

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

LUẬN ÁN TIẾN SĨ HÓA PHÂN TÍCH

Hà Nội - 2026

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

LUẬN ÁN TIẾN SĨ HÓA PHÂN TÍCH

NGƯỜI HƯỚNG DẪN KHOA HỌC:

- PGS.TS Tạ Thị Thảo
- PGS.TS. Lê Thị Hồng Hảo

Hà Nội - 2026

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn khoa học của PGS.TS. Tạ Thị Thảo và PGS.TS. Lê Thị Hồng Hảo. Các số liệu, kết quả trình bày trong luận án là trung thực và chưa từng được công bố trong bất kỳ công trình nào khác.

Hà Nội, ngày tháng năm 2026

Nghiên cứu sinh

Nguyễn Đức Thanh

LỜI CẢM ƠN

Luận án này được hoàn thành tại các phòng thí nghiệm của Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia, Học viện Quân Y, Học viện Y Dược học cổ truyền Việt Nam, Viện Khoa học và Công nghệ Quân sự và Khoa Hóa học, Trường Đại học Khoa học Tự nhiên - Đại học Quốc gia Hà Nội.

Cảm ơn sự hỗ trợ kinh phí từ nhiệm vụ khoa học và công nghệ cấp ĐHQGHN: “Nghiên cứu xây dựng hệ thống trí tuệ nhân tạo phân loại chất lượng cam trong công đoạn sơ chế và phát triển chế phẩm bảo quản, tích hợp trên dây chuyền chế biến hướng tới xuất khẩu” mã số QG 23.68.

Để hoàn thành luận án này, tôi bày tỏ lòng biết ơn sâu sắc tới:

PGS.TS Tạ Thị Thảo, Giảng viên Bộ môn Hóa học Phân tích, Khoa Hóa học Trường Đại học Khoa học Tự nhiên - Đại học Quốc gia Hà Nội và PGS.TS Lê Thị Hồng Hảo, Viện trưởng Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia, hai người thầy đã giúp đỡ, định hướng cho tôi đến một hướng nghiên cứu rất mới, gắn liền nghiên cứu cơ bản với ứng dụng thực tiễn.

Các thầy cô Bộ môn Hóa học Phân tích, Khoa Hóa học, Trường Đại học Khoa học Tự nhiên - Đại học Quốc gia Hà Nội đã giúp đỡ, tạo điều kiện thuận lợi cho tôi trong quá trình học tập và nghiên cứu. Các anh, chị cán bộ, giảng viên, nhân viên Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia, Khoa Dược - Học viện Y Dược học cổ truyền Việt Nam, Viện Khoa học và Công nghệ Quân sự, các sinh viên, học viên đã hỗ trợ và phối hợp với tôi trong quá trình thực nghiệm.

Lãnh đạo, Ban Giám đốc Học viện Quân Y, lãnh đạo chỉ huy và tập thể Khoa Hóa - Học viện Quân Y đã tạo điều kiện giúp đỡ, động viên trong công việc để tôi có thể hoàn thành luận án.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc tới gia đình, người thân, bạn bè đã quan tâm, động viên, khích lệ và ủng hộ tôi trong suốt quá trình học tập và nghiên cứu.

Nghiên cứu sinh

Nguyễn Đức Thanh

MỤC LỤC

Trang

LỜI CAM ĐOAN

LỜI CẢM ƠN

MỤC LỤC.....	1
DANH MỤC CÁC CHỮ VIẾT TẮT	4
DANH MỤC CÁC BẢNG.....	6
DANH MỤC CÁC HÌNH.....	8
MỞ ĐẦU.....	11
Chương 1. TỔNG QUAN	13
1.1. Tổng quan về cam và các sản phẩm nước cam trên thị trường.....	13
1.1.1. Các giống cam ở Việt Nam	13
1.1.2. Thành phần hóa học có trong nước cam.....	17
1.1.3. Các sản phẩm nước cam	23
1.2. Các phương pháp xác định thành phần hóa học chính của nước cam	27
1.2.1. Một số nghiên cứu định lượng các acid hữu cơ trong nước cam	27
1.2.2. Một số nghiên cứu định lượng các carotenoid trong nước cam	28
1.2.3. Một số nghiên cứu định lượng các flavonoid trong nước cam.....	29
1.3. Học máy trong phân biệt và phân loại cam	31
1.3.1. Một số khái niệm về trí tuệ nhân tạo	31
1.3.2. Các thuật toán học máy ứng dụng trong phân loại và xác thực nước cam	31
1.3.3. Phương pháp đánh giá độ chính xác mô hình	37
1.4. Một số nghiên cứu phân biệt, phân loại cam	39
Chương 2. THỰC NGHIỆM.....	43
2.1. Hóa chất, dụng cụ và thiết bị	43
2.1.1. Hóa chất.....	43
2.1.2. Thiết bị và dụng cụ	44
2.1.3. Phần mềm xử lý số liệu	44

2.2. Đối tượng nghiên cứu	45
2.3. Phương pháp nghiên cứu	46
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	46
2.3.2. Phân biệt các sản phẩm nước cam dựa trên thành phần hóa học	51
2.3.3. Phân biệt các sản phẩm nước cam dựa trên dữ liệu phổ UV - Vis.....	52
2.3.4. Xác định sự pha trộn trong nước cam bằng phân tích Volt-Ampere và học máy	55
Chương 3. KẾT QUẢ VÀ THẢO LUẬN	57
3.1. Nghiên cứu xây dựng và thẩm định phương pháp định lượng các thành phần hóa học chính có trong nước cam	57
3.1.1. Xây dựng và thẩm định phương pháp định lượng các acid trong nước cam	57
3.1.2. Xây dựng và thẩm định phương pháp định lượng các carotenoid trong nước cam.....	64
3.1.3. Xây dựng và thẩm định phương pháp định lượng các flavonoid trong nước cam.....	76
3.2. Phân loại các sản phẩm nước cam dựa trên thành phần hóa học.....	87
3.2.1. Phân tích thống kê	87
3.2.2. Sự khác biệt về thành phần hóa học trong nước cam.....	92
3.2.3. Phân loại bằng học máy dựa trên thành phần hóa học	93
3.3. Phân loại nguồn gốc địa lý cam dựa trên dữ liệu phổ UV - Vis.....	98
3.3.1. Phổ UV-Vis của nước cam và tiền xử lý dữ liệu phổ.....	98
3.3.2. Phân tích thăm dò dữ liệu bằng PCA	102
3.3.3. Phân loại nguồn gốc địa lý của cam bằng các kỹ thuật học giám sát.....	104
3.4. Phân biệt nước cam bằng dữ liệu Volt-Ampere	110
3.4.1. Đặc trưng điện hóa (dấu vân tay volt-ampere) của các mẫu nước cam .	110
3.4.2. Nhận diện và xác thực nước cam	112
3.4.3. Ứng dụng học giám sát trong phát hiện pha trộn nước cam	114

KẾT LUẬN	121
DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN	124
TÀI LIỆU THAM KHẢO.....	126
PHỤ LỤC	

DANH MỤC CÁC CHỮ VIẾT TẮT

TT	Viết tắt	Tiếng Anh	Tiếng Việt
1	ANN	Artificial neural network	Mạng noron nhân tạo
2	ACN	Acetonitrile	Acetonitrile
3	CA	Cluster analysis	Phân tích cụm
4	DA	Discriminant analysis	Phân tích phân biệt
5	DAD	Diode array detection	Detector mảng diod
6	DT	Decision tree	Cây quyết định
7	EtOH	Ethanol	Ethanol
8	FC	From concentrate	Hoàn nguyên
9	HPLC	High performance liquid chromatography	Sắc ký lỏng hiệu năng cao
10	iPLS	Interval partial least squares	Bình phương tối thiểu từng phần theo khoảng
11	IR	Infrared spectroscopy	Phổ hồng ngoại
12	LDA	Linear discriminant analysis	Phân tích phân biệt tuyến tính
13	MC	Mean centreing	Chuẩn hóa trung bình
14	MDL	Method detection limit	Giới hạn phát hiện
15	MeOH	Methanol	Methanol
16	ML	Machine learning	Học máy
17	MLP	Multi layer perceptron	Mạng nơ ron lan truyền đa lớp
18	MLR	Multivariate linear regression	Hồi quy đa biến tuyến tính
19	MQL	Method quantification limit	Giới hạn định lượng
20	MS	Mass spectrometry	Phổ khối
21	NFC	Not from concentrate	Nguyên chất
22	NIR	Near infrared spectroscopy	Phổ hồng ngoại gần

TT	Viết tắt	Tiếng Anh	Tiếng Việt
23	PC	Principal component	Thành phần chính
24	PCA	Principal component analysis	Phân tích thành phần chính
25	PCR	Principal component regression	Hồi quy thành phần chính
26	PDA	Photo diode array	Mảng diod quang
27	PLS	Partial least squares	Bình phương tối thiểu từng phần
28	ppm	Parts per million	Phần triệu
29	R	Correlation coefficient	Hệ số tương quan
30	RF	Random forest	Rừng ngẫu nhiên
31	RMSE	Root mean square error	Căn bậc hai trung bình bình phương sai số
32	SG	Savitzky-Golay filter	Bộ lọc Savitzky-Golay
33	SNV	Standard normal transformation	Chuyển dạng thành phân bố chuẩn
34	SVN	Standard normal variate	Biến chuẩn hóa
35	SVM	Support vector machine	Máy véc tơ hỗ trợ
36	t_R	Retention time	Thời gian lưu
37	UV	Ultraviolet	Tử ngoại
38	Vis	Visible	Khả kiến
39	v/v	Volume/volume	Thể tích/thể tích

DANH MỤC CÁC BẢNG

Bảng 1.1. Một số acid chính trong nước cam	17
Bảng 1.2. Một số carotenoid chính trong nước cam	19
Bảng 1.3. Một số flavonoid chính trong nước cam.....	21
Bảng 1.4. Một số đường chính trong nước cam.....	22
Bảng 1.5. Một số nghiên cứu phân biệt và xác thực nước cam	39
Bảng 2.1. Chương trình gradient pha động định lượng flavonoid	49
Bảng 3.1. Các thông số khảo sát pha động	59
Bảng 3.2. Khoảng tuyến tính và đường chuẩn của các chất phân tích	60
Bảng 3.3. Kết quả thẩm định độ ổn định hệ thống.....	62
Bảng 3.4. Kết quả thẩm định MDL, MQL của 5 acid.....	63
Bảng 3.5. Kết quả độ lặp lại của 5 acid.....	64
Bảng 3.6. Kết quả độ thu hồi của 5 acid	64
Bảng 3.7. Các thông số khảo sát tách 3 carotenoid với các tốc độ dòng.....	67
Bảng 3.8. Khoảng tuyến tính và đường chuẩn của 3 carotenoid	68
Bảng 3.9. Kết quả thẩm định độ ổn định hệ thống.....	73
Bảng 3.10. Kết quả thẩm định MDL, MQL của 3 carotenoid	74
Bảng 3.11. Kết quả độ lặp lại của 3 carotenoid	75
Bảng 3.12. Kết quả độ thu hồi của 3 carotenoid	75
Bảng 3.13. Các thông số khảo sát tách 4 falvonoid với các tốc độ dòng.....	79
Bảng 3.14. Đường chuẩn của các chất phân tích	80
Bảng 3.15. Kết quả độ ổn định hệ thống.....	84
Bảng 3.16. Kết quả thẩm định MDL, MQL của 4 flavonoid.....	85
Bảng 3.17. Kết quả độ lặp lại của 4 flavonoid.....	86
Bảng 3.18. Kết quả độ thu hồi của 4 flavonoid.....	86
Bảng 3.19. Hàm lượng các hợp chất chính trong nước cam tươi ở các giống và được trồng ở các tỉnh khác nhau	88
Bảng 3.20. Kết quả ANOVA thành phần hóa học	92
Bảng 3.21. Kết quả phân loại theo giống cam.	95

Bảng 3.22. Kết quả phân loại cam theo nguồn gốc địa lý.....	97
Bảng 3.23. Hiệu suất phân loại của các mô hình (%) LDA, RF, ANN	105
Bảng 3.24. Phần trăm mỗi loại mẫu trong nhóm phân loại	114

DANH MỤC CÁC HÌNH

Hình 1.1. Cây cam.....	13
Hình 1.2. Cam Xã Đoài.....	14
Hình 1.3. Cam Sành	14
Hình 1.4. Cam Vinh	15
Hình 1.5. Cam Cao Phong.....	15
Hình 1.6. Cam Soàn	16
Hình 1.7. Các sản phẩm nước cam tự nhiên	25
Hình 1.8. Các sản phẩm đồ uống từ cam	26
Hình 1.9. Mối quan hệ của trí tuệ nhân tạo, học máy và học sâu	31
Hình 1.10. Mô hình thuật toán cây quyết định.....	33
Hình 1.11. Mô hình thuật toán SVM	35
Hình 1.12. Mô hình thuật toán MLP	35
Hình 1.13. Mô hình thuật toán downsampling.....	36
Hình 1.14. Mô hình thuật toán SMOTE.....	37
Hình 1.15. Ma trận nhầm lẫn của phép phân loại ba lớp	38
Hình 2.1. Sơ đồ xử lý mẫu định lượng acid hữu cơ.....	46
Hình 2.2. Dự kiến sơ đồ xử lý mẫu định lượng carotenoid	48
Hình 2.3. Dự kiến sơ đồ xử lý mẫu định lượng flavonoid	50
Hình 3.1. Sắc ký đồ 5 acid với dung môi pha động có pH khác nhau	57
Hình 3.2. Sắc ký đồ 5 acid trên 2 cột khảo sát.....	58
Hình 3.3. Sắc ký đồ 5 acid với các tốc độ dòng khảo sát.....	59
Hình 3.4. Đường chuẩn của 5 acid nghiên cứu	61
Hình 3.5. Sắc ký đồ khảo sát độ đặc hiệu	63
Hình 3.6. Phổ UV-Vis của lutein (a), lycopene (b), β -caroten (c).....	65
Hình 3.7. Sắc ký đồ 3 carotenoid với 4 cột pha tĩnh.....	66
Hình 3.8. Sắc ký đồ 3 carotenoid với các hệ pha động.....	67
Hình 3.9. Đường chuẩn của 3 carotenoid.....	69
Hình 3.10. Kết quả khảo sát ảnh hưởng của dung môi chiết mẫu	69

Hình 3.11. Sắc ký đồ khảo sát dung môi chiết.....	70
Hình 3.12. Kết quả khảo sát dung môi hòa cần	71
Hình 3.13. Sắc ký đồ dung môi hòa cần	71
Hình 3.14. Kết quả khảo sát số lần chiết lặp lại.....	72
Hình 3.15. Quy trình xử lý mẫu	73
Hình 3.16. Sắc ký đồ khảo sát độ đặc hiệu	74
Hình 3.17. Phổ hấp thụ UV-Vis của các flavonoid.....	76
Hình 3.18. Sắc ký đồ hỗn hợp chuẩn của 3 cột khảo sát.....	77
Hình 3.19. Sắc ký đồ khảo sát pha động.....	78
Hình 3.20. Sắc ký đồ khảo sát tốc độ dòng tách 4 flavonoid.....	79
Hình 3.21. Đường chuẩn của 4 flavonoid trong nước cam.....	81
Hình 3.22. Kết quả khảo sát dung môi hòa tan	82
Hình 3.23. Sắc ký đồ mẫu nước cam được hòa tan trong 3 dung môi.....	82
Hình 3.24. Kết quả khảo sát kỹ thuật hòa tan	83
Hình 3.25. Kết quả khảo sát thời gian hòa tan	83
Hình 3.26. Quy trình xử lý mẫu	84
Hình 3.27. Sắc ký đồ khảo sát độ đặc hiệu phân tích flavonoid	85
Hình 3.28. Sự phân bố hàm lượng carotenoid trong nước cam	89
Hình 3.29. Sự phân bố hàm lượng đường trong nước cam.....	90
Hình 3.30. Sự phân bố hàm lượng acid trong nước cam	91
Hình 3.31. Sự phân bố hàm lượng flavonoid trong nước cam.....	91
Hình 3.32. Kết quả biểu đồ PCA biplot dựa trên tiêu chí phân loại theo giống	94
Hình 3.33. Kết quả biểu đồ PCA biplot dựa trên tiêu chí phân loại theo nguồn gốc địa lý.....	96
Hình 3.34. Phổ UV-Vis của các mẫu cam ở 3 tỉnh Việt Nam	101
Hình 3.35. Phổ UV - Vis của các mẫu nghiên cứu sau khi xử lý bằng SNV+SG+FD.	102
Hình 3.36. Biểu đồ PCA của phổ UV - Vis các mẫu nước cam với 2 PC đầu tiên	103
Hình 3.37. Ma trận nhầm lẫn của các mô hình	108

Hình 3.38. Volt-ampere đồ của nước cam	111
Hình 3.39. Volt-ampere đồ của nước cam sau khi chuẩn hóa bằng SNV, đạo hàm bậc hai và làm mượt với Savitzky-Golay	112
Hình. 3.40. Phân cụm phân cấp sử dụng tương đồng cosin.....	113
Hình 3.41. Lựa chọn các đặc trưng bằng iPLS	115
Hình 3.42. Biểu đồ LDA của tập huấn luyện và tập kiểm tra được chiếu lên LDA	116
Hình 3.43. Ma trận nhầm lẫn mô hình iPLS	118
Hình 3.44. Mối tương quan giữa tín hiệu đạo hàm bậc hai và các đặc trưng của mẫu trong mô hình iPLS-LDA.....	119

MỞ ĐẦU

Gian lận thực phẩm là một vấn đề mang tính toàn cầu, ảnh hưởng trực tiếp đến quyền lợi người tiêu dùng, tính minh bạch của thị trường và hiệu quả quản lý chất lượng sản phẩm. Trong bối cảnh đó, nước ép trái cây được xếp vào nhóm thực phẩm có nguy cơ gian lận cao do mức tiêu thụ lớn, giá trị thương mại đáng kể và khả năng bị pha trộn, thay thế hoặc mô phỏng thành phần tương đối dễ dàng. Các hình thức gian lận thường gặp như: pha loãng bằng nước, bổ sung đường, acid, hương liệu tổng hợp hoặc pha trộn với các loại nước ép khác có giá thành thấp hơn nhưng đặc tính cảm quan tương tự [106, 132].

Về bản chất hóa học, nước cam tự nhiên là một hệ phức hợp chứa nhiều nhóm hợp chất, trong đó quan trọng nhất là các acid hữu cơ (acid citric, malic, ascorbic...), các flavonoid đặc trưng của họ *Citrus* (hesperidin, naringin...) và các carotenoid (β -carotene, lutein...). Các hợp chất này không chỉ quyết định các đặc tính cảm quan nguyên bản như màu sắc, mùi vị mà còn phản ánh đặc trưng của giống cam, điều kiện khí hậu, thổ nhưỡng và kỹ thuật canh tác [58, 64]. Nhiều nghiên cứu đã chỉ ra rằng, thành phần hóa học của nước cam tự nhiên biến thiên rõ rệt theo giống và vùng địa lý; trong khi đó, nước cam nhân tạo hoặc nước cam pha trộn - dù được sản xuất thông qua các quy trình công nghiệp như hoàn nguyên từ nước cam cô đặc, bổ sung phụ gia nhằm tái tạo đặc tính cảm quan - vẫn không thể bảo toàn đầy đủ cấu trúc thành phần hóa học tự nhiên [40, 55]. Sự khác biệt có hệ thống này chính là cơ sở khoa học cốt lõi cho các nghiên cứu phát hiện gian lận và phân loại nước cam.

Trước những hạn chế về yêu cầu công bố các thành phần đơn lẻ, xu hướng hiện đại trong kiểm nghiệm thực phẩm là kết hợp giữa các dữ liệu hóa học và học máy. Các kỹ thuật phân tích như HPLC, GC, GC-MS cho phép định lượng và định tính chính xác các hợp chất đặc trưng trong nước cam, tạo ra bộ dữ liệu hóa học có độ tin cậy cao [42]. Trên cơ sở đó, các phương pháp chemometrics và học máy đã được áp dụng để khai thác mối quan hệ đa biến giữa hàm lượng các chất hoặc tín hiệu đo trên thiết bị phân tích và thông tin nguồn gốc mẫu, phục vụ phân loại và xác thực nước cam [57, 91]. Song song với dữ liệu thành phần hóa học, các dữ liệu tín hiệu phân

tích tổng thể như phổ UV-Vis, phổ hồng ngoại gần (NIR), phổ khối và tín hiệu điện hóa ngày càng được sử dụng rộng rãi trong các nghiên cứu xác thực theo hướng không mục tiêu. Các nghiên cứu gần đây cho thấy dữ liệu phổ và dấu vân tay điện hóa kết hợp với chemometrics cho phép phát hiện nước cam bị pha trộn với độ chính xác cao, đồng thời phù hợp cho sàng lọc nhanh trong kiểm nghiệm thực tế [95, 112].

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. Cho đến nay 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 tổng quát là xây dựng được phương pháp phân biệt nước cam theo vùng địa lý, theo giống cam, nước cam tự nhiên với nước cam hóa học, với các mục tiêu cụ thể như sau:

1. Phát triển phương pháp định lượng một số chất chính thuộc các nhóm chất flavonoid, carotenoid, acid hữu cơ trong các mẫu nước cam.
2. 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.
3. 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.
4. 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.

Chương 1. TỔNG QUAN

1.1. Tổng quan về cam và các sản phẩm nước cam trên thị trường

1.1.1. Các giống cam ở Việt Nam

Cam là cây ăn quả lâu năm, cùng họ với bưởi, có quả nhỏ hơn quả bưởi, vỏ mỏng, khi chín thường có màu da cam hoặc xanh bóng, sinh trưởng rất tốt ở điều kiện khí hậu nhiệt đới hoặc cận nhiệt đới để cho quả ngọt. Cam là một cây lai được trồng từ xa xưa, có thể lai giống giữa loài bưởi (*Citrus maxima*) và quýt (*Citrus reticulata*) [19]. Đây là cây nhỏ, cao đến khoảng 10m, có cành gai và lá thường xanh dài khoảng 4 - 10 cm, vị có thể biến đổi từ ngọt đến chua.



Hình 1.1. Cây cam

Việt Nam có thể được coi là một trong những trung tâm đa dạng cao với nhiều giống cam bản địa quý. Ở nước ta các giống cam chủ yếu được đặt tên theo vùng, miền trồng như cam Xã Đoài, cam Cao Phong, cam Vinh, cam Vân Du, cam Sông Con,... hoặc được gọi tên theo đặc điểm từng loại như cam Sành, cam Mật,... Tùy thuộc vào điều kiện địa lý, khí hậu của từng nơi để phát triển mỗi giống cam khác nhau. Một số giống cam được trồng phổ biến ở nước ta gồm [4, 8]:

❖ Cam Xã Đoài

Là giống cam có nguồn gốc ở vùng Nghi Lộc - Nghệ An, có lá màu xanh đậm, hình thái lá thuôn dài, cành có gai, lá đứng, eo lá rộng. Cam Xã Đoài thích ứng rộng, có 2 dạng quả: dạng có quả tròn và dạng có quả tròn dài. Dạng có quả tròn dài cho

năng suất cao hơn, trọng lượng quả trung bình 180 - 200g, hương vị thơm ngon nhưng có nhược điểm là nhiều hạt, xơ bã nhiều.



