

GEOGRAPHICAL ORIGIN CLASSIFICATION OF VIETNAMESE TURMERIC (*CURCUMA LONGA* L.) BASED ON UV-VIS AND FTIR SPECTRA COMBINED WITH MACHINE LEARNING

Nguyen Thi Van Anh¹, Nguyen Thu Ha¹, Nguyen Duc Phong², Nguyen Duc Thanh^{3*}

¹VietNam University of Traditional Medicine and Pharmacy

²TRAPHACO joint stock company, ³Vietnam Military Medical University

ARTICLE INFO		ABSTRACT
Received:	06/8/2025	A simple technique was developed to determine the geographical origin of Vietnamese turmeric (<i>Curcuma longa</i> L.) using UV-Vis and FTIR spectral data combined with machine learning models. A total of 160 turmeric samples collected from four northern provinces of Vietnam were measured UV-Vis and IR spectra. Spectral data was preprocessed using SNV, Savitzky-Golay and SPA algorithms to minimize measurement variability and reduce data dimensionality. Both supervised and unsupervised machine learning models were developed for geographic classification. The results showed that supervised models yielded high classification accuracy: the Linear Discriminant Analysis (LDA) model achieved the highest accuracy with 97.92% for UV-Vis and 95.83% for IR data. The SVM-LDA hybrid model also performed well, with accuracies of 95.83% (UV-Vis) and 93.75% (IR). The findings highlight the potential of spectral data and machine learning in the traceability of medicinal herbs in Vietnam.
Revised:	26/12/2025	
Published:	31/12/2025	
KEYWORDS		
Geographical origin		
Machine learning		
<i>Curcuma longa</i> L.		
UV-Vis		
FTIR		

PHÂN LOẠI NGUỒN GỐC ĐỊA LÝ NGHỆ VÀNG VIỆT NAM (*CURCUMA LONGA* L.) DỰA TRÊN DỮ LIỆU PHỔ UV-VIS VÀ FTIR KẾT HỢP HỌC MÁY

Nguyễn Thị Vân Anh¹, Nguyễn Thu Hà¹, Nguyễn Đức Phong², Nguyễn Đức Thanh^{3*}

¹Học viện Y Dược học cổ truyền Việt Nam, ²Công ty cổ phần TRAPHACO, ³Học viện Quân Y

THÔNG TIN BÀI BÁO		TÓM TẮT
Ngày nhận bài:	06/8/2025	Một kỹ thuật đơn giản đã được phát triển để xác định nguồn gốc địa lý nghệ vàng Việt Nam (<i>Curcuma longa</i> L.) thông qua việc sử dụng dữ liệu phổ UV-Vis và IR kết hợp với các mô hình học máy. Các dữ liệu phổ UV-Vis và IR được đo từ 160 mẫu nghệ vàng trồng tại bốn tỉnh khác nhau của miền Bắc Việt Nam. Các dữ liệu phổ được thực hiện các thuật toán tiền xử lý SNV, Savitzky-Golay, SPA để giảm các yếu tố ảnh hưởng do phép đo và làm giảm chiều của ma trận tín hiệu, sau đó phát triển các mô hình học máy có giám sát và không giám sát để xác định nguồn gốc địa lý. Kết quả cho thấy, các mô hình học máy có giám sát cho khả năng phân loại tốt: mô hình phân tích phân biệt tuyến tính (LDA) đạt độ chính xác phân loại cao nhất với cả hai bộ dữ liệu UV-Vis và IR lần lượt 97,92% và 95,83%. Mô hình SVM-LDA cũng cho kết quả khả quan với độ chính xác tương ứng 95,83% và 93,75%. Kết quả của nghiên cứu này mở ra triển vọng ứng dụng dữ liệu phổ kết hợp với học máy trong việc truy xuất nguồn gốc dược liệu tại Việt Nam.
Ngày hoàn thiện:	26/12/2025	
Ngày đăng:	31/12/2025	
TỪ KHÓA		
Nguồn gốc địa lý		
Học máy		
Nghệ vàng		
UV – Vis		
FTIR		

DOI: <https://doi.org/10.34238/tnu-jst.13380>

* Corresponding author. Email: nguyenducthanh@vmmu.edu.vn

1. Đặt vấn đề

Nghệ vàng (*Curcuma longa* L., họ Zingiberaceae) là một trong những dược liệu được sử dụng phổ biến ở nhiều nước trong đó có Việt Nam do có nhiều công dụng như: kháng viêm, chống oxy hóa, kháng khuẩn, kháng nấm, bảo vệ hệ tim mạch... [1]. Thành phần hóa học của nghệ vàng chịu ảnh hưởng bởi các yếu tố môi trường như thổ nhưỡng, khí hậu và kỹ thuật canh tác, dẫn đến sự khác nhau về chất lượng của dược liệu ở những vùng trồng khác nhau [2], [3]. Vì vậy, việc xác định nguồn gốc địa lý của nghệ vàng là một yêu cầu cấp bách trong việc chỉ dẫn địa lý, bảo hộ vùng nguyên liệu và góp phần xây dựng vùng trồng đặc hữu dược liệu.

