

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nguyễn Đức Thanh

NGHIÊN CỨU PHÂN BIỆT VÀ PHÂN LOẠI
CÁC SẢN PHẨM NƯỚC CAM TRÊN CƠ SỞ
PHÂN TÍCH THÀNH PHẦN HÓA HỌC, TÍN HIỆU ĐO
KẾT HỢP VỚI MÁY HỌC

Chuyên ngành: **Hóa phân tích**

Mã số: **9440118**

TÓM TẮT LUẬN ÁN TIẾN SĨ HÓA PHÂN TÍCH

Hà Nội - 2026

Công trình được hoàn thành tại:
Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học: **1. PGS.TS Tạ Thị Thảo**
2. PGS.TS Lê Thị Hồng Hảo

Phản biện: **PGS.TS. Phạm Thị Thanh Hà**

Trường Đại học Dược Hà Nội

Phản biện: **PGS. TS. Trịnh Thu Hà**

Viện Hóa học, Viện Hàn lâm KH&CN Việt Nam

Phản biện: **PGS.TS. Nguyễn Xuân Trường**

Trường Hóa và KHSS, ĐH Bách Khoa Hà Nội

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ họp tại Trường Đại học Khoa học Tự nhiên - ĐHQGHN vào
hồi giờ ngày tháng năm 2026.

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam;
- Trung tâm Thư viện và Tri thức số, Đại học Quốc gia Hà Nội

MỞ ĐẦU

1. Lý do chọn đề tài

Các hành vi gian lận thực phẩm nói chung và nước trái cây nói riêng bao gồm pha trộn với các thành phần rẻ hơn, tuyên bố sai xuất xứ, làm giảm chất lượng của sản phẩm, đánh lừa người tiêu dùng và thậm chí có thể tiềm ẩn nguy cơ về sức khỏe. Một trong những loại hoa quả phổ biến ở Việt Nam là Cam với nhiều giống và nhiều vùng trồng đặc hữu, việc phân loại và xác thực cam Việt Nam là một vấn đề quan trọng trong quản lý chất lượng cũng như an toàn thực phẩm. Phương pháp xác thực hiện nay tại Việt Nam thường dựa trên cảm quan theo kinh nghiệm và uy tín của nhà cung cấp.

Các phương pháp đánh giá truyền thống dựa trên một số chỉ tiêu hóa lý riêng lẻ như độ Brix, pH hoặc hàm lượng acid hữu cơ thường không đủ khả năng phản ánh toàn diện sự khác biệt về thành phần hóa học giữa các loại nước cam. Trong khi đó các phương pháp phân tích phổ cho phép thu nhận dữ liệu phổ và dấu vân tay hóa học của nước cam, phản ánh đồng thời nhiều nhóm hợp chất quan trọng như đường, acid hữu cơ và flavonoid. Tuy nhiên, dữ liệu phổ có tính đa chiều cao và mối tương quan phức tạp giữa các biến, gây khó khăn cho việc xử lý và diễn giải bằng các phương pháp thống kê truyền thống.

Học máy được xem là công cụ phù hợp, các thuật toán học máy cho phép xử lý đồng thời nhiều biến, nhận diện các mô hình phân bố ẩn trong dữ liệu và thực hiện phân loại mẫu với độ chính xác cao. Việc kết hợp học máy với dữ liệu phân tích hóa học giúp xây dựng các mô hình phân biệt có cơ sở khoa học rõ ràng và khả năng diễn giải tốt.

Tại Việt Nam, mặc dù cam là loại cây ăn quả quan trọng với nhiều giống và vùng trồng đặc hữu, các nghiên cứu về phát hiện gian lận và phân loại nước cam vẫn còn hạn chế và thiếu tính hệ thống. Chưa có công trình nào xây dựng đồng bộ cơ sở dữ liệu thành phần hóa học kết hợp dữ liệu tín hiệu phân tích của nước cam Việt Nam phục vụ cho mục tiêu phát hiện gian lận, phân loại giống và truy xuất nguồn gốc sản phẩm.

Xuất phát từ thực tiễn đó, luận án “**Nghiên cứu phân biệt và phân loại các sản phẩm nước cam trên cơ sở phân tích thành phần hóa học, tín hiệu đo kết hợp với máy học**” hướng đến mục tiêu xây dựng được cơ sở dữ liệu để phân biệt nước cam theo vùng địa lý, theo giống cam, nước cam tự nhiên với nước cam hóa học

2. Mục tiêu của luận án:

Luận án được thực hiện hướng đến các mục tiêu:

- Phát triển phương pháp định lượng một số chất chính thuộc các nhóm chất flavonoid, carotenoid, acid hữu cơ trong các mẫu nước cam.

- Nghiên cứu xây dựng phương pháp phân loại nước cam theo giống cam và theo nguồn gốc địa lý dựa trên thành phần hóa học kết hợp máy học.

- Nghiên cứu xây dựng phương pháp xác thực nguồn gốc địa lý của nước cam dựa trên dữ liệu phổ UV kết hợp máy học.

- Nghiên cứu xây dựng phương pháp phân biệt nước cam tự nhiên với nước cam pha chế dựa trên dấu vân tay điện hóa kết hợp máy học.

