DOI: 10.6060/ivkkt.20256807.7246

ГЕОГРАФИЧЕСКИЙ ДИСКРИМИНАНТ И КЛАССИФИКАЦИЯ КОРИЦЫ КАССИИ, СОБРАННОЙ ВО ВЬЕТНАМЕ, С ИСПОЛЬЗОВАНИЕМ ИК-ФУРЬЕ-СПЕКТРОСКОПИИ ATR В СОЧЕТАНИИ С АЛГОРИТМАМИ МАШИННОГО ОБУЧЕНИЯ

Буи Тхи Лан Фуонг, Хоанг Тхи Бич, Нгуен Ван Фуонг, Буи Ан Дуй, Нгуен Дук Фонг, Нгуен Мань Сон, Фам Зиа Бах, Нгуен Тхи Кам Ха, Та Тхи Тхао, Нгуен Тхи Киеу Ань

Буй Тхи Лан Фуонг, Хоанг Тхи Бич, Нгуен Ван Фуонг, Буй Ан Дуй, Нгуен Тхи Кьеу Ань*

Ханойский фармацевтический университет, Ханой, Вьетнам

Нгуен Мань Сон, Фам Гиа Бах, Нгуен Тхи Кам Ха, Та Тхи Тао

Факультет химии, Научный университет ВНУ, Ханой, Вьетнам

Нгуен Дык Фонг

Компания TRAPHACO, Ханой, Вьетнам

Во Вьетнаме Cinnamomum cassia широко используется не только как пряность, но и как ключевой ингредиент в медицине, косметике и перерабатывающей промышленности. Различные климатические условия в различных географических регионах, где выращивается корица, значительно влияют на ее качество, что часто приводит к непреднамеренному смешиванию корицы из разных источников. Для анализа образцов корицы использовалась инфракрасная спектроскопия с нарушенным полным внутренним отражением и преобразованием Фурье (ATR-FTIR) для измерения 139 образцов, собранных в четырех различных провинциях Вьетнама – Йенбай, Куангнинь, Тханьхоа и Куангнам – каждая из которых представляла уникальные климатические условия. Для повышения качества данных был применен метод сглаживания второй производной Савицкого-Голея, что улучшило спектральное разрешение и снизило шум. Для оценки потенциала классификации использовались неконтролируемые хемометрические методы, включая анализ главных компонентов (PCA) и иерархический кластерный анализ (HCA). Кроме того, контролируемые модели машинного обучения, такие как Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), GausianNB, BernoulliNB, AdaBoost и Gradient Boosting, были объединены с результатами PCA и обучены с использованием сглаженных спектров первой производной Савицкого-Голея для классификации. Результаты показали, что модель PCA-SVM достигла наивысшей точности классификации (96.88%) со спектрами второй производной, тогда как первая производная и необработанные данные показали более низкую производительность. В этом исследовании подчеркивается, что спектроскопия ATR-FTIR в сочетании с предварительной обработкой второй производной и PCA-SVM обеспечивает простой, быстрый, неразрушающий и экономически эффективный подход для различения образцов коры корицы на основе их географического происхождения, предлагая ценную информацию для оценки качества.

Ключевые слова: ATR-FTIR, Cinnamon cassia, производные спектров, географическая классификация, машинное обучение

Для цитирования:

Буи Тхи Лан Фуонг, Хоанг Тхи Бич, Нгуен Ван Фуонг, Буи Ан Дуй, Нгуен Дук Фонг, Нгуен Мань Сон, Фам Зиа Бах, Нгуен Тхи Кам Ха, Та Тхи Тхао, Нгуен Тхи Киеу Ань Географический дискриминант и классификация корицы кассии, собранной во Вьетнаме, с использованием ИК-Фурье-спектроскопии ATR в сочетании с алгоритмами машинного обучения. *Изв. вузов. Химия и хим. технология*. 2025. Т. 68. Вып. 7. С. 102–113. DOI: 10.6060/ivkkt.20256807.7246.

For citation:

Bui Thi Lan Phuong, Hoang Thi Bich, Nguyen Van Phuong, Bui An Duy, Nguyen Duc Phong, Nguyen Manh Son, Pham Gia Bach, Nguyen Thi Cam Ha, Ta Thi Thao, Nguyen Thi Kieu Anh Geographical discriminant and classification of cinnamomum cassia collected in Vietnam using ATR-FTIR coupled with machine learning algorithms. *ChemChemTech* [*Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol*.]. 2025. V. 68. N 7. P. 102–113. DOI: 10.6060/ivkkt.20256807.7246.

GEOGRAPHICAL DISCRIMINANT AND CLASSIFICATION OF CINNAMOMUM CASSIA COLLECTED IN VIETNAM USING ATR-FTIR COUPLED WITH MACHINE LEARNING ALGORITHMS

Bui Thi Lan Phuong, Hoang Thi Bich, Nguyen Van Phuong, Bui An Duy, Nguyen Duc Phong, Nguyen Manh Son, Pham Gia Bach, Nguyen Thi Cam Ha, Ta Thi Thao, Nguyen Thi Kieu Anh

Bui Thi Lan Phuong, Hoang Thi Bich, Nguyen Van Phuong, Bui An Duy, Nguyen Thi Kieu Anh*
Hanoi University of Pharmacy, Hanoi, Vietnam

Nguyen Manh Son, Pham Gia Bach, Nguyen Thi Cam Ha, Ta Thi Thao
Faculty of Chemistry, VNU University of Science, Hanoi, Vietnam

Nguyen Duc Phong
TRAPHACO Company, Hanoi, Vietnam

In Vietnam, Cinnamomum cassia is widely utilized not only as a spice but also as a key ingredient in medicine, cosmetics, and the processing industry. The varying climatic conditions across different geographic regions where cinnamon is cultivated significantly influence its quality, often leading to the unintentional mixing of cinnamon from multiple sources. To analyze cinnamon samples, Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy (ATR-FTIR) was employed to measure 139 samples collected from four distinct provinces in Vietnam – Yen Bai, Quang Ninh, Thanh Hoa, and Quang Nam – each representing unique climatic conditions. To enhance data quality, the second derivative Savitzky-Golay smoothing method was applied, improving spectral resolution and reducing noise. Unsupervised chemometric techniques, including Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA), were used to assess classification potential. Additionally, supervised machine learning models, such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), GausianNB, BernoulliNB, AdaBoost, and Gradient Boosting were combined with PCA results and trained using first derivative Savitzky-Golay smoothed spectra for classification. The results demonstrated that the PCA-SVM model achieved the highest classification accuracy (96.88%) with the second derivative spectra, whereas first derivative and raw data exhibited lower performance. This study underscores that ATR-FTIR spectroscopy, when combined with second-derivative preprocessing and PCA-SVM, provides a simple, rapid, non-destructive, and cost-effective approach for distinguishing cinnamon bark samples based on their geographical origin, offering valuable insights for quality assessment.

Keywords: ATR-FTIR, Cinnamon cassia, derivative spectra, geographical classification, machine learning

INTRODUCTION

Cinnamon (*Cinnamomum spp.)* is cultivated worldwide and widely used in food, pharmaceuticals, and cosmetics. As a member of the *Lauraceae* family [1], it thrives in tropical and subtropical climates, making Vietnam an ideal location for its cultivation. The primary species grown in Vietnam, *Cinnamomum cassia*, is distributed across multiple regions [2, 3].