Việc phân loại dược liệu thường dựa trên 2 phương thức: phân tích có mục tiêu và phân tích không mục tiêu. Để phân loại nghệ bằng phân tích có mục tiêu, các phương pháp hóa học kinh điển như sắc ký lỏng hiệu năng cao, sắc ký khối phổ đã được nghiên cứu và áp dụng trong việc đánh giá hàm lượng curcuminoid trong nghệ vàng. Phân tích không mục tiêu sử dụng các dữ liệu phổ như UV-Vis, IR, điện hóa... để phân loại mà không cần phân tích cụ thể chất nào [4], [5]. Phương pháp này đơn giản, nhanh và kinh tế hơn so với các phương pháp phân tích có mục tiêu.

Một số mô hình học máy phổ biến có thể được sử dụng trong phân loại bao gồm phân tích thành phần chính (PCA - Principal Component Analysis) [6], phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis) [7] và máy vector hỗ trợ (SVM - Support Vector Machine) [8]. Các phương pháp này cho phép phân loại và dự đoán nguồn gốc địa lý của dược liệu từ các đặc trưng của dữ liệu phổ UV, IR, điện hóa... mà không cần phải dựa vào hàm lượng các chất với các phương pháp hóa học phức tạp.

Pauzi và cộng sự [9] đã sử dụng bộ dữ liệu thu được bằng mũi điện tử kết hợp với PCA để phân loại các dược liệu họ Zingiberaceae, kết quả đã phân biệt được chín loại thảo dược với độ chính xác trên 95%. Để nâng cao hiệu suất của mô hình phân loại, nhiều mô hình tiền xử lý dữ liệu đã được sử dụng nhằm chuẩn hóa và chọn lọc các thông tin đặc trưng của dữ liệu. Một số mô hình thường được sử dụng là bộ lọc Savitzky-Golay, biến thiên chuẩn (SNV - Standard Normal Variate), mô hình chiếu liên tiếp (SPA - Successive Projections Algorithm),... D. Yaoyao và cộng sự [10] sử dụng SPA để chọn các bước sóng đặc trưng nhằm dự đoán hàm lượng tanshinone và phân loại nguồn gốc địa lý của *Salvia miltiorrhiza*. Khi sử dụng SPA, số lượng bước sóng có giá trị đều giảm trên 90% so với toàn phổ, hiệu suất của các mô hình đều rất tốt (RPD > 3). Rohaeti và cộng sự [11] nghiên cứu phân biệt *Curcuma longa*, *Curcuma xanthorrhiza* và *Zingiber cassumunar* dựa trên dữ liệu phổ FTIR. Kết quả cho thấy việc sử dụng SNV để chuẩn hóa dữ liệu phổ, loại bỏ ảnh hưởng của tán xạ ánh sáng cho kết quả tối ưu phân biệt 3 loài.

Nghiên cứu này kết hợp dữ liệu phổ UV và IR với kỹ thuật học máy để xác định nguồn gốc địa lý của nghệ vàng trồng tại Việt Nam. Đây là nghiên cứu đầu tiên tại Việt Nam ứng dụng học máy vào việc truy xuất nguồn gốc nghệ vàng, góp phần cung cấp một công cụ hiệu quả cho kiểm soát chất lượng và truy xuất nguồn gốc nghệ vàng.

2. Đối tượng và phương pháp nghiên cứu

2.1. Hóa chất và thiết bị

Dung môi methanol tinh khiết 99,7% của hãng Xilong (Trung Quốc). Nước cất 2 lần được lọc bằng hệ thống Milli - Q (Mỹ).

Hệ thống UV-Vis U-3900/3900H (Hitachi, Nhật Bản), cuvet thạch anh có chiều dày 1 cm, phần mềm PC OS WindowsTM.

Hệ thống quang phổ hồng ngoại FTIR Cary 630 (Agilent, Mỹ), phần mềm Microlab.

2.2. Lấy mẫu và chuẩn bị mẫu

Lấy mẫu: 160 mẫu nghệ tươi được thu thập từ 4 tỉnh miền Bắc Việt Nam gồm: Hà Nội, Bắc Ninh, Hưng Yên, Tuyên Quang, mỗi tỉnh 40 mẫu. Thời gian lấy mẫu từ tháng 11/2023 đến tháng

11/2024, mẫu được lấy theo tiêu chuẩn quốc gia TCVN 8958:2011 (ISO 5562:1983) và được mã hóa với các thông tin về thời gian và địa điểm lấy mẫu.

Xử lý mẫu: Sử dụng các củ nghệ còn nguyên vẹn, không đập nát, không thổi, hư hỏng, sâu bệnh và có mùi thơm đặc trưng. Nghệ tươi được rửa sạch, gọt vỏ, thái lát mỏng và đem sấy khô ở nhiệt độ 50°C, sau đó nghiền thành bột thô, đo hàm ẩm theo Dược điển Việt Nam 5. Các mẫu được bảo quản trong bình hút ẩm ở nhiệt độ phòng.

Chuẩn bị mẫu: Cân khoảng 0,3000 g bột nghệ vào bình định mức 10 ml, định mức bằng methanol, trộn đều bằng thiết bị lắc vortex khoảng 10 giây. Ngâm mẫu trong 24 giờ ở nhiệt độ phòng, thỉnh thoảng lắc vortex, thu được dịch chiết.

Đo UV-Vis: pha loãng dịch chiết 1250 lần vào bình định mức 50 ml. Quét phổ UV-Vis trong dải bước sóng từ 200 nm đến 800 nm với độ rộng khe phổ 5 nm.

Đo FTIR: Dịch chiết được lọc qua giấy lọc 0,45 μm rồi cô đến cạn. Tiến hành đo FTIR trong dải bước sóng từ 648 cm^{-1} đến 4000 cm^{-1} , khoảng cách 2 cm^{-1} .