3. Nội dung của luận án

- Phát triển phương pháp HPLC để định lượng đồng thời các chất nhóm chất flavonoid, carotenoid, acid hữu cơ trong các mẫu nước cam: khảo sát điều kiện phân tích, xác nhận giá trị sử dụng của phương pháp và phân tích mẫu nước cam thực tế.

- Xây dựng phương pháp xác thực và phân loại các sản phẩm nước cam dựa trên thành phần hóa học kết hợp máy học.

- Xây dựng phương pháp xác thực và phân loại các sản phẩm nước cam dựa trên dữ liệu phổ UV và dữ liệu điện hóa kết hợp máy học.

4. Ý nghĩa khoa học và thực tiễn của đề tài nghiên cứu

Ở Việt Nam, những năm gần đây, ứng dụng chemometrics trong nhận dạng, phân loại đối tượng phân tích cũng như phân tích đồng thời các chất sử dụng trực tiếp tín hiệu phân tích trên thiết bị hoặc từ bộ số liệu kết quả phân tích thường được tiến hành với sự hỗ trợ của phần mềm Matlab. Tuy nhiên, phần mềm này có giá thành đắt, nếu không có sự hỗ trợ của các đơn vị đào tạo hoặc thương mại thì người dùng thường sử dụng phiên bản bẻ khóa, không chính thống. Vì vậy, khi nghiên cứu phát triển cơ sở dữ liệu hoặc phát triển thiết bị sẽ khó có khả năng mở rộng và liên kết với các phần mềm khác khiến cho việc thương mại hóa các sản phẩm trở nên khó khăn. Vì vậy, nhu cầu đặt ra cần sử dụng các ngôn ngữ lập trình mã nguồn mở như Python, R, Java... hoàn toàn miễn phí, có các thư viện hỗ trợ xử lý dữ liệu chất lượng khiến cho việc mở rộng và thương mại hóa trở nên dễ dàng hơn bao giờ hết.

Chất lượng của cam được quyết định bởi cả đặc tính cảm quan (mùi, màu, độ chua, độ ngọt) và thành phần hóa học tự nhiên.

Hàm lượng đường quyết định vị ngọt, trong khi acid hữu cơ mang lại vị chua và độ tươi; flavonoid góp phần tạo vị đắng nhẹ đặc trưng - đồng thời là chỉ tiêu đánh giá độ chín và chất lượng hương vị.

Hiện tại chưa có nghiên cứu nào xây dựng bộ dữ liệu về thành phần hóa học trong Cam Việt Nam. Cho đến nay cũng chưa có công bố nào xác thực và phân loại nước cam Việt Nam dựa trên các thành phần hóa học hay các tín hiệu phân tích. Vì vậy, đề tài góp phần giải quyết hai vấn đề trên.

Đề tài đã xây dựng các phương pháp định lượng 3 nhóm hợp chất chính có mặt trong nước cam bao gồm: acid hữu cơ, carotenoid, flavonoid, các phương pháp này đã được đưa vào ứng dụng thực tế, đang được triển khai thường quy tại Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia. Trong đó có 02 phép thử (acid, carotenoid) đã được công nhận phù hợp theo tiêu chuẩn ISO 17025:2017. Số hiệu phòng thử nghiệm VILAS 203.

Đề tài thành công sẽ mở ra cơ hội ứng dụng lớn cho các cơ sở kiểm nghiệm trong việc xây dựng được bộ dữ liệu nhận dạng, phân biệt các sản phẩm nước cam trên thị trường nhằm đánh giá tính xác thực các công bố của nhà sản xuất. Đồng thời, việc truy xuất nguồn gốc của các sản phẩm Cam tươi một cách chính xác cũng có thể triển khai theo nhiều cách tùy thuộc điều kiện phòng thí nghiệm. Hơn nữa có thể mở rộng ứng dụng cho nhiều loại thực phẩm, dược liệu, dược phẩm..., góp phần giải quyết các vấn đề thực tế cấp bách hiện nay tại Việt Nam.

Kết quả của luận án có thể áp dụng trên thực tế kiểm nghiệm ở nhiều đơn vị, do đề tài tiến hành thử nghiệm trên nhiều loại thiết bị như UV, HPLC hay cực phổ. Các cơ sở có thể lựa chọn

phương pháp phù hợp với điều kiện thực tế tại đơn vị mình. Hơn nữa, để phân loại phân biệt, các thuật toán được xây dựng trên phần mềm mở Python nên không bị giới hạn bởi bản quyền, kinh phí và dễ dàng thay đổi thuật toán cho người sử dụng.

5. Những đóng góp mới của Luận án

Luận án có nhiều đóng góp mới cho chuyên ngành như sau:

- Đã xây dựng và thẩm định được phương pháp định lượng các thành phần hóa học chính có trong nước cam (acid hữu cơ, carotenoid, flavonoid). Đây là nghiên cứu toàn diện, đóng góp nâng cao chất lượng công tác kiểm nghiệm an toàn thực phẩm.