The quality of cinnamon depends on its chemical composition, which is influenced by factors such as plant variety, cultivation practices, processing methods, and environmental conditions, including climate and soil composition [4, 5]. Recognizing this, Vietnam has designated four regions – Thuong Xuan, Van Yen, Tra Bong, and Tra My – as geographically protected cinnamon-growing areas. Vietnam's diverse climate and topography contribute to regional variations in cinnamon composition [6]. The four regions examined in this study represent distinct climatic conditions. Yen Bai (Northwest Vietnam) has a warm, mountainous climate influenced by cold winds and flash floods. Quang Ninh (Northeast Vietnam) experiences colder weather with a humid monsoon climate and frequent tropical storms. Thanh Hoa (North Central Vietnam) has mild winters, high rainfall, and hot summers affected by foehn winds. Quang Nam (South Central Vietnam) features a dry climate with mild winters and low precipitation. These environmental differences significantly impact cinnamon's chemical profile, further underscoring the importance of classifying it based on geographical origin to ensure authenticity and product integrity [7].

The geographical origin of cinnamon is often undisclosed by industries, especially when different varieties or sources are blended [8]. With consumers becoming more conscious of food quality and safety, there is increasing scrutiny over food authenticity. Media reports on food fraud have heightened concerns, leading to growing demands for stricter regulatory oversight [9-10]. The most common forms of economically motivated adulteration include the mixing of premium with lower quality such as *Cinnamomum verum* and *Cinnamomum cassia* [11] or *Bunium persicum* mixed with *Cuminum cyminum, Safed zeera* [12], or mixing products of different origins such as green tea products [13], ... practices that occur frequently in the supply chain.

Analytical approaches for cinnamon authentication fall into two categories: targeted and non-targeted analysis. Targeted analysis focuses on detecting or quantifying specific compounds to verify authenticity, often relying on chromatographic techniques due to the complexity of cinnamon's chemical composition. For example, HPLC-UV has been used to quantify key cinnamon compounds for species discrimination [14-16], while UPLC-MS and GC-MS [17-19], and nuclear magnetic resonance (NMR) [18]. Non-targeted analysis, on the other hand, does not depend on specific compounds but instead utilizes signals from unidentified components (e.g., chromatographic peak areas) or instrumental data (e.g., selected NIR spectral regions) [20]. Among the various methods used in non-targeted analysis, FTIR spectroscopy stands out for its rapid analysis, high accuracy, and ease of use. When integrated with machine learning techniques, FTIR becomes a powerful tool for classifying samples based on their geographic origin. Previous studies have successfully applied FTIR-based classification to herbal products such as *Paris yunnanensis* [21], *Ganoderma lucidum* [22], and *Cinnamomum verum* [23]. In addition to classifying plant-based products, FTIR spectra are also combined with machine learning to identify synthetic products such as food colors E110, E124 [24]. However, no study has yet explored the classification of Vietnamese cinnamon based on its geographical origin using the combined approach of FTIR and machine learning.

While targeted analytical methods necessitate extensive sample preparation, gradient elution protocols, and may lack specific biomarkers associated with geographic origin, non-targeted analysis accounts for the inherent variability in the chemical composition of cinnamon from different regions. Consequently, non-targeted analysis represents a more suitable approach for the authentication of cinnamon. This approach broadens detection scope, it often fails to provide detailed chemical information directly related to sample authenticity, particularly when spectroscopic techniques are employed. This study aims to develop a rapid and accurate classification method for determining the geographical origin of cinnamon by integrating ATR-FTIR spectroscopy with various machine learning algorithms. Both linear models (including SVM with a linear kernel, GausianNB, and BernoulliNB) and non-linear models (including RF, DT, AdaBoost, and GradientBoost) were employed. These models were combined with dimensionality reduction techniques and preprocessing methods, such as the derivative Savitzky-Golay method, to optimize variable inputs through spectrum preprocessing, hyperparameter tuning, and wavelength selection. The entire data processing and model development were carried out using open-source Python, ensuring scalability and potential integration into broader classification systems in factories.

METHODOLOGY

Geographical location and climate of the studied sites

Yen Bai province, located in the Northeast region, experiences cold, cloudy winters with minimal sunshine and frequent drizzle. Summers are hot and rainy, aligning with the monsoon season. The northeast monsoon influences the region, causing early cold spells compared to other provinces. Quang Ninh province, also in the Northeast, shares similar climatic characteristics with Yen Bai. Winters are cold, cloudy, and drizzly, while summers are hot and rainy, coinciding with the monsoon period. The northeast monsoon significantly impacts Quang Ninh, leading to an early onset of cold weather. Thanh Hoa province, situated in the North Central Coast region, has mildly cold winters with occasional drizzles. Summers are hot and humid, with substantial rainfall during the monsoon season. The province's climate is influenced by both northern and central weather patterns, resulting in diverse conditions. Quang Nam Province, in the South-Central Coast region, experiences mild winters with less pronounced cold spells. Summers are hot and humid, with the rainy season typically occurring from September to December [25].

Sample Collection and Preparation

Cinnamomum cassia bark samples were collected from four provinces in Vietnam, including: Yen Bai (48 samples), Quang Ninh (20 samples), Thanh Hoa (22 samples), and Quang Nam (49 samples). Each sample was labeled and stored in a dry environment. A 10 g portion of each sample was ground and sieved to obtain a particle size of 125–250 μm. The powdered

samples were vacuum-sealed and stored at -30 °C for no longer than two weeks before analysis. Moisture content was measured before analysis to ensure levels remained below 15%.

IR Measurements

FTIR spectra were obtained using an IR Affinity-1S spectrometer (Shimadzu, Japan) equipped with a diamond ATR crystal. Measurements were conducted at room temperature over the spectral range of 4000-600 cm^{-1}, with a data interval of 2 cm^{-1} and an average spectrum derived from 40 scans.

Data Preprocessing

In practice, FTIR spectroscopy is highly sensitive to background interference, light scattering, fluctuating noise levels, and other unpredictable factors. Preprocessing plays a vital role in transforming raw data into clean, structured data, thereby improving the accuracy of the developed model. Multivariate data processing is essential for minimizing variations caused by instrumental differences and environmental conditions during spectral acquisition. In this study, normalization using Min-Max scaling and first- and second-order Savitzky-Golay derivative filtering [26] were applied.

Exploratory Data Analysis

Unsupervised learning techniques were utilized to assess the data structure and clustering tendencies. Hierarchical Cluster Analysis (HCA) [27] was applied to investigate relationships between samples and identify similarities, while Principal Component Analysis (PCA) [28] was used for dimensionality reduction and pattern recognition within the dataset. Based on the results from unsupervised learning in identification, it is a fundamental step to proceed with supervised learning algorithms.

Supervised machine learning models were applied to classify the *Cinnamomum cassia* samples. Before training the model, Savitzky-Golay smoothing was applied to raw data to remove noises. The standard normal variate (SNV) was used correctly for the light scattering effect in spectral data. Savitsky-Golay first and second derivative techniques are used to enhance the signal-to-noise ratio. PCA was also integrated into studied models to improve classification efficiency.

The tested models included both linear and non-linear approaches. The linear models comprised Support Vector Machine (SVM) with a linear kernel [29], Gaussian Naïve Bayes (GNB) [30], and Bernoulli Naïve Bayes (BNB) [31]. The non-linear models included Random Forest (RF) [32], Decision Tree (DT) [33], AdaBoost, and Gradient Boosting [34]. The dataset was split into a training set (70%) and a test set (30%). All data processing and model training were performed using Python 3.11.6, following the workflow illustrated in Fig. 1.