Hình 1.2. Cam Xã Đoài

❖ *Cam Sành*

Cam Sành là một giống cây ăn quả thuộc chi cam chanh, có nguồn gốc từ Việt Nam. Cam Sành được gán nhiều tên khoa học khác nhau như *Citrus nobilis*, *Citrus reticulata*, hay *Citrus sinensis*, trên thực tế nó là giống lai tự nhiên giữa *C. reiculata* và *C. sinensis*. Quả cam Sành rất dễ nhận ra nhờ lớp vỏ dày, sần sùi giống bề mặt mảnh sành, các múi thịt có màu cam, nhiều nước, hương vị chua ngọt. Chu kỳ khai thác 10 - 15 năm. Cam Sành nước ta hiện trồng chủ yếu ở khu vực Đồng bằng sông Cửu Long như: Vĩnh Long, Tiền Giang, Cần Thơ,...



Hình 1.3. Cam Sành

❖ *Cam Vinh*

Cam Vinh được trồng trên chất đất đỏ của vùng miền Tây Nghệ An, thuộc chi cam chanh, quả tròn đều, nhỏ, nhiều nước, ít hạt, màu xanh vàng đều, vỏ mỏng, thường bị nám, thơm có vị ngọt thanh, đậm đà. Màu vàng của cam Vinh là màu vàng tươi chanh pha với màu xanh, chứ không phải màu vàng da cam. Kể cả phần tép cam cũng vàng nhẹ không phải màu vàng cam.



Hình 1.4. Cam Vinh

❖ *Cam Cao Phong*

Loại cam nổi tiếng ở thị trấn Cao Phong - Hòa Bình. Cam Cao Phong bao gồm 4 giống: cam CS1 (cam lòng vàng), cam Xã Đoài lùn, cam Xã Đoài cao và cam Canh. Trong đó, giống cam Xã Đoài được di thực và trồng tại Cao Phong. Cam Cao Phong nổi tiếng bởi vị ngọt thanh, thơm đặc trưng, vỏ mỏng, màu vàng xanh. Cam Cao Phong bắt đầu thu hoạch vào giữ tháng 10.



Hình 1.5. Cam Cao Phong

❖ *Cam Soàn*

Cam Soàn là giống được trồng lâu đời ở một số tỉnh đồng bằng sông Cửu Long. Cam Soàn có đặc điểm là trái cam càng nhỏ thì vị càng ngọt thanh, có mùi thơm nhẹ. Tuy nhiên, dựa vào từng vùng đất và từng độ tuổi mà cho những cây cam Soàn có độ ngon ngọt khác nhau.



Hình 1.6. Cam Soàn

Bộ phận dùng quan trọng nhất của loài cam là quả cam. Nước cam có giá trị dinh dưỡng cao dùng để bồi bổ và làm mát cơ thể. Ngoài ra vỏ quả, hoa và lá cam đều có chứa lượng tinh dầu lớn nên để cất tinh dầu dùng để làm thơm, xua đuổi côn trùng... [5]. Cam được sử dụng nhiều do giá trị dinh dưỡng cao và nhiều hoạt tính sinh học quý [9]. Trong nước cam có thành phần acid ascorbic (vitamin C) tăng tổng hợp collagen giúp tránh các bệnh như: bệnh còi, gây rụng răng, chảy máu chân răng, chảy máu da và lở loét [2]. Vitamin C trong cam có tác dụng chống ung thư do nó có thể trung hòa các gốc tự và vô hiệu hóa các gốc tự do này trong cơ thể [2]. Carotenoid, đặc biệt là β -caroten có thể chuyển đổi thành vitamin A trong cơ thể. Vitamin A đóng vai trò quan trọng trong việc duy trì sức khỏe thị lực, đặc biệt là trong điều kiện ánh sáng yếu [37]. Hoạt chất flavonoid chính trong nước cam là hesperidin và naringin có tác dụng ảnh hưởng đến một số con đường trao đổi chất giúp cải thiện khả năng chống oxy hóa trong huyết thanh và hiệu quả chống viêm. Ngoài ra, nước cam cung cấp đủ chất dinh dưỡng, giúp no lâu làm duy trì cân nặng phù hợp [43].

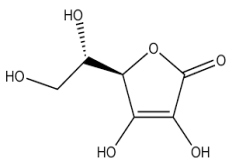
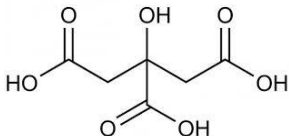
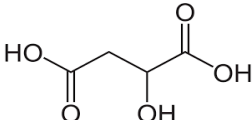
1.1.2. Thành phần hóa học có trong nước cam

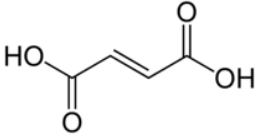
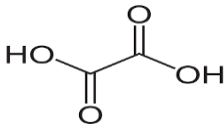
Thành phần hóa học của nước cam rất phong phú bao gồm: Các loại đường, acid hữu cơ, flavanoid và đặc biệt tất cả các loại trái cây họ cam quýt đều chứa nhiều chất chống oxy hóa carotenoid, đây cũng là nguyên nhân tạo ra màu cam, đỏ và vàng phong phú của chúng [16, 59, 60]. Ngoài ra còn có các nguyên tố vi lượng như sắt, photpho, kẽm ...

1.1.2.1. Một số acid hữu cơ chính trong nước cam

Nước cam chứa nhiều acid hữu cơ tuy nhiên chỉ có 5 acid hữu cơ là acid oxalic, acid ascorbic, acid malic, acid citric, acid fumaric đóng vai trò quan trọng nhất vì chúng đem lại nhiều lợi ích đến sức khỏe con người [108]. Một số acid hữu cơ chính trong nước cam được trình bày trong Bảng 1.1.

Bảng 1.1. Một số acid chính trong nước cam

Acid	Công thức cấu tạo	Lý tính	Tác dụng sinh học
Ascorbic $C_6H_8O_6$		Tinh thể không màu hay bột kết tinh trắng hoặc gần như trắng. $t_{\text{nóng chảy}}: 190^{\circ}\text{C}$	- Tổng hợp collagen, norepinephrin [96, 62] - Hoạt hóa hormon [62] - Đóng vai trò như một chất chống oxy- hóa khử [46]
Citric $C_6H_8O_7$		Bột kết tinh trắng hoặc tinh thể không màu $t_{\text{nóng chảy}}: 153^{\circ}\text{C}$	- Giảm quá trình viêm [123] - Chống oxy hóa [17]
Malic $C_4H_6O_5$		Bột màu trắng $t_{\text{nóng chảy}}: 130^{\circ}\text{C}$	Tăng tổng hợp glycosaminoglycans và làm dày da [106]

<i>Acid</i>	<i>Công thức cấu tạo</i>	<i>Lý tính</i>	<i>Tác dụng sinh học</i>
Fumaric $C_4H_4O_4$		Chất rắn kết tinh không màu $t_{\text{nóng chảy}}: 287^{\circ}\text{C}$	- Chống bệnh vẩy nến [19] - Giảm quá trình viêm bạch cầu ái toan [48]
Oxalic $C_2H_2O_4$		Chất rắn kết tinh không màu $t_{\text{nóng chảy}}: 101^{\circ}\text{C}$	Điều trị các tổn thương không sâu của ngà răng [142]

Các công bố đã khẳng định trong nước cam tự nhiên có chứa các acid hữu cơ với ba acid chính: acid citric, acid malic và acid ascorbic (vitamin C) [75, 108]. Bên cạnh đó, nước cam cũng chứa các acid thứ cấp ở dạng vết như acid oxalic, fumaric và succinic [75]. Trong đó, acid citric đóng vai trò là thành phần chủ đạo, quyết định trực tiếp đến độ chua và tính tươi mát đặc trưng của nước cam. Hàm lượng của các acid này có sự biến thiên nhất định phụ thuộc vào giống và điều kiện thổ nhưỡng. Acid citric chiếm ưu thế tuyệt đối với khoảng hàm lượng thường dao động từ 6,05 g/L đến 73,93 g/L đối với các dòng cam quýt nói chung, trong đó trung bình đạt khoảng 13,28 g/L đối với các giống cam ngọt [75, 108]. Đứng thứ hai là acid malic, góp phần tạo nên vị chua dịu và hương vị trái cây, với hàm lượng phân bố trong khoảng từ 1,27 g/L đến 12,15 g/L [75]. Mặc dù chiếm tỷ lệ nhỏ hơn, acid ascorbic lại có tác dụng chống oxy hóa mạnh, mức nồng độ từ 0,215 g/L đến 0,718 g/L [108]. Sự ổn định tương đối về tỷ lệ và khoảng phân bố nồng độ của các acid hữu cơ chủ đạo này được xem như một "dấu vân tay" hóa học quan trọng, cung cấp cơ sở dữ liệu tin cậy phục vụ công tác xác thực giống và phát hiện gian lận pha trộn trong nước cam thương mại [137].

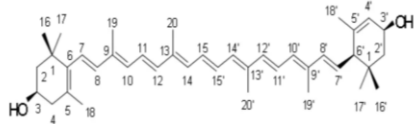
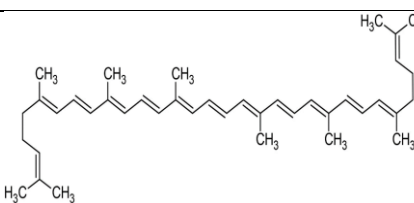
1.1.2.2. Một số carotenoid chính trong nước cam

Cho đến nay, các carotenoid được biết đến như là một nhóm chất có vai trò trong việc phòng tránh một số bệnh như thoái hóa điểm vàng, viêm họng, đục thủy tinh thể, ung thư cổ tử cung, nhiễm khuẩn, các bệnh về tim mạch, ung thư phổi, viêm phổi,

ung thư thanh quản, viêm khớp, ung thư tuyến tiền liệt, ung thư da, AIDS [101, 140]...

Một số carotenoid chính trong nước cam được trình bày trong Bảng 1.2.

Bảng 1.2. Một số carotenoid chính trong nước cam

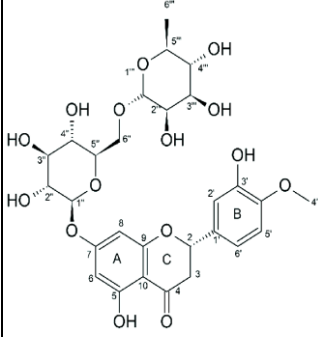
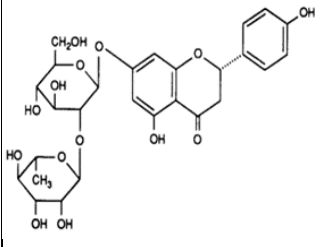
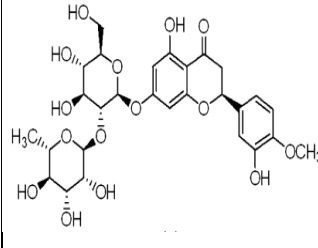
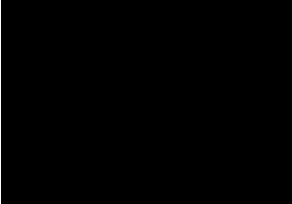
Chất	Công thức cấu tạo	Lý tính	Tác dụng sinh học
Lutein $C_{40}H_{56}O_2$		Bột vô định hình hoặc tinh thể màu đỏ cam $t_{\text{nóng chảy}}: 190^{\circ}\text{C}$	- Ngăn ngừa ung thư da, thoái hóa điểm vàng [1] - Hạn chế sự suy giảm khả năng miễn dịch [35]
Lycopene $C_{40}H_{56}$		Bột hoặc tinh thể hình kim màu đỏ $t_{\text{nóng chảy}}: 167 - 168^{\circ}\text{C}$	- Kìm hãm sự phát triển của một số căn bệnh ung thư và nhồi máu cơ tim [6, 33] - Có tác dụng như một chất chống oxy hóa [33]
β -caroten $C_{40}H_{56}$		Bột màu đỏ tía $t_{\text{nóng chảy}}: 180 - 182^{\circ}\text{C}$	- Phòng tránh được tình trạng thiếu vitamin A [7] - Chống oxy hóa và tăng cường miễn dịch [3, 93] - Phòng tránh được các vấn đề về trí nhớ như Alzheimer và các bệnh mất trí nhớ khác [47].

Các nghiên cứu đã chỉ ra rằng, hệ carotenoid trong nước cam là một hỗn hợp phức tạp, trong đó nổi bật lên vai trò cốt lõi của các hợp chất chủ yếu bao gồm β -caroten, lutein và lycopene [82]. Sự hiện diện và tỷ lệ phân bố của các hợp chất này không chỉ quyết định dải màu sắc đa dạng từ vàng tươi, cam đến đỏ rực của sản phẩm, mà còn mang lại giá trị sinh học cao. Hàm lượng của các chất này có sự phân hóa sâu sắc và mang tính đặc thù sinh học rất cao. Đối với các giống cam ngọt truyền thống (*Citrus sinensis*), lutein và β -carotene là những thành phần nền tảng. Lutein thường duy trì ở khoảng nồng độ đáng kể, dao động từ 0,1 đến 0,3 mg/L [36, 50], đóng góp quan trọng vào sắc vàng tươi của nước ép. Trong khi đó, β -carotene chỉ chiếm một tỷ lệ khiêm tốn, tồn tại ở mức nồng độ dạng vết, thường nằm trong khoảng từ 0,02 đến 0,08 mg/L [50]. Ở các giống cam ngọt thông thường, lycopene hầu như không được phát hiện hoặc chỉ tồn tại ở mức vết [50]. Lycopene có thể chiếm tỷ lệ đáng kể trong các giống cam đột biến, điển hình là giống *Cara Cara* góp phần tạo nên màu đỏ cam đến đỏ ruby của thịt quả [50, 113].

1.1.2.3. Một số flavonoid chính trong nước cam

Các flavonoid là hợp chất chính và có nhiều trong vỏ quả hay nước quả của các loài *Citrus* [11]. Trong đó, hesperidin được xác định là flavonoid đặc trưng cho họ cam quýt. Ngoài ra một số flavonoid khác cũng được tìm thấy trong cam như naringin, narirutin, neohesperidin. Ở một số nghiên cứu khác, quercetin cũng được xác định là có mặt trong nước cam dưới dạng rutin mặc dù đây là flavonoid đặc trưng của loài bưởi [135]. Một số carotenoid chính trong nước cam được trình bày trong Bảng 1.3

Bảng 1.3. Một số flavonoid chính trong nước cam

Chất	Công thức cấu tạo	Lý tính	Tác dụng sinh học
Hesperidin $C_{28}H_{34}O_{15}$		Tinh thể màu vàng $t_{\text{nóng chảy}}: 262^{\circ}\text{C}$	- Giảm phù nề sưng tấy [117] - Chống viêm [39] - Hạ huyết áp [66] - Bảo vệ dạ dày do rượu [34]
Naringin $C_{27}H_{32}O_{14}$		$t_{\text{nóng chảy}}: 166^{\circ}\text{C}$	- Tác dụng làm giảm mỡ máu, hạ huyết áp, bệnh tiểu đường, chống viêm [111]
Neohesperidin $C_{28}H_{23}O_{15}$		$t_{\text{nóng chảy}}: 156 - 158^{\circ}\text{C}$	- Chống viêm [12]
Quercetin $C_{15}H_{10}O_7$		Tinh thể hoặc bột màu vàng $t_{\text{nóng chảy}}: 316^{\circ}\text{C}$	- Tác dụng kháng virus [148] - Tác dụng đối với béo phì và bệnh tiểu đường [110]

Flavonoid trong nước cam tự nhiên chủ yếu tồn tại dưới dạng các glycosid của flavanon và flavonol. Trong đó, bốn hợp chất hesperidin, naringin, quercetin và neohesperidin được xem là cơ sở dữ liệu quan trọng nhất để xây dựng các mô hình xác thực nguồn gốc và kiểm soát gian lận thực phẩm [53, 111]. Hesperidin là flavonoid đặc trưng và chiếm hàm lượng cao nhất trong nước cam, thường dao động trong khoảng từ 200 đến 600 mg/L [53, 44]. Nồng độ hesperidin phụ thuộc chặt chẽ

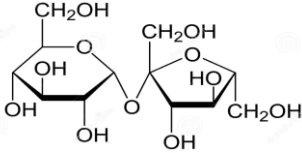
vào kỹ thuật chiết tách và giống cam [44]. Trong các giống cam ngọt truyền thống naringin và neohesperidin chỉ tồn tại ở nồng độ dạng vết ($< 5 \text{ mg/L}$) hoặc hoàn toàn vắng mặt [111]. Tuy nhiên, naringin lại là thành phần chủ đạo tạo vị đắng trong bưởi (*C. paradisi*), và neohesperidin chiếm ưu thế trong cam đắng (*C. aurantium*) [111, 72]. Quercetin đại diện cho nhóm flavonol, dù chỉ hiện diện ở hàm lượng thấp (thường $< 10 \text{ mg/L}$) nhưng lại mang giá trị cao trong việc đánh giá hoạt tính chống oxy hóa và độ tươi của sản phẩm [72]. Tỷ lệ tương quan giữa quercetin và nhóm flavonoid chính giúp làm giàu thêm bộ dữ liệu đa biến, tăng cường độ chính xác cho các thuật toán học máy trong việc phân biệt giữa nước cam tự nhiên và các sản phẩm hóa học. Việc sử dụng dữ liệu định lượng của bốn hợp chất này vào các mô hình máy học (như PCA, SVM hay ANN) cho phép thiết lập những "dấu vân tay" hóa học (chemical fingerprints), giúp phân loại các giống cam và truy xuất nguồn gốc địa lý của sản phẩm nước cam trên thị trường [53, 111].

1.1.2.4. Một số đường chính trong nước cam

Đường là nhóm hợp chất chiếm tỷ lệ lớn nhất trong nước cam, đóng vai trò quyết định đến vị ngọt của sản phẩm, chủ yếu bao gồm các monosaccharid và disaccharid, trong đó ba đường chính là glucose, fructose và sucrose. Ngoài ra, có thể có một số oligosaccharide và polysaccharide với hàm lượng thấp [128]. Tổng hàm lượng đường trong nước cam thường dao động trong khoảng $80 - 120 \text{ g/L}$ [121]. Một số đường chính trong nước cam được trình bày trong Bảng 1.4.

Bảng 1.4. Một số đường chính trong nước cam

Chất	Công thức cấu tạo	Lý tính	Tác dụng sinh học
Glucose $\text{C}_6\text{H}_{12}\text{O}_6$	$\text{CH}_2\text{OH}-(\text{CHOH})_4-\text{CHO}$	Tinh thể không màu; $t_{\text{nóng chảy}}: 146$ -150°C	- Nguồn năng lượng chính của tế bào; tham gia chu trình Krebs [67] - Duy trì chuyển hóa năng lượng trong cơ thể [68].

<i>Chất</i>	<i>Công thức cấu tạo</i>	<i>Lý tính</i>	<i>Tác dụng sinh học</i>
Fructose $C_6H_{12}O_6$	$CH_2OH-CO-(CHOH)_3-CH_2OH$	Tinh thể không màu; $t_{\text{nóng chảy}}: 103-105\text{ }^{\circ}\text{C}$	- Tham gia chuyển hóa carbohydrat tại gan [98]. - Tiêu thụ cao liên quan đến rối loạn chuyển hóa, kháng insulin và bệnh gan nhiễm mỡ [138].
Sucrose $C_{12}H_{22}O_{11}$		Tinh thể trắng; $t_{\text{nóng chảy}}: 185 - 189\text{ }^{\circ}\text{C}$	- Bị thủy phân thành glucose và fructose [67]. - Tiêu thụ quá mức liên quan đến béo phì, đái tháo đường và sâu răng [122, 138].

Trong nước cam, sucrose thường chiếm ưu thế, trong khi fructose và glucose có hàm lượng thấp hơn; tuy nhiên, nồng độ của ba đường này biến thiên đáng kể tùy theo giống, điều kiện canh tác và độ chín của quả [83, 103]. Các nghiên cứu chỉ ra rằng hàm lượng sucrose thường nằm trong khoảng 30 - 60 g/L, còn fructose và glucose lần lượt dao động khoảng 15 - 30 g/L và 15 - 25 g/L, với tổng chất rắn hòa tan ($^{\circ}\text{Brix}$) thường ở mức 10 - 14 $^{\circ}$ [83, 103]. Sự gia tăng bất thường của sucrose có thể liên quan đến việc bổ sung đường mía, hoặc sự sai lệch đáng kể của tỷ lệ fructose/glucose có thể phản ánh việc sử dụng siro ngô giàu fructose [111].

1.1.3. Các sản phẩm nước cam

Hiện nay, trên thị trường hầu như không có tiêu chuẩn về việc ghi nhãn các sản phẩm nước cam. Nước cam thương mại thường được phân thành hai nhóm chính là nước cam ép trực tiếp (Not From Concentrate - NFC) và nước cam hoàn nguyên từ dịch cô đặc (From Concentrate - FC), khác nhau chủ yếu ở quy trình chế biến và mức độ xử lý nguyên liệu. NFC là loại nước cam được ép trực tiếp từ quả tươi và có thể trải qua một số bước xử lý cơ học như lọc thô, loại bỏ bã hoặc đồng hóa nhằm ổn định cấu trúc sản phẩm. Do nước cam có thể chứa vi sinh vật từ nguyên liệu và trong

quá trình chế biến, sản phẩm thường được áp dụng các biện pháp xử lý thích hợp như thanh trùng hoặc các công nghệ hiện đại như xử lý áp suất cao (High Pressure Processing - HPP) nhằm tiêu diệt hoặc làm giảm số lượng vi sinh vật gây bệnh xuống mức an toàn theo quy định. Sau khi xử lý, sản phẩm được bảo quản trong điều kiện thích hợp, thường là bảo quản lạnh, để duy trì chất lượng và kéo dài thời gian sử dụng [45, 54, 78]. Các phương pháp xử lý khác nhau có thể ảnh hưởng đáng kể đến thành phần hóa học và đặc tính cảm quan của nước cam. Các quá trình xử lý nhiệt như thanh trùng hoặc tiệt trùng thường được thực hiện ở nhiệt độ cao, có thể gây biến đổi hoặc làm suy giảm các hợp chất bay hơi nhạy nhiệt, những thành phần quan trọng tạo nên hương thơm đặc trưng của nước cam, từ đó làm giảm mức độ tự nhiên của hương vị sản phẩm. Ngược lại, công nghệ xử lý áp suất cao HPP được thực hiện ở nhiệt độ thấp hơn nên hạn chế sự biến đổi của các hợp chất tạo hương và các thành phần dinh dưỡng. Sự khác biệt này được nhận diện bằng dữ liệu các chất chuyển hóa dựa trên các kỹ thuật phân tích kết hợp các thuật toán thay vì chỉ sử dụng các thông số hóa học truyền thống như đánh giá cảm quan, pH, chất khô hoà tan tổng số (°Brix)... [32, 78, 81]. FC là loại nước cam được ép từ cam tươi, cô đặc, sau đó hoàn nguyên trước khi đóng gói. Quá trình cô đặc - hoàn nguyên và xử lý nhiệt kèm theo có thể làm thay đổi nhiều chất nhạy cảm nhiệt, đặc biệt là các chất bay hơi. Vì vậy, cần thiết phải xây dựng được các phương pháp phân biệt NFC và FC phục vụ cho nghiên cứu và giám sát thị trường. Phương pháp phân tích không mục tiêu các chất chuyển hóa đã xác định được các marker tiềm năng giúp phân biệt NFC và FC [84]. NIR kết hợp các thuật toán học máy được sử dụng để xác thực các mẫu nước cam và phát hiện sự hiện diện của các chất pha trộn, phù hợp cho sàng lọc hiện trường [102]. Việc phân biệt các phương pháp xử lý mẫu nước cam (nhiệt độ cao và không dùng nhiệt) bằng dấu vân tay các chất bay hơi cũng đã được chứng minh hiệu quả bằng GC-MS kết hợp phân tích đa biến, kết quả cho thấy có sự thay đổi đáng kể ở các nhóm terpenoid, ester, aldehyd... [131, 132].