- Ứng dụng học máy xây dựng được các mô hình phân loại nguồn gốc địa lý của mẫu cam, phân biệt được nước cam tự nhiên và nước cam pha chế dựa trên thành phần hóa học và các dữ liệu phân tích (UV, Volt-ampere). Các mô hình nhận dạng, phân loại được xây dựng trên phần mềm Visual Studio Code với ngôn ngữ Python phiên bản 3.12, phiên bản mới nhất hiện có. Đồng thời nguồn dữ liệu đầu vào đa dạng từ những kỹ thuật đơn giản như UV đến LC.

6. Bố cục của luận án

Luận án gồm 143 trang với 62 hình, 30 bảng số liệu và 151 tài liệu tham khảo. Luận án được cấu tạo gồm: 10 trang danh mục các bảng, hình vẽ, chữ viết tắt và mục lục, 2 trang mở đầu, 30 trang tổng quan tài liệu, 14 trang thực nghiệm và phương pháp nghiên cứu, 64 trang kết quả và thảo luận, 3 trang kết luận. Ngoài ra, luận án còn có 2 trang danh mục công trình của tác giả có liên quan đến luận án, 18 trang tài liệu tham khảo.

CHƯƠNG 1. TỔNG QUAN

Chương này đề cập đến các nội dung như sau:

Tổng quan về Cam: các loài Cam ở Việt Nam, thành phần hóa học chính có trong nước cam.

Tổng quan về các phương pháp xác định thành phần hóa học của nước cam: acid hữu cơ, carotenoid và flavonoid.

Tổng quan về học máy trong phân biệt, phân loại cam. Một số nghiên cứu phân biệt phân loại Cam đã triển khai trên thế giới.

CHƯƠNG 2. THỰC NGHIỆM VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Hóa chất, dụng cụ và thiết bị

2.1.1. Hóa chất

- Các chất chuẩn acid của hãng Merck (Đức), carotenoid của hãng Sigma Aldrich (Mỹ), flavonoid của Merck (Đức).

- Các hóa chất, dung môi khác có nguồn gốc từ Merck (Đức).

2.1.2. Thiết bị và dụng cụ

- Hệ thống sắc ký lỏng hiệu năng cao, thiết bị cực phổ, thiết bị UV-Vis, sắc ký lỏng khối phổ phân giải cao LC-HR/MS.

2.1.3. Xử lý số liệu

Phần mềm Python phiên bản 3.12

2.2. Đối tượng nghiên cứu

Đối tượng nghiên cứu là quả cam Việt Nam và nước cam đang được bán trên thị trường Hà Nội.

2.3. Phương pháp nghiên cứu

2.3.1. Xây dựng và xác nhận giá trị sử dụng các quy trình định lượng thành phần hóa học chính trong nước cam bằng phương pháp HPLC

- Định lượng các acid hữu cơ

- Định lượng các carotenoid

- Định lượng các flavonoid

*Bảng 2.1. Chương trình gradient pha động định lượng
flavonoid*

Chương trình Gradient 1			
<i>Thời gian (phút)</i>	<i>Tốc độ dòng (mL/phút)</i>	<i>Kênh A (%, v : v)</i>	<i>Kênh B (%, v : v)</i>
0,00	1,00	90	10
7,00	1,00	80	20
16,00	1,00	80	20
18,00	1,00	0	100
22,00	1,00	0	100
22,10	1,00	90	10
26,00	1,00	90	10
Chương trình Gradient 2			
<i>Thời gian (phút)</i>	<i>Tốc độ dòng (mL/phút)</i>	<i>Kênh A (%, v : v)</i>	<i>Kênh B (%, v : v)</i>
0,00	1,00	90	10
7,00	1,00	80	20
16,00	1,00	65	35
18,00	1,00	0	100
22,00	1,00	0	100
22,10	1,00	90	10
26,00	1,00	90	10

2.3.2. Phân biệt các sản phẩm nước cam dựa trên thành phần hóa học

Mô tả các phương pháp lấy mẫu, chuẩn bị mẫu, phương pháp phân tích thành phần hóa học chính và phương pháp xử lý dữ liệu.

2.3.3. Phân biệt các sản phẩm nước cam dựa trên dữ liệu phổ UV – Vis

Mô tả các phương pháp lấy mẫu, chuẩn bị mẫu, phương pháp quét phổ UV - Vis và phương pháp xử lý dữ liệu.

2.3.4. Xác định sự pha trộn trong nước cam bằng phân tích Volt-Ampere và học máy

Mô tả các phương pháp lấy mẫu, chuẩn bị mẫu, phương pháp đo volt-ampere và phương pháp xử lý dữ liệu.

CHƯƠNG 3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Nghiên cứu xây dựng và thẩm định phương pháp định lượng các thành phần hóa học có trong nước cam

3.1.1. Xây dựng và thẩm định phương pháp định lượng các acid trong nước cam

3.1.1.1. Khảo sát điều kiện tối ưu của phương pháp phân tích

Điều kiện tối ưu để định lượng 5 acid: oxalic, malic, ascorbic, citric, fumaric trong nước cam bằng phương pháp HPLC là:

- Cột sắc ký: cột ReliantTM C18 5 μ m (4,6 x 250mm)
- Pha động gồm hỗn hợp KH₂PO₄ nồng độ 10mM và H₃PO₄ 10% có pH 2,2
- Tốc độ dòng 0,6ml/phút
- Detector DAD
- Bước sóng phát hiện 254nm đối với acid ascorbic và 214nm với các acid còn lại.