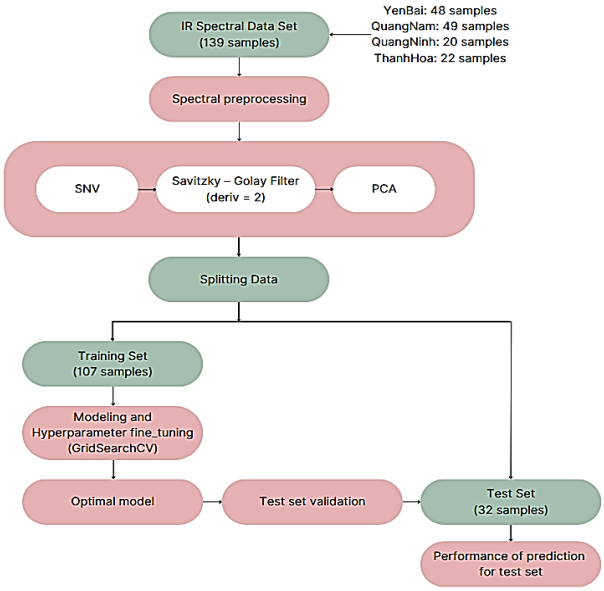


Fig. 1. Flow diagram for preprocessing of FTIR spectra and supervised learning models

Рис. 1. Блок-схема предварительной обработки спектров FTIR и контролируемых моделей обучения

RESULT AND DISCUSSION

FTIR fingerprints of Vietnamese Cinnamomum cassia samples

The IR spectral signals within the range of 4000-600 cm^{-1} were extracted into an Excel dataset, resulting in a dataset comprising 139 samples and 1701 columns representing intensity values. The output data was labeled as discrete variables: Yen Bai – YB (Label: 1); Quang Ninh – Qni (Label: 2); Thanh Hoa – TH (Label: 3); Quang Nam – Qna (Label: 4) for the purpose of constructing classification models.

The *Cinnamomum cassia* samples are complex mixtures, with their IR spectra displaying a total overlap of absorption bands from various components (Fig. 2a). Most characteristic fingerprint peaks for *Cinnamomum cassia* are concentrated within the 1800-600 cm^{-1} range (Fig. 2a). Analysis of these peaks confirmed consistency with previous studies [23]. For instance, the peaks at 1679 cm^{-1} and 1626 cm^{-1} correspond to the stretching vibrations of an aldehyde carbonyl (C=O), indicating high levels of cinnamaldehyde and aldehydes in *Cinnamomum cassia's* volatile oil. The peaks at 1124 cm^{-1} and 1070 cm^{-1} are attributed to C=O stretching and C-OH deformation vibrations, while the peak at 685 cm^{-1} corresponds to alkene vibrational absorption. Despite the complexity and diversity of *Cinnamomum cassia* 's chemical composition, the spectra

of all samples within the 1800-600 cm^{-1} range exhibit significant similarities.

Significant spectral noise was observed in the regions below 1749 cm^{-1} and above 3581 cm^{-1}, and the data was further influenced by variations in measurement conditions. Therefore, before further analysis, Min-Max scaling and first- and second-order Savitzky-Golay derivative filtering were implemented to normalize the spectral data for each sample. The IR spectra, preprocessed using Min-Max Scaling – the most effective algorithm, are shown alongside the first derivative (Fig. 2b) and the second derivative (Fig. 2c).

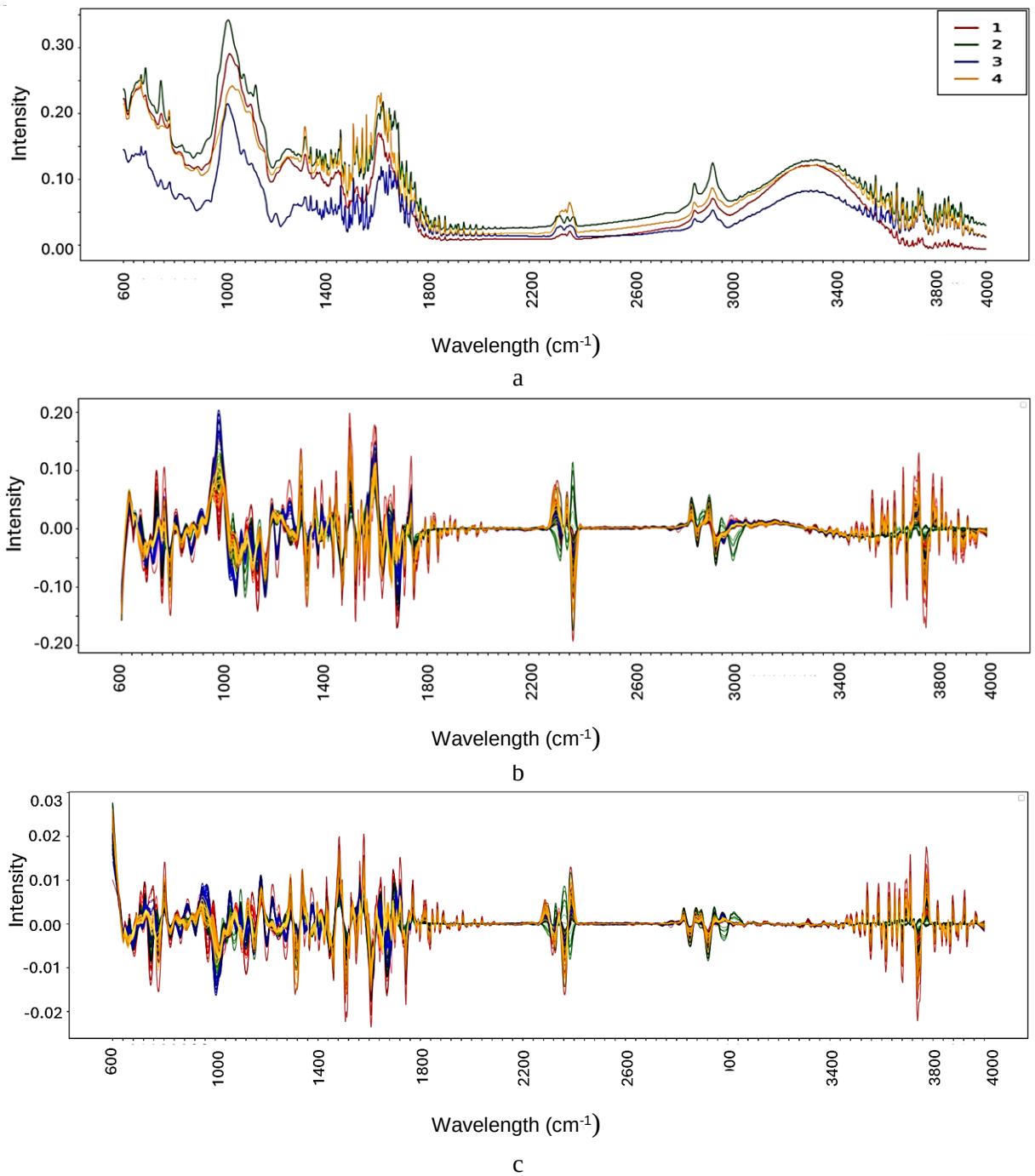


Fig. 2. IR spectra data of *cinnamomum cassia* with different processing techniques. (a) FTIR of 139 samples before reprocessing. (b) FTIR spectra after preprocessed by first derivative and Savitsky- Goley. (c) Further processed IR spectra with 2nd derivative and Savitsky- Goley to enhance spectra features coesdondig with 4 samples collected in 4 geographical regions

Рис. 2. Данные ИК-спектров *cinnamomum cassia* с различными методами обработки. (a) ИК-спектры 139 образцов до повторной обработки. (b) Спектры ИК-спектров после предварительной обработки первой производной и Савицким-Голеем. (c) Дальнейшая обработка ИК спектров со второй производной и Савицким-Голеем для улучшения характеристик спектров совместно с 4 образцами, собранными в 4 географических регионах