Hình 1.7. Các sản phẩm nước cam tự nhiên

Trên thị trường, không phải tất cả các sản phẩm mang tên đồ uống từ cam đều là nước ép cam nguyên chất; nhiều sản phẩm được pha chế từ nước, đường, acid hữu cơ, hương liệu, chất điều chỉnh độ đục, vitamin và đôi khi bổ sung dịch quả khác. Các sản phẩm này thường được xếp vào nhóm đồ uống hương cam hoặc đồ uống pha chế từ nước cam, trong đó hàm lượng nước cam tự nhiên có thể khác nhau tùy theo công thức sản phẩm. Do thành phần phức tạp và có thể được điều chỉnh để đạt các chỉ tiêu cảm quan hoặc hóa lý tương tự nước cam tự nhiên, nhóm sản phẩm này tiềm ẩn nguy cơ gian lận như mạo danh nước cam 100% hoặc pha trộn nhưng không ghi nhãn đầy đủ. Vì vậy, việc xác thực nguồn gốc và thành phần của các sản phẩm này cần dựa trên các phương pháp phân tích tiên tiến, bao gồm phân tích đa biến và các phương pháp phân tích không mục tiêu, thay vì chỉ dựa vào các chỉ tiêu hóa học truyền thống [66]. Các nghiên cứu gần đây cho thấy mặc dù các đồ uống từ cam có thể được điều chỉnh để đạt các chỉ tiêu hóa lý cơ bản tương tự nước cam tự nhiên (pH, hàm lượng chất khô hoà tan tổng số, chất lượng cảm quan về màu sắc, mùi, vị, và trạng thái) nhưng thành phần các hợp chất bay hơi của chúng vẫn có sự khác biệt đáng kể. Khi kết hợp phân tích đa biến với các thành phần bay hơi, có sự phân cụm rõ ràng giữa các sản phẩm đồ uống từ cam và nước cam hoàn toàn tự nhiên dù khi sử dụng các dữ liệu phân tích truyền thống không có sự khác biệt rõ ràng. Kết quả này cho thấy các phân tích thành phần hóa học đơn lẻ không đủ độ nhạy để phân biệt nhóm sản phẩm nước cam nhân tạo với nước cam tự nhiên [24]. Phổ Raman kết hợp phân tích đa biến cũng

cho khả năng phát hiện và định lượng pha trộn nước ép khác trong nền sản phẩm cam, kể cả khi có sự hiện diện của hương liệu bổ sung. Điều này cho thấy các phương pháp phổ có ưu thế rõ rệt trong việc phát hiện pha trộn ở các hệ sản phẩm có công thức phức tạp [144]. Phương pháp phổ hồng ngoại gần (NIR) có thể phát hiện việc bổ sung các chất cũng như phát hiện bất thường trong các sản phẩm đồ uống cam trên thị trường [41].



Hình 1.8. Các sản phẩm đồ uống từ cam

Như vậy, các kiểu gian lận phổ biến của các sản phẩm nước cam bao gồm: pha loãng sau đó bổ sung đường và acid, bổ sung phụ phẩm nhằm tăng sản lượng, pha trộn dịch quả khác, công bố sai loại sản phẩm (NFC/FC) hoặc sai nguồn gốc. Các nghiên cứu gần đây cũng đã sử dụng H-NMR cùng học máy để định lượng nhiều cấu tử và phân loại nguồn gốc của nước cam [71]. Sự khác biệt giống cũng tạo ra sai khác đáng kể về hàm lượng flavonoid và hợp chất bay hơi, do đó khi đánh giá nguồn gốc địa lý hay sự pha trộn cần phải kiểm soát hoặc mô hình hóa đồng thời yếu tố giống để tránh nhầm lẫn tín hiệu tự nhiên với tín hiệu gian lận [100].

Về công nghệ chế biến, các công nghệ không dùng nhiệt có xu hướng giữ đặc tính cam tươi tốt hơn so với dùng nhiệt, nhưng có các khác biệt đủ để tạo dấu vân tay phân biệt giữa các loại [45, 54, 78].

1.2. Các phương pháp xác định thành phần hóa học chính của nước cam

1.2.1. Một số nghiên cứu định lượng các acid hữu cơ trong nước cam

Hasim Kelebek và cộng sự đã định lượng 3 acid hữu cơ trong nước cam gồm: acid ascorbic, acid malic, acid citric bằng phương pháp HPLC. Cột sắc ký được sử dụng là Aminex HPX-97 (300 x 7,8mm), pha động là 0,045N H₂SO₄ với 6% ACN. Tốc độ dòng là 0,3ml/phút, detector UV-Vis ở bước sóng 210nm. Mẫu nước cam được ly tâm với tốc độ 4000 vòng/phút trong 20 phút, phần dịch trên được lọc qua màng lọc có kích thước lỗ lọc 0,45 µm và bảo quản ở -18°C cho đến khi phân tích. Kết quả cho thấy acid citric là acid hữu cơ chiếm tỉ lệ lớn nhất trong tổng hàm lượng acid, xấp xỉ 12,66 g/L trong 14,21 g/L tổng acid (tương đương hơn 85%). Điều này khẳng định vai trò chủ đạo của acid citric trong đặc tính hóa học của nước cam [60].

Để bổ sung việc định lượng acid tartaric, Rodrigo Scherer và cộng sự sử dụng cột sắc ký RP- C18, pha động gồm 0,01ml/L dung dịch đệm KH₂PO₄, pH= 2,6 (điều chỉnh bằng acid o-phosphoric). Tốc độ dòng 0,5ml/phút, thể tích tiêm mẫu 20ml, detector DAD, bước sóng 210nm với acid tartaric, malic, citric, 250nm với acid ascorbic. Độ lệch chuẩn tương đối của phương pháp từ 0,4% đến 2,3% (n = 10), độ lặp lại từ 1,2% đến 5,0% (n = 18) đối với độ thu hồi, giới hạn phát hiện (MDL) nằm trong khoảng từ 0,03 đến 3,31 µg/mL và định lượng (MQL) nằm trong khoảng từ 0,10 đến 11,03 µg/mL, độ thu hồi nằm trong khoảng từ 82% đến 110%. Nghiên cứu này chứng minh phương pháp có độ chính xác cao và phù hợp cho phân tích định lượng acid hữu cơ trong nước cam [123].

Tương tự, Nour và cộng sự tiến hành định lượng các acid hữu cơ trong nước cam bằng sắc ký pha đảo với cột Hypersil Gold Aq, pha động chuẩn bị bằng cách hòa tan 6,8 g Kali dihydro phosphate thành 1L, pH = 2,8 (điều chỉnh bằng acid phosphoric). Tốc độ dòng 0,7ml/phút, detector DAD với bước sóng 254nm đối với acid ascorbic và 214nm với các acid hữu cơ khác. Phương pháp có RSD nằm trong khoảng từ 0,154 đến 3,18%, đặc biệt đối với acid citric, ascorbic, lactic và oxalic (RSD <2%). Nghiên cứu cũng khẳng định acid citric là thành phần chiếm ưu thế trong tất cả các mẫu khảo sát [108].

Để khẳng định vai trò quan trọng của acid citric trong kiểm soát chất lượng nước cam, Katie Weikle và cộng sự tiến hành định lượng acid

citric trong nước cam bằng HPLC: cột Kinetex XB-C18, pha động là dung dịch đệm 100% 50 mM phosphate ở pH = 2,80. Tốc độ dòng 0,5 mL/phút, detector UV-Vis ở 214nm. Phương pháp này cho phương trình tuyến tính là $y = 0,5551x$ với $R^2 = 0,99909$, chứng minh độ tin cậy cao trong định lượng acid citric, một chỉ thị quan trọng trong kiểm soát chất lượng nước cam [76].

Các nghiên cứu đều cho thấy acid citric luôn là acid chiếm ưu thế, chiếm lượng lớn trong tổng lượng acid hữu cơ trong nước cam, bất kể phương pháp phân tích hay giống cam khảo sát, ở nhiều vùng địa lý khác nhau. Phương pháp HPLC được sử dụng phổ biến, và thuận lợi cho quá trình định lượng với pha động trong môi trường acid mạnh. Tuy nhiên, sự khác biệt về pha động, tốc độ dòng và cột sắc ký dẫn đến sự khác biệt về thời gian phân tích và độ nhạy của phương pháp.

Các nghiên cứu chủ yếu xây dựng các quy trình định lượng nhóm acid hữu cơ, rất ít nghiên cứu kết hợp dữ liệu định lượng acid hữu cơ với các mô hình học máy để phục vụ phân loại giống, truy xuất nguồn gốc hoặc phát hiện pha trộn. Và đặc biệt các nghiên cứu đều trên các giống cam ở các nước khác, chưa có bất kỳ nghiên cứu nào trên các giống cam Việt nam. Trong bối cảnh đó, việc xây dựng một cơ sở dữ liệu định lượng acid hữu cơ trên bộ mẫu đa giống, đa vùng tại Việt Nam, đồng thời tích hợp với các nhóm hợp chất khác và các thuật toán học máy nhằm xây dựng bộ dữ liệu cho Cam Việt Nam cũng như phục vụ mục đích kiểm soát chất lượng, là một hướng nghiên cứu cần thiết và có ý nghĩa thực tiễn cao.

1.2.2. Một số nghiên cứu định lượng các carotenoid trong nước cam

Pupin và cộng sự đã định lượng carotenoid trong nước cam bằng phương pháp HPLC pha đảo sử dụng cột C18 Vydac ($4,6 \times 250$ mm; 5 μ m). Pha động là hỗn hợp acetonitril:methanol:1,2-dichloroethan (60:35:5), tốc độ dòng 1,0 mL/phút, detector UV-Vis bước sóng 450 nm. Mẫu nước cam được chiết bằng ethyl acetat, cô quay và hòa tan lại trong pha động trước khi bơm mẫu. Kết quả cho thấy tổng hàm lượng carotenoid dao động từ 0,11 - 1,21 mg/L. Phương pháp có khoảng tuyến tính trong khoảng 0,03 - 0,4 mg/L, giới hạn phát hiện 0,01 - 0,02 mg/L nhưng độ thu hồi chỉ đạt khoảng 71 %, cho thấy ảnh hưởng đáng kể của quá trình chiết mẫu đến hiệu suất thu

hồi của phương pháp định lượng [14]. Gama và cộng sự đã phát triển phương pháp HPLC-DAD với cột C18 ($4,6 \times 250$ mm; $5 \mu\text{m}$) và hệ pha động gradient gồm acetonitril, methanol và ethyl acetat. Detector được đặt tại bước sóng 450 nm, tốc độ dòng 0,7 mL/phút. Mẫu nước cam được chiết bằng acetone, cô quay và hòa tan lại trong methanol trước khi phân tích. Kết quả cho thấy tổng hàm lượng carotenoid trong nước cam đạt khoảng $12 \pm 6,7$ mg/L. Phương pháp cho độ thu hồi đạt 91 - 99%, chứng minh độ tin cậy tốt hơn so với các phương pháp chiết đơn giản. Nghiên cứu chỉ ra rằng có sự khác biệt rõ ràng về hàm lượng carotenoid giữa các loại cam [65].

Các nghiên cứu đều cho kết quả HPLC pha đảo với cột C18 phù hợp để định lượng carotenoid. Tuy nhiên, do carotenoid là các hợp chất nhạy với ánh sáng, nhiệt độ và oxy, quá trình xử lý mẫu (chiết, cô quay, bảo quản) có ảnh hưởng rất lớn đến kết quả định lượng. Do đó, việc tối ưu hóa điều kiện chiết và bảo quản mẫu là yếu tố quan trọng quyết định độ chính xác của phương pháp. Đặc biệt, chưa có nghiên cứu nào xây dựng cơ sở dữ liệu carotenoid với mẫu cam Việt Nam và tích hợp với học máy để phục vụ phân biệt và xác thực nước cam.

1.2.3. Một số nghiên cứu định lượng các flavonoid trong nước cam

Belajova và cộng sự đã định lượng flavonoid trong nước cam bằng HPLC-DAD với cột C6-phenyl (3×150 mm; $5 \mu\text{m}$). Pha động gồm kênh A là acid phosphoric 0,01 M/methanol (80:20, v/v) và kênh B là methanol, chạy gradient trong 55 phút, tốc độ dòng 0,8 mL/phút. Các mẫu được lọc qua giấy lọc rồi tiêm vào hệ thống. Kết quả cho thấy hesperidin có hàm lượng cao nhất (92,7 mg/L), trong khi naringin (7,0 mg/L), neohesperidin (11,5 mg/L) và quercetin (23,0 mg/L) có hàm lượng thấp hơn đáng kể. Điều này khẳng định vai trò đặc trưng của hesperidin trong nước cam. Một số mẫu nước cam thương mại không phát hiện naringin và quercetin [22]. Silva và cộng sự đã chiết mẫu nước cam bằng ethyl acetat, cô quay chân không ở nhiệt độ 40°C , hòa lẫn bằng methanol, lọc qua màng lọc $0,22 \mu\text{m}$, rồi bơm mẫu vào hệ HPLC-DAD với cột C18 (4×55 mm; $3 \mu\text{m}$). Pha động gồm methanol và acid acetic 1%, chạy gradient trong 50 phút với tốc độ dòng 0,3 mL/phút. Kết quả cho thấy hàm lượng hesperidin dao động từ 18,80 - 139,00 mg/100 g, cao hơn nhiều so với các flavonoid

khác. Phương pháp này cho độ nhạy cao với giới hạn phát hiện ở mức $\mu\text{g}/100\text{ g}$, phù hợp cho phân tích định lượng chính xác [130].

Các nghiên cứu cho thấy HPLC-DAD phù hợp để định lượng flavonoid. Mẫu nước cam được chiết bằng kỹ thuật phù hợp cho độ nhạy tốt hơn so với định lượng trực tiếp. Các nghiên cứu đều thống nhất rằng hesperidin là flavonoid chiếm ưu thế, trong khi các flavonoid khác có hàm lượng thấp hơn đáng kể. Tại Việt Nam, chưa có nghiên cứu hệ thống nào xây dựng cơ sở dữ liệu flavonoid nước cam Việt Nam, cũng như chưa có nghiên cứu kết hợp bộ dữ liệu này với học máy để phục vụ truy xuất nguồn gốc và phát hiện gian lận.

Tổng hợp các nghiên cứu đã công bố cho thấy HPLC là kỹ thuật hiệu quả trong phân tích định lượng các nhóm chất chính của nước cam như acid hữu cơ, carotenoid và flavonoid. Đối với nhóm acid hữu cơ, phần lớn các nghiên cứu cho thấy mẫu nước cam có thể được phân tích trực tiếp sau các bước xử lý đơn giản như ly tâm và lọc mà không cần chiết tách hay làm giàu mẫu, do các acid hữu cơ có hàm lượng tương đối cao và tan tốt trong nước. Đối với carotenoid và flavonoid, kỹ thuật chiết ảnh hưởng trực tiếp đến độ nhạy và hiệu suất thu hồi, do các hợp chất này có độ tan phức tạp, từ tan tốt trong nước đến rất ít tan trong nước và tồn tại ở hàm lượng thấp hơn, chịu ảnh hưởng mạnh bởi nền mẫu. Điều này cho thấy tùy theo đặc tính hóa học và nồng độ của từng nhóm hợp cần lựa chọn phương pháp phù hợp để xử lý mẫu.

Các nghiên cứu cũng cho thấy hàm lượng các hợp chất hóa học trong nước cam có sự biến thiên đáng kể theo giống cam, vùng địa lý và mức độ tự nhiên của sản phẩm. Điều này cho thấy thành phần hóa học chịu ảnh hưởng mạnh của yếu tố di truyền, điều kiện khí hậu, ánh sáng và kỹ thuật canh tác. Ngoài ra, nhiều nghiên cứu về kiểm soát tính xác thực cho thấy thành phần hóa học của nước cam không chỉ phản ánh đặc điểm sinh hóa của từng giống và vùng trồng mà còn có tiềm năng lớn trong truy xuất nguồn gốc và phát hiện gian lận thực phẩm.

Hiện chưa có nghiên cứu hệ thống nào xây dựng cơ sở dữ liệu tổng hợp về thành phần hóa học trên các giống cam Việt Nam nhằm phục vụ phân loại giống, truy xuất

nguồn gốc và phát hiện pha trộn. Đây là khoảng trống nghiên cứu quan trọng và cũng là cơ sở khoa học cho hướng tiếp cận của luận án này.

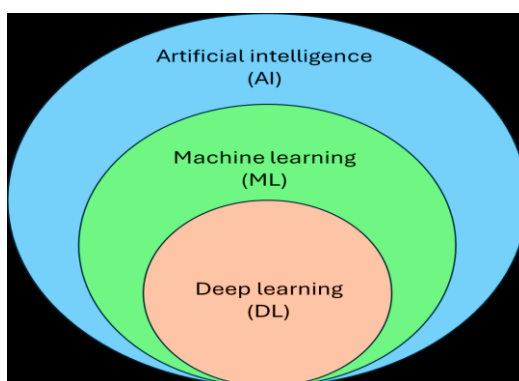
1.3. Học máy trong phân biệt và phân loại cam

1.3.1. Một số khái niệm về trí tuệ nhân tạo

Trí tuệ nhân tạo (AI - Artificial intelligence): là một lĩnh vực khoa học máy tính rộng lớn tập trung vào việc tạo ra các tác nhân thông minh, có thể suy nghĩ, học hỏi và hành động tự chủ. Mục tiêu của AI là tạo ra các máy móc có thể mô phỏng các khả năng nhận thức của con người như học tập, lý luận, giải quyết vấn đề, nhận thức, lập kế hoạch, di chuyển và thao tác với môi trường xung quanh.

Học máy (ML - Machine Learning): là một nhánh con của AI, tập trung vào việc cho phép máy móc học hỏi từ dữ liệu mà không cần được lập trình rõ ràng. Các thuật toán học máy có thể phân tích dữ liệu, nhận dạng mô hình và đưa ra dự đoán hoặc quyết định dựa trên những gì đã học được.

Học sâu (DL - Deep Learning): là một nhánh con của học máy sử dụng các Biến thể của mạng nơ-ron nhân tạo (artificial neural networks) để học hỏi từ dữ liệu, cho phép mạng học hỏi các mô hình phức tạp từ dữ liệu và đưa ra dự đoán chính xác hơn so với các phương pháp học máy truyền thống.



Hình 1.9. Mối quan hệ của trí tuệ nhân tạo, học máy và học sâu

1.3.2. Các thuật toán học máy ứng dụng trong phân loại và xác thực nước cam

1.3.2.1. Tiền xử lý dữ liệu phổ và dữ liệu phân tích

Biến chuẩn hóa (SVN - Standard Normal Variate) có tác dụng loại bỏ ảnh hưởng của sự thay đổi cường độ tín hiệu tổng thể và các sai lệch do điều kiện đo, đặc biệt

phù hợp với dữ liệu phổ UV-Vis, NIR và IR [23]. SNV làm giảm các biến thiên không mong muốn, tăng khả năng so sánh giữa các mẫu và cải thiện độ ổn định của các mô hình phân loại. Tuy nhiên, SNV không xử lý được các nhiễu phi tuyến hoặc biến dạng nền phức tạp, do đó thường được kết hợp với các kỹ thuật lọc và biến đổi khác.

Bộ lọc Savitzky-Golay kết hợp với biến đổi đạo hàm bậc nhất được sử dụng trong trường hợp dữ liệu phổ có nhiễu cao hoặc các đỉnh hấp thụ có sự chồng lấn. Phương pháp này cho phép làm mượt dữ liệu, làm nổi bật các đặc trưng phổ liên quan đến thành phần hóa học, qua đó nâng cao khả năng phân biệt của mô hình. Việc lựa chọn cửa sổ lọc và bậc đa thức cần được tối ưu nhằm tránh làm mất thông tin hóa học quan trọng [25].

1.3.2.2. Phương pháp học máy không giám sát

Phân tích thành phần chính (PCA - Principal Component Analysis) được sử dụng để giảm chiều dữ liệu và khảo sát cấu trúc tổng thể của bộ dữ liệu [70]. Trong nghiên cứu nước cam, PCA đóng vai trò thăm dò ban đầu, giúp đánh giá mức độ phân tách tự nhiên của các mẫu theo giống hoặc nguồn gốc và phát hiện các điểm ngoại lai có thể ảnh hưởng đến độ tin cậy của mô hình. Tuy nhiên, PCA không trực tiếp tối ưu cho mục tiêu phân loại do không sử dụng thông tin nhãn.

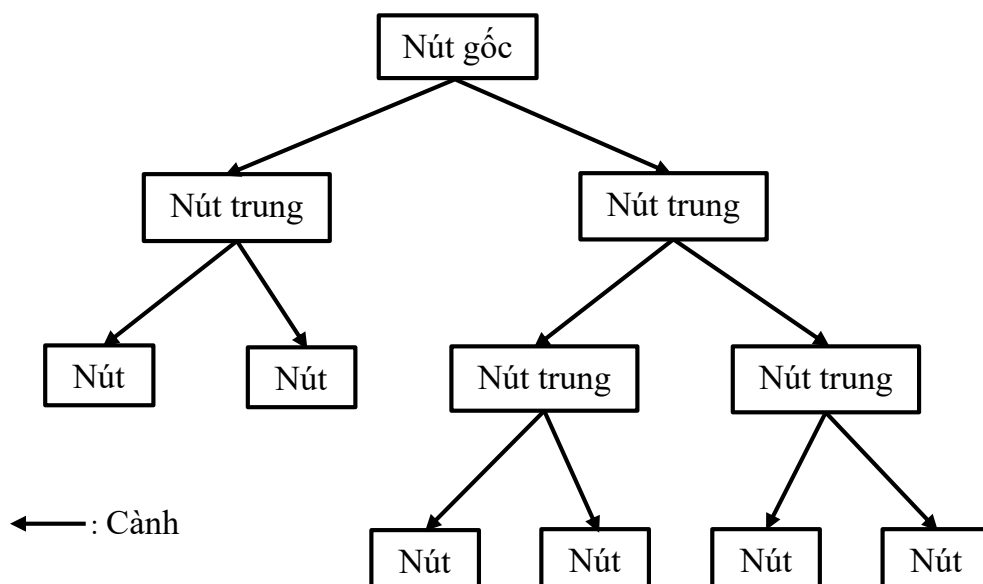
Phân tích cụm (Cluster Analysis - CA) và độ tương đồng cosin được sử dụng để đánh giá mức độ tương đồng giữa các mẫu khi chưa có thông tin nhãn. Các phương pháp này hỗ trợ kiểm tra giả thuyết về sự đồng nhất nội tại của các nhóm mẫu, song kết quả phụ thuộc mạnh vào thước đo khoảng cách và phương pháp liên kết, do đó chủ yếu mang tính hỗ trợ [77].