3.1.1.2. Thẩm định quy trình phân tích

Độ ổn định của hệ thống

Độ lệch chuẩn tương đối của thời gian lưu và diện tích peak đối với 5 chất đều nhỏ hơn 2%, đáp ứng theo yêu cầu của AOAC

Độ đặc hiệu

Trên sắc ký đồ mẫu trắng, không xuất hiện các peak có cùng thời gian lưu với các peak của các chất phân tích trong mẫu chuẩn và thử. Trên sắc ký đồ mẫu thử thêm chuẩn có các peak có cùng thời gian lưu tương tự như trên sắc ký đồ của mẫu chuẩn.

Giới hạn phát hiện (MDL) và giới hạn định lượng (MQL)

Bảng 3.4. Kết quả thẩm định MDL, MQL của 5 acid

Chất phân tích	Acid oxalic	Acid malic	Acid ascorbic	Acid citric	Acid fumaric
MDL ($\mu\text{g/g}$)	0,60	0,9	3,9	14,4	0,15
MQL ($\mu\text{g/g}$)	2,0	3,0	13,0	48,0	0,5
R	4,3	8,2	4,2	4,3	8,7

Giá trị R của cả 5 acid đều nằm trong khoảng $4 < R < 10$.

Độ lặp lại và độ thu hồi

Kết quả cho thấy, khi phân tích hàm lượng các acid trong nước cam đều có giá trị $\text{RSD} < 5,3\%$, độ thu hồi và RSD đều nằm trong khoảng cho phép đáp ứng đúng yêu cầu của AOAC.

3.1.2. Xây dựng và thẩm định phương pháp định lượng các carotenoid trong nước cam

3.1.2.1. Khảo sát điều kiện tối ưu của phương pháp phân tích

Điều kiện phân tích đồng thời β -caroten, lycopene và lutein trong nước cam bằng HPLC cụ thể như sau:

- Cột sắc ký: Cột XBridge® C18 (4,6 x 250 mm, 5 µm) và tiền cột tương ứng.

- Pha động: ACN: MeOH: H₂O: CH₃COOC₂H₅ = 70: 9,6: 0,4: 20 (v/v/v/v).

- Tốc độ dòng: 1,0 mL/phút.

- Thời gian phân tích: 20 phút.

- Thể tích tiêm mẫu: 10 µL.

- Detector PDA.

- Bước sóng phát hiện: 455 nm.

3.1.2.2. *Xây dựng quy trình xử lý mẫu*

Quy trình chiết để phân tích hàm lượng carotenoid trong nước cam như sau:

Lấy 20,00 mL mẫu nước cam đã đồng nhất vào ống ly tâm 50 mL, thêm 15 mL dung môi chiết n-hexan, lắc ngang trong 10 phút. Ly tâm mẫu 5 phút ở tốc độ 6000 vòng/phút. Gạn dịch chiết, cô quay chân không sau đó định mức bằng 5 mL dung môi pha động (acetonitril: methanol: nước: ethyl acetat = 70: 9,6: 0,4: 20 (v/v/v/v)); lọc qua màng lọc 0,22 µm rồi bơm vào hệ thống sắc ký HPLC - PDA. Thực hiện chiết lặp 2 lần.

3.1.2.3. *Thẩm định quy trình phân tích*

Độ ổn định của hệ thống

Độ lệch chuẩn tương đối của thời gian lưu và diện tích peak đối với 3 chất đều nhỏ hơn 2%, đáp ứng theo yêu cầu của AOAC.

Độ đặc hiệu

Mẫu trắng không cho tín hiệu phân tích, mẫu chuẩn, mẫu thực và mẫu thực thêm chuẩn cho tín hiệu chất phân tích tại cùng thời gian lưu với sai lệch nhỏ hơn 3%.

Giới hạn phát hiện (MDL) và giới hạn định lượng (MQL)

Bảng 3.10. Kết quả thẩm định MDL, MQL của 3 carotenoid

Chất phân tích	MDL (mg/L)	MQL (mg/L)	R
Lutein	0,002	0,006	8,59
Lycopene	0,003	0,010	4,11
β -caroten	0,006	0,019	4,17

Giá trị R của 3 carotenoid đều nằm trong khoảng $4 < R < 10$.

Độ lặp lại và độ thu hồi

Phương pháp có độ lặp lại với RSD nằm trong khoảng 0,3 - 2,4% ($n = 6$) nhỏ hơn 11%, hiệu suất thu hồi và RSD nằm trong giới hạn cho phép đáp ứng yêu cầu của AOAC khi phân tích đồng thời β -caroten, lutein và lycopene trên nền mẫu nước cam.

3.1.3. Xây dựng và thẩm định phương pháp định lượng các flavonoid trong nước cam

3.1.3.1. Khảo sát điều kiện tối ưu của phương pháp phân tích

Điều kiện phân tích đồng thời naringin, hesperidin, neohesperidin và quercetin trong nước cam bằng HPLC cụ thể như sau:

- Cột sắc ký: Cột XBridge® Water C18 (4,6mm x 150mm; 5 μ m).