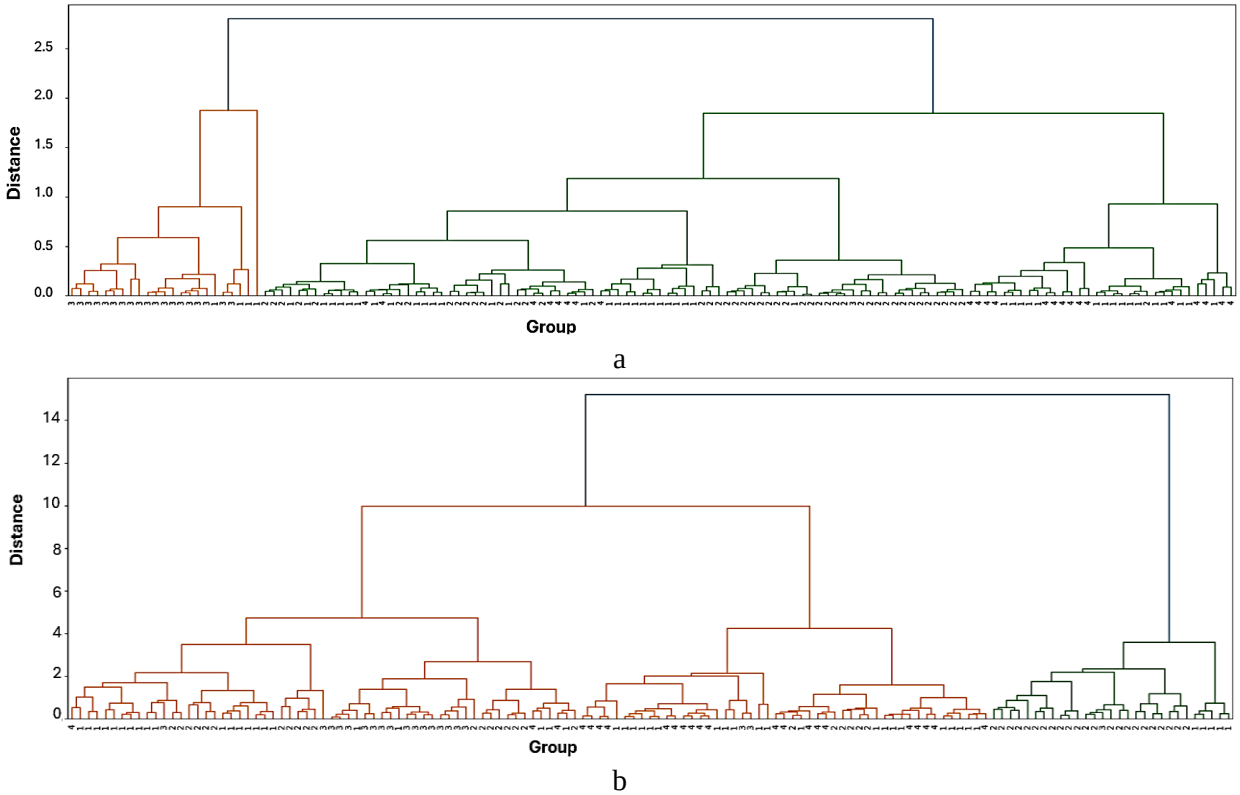


Fig. 3. HCA on the raw data (a) and first derivative spectra (b) of *Cinnamomum cassia samples*
Рис. 3. HCA на необработанных данных (a) и спектрах первой производной (b) образцов *Cinnamomum cassia*

The raw data displays the original spectra with distinct features but is impacted by baseline drift and linear trends, which obscure finer details. The first derivative enhances local variations, particularly in the 600-1000 and 2800-3000 cm^{-1} regions, though it also amplifies random noise. The second derivative further improves sensitivity to subtle features and effectively removes parabolic baseline effects, albeit with considerable noise amplification. Despite this, the second derivative remains the most effective for classification tasks.

Exploratory Data Analysis Using Unsupervised Learning for Discrimination of Cinnamomum cassia

Due to the high overlap in chemical composition fingerprints, distinguishing the geographical origin among the 139 samples is challenging. Therefore, applying chemometric techniques, such as hierarchical clustering analysis (HCA) and principal component analysis (PCA), enhances the ability to extract meaningful spectral patterns, improving the classification accuracy of *Cinnamomum cassia* samples based on their IR spectra.

Cluster analysis of FTIR spectra of Cinnamomum cassia

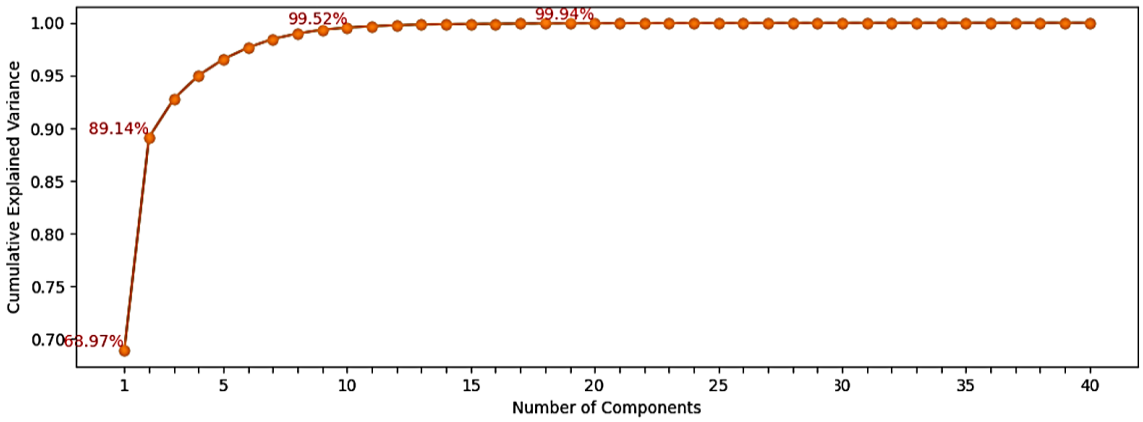
To better visualize the relationships and genetic similarities among *Cinnamomum cassia* samples from different regions, hierarchical cluster analysis (HCA) was conducted on the processed IR spectra of 139 samples. The quantity and proportion of each sample type within the resulting clusters are detailed in Fig. 3 and summarized in Table 1. These results provide insights into the clustering structure and the distribution of sample types across different clusters. The hierarchical clustering diagram illustrates the hierarchical relationships among samples, grouping similar samples into clusters. The clusters correspond to the target categories identified in the study, with each region assigned a specific label: Yen Bai (YB) – 1, Quang Ninh (QNi) – 2, Thanh Hoa (TH) – 3, and Quang Nam (QNa) – 4. This classification aligns with the regional variations in climate and environmental conditions, supporting the differentiation of *Cinnamomum cassia* samples based on their geographic origin. The HCA analysis revealed four main clusters. However, some clusters exhibit overlaps, particularly in cases where sample types share similar spectral characteristics. The distance between the two main groups in the raw data is only about 2.0, significantly lower than the 14 observed with the first derivative. This substantial overlap between Quang Ninh and Thanh Hoa presents challenges for linear classification models such as Linear Discriminant Analysis (LDA) or distance-based methods in distinguishing between regions. The issue arises

from the raw data containing excessive redundant information and features with low discriminative power. Samples from Region 3 formed a distinct cluster, indicating clear differentiation from other regions. Meanwhile, samples from Regions 1 and 2 clustered together, reflecting their compositional similarities. The final cluster primarily consisted of samples from Region 4, with a few from Region 1, suggesting a certain degree of overlap.

This clustering pattern clearly indicates that *Cinnamomum cassia* samples have gradually adapted to their local climates and environmental conditions over time. As previously reported, the chemical composition of *Cinnamomum cassia* varies significantly with geographical proximity, leading to similar chemical profiles among different species collected within the same region.

