Vietnam Journal of Food Control

A Scientific Journal of National Institute for Food Control
Journal homepage: http://vjfc.vn/

Research Article

Machine learning based on UV-Vis full spectra for the simultaneous determination of curcumin, demethoxycurcumin, bisdemethoxycurcumin in *Curcuma longa* L.

Nguyen Thi Van Anh[1], Nguyen Ha Anh[1], Nguyen Duc Phong[2], Nguyen Duc Thanh[3*]
[1]*Vietnam University of Traditional Medicine, Hanoi, Vietnam*
[2]*TRAPHACO Joint Stock Company, Hung Yen, Vietnam*
[3]*Vietnam Military Medical Academy, Hanoi, Vietnam*

(Received: 12 Nov 2025; Revised: 28 Dec 2025; Accepted: 29 Dec 2025)

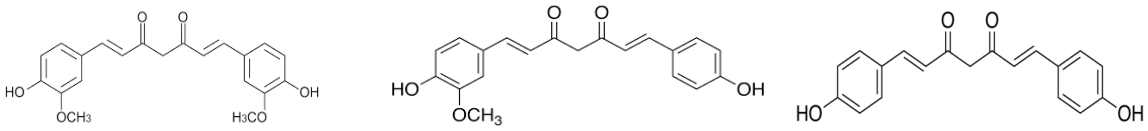
Abstract

This study has developed a rapid and simple method based on the UV-Vis full spectra coupled with machine learning for the quantitative prediction of curcumin, demethoxycurcumin and bisdemethoxycurcumin in turmeric rhizomes. The UV-Vis spectral data and HPLC quantification results of the individual components from 55 turmeric rhizome samples were utilized for model development and training. Analogous data matrices from an independent set of 24 samples were used for external validation of the developed models. Four machine learning models were investigated, comprising two linear multivariate regression algorithms: principal component regression (PCR) and partial least squares regression (PLSR), two non-linear multivariate regression algorithms: artificial neural network (ANN) and random forest (RF). The results demonstrated that the linear multivariate regression models exhibited superior analytical performance. Specifically, PCR yielded a coefficient of determination (R^2) values from 0.957 to 0.982 with root mean square error (RMSE) values from 3.086 to 1.295, while PLSR achieved R^2 values from 0.956 to 0.979 and RMSE values from 3.116 to 1.139. However, when comparing the HPLC-quantified contents with the values predicted by the two models, some samples still exhibited relative errors exceeding 20%. This study confirms the feasibility of rapidly and simultaneously predicting the contents of curcumin, demethoxycurcumin and bisdemethoxycurcumin in turmeric rhizomes using UV-Vis spectral data coupled with either PLSR or PCR models, offering an efficient alternative to conventional methods.

Keywords: *Machine Learning, UV-Vis, curcumin, turmeric.*

1. INTRODUCTION

In Vietnam, turmeric (*Curcuma longa* L.) is widely cultivated across many provinces. Turmeric rhizome is a traditional herbal medicine whose primary active constituents are curcuminoids, including curcumin (CUR), demethoxycurcumin (DMC) and bisdemethoxycurcumin (BDMC) [1]. The molecular structure of these 3 compounds are shown in **Figure 1**.



Curcumin Demethoxycurcumin Bisdemethoxycurcumin

Figure 1. *Structure of 3 curcuminoids*

*Corresponding author: Nguyen Duc Thanh (E-mail: nguyenducthanh@vmmu.edu.vn)
https://doi.org/10.47866/2615-9252/vjfc.4633

