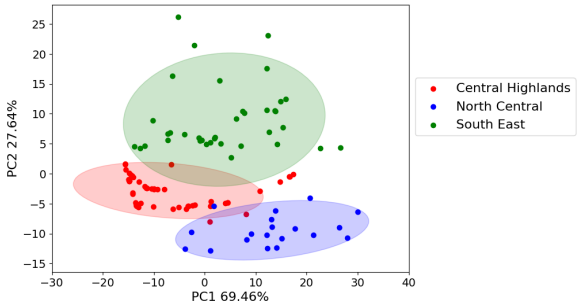

Fig. 4. PCA scores plot of NIR spectra with SNV with confidence ellipses at the 95% confidence interval

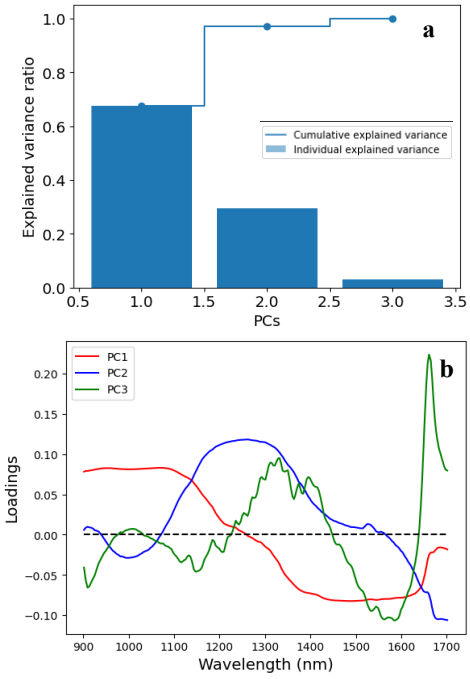

Fig. 5. (a) Cumulative variance contribution rate curve for the first 3 principle components (b) Loading plot from PCA of spectral data of NIR spectra

In Fig. 5b, it is demonstrated that spectral loadings show the wavelengths that contributed the most to the formation of the first three principal components. In this plot, the wavelengths between 1300 - 1450 nm have large loadings score for both

PC1, PC2, and PC3, implying that they contain a significant amount of information in describing the differences in the geographic origin of the samples. Furthermore, a shift in the loadings score for 3 PCs is observed in 1600-1700 nm band, which is related to the resonance band of the O-H bond.

3.3. Geographical Origin Classification of Black Pepper

To optimize the classification models, different types of pre-processing such as SG1, SG2, and SNV were used independently or in combination to generate common transformations before putting into model classification. From 3 classifiers and 5 pre-processing, a total of 18 models were trained through calibration sets and validation sets.

3.3.1. Classification with SVM algorithm

SVM algorithms were used to classify peppers according to their geographical origin. In this study, the RBF (Radial Basic Function) kernel is used as the kernel for the SVM algorithm due to its popularity for the classification problem. Hyperparameters were optimized through a cross-validated grid search of the training set, with the 'C' parameter from 10^{-2} to 10^{5} and the 'gamma' parameter from 10^{-5} to 1. The results of the 5-fold cross-validation, carried out on both calibration and validation sets, are presented in Table 1.

The results of SVM analysis showed that the ability to recognize the geographical origin of pepper is very effective. An accuracy rate higher than 90% in the validation sets was achieved for different preprocessing methods. In particular, the accuracy of the calibration set for SNV and SG1 pre-processing was recorded at 98.7% and 97.3% respectively and reached 100% in the validation sets.

The Confusion matrix in Fig. 6 details the predicted results and actual results of the observations in both calibration and validation sets. There was only one sample from the North Central that was predicted to be from the Central Highlands region.

3.3.2. Classification with LDA algorithm

LDA is widely used for data classification, which attempts to find linear combinations of features to find the best linear fit for separating two or more classes. The LDA model in the study uses the "SVD" solver and the number of components examined is from 1 to 2. The results are presented in Table 2.