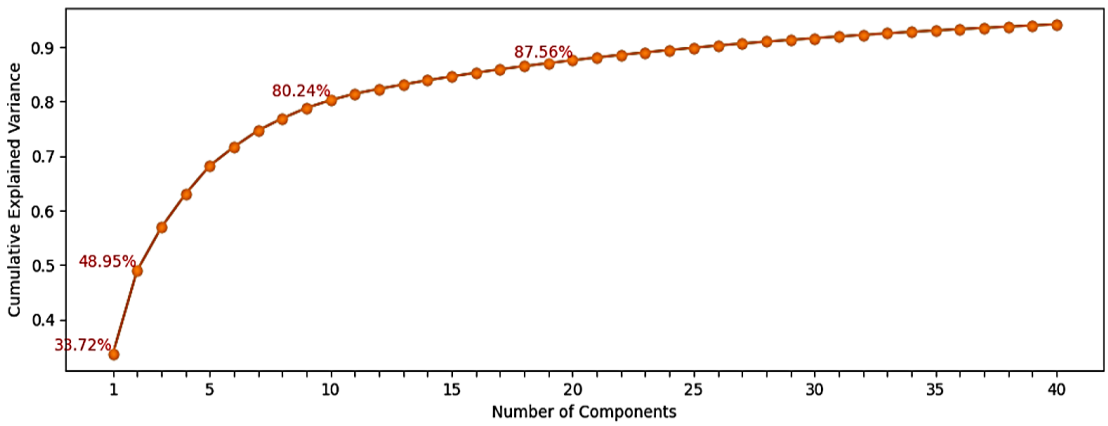
Principal Components Analysis of FTIR spectra of Cinnamomum cassia

PCA was conducted on the raw spectra of cinamon samples indicate the cumulative explained variance of the first 20 PCs which is presented in Fig. 4(a). The five most significant PCs accounted for 96.52% of the total variation in the *Cinnamomum cassia* sample set (PC-1 = 68.97%, PC-2 = 89.14%, PC-3 = 92.76%, PC-4 = 94.99%, and PC-5 = 96.52%) with their cumulative contribution exceeding 99.94%. In contrast, the first derivative, despite having a lower cumulative variance (70% with 3 components and 90% with 10 components), significantly enhances group separation by highlighting local variations in the spectra. As a result, the boundaries between regions become more distinct, reducing overlap between groups and improving classification performance (Fig. 5b). Meanwhile, the second derivative does not provide substantial improvement over the first derivative, with a similar cumulative variance (70% with 3 components and 90% with 10 components), but it may amplify noise due to its high sensitivity to errors in the original data. This results in less effective group separation compared to the first derivative.



a

Raw Data



b

First derivative spectra

Fig. 4. Cumulative explained variance rates of first 40 PCs on the raw data (a) and (b)- first derivative spectra

Рис. 4. Кумулятивные объясненные скорости дисперсии первых 40 ПК на необработанных данных (a) и (b) - спектры первой производной

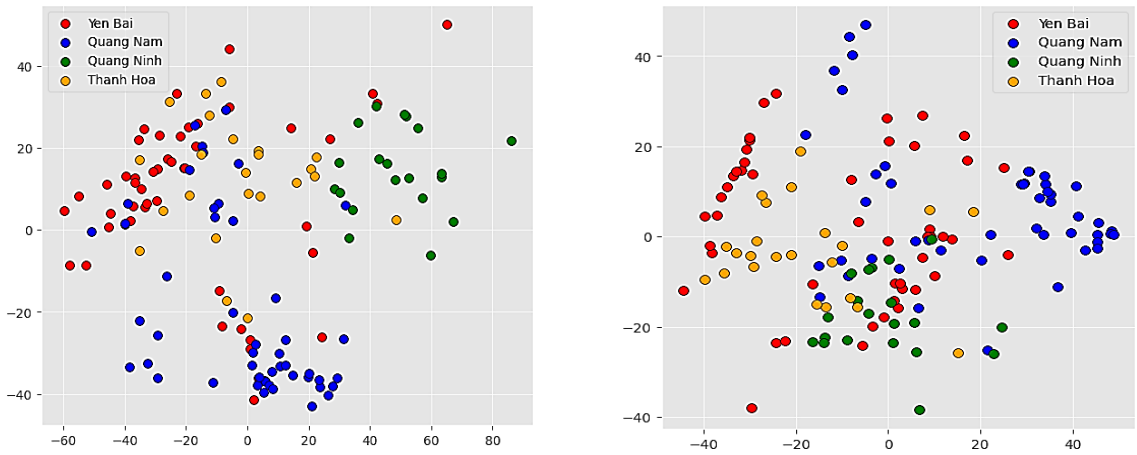


Fig. 5. The score plot of PC-1 and PC-2 analyzing *Cinnamomum cassia* from different provinces based on raw data (a) and first derivative spectra (b)

Рис. 5. График оценок ПК-1 и ПК-2, анализирующих *Cinnamomum cassia* из разных провинций на основе необработанных данных (a) и спектров первой производной (b)

The score plots of FTIR raw data, first derivative, and second derivative indicate that the first derivative provides the most effective group separation. with PC-1 and PC-2 providing discrimination in the 2D scores plot (Fig. 5a). *Cinnamomum cassia* samples from the Quang Ninh region were distinctly clustered on the right axis, while many samples from the Quang Nam region appeared on the left of the 2D scores plot. However, the distribution of northeastern samples slightly overlapped with that of *Cinnamomum cassia* from the northern region. It achieves greater distances between major groups and sharper boundaries, particularly between Yen Bai and Quang Nam. However, approximately 10-15% of samples from Quang Ninh and Thanh Hoa remain intermixed due to their similar spectral characteristics.

Supervised learning Algorithm for original Classification

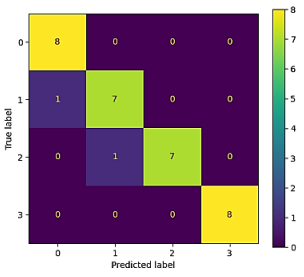
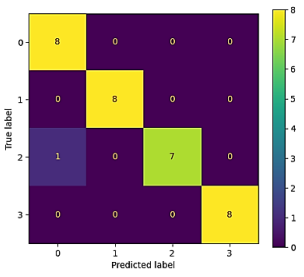
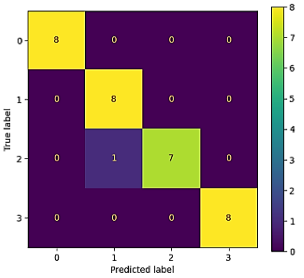
The supervised learning algorithms were developed and trained on 80% of the total samples, while the model's performance was evaluated on the remaining 20%, as described in Fig. 1 (actual number of samples in the study).

The confusion matrix of PCA-SVM reveals its superior classification performance across all three datasets with high density along the diagonal elements (where predicted labels match true labels), particularly for labels 0 and 8. On the no derivative and first derivative datasets, PCA-SVM accurately predicts the majority of samples for these labels while effectively minimizing confusion among intermediate labels such as 1, 2, and 7. Even on the second derivative dataset, despite some minor misclassifications between labels 1 and 7, PCA-SVM maintains strong performance with label 8, demonstrating consistent classification stability.