Dữ liệu được dán nhãn theo 4 tỉnh: HN - Hà Nội, BN - Bắc Ninh, TQ - Tuyên Quang, HY - Hưng Yên.

2.3. Phân tích dữ liệu

2.3.1. Tiền xử lý dữ liệu

Dữ liệu phổ thường chứa nhiễu do một số yếu tố như môi trường, thiết bị đo, ánh sáng tán xạ,... vì vậy, tiền xử lý dữ liệu giúp chuẩn hóa dữ liệu, đồng nhất các mẫu về cùng một thang đo, làm tăng độ chính xác của mô hình phân loại. Các mô hình tiền xử lý dữ liệu được sử dụng trong nghiên cứu này gồm: Savitzky-Golay làm mượt dữ liệu mà vẫn giữ được các đặc trưng quan trọng, điều này giúp giảm nhiễu mà không làm mất đi hình dạng tổng thể của tín hiệu. Thuật toán SNV loại bỏ ảnh hưởng của sự tán xạ và khác biệt về độ dày mẫu, loại bỏ sự khác biệt không mong muốn và làm nổi bật các đặc điểm quan trọng trong dữ liệu phổ. SPA làm giảm hiện tượng đa cộng tuyến giữa các biến trong dữ liệu đa chiều và nhưng vẫn giữ được thông tin chính của tập dữ liệu gốc.

2.3.2. Mô hình học máy

Để phân loại nguồn gốc nghệ, trong nghiên cứu này khảo sát các mô hình học máy gồm:

Phân biệt không giám sát: mô hình PCA - phương pháp thống kê tuyến tính được sử dụng để giảm chiều dữ liệu và tìm kiếm các mẫu quan trọng trong dữ liệu phức tạp.

Phân biệt có giám sát, bao gồm 2 mô hình: LDA là một phương pháp phân loại giảm chiều dữ liệu tương tự như PCA. LDA tìm ra các vectơ và giá trị riêng để xác định các trục phân biệt và SVM đề xuất cách tính toán một siêu phẳng (hyperplane) có khả năng phân chia các lớp dữ liệu.

Các mô hình khảo sát được đánh giá thông qua độ chính xác (1):

$$\text{Độ chính xác} = \frac{\text{Số mẫu phân loại đúng}}{\text{Tổng số mẫu}} \times 100 \quad (1)$$

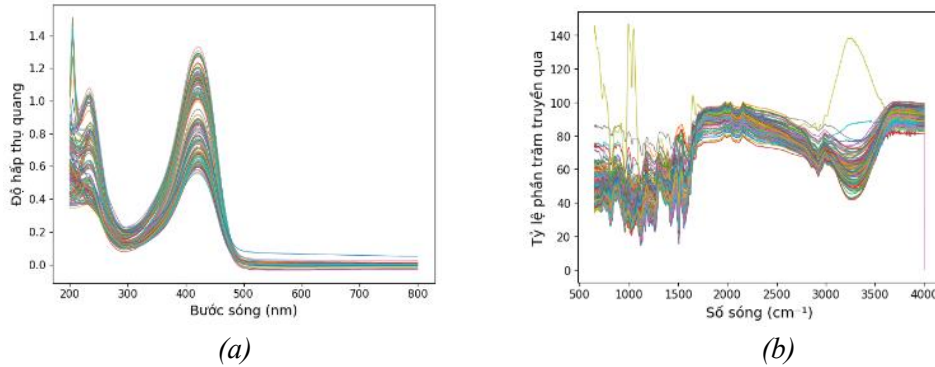
3. Kết quả nghiên cứu

3.1. Dữ liệu phổ và tiền xử lý

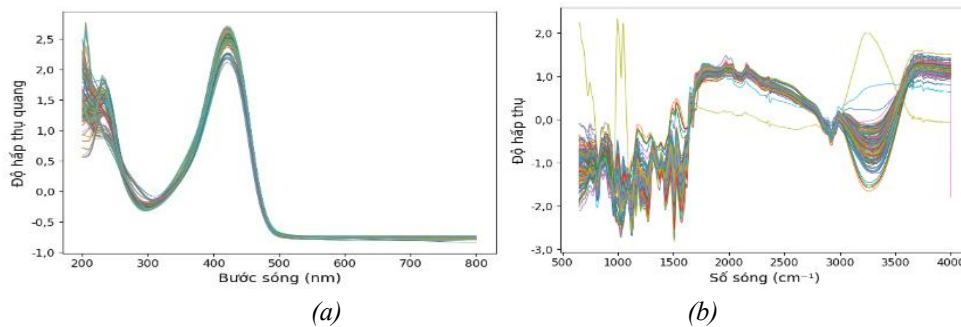
Phổ đồ của 160 mẫu nghệ được trình bày trong Hình 1. Phổ hồng ngoại (FTIR) được đo trong khoảng 648 cm^{-1} - 4000 cm^{-1} với khoảng cách đo 2 cm^{-1} thu được ma trận 160 mẫu x 1798 cường độ. Phổ UV được quét trong khoảng 800 nm - 200 nm, khoảng cách đo 5 nm thu được ma trận 160 mẫu x 121 cường độ.