In vitro and *in vivo* studies have demonstrated that curcuminoids derived from turmeric exhibit various biological activities, such as antioxidant, anticancer, antibacterial, antiviral, antihypertensive, and hypocholesterolemic effects. However, most studies have focused on CUR, whereas DMC and BDMC have been less investigated. Some reports indicate that DMC and BDMC also exhibit anti-inflammatory, antioxidant, and anticancer activities; BDMC even demonstrated a stronger inhibitory effect on cancer cell invasion than CUR [2]. Currently, chromatographic methods are commonly used for curcuminoid quantification, particularly high-performance liquid chromatography (HPLC). HPLC systems equipped with a UV-Vis or photodiode array (PDA) detector and a C18 column are frequently used to achieve good analytical performance [3]. These methods offer high specificity, sensitivity and accuracy, and are capable of simultaneously separating and quantifying multiple compounds; however, they require long analysis times, high costs and complex instrumentation. To meet the demand for rapid, simple, and cost-effective analysis, UV-Vis spectroscopy is considered a promising alternative. Nevertheless, due to the overlap of absorption spectra, UV-Vis spectral data alone cannot be used to individually quantify each component in a multicomponent mixture. The integration of multivariate regression algorithms enables the extraction of relevant spectral information, thereby allowing the prediction of individual compound contents. Some studies have developed simultaneous quantification models based on UV-Vis spectral data combined with multivariate regression algorithms and achieved good predictive performance for example, caffeine, theobromine and theophylline in tea [4]; ciprofloxacin, lomefloxacin and enrofloxacin in laboratory mixture, dosage forms and water samples [5]; p-hydroxybenzoic acid, vanillic acid and caffeic acid in fruit juice samples [6]. Multivariate regression methods thus enable the rapid and simple simultaneous analysis of multiple compounds without the need for separation techniques.

This study develops a method for the simultaneous quantification of three curcuminoids in turmeric rhizome - CUR, DMC and BDMC - using UV-Vis spectral data combined with multivariate regression algorithms implemented in Python open software. This approach offers a rapid and efficient alternative for the simultaneous determination of three curcuminoids without the need for separation techniques.

2. MATERIALS AND METHODS

2.1. Chemicals and equipment

Reference standards of CUR, DMC, and BDMC were supplied by Biopurify (China) with purities of 96.83%, 98.0% and 98.5%, respectively. Stock standard solutions of CUR (1000 µg/mL), DMC (500 µg/mL) and BDMC (500 µg/mL) were prepared in methanol and stored at 0-5°C for up to six months.

Methanol, acetonitrile and acetic acid (Merck, Germany) with purities of 99.9%, 99.8% and 100%, respectively, were used. Double-distilled water was purified using a Milli-Q water purification system (USA).

UV-Vis spectra were recorded using an U-3900/3900H spectrophotometer (Hitachi, Japan) over a wavelength range of 190-900 nm with a 1 cm quartz cuvette.

The HPLC analysis was performed using a Waters Alliance e2695 system equipped with a 2998 PDA detector (Water, Singapore) and an Xbridge® BEH C18 column (4.6 x 250 mm, 5 µm).

2.2. Sample collection and preparation

Sample collection: 79 turmeric rhizome samples were collected from November 2023 to November 2024 in 17 provinces in Northern Vietnam (Bac Giang, Bac Kan, Bac Ninh, Ha Giang, Ha Nam, Hanoi, Hai Duong, Hoa Binh, Hung Yen, Nam Dinh, Ninh Binh, Phu Tho, Quang Ninh, Thanh Hoa, Thai Nguyen, Tuyen Quang and Vinh Phuc) in accordance with the Vietnamese national standard TCVN 8958:2011 (ISO 5562:1983). The samples were coded and fully documented with information regarding sampling location and time.

Sample processing: Fresh turmeric rhizomes were washed, peeled and sliced. The samples were then dried at 50°C, ground and sieved to obtain coarse powder. The powder was evaluated to ensure that its moisture content was below 5%. The samples were labeled, packaged, and stored at room temperature in a desiccator.