Table 1

Confusion Matrix of the selected supervised learning Methods

***Таблица 1.* Матрица путаницы выбранных контролируемых методов обучения**

Algorithm	Raw Data	First Derivative spectra	Second Derivative spectra
PCA-SVM			

Continuation of the table

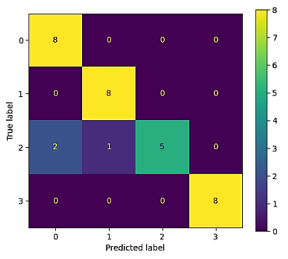
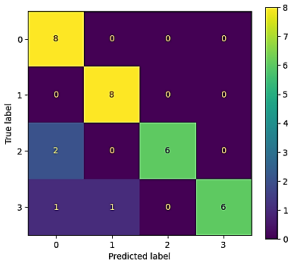
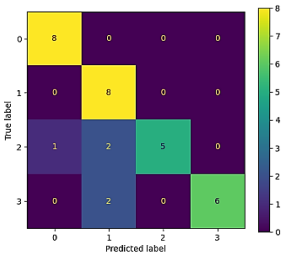
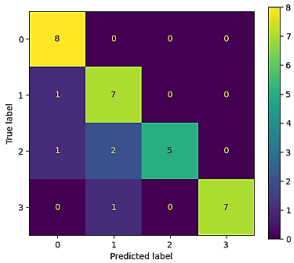
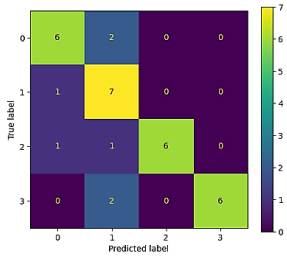
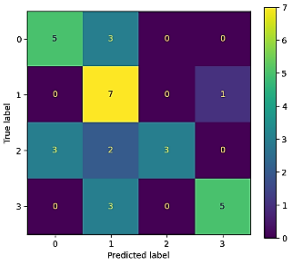
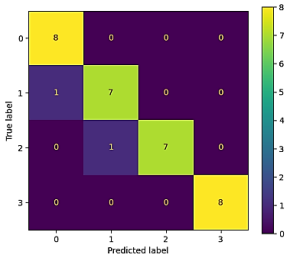
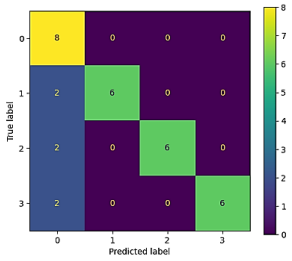
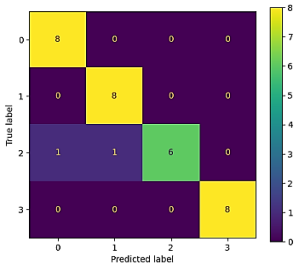
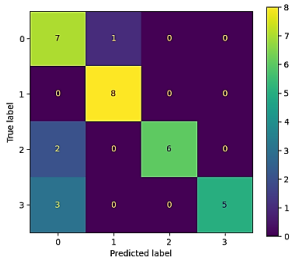
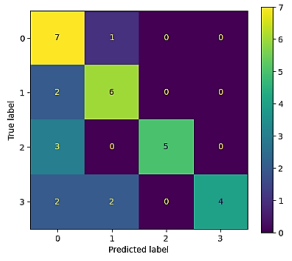
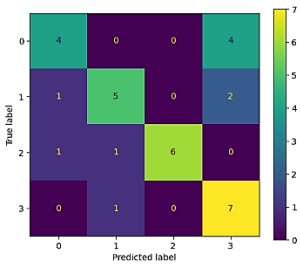
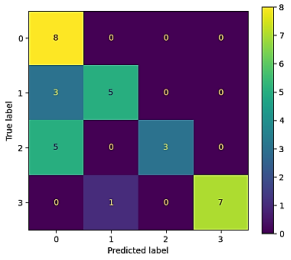
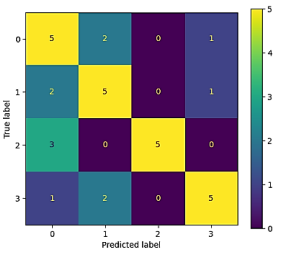
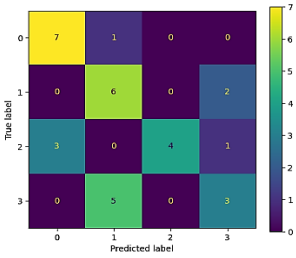
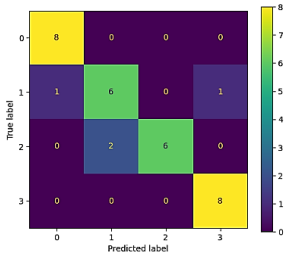
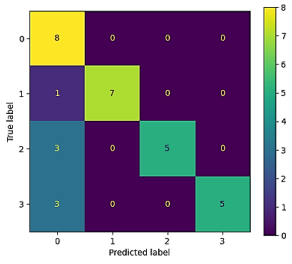
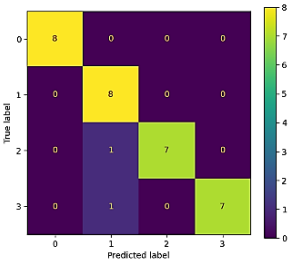
PCA-GaussianNB			
PCA-BernoulliNB			
PCA-RF			
PCA-DT			
PCA-AdaBoost			
PCA-GradientBoost			

Table 2

Summary of Classification Results of the studied supervised learning models

Таблица 2. **Сводка результатов классификации изученных моделей контролируемого обучения**

Model	Pre-processing	#PCs	Accuracy (%)
SVM	None	35	93.75
	1st der.	28	96.88
	2nd der.	33	96.88
GausianNB	None	15	90.63
	1st der.	41	87.50
	2nd der.	20	84.38
BernoulliNB	None	36	84.38
	1st der.	10	78.13
	2nd der.	17	62.50
RF	None	24	93.75
	1st der.	19	81.25
	2nd der.	10	93.75
DT	None	8	81.25
	1st der.	8	68.75
	2nd der.	13	68.75
AdaBoost	None	15	81.25
	1st der.	20	68.75
	2nd der.	8	68.75
GradientBoosting	None	19	87.50
	1st der.	17	78.13
	2nd der.	16	93.75

In contrast, PCA-GaussianNB and PCA-Bernoulli NB exhibit a more uniform distribution of predictions but encounter significant misclassifications, especially with intermediate labels like 2, 5, and 6, particularly on the second derivative dataset, where these labels are more frequently mispredicted.

PCA-RF and PCA-DT also demonstrate reasonable predictive power for label 8 across all datasets; however, they exhibit a notable tendency to confuse labels 0 and 1, which undermines their overall classification efficacy.

Meanwhile, boosting models such as PCA-AdaBoost and PCA-GradientBoost show improvement on the "Second derivative" dataset, with better predictions for labels 7 and 8, but they still suffer from substantial misclassifications across other labels, notably on the no derivative and first derivative datasets, where labels 5 and 6 are frequently misclassified.

The results, presented as accuracy (%) of the predicted models applied to the tested samples using raw data, first derivative, and second derivative FTIR after PCA reduction, are listed in Table 2.

The results in Table 2 suggest that the Support Vector Machine (SVM) and Random Forest (RF) models exhibit strong performance, with accuracy ranging from 93.75% to 96.88%, due to their ability to handle nonlinear data. Specifically, SVM improves from 93.75% (without preprocessing, using 35 principal components) to 96.88% when applying first or second derivatives (with 28-33 principal components). This improvement indicates that derivative preprocessing enhances spectral peaks and valleys, aiding the model in defining nonlinear classification boundaries more effectively.

Similarly, RF maintains a stable accuracy of 93.75%, demonstrating its robustness against preprocessing variations. This stability highlights the strength of its ensemble decision tree mechanism in managing complex feature interactions, ensuring reliable classification across different preprocessing methods.

In contrast, Naïve Bayes models, including GaussianNB and BernoulliNB, exhibit limited performance, with accuracies ranging from 62.50% to 90.63%. GaussianNB achieves its highest accuracy of 90.63% without preprocessing but drops to 84.38% with the second derivative. This decline is due to the model's assumption of a normal distribution and feature independence, which do not align with spectral data, where strong correlations between wavelengths are present.