Phổ UV được quét với khoảng cách đo 5 nm vẫn thể hiện rõ các cực đại hấp thụ tại 265 nm (vòng phenolic) và 425 nm (hệ liên hợp $\pi \rightarrow \pi^*$ của curcumin), phổ đồ không bị biến dạng khi so sánh với phép đo có khe phổ nhỏ hơn. Với phổ FTIR được quét với khoảng cách đo 2 cm^{-1} , các đỉnh 1627 cm^{-1} (nhóm C=O liên hợp), 1270 cm^{-1} (nhóm C-O aryl) và 3350 cm^{-1} (nhóm O-H) đều

được xác định chính xác, đồng thời hình dạng phổ không có sự khác biệt với các phép đo khác. Phổ IR của các mẫu nhiều đáng kể trong vùng phổ $648\text{ cm}^{-1} - 1600\text{ cm}^{-1}$. Phổ UV ít nhiễu hơn, tuy nhiên sự phân tán về cường độ giữa các mẫu nhiều. Để khắc phục, các mô hình tiền xử lý đã được sử dụng để chuẩn hóa dữ liệu, mô hình SNV giúp giảm nhiễu, các mẫu trở nên đồng nhất hơn, giảm sự phân tán. Mô hình làm mịn Savitzky-Golay tăng cường sự khác biệt giữa các mẫu, giúp mô hình hội tụ nhanh hơn. Phổ sau tiền xử lý được trình bày ở Hình 2.

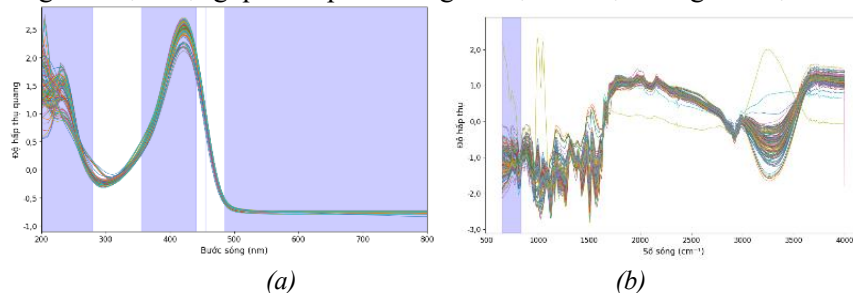


Hình 1. Phổ của các mẫu nghệ: (a) Phổ UV-Vis, (b) Phổ IR



Hình 2. Phổ các mẫu nghệ sau tiền xử lý: (a) Phổ UV-Vis, (b) Phổ IR

Sau khi đã làm mịn phổ bằng các mô hình trên, mô hình SPA chọn lọc ra các bước sóng chứa nhiều thông tin nhất mà không bị trùng lặp, loại bỏ các vùng phổ không cần thiết, giúp các mô hình phân loại hoạt động nhanh và chính xác hơn, ma trận tín hiệu giảm còn 160 mẫu x 100 cường độ đối với cả phổ UV-Vis và phổ IR như Hình 3, vùng phổ được lựa chọn biểu diễn bằng màu xanh nhạt. Bước giảm chiều bằng SPA không phải là phép nén ngẫu nhiên dữ liệu mà là quá trình lựa chọn bước sóng mang thông tin cao nhất, loại bỏ các điểm phổ dư thừa hoặc đồng tuyến tính cao, từ đó giúp mô hình giảm hiện tượng quá khớp mà vẫn giữ được các đặc trưng hóa học chính.



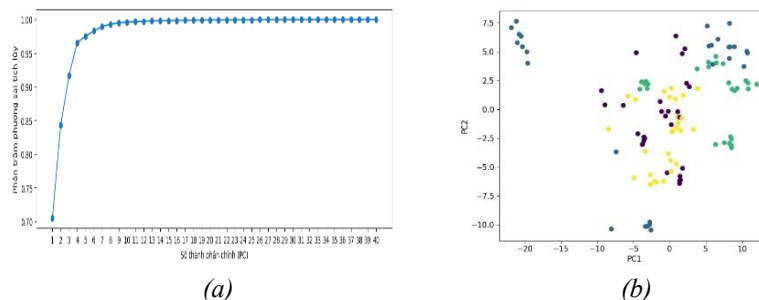
Hình 3. Vùng phổ được lựa chọn sau khi dùng SPA: (a) Phổ UV-Vis, (b) Phổ IR

Như vậy, các mô hình phân loại được khảo sát trên dữ liệu phổ đã được tiền xử lý. Các bộ tín hiệu được chia ngẫu nhiên thành tập huấn luyện và tập kiểm tra theo tỷ lệ tương ứng 70:30, tương ứng với 112 mẫu huấn luyện và 48 mẫu kiểm tra độ chính xác của mô hình.

3.2. Phân biệt không giám sát

3.1.1. Phân biệt không giám sát dựa trên phổ UV-Vis

Sử dụng mô hình PCA vào bộ dữ liệu phổ đã tiền xử lý, biểu đồ biểu diễn các giá trị phương sai được giải thích (explained variance) theo số lượng thành phần chính (PCs) thể hiện trong Hình 4.