1.3.2.3. Các phương pháp phân loại tuyến tính

Phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis) là phương pháp phân loại có giám sát dựa trên giả thiết tuyến tính, nhằm tối đa hóa khoảng cách giữa các lớp trong khi tối thiểu hóa độ phân tán nội lớp. LDA có ưu điểm về khả năng diễn giải và ổn định khi số biến không quá lớn, nhưng hiệu năng có thể suy giảm đối với dữ liệu phi tuyến hoặc chồng lấn mạnh [51].

1.3.2.4. Các phương pháp học máy phi tuyến

Cây quyết định (DT - Decision Tree) là thuật toán sử dụng cấu trúc mô hình cây phân cấp dùng để phân loại các lớp của đối tượng với biến đầu vào là các giá trị đặc trưng của đối tượng đó. Cây quyết định sẽ tạo ra các điều kiện phù hợp để phân loại được các đối tượng trong tập train, và từ đó sau khi có được các quyết định sẽ sử dụng chúng để dự đoán các đối tượng trong tập test. Cây quyết định có thể được ứng dụng trong cả bài toán hồi quy và bài toán phân loại, phù hợp cho khảo sát ban đầu. Tuy nhiên, DT đơn lẻ dễ bị quá khớp khi dữ liệu phức tạp [25]. Cấu trúc của cây quyết định được biểu diễn như hình 1.10.



Hình 1.10. Mô hình thuật toán cây quyết định

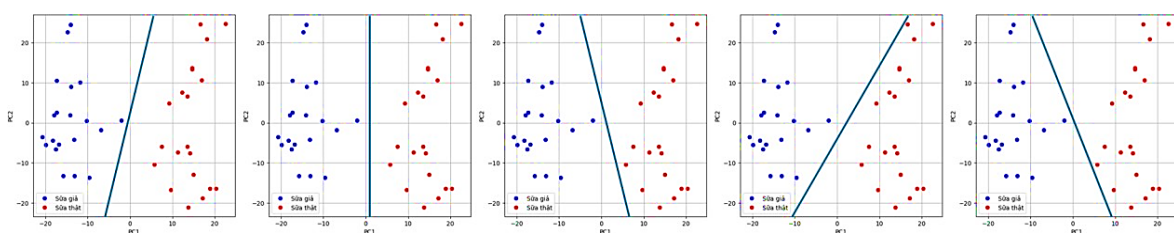
Rừng ngẫu nhiên (RF - Random Forest) là thuật toán học máy có giám sát, được xây dựng dựa trên nhiều cây quyết định. Ý tưởng của thuật toán này là kết hợp nhiều cây quyết định độc lập làm cơ sở để đưa ra dự đoán thay vì chỉ sử dụng một cây quyết định. Mỗi cây quyết định trong rừng ngẫu nhiên được xây dựng dựa trên các thuộc tính ngẫu nhiên, các yếu tố khác nhau. Mỗi nút trung gian của mỗi cây quyết định có một tiêu chí phân loại khác nhau. Sự ngẫu nhiên này làm giảm hiện tượng quá khớp (overfitting) và làm giảm mức độ tương quan giữa các cây quyết định. Đối với bài toán phân loại, RF sử dụng phương pháp "bỏ phiếu" để xác định kết quả cuối cùng. Nhãn dự đoán của mô hình là nhãn được đa số các cây quyết định lựa chọn. Với bài

toán hồi quy, mô hình tính trung bình giá trị dự đoán từ tất cả các cây để đưa ra kết quả. Các tập dữ liệu con dùng để huấn luyện được chọn ngẫu nhiên bằng phương pháp bootstrap sampling (lấy mẫu có hoàn lại) từ tập dữ liệu tổng thể. RF xử lý tốt dữ liệu nhiều chiều, đồng thời cung cấp thông tin về tầm quan trọng của biến, hỗ trợ xác định các chỉ thị hóa học. Nhược điểm chính là khó diễn giải trực quan so với các mô hình tuyến tính [27].

XGBoost (Extreme Gradient Boosting) là một thư viện học máy được thiết kế đặc biệt để cải thiện hiệu suất và tốc độ của các mô hình dựa trên thuật toán gradient boosting. XGBoost xây dựng một tập hợp các cây quyết định một cách tuần tự. Trong bài toán hồi quy, mỗi cây quyết định mới được tạo ra nhằm giảm thiểu lỗi dự đoán từ các cây trước đó bằng cách dựa trên mức độ sai lệch của dự đoán, thể hiện bằng hàm lỗi. Ban đầu, mô hình sẽ huấn luyện cây quyết định đầu tiên để tính toán lỗi sai, sau đó tiếp tục đào tạo cây quyết định thứ hai sử dụng cùng dữ liệu đầu vào như cây đầu tiên nhưng khác đầu ra để tiếp tục tính toán lỗi sai của dự đoán so với lỗi sai của dự đoán trước đó. Quy trình này được lặp đi lặp lại cho đến khi lỗi sai của hai lần dự đoán không đổi hoặc số lượng cây đạt được giá trị điều chỉnh. XGBoost đặc biệt hiệu quả trong các bài toán dữ liệu phi tuyến và số biến lớn. Mô hình này thường cho hiệu năng phân loại cao, song yêu cầu tinh chỉnh tham số cẩn trọng và khả năng diễn giải trực tiếp còn hạn chế [30]. Trong nghiên cứu nước cam, XGBoost được sử dụng để phân loại giống và truy xuất nguồn gốc địa lý dựa trên dữ liệu thành phần hóa học và dữ liệu phổ có tính phi tuyến cao. Thuật toán này cho phép khai thác hiệu quả mối quan hệ phức tạp giữa các biến, đồng thời cung cấp thông tin về mức độ đóng góp của từng biến thông qua chỉ số tầm quan trọng đặc trưng.

1.3.2.5. Máy vector hỗ trợ (SVM - Support Vector Machine)

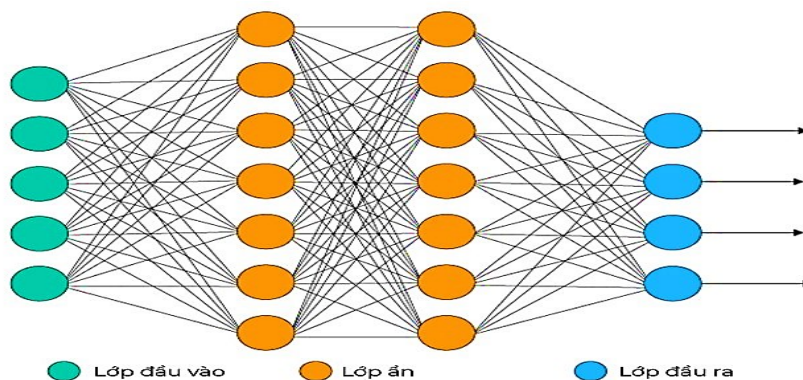
SVM là phương pháp phân loại mạnh dựa trên việc tối ưu hóa biên phân cách giữa các lớp trong không gian đặc trưng. Với các hàm nhân phi tuyến, SVM xử lý hiệu quả dữ liệu phổ có tính chồng lấn cao. Nhược điểm của phương pháp là chi phí tính toán lớn khi số mẫu tăng và khó diễn giải ý nghĩa hóa học của mô hình [31]. Các cách phân chia dữ liệu trong không gian được biểu thị trên hình 1.11.



Hình 1.11. Mô hình thuật toán SVM

1.3.2.6. Mạng nơ-ron lan truyền đa lớp (MLP/ANN - MultiLayer Perceptron)

MLP cho phép mô hình hóa các mối quan hệ phi tuyến phức tạp giữa dữ liệu đầu vào và kết quả phân loại. Trong nghiên cứu nước cam, MLP thường đạt độ chính xác cao khi có đủ dữ liệu huấn luyện. Tuy nhiên, mô hình có nguy cơ quá khớp nếu không được kiểm soát chặt chẽ [25].



Hình 1.12. Mô hình thuật toán MLP

Mô hình MLP được thể hiện trên hình 1.12, gồm một lớp đầu vào, một hoặc nhiều lớp ẩn và một lớp đầu ra. Dữ liệu đầu vào (các biến hóa học hoặc đặc trưng phổ sau tiền xử lý) được lan truyền theo một chiều qua các lớp ẩn thông qua các hàm kích hoạt phi tuyến, cho phép mô hình hóa các mối quan hệ phức tạp giữa thành phần hóa học và tiêu chí phân loại. Lớp đầu ra cung cấp xác suất hoặc nhãn phân loại tương ứng với giống hoặc nguồn gốc nước cam.

1.3.2.7. Bình phương tối thiểu từng phần theo khoảng (iPLS - Interval Partial Least Squares)

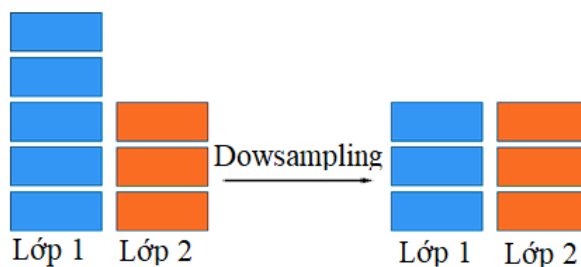
iPLS được sử dụng nhằm xác định các biến đóng góp mạnh nhất cho mô hình phân loại, đặc biệt hữu ích với dữ liệu phổ. Phương pháp này giúp giảm chiều dữ liệu, cải thiện hiệu năng và tăng khả năng diễn giải hóa học, song phụ thuộc vào cách chia

khoảng và tiêu chí lựa chọn [107]. iPLS lựa chọn biến dựa trên mô hình hồi quy bình phương tối thiểu từng phần, trong đó miền biến ban đầu (đặc biệt là dữ liệu phổ) được chia thành nhiều khoảng (intervals) liên tiếp. Với mỗi khoảng, một mô hình PLS riêng biệt được xây dựng và đánh giá hiệu năng dự đoán. Các khoảng biến thể hiện khả năng dự đoán tốt nhất sẽ được lựa chọn để xây dựng mô hình cuối cùng. Trong nghiên cứu nước cam, iPLS được sử dụng nhằm xác định các vùng phổ hoặc nhóm biến hóa học đóng góp quan trọng nhất cho việc phân biệt giống và truy xuất nguồn gốc địa lý, từ đó giảm chiều dữ liệu, tăng độ ổn định của mô hình học máy và nâng cao khả năng diễn giải ý nghĩa hóa học.

1.3.2.8. Thuật toán cân bằng dữ liệu

Một vấn đề thường gặp trong các bài toán phân loại đó là mất cân bằng dữ liệu, khi một lớp dữ liệu có số lượng mẫu lớn hơn rất nhiều so với các lớp dữ liệu. Điều này gây ra việc các mô hình phân loại sẽ có xu hướng thiên vị hơn cho lớp dữ liệu có nhiều mẫu hơn, khiến cho các mô hình trở nên không khách quan trong việc phân loại. Các hướng giải quyết phổ thông nhất thường sẽ cố làm cho dữ liệu ở các lớp trở nên đồng đều hơn, sử dụng các thuật toán cân bằng dữ liệu.

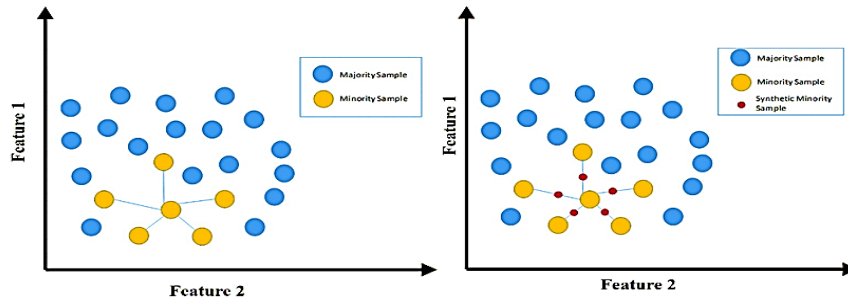
Downsampling là kỹ thuật khiến cho các lớp dữ liệu có số mẫu đồng đều nhau bằng cách cắt giảm số mẫu ở các lớp có số mẫu lớn hơn một cách ngẫu nhiên. Mô hình hóa thuật toán downsampling dữ liệu được thể hiện trên hình 1.13.



Hình 1.13. Mô hình thuật toán downsampling

Việc cắt giảm dữ liệu ở các lớp có số mẫu lớn hơn khiến cho dữ liệu đồng đều hơn, nhưng đồng thời cũng có thể làm vô tình mất đi các mẫu quan trọng, nên việc sử dụng thuật toán downsampling thường được áp dụng khi toàn bộ tập dữ liệu có số mẫu khá lớn.

Upsampling tạo ra các điểm dữ liệu mới bằng các phương pháp nội suy dựa trên dữ liệu đã có. Một trong các thuật toán phổ biến nhất của kỹ thuật này là SMOTE - Synthetic Minority Over-sampling Technique. SMOTE tạo các mẫu lớp thiểu số mới bằng cách thực hiện các bước nhỏ từ một trong các mẫu lớp thiểu số đến một trong k hàng xóm gần nhất của nó trong không gian đặc trưng. Thuật toán tạo ra một mẫu mới bằng cách chọn ngẫu nhiên một trong k hàng xóm gần nhất và sau đó thêm một nhiễu loạn nhỏ vào vector đặc trưng giữa mẫu ban đầu và hàng xóm được chọn. Điều này tạo ra dữ liệu tổng hợp mới tương tự như các mẫu lớp thiểu số trong không gian đặc trưng nhưng không phải là bản sao chính xác của bất kỳ mẫu hiện có nào. Cách tạo ra dữ liệu mới được hiển thị ở hình 1.14.



Hình 1.14. Mô hình thuật toán SMOTE

1.3.3. Phương pháp đánh giá độ chính xác mô hình

1.3.3.1 Các phép đo hồi quy

- Giá trị RMSE:
$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$
 (1.1)

Trong đó: y_i : giá trị mẫu thứ i thực tế

$\hat{y}_i$: giá trị mẫu thứ i được dự đoán

n : số mẫu

RMSE tính toán trung bình sai khác của giá trị dự đoán và giá trị thực tế, tuy nhiên do đã khử được bậc hai nên sai số tính toán không bị phóng đại, điều đó giúp cho thuật toán RMSE đáng tin cậy hơn khi biểu thị sai số.

- Hệ số xác định R^2 :
$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
 (1.2)

Trong đó: $\bar{y}$: giá trị trung bình của biến mục tiêu

Hệ số xác định R^2 cho biết tỉ lệ phương sai của biến mục tiêu được giải thích bởi các biến độc lập trong mô hình hồi quy tuyến tính. Giá trị của hệ số xác định nằm trong khoảng từ 0 đến 1. Nếu giá trị của hệ số xác định càng gần 1, mô hình hồi quy giải thích càng nhiều sự biến thiên của biến mục tiêu.

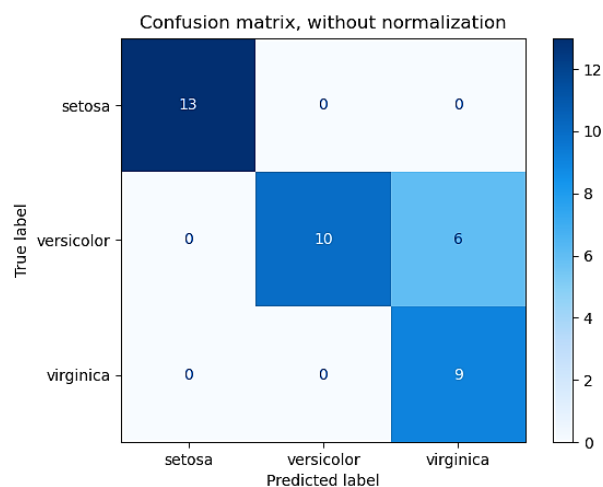
$$\text{- Phần trăm sai số: } \%Error = \left| \frac{y_i - \hat{y}}{\hat{y}} \right| \cdot 100\% \quad (1.3)$$

Giá trị phần trăm sai số thể hiện sai số tương đối giữa giá trị dự đoán và giá trị thực tế, không như RMSE giá trị phần trăm sai số được tính trên từng giá trị thay vì lấy trung bình cộng. Phần trăm sai số càng gần 0 thì sai số càng thấp chứng tỏ độ chính xác của mô hình càng cao.

1.3.3.2 Các phép đo phân loại

$$\text{Độ chính xác: } \text{Độ chính xác} = \frac{\text{Số mẫu đoán đúng}}{\text{Tổng số mẫu dự đoán}} \quad (1.4)$$

Ma trận nhầm lẫn (confusion matrix): là một bảng dùng để đánh giá hiệu suất của mô hình phân loại, cho biết mức độ chính xác của các dự đoán. Bảng này so sánh giữa các dự đoán của mô hình và kết quả thực tế, giúp xác định số lần mô hình dự đoán đúng hay sai cho từng loại cụ thể. Trục hoành của ma trận đại diện cho các giá trị thực tế, trong khi trục tung của ma trận đại diện cho các giá trị dự đoán. Ví dụ về ma trận nhầm lẫn được biểu thị ở hình 1.15.



Hình 1.15. Ma trận nhầm lẫn của phép phân loại ba lớp

Các giá trị nằm trên đường chéo chính của ma trận nhầm lẫn chính là số mẫu dự đoán đúng của các lớp phân loại, trong khi đó tất cả các giá trị nằm lệch khỏi đường chéo đều là số mẫu dự đoán sai.

1.4. Một số nghiên cứu phân biệt, phân loại cam

Nước cam có thành phần hóa học phức tạp, trong đó nhiều hợp chất cùng đóng góp vào tín hiệu phân tích. Đồng thời, sự biến thiên tự nhiên theo giống, theo vùng địa lý cũng như công nghệ chế biến tạo ra mức chồng lấn đáng kể giữa các mẫu, khiến việc phân biệt và xác thực bằng các chỉ tiêu hóa học đơn biến gặp nhiều hạn chế. Để đáp ứng nhu cầu phân biệt, phân loại các sản phẩm nước cam cần thực hiện bằng các phương pháp phân tích dữ liệu đa biến kết hợp học máy nhằm khai thác đồng thời thông tin từ nhiều biến đầu vào, từ đó nâng cao khả năng xác thực và truy xuất nguồn gốc các sản phẩm nước cam. Một số nghiên cứu về phân loại và xác thực nước cam được trình bày trong Bảng 1.4.

Bảng 1.5. Một số nghiên cứu phân biệt và xác thực nước cam

TT	Mục tiêu	Đối tượng	Kỹ thuật phân tích	Thuật toán	Kết quả chính	TLTK
1	Phân biệt NFC/FC	Nước cam NFC và FC	HS-SPME-GC-MS	PCA, PLS-DA	Các hợp chất bay hơi cho phép phân biệt NFC/FC.	[134]
2	Phân biệt NFC/FC	NFC/FC mô phỏng và thương mại	UPLC-QTOF-MS	PCA, OPLS-DA	Phân tích các chất chuyển hóa không mục tiêu cho phép phân loại NFC/FC.	[135]
3	Biến đổi trong bảo quản	Nước cam xử lý bằng áp suất cao	GC-MS, LC-MS	PLS	Sau 21 ngày, hơn 60% các chất chuyển hóa biến đổi	[133]

TT	Mục tiêu	Đối tượng	Kỹ thuật phân tích	Thuật toán	Kết quả chính	TLTK
4	Phát hiện pha trộn quýt	Nước cam và nước quýt	HS-SPME-GC-MS	PCA, LDA	Các chất bay hơi cho phép phát hiện pha trộn nước quýt vào nước cam	[151]
5	Phân biệt giống	Nước cam tươi nhiều giống	HPLC-DAD GC-MS	MANOVA, LDA	Kết hợp các thành phần hóa học cho phép phân biệt giống cam.	[145]
6	Truy xuất nguồn gốc	Nước cam ở các vùng	¹ H-NMR	PCA, PLS-DA	NMR định lượng cho phép phân biệt nguồn gốc địa lý.	[85]
7	Phát hiện nhanh sự pha trộn chất hóa học	Mẫu nước cam thương mại	NIR cầm tay	SIMCA, PLS-DA	NIR + SIMCA phù hợp cho sàng lọc xác thực nhanh hiện trường.	[129]
8	Phát hiện pha trộn bưởi	Nước cam và nước bưởi	Raman cầm tay	PLS-DA	Raman kết hợp PLS-DA phát hiện được sự pha trộn nước bưởi vào nước cam.	[143]
9	Xác thực nước cam	Nước cam và nước bưởi	Điện hóa	PCA, PLS-DA	Dấu vân tay điện hóa kết hợp PLS cho khả năng phát hiện sự pha trộn.	[104]

TT	Mục tiêu	Đối tượng	Kỹ thuật phân tích	Thuật toán	Kết quả chính	TLTK
10	Phân biệt các loại nước ép họ Cam quýt	Nước quả họ Citrus	HS-GC-MS-IMS	LDA, SVM, kNN	Các chất dễ bay hơi kết hợp thuật toán có thể phân biệt các loại nước họ Cam.	[28]
11	Phân biệt nước hoa quả	Nước cam, chanh, chanh leo, bưởi	¹ H-NMR, HPLC-DAD	PCA, PLS	Kết hợp dữ liệu NMR và HPLC cho khả năng phân biệt các loại nước ép.	[71]

*** Nhận xét phân tổng quan**

Thông qua tổng quan các công bố trong và ngoài nước cho thấy việc xác thực, phân loại và truy xuất nguồn gốc nước cam là một hướng nghiên cứu có ý nghĩa khoa học và là nhu cầu thực tế của thị trường nước hoa quả nói chung và nước cam nói riêng. Nhiều nghiên cứu đã khẳng định thành phần hóa học của nước cam, bao gồm đường, acid hữu cơ, flavonoid, carotenoid và các hợp chất bay hơi, chịu ảnh hưởng đồng thời của giống, nguồn gốc địa lý và điều kiện canh tác, từ đó có thể được khai thác như các dấu vân tay để phân biệt và truy xuất nguồn gốc. Xu hướng nghiên cứu chuyển mạnh từ việc sử dụng các thành phần hóa học đơn lẻ sang khai thác dữ liệu tổng thể đa biến hay các dấu vân tay hóa học, kết hợp với các phương pháp học máy. Các thuật toán học máy như PCA, LDA, SVM, RF, XGBoost ... đã được các nghiên cứu chỉ ra rằng có thể được sử dụng để phân loại nước cam theo loại sản phẩm, giống hoặc nguồn gốc địa lý. Tuy nhiên, các nghiên cứu này được thực hiện trên bộ mẫu thu thập tại các quốc gia với giống và điều kiện sinh thái khác biệt đáng kể so với Việt Nam.