Sample preparation: Approximately 0.5 g of turmeric powder (accurate to 0.0001 g) was weighed into a 50 mL falcon tube. Water was added to the 5 mL mark, and the mixture was subjected to ultrasonic extraction for 2 min. Distilled water was added to 5 mL, followed by ultrasonic treatment for 2 min. Methanol was then added to 50 mL and the mixture was sonicated for 20 min. After centrifugation, the supernatant was transferred into

a 100 mL volumetric flask. The residue was re-extracted twice with methanol up to 50 mL each time under the same conditions. All extracts were combined and diluted to volume with methanol in the volumetric flask. The combined extracts were collected in the volumetric flask and diluted to volume with methanol. The solution was filtered through a 0.45 µm membrane filter to obtain the filtrate.

HPLC analysis: The filtrate was diluted 10 times with methanol, filtered through a 0.2 µm membrane filter and then injected into the HPLC system. Chromatographic analysis was performed under the following conditions: the mobile phase consisted of 1% acetic acid and acetonitrile (55 : 45, v/v) at a flow rate of 1.0 mL/min, with an injection volume of 10 µL and detection at 420 nm.

UV-Vis spectral measurement: The filtrate was diluted 100 times with methanol and scanned using the UV-Vis spectrophotometer over a wavelength range of 200 - 800 nm, with methanol as the blank. Spectral data were recorded in Excel format with a sampling interval of 5 nm to obtain the input dataset.

2.3. Multivariate regression and machine learning

2.3.1. UV-Vis spectral data preprocessing

To standardize the input UV-Vis spectral data, Savitzky-Golay filtering, first derivative transformation, the successive projections algorithm (SPA) and standard scaling were applied. The dataset was randomly divided into 55 samples for the training set to develop the models and 24 samples for the test set to evaluate model performance. Data were standardized to a normal distribution using the StandardScaler method. The statistical parameters (mean and standard deviation) were calculated from the training set and then applied to the test set.

2.3.2. Exploratory data analysis

To make appropriate assumptions about the obtained dataset for the development of models, an overall evaluation of the dataset was required. In this study, principal component analysis (PCA) was used to provide insights into the correlations and distribution patterns of the data.

2.3.3. Regression model development

Linear regression modeling was conducted using two algorithms: principal component regression (PCR) and partial least squares regression (PLSR). PCR reduces the dimensionality of the independent variables using PCA before performing linear regression on the extracted components. PLSR also performs dimensionality reduction prior to regression; however, it applies a supervised transformation by optimizing latent components to achieve the strongest correlation with the dependent variables [7].

Two nonlinear regression algorithms were also investigated: artificial neural network (ANN) and random forest (RF). ANN is capable of modeling complex nonlinear relationships. In contrast, RF is an ensemble method consisting of multiple decision trees, where the final prediction is obtained by averaging the outputs of individual trees to reduce variance and improve predictive accuracy [8].

Model hyperparameters were optimized on the training dataset using cross-validation with the GridSearchCV module from the scikit-learn library. Five-fold cross-validation (cv = 5) was applied. After selecting the optimal hyperparameters, the final models were trained on the full training set and subsequently evaluated on the test set.

In this study, the input variables consisted of UV-Vis spectral data from 79 samples, while the target variables were the concentrations of three curcuminoids quantified by HPLC. Model performance was evaluated using the coefficient of determination (R^2) and the root mean square error (RMSE). R^2 indicates the proportion of variance in the dependent variable explained by the independent variables. RMSE measures the average deviation between predicted and actual values, providing an intuitive assessment of model prediction accuracy. Lower RMSE values and R^2 values closer to 1 indicate better model performance [9].

2.3.4. Software

All data processing and model development were implemented in Python (version 3.11.6), using Visual Studio Code as the programming environment.

3. RESULTS AND DISCUSSION

3.1. Determination of CUR, DMC and BDMC in turmeric samples by HPLC

The method was validated in accordance with AOAC requirements, and the results are presented in **Table 1.**

Table 1. *Validation results of the HPLC quantification method*

Analytes	t_R of standard (min)	t_R of sample (min)	Working range (µg/mL)	Repeatability RSD (%)	Recovery (%)
CUR	11.2	11.3	0.5 - 100.0	3.0	93.1 - 104.5
DMC	10.1	10.1	0.5 - 100.0	1.4	91.2 - 103.8
BDMC	9.1	9.1	0.5 - 100.0	3.0	91.7 - 103.0

The contents of CUR, DMC, BDMC in 79 turmeric samples were determined using the HPLC method described in Section 2.2, with the results shown in **Figure 2**. These values were used as reference data for model construction.