In Table 2, the results of the classification using the LDA model were much worse compared to those of the SVM, as demonstrated by the fact that none of the validation set results have an accuracy of more than 90%. The best accuracy levels achieved by the LDA model, when using SG1 pre-processor with 1 component, were 79.3% and 88.6% for the calibration and validation sets, respectively.

Table 1. Compare the classification results of the SVM model for different preprocessing methods.

Pre-process	Parameter		Accuracy	
	C	Gamma	Calibration set	Validation set
Raw	1000	0.01	0.987 ± 0.027	0.943 ± 0.114
SNV	10	0.01	0.987 ± 0.027	1.000 ± 0.000
SG1	100000	1	0.973 ± 0.033	1.000 ± 0.000
SG2	100000	1	0.934 ± 0.042	0.914 ± 0.114
SNV+SG1	100	0.01	0.973 ± 0.033	0.971 ± 0.057
SNV+SG2	1000	0.001	0.960 ± 0.053	0.943 ± 0.114

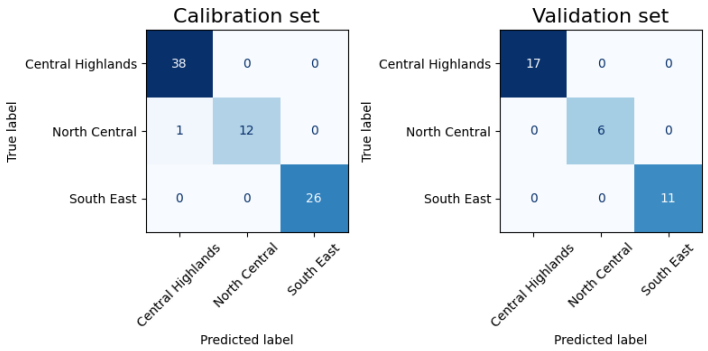

Fig. 6. Confusion matrix of black pepper classification based on NIR spectra with SVM and SNV

Table 2. Compare the classification results of the LDA model for different preprocessing methods

Pre-process	Parameter	Calibration set	Validation set
	n_components		
Raw	1	0.807 ± 0.067	0.886 ± 0.107
SNV	1	0.908 ± 0.052	0.829 ± 0.277
SG1	1	0.793 ± 0.083	0.886 ± 0.14
SG2	1	0.832 ± 0.094	0.800 ± 0.114
SNV+SG1	1	0.922 ± 0.028	0.824 ± 0.139
SNV+SG2	1	0.844 ± 0.031	0.829 ± 0.167

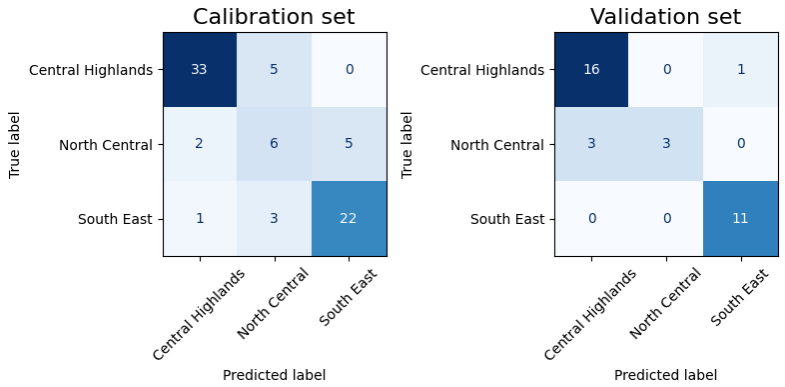

Fig. 7. Confusion matrix of black pepper classification based on NIR spectra with LDA and SNV+SG2

In Fig. 7, it can be seen that many samples from the North Central were misclassified and predicted as belonging to other regions, in both calibration set and validation set. The results showed that the sensitivity and precision of the North Central region samples on LDA model were poor. It was difficult for the model to recognize these regional patterns in the data set.

3.3.3. Classification with KNN algorithm

KNN is an easy, simple, and low-cost algorithm that also works well with small and multivariate datasets. The KNN model in this study is used and optimized on two parameters: n_neighbors (from 1 to 10) and weight function ("uniform" or "distance"). The obtained results are presented in Table 3.