Cũng qua tổng quan tài liệu, nhận thấy có một số phương pháp phân tích đã được áp dụng để định lượng các thành phần hóa học chính trong nước cam trong đó HPLC là kỹ thuật đang được sử dụng phổ biến nhất. Tuy nhiên, khi so sánh các

công bố trên thế giới, các điều kiện phân tích cũng như điều kiện xử lý mẫu khác nhau rõ rệt.

Việt Nam có sự đa dạng đáng kể về giống cam và nhiều vùng trồng cam đặc sản với sự đa dạng về khí hậu, thổ nhưỡng cũng như kinh nghiệm trồng và chăm sóc. Cùng một giống cam nhưng được trồng ở những vùng khác nhau có thể cho chất lượng rất khác nhau. Tuy nhiên, hiện tại chưa có nghiên cứu nghiên cứu mang tính hệ thống về xây dựng và áp dụng phương pháp phân tích thành phần hóa học cho nước cam dựa trên bộ mẫu là các giống cam và vùng trồng cam tại Việt Nam. Nhiều nghiên cứu tập trung vào bài toán phát hiện gian lận hoặc phân biệt NFC/FC trong bối cảnh kiểm soát thương mại, trong khi chưa có báo cáo nào về việc xác lập cơ sở khoa học cho phân biệt giống và nguồn gốc phục vụ bảo hộ chỉ dẫn địa lý đối với cam Việt Nam.

Xuất phát từ các khoảng trống nghiên cứu nêu trên, luận án lựa chọn hướng nghiên cứu xây dựng và áp dụng phương pháp phân tích thành phần hóa học, các tín hiệu đo kết hợp với học máy nhằm phân biệt giống, truy xuất nguồn gốc và phát hiện pha trộn nước cam Việt Nam. Hướng tiếp cận này không chỉ cho phép thu nhận bộ dữ liệu phân tích có độ tin cậy cao, phù hợp với điều kiện trong nước, mà còn tạo nền tảng để đánh giá một cách có hệ thống hiệu quả của các thuật toán học máy trong bối cảnh thực tiễn sản xuất và kiểm soát chất lượng tại Việt Nam. Kết quả nghiên cứu kỳ vọng góp phần bổ sung cơ sở khoa học cho công tác kiểm soát chất lượng, bảo vệ chỉ dẫn địa lý và phát triển bền vững các sản phẩm nước cam tại Việt Nam.

Chương 2. THỰC NGHIỆM

2.1. Hóa chất, dụng cụ và thiết bị

2.1.1. Hóa chất

- Các chất chuẩn acid gồm: ascorbic (47863), oxalic (41706), malic (94916), citric (46933), fumaric (1286708), độ tinh khiết >99% của Merck (Đức). Dung dịch chuẩn gốc từng acid nồng độ 20 g/L được pha bằng cách cân chính xác khoảng 200 mg chất chuẩn (có tính đến độ tinh khiết) trên cân phân tích có độ chính xác 0,1 mg, hòa tan vào nước cất và định mức đến thể tích 10,00 mL bằng nước cất. Dung dịch hỗn hợp chuẩn trung gian có nồng độ 3000 mg/L được pha loãng từ dung dịch chuẩn gốc bằng nước cất. Các dung dịch hỗn hợp chuẩn làm việc nồng độ từ 5 đến 3000 mg/L mỗi chất được pha loãng từ dung dịch hỗn hợp chuẩn trung gian bằng nước cất.

- Các chất chuẩn carotenoid gồm: lutein (07168), lycopene (SMB00706), β -caroten (22040), độ tinh khiết >96% của Sigma Aldrich (Mỹ). Dung dịch chuẩn gốc từng chất nồng độ 100 mg/L được pha bằng cách cân chính xác khoảng 1 mg chất chuẩn (có tính đến độ tinh khiết) trên cân phân tích có độ chính xác 0,1 mg, hòa tan và định mức đến thể tích 10,00 mL bằng dung môi dichloromethane. Các dung dịch hỗn hợp chuẩn làm việc được pha loãng từ dung dịch hỗn hợp chuẩn gốc bằng dung môi dichloromethane để thu được dãy các dung dịch chuẩn làm việc có nồng độ từ 0,01 đến 50 mg/L.

- Các chất chuẩn flavonoid gồm: naringin (71162), hesperidin (50162), neohesperidin (44247), quercetin (47959), độ tinh khiết >95% của Merck (Đức). Dung dịch chuẩn gốc từng chất nồng độ 1000 ppm được chuẩn bị bằng cách cân chính xác khoảng 25 mg chất chuẩn (có tính đến độ tinh khiết) trên cân phân tích có độ chính xác 0,1 mg, hòa tan và định mức đến thể tích 25,00 mL bằng methanol. Dung dịch hỗn hợp chuẩn trung gian có nồng độ 100 ppm được pha loãng từ dung dịch chuẩn gốc với dung môi methanol. Các dung dịch hỗn hợp chuẩn làm việc nồng độ từ 0,1 đến 100 ppm mỗi chất được pha loãng từ dung dịch hỗn hợp chuẩn trung gian bằng dung môi methanol.

- Các hóa chất, dung môi khác: muối kali dihydrophosphat, acid phosphoric đặc, ACN, n-hexan, ethyl acetate, dichloromethan, MeOH, EtOH, acid hydrochloric (HCl, 37%), tất cả đều thuộc loại tinh khiết HPLC có nguồn gốc từ Merck (Đức).

- Nước cất hai lần được chuẩn bị bằng hệ thống Milli-Q tại phòng thí nghiệm Viện kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia.

2.1.2. Thiết bị và dụng cụ

- Hệ thống sắc ký lỏng hiệu năng cao HPLC (Alliance - Waters e2695) trang bị detector 2998 PDA - Waters, Singapore. Cột và tiền cột tương ứng; thể tích tiêm mẫu 10 μ L. Thiết bị được đặt tại Viện kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia.

- Hệ thống sắc ký lỏng hiệu năng cao HPLC (Shimadzu GC2014) trang bị detector UV, Nhật Bản. Cột sắc ký và tiền cột tương ứng, thể tích tiêm mẫu 10 μ L. Thiết bị được đặt tại Học viện Y dược học cổ truyền Việt Nam.

- Thiết Bị cực phổ 797 VA Computrace (Metrohm), Thụy Sĩ, với hệ điện cực: AE: Pt, WE: Au, RE: Ag/AgCl (KCl 3M). Thiết bị được đặt tại Viện Khoa học và công nghệ Quân sự.

- Hệ thống đo hồng ngoại FTIR Cary 630 của hãng Agilent Technologies (Mỹ). Thiết bị được đặt tại Học viện Y dược học cổ truyền Việt Nam.

- Hệ thống thiết bị UV-Vis UV 2400PC của hãng Shimadzu (Nhật Bản). Thiết bị được đặt tại Học viện Quân Y.

- Hệ thống sắc ký lỏng khối phổ phân giải cao LC-HR/MS Dionex Ultimate UHPLC kết hợp với khối phổ Q-Exactive Orbitrap MS được trang bị bơm tứ phân và bộ lấy mẫu tự động của hãng Thermo Scientific. Thiết bị được đặt tại Viện kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia.

- Các dụng cụ cơ bản khác: cân, máy lắc, máy ly tâm, máy voltex, micropipet, máy siêu âm, bộ lọc hút chân không, máy đo pH....

2.1.3. Phần mềm xử lý số liệu

- Các giá trị phân tích được xử lý bằng phần mềm tương ứng đi kèm thiết bị.

- Các phép thống kê được xử lý bằng Microsoft Excel

- Các mô hình học máy được xây dựng bằng phần mềm Visual Studio Code với ngôn ngữ Python phiên bản 3.12, sử dụng thư viện Tkinter để tạo cửa sổ giao diện.

2.2. Đối tượng nghiên cứu

Đối tượng nghiên cứu là nước ép quả cam tươi từ các loài cam trồng ở các vùng của Việt Nam và các mẫu nước cam thương mại đang được bán trên thị trường Hà Nội.

Phương pháp lấy mẫu:

- *Với quả cam*: Cam được lấy mẫu thuộc các giống ở các tỉnh khác nhau. Các mẫu được ghi đầy đủ thông tin về giống cam, địa điểm trồng cam, mỗi địa điểm được lấy ở nhiều vụ khác nhau. Các mẫu được thu thập từ tháng 11/2022 đến tháng 1/2025, theo TCVN 9994 : 2013. Mẫu được bảo quản ở nhiệt độ từ 2-8°C trong vòng 7 ngày. Các mẫu được mã hóa với tên mẫu, thời gian và địa điểm thu hái được trình bày trong phụ lục 3. Lấy mẫu phải đảm bảo như sau:

Nguyên vẹn, lành lặn, không bị dập nát hoặc hư hỏng;

Sạch, hầu như không có bất kỳ tạp chất lạ nào nhìn thấy bằng mắt thường;

Hầu như không chứa côn trùng ảnh hưởng đến hình thức bên ngoài của sản phẩm;

Không có bất kỳ mùi và/hoặc vị lạ nào;

Hầu như không có vết thâm và/hoặc nhiều vết cắt đã lành.

- *Với mẫu nước cam thương mại*: Thu thập các mẫu nước cam đóng chai, lọ, lon của các thương hiệu khác nhau đang được bày bán tại các siêu thị trên địa bàn Hà Nội, Việt Nam. Các mẫu được ghi đầy đủ thông tin về số lô, hạn dùng của sản phẩm. Các mẫu đưa vào thí nghiệm là các mẫu còn hạn sử dụng và được bảo quản theo đúng điều kiện của nhà sản xuất.

Số lượng mẫu, thời gian lấy mẫu, xử lý mẫu và bảo quản mẫu được trình bày cụ thể ở từng nội dung tương ứng.

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

2.3.1.1. Định lượng các acid hữu cơ

a. Khảo sát điều kiện HPLC

Tiến hành khảo sát điều kiện HPLC được tiến hành dự kiến như sau [76, 108, 123]:

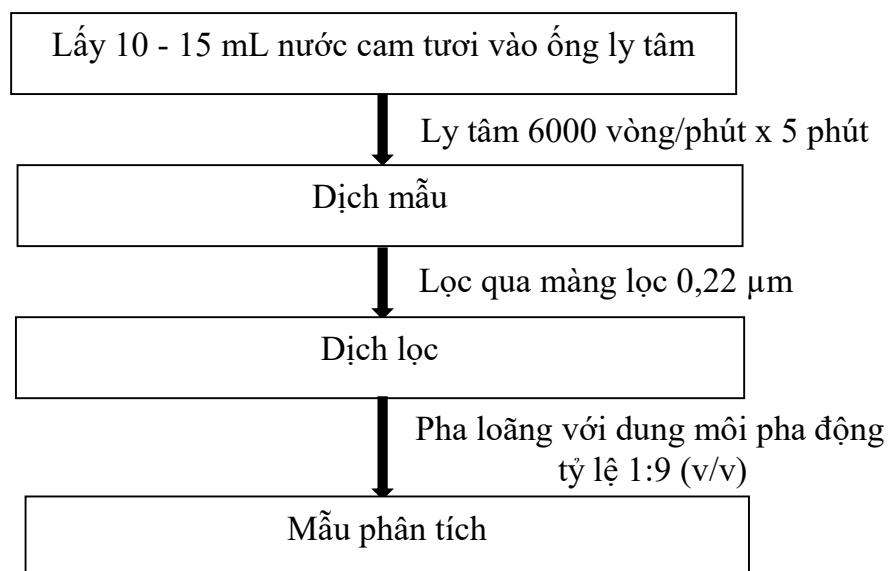
- Khảo sát pha động: Pha động được lựa chọn là dung dịch KH_2PO_4 với nồng độ 10mM, tiến hành khảo sát với dung dịch có pH khác nhau: 2; 2,2; 2,5; 3; 4 bằng cách dùng dung dịch acid H_3PO_4 10% để điều chỉnh đến các giá trị pH mong muốn.

- Khảo sát tốc độ dòng: Tiến hành với 3 tốc độ dòng: 0,6 ml/phút, 0,7 ml/phút và 0,8 ml/phút.

- Khảo sát pha tĩnh: Tiến hành với 2 loại cột: Cột C18, cột C610H Supelcogel

b. Phương pháp xử lý mẫu

Mẫu nước cam ép từ quả cam tươi, tiến hành quy trình xử lý mẫu theo sơ đồ dưới đây:



Hình 2.1. Sơ đồ xử lý mẫu định lượng acid hữu cơ

2.3.1.2. Định lượng các carotenoid

Về mặt hóa học, các hợp chất carotenoid đa số không tan trong nước do cấu trúc mạch carbon dài. Hơn nữa, hàm lượng các chất trong nước cam cũng không phải quá lớn, do vậy cần lựa chọn dung môi chiết và pha động có độ phân cực phù hợp để tăng hiệu suất thu hồi của phương pháp. Để tối ưu hóa phương pháp định lượng 3 carotenoid gồm lutein, lycopene, β -caroten, tiến hành khảo sát các điều kiện sau:

a. Khảo sát điều kiện HPLC

- Khảo sát bước sóng phát hiện: Quét phổ UV-Vis của 3 carotenoid trong khoảng 190 - 800 nm, chọn cực đại hấp thụ cho mỗi chất.

- Khảo sát pha tĩnh: Tiến hành khảo sát với 4 loại cột khác nhau:

+ (a) Symmetry® C18 (3,0 x 150 mm, 3,6 μ m);

+ (B) SunFire® C18 (4,6 x 250 mm, 5 μ m);

+ (c) Symmetry® C18 (6,0 x 250 mm, 5 μ m);

+ (d) XBridge® C18 (4,6 x 250 mm, 5 μ m).

- Khảo sát pha động: Lựa chọn khảo sát thành phần pha động với chương trình đẳng dòng, các dung môi pha động gồm:

+ MeOH;

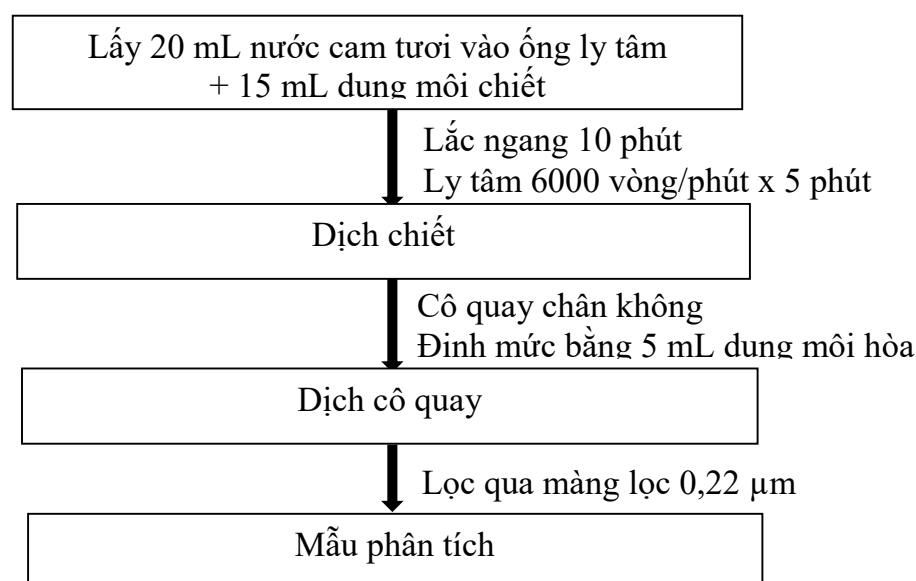
+ Hệ dung môi 1: ACN: MeOH: CH_2Cl_2 = 70: 27: 3 (v/v/v);

+ Hệ dung môi 2: ACN: MeOH: H_2O : $\text{CH}_3\text{COOC}_2\text{H}_5$ = 70: 9,6: 0,4: 20 (v/v/v/v).

- Khảo sát tốc độ dòng: Tiến hành với 3 tốc độ dòng 0,8 mL/phút; 1,0 mL/phút và 1,2 mL/phút.

b. Phương pháp xử lý mẫu

Mẫu nước cam ép từ quả cam tươi, tiến hành quy trình xử lý mẫu dự kiến theo sơ đồ dưới đây:



Hình 2.2. Dự kiến sơ đồ xử lý mẫu định lượng carotenoid

Các thông số khảo sát cụ thể như sau:

- Đo trực tiếp (không chiết);
- Khảo sát dung môi chiết: dichloromethan, n-hexan, ethyl acetat;
- Khảo sát dung môi hòa cần: MeOH hoặc dung môi pha động (ACN: MeOH: nước: ethyl acetat = 70: 9,6: 0,4: 20 (v/v/v/v));
- Khảo sát số lần chiết: 1 lần, 2 lần và 3 lần.

2.3.1.3. Định lượng các flavonoid

Về mặt hóa học, các hợp chất flavonoid có độ hòa tan tương đối phức tạp, có thể tan tốt nhưng cũng có những chất rất kém tan. Hơn nữa, hàm lượng các chất trong nước cam cũng không phải quá lớn, do vậy cần lựa chọn dung môi chiết và pha động có độ phân cực phù hợp để tăng hiệu suất thu hồi của phương pháp. Để tối ưu hóa phương pháp định lượng 4 flavonoid gồm naringin, hesperidin, neohesperidin; quercetin, tiến hành khảo sát các điều kiện sau:

a. Khảo sát điều kiện HPLC

- Khảo sát bước sóng phát hiện: quét phổ UV-Vis của 4 flavonoid trong khoảng 200 - 800 nm, chọn cực đại hấp thụ cho mỗi chất.
- Khảo sát pha tĩnh: Tiến hành khảo sát với 3 loại cột:
 - + Waters Bridge C₁₈ (4,6 mm x 150 mm; 5 μm);

+ Cột Sunfire TM C18 (4,6 mm x 150 mm; 5 μ m);

+ Cột Symmetry® C18 (4,6 mm x 150 mm; 5 μ m).

- Khảo sát pha động: hệ dung môi gồm Kênh A là acid phosphoric 0,1% và Kênh B là ACN với 3 chương trình đẳng dòng: (a) 100% B; (b) A/B = 70/30; (c) A/B = 50/50 và với 2 chương trình gradient như Bảng 2.1

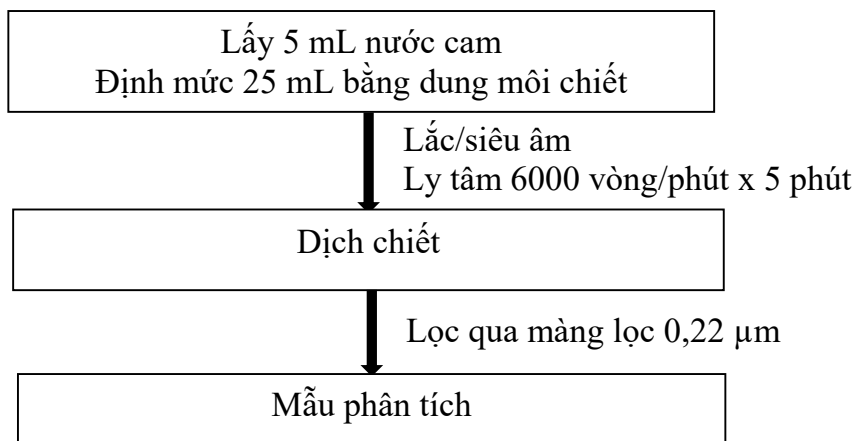
Bảng 2.1. Chương trình gradient pha động định lượng flavonoid

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
<i>Thời gian (phút)</i>	<i>Tốc độ dòng (mL/phút)</i>	<i>Kênh A(% , v : v)</i>	<i>Kênh B (% , v : v)</i>
22,00	1,00	0	100
22,10	1,00	90	10
26,00	1,00	90	10

- Khảo sát tốc độ dòng: Tiến hành với 3 tốc độ dòng: 0,8 mL/phút; 1,0 mL/phút và 1,2 mL/phút.

b. Khảo sát điều kiện xử lý mẫu

Mẫu nước cam ép từ quả cam tươi, tiến hành quy trình xử lý mẫu dự kiến theo sơ đồ dưới đây:



Hình 2.3. Dự kiến sơ đồ xử lý mẫu định lượng flavonoid

Các thông số khảo sát cụ thể như sau:

- Đo trực tiếp (không chiết);
- Khảo sát dung môi chiết: MeOHl, ethyl acetat, ACN;
- Khảo sát kỹ thuật hòa tan mẫu: lắc tay, lắc ngang, lắc vortex, rung siêu âm;
- Khảo sát thời gian hòa tan mẫu: 5, 10, 15 phút.

2.3.1.4. Xác nhận giá trị sử dụng của phương pháp phân tích

Phương pháp định lượng sau khi xây dựng được xác nhận giá trị sử dụng thông qua việc đánh giá các thông số sau [10, 20]:

- Độ ổn định của hệ thống sắc ký: được đánh giá qua độ lệch chuẩn tương đối (RSD) của thời gian lưu và diện tích peak với 6 lần tiêm lặp dung dịch chuẩn có nồng độ mỗi chất 20 mg/L vào hệ thống, yêu cầu RSD không quá 2%.

- Độ đặc hiệu của phương pháp: được đánh giá dựa trên thời gian lưu khi phân tích mẫu trắng (dung môi pha động; dung môi chiết mẫu), dung dịch chuẩn, mẫu thực và mẫu thực thêm chuẩn.

- Giới hạn phát hiện (MDL) và giới hạn định lượng (MQL) của phương pháp: được xác định qua phân tích lặp lại 10 lần mẫu thực có nồng độ thấp (trong khoảng 5-7 lần MDL ước lượng), tính giá trị trung bình và độ lệch chuẩn (SD) của nồng độ xác định được. MDL và MQL tương ứng bằng 3SD và 10SD, hệ số $R = \bar{x}/MDL$ phải trong khoảng từ 4 - 10.

- Độ chính xác của phương pháp: được đánh giá qua độ lặp lại (phân tích lặp 6 lần mẫu nước cam) với yêu cầu RSD không vượt quá 11 % và độ thu hồi được xác định bằng cách thêm 1 lượng chất chuẩn xác định ở 3 mức 50 %, 100 %, 200 % so với hàm lượng chất có trong mẫu nước cam, thực hiện lặp lại 12 lần vào mẫu thực với yêu cầu độ thu hồi tại mức hàm lượng 100 µg/L - 10 mg/L phải đạt từ 80 - 110 %.