- Pha động: acid phosphoric 0,1% và ACN với chương trình gradient 2 ở Bảng 2.1.

- Tốc độ dòng 1,0 mL/phút.

- Thể tích tiêm mẫu: 10 μ L.

- Detector PDA

- Bước sóng phát hiện tại 284 nm đối với naringin, hesperidin, neohesperidin; tại 254 nm đối với quercetin.

3.1.3.2. Xây dựng quy trình xử lý mẫu

Lấy 5,00 mL mẫu nước cam đã đồng nhất vào bình định mức 25 mL, thêm dung môi chiết methanol đến vạch, lắc ngang trong 10 phút. Ly tâm mẫu và lọc qua màng lọc 0,45 μ m rồi bơm vào hệ thống sắc ký HPLC. Thực hiện chiết lặp 2 lần.

3.1.3.3. Thẩm định phương pháp

Độ ổn định hệ thống

Giá trị RSD theo thời gian lưu và theo diện tích pic của cả 4 chất chất đều nhỏ hơn 2%, đáp ứng theo yêu cầu của AOAC.

Độ đặc hiệu

Trên sắc ký đồ của mẫu chuẩn xuất hiện peak tại thời điểm 10,2 phút của naringin, 10,7 phút của hesperidin, 11,3 phút của neohesperidin và 13,8 phút của quercetin. Mẫu thêm chuẩn cho tín hiệu chất phân tích tại cùng thời gian lưu của mẫu chuẩn và sai lệch thời gian lưu không quá 1%. Mẫu trắng không cho tín hiệu của chất phân tích.

Giới hạn phát hiện (MDL) và giới hạn định lượng (MQL)

Bảng 3.16. Kết quả thẩm định MDL, MQL của 4 flavonoid

Chất phân tích	MDL (mg/L)	MQL (mg/L)	R
Naringin	0,12	0,40	8,07
Hesperidin	0,17	0,57	6,87
Neohesperidin	0,13	0,44	6,48
Quercetin	0,18	0,61	4,18

Các giá trị của hệ số R đều nằm trong giới hạn yêu cầu $4 \leq R \leq 10$ của AOAC.

Độ lặp lại và độ thu hồi

Kết quả cho thấy phương pháp có độ lặp lại nằm trong khoảng giới hạn cho phép với RSD đạt từ 1,35 - 2,48% ($n = 6$) nhỏ hơn 11%, hiệu suất thu hồi và RSD nằm trong giới hạn cho phép đáp ứng yêu cầu của AOAC.

3.2. Phân loại các sản phẩm nước cam dựa trên thành phần hóa học

3.2.1. Phân tích thống kê

Kết quả phân tích hàm lượng các hợp chất chính cho thấy nồng độ các chất này biến thiên đáng kể theo giống cam và vùng địa lý trồng, với sự khác biệt có ý nghĩa thống kê.

3.2.2. Sự khác biệt về thành phần hóa học trong nước cam

Kết quả ANOVA 1 chiều cho thấy đa số các hợp hóa học có mối liên hệ chặt chẽ với cả hai tiêu chí phân loại (nguồn gốc và giống), trong đó lutein, β -caroten, quercetin, hesperidin, glucose/fructose) là các ‘marker’ tiềm năng. Các biến ít đặc trưng theo giống (sucrose, fumaric acid) vẫn có giá trị cho phân loại theo địa phương. Như vậy, với các thành phần hóa học trong nước cam Việt Nam khả năng phân loại theo nguồn gốc địa lý có nhiều dấu vân tay mạnh hơn so với phân loại theo giống.

3.2.3. Phân loại bằng học máy dựa trên thành phần hóa học

3.2.3.1. Phân loại cam theo giống

Kết quả PCA cho thấy phương pháp này có khả năng phân biệt hiệu quả các giống cam dựa trên dấu vân tay thành phần hóa học, đồng thời làm nổi bật những chỉ thị sinh hóa chủ chốt đóng vai trò quyết định trong sự khác biệt giữa các giống

Bảng 3.21. Kết quả phân loại theo giống cam.

	SVC	DT	RF	MLP	XGBoost
Không SMOTE	0,64	0,588	0,765	0,706	0,706
SMOTE	0,896	0,897	0,897	0,931	0,897

Các kết quả phân loại thể hiện trên Bảng 3.21, độ chính xác tổng thể cao nhưng độ thu hồi (recall) thấp đối với lớp thiểu số. Su khi áp dụng SMOTE đã giảm thiểu một phần hạn chế này. Sự cải thiện nhất quán về hiệu năng của các mô hình SVC, DT, RF và XGBoost từ khoảng 0,6 - 0,7 lên gần 0,9 cho thấy SMOTE không chỉ làm giảm thiên lệch mà còn làm trơn ranh giới quyết định, đặc biệt có lợi đối với các mô hình dựa trên phân hoạch như DT và RF.

Quan sát đáng chú ý là mô hình MLP đạt hiệu năng cao nhất (0,931), phản ánh khả năng học các ánh xạ phi tuyến tổng quát vượt trội so với các mô hình dựa trên phân loại rời rạc hoặc kernel cố định.