BernoulliNB performs even worse, with accuracy decreasing from 84.38% to just 62.50% when applying the second derivative. Since BernoulliNB is designed for binary data, it is inherently unsuitable for continuous spectral data. Moreover, derivative preprocessing alters the data structure, further reducing its class discrimination capability.

Among boosting models, Gradient Boosting achieves 93.75% accuracy with the second derivative, outperforming AdaBoost, which reaches a maximum of 81.25%. This advantage is attributed to Gradient Boosting's ability to directly optimize the loss function (typically log-loss), whereas AdaBoost primarily focuses on reweighting misclassified samples, making it less effective for spectral classification.

The Decision Tree (DT) model, however, shows lower performance, ranging from 68.75% to 81.25%. This is due to its tendency to overfit, particularly when spectral data contains environmental or instrumental noise. Derivative preprocessing amplifies this noise, further reducing classification accuracy by leading to less precise decision boundaries.

Regarding the impact of preprocessing, applying first or second derivatives benefits certain models like SVM and GradientBoosting by enhancing key spectral features, but it proves detrimental to noise-sensitive models like DT and BernoulliNB, significantly reducing their performance. Meanwhile, LDA

and RF show less dependence on preprocessing, highlighting their flexibility. Consequently, LDA emerges as the optimal choice for this problem due to its high performance, low computational resource demand (only three principal components), and stability. If nonlinear data processing is required, SVM is a strong contender, particularly when paired with derivative preprocessing.

CONCLUSION

This study demonstrates that FTIR spectroscopy, combined with machine learning, enables effective classification of *Cinnamomum cassia samples* by geographic origin. The ATR-FTIR spectra data coupled with SVM and LDA models after using reduciton techcniques (PCA) provided the highest classification accuracy (96.88%), confirming the potential of FTIR-based authentication for *Cinnamomum cassia* quality assessment. The classification methods be the robust protocol which is the fast and reliable approach for identifying geographical origin identification of Vietnamese *Cinnamomum cassia*.

ACKNOWLEDGEMENT

This study is supported by Hanoi University of Pharmacy, Vietnam in the project number ĐTTĐCT.24.02.

Данное исследование поддержано Ханойским фармацевтическим университетом, Вьетнам, в рамках проекта под номером ĐTTĐCT.24.02.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Bui Thi Lan Phuong: Experimental performance, Initial data processing, Funding acquisition, Writing – review & editing. Hoang Thi Bich: Experimental performance, and Methodology. Nguyen Van Phuong and Bui An Duy: Formal analysis, Data curation. Nguyen Duc Phong: Software. Nguyen Manh Son: Investigation, Formal analysis, Data curation, Funding acquisition, and Conceptualization, Pham Gia Bach: Writing – review. Nguyen Thi Cam Ha: Investigation, Funding acquisition, and Russian writing of Abstract. Ta Thi Thao: Conceptualization. Nguyen Thi Kieu Anh[1]*: Corresponding author.

CONFLICT OF INTEREST

The authors declare the absence of a conflict of interest warranting disclosure in this article.

Авторы заявляют об отсутствии конфликта интересов, требующего раскрытия в данной статье.

REFERENCES
ЛИТЕРАТУРА

1. **Hajimonfarednejad M., Ostovar M., Raee M.J., Hashempur M.H., Mayer J.G., Heydari M.** Cinnamon: A systematic review of adverse events. *J. Clin. Nutr.* 2019. V. 38. N 2. P. 594-602. DOI: 10.1016/j.clnu.2018.03.013.
2. **Derks A., Turner S., Thúy Hạnh N.** Bastard Spice or Champagne of Cinnamon? Conflicting Value Creations along Cinnamon Commodity Chains in Northern Vietnam. *Dev. Change.* 2020. V. 51. N. 3. P. 895-920. DOI: 10.1111/dech.12582.
3. **Luan D. X.** Motivation and barriers to access to formal credit of primary cinnamon producers from the perspective of value chain development in Northwestern Vietnam. *J. Agribusiness Dev. Emerg. Econom.* 2020. V. 10. N 2. P. 117-138. DOI: 10.1108/JADEE-01-2019-0003.
4. **Samaraweera D.N., Weerasuriya S.N., Karunaratne A.S., Subasinghe S., Senaratne R.** Ecology, Agronomy and Management of Cinnamon (Cinnamomum zeylanicum Blume). In: Cinnamon: Botany, Agronomy, Chemistry and Industrial Applications. Ed. by R. Senaratne, R. Pathirana. Cham: Springer Internat. Publ. 2020. P. 171-200. DOI: 10.1007/978-3-030-54426-3_7.
5. **Hashitha Nayananjalee Aluthgamage, Kumari Fonseka D.L.C., Niluka Nakandalage.** Enhancement of high-quality cinnamon quill production through agronomic approaches: a review. *Acad. Biol.* 2023. V. 1. N 1. DOI: 10.20935/AcadBiol6025.
6. **Nguyen A.T., Ta V.H., Nguyen V.H., Pham A.T., Monnerat M., Hens L.** Shifting challenges for Cinnamomum cassia production in the mountains of Northern Vietnam: spatial analysis combined with semi-structured interviews. *Environ. Dev. Sustain.* 2022. V. 24. N 5. P. 7213-7235. DOI: 10.1007/s10668-021-01745-x.
7. **Saha S., Bhattacharya R., Chaudhary M., Hazarika T.K., Mitra A.** Impact of geographical locations on essential oil composition and leaf histochemistry in Cinnamomum verum J. S. Presl. *J. Essent. Oil-Bear. Plants.* 2023. V. 26. N 6. P. 1546 - 1562. DOI: 10.1080/0972060X.2023.2256779.
8. **Woehrlin F., Fry H., Abraham K., Preiss-Weigert A.** Quantification of flavoring constituents in cinnamon: high variation of coumarin in cassia bark from the German retail market and in authentic samples from indonesia. *J Agric Food Chem.* 2010. V. 58. N 19. P. 10568-75. DOI: 10.1021/jf102112p.
9. **Pages-Rebull J., Pérez-Ràfols C., Serrano N., Díaz-Cruz J.M.** Analytical methods for cinnamon authentication. *Trends Food Sci. Technol.* 2024. V. 146. 104388. DOI: 10.1016/j.tifs.2024.104388.
10. **Castro R.C., Ribeiro D.S.M., Santos J.L.M., Páscoa R.N.M.J.** Authentication/discrimination, identification and quantification of cinnamon adulterants using NIR spectroscopy and different chemometric tools: A tutorial to deal with counterfeit samples. *Food Control.* 2023. V. 147. 109619. DOI: 10.1016/j.foodcont.2023.109619.
11. **Bansal S., Thakur S., Mangal M., Mangal A.K., Gupta R.K.** DNA barcoding for specific and sensitive detection of Cuminum cyminum adulteration in Bunium persicum. *Phytomedicine.* 2018. V. 50. P. 178-83. DOI: 10.1016/j.phymed.2018.04.023.
12. **Feltes G., Ballen S.C., Steffens J., Paroul N., Steffens C.** Differentiating True and False Cinnamon: Exploring Multiple Approaches for Discrimination. *Micromachines.* 2023. V. 14. N 10. 1819. DOI: 10.3390/mi14101819.