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

2.3.2.1. Lấy mẫu và chuẩn bị mẫu

Tổng cộng 82 mẫu được thu thập từ 10 tỉnh thuộc 3 vùng miền Bắc, Trung và Nam Việt Nam. Cam được thu thập từ 4 giống gồm: cam Vinh (Bắc Giang: 6 mẫu; Hưng Yên: 16 mẫu; Quảng Ninh: 4 mẫu; Tuyên Quang: 10 mẫu), cam Cao Phong (Hòa Bình: 6 mẫu), cam Xã Đoài (Hà Giang: 5 mẫu; Nghệ An: 10 mẫu), cam Sành (Tây Ninh: 9 mẫu; Tuyên Quang: 2 mẫu; Trà Vinh: 4 mẫu; Vĩnh Long: 10 mẫu). Các mẫu cam được thu hái từ tháng 11 năm 2022 đến tháng 01 năm 2025 tại giai đoạn chín thương phẩm, dựa trên kinh nghiệm của người trồng. Sau khi đưa về phòng thí nghiệm, cam được rửa sạch, gọt vỏ và ép cẩn thận bằng máy ép để thu nước cam tươi. Bã tép và hạt được loại bỏ bằng rây/lọc. Nước cam tươi được ly tâm ở 3.000 vòng/phút (rpm) trong 10 phút và lọc qua màng polytetrafluoroethylene (PTFE) 0,45 µm để thu dịch nổi (supernatant). Mỗi mẫu dịch nổi được thu và chuẩn bị lặp lại 3 lần. Dịch nổi sau đó được bảo quản ở 2 - 8°C. Các dịch nổi này được phân tích các thành phần hóa học chính theo các quy trình tương ứng.

2.3.2.2. Phân tích các thành phần hóa học chính

Tiến hành phân tích hàm lượng acid, flavonoid, carotenoid theo các quy trình từ kết quả nghiên cứu phát triển phương pháp.

Hàm lượng đường được phân tích theo quy trình của Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia: mẫu nước cam tươi và nước cam thương mại đều được ly tâm ở 6000 vòng/phút trong 5 phút, sau đó lọc qua màng lọc 0,22 μm . Dịch lọc được pha loãng với pha động theo tỷ lệ 1:10 (v/v) trước khi tiến hành phân tích sắc ký. Hệ thống HPLC được trang bị detector khúc xạ RID, cột HILIC pak và SZ, thời gian phân tích lần lượt là 20 phút và 30 phút. Pha động gồm ACN : nước = 8: 2, chế độ đẳng dòng. Tốc độ dòng 0,8 mL/phút, nhiệt độ cột 60°C, thể tích bơm mẫu 10 μL .

Các kết quả phân tích được nhập vào Microsoft Excel để xử lý thống kê.

2.3.2.3. Xử lý dữ liệu

Ma trận dữ liệu đầu vào có kích thước 82 mẫu x 10 chỉ tiêu hàm lượng chất (fructose, glucose, sucrose, lutein, β -caroten, acid citric, acid ascorbic, acid fumaric, quercetin, hesperidin) được chia thành tập số liệu dùng để huấn luyện, kiểm tra và xác thực với tỷ lệ lần lượt là 70 : 20 : 10 (%).

Các thuật toán được sử dụng bao gồm: Máy vector hỗ trợ (Support Vector Machine - SVM), Cây quyết định (Decision Tree - DT), Rừng ngẫu nhiên (Random Forest), XGBoost, Mạng nơ-ron lan truyền đa lớp (Multi-Layer Perceptron - MLP). Lựa chọn thuật toán tối ưu dựa trên độ chính xác của từng thuật toán.

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

2.3.3.1. Lấy mẫu và chuẩn bị mẫu

Tổng cộng 111 mẫu cam cùng giống cam Vinh (*Citrus sinensis*) được thu hoạch tại ba tỉnh ở Việt Nam trong khoảng thời gian từ tháng 11/2023 đến tháng 11/2024: Hòa Bình (n = 54), Quảng Ninh (n = 24) và Tuyên Quang (n = 33). Cam được thu hái từ nhiều cây cam chín trong cùng một tỉnh, với nhiều đợt thu hoạch trong năm, cam thu hoạch ở mỗi vườn là một mẫu. Ban đầu, số lượng mẫu được thu thập tương đương giữa các tỉnh, tuy nhiên, nhiều mẫu không đạt yêu cầu bị loại bỏ dẫn tới sự mất cân đối số mẫu giữa các tỉnh.

Sau khi đưa về phòng thí nghiệm, cam được rửa sạch, gọt vỏ và ép cẩn thận bằng máy ép để thu nước cam tươi. Bã tép và hạt được loại bỏ bằng rây/lọc. Nước cam tươi được ly tâm ở 6000 vòng/phút trong 5 phút và lọc qua màng

polytetrafluoroethylene (PTFE) 0,45 μm để thu dịch nổi (supernatant). Mỗi mẫu dịch nổi được thu và chuẩn bị lặp lại 3 lần. Dịch nổi sau đó được bảo quản ở 2 - 8°C cho đến khi phân tích bằng phổ hấp thụ phân tử UV-Vis.

2.3.3.2. Quét phổ UV - Vis

Dịch nổi được pha loãng tỷ lệ 1:20 bằng hỗn hợp dung môi $\text{CH}_2\text{Cl}_2:\text{EtOH} = 2:3$ (v/v). Hệ dung môi chiết $\text{CH}_2\text{Cl}_2:\text{EtOH}$ (2:3, v/v) được lựa chọn nhằm tạo dung dịch ổn định, độ đục thấp, đồng thời thu nhận các chromophore có dải phân cực rộng trong nền mẫu cam phức tạp.

Mẫu sau xử lý được đo UV- Vis bằng cuvet thạch anh với bề dày 10 mm ở nhiệt độ phòng. Phép đo thực hiện ở chế độ quét đơn (single-scan) trong vùng bước sóng 250 nm đến 1,00 μm , với khoảng quét 5 nm. Độ rộng khe phổ (spectral bandwidth/slit width) được cài đặt 1,0 nm, tốc độ quét ở mức trung bình. Thiết bị sử dụng đèn deuteri (D_2) cho vùng tử ngoại và đèn tungsten-halogen (W) cho vùng khả kiến, với cơ chế tự động chuyển nguồn sáng tại 360 nm. Phổ hấp thụ thu được được xuất sang Microsoft Excel để xử lý thống kê.

2.3.3.3. Xử lý dữ liệu

❖ Tiền xử lý dữ liệu

Để tăng độ phân giải phổ và tối ưu độ chính xác dự đoán của các mô hình học có giám sát, nghiên cứu áp dụng các phương pháp tiền xử lý gồm: chuẩn hóa standard normal variate (SNV), lọc Savitzky-Golay (SG) và biến đổi đạo hàm bậc nhất (FD, $\text{Deriv} = 1$) nhằm loại bỏ biến thiên có hệ thống của đường nền.

❖ Phân tích thành phần chính (PCA)

Tập dữ liệu ban đầu dạng 111 mẫu $\times$ 161 bước sóng được tiền xử lý cho từng mẫu riêng lẻ và phổ trung bình của 3 lần lặp. Tập dữ liệu sau đó được phân tích thăm dò bằng PCA nhằm xác định các đặc trưng dữ liệu chính trước khi áp dụng các mô hình nhận dạng có giám sát.

❖ *Thuật toán học có giám sát và đánh giá hiệu năng mô hình*

Các mô hình được khảo sát gồm PCA-LDA, PCA-RF và PCA-ANN; để đơn giản, các mô hình này về sau được ký hiệu lần lượt là LDA, RF và ANN. Ngoài ra, các mô hình được tối ưu siêu tham số bằng GridSearch kết hợp xác thực chéo 5-fold nhằm giảm quá khớp và đánh giá minh bạch hiệu năng mô hình.

Để đánh giá hiệu quả phương pháp, các ma trận nhầm lẫn (confusion matrix) được xây dựng và hiệu năng mô hình được xác định thông qua các chỉ số: độ chính xác (A) hoặc tỷ lệ phân loại đúng (CCR, %), tỷ lệ sai (E), độ chính xác dương tính (precision, P), độ nhạy/thu hồi (recall/sensitivity, R), độ đặc hiệu (specificity, S) và điểm F1 (F1) [79, 80]. Nhìn chung, một mô hình phân loại tốt được đặc trưng bởi A, P, R, S, F1 cao và E thấp, và ngược lại [64]. Các tiêu chí này được tính theo các phương trình:

$$A = \frac{TP + TN}{Total} \cdot 100 = \frac{TP + TN}{TP + FP + FN + TN} \cdot 100(\%) \quad (2.1)$$

$$E = \frac{FP + FN}{Total} \cdot 100 = \frac{FP + FN}{TP + FP + FN + TN} \cdot 100 = 100 - A (\%) \quad (2.2)$$

$$P = \frac{TP}{Predicted\ positive} \cdot 100 = \frac{TP}{TP + FP} \cdot 100(\%) \quad (2.3)$$

$$R = \frac{TP}{Actual\ positive} \cdot 100 = \frac{TP}{TP + FN} \cdot 100 (\%) \quad (2.4)$$

$$S = \frac{TN}{Actual\ negative} \cdot 100 = \frac{TN}{TN + FP} \cdot 100 (\%) \quad (2.5)$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \cdot 100 (\%) \quad (2.6)$$

Trong đó TP, TN, FP và FN lần lượt là số lượng dương tính đúng, âm tính đúng, dương tính giả và âm tính giả thu được từ ma trận nhầm lẫn. Cụ thể, A biểu thị tỷ lệ tổng số mẫu được dự đoán đúng (TP + TN) trên tổng số mẫu (Phương trình 2.1). Ngược lại, E được tính là tỷ lệ mẫu bị phân loại sai (FP + FN) trên tổng số mẫu (Phương trình 2.2). P phản ánh tỷ lệ mẫu được gán đúng vào lớp thực (TP) trên tổng số mẫu được dự đoán thuộc lớp đó (TP + FP) (Phương trình 2.3). R là tỷ lệ TP, cho

biết phần trăm TP trên tổng số mẫu thuộc lớp dương tính thực (TP + FN) (Phương trình 2.4). S hay tỷ lệ TN thể hiện phần trăm mẫu được xác định đúng là không thuộc lớp tương ứng (TN) trên tổng số mẫu thuộc lớp âm tính thực (TN + FP) (Phương trình 2.5). Cuối cùng, F1 (Phương trình 2.6) là trung bình điều hòa giữa P và R, hữu ích khi đánh giá hiệu quả phân loại trong trường hợp tập dữ liệu bị mất cân bng.

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

2.3.4.1. Lấy mẫu và chuẩn bị mẫu

Tất cả các mẫu cam trong nghiên cứu thuộc loài *Citrus sinensis* và được thu thập trong năm 2024. Tổng cộng 80 mẫu cam được lấy từ 8 tỉnh gồm: Hòa Bình, Hà Giang, Bến Tre, Vĩnh Long, Nghệ An, Bắc Giang, Hưng Yên và Quảng Ninh. Sau khi đưa về phòng thí nghiệm, cam được rửa sạch, gọt vỏ và ép cẩn thận bằng máy ép để thu nước cam tươi. Bã tép và hạt được loại bỏ bằng rây/lọc.

Tổng 50 mẫu nước cam thương mại được thu thập từ các siêu thị tại Hà Nội, Việt Nam, bao gồm các sản phẩm đóng chai/đóng lon của một số nhãn hàng như Mirinda, Fanta, TH, Splash.

Trong nghiên cứu pha trộn (adulteration), nước cam tự nhiên nguyên chất được phối trộn với nước cam thương mại theo 9 mức pha trộn khác nhau tương ứng với tỷ lệ nước cam tự nhiên: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% và 90%. Mỗi mức pha trộn được chuẩn bị lặp lại 3 lần.

Các mẫu nước cam được ly tâm 10 phút ở 13.500 vòng/phút. Sau đó, mẫu được pha loãng theo tỷ lệ 1:4 bằng dung dịch HCl 0,1 M để điều chỉnh pH xấp xỉ 1,2. Mỗi mẫu được chuẩn bị lặp lại 3 lần.

2.3.4.2. Đo Volt-ampere

Mẫu được đo bằng A, điện cực so sánh Ag/AgCl và điện cực phụ trợ bằng platin. Với mỗi mẫu, thể tích sử dụng là 100 μ L (dạng giọt). Kỹ thuật đo xung vi phân (Differential Pulse Voltammetry - DPV) được áp dụng để ghi voltammogram với các

điều kiện thực nghiệm như sau: khoảng quét: $-0,2\text{ V}$ đến $+1,2\text{ V}$ (so với Ag/AgCl), biên độ xung: 50 mV , bước thế: 5 mV , tốc độ quét: 10 mV/s .

2.3.4.3. Xử lý dữ liệu

Để phân tích thăm dò, nghiên cứu sử dụng phân tích cụm (Cluster Analysis - CA). Dữ liệu thô của tất cả các mẫu được sử dụng, trong đó giá trị trung bình của 3 lần lặp cho mỗi mẫu được tính toán và đưa vào phân tích. Các thuật toán được sử dụng bao gồm: độ tương đồng cosin (Cosine similarity), chuẩn hóa SNV (Standard Normal Variate): Dữ liệu được chuẩn hóa theo phương pháp SNV nhằm loại bỏ biến thiên do hiệu ứng tán xạ và hiệu chỉnh dịch chuyển đường nền. Sau chuẩn hóa, dữ liệu tiếp tục được xử lý bằng bộ lọc Savitzky - Golay với độ dài cửa sổ 15, bậc đa thức 3 và tính đạo hàm bậc hai. Bước này giúp làm trơn dữ liệu, giảm nhiễu và làm nổi bật các đặc trưng phổ quan trọng phục vụ cho phân tích tiếp theo bằng iPLS, RF.

Để đánh giá hiệu quả phương pháp, ma trận nhầm lẫn (confusion matrix) được xây dựng và các chỉ số hiệu năng như độ chính xác (accuracy) được xác định. Độ chính xác được tính bằng tỷ lệ số mẫu được phân loại đúng trên tổng số mẫu, qua đó phản ánh trực quan hiệu quả phân loại của mô hình.

Chương 3. KẾT QUẢ VÀ THẢO LUẬN

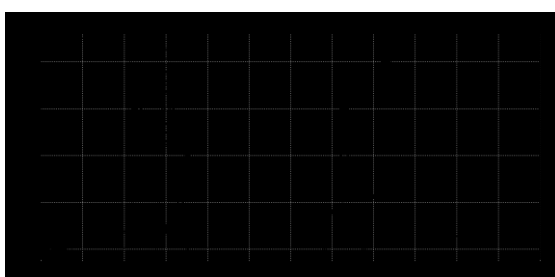
3.1. Nghiên cứu xây dựng và thẩm định phương pháp định lượng các thành phần hóa học chính có trong nước cam

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

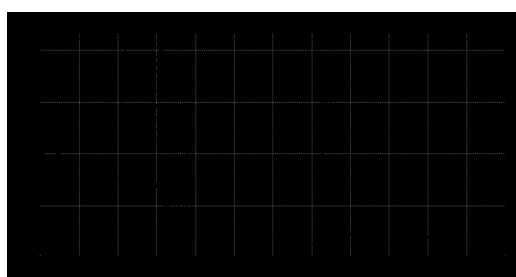
3.1.1.1. Khảo sát điều kiện tốt ưu của phương pháp phân tích

a. Khảo sát pha động

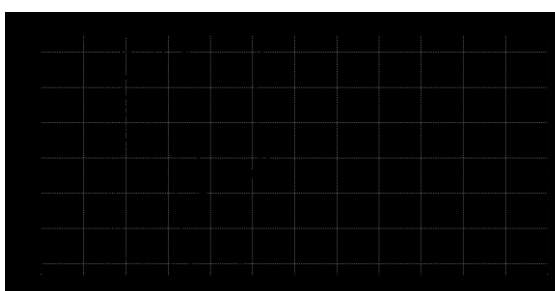
Để xác định hệ dung môi phù hợp nhất cho quá trình phân tích các acid, tiến hành khảo sát các pha động có pH khác nhau, sắc ký đồ được thể hiện ở hình 3.1.



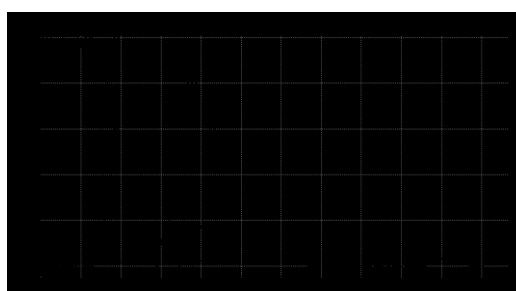
$pH = 2$



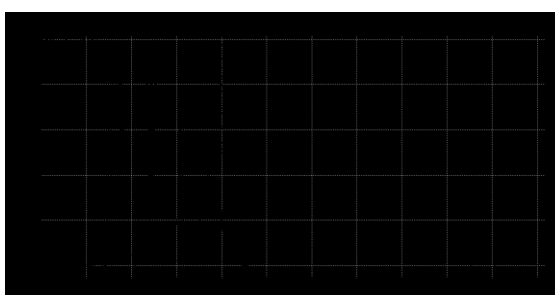
$pH = 2,2$



$pH = 2,5$



$pH = 3,0$



$pH = 4,0$

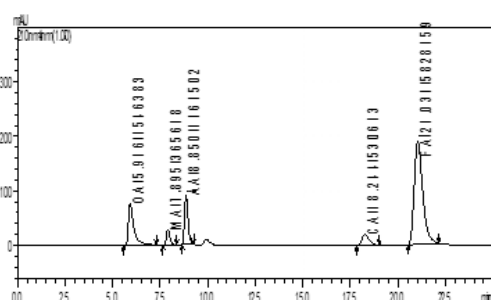
Hình 3.1. Sắc ký đồ 5 acid với dung môi pha động có pH khác nhau

Kết quả cho thấy pH ảnh hưởng nhiều đến việc tách 5 acid nghiên cứu. Ở pH 2,5; 3 và 4 tỏ ra thua kém trong việc phân tách các chất, peak phân tích của acid fumaric

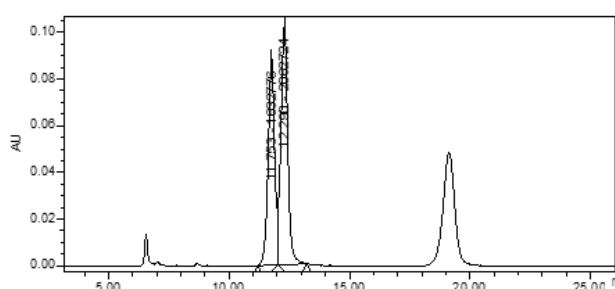
đã nhập với peak phân tích acid citric. Đối với pH 3 và 4 cho thấy không có độ đối xứng của peak của các chất phân tích. Ở điều kiện pH 2 và 2,2 cho thấy khả năng tách các chất hiệu quả nhất, tuy nhiên ở điều kiện pH 2, cột không bền nếu chạy thời gian dài. Như vậy, điều kiện pH 2,2 là tối ưu nhất.

b. Khảo sát pha tĩnh

Kết quả sắc ký đồ khi tách các acid trên 2 loại cột khảo sát được thể hiện ở hình 3.2.



Cột C18



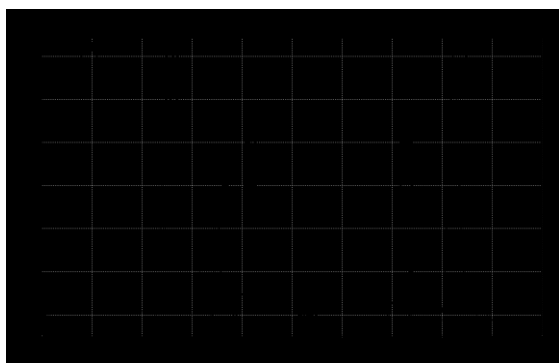
Cột C610H Supelcogel

Hình 3.2. Sắc ký đồ 5 acid trên 2 cột khảo sát

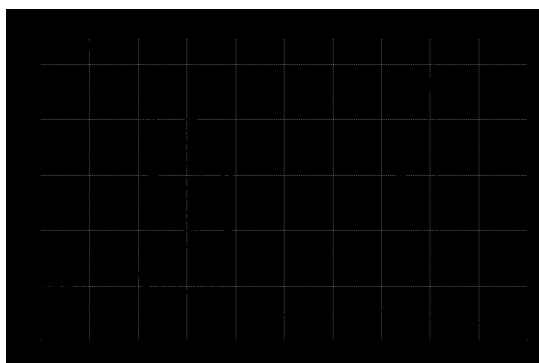
Trên cột C18 các acid tách nhau tốt đều có độ phân giải R(Resolution) lớn hơn 1,5, các peak nhọn, đối xứng và không dính nhau. Trong khi đó acid malic và acid ascorbic khi phân tích trên cột C610H Supelcogel lại có độ phân giải R nhỏ hơn 1,5, peak acid malic và acid ascorbic dính nhau, điều này cho thấy 2 acid này không tách tốt trên cột C610H Supelcogel

c. Khảo sát tốc độ dòng

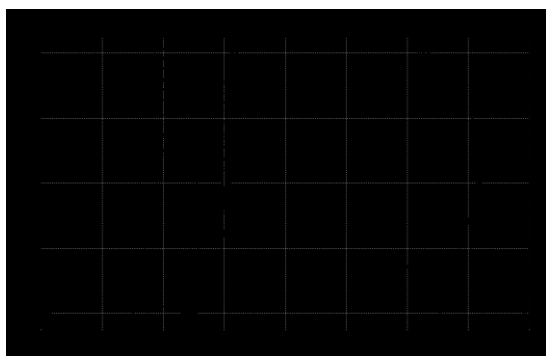
Tốc độ dòng pha động quyết định đến quá trình rửa giải các chất trong cột sắc ký vì nó ảnh hưởng đến khả năng thiết lập cân bằng của chất tan giữa pha tĩnh và pha động. Tốc độ pha động quá nhỏ sẽ gây doãng peak, thời gian lưu kéo dài làm tăng chi phí và thời gian phân tích, tốc độ pha động quá lớn làm giảm khả năng tách các chất dẫn tới các peak không tách nhau hoàn toàn. Tiến hành khảo sát với 3 tốc độ dòng, kết quả được thể hiện trên hình 3.3 và bảng 3.1.



Tốc độ dòng 0,6 ml/phút



Tốc độ dòng 0,7 ml/phút



Tốc độ dòng 0,8 ml/phút

Hình 3.3. Sắc ký đồ 5 acid với các tốc độ dòng khảo sát

Bảng 3.1. Các thông số khảo sát pha động

Tốc độ dòng	Chỉ tiêu	S/N	Hệ số đối xứng	Độ rộng chân peak
0,6	Acid oxalic	510,5512	1,1422	0,092
	Acid malic	573,1321	1,1131	0,096
	Acid ascorbic	560,1132	1,1012	0,095
	Acid citric	410,0214	1,1435	0,091
	Acid fumaric	408,7452	1,1080	1,019
0,7	Acid oxalic	498,2139	1,1630	0,110
	Acid malic	473,9076	1,2010	0,098
	Acid ascorbic	502,4412	1,2109	0,091
	Acid citric	420,7802	1,19812	0,103
	Acid fumaric	406,4216	1,12091	1,145

Tốc độ dòng	Chỉ tiêu	S/N	Hệ số đối xứng	Độ rộng chân peak
0,8	Acid oxalic	502,0981	1,1422	0,210
	Acid malic	473,9156	1,2131	0,096
	Acid ascorbic	460,3981	1,2012	0,095
	Acid citric	310,1209	1,1835	1,191
	Acid fumaric	308,6791	1,1980	1,221

Kết quả cho thấy, với tốc độ dòng 0,6ml/phút các chất phân tách khá tốt, ở tốc độ dòng 0,7ml/phút và 0,8 ml/phút peak của 2 acid citric và fumaric không cân đối và dính peak. Tại tốc độ dòng 0,6 mL/phút tỷ số S/N lớn, hệ số đối xứng gần 1 và chân peak hẹp. Do đó, tốc độ dòng 0,6ml/phút được lựa chọn là tốc độ dòng tối ưu.