3.2.3.2. Phân loại cam theo nguồn gốc địa lý

Kết quả phân loại được thể hiện trong bảng 3.22.

Bảng 3.22. Kết quả phân loại cam theo nguồn gốc địa lý

	SVC	DT	RF	MLP	XGBoost
Không SMOTE	0,823	0,529	0,882	0,765	0,647
SMOTE	1,0	0,906	1,0	0,875	0,969

Các mô hình cho thấy sự cải thiện đáng kể về độ chính xác khi áp dụng SMOTE. Trong trường hợp chưa áp dụng SMOTE, RF (0,882) và SVC (0,823) thể hiện hiệu năng vượt trội so với các phương pháp học máy khác. Sau khi áp dụng SMOTE, tất cả các mô hình đều được cải thiện rõ rệt, trong đó SVC và RF đạt độ chính xác tuyệt đối (1,0), còn XGBoost đạt 0,969.

Như vậy, việc phân loại giống cam và xác định nguồn gốc địa lý có thể được thực hiện hiệu quả dựa trên thành phần hóa học của nước cam khi kết hợp với các thuật toán học máy.

3.3. Phân loại nguồn gốc địa lý cam dựa trên dữ liệu phổ UV - Vis

3.3.1. Phổ UV-Vis của nước cam và tiền xử lý dữ liệu phổ

Tồn tại sự khác biệt trong các phổ UV-Vis thu được - liên quan đến sự khác nhau về hàm lượng các hợp chất hoạt tính sinh học - giữa các quần thể cam cùng giống nhưng có nguồn gốc địa lý khác nhau, không thể phân biệt trực quan nguồn gốc địa lý chỉ dựa trên các phổ UV-Vis thô.

Do đó, một bước tiền xử lý dữ liệu sử dụng tổ hợp SNV + SG + FD đã được áp dụng nhằm chuẩn hóa cường độ phổ và nâng cao độ tin cậy của phân tích phổ. Kết quả hiệu chỉnh phổ bằng các kỹ thuật tiền xử lý này đã làm nổi bật dạng “dấu vân tay” của các nhóm cinnamoyl với đỉnh đặc trưng tại vùng khoảng 330 nm.

3.3.2. Phân tích thăm dò dữ liệu bằng PCA

Phân tích PCA cho thấy dữ liệu phổ UV-Vis của các mẫu cam có xu hướng phân cụm theo nguồn gốc địa lý, tuy nhiên mức độ phân tách còn hạn chế khi sử dụng phương pháp không giám sát. Kết quả này khẳng định sự cần thiết phải áp dụng các mô hình

học giám sát để khai thác hiệu quả hơn thông tin tiềm ẩn trong dữ liệu phổ.

3.3.3. Phân loại nguồn gốc địa lý của cam bằng các kỹ thuật học giám sát

Ảnh hưởng của việc lựa chọn dải bước sóng đến khả năng nhận dạng của các mô hình phân loại

Dựa trên các giá trị chỉ số A so sánh giữa dải UV hẹp (250 - 400 nm) và toàn bộ dải phổ (250 - 1000 nm), có thể nhận thấy rằng việc sử dụng dữ liệu phổ toàn dải mang lại độ chính xác cao hơn trong nhận dạng nguồn gốc cam

So sánh hiệu năng giữa các thuật toán phân loại có giám sát

Thuật toán LDA đủ khả năng dự đoán và nhận dạng nguồn gốc địa lý của cam trong trường hợp dữ liệu có lượng mẫu hạn chế, việc tiền xử lý dữ liệu bằng tổ hợp SNV + SG + FD đã nâng cao đáng kể hiệu năng phân loại của hai mô hình học máy RF và ANN. Mặc dù các mô hình học máy chỉ đạt độ chính xác trên 80% khi sử dụng dải phổ hẹp, các giá trị này vẫn đáp ứng ngưỡng chấp nhận trong phân loại. Các mô hình có A, E, P, R, S và F1 đều lớn hơn 90% khi phân tích trên toàn dải phổ được xem là mô hình hiệu năng cao. So với các nghiên cứu đã công bố trước đây, phương pháp trong nghiên cứu này cho thấy kết quả phân loại nguồn gốc cam tương đương hoặc vượt trội, đồng thời sở hữu các ưu điểm nổi bật về chi phí thấp, tính đơn giản và tốc độ phân tích nhanh khi kết hợp quang phổ UV-Vis với học giám sát.

3.4. Phân biệt nước cam bằng dữ liệu Volt-Ampere

3.4.1. Đặc trưng điện hóa (dấu vân tay volt-ampe) của các mẫu nước cam

Thuật toán Standard Normal Variate (SNV) được sử dụng để chuẩn hóa dữ liệu, sau đó tín hiệu được lấy đạo hàm bậc hai và làm trơn bằng thuật toán Savitzky-Golay.