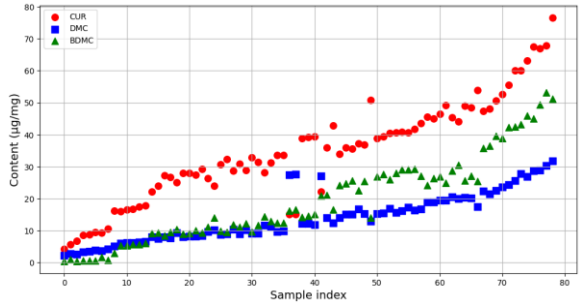
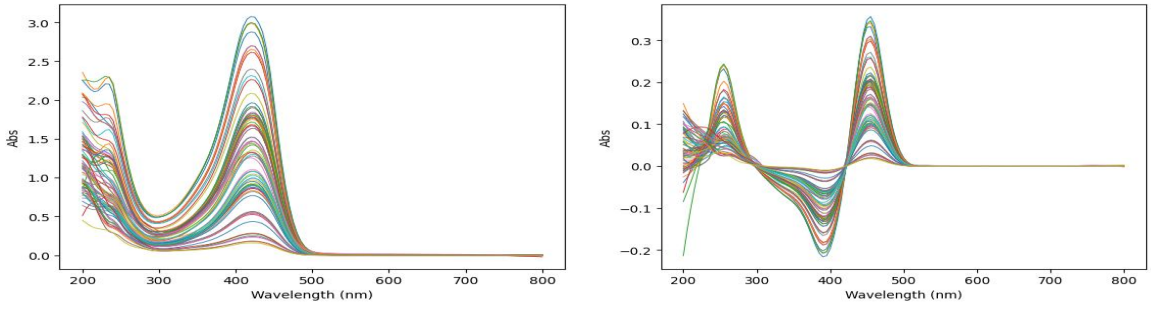


Figure 2*. The contents of three curcuminoids were determined by HPLC*

The total curcuminoid content in the 79 samples analyzed by HPLC ranged from 6.89 to 159.64 µg/mg, indicating a wide variation among the samples. For each individual analyte, substantial variability was also observed: the CUR content ranged from 4.22 to 76.64 µg/mg, DMC from 2.16 to 31.82 µg/mg, and BDMC from 0.48 to 53.24 µg/mg.

3.2. UV-Vis spectral results and data preprocessing

The UV-Vis spectra of turmeric samples recorded over the wavelength range of 200 - 800 nm are presented in **Figure 3a**. The absorption maximum of the samples was observed at 420 nm, consistent with the specification in the Vietnamese Pharmacopoeia V. The spectral data were subsequently exported to Excel, generating a 79 × 121 data matrix corresponding to 79 samples and 121 wavelength variables.



a - UV-Vis spectra before preprocessing

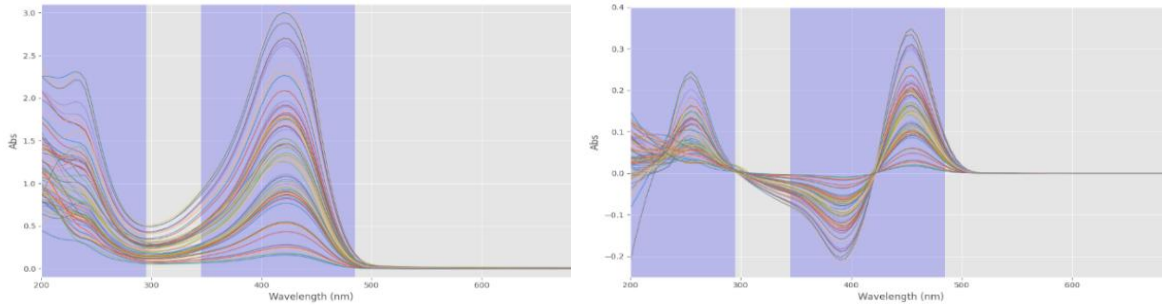
b - UV-Vis spectra processed by first derivative transformation and Savitzky-Golay smoothing