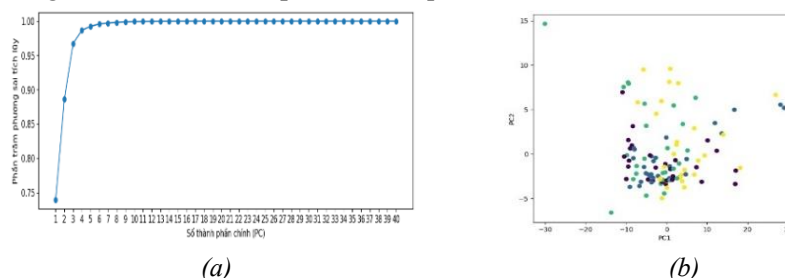
Hình 4. Kết quả PCA từ dữ liệu phổ UV: (a) Giá trị phương sai giải thích theo số PC, (b) Biểu đồ phân tán của 160 mẫu trên 2 PC đầu tiên

Kết quả cho thấy, với hai thành phần chính đầu tiên giá trị phương sai giữ lại là khoảng 85%. Điều này có nghĩa là chỉ với hai thành phần chính đầu tiên, phần lớn thông tin đặc trưng của dữ liệu phổ đã được bảo toàn, đảm bảo tính đại diện khi biểu diễn và phân tích trong không gian hai chiều. Phân tích tải trọng PCA cho thấy, các bước sóng có hệ số tải trọng tuyệt đối cao nhất của PC1 tập trung tại vùng 410 - 430 nm, trong khi PC2 tập trung tại vùng 200 - 220 nm. Các vùng này tương ứng với chuyển dời điện tử $\pi \rightarrow \pi^*$ của hệ liên hợp curcuminoid (vùng khả kiến) và $n \rightarrow \pi^*$ của liên kết C=O trong các nhóm phenolic (vùng tử ngoại), phản ánh sự biến thiên hàm lượng giữa curcumin và flavonoid trong các mẫu nghệ khác nhau. Phân tích 160 mẫu trên hai thành phần chính đầu tiên, bộ dữ liệu phân tách thành bốn nhóm mẫu tương ứng với 4 tỉnh. Các mẫu nghệ từ bốn tỉnh không phân tách rõ ràng với nhau trên không gian hai thành phần chính. Các mẫu cùng màu (cùng vùng địa lý) không tập trung thành cụm riêng biệt mà phân bố rải rác và có sự chồng lấn đáng kể với các nhóm khác. Điều này phản ánh mức độ tương đồng cao giữa các mẫu hoặc độ biến thiên lớn trong cùng một vùng. Điều này khẳng định rằng sự khác biệt địa lý của mẫu nghệ chủ yếu xuất phát từ biến thiên hàm lượng curcuminoid và mức độ liên hợp

Có thể thấy, sử dụng PCA dựa trên phổ UV-Vis đã có thể chia thành 4 nhóm theo vùng địa lý, tuy nhiên khả năng phân loại riêng biệt bốn nhóm mẫu đạt hiệu quả không cao.

3.1.2. Phân biệt không giám sát dựa trên phổ IR

Kết quả sử dụng mô hình PCA với phổ IR, kết quả được biểu diễn trên Hình 5.



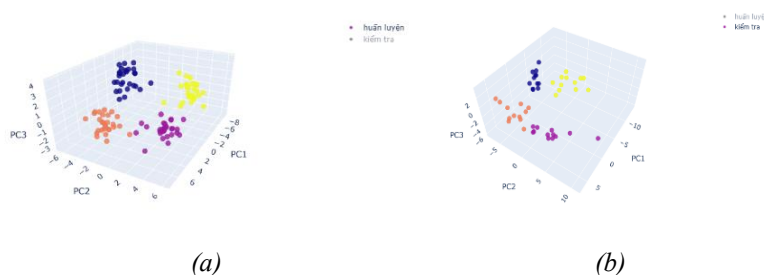
Hình 5. Kết quả PCA từ dữ liệu phổ IR: (a) Giá trị phương sai giải thích theo số PC, (b) Biểu đồ phân tán của 160 mẫu trên 2 PC đầu tiên

Kết quả PCA với tập dữ liệu IR cho thấy, với hai thành phần chính có thể giữ lại hơn 88% giá trị phương sai được giải thích. Phân tích tải trọng PCA cho thấy, các số sóng có hệ số tải trọng lớn nhất của PC1 nằm tại vùng 650 - 810 cm^{-1} (dao động uốn ngoài mặt phẳng $=\text{C}-\text{H}$ trong hệ

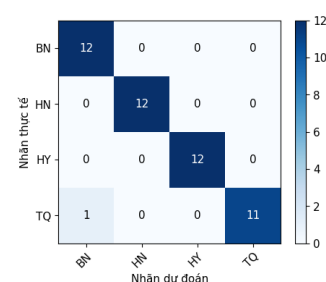
thơm liên hợp), trong khi PC2 có hệ số tải trọng cao nhất tại vùng 1503–1511 cm^{-1} (dao động kéo giãn C=C trong vòng thơm). Như vậy, hai thành phần chính đầu tiên chủ yếu phản ánh sự khác biệt về mức độ liên hợp và hàm lượng curcuminoid giữa các vùng địa lý, thay vì dao động C=O riêng lẻ ở 1620 cm^{-1} . Biểu diễn dữ liệu phổ FTIR trên hai thành phần chính đầu tiên cho thấy các mẫu nghệ từ bốn tỉnh không hình thành các cụm rõ ràng. Tương tự kết quả của phổ UV-Vis, các điểm dữ liệu cùng màu đại diện cho các mẫu cùng vùng địa lý phân bố rải rác, không tập trung và bị chồng lấn nhiều với các nhóm khác. Sự chồng lấn này phản ánh phổ FTIR giữa các vùng có mức độ tương đồng cao hoặc có sự biến thiên lớn trong nội bộ từng vùng, làm giảm khả năng phân biệt nguồn gốc địa lý chỉ dựa trên PCA.