3.4.2. Nhận diện và xác thực nước cam

Kết quả dendrogram cho thấy, ngay cả khi áp dụng phương pháp không giám sát, vẫn có thể phân biệt được ba cụm dữ liệu chính, tương ứng với nước cam tự nhiên, nước cam thương mại và nước cam pha trộn

3.4.3. Ứng dụng học giám sát trong phát hiện pha trộn nước cam

Mô hình iPLS-LDA

Nhìn chung, các kết quả trên chỉ ra rằng giả định phân tách tuyến tính của LDA là hạn chế trong trường hợp dữ liệu có ranh giới lớp mờ và phi tuyến, đặc biệt khi các mẫu pha trộn tạo thành dải chuyển tiếp về thành phần. Đây cũng là cơ sở quan trọng để xem xét và ưu tiên các mô hình học máy phi tuyến trong các phân tích tiếp theo nhằm nâng cao hiệu năng phát hiện và phân loại pha trộn nước cam.

Mô hình iPLS-RF

Kết quả này làm nổi bật tiềm năng ưu việt của thuật toán RF, iPLS-RF thể hiện là một giải pháp mạnh mẽ và đáng tin cậy cho bài toán phát hiện và phân loại pha trộn nước cam dựa trên dữ liệu phổ điện hóa.

KẾT LUẬN

Luận án đã xây dựng thành công các quy trình định lượng các nhóm chất carotenoid, flavonoid và acid có mặt trong nước cam, các quy trình định lượng đã được áp dụng thành quy trình thường quy tại cơ sở kiểm nghiệm. Đây cũng là công trình đầu tiên phân biệt, phân loại cam Việt Nam dựa trên các dữ liệu hóa học. Các kết quả chính luận án đạt được như sau:

1. Đã phát triển phương pháp HPLC định lượng đồng thời 5 acid hữu cơ, 3 carotenoid và 4 flavonoid trong nước cam:

Năm acid hữu cơ gồm: acid oxalic, acid malic, acid ascorbic, acid citric, acid fumaric được phân tích với các điều kiện HPLC: pha động gồm hỗn hợp KH_2PO_4 nồng độ 10mM và H_3PO_4 10%, pH 2,2; tốc độ dòng 0,6ml/phút; cột sắc ký ReliantTM C18 5 μm (4,6 x 250mm); detector DAD. Phương pháp đã được xác nhận giá trị sử dụng với RSD từ 1,3% - 3,5% và độ thu hồi của năm trong khoảng 90% - 107%.

Ba carotenoid gồm: lutein, lycopene và β -caroten được phân tích với các điều kiện HPLC: pha động gồm hỗn hợp dung môi ACN: MeOH: H₂O: CH₃COOC₂H₅ = 70: 9,6: 0,4: 20 (v/v/v/v); tốc độ dòng 1,0 mL/phút; cột sắc ký: XBridge[®] C18 (4,6 x 250 mm, 5 μm); detector PDA. Phương pháp đã được xác nhận giá trị sử dụng với RSD từ 0,3% - 2,4% và độ thu hồi nằm trong khoảng 80 - 110%.

Bốn flavonoid gồm: naringin, hesperidin, neohesperidin và quercetin được phân tích với các điều kiện HPLC : pha động gồm acid phosphoric 0,1% và ACN sử dụng chương trình gradient, tốc độ dòng 1,0 mL/phút; cột sắc ký: Waters Bridge C18 (4,6mm x 150mm; 5 μm); detector PDA. Phương pháp đã được xác nhận

giá trị sử dụng với RSD từ 1,35% - 3,32% và độ thu hồi nằm trong khoảng 80 -110%.

2. Đã xây dựng được phương pháp phân loại nước cam theo giống và theo nguồn gốc địa lý dựa trên thành phần hóa học kết hợp máy học. Áp dụng các quy trình định lượng được phát triển (acid, carotenoid, flavonoid) và quy trình nội bộ của Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia (đường) để định lượng hàm lượng đường, acid hữu cơ, carotenoid và flavonoid trong 82 mẫu nước cam đại diện cho bốn giống cam được thu thập từ mười tỉnh của Việt Nam. Phân tích thống kê mô tả cho thấy có sự khác biệt có ý nghĩa về hàm lượng các chất giữa các giống và giữa các vùng địa lý, tuy nhiên hàm lượng sucrose và acid fumaric sự khác biệt không có ý nghĩa giữa các giống. Để giảm thiểu sự mất cân bằng dữ liệu, bộ dữ liệu hàm lượng được tiền xử lý với SMOOTE. Với tiêu chí phân loại theo giống, thuật toán MLP cho khả năng phân loại đạt 93,1%. Với tiêu chí phân loại theo vùng địa lý, các thuật toán SVC và RF đạt độ chính xác phân loại tuyệt đối.

3. Đã xây dựng được phương pháp xác thực nguồn gốc địa lý của nước cam dựa trên dữ liệu phổ UV - Vis kết hợp máy học. Đã xây dựng được bộ dữ liệu phổ UV - Vis cho giống cam Vinh được thu thập từ ba vùng trồng tại miền Bắc Việt Nam gồm Hòa Bình, Quảng Ninh và Tuyên Quang. Bộ dữ liệu phổ được tiền xử lý bằng chuẩn hóa standard normal variate (SNV) kết hợp làm trơn Savitzky-Golay (SG). Khi sử dụng dải phổ hẹp 250 nm - 400 nm, LDA cho khả năng phân loại 100%, khi sử dụng toàn phổ (250 - 1000 nm) cả 3 thuật toán LDA, RF, ANN cho khả năng phân loại 91,3 - 95,6%. Khi bộ dữ liệu toàn phổ được tiền xử lý

bằng SNV + SG kết hợp thêm đạo hàm bậc nhất, hiệu quả phân loại tăng lên đáng kể, ANN đạt 100% trên cả tập huấn luyện và tập kiểm tra.