Figure 3. *UV-Vis spectra of turmeric samples*

In the UV-Vis spectra, the most relevant information was concentrated in the 300 - 500 nm region. The region below 300 nm exhibited high noise and considerable variability, while the region above 500 nm showed little analytical signal. To enhance selectivity and reduce spectral noise, first-derivative transformation combined with the Savitzky-Golay algorithm was applied. The results are shown in **Figure 3b**.

After preprocessing, the spectra exhibited reduced noise, and the spectral profile changed from a single main absorption peak to two derivative features (peak-trough). The absorption maximum at 420 nm in the original UV-Vis spectrum was transformed into a zero-crossing point after derivative processing. The baseline of the processed spectra became more stable, and subtle variations in the data were amplified, which were difficult to observe in the original spectra. In addition, the influences of instrumental and environmental conditions were minimized, thereby improving model stability.

As shown in **Figure 4**, 50 wavelengths with the strongest correlations to the target variables were selected for model training. The most informative wavelengths (highlighted in blue) were located in the ranges of 200 - 295 nm and 345 - 485 nm. Wavelengths with weak correlations were excluded to reduce noise and redundant information, simplify model training, and improve predictive performance.



a - Selected wavelengths by SPA on original UV-Vis spectra

b - Selected wavelengths by SPA on pre-processed UV-Vis spectra

***Figure 4**. Spectral feature selection using SPA*

3.3. Exploratory data analysis

The PCA algorithm was applied to the pre-processed spectral dataset, resulting in a 55 × 50 matrix corresponding to 55 training samples and 50 selected wavelengths used for model construction. As shown in **Figure 5**, the first four principal components (PCs) retained more than 95% of the total variance. Therefore, the study initially considered four PCs when determining the optimal number of PCs for model development.

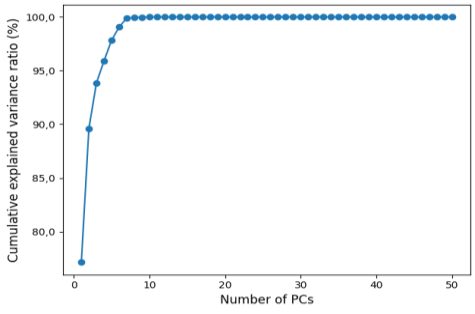


***Figure 5.** Percentage of variance explained by the number of PCs*

3.4. Regression model development

3.4.1. Development of linear regression models

Multivariate regression models were developed for the simultaneous determination of CUR, DMC, and BDMC in the training set samples, for which the reference concentrations were determined by the HPLC method. The signal matrix of the 55 samples consisted of absorbance values recorded in the wavelength ranges of 200 - 295 nm and 345 - 485 nm. The cross-validation results presented in **Table 2** indicate that eight PCs were sufficient for both PCR and PLSR models to achieve optimal performance, demonstrating the effectiveness of dimensionality reduction compared with using all 50 PCs. Therefore, the regression models were constructed in the reduced-dimensional space defined by the optimal number of PCs for each model.