Như vậy, điều kiện tối ưu để định lượng 5 acid trong nước cam bằng phương pháp HPLC là:

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

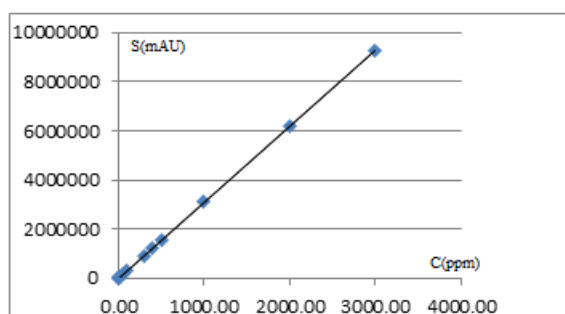
d. Khoảng tuyến tính và đường chuẩn

Sau khi đã tối ưu hóa các điều kiện phân tích, phương trình đường chuẩn và hệ số tương quan của các acid được trình bày trong bảng 3.2. và hình 3.4.

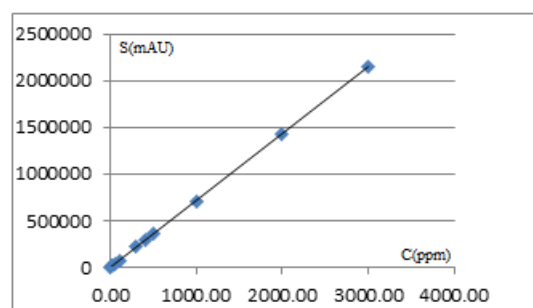
Bảng 3.2. Khoảng tuyến tính và đường chuẩn của các chất phân tích

Chất phân tích	Khoảng tuyến tính	Đường chuẩn
Acid oxalic	5 - 3000 ppm	$y = 3085x - 1399,3$ $R^2 = 1$
Acid malic	5 - 3000 ppm	$y = 714,38x + 828,12$ $R^2 = 0,9999$

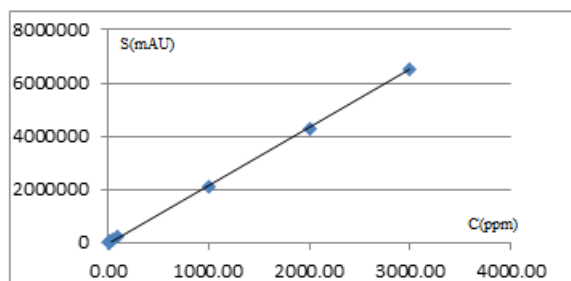
Chất phân tích	Khoảng tuyến tính	Đường chuẩn
Acid ascorbic	5 - 3000 ppm	$y = 2167,1x - 2246$ $R^2 = 0,9999$
Acid citric	5 - 3000 ppm	$y = 1065,6x - 3,3589$ $R^2 = 0,9999$
Acid fumaric	5-1000ppm	$y = 10558x - 11333$ $R^2 = 0,9997$



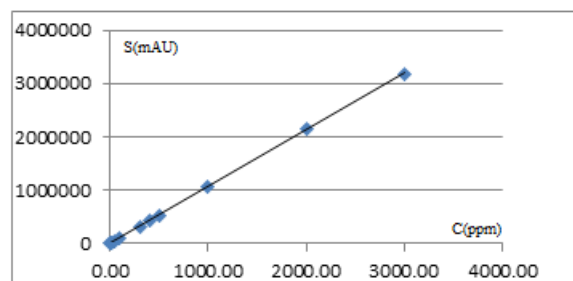
Oxalic



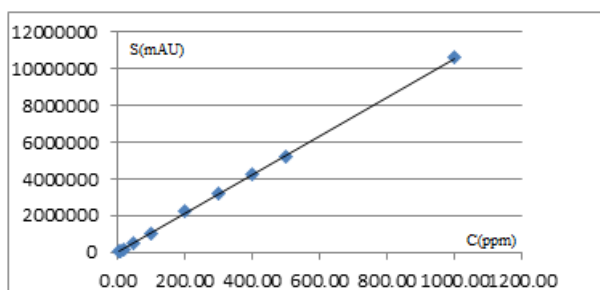
Malic



Ascorbic



Citric



Fumaric

Hình 3.4. Đường chuẩn của 5 acid nghiên cứu

Kết quả cho thấy khoảng tuyến tính của 4 acid: acid oxalic, acid malic, acid ascorbic, acid citric trong khoảng nồng độ từ 5 - 3000 ppm và acid fumaric có khoảng tuyến tính 5-1000ppm, các phương trình tuyến tính đều có hệ số tương quan $R^2 > 0,99$, đạt yêu cầu của AOAC.

3.1.1.2. Thẩm định quy trình phân tích

a. Độ ổn định của hệ thống

Độ ổn định hệ thống được xác định bằng cách thực hiện tiêm lặp lại 6 lần dung dịch mẫu chuẩn có nồng độ 20 µg/mL.

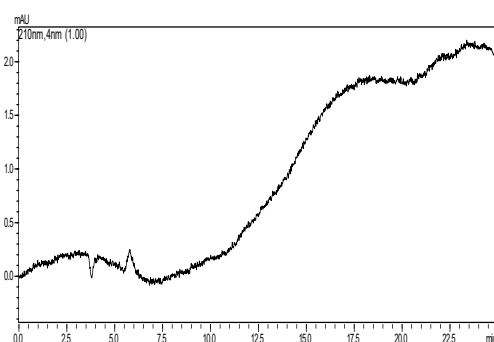
Bảng 3.3. Kết quả thẩm định độ ổn định hệ thống

Chất phân tích	RSD _{IR} (%)	RSD _{Peak} (%)
Acid oxalic	0,11	0,87
Acid malic	0,64	1,03
Acid ascorbic	0,09	0,75
Acid citric	0,87	0,09
Acid fumaric	1,02	0,83

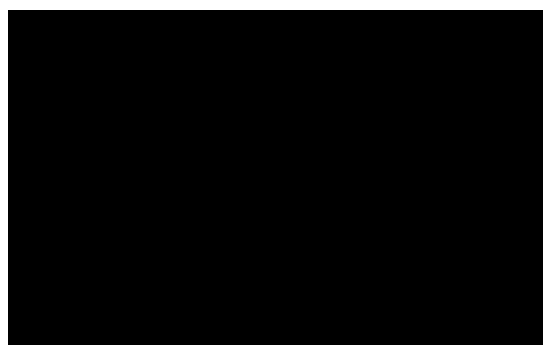
Kết quả thẩm định độ ổn định của hệ thống sắc ký được thể hiện tại Bảng 3.3. Độ lệch chuẩn tương đối của thời gian lưu và diện tích peak đối với 5 chất đều nhỏ hơn 2%, đáp ứng theo yêu cầu của AOAC. Như vậy, hệ thống sắc ký đáp ứng yêu cầu về độ ổn định.

b. Độ đặc hiệu

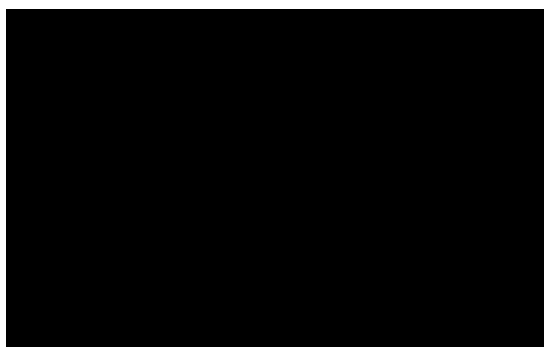
Tiến hành phân tích mẫu trắng, mẫu chuẩn và mẫu thử thêm chuẩn. Các sắc ký đồ được thể hiện trên hình 3.5.



Mẫu trắng



Mẫu chuẩn



Mẫu thử thêm chuẩn

Hình 3.5. Sắc ký đồ khảo sát 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. Như vậy, phương pháp có độ đặc hiệu tốt với 5 acid nghiên cứu.

c. MDL và MQL

Các kết quả xác định MDL và MQL của các acid được trình bày ở phụ lục 2.1. Kết quả tổng hợp được trình bày ở bảng 3.4.

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 (mg /L)	0,60	0,9	3,9	14,4	0,15
MQL (mg /L)	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 $4 < R < 10$, như vậy các giá trị R đều nằm trong khoảng cho phép, từ đó đánh giá kết quả MDL và MQL của các chất phân tích là phù hợp

d. Độ lặp lại và độ thu hồi

Các kết quả xác định độ lặp lại và độ thu hồi của các acid được trình bày ở phụ lục 2.1. Kết quả tổng hợp được trình bày ở bảng 3.5 và 3.6.

Bảng 3.5. Kết quả độ lặp lại của 5 acid

Chất phân tích	Acid oxalic	Acid malic	Acid ascorbic	Acid citric	Acid fumaric
Hàm lượng trung bình (mg/L)	123,5	712,8	1361,6	16054	31,9
RSD	3,0	1,7	3,5	3,5	1,3

Kết quả cho thấy, khi phân tích hàm lượng các acid trong nước cam đều có giá trị RSD < 5,3%, đáp ứng đúng yêu cầu của AOAC. Như vậy, phương pháp có độ lặp lại đạt yêu cầu.

Bảng 3.6. Kết quả độ thu hồi của 5 acid

Chất phân tích		Acid oxalic	Acid malic	Acid ascorbic	Acid citric	Acid fumaric
Thêm chuẩn mức thấp	Hiệu suất (%)	97,4 ÷ 102,4	101,4 ÷ 102,6	101,8 ÷ 104,0	101,2 ÷ 103,6	98,4 ÷ 100,1
	RSD (%)	2,3	0,6	1,0	1,1	0,8
Thêm chuẩn mức trung bình	Hiệu suất (%)	97,3 ÷ 99,1	97,9 ÷ 9,6	102,5 ÷ 104,1	103,4 ÷ 104,0	97,8 ÷ 102,8
	RSD (%)	0,9	0,8	0,7	0,3	2,2
Thêm chuẩn mức cao	Hiệu suất (%)	97,8 ÷ 102,3	97,3 ÷ 99,1	99,5 ÷ 104,0	97,3 ÷ 103,8	97,7 ÷ 99,7
	RSD (%)	2,1	0,8	2,1	2,9	0,9

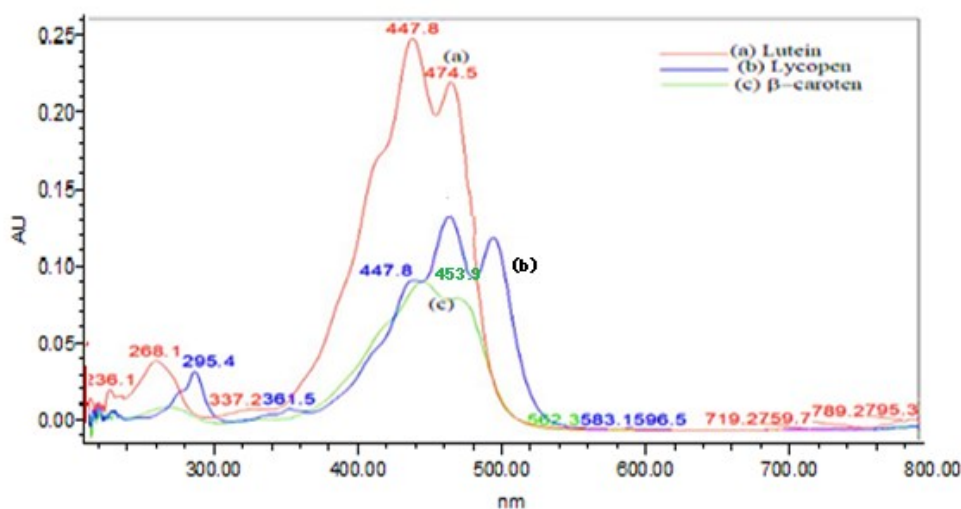
Hướng dẫn của AOAC về độ thu hồi chấp nhận ở mức nồng độ 100 ppm là 90 - 107 %. Kết quả thu được cho thấy độ thu hồi và RSD đều nằm trong khoảng cho phép, từ đó cho thấy phương pháp có độ thu hồi phù hợp.

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

a. Lựa chọn bước sóng phát hiện

Để lựa chọn bước sóng thích hợp ghi sắc ký đồ khi phân tích đồng thời lutein, lycopene và β -caroten bằng detector PDA, quét phổ UV-Vis trong khoảng 190 - 800 nm, chọn cực đại hấp thụ cho mỗi chất. Kết quả được thể hiện trên hình 3.6.

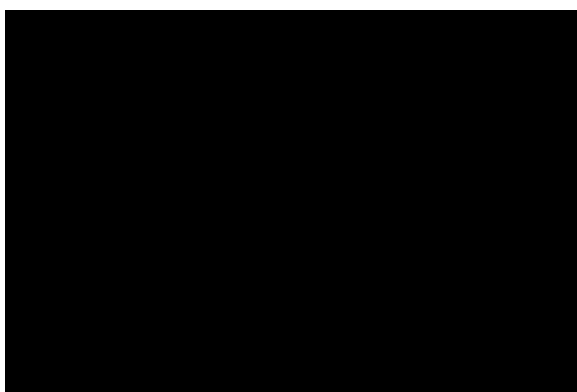


Hình 3.6. Phổ UV-Vis của lutein (a), lycopene (b), β -caroten (c)

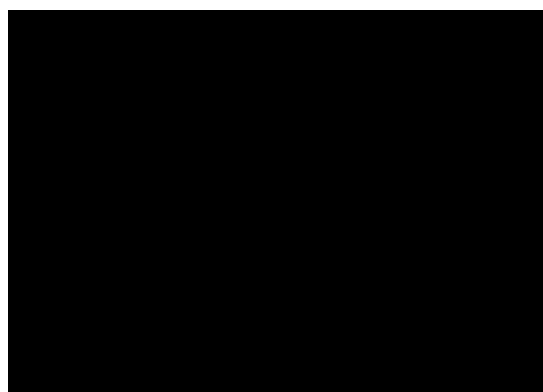
Phổ hấp thụ phân tử UV-Vis của 3 chất phân tích cho thấy lutein, lycopene và β -caroten có cực đại hấp thụ chính ở các bước sóng tương ứng là 447,8; 474,5 và 453,9 nm. Để giảm ảnh hưởng của dung môi và tạp chất nhưng vẫn đảm bảo được độ nhạy của phương pháp khi phân tích đồng thời cả 3 carotenoid bước sóng 455 nm được chọn.

b. Khảo sát pha tĩnh

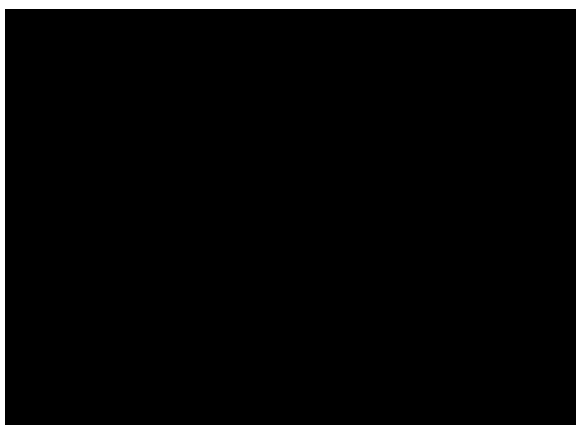
Pha tĩnh ảnh hưởng rất lớn đến khả năng tách của các chất, tiến hành thí nghiệm với 4 loại cột sắc ký C18 khác nhau, kết quả thu được như Hình 3.7.



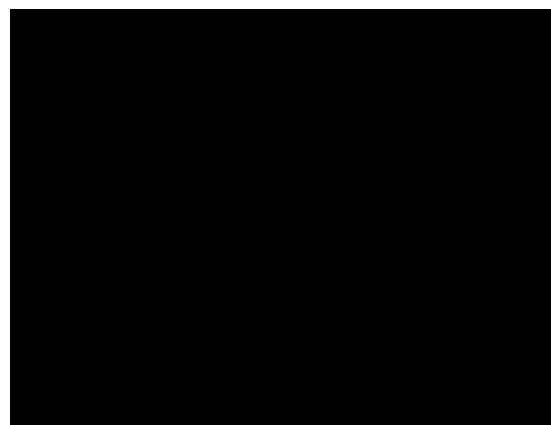
Cột Symmetry® C18 150



Cột SunFire® C18



Cột Symmetry® C18 250



Cột XBridge® C18

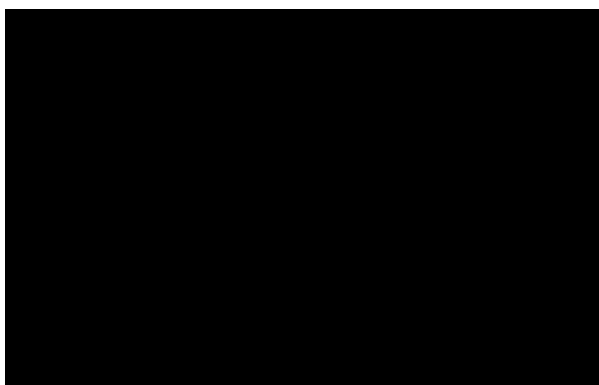
Hình 3.7. Sắc ký đồ 3 carotenoid với 4 cột pha tĩnh

Trên các sắc ký đồ thấy rằng khi sử dụng cột Symmetry® C18 150 thời gian lưu của các carotenoid ra quá sớm, cột SunFire® C18 thời gian lưu dài so với cột Symmetry® C18 250 và cột XBridge® C18. Cột Symmetry® C18 250 và cột XBridge® C18 thời gian lưu tương đương nhau, tuy nhiên, peak lutein khi dùng của cột Symmetry® C18 250 dính tạp còn khi dùng cột XBridge® C18 các peak hoàn toàn tách ra khỏi nhau.

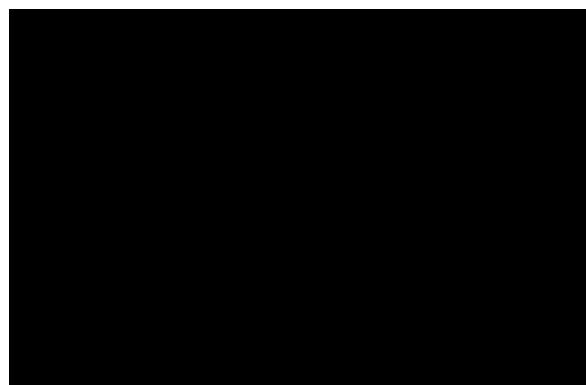
Do đó, cột XBridge® C18, (4,6 x 250 mm, 5 µm) được lựa chọn sử dụng trong nghiên cứu này.

c. Khảo sát điều kiện pha động

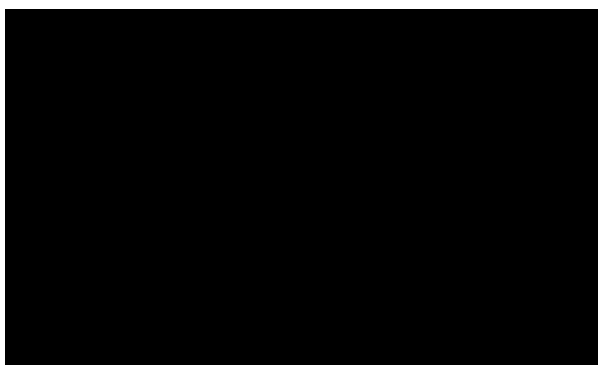
Với 3 hệ dung môi pha động nghiên cứu, khi sử dụng chế độ pha động đẳng dòng đã tách hoàn toàn được 3 chất phân tích ra khỏi nhau (Hình 3.8) và ra khỏi các thành phần nền mẫu khác, đường nền ổn định và thời gian lưu phù hợp.



Methanol



Hệ dung môi 1



Hệ dung môi 2

Hình 3.8. Sắc ký đồ 3 carotenoid với các hệ pha động

Với dung môi methanol peak thu được có tín hiệu thấp, chân peak doãng, hệ dung môi 1 cho thời gian lưu quá ngắn. Với hệ dung môi 2 thời gian lưu vừa phải, peak nhọn, tín hiệu tốt hơn nhiều so với methanol và hệ dung môi 1. Do vậy, hệ dung môi 2 có thành phần ACN: MeOH: H₂O: CH₃COOC₂H₅ = 70: 9,6: 0,4: 20 (v/v/v/v) được lựa chọn cho nghiên cứu này.

d. Lựa chọn tốc độ dòng pha động

Tốc độ dòng quyết định đến thời gian lưu của các chất cũng như tuổi thọ của cột, vì vậy nghiên cứu thực hiện khảo sát với các tốc độ dòng 0,8; 1,0 và 1,2 mL/phút.

Bảng 3.7. Các thông số khảo sát tách 3 carotenoid với các tốc độ dòng

<i>Tốc độ dòng (mL/phút)</i>	<i>Chất phân tích</i>	<i>Thời gian lưu (phút)</i>	<i>S/N</i>	<i>Hệ số đối xứng</i>	<i>Độ rộng chân peak</i>
0,8	Lutein	4,356	389,309	1,336	0,136
	Lycopene	9,014	144,351	1,497	0,297
	β-caroten	14,389	177,661	1,212	0,261
1,0	Lutein	3,704	463,320	1,275	0,086
	Lycopene	8,279	170,80	1,289	0,266
	β-caroten	13,060	205,80	1,049	0,183
1,2	Lutein	3,062	377,700	1,280	0,088
	Lycopene	6,816	140,327	1,337	0,229
	β-caroten	10,759	132,100	1,063	0,185

Tốc độ dòng 1,2 mL/phút diện tích peak của 3 carotenoid thấp, làm giảm độ nhạy của phương pháp. Tốc độ 0,8 mL/phút và 1,0 mL/phút cho thời gian lưu vừa phải, diện tích peak lớn. Tuy nhiên, tốc độ dòng 1,0 mL/phút tỉ lệ S/N lớn hơn, hệ số đối xứng gần 1 hơn và độ rộng chân peak nhỏ hơn của cả 3 chất so với tốc độ dòng 0,8 mL/phút.

Do đó, để thu được sắc ký đồ với các peak thon gọn, không dính tạp, thời gian phân tích vừa phải, tốc độ 1,0 mL/phút được lựa chọn.