4. Đã xây dựng được phương pháp phân biệt nước cam tự nhiên với nước cam hóa học dựa trên dấu vân tay điện hóa kết hợp máy học. 198 mẫu nước cam đã được xây dựng bộ dấu vân tay điện hóa. Các dữ liệu điện hóa được tiền xử lý bằng SVN kết hợp với SG, sau đó được giảm chiều bằng iPLS. Thuật toán RF cho khả năng phân loại nước cam tự nhiên bị pha trộn đạt 95%, tuy nhiên khả năng phân biệt hạn chế khi mức độ pha trộn ở mức tỉ lệ ngang nhau cỡ 45 - 50% pha trộn.

Các kết quả của luận án đã góp phần vào việc phát triển các phương pháp xác định đồng thời các chất trong nền mẫu thực phẩm. Kết quả của 03 quy trình định lượng acid hữu cơ, carotenoid và flavonoid trong nước cam đều đã được xây dựng thành quy trình thường quy tại Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia, trong đó có 02 phương pháp được công nhận phù hợp tiêu chuẩn ISO 17025:2017. Đây là nghiên cứu đầu tiên tại Việt Nam ứng dụng trí tuệ nhân tạo trong phân loại, phân biệt các sản phẩm cam Việt Nam. Đặc biệt, với bộ dữ liệu đầu vào đa dạng, có thể phù hợp với nhiều mục đích, nhiều cơ sở kiểm nghiệm tuyến cơ sở, có thể sử dụng những thiết bị đơn giản như UV mà hầu như cơ sở nào cũng được trang bị. Các kết quả này có thể triển khai được trong thực tế và mang lại những lợi ích thiết thực cho nhà sản xuất và người tiêu dùng.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

1. Nguyễn Mạnh Sơn, Nguyễn Đức Phong, Bùi Xuân Thành, Tạ Thị Thảo, Lê Thị Hồng Hào, **Nguyễn Đức Thanh** (2024), “Một số ứng dụng của học máy kết hợp với thị giác máy tính, mũi điện tử và phân tích không mục tiêu trong kiểm soát chất lượng thực phẩm”, *Kỷ yếu Hội nghị khoa học quốc tế về Kiểm nghiệm thực phẩm*, tr. 313-328.

2. Nguyen Manh Son, Nguyen Duc Phong, Nguyen Quoc Hieu, Nguyen Van Dai, Tran Khanh Du, Vu Tien Dung, Ta Thi Thao, Bui Xuan Thanh, Nguyen Trong Hieu, Vu Duong, Nguyen Thi Van Anh, **Nguyen Duc Thanh** (2024), “Predicting orange’s sweetness using deep learning coupled with computer vision”, Báo cáo poster, *Hội nghị khoa học quốc tế về Kiểm nghiệm thực phẩm*.

3. **Nguyễn Đức Thanh**, Dương Thị Dung, Nguyễn Thị Vân Anh, Vũ Thị Trang, Lê Thị Hồng Hào, Tạ Thị Thảo (2024), “Nghiên cứu định lượng đồng thời một số carotenoid trong nước cam tự nhiên”, *Tạp chí Phân tích Hóa, Lý và Sinh học* Tập 30(1), tr. 8 - 17.

4. **Nguyễn Đức Thanh**, Nguyễn Thị Quyên, Nguyễn Thị Vân Anh, Lê Việt Ngân, Lê Thị Hồng Hào, Tạ Thị Thảo (2024), “Nghiên cứu định lượng đồng thời một số flavonoid trong nước cam Việt Nam Bằng phương pháp sắc ký lỏng hiệu năng cao”, *Tạp chí Khoa học và Công nghệ Trường Đại học Công nghiệp Hà nội* Tập 60(11) , tr. 245 - 253.

5. **Nguyen Duc Thanh**, Nguyen Manh Son, Nguyen Duc Phong, Pham Huu Vang, Ninh Duc Ha, Nguyen Thi Van Anh, Le Thi Hong Hao, Nguyen Manh Ha, Nguyen Thi Kim Thuong, Ta Thi Thao (2026), “Rapid Identification of Orange Juice Adulteration Using Voltammetric Profiling and Machine Learning”, *ASEAN Journal of Scientific and Technological Report* Vol. 29(2), e260129, DOI: <https://doi.org/10.55164/ajstr.v29i2.260129>, Q4, Scopus.

6. **Duc Thanh Nguyen**, Manh Son Nguyen, Huu Vang Pham, Manh Ha Nguyen, Duc Phong Nguyen, Thi Van Anh Nguyen, Thi Hong Hao Le, Ba Lich Pham, Thi Thao Ta (2026), “Enhancing geographical origin classification of Vietnamese oranges via hyBrid preprocessing and supervised learning of UV-Vis spectral data”, *Talanta Open* (2026), DOI: <https://doi.org/10.1016/j.talo.2026.100625>, Q2, IF 3.7.