The accuracy of the validation set obtained from different pre-processing methods was found to be similar. The KNN model achieved the best performance when both SNV and SG1 pre-processing were applied with n_neighbors equal 2 and a weight function set to uniform. The accuracy was recorded 97.4% and 97.1% in the calibration set and validation set, respectively.

In Fig. 8, one sample from the North Central region was predicted to be Central Highlands in both the calibration and validation sets. However, only one sample in the Central Highlands was predicted to be from the North Central in the calibration sets.

Table 3. Compare the classification results of the KNN model for different preprocessing methods

Pre-process	Parameter		Calibration set	Validation set
	n_neighbors	weight		
Raw	1	uniform	0.897 ± 0.052	0.886 ± 0.167
SNV	3	distance	0.961 ± 0.053	0.971 ± 0.057
SG1	1	uniform	0.961 ± 0.032	0.971 ± 0.057
SG2	1	uniform	0.962 ± 0.031	0.943 ± 0.114
SNV+SG1	1	uniform	0.974 ± 0.032	0.971 ± 0.057
SNV+SG2	1	uniform	0.961 ± 0.053	0.943 ± 0.114

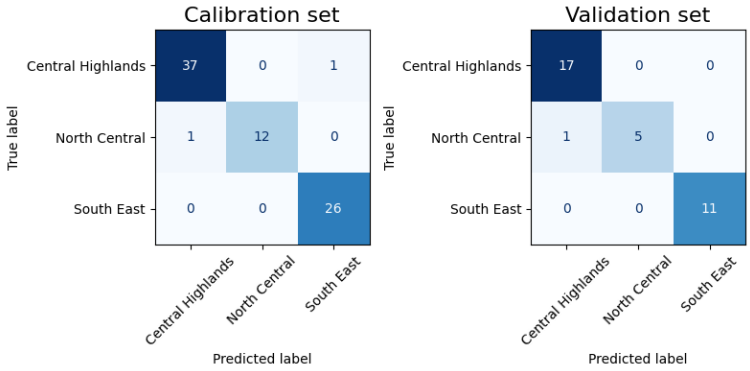

Fig. 8. Confusion matrix of black pepper classification based on NIR spectra with KNN and SG1

4. Conclusion

In this study, the feasibility of using a combination of near-infrared spectroscopy and classical statistical machine learning algorithms to classify the geographical origin of black pepper from three main regions in Vietnam was demonstrated. The differences between the spectral plots of black pepper samples grown from different regions were attributed to differences in climate and soil. Three classification models and five different spectral pre-treatment methods were investigated. The results showed that SNV pre-processing combined with SVM, the accuracy of the model reached 98.7% and 100% after 5-fold cross-validated of the calibration set and validation sets, respectively. This outcome offers a useful, simple, and fast solution to verify the geographical origin of black pepper in Vietnam.

Acknowledgments

This work was supported by the Ministry of Science and Technology of Vietnam in the project "Research and application of Blockchain technology to manage the production and consumption chain of Vietnamese pepper" (No: KC-4.0-24/19-25).

References

[1] S. Shityakov *et al.*, Phytochemical and pharmacological attributes of piperine: A bioactive ingredient of black pepper, Eur. J. Med. Chem., vol. 176, pp. 149-161, 2019, https://doi.org/10.1016/j.ejmech.2019.04.002.

[2] M. Nikolić *et al.*, Could essential oils of green and black pepper be used as food preservatives?, J. Food Sci. Technol., vol. 52, no. 10, pp. 6565-6573, 2015, https://doi.org/10.1007/s13197-015-1792-5.

[3] Y. Wang, L. Chen, K. Chaisiwamongkhol, R. G. Compton, Electrochemical quantification of piperine in black pepper, Food Chem., vol. 309, 125-606, 2020, https://doi.org/10.1016/j.foodchem.2019.125606.

[4] M. Meghwal and T. K. Goswami, Piper nigrum and piperine: An update, Phytotherapy Research, vol. 27, no. 8. pp. 1121-1130, 2013. https://doi.org/10.1002/ptr.4972.