3.3. Phân biệt có giám sát dựa trên phổ UV-Vis

3.3.1. Mô hình LDA



Hình 6. Biểu đồ phân tán mô hình LDA dựa trên phổ UV: (a) Tập huấn luyện, (b) Tập kiểm tra



Hình 7. Ma trận nhầm lẫn mô hình LDA trên phổ UV-Vis

Hình 6 thể hiện khi sử dụng mô hình LDA kết hợp với kết quả giảm chiều của PCA trên bộ dữ liệu phổ UV, sự phân tách thành các cụm rõ ràng ở cả tập huấn luyện và tập kiểm tra cho thấy mô hình đã học được các thành phần phân biệt tốt từ tập huấn luyện và có thể áp dụng hiệu quả cho tập test. Đồng thời, trên tập kiểm tra, các cụm dữ liệu vẫn duy trì được sự sắp xếp tương đồng với tập huấn luyện, chứng tỏ mô hình có khả năng tổng quát hóa tốt. Một số điểm dữ liệu trong tập kiểm tra có sự phân tán nhẹ, điều này có thể là do sự biến thiên tự nhiên của dữ liệu chưa được quan sát trong quá trình huấn luyện. Tuy nhiên, sự phân tán này không quá lớn và không làm giảm đáng kể độ phân tách giữa các nhóm, có thể kết luận rằng mô hình LDA hoạt động ổn định và không gặp hiện tượng chồng lấn. Ma trận nhầm lẫn của mô hình LDA được biểu diễn trên Hình 7.

Kết quả phân loại trên 48 mẫu kiểm tra cho thấy mô hình phân loại bằng LDA cho độ chính xác lên đến 97,92% (47/48 mẫu đúng). Cụ thể, hầu hết các vùng đều được phân loại chính xác, ngoại trừ một trường hợp nhầm lẫn mẫu Tuyên Quang thành Bắc Ninh. Điều này cho thấy LDA đã tối ưu hóa không gian đặc trưng hiệu quả, giúp phân tách các nhóm dữ liệu một cách rõ ràng.

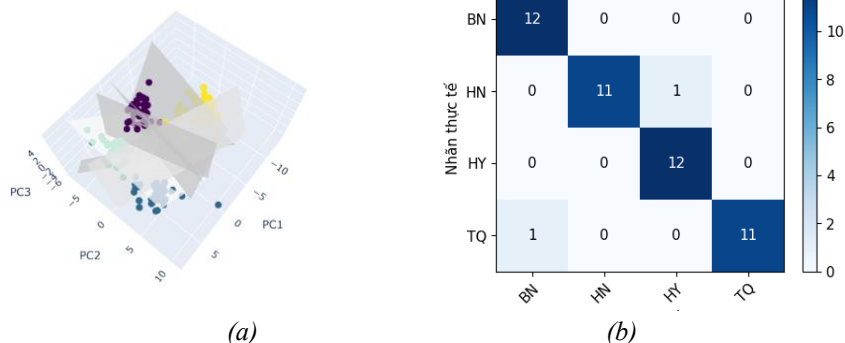
3.3.2. Mô hình SVM-LDA

Mô hình SVM được huấn luyện trên không gian đặc trưng giảm chiều bằng LDA để thực hiện phân loại nhiều vùng. Quá trình huấn luyện sử dụng phương pháp one-vs-one (OVO) với hàm nhân tuyến tính, cho phép xác định các mặt phẳng phân tách giữa các nhóm dữ liệu. Sau khi huấn luyện, mô hình được đánh giá bằng cách tạo một lưới điểm 3D bao phủ không gian đặc trưng và các bề mặt quyết định được biểu diễn bằng phương pháp bề mặt đẳng mức, giúp hiển thị trực tiếp các vùng biên giữa các vùng. Kết quả được chỉ ra trên Hình 8.

Kết quả cho thấy các điểm dữ liệu trong tập huấn luyện và kiểm tra đều được phân tách rõ ràng trong không gian 3D, các dữ liệu tập trung thành từng cụm theo vùng địa lý. Việc sử dụng kernel tuyến tính trong SVM trên không gian LDA đã đảm bảo quá trình phân loại hiệu quả.

Ma trận nhầm lẫn cho thấy, hầu hết các mẫu được phân loại chính xác, chỉ có một mẫu nghệ Tuyên Quang bị dự đoán sai thành Bắc Ninh và một mẫu nghệ của Hà Nội bị nhầm thành Hưng

Yên. Độ chính xác của mô hình đạt 95,83%, chứng tỏ hiệu quả của việc kết hợp SVM với LDA trong phân loại nghệ theo vùng địa lý. Tuy nhiên, độ chính xác thấp hơn so với mô hình LDA (97,92%), cho thấy sử dụng LDA để phân loại dữ liệu UV có vẻ phù hợp hơn.