Table 2. *Accuracy of the PCR and PLSR models evaluated by cross-validation*

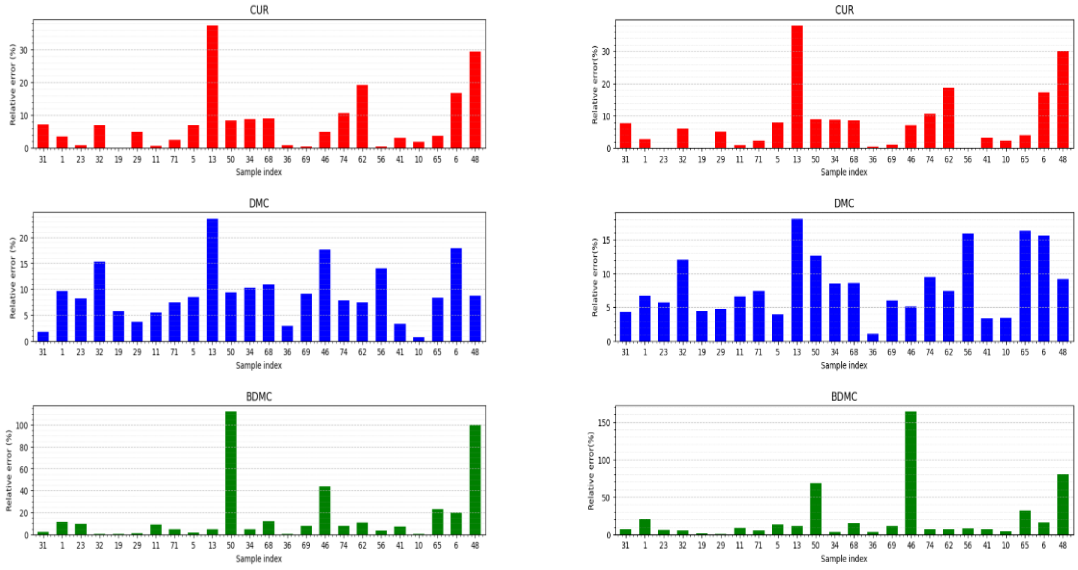
Model	Optimal PCs	R^2	RMSE
PCR	8	0.969	2.078
PLSR	8	0.970	2.077

The model development results presented in **Table 3** indicate that both PCR and PLSR provided good predictive performance for the quantification of three analytes. PLSR showed superior performance in predicting DMC, whereas PCR yielded better predictions for CUR and BDMC.

Table 3. *Results of regression model development using linear models*

Model		PCR		PLSR	
		Training set	Test set	Training set	Test set
CUR	R^2	0.976	**0.957**	0.976	0.956
	RMSE	2.584	**3.086**	2.579	3.116
DMC	R^2	0.969	0.968	0.973	**0.975**
	RMSE	1.372	1.295	1.282	**1.139**
BDMC	R^2	0.966	**0.982**	0.970	0.979
	RMSE	2.468	**1.854**	2.337	1.978

The predicted contents of CUR, DMC, and BDMC for the 24 samples in the test set obtained from the PCR and PLSR models are presented in **Figure 6** and compared with the reference values determined by HPLC.



a - PCR Model *b - PLSR Model*

Figure 6. *The relative error between the predicted contents of CUR, DMC, and BDMC obtained from the linear models and the reference values determined by HPLC*

The error analysis results showed that, for CUR predicted by the PCR model, 13 samples had relative errors below 5%, 6 samples had errors in the range of 5 - 10%, and 2 samples exhibited errors greater than 20%. For BDMC, the PCR model yielded 8 samples with relative errors below 5%, 4 samples within the range of 5 - 10%, and 4 samples with errors exceeding 20%. Meanwhile, for DMC predicted by the PLSR model, 7 samples showed relative errors below 5%, 11 samples were within the 5 - 10% range, and no sample exhibited a relative error greater than 20%.