13. **Фам Куанг Трунг, Хоанг Бич Нгок, Нгуен Ван Тхык, Чан Тхи Хюэ, Фам Гиа Бах, Та Тхи Тхао.** Химическая характе-ристика и классификация географического происхождения вьетнамских зеленых чаев на основе данных ^{1}H-ЯМР в сочетании с машинным обучением. *Изв. вузов. Химия и хим. технология.* 2023. Т. 66. Вып. 12. С. 56-64. **Pham Quang Trung, Hoang Bich Ngoc, Nguyen Van Thuc, Tran Thi Hue, Pham Gia Bach, Ta Thi Thao.** Chemical characterization and classification of geographical origin of Vietnamese green teas based on ^{1}H-NMR data combined with machine learning. *ChemChemTech [Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.]*. 2023. V. 66. N 12. P. 56-64. DOI: 10.6060/ivkkt.20236612.6874.
14. **Ding Y., Wu E.Q., Liang C., Chen J., Tran M.N., Hong C.H., Jang Y., Park K.L., Bae K., Kim Y.H., Kang J.S.** Discrimination of cinnamon bark and cinnamon twig samples sourced from various countries using HPLC-based fingerprint analysis. *Food Chem.* 2011. V. 127. N 2. P. 755-760. DOI: 10.1016/j.foodchem.2011.01.011.
15. **Liu C., Long H., Wu X., Hou J., Gao L., Yao S., Lei M., Zhang Z., Guo D.-a., Wu W.** Quantitative and fingerprint analysis of proanthocyanidins and phenylpropanoids in Cinnamomum verum bark, Cinnamomum cassia bark, and Cassia twig by UPLC combined with chemometrics. *Eur. Food Res. Technol.* 2021. V. 247. N 11. P. 2687-2698. DOI: 10.1007/s00217-021-03795-x.
16. **Pages-Rebull J., Sagristà G., Pérez-Ràfols C., Serrano N., Díaz-Cruz J.M.** Application of HPLC-UV combined with chemometrics for the detection and quantification of 'true cinnamon' adulteration. *Talanta*. 2024. V. 271. 125676. DOI: 10.1016/j.talanta.2024.125676.
17. **Farag M.A., Kabbash E.M., Mediani A., Döll S., Esatbeyoglu T., Afifi S.M.** Comparative Metabolite Fingerprinting of Four Different Cinnamon Species Analyzed via UPLC–MS and GC–MS and Chemometric Tools. *Molecules*. 2022. V. 27. N 9. 2935. DOI: 10.3390/molecules27092935.
18. **Suzuki R., Kasuya Y., Sano A., Tomita J., Maruyama T., Kitamura M.** Comparison of various commercially available cinnamon barks using NMR metabolomics and the quantification of coumarin by quantitative NMR methods. *J. Nat. Med.* 2022. V. 76. N 1. P. 87-93. DOI: 10.1007/s11418-021-01554-6.
19. **Pan Y., Qiao L., Liu S., He Y., Huang D., Wu W., Liu Y., Chen L., Huang D.** Explorative Study on Volatile Organic Compounds of Cinnamon Based on GC-IMS. *Metabolites*. 2024. V. 14. N 5. P. 274. DOI: 10.3390/metabo14050274.
20. **Ballin N.Z., Laursen K.H.** To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends Food Sci. Technol.* 2019. V. 86. P. 537-543. DOI: 10.1016/j.tifs.2018.09.025.
21. **Pei Y.-F., Zhang Q.-Z., Zuo Z.-T., Wang Y.-Z.** Comparison and Identification for Rhizomes and Leaves of Paris yunnanensis Based on Fourier Transform Mid-Infrared Spectroscopy Combined with Chemometrics. *Molecules*. 2018. V. 23. N 12. 3343. DOI: 10.3390/molecules23123343.
22. **Ying Zhu, Augustine Tuck Lee Tan.** Chemometric Feature Selection and Classification of Ganoderma lucidum Spores and Fruiting Body Using ATR-FTIR Spectroscopy. *Am. J. Anal. Chem.* 2015. V. 6. N P. 830-840. DOI: 10.4236/ajac.2015.610079.
23. **Lixourgioti P., Goggin K.A., Zhao X., Murphy D.J., van Ruth S., Koidis A.** Authentication of cinnamon spice samples using FT-IR spectroscopy and chemometric classification. *LWT*. 2022. V. 154. 112760. DOI: 10.1016/j.lwt.2021.112760.
24. **Силаев Д.В., Шестопалова Н.Б., Фомина Ю.А., Русанова Т.Ю.** Применение хемометрических алгоритмов для спектрофотометрического определения синтетических пищевых красителей Е110 и Е124. *Изв. вузов. Химия и хим. технология.* 2022. Т. 65. Вып. 2. С. 50-59. **Silaev D.V., Shestopalova N.B., Fomina Yu.A., Rusanova T.Yu.** Application of chemometric algorithms for spectrophoto-metric determination of synthetic food colors. *ChemChemTech [Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.]*. 2022. V. 65. N 2. P. 50-59. DOI: 10.6060/ivkkt.20226502.6497.
25. **Chu Thi Thu Ha.** State of the environment and natural resources in Vietnam. *J. Viet. Env.* 2014. V. 6. N 1. P. 1-3. DOI: 10.13141/jve.vol6.no1.pp1-3.
26. **Luo J., Ying K., Bai J.** Savitzky–Golay smoothing and differentiation filter for even number data. *Signal Process*. 2005. V. 85. N 7. P. 1429-1434. DOI: 10.1016/j.sigpro.2005.02.002.
27. **Smoliński A., Walczak B., Einax J.W.** Hierarchical clustering extended with visual complements of environmental data set. *Chemometr. Intell. Lab. Syst.* 2002. V. 64. N 1. P. 45-54. DOI: 10.1016/S0169-7439(02)00049-7.
28. **Wold S., Esbensen K., Geladi P.** Principal component analysis. *Chemometr. Intell. Lab. Syst.* 1987. V. 2. N 1. P. 37-52. DOI: 10.1016/0169-7439(87)80084-9.
29. **Luts J., Ojeda F., Van de Plas R., De Moor B., Van Huffel S., Suykens J.A.K.** A tutorial on support vector machine-based methods for classification problems in chemometrics. *Anal. Chim. Acta*. 2010. V. 665. N 2. P. 129-45. DOI: 10.1016/j.aca.2010.03.030.
30. **Shobha G., Rangaswamy S.** Chapter 8 - Machine Learning. *Handbook of Statistics*. 2018. V. 38. P. 197-228. DOI: 10.1016/bs.host.2018.07.004.
31. **Scott I.M., Lin W., Liakata M., Wood J.E., Vermeer C.P., Allaway D., Ward J.L., Draper J., Beale M.H., Corol D.I., Baker J.M., King R.D.** Merits of random forests emerge in evaluation of chemometric classifiers by external validation. *Anal. Chim. Acta*. 2013. V. 801. P. 22-33. DOI: 10.1016/j.aca.2013.09.027.
32. **Liu S., McGree J., Ge Z., Xie Y.** 2 - Classification methods. In: Computational and Statistical Methods for Analysing Big Data with Applications. Ed. by S. Liu, J. McGree, Z. Ge, Y. Xie. San Diego: Academic Press. 2016. P. 7-28. DOI: 10.1016/B978-0-12-803732-4.00002-7.
33. **de Giorgio A., Cola G., Wang L.** Systematic review of class imbalance problems in manufacturing. *J. Manuf. Syst.* 2023. V. 71. P. 620-44. DOI: 10.1016/j.jmsy.2023.10.014.
34. **Belyadi H., Haghighat A.** Chapter 5 - Supervised learning. In: Machine Learning Guide for Oil and Gas Using Python. Ed. by H. Belyadi, A. Haghighat. Gulf Professional Publ. 2021. P. 169-295. DOI: 10.1016/B978-0-12-821929-4.00004-4.

Поступила в редакцию (Received) 24.03.2025
Принята к опубликованию (Accepted) 07.04.2025