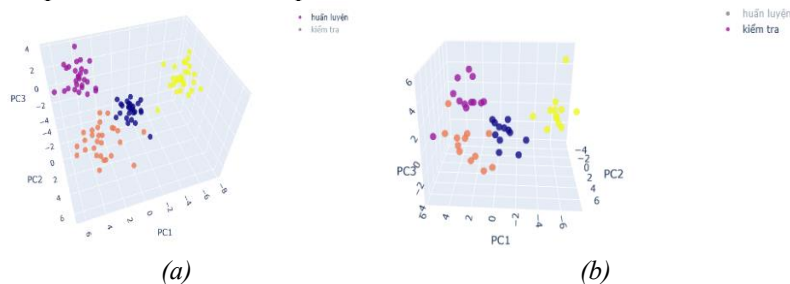


Hình 8. Mô hình SVM-LDA dựa trên phổ UV - Vis: (a) Ranh giới quyết định, (b) Ma trận nhầm lẫn

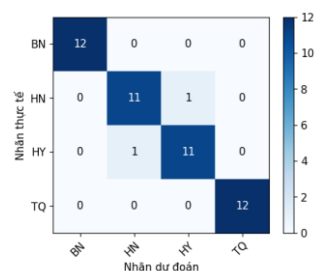
3.4. Phân biệt có giám sát trên phổ IR

3.4.1. Mô hình LDA

Tương tự với dữ liệu phổ UV, áp dụng LDA kết hợp với kết quả giảm chiều của PCA cho dữ liệu phổ IR thu được kết quả như Hình 9.



Hình 9. Biểu đồ phân tán mô hình LDA dựa trên phổ IR: (a) Tập huấn luyện, (b) Tập kiểm tra

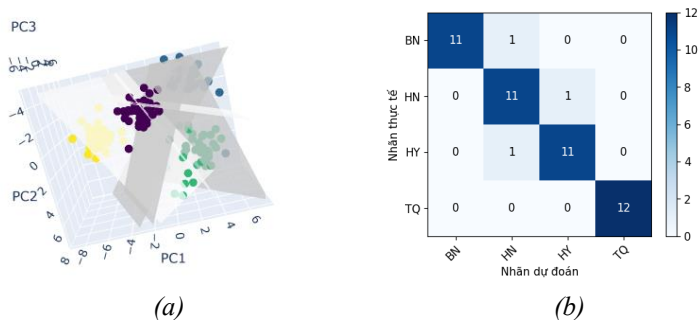


Hình 10. Ma trận nhầm lẫn mô hình LDA trên phổ IR

Trong tập huấn luyện, các cụm dữ liệu được phân tách rõ ràng, cho thấy LDA đã thành công trong việc giảm chiều dữ liệu và tối ưu hóa khả năng phân biệt giữa các vùng. Đối với tập kiểm tra, phân bố dữ liệu vẫn giữ được cấu trúc tương tự nhưng có một số điểm bị phân tán.

Ma trận nhầm lẫn của mô hình LDA trên phổ IR được thể hiện trên Hình 10. Kết quả phân loại 48 mẫu kiểm tra cho thấy mô hình phân loại đúng phần lớn các mẫu, tuy nhiên vẫn còn hiện tượng dự đoán nhầm giữa nghệ của Hưng Yên và Hà Nội. Mô hình có độ chính xác đạt 95,83%.

3.4.2. Mô hình SVM-LDA



Hình 11. Mô hình SVM-LDA dựa trên phổ IR: (a) Ranh giới quyết định, (b) Ma trận nhầm lẫn

Kết quả áp dụng mô hình SVM-LDA cho bộ dữ liệu phổ IR được thể hiện trên Hình 11.

Trên không gian ba chiều của mô hình SVM-LDA, cho thấy dữ liệu thuộc các vùng khác nhau được phân tách tương đối rõ ràng. Tuy nhiên vẫn có một số điểm dữ liệu gần ranh giới phân tách, khả năng tồn tại một số lỗi phân loại. Ma trận nhầm lẫn của mô hình cho thấy, có 3 mẫu dự đoán nhầm, độ chính xác của mô hình đạt 93,75%. Độ chính xác thấp hơn so với mô hình LDA (95,83%), chứng tỏ mô hình LDA phù hợp hơn để phân loại cho bộ dữ liệu IR.

4. Kết luận

Dữ liệu phổ của các mẫu nghệ bao gồm phổ UV-Vis và IR thể hiện sự khác biệt về cường độ và hình dạng giữa các vùng địa lý. Sau khi xử lý dữ liệu phổ bằng các mô hình tiền xử lý như Savitzky-Golay, SNV và SPA, ma trận tín hiệu đã giảm từ 160 mẫu x 1798 cường độ IR và 160 mẫu x 121 cường độ UV còn 160 mẫu x 100 cường độ với cả hai loại phổ. Phân tích thành phần chính (PCA) cho kết quả phân tách giữa các cụm mẫu từ các tỉnh chưa rõ ràng, cho thấy PCA chỉ phù hợp để giảm chiều dữ liệu và khám phá cấu trúc phân bố sơ bộ, chưa đủ mạnh để thực hiện phân loại độc lập theo tỉnh. Tuy nhiên, PCA góp phần xác định hai vùng phổ có giá trị phân biệt cao nhất là 420 – 425 nm của phổ UV-Vis và 1503–1511 cm^{-1} của phổ IR. Hai vùng này lần lượt đặc trưng cho chuyển dời điện tử $\pi \rightarrow \pi^*$ trong hệ liên hợp curcuminoid và dao động kéo giãn C=C thơm trong khung polyphenol. Sự khác biệt cường độ tín hiệu ở các vùng này phản ánh thay đổi về hàm lượng và tỷ lệ giữa các dạng curcuminoid trong mẫu nghệ. Cụ thể, nghệ Hưng Yên có tín hiệu hấp thụ mạnh hơn tại 420 nm, cho thấy hàm lượng curcumin cao hơn, trong khi nghệ Tuyên Quang thể hiện tín hiệu mạnh hơn tại 1507–1511 cm^{-1} , phản ánh tỷ lệ demethoxycurcumin lớn hơn. Những khác biệt này nhiều khả năng xuất phát từ điều kiện thổ nhưỡng và khí hậu khác nhau.