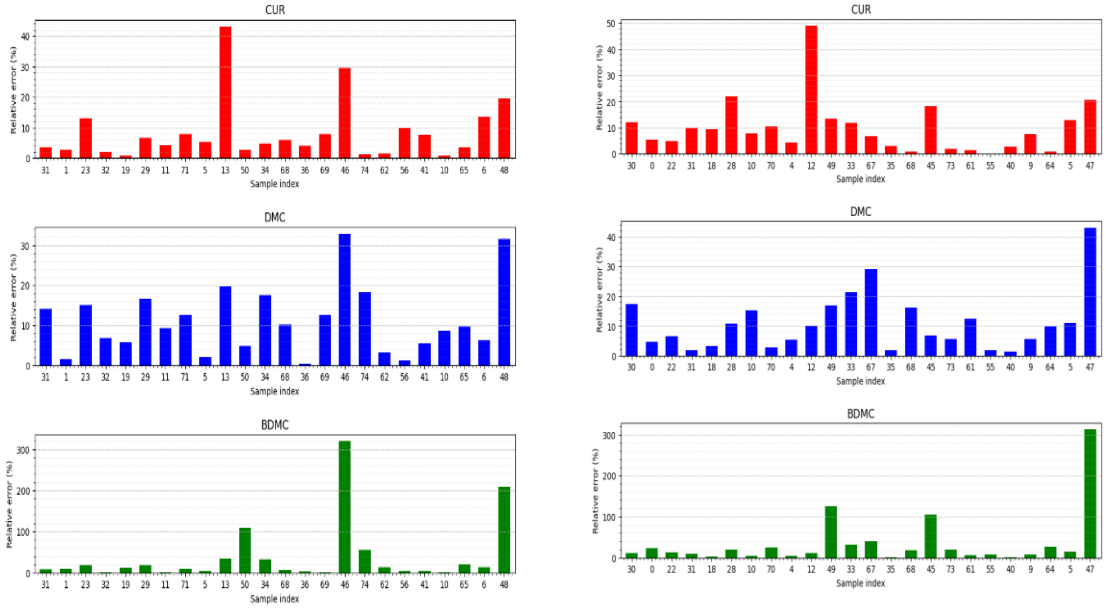
3.4.2. Development of nonlinear regression models

ANN model was investigated using one or two hidden layers, with the number of neurons tested at 64, 64 and 32, 64 and 128. The L2 regularization coefficient (alpha) was evaluated at 0.0001, 0.001, and 0.01, while

the learning rate was tested at 0.01 and 0.1. The hidden layers employed the nonlinear ReLU activation function, and the output layer used a linear activation function. The model was trained using the Adam optimization algorithm. The prediction results showed that the ANN model achieved an R^2 of 0.921 and an RMSE of 3.597.

RF algorithm was combined with PCA to reduce the data dimensionality to seven PCs. The RF model demonstrated superior predictive performance compared with the ANN model, achieving an R^2 of 0.934 and an RMSE of 3.084.

The predicted contents of CUR, DMC, and BDMC for the 24 samples in the test set obtained from the RF and ANN models were compared with the reference values determined by HPLC, as shown in **Figure 7**.



a - RF Model *b - ANN Model*

Figure 7. *The relative error between the predicted contents of CUR, DMC, and BDMC obtained from the nonlinear models and the reference values determined by HPLC*

Cross-validation was performed, and the results presented in **Table 4** indicate that the RF model provided better predictive performance for CUR and BDMC, whereas the ANN model yielded superior predictions for DMC. However, both ANN and RF exhibited lower predictive performance compared with the PCR and PLSR models.

The results reveal differences in predictive performance among the regression models for estimating curcuminoid contents from UV-Vis spectral data. Specifically, the linear regression models outperformed the nonlinear models. This finding confirms that the relationship between UV-Vis absorbance spectra and the contents of the three curcuminoids is highly linear, in agreement with the Beer-Lambert law.

Table 4. *Results of regression model development using nonlinear models*

Model		ANN		RF	
		Training set	**Test set**	**Training set**	**Test set**
CUR	R^2	0.991	0.912	0.986	**0.932**
	RMSE	1.599	4.390	1.944	**3.867**
DMC	R+	0.996	**0.935**	0.986	0.933
	RMSE	0.503	**1.840**	0.922	1.874
BDMC	R^2	0.992	0.915	0.993	**0.934**
	RMSE	1.171	4.020	1.140	**3.554**

In contrast, nonlinear models generally require large datasets to achieve optimal performance, as they tend to learn noise when the training data are limited, leading to overfitting. The UV-Vis spectral dataset used in this study comprised a relatively small number of samples, exhibited fairly simple characteristics, and contained limited nonlinear features. Therefore, the application of complex nonlinear models such as ANN or RF did not provide substantial benefits and in some cases, even reduced predictive performance.