[5] L. Hu, C. Yin, S. Ma, and Z. Liu, Assessing the authenticity of black pepper using diffuse reflectance mid-infrared Fourier transform spectroscopy coupled with chemometrics, Comput. Electron. Agric., vol. 154, no. February, pp. 491-500, 2018, https://doi.org/10.1016/j.compag.2018.09.029.

[6] F. R. Huang *et al.*, Determination of chinese honey adulterated with syrups by near infrared spectroscopy combined with chemometrics, Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy Spectr. Anal., vol. 39, no. 11, pp. 3560-3565, Nov. 2019,

[7] M. J. Aliaño-González, M. Ferreiro-González, E. Espada-Bellido, M. Palma, and G. F. Barbero, A screening method based on Visible-NIR spectroscopy for the identification and quantification of different adulterants in high-quality honey, Talanta, vol. 203, no. January, pp. 235-241, 2019, https://doi.org/ 10.1016/j.talanta.2019.05.067.

[8] P. Theanjumpol *et al.*, Non-destructive identification and estimation of granulation in 'sai Num Pung' tangerine fruit using near infrared spectroscopy and chemometrics, Postharvest Biol. Technol., vol. 153, pp. 13-20, Jul. 2019, https://doi.org/ 10.1016/j.postharvbio.2019.03.009.

[9] S. Li, Q. Shao, Z. Lu, C. Duan, H. Yi, and L. Su, Rapid determination of crocins in saffron by near-infrared spectroscopy combined with chemometric techniques, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc., vol. 190, pp. 283-289, Feb. 2018, https://doi.org/ 10.1016/j.saa.2017.09.030.

[10] A. Giraudo, S. Grassi, F. Savorani, G. Gavoci, E. Casiraghi, and F. Geobaldo, Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis, Food Control, vol. 99, pp. 137-145, May 2019, https://doi.org/ 10.1016/J.FOODCONT.2018.12.033.

[11] L. Wang *et al.*, Fast discrimination and quantification analysis of Curcumae Radix from four botanical origins using NIR spectroscopy coupled with chemometrics tools, Spectrochim. Acta Part A Mol. Biomol. Spectrosc., vol. 254, p. 119626, Jun. 2021, https://doi.org/ 10.1016/J.SAA.2021.119626.

[12] F. Mabood *et al.*, Robust Fourier transformed infrared spectroscopy coupled with multivariate methods for

detection and quantification of urea adulteration in fresh milk samples, Food Sci. Nutr., vol. 8, no. 10, pp. 5249-5258, 2020, https://doi.org/ 10.1002/fsn3.987.

[13] M. Zhu *et al.*, Fast determination of lipid and protein content in green coffee beans from different origins using NIR spectroscopy and chemometrics, J. Food Compos. Anal., vol. 102, p. 104055, Sep. 2021, https://doi.org/ 10.1016/J.JFCA.2021.104055.

[14] B. M. Nicolaï *et al.*, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review, Postharvest Biol. Technol., vol. 46, no. 2, pp. 99-118, Nov. 2007, https://doi.org/10.1016/J.POSTHARVBIO.2007.06.024.

[15] J. Engel *et al.*, Breaking with trends in pre-processing?, TrAC - Trends Anal. Chem., vol. 50, pp. 96-106, 2013, https://doi.org/ 10.1016/j.trac.2013.04.015.

[16] Å. Rinnan, F. van den Berg, and S. B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem., vol. 28, no. 10, pp. 1201-1222, 2009, https://doi.org/ 10.1016/j.trac.2009.07.007.

[17] A. S. Wilde, S. A. Haughey, P. Galvin-King, and C. T. Elliott, The feasibility of applying NIR and FT-IR fingerprinting to detect adulteration in black pepper, Food Control, vol. 100, pp. 1-7, Jun. 2019, https://doi.org/ 10.1016/J.FOODCONT.2018.12.039.

[18] I. Orrillo *et al.*, Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper, Food Control, vol. 101, pp. 45-52, Jul. 2019, https://doi.org/ 10.1016/j.foodcont.2019.02.036.