Hai mô hình học máy có giám sát là LDA và SVM-LDA cho thấy hiệu quả phân loại cao. Mô hình LDA trên bộ dữ liệu UV - Vis và IR đạt độ chính xác tương ứng 97,92% và 95,83%, cho thấy khả năng phân biệt tốt giữa các nhóm mẫu nghệ theo tỉnh thành. Mô hình SVM-LDA cũng cho kết quả khả quan với độ chính xác đạt 95,83% với dữ liệu UV-Vis và 93,75% với dữ liệu IR. Tuy có độ chính xác thấp hơn một chút so với LDA, nhưng SVM-LDA vẫn chứng minh khả năng xử lý dữ liệu phổ sau khi tiền xử lý dữ liệu.

Như vậy, để phân loại nghệ theo vùng địa lý, phương pháp tối ưu là sử dụng bộ dữ liệu phổ UV-Vis kết hợp với mô hình LDA. Trong thời gian tới, để tăng thêm độ chính xác của mô hình cần xây dựng một tập dữ liệu phổ phong phú hơn bằng cách thu thập số lượng mẫu lớn và từ nhiều vùng khác để kiểm chứng tính ổn định và khả năng ứng dụng của phương pháp trong thực tế.

Nhìn chung, nghiên cứu này đã phát triển mô hình xác định nguồn gốc địa lý trồng nghệ vàng (*Curcuma longa* L.) dựa trên dữ liệu phổ UV-Vis và IR kết hợp với các phương pháp học máy. Phương pháp này giúp xác định chính xác nguồn gốc địa lý của nghệ vàng, mang lại hiệu quả cao với chi phí thấp, thời gian phân tích nhanh và độ chính xác cao. Kết quả của nghiên cứu này cho thấy sự kết hợp giữa phổ UV-Vis và học máy mang lại một hướng tiếp cận mới trong đánh giá chất lượng và truy xuất nguồn gốc nghệ và có thể triển khai trên các dược liệu khác.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] C. W. Nita, "Safety and anti-inflammatory activity of curcumin: a component of tumeric (*Curcuma longa*)," *The Journal of Alternative & Complementary Medicine*, vol. 9, no. 1, pp. 161-168, 2003, doi: 10.1089/107555303321223035.
- [2] B. Kumar and B. Gill, "Effect of method of planting and harvesting time on growth, yield and quality of turmeric (*Curcuma longa* L.)," *Himachal Journal of Agricultural Research*, vol. 49, no. 2, pp. 253-256, 2011.
- [3] I. S. Sandeep, A. Kuanar, A. Akbar, *et al.*, "Agroclimatic zone based metabolic profiling of turmeric (*Curcuma Longa* L.) for phytochemical yield optimization," *Industrial Crops and Products*, vol. 85, pp. 229-240, 2016, doi: 10.1016/j.indcrop.2016.03.007.

-
- [4] T. H. Nguyen and T. T. H. Tran, "Preliminary investigation of curcumin contents in the rhizome of the yellow turmeric (*Curcuma longa* L. Zingiberaceae) according to cultivated regions by HPLC method," (in Vietnamese), *Journal Of Science of Lac Hong University*, vol. 12, pp. 029-032, 2021.
- [5] A.C. Gören, S. Çıkırıkçı, M. Çergel, *et al.*, "Rapid quantitation of curcumin in turmeric via NMR and LC–tandem mass spectrometry," *Food Chemistry*, vol. 113, no. 4, pp. 1239-1242, 2009, doi: 10.1016/j.foodchem.2008.08.014.
- [6] M. Greenacre, P. J. Groenen, T. Hastie, *et al.*, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, 2022, Art. no. 100, doi: 10.1038/s43586-022-00184-w.
- [7] S. Zhao, B. Zhang, J. Yang, *et al.*, "Linear discriminant analysis," *Nature Reviews Methods Primers*, vol. 4, no. 1, 2024, Art. no. 70, doi: 10.1038/s43586-024-00346-y.
- [8] S. Suthaharan, "Support Vector Machine," *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, Springer US, Boston, MA, 2016, pp. 207-235, doi: 10.1007/978-1-4899-7641-3_9.
- [9] A. Pauzi, N. Muhammad, N. Abdullah, *et al.*, "Discrimination of herbal products from Zingiberaceae family using electric nose combined with chemometric techniques," *Malays J. Chem.*, vol. 23, no. 2, pp. 205-212, 2021.
- [10] Y. Dai, B. Yan, F. Xiong, *et al.*, "Tanshinone content prediction and geographical origin classification of *Salvia miltiorrhiza* by combining hyperspectral imaging with chemometrics," *Foods*, vol. 13, no. 22, 2024, Art. no. 3673, doi: 10.3390/foods13223673.
- [11] E. Rohaeti, M. Rafi, U. D. Syafitri, *et al.*, "Fourier transform infrared spectroscopy combined with chemometrics for discrimination of *Curcuma longa*, *Curcuma xanthorrhiza* and *Zingiber cassumunar*," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 137, pp. 1244-1249, 2015, doi: 10.1016/j.saa.2014.08.139.