Some predicted values still showed relatively large errors, particularly five cases with relative errors exceeding 20%. This may be attributed to the considerable variation in curcuminoid contents among samples, which makes it challenging for the models to fully capture the variability of the data. In addition, background interference also contributed to the reduced predictive performance. Turmeric extracts contain various UV-Vis absorbing compounds that may interfere with the signals of curcuminoids. Although preprocessing techniques and multivariate regression methods can partially mitigate these effects, the overall model accuracy remains affected. To further improve model performance, increasing the sample size in future studies is recommended.

Overall, the predicted results obtained from the two linear regression models, PCR and PLSR, showed strong correlation with the quantitative results determined by HPLC. This suggests that the proposed approach has potential for rapid estimation of curcuminoid content. However, due to the limited number of samples used for model development and the inherently lower sensitivity and selectivity of UV-Vis spectroscopy compared with HPLC, this method cannot yet replace HPLC in applications requiring high analytical sensitivity.

4. CONCLUSION

By measuring the UV-Vis full spectra of turmeric extract solutions in combination with linear regression models (PLSR and PCR), the contents of the three curcuminoids can be predicted without prior separation from the sample matrix or from each other. This approach offer a rapid and cost-effective analytical technique with simple sample preparation, which helps minimize operational errors. In addition, it shows potential for the rapid quantification of target compounds in complex samples without the need for reference standards. However, a larger sample size is recommended to enhance model robustness.

REFERENCES

[1]. D.H Bich, D.Q. Chung, B.X. Chuong, "The medicinal plants and animals in Vietnam," Hanoi for science and technology, vol. 2, pp. 1256, 2006 [in Vietnamese].

[2]. A. M. Araya-Sibaja, F. Vargas-Huertas, S. Quesada *et al.*, "Characterization, antioxidant and cytotoxic evaluation of demethoxycurcumin and bisdemethoxycurcumin from *Curcuma longa* cultivated in Costa Rica," *Separations*, vol. 11, no. 1, pp. 23, 2024.

[3]. V. S. R. Kotra, L. Satyabanta, and T. K. Goswami, "A critical review of analytical methods for determination of curcuminoids in turmeric," *Journal of Food Science and Technology*, vol. 56, no. 12, pp. 5153-5166, 2019.

[4]. T. T. Hue, T. T. T. Dung, N. V. Ri *et al.*, "Simultaneous determination of caffeine, theobromine, theophylline in tea using ultraviolet-visible spectroscopy combined with multivariate analysis," *Journal of Analytical Sciences*, vol. 24, no. 1, pp. 192-196, 2019 [in Vietnamese].

[5]. A. Alqahtani, T. Alqahtani, A. A. Fatease *et al.*, "Rapid UV-Vis spectrophotometric method aided by firefly-PLS models for the simultaneous quantification of ciprofloxacin, lomefloxacin, and enrofloxacin in their laboratory mixture, dosage forms and water samples: greenness and blueness assessment," *BMC Chemistry*, vol. 18, no. 1, 2024.

[6]. R. Khani, R. Rahmanian, and N. V. Motlagh, "UV-Visible spectrometry and multivariate calibration as a rapid and reliable tool for simultaneous quantification of ternary mixture of phenolic acids in fruit juice samples," *Food Analytical Methods*, vol. 9, no. 5, pp. 1112-1119, 2016.

[7]. R. Ergon, "Principal component regression (PCR) and partial least squares regression (PLSR)," in *Mathematical and Statistical Methods in Food Science and Technology*, pp. 143-174, 2013.

[8]. K. Upreti, M. Verma, M. Agrawal *et al.*, "Prediction of mechanical strength by using an artificial neural network and random forest algorithm," *Journal of Nanomaterials*, vol. 2022, no. 1, 2022.

[9]. D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, 2021.