Như vậy, sau khi tiến hành khảo sát điều kiện phân tích đồng thời β -caroten, lycopene và lutein trong nước cam bằng HPLC cụ thể như sau:

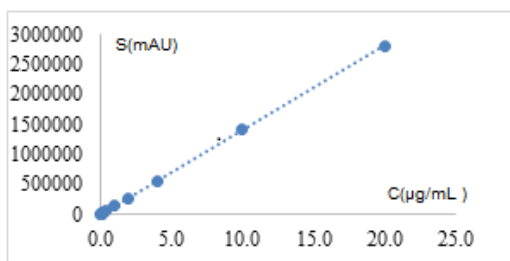
- Cột sắc ký: Cột XBridge® C18 (4,6 x 250 mm, 5 μ m) và tiền cột tương ứng.
- Pha động: ACN: MeOH: H₂O: CH₃COOC₂H₅ = 70: 9,6: 0,4: 20 (v/v/v/v).
- Tốc độ dòng: 1,0 mL/phút.
- Thời gian phân tích: 20 phút.
- Thể tích tiêm mẫu: 10 μ L.
- Detector PDA.
- Bước sóng phát hiện: 455 nm.

e. Xây dựng đường chuẩn

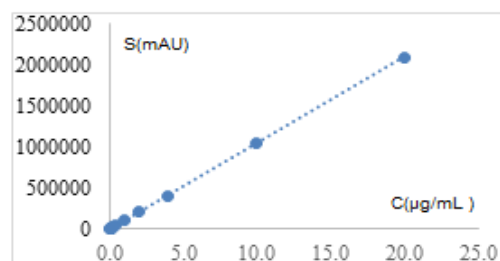
Tiến hành đo dung dịch chuẩn có nồng độ thay đổi từ 0,025 μ g/mL đến 20 μ g/mL với cả 3 chất. Xây dựng mối tương quan giữa nồng độ chất phân tích và diện tích peak tương ứng thu được trên sắc ký đồ bằng phương pháp hồi quy tuyến tính. Kết quả được thể hiện ở bảng 3.8 và hình 3.9.

Bảng 3.8. Khoảng tuyến tính và đường chuẩn của 3 carotenoid

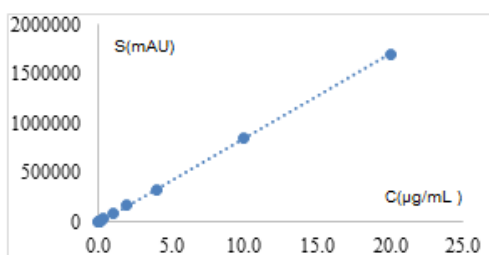
Chất phân tích	Khoảng tuyến tính	Đường chuẩn	R ²
Lutein	0,025 - 20 μ g/mL	$y = 140541x - 2036,8$	1,0
Lycopene	0,025 - 20 μ g/mL	$y = 104597x - 1205,9$	1,0
β -caroten	0,025 - 20 μ g/mL	$y = 85195x - 797,64$	0,9999



Lutein



Lycopene



β - caroten

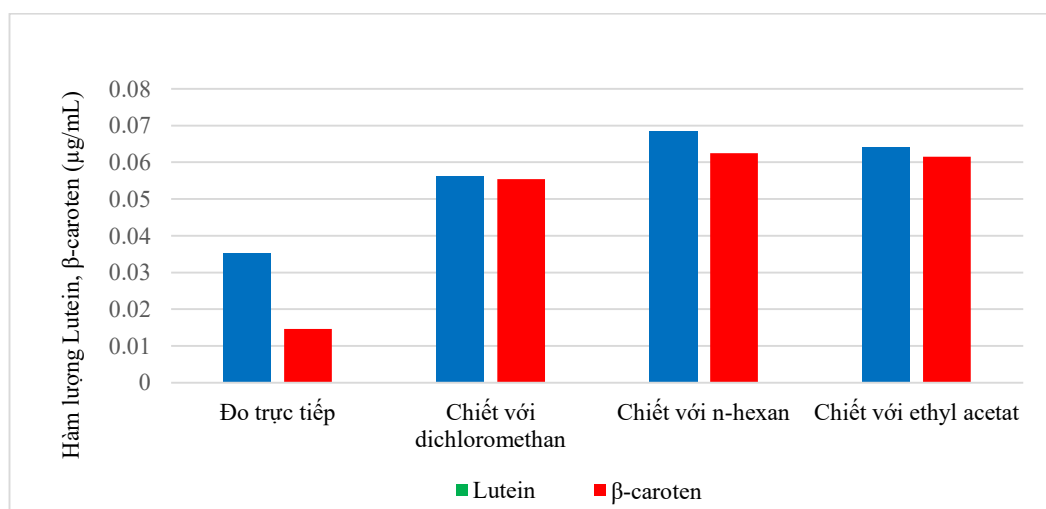
Hình 3.9. Đường chuẩn của 3 carotenoid

Kết quả cho thấy đường chuẩn của cả 3 chất đều có hệ số tương quan $R^2 > 0,9997$, đạt yêu cầu của AOAC.

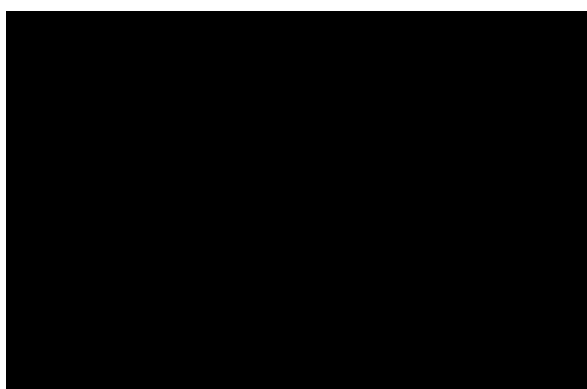
3.1.2.2. Xây dựng quy trình xử lý mẫu

a. Khảo sát lựa chọn dung môi chiết

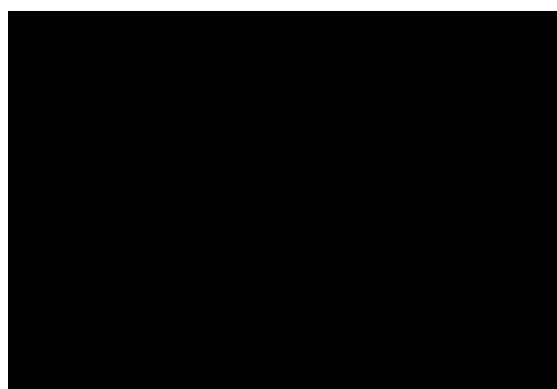
Nghiên cứu tiến hành đo trực tiếp và chiết mẫu với 3 dung môi ethylacetat, n-hexan và dichloromethan. Kết quả thể hiện trên hình 3.10 và hình 3.11.



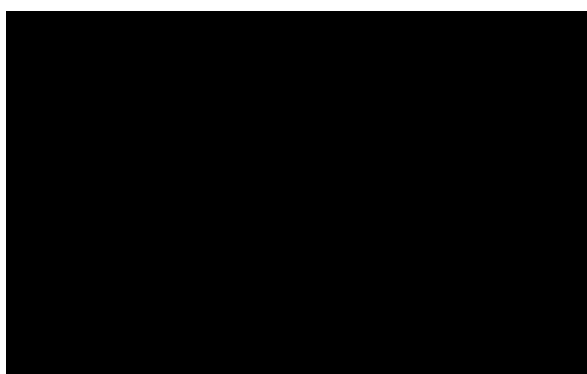
Hình 3.10. Kết quả khảo sát ảnh hưởng của dung môi chiết mẫu



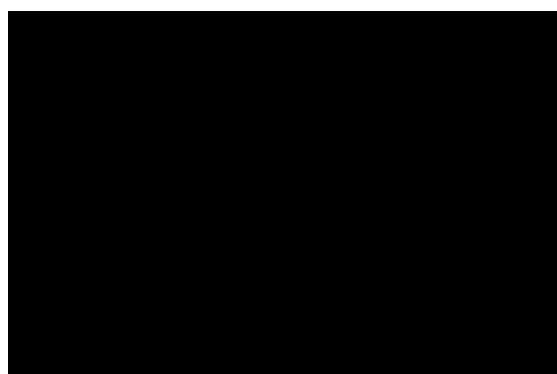
Đo mẫu trực tiếp



Chiết với dichloromethan



Chiết với n-hexan



Chiết với ethyl acetat

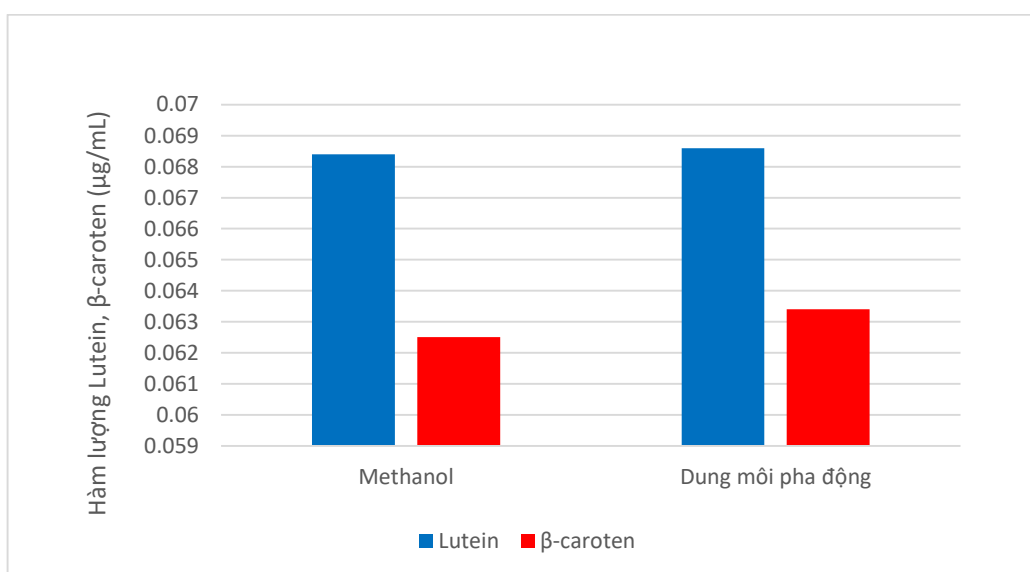
Hình 3.11. Sắc ký đồ khảo sát dung môi chiết

Kết quả khảo sát cho thấy trong nước cam không phát hiện được lycopene, hai chất còn lại tách tốt ra khỏi nhau. Tuy nhiên, hàm lượng của lutein và β -caroten thu được khi sử dụng dung môi chiết n-hexan lớn nhất, khi đo trực tiếp hàm lượng lutein và β -caroten thu được rất thấp so với các dung môi khác (chỉ đạt khoảng 30 - 70% so với hàm lượng chiết bằng dung môi n-hexan).

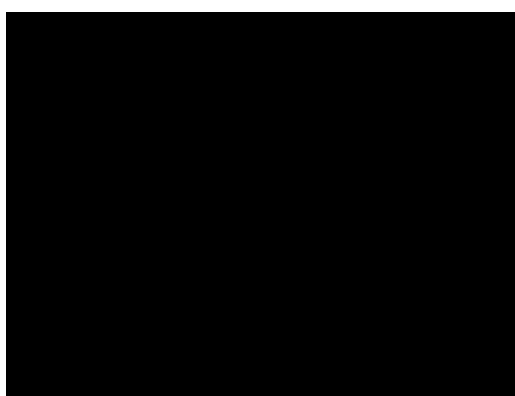
Do đó, dung môi n-hexan được lựa chọn để thực hiện các khảo sát tiếp theo.

b. Khảo sát lựa chọn dung môi hòa cần

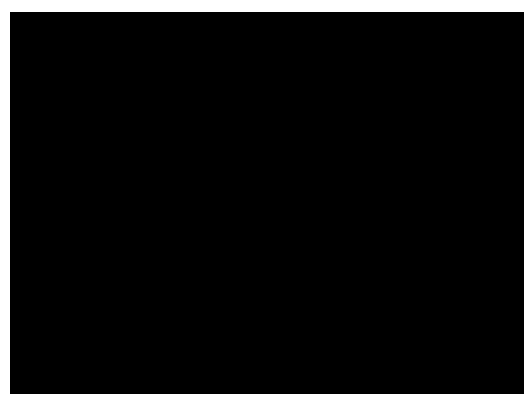
Việc lựa chọn dung môi hòa cần rất quan trọng vì nó ảnh hưởng tới khả năng hòa tan của chất phân tích vào dung môi đó. Tiến hành khảo sát với hai hệ dung môi là methanol và dung môi pha động.



Hình 3.12. Kết quả khảo sát dung môi hòa tan



Hòa tan với methanol



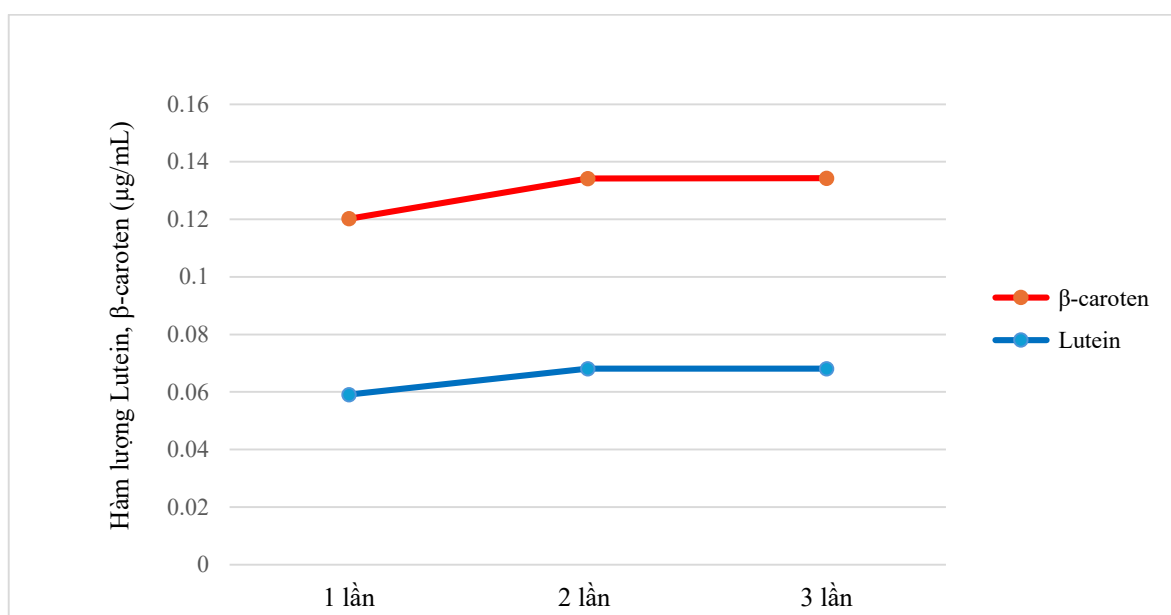
Hòa tan với dung môi pha động

Hình 3.13. Sắc ký đồ dung môi hòa tan

Với cả hai loại dung môi lutein và β-caroten đều tách tốt ra khỏi nhau, peak đẹp. Tuy nhiên, hàm lượng lutein và β-caroten thu được khi sử dụng dung môi pha động để hòa tan đều cao hơn khi sử dụng methanol. Vì vậy, hệ dung môi pha động gồm acetonitril: methanol: nước: ethyl acetat = 70: 9,6: 0,4: 20 (v/v/v/v)) được sử dụng làm dung môi hòa tan.

c. Khảo sát số lần chiết

Chiết lặp lại là một cách để đảm bảo chất phân tích được chiết tối đa ra khỏi nền mẫu. Nghiên cứu này đã tiến hành khảo sát chiết 1, 2, 3 lần.

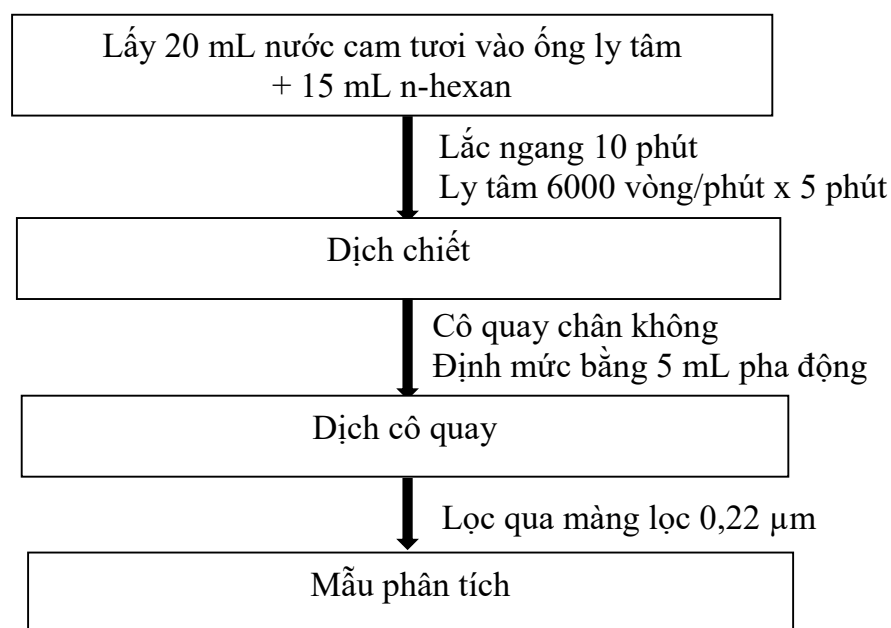


Hình 3.14. Kết quả khảo sát số lần chiết lặp lại

Khi số lần chiết tăng lên, hàm lượng cả 2 chất có mặt trong mẫu nước cam tự nhiên là lutein và β-caroten đều cao hơn hơn chiết 1 lần. Hàm lượng chất chiết được sau khi chiết 2 lần và 3 lần khác biệt không đáng kể. Vì vậy, để kết quả chiết mẫu đảm bảo đồng thời tiết kiệm thời gian và dung môi, nghiên cứu lựa chọn chiết 2 lần.

Sau khi khảo sát các điều kiện tối ưu của 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.



Hình 3.15. Quy trình xử lý mẫu

3.1.2.3. Thẩm định quy trình phân tích

a. Độ ổn định của hệ thống

Độ ổn định hệ thống được xác định bằng cách thực hiện tiêm lặp lại 6 lần dung dịch mẫu chuẩn có nồng độ 20 µg/mL.

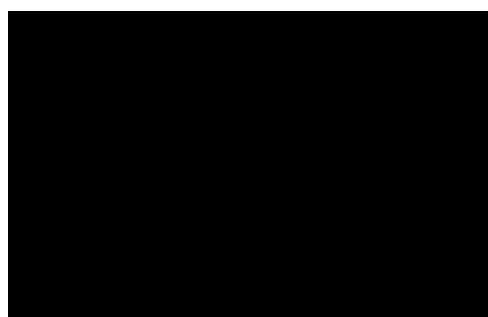
Bảng 3.9. Kết quả thẩm định độ ổn định hệ thống

Chất phân tích	RSD _{IR} (%)	RSD _{Speak} (%)
Lutein	0,06	0,76
Lycopene	0,05	1,53
β-caroten	0,55	0,58

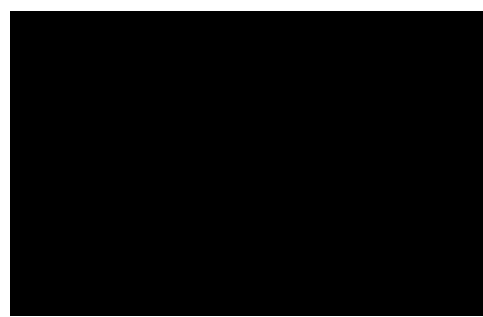
Độ lệch chuẩn tương đối của thời gian lưu và diện tích peak đối với 3 chất đều nhỏ hơn 2%, đáp ứng theo yêu cầu của AOAC. Như vậy, hệ thống sắc ký đáp ứng yêu cầu về độ ổn định.

b. Độ đặc hiệu

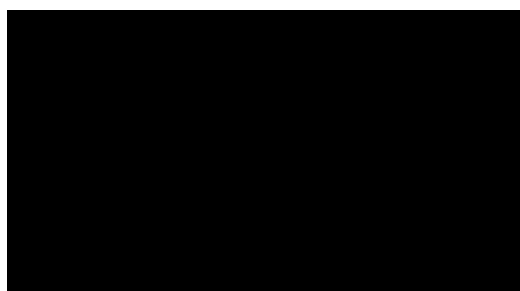
Độ đặc hiệu của phương pháp được đánh giá thông qua việc phân tích mẫu trắng, mẫu chuẩn, mẫu thực và mẫu thực thêm chuẩn.



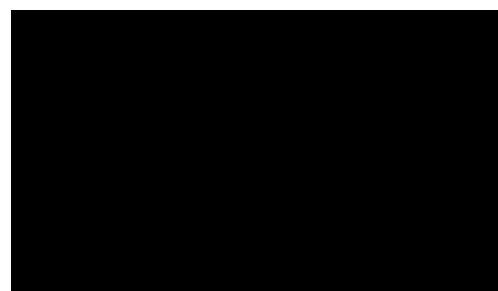
Mẫu trắng



Mẫu chuẩn



Mẫu thực



Mẫu thực thêm chuẩn

Hình 3.16. Sắc ký đồ khảo sát 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%. Như vậy phương pháp có độ đặc hiệu tốt.

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

Các kết quả xác định MDL và MQL của các carotenoid được trình bày ở phụ lục 2.2. Kết quả tổng hợp được trình bày ở bảng 3.10.

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 $4 < R < 10$, như vậy các giá trị R đều nằm trong khoảng cho phép, từ đó đánh giá kết quả MDL và MQL đáng tin cậy.

e. Độ lặp lại và độ thu hồi

Khi phân tích lặp lại 6 lần mẫu nước cam Vinh phát hiện được lutein và β -caroten nhưng không phát hiện được lycopene. Vì vậy nghiên cứu thêm một lượng chuẩn (0,5 $\mu\text{g/mL}$) của lycopene vào mẫu thực để đánh giá độ thu hồi ở mức lặp lại của lycopene.

Bảng 3.11. Kết quả độ lặp lại của 3 carotenoid

Chất phân tích	Lutein	Lycopene	β -caroten
Hàm lượng trung bình ($\mu\text{g/mL}$)	0,047	0,025	0,147
RSD (%)	2,4	1,3	0,3

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

Độ thu hồi được thực hiện bằng cách thêm 1 lượng chuẩn xác định ở 3 mức nồng độ thấp, trung bình và cao, mỗi mức thực hiện thêm chuẩn 3 lần.

Bảng 3.12. Kết quả độ thu hồi của 3 carotenoid

Chất phân tích		Lutein	Lycopene	β -caroten
Thêm chuẩn mức thấp	Hiệu suất (%)	90,8 $\div$ 93,6	81,0 $\div$ 87,1	95,4 $\div$ 96,8
	RSD (%)	0,35	0,73	0,18
Thêm chuẩn mức trung bình	Hiệu suất (%)	80,4 $\div$ 81,6	87,9 $\div$ 90,5	95,7 $\div$ 97,1
	RSD (%)	0,33	0,78	0,36
Thêm chuẩn mức cao	Hiệu suất (%)	103 $\div$ 106	99,9 $\div$ 105	83,7 $\div$ 86,9
	RSD (%)	1,12	1,62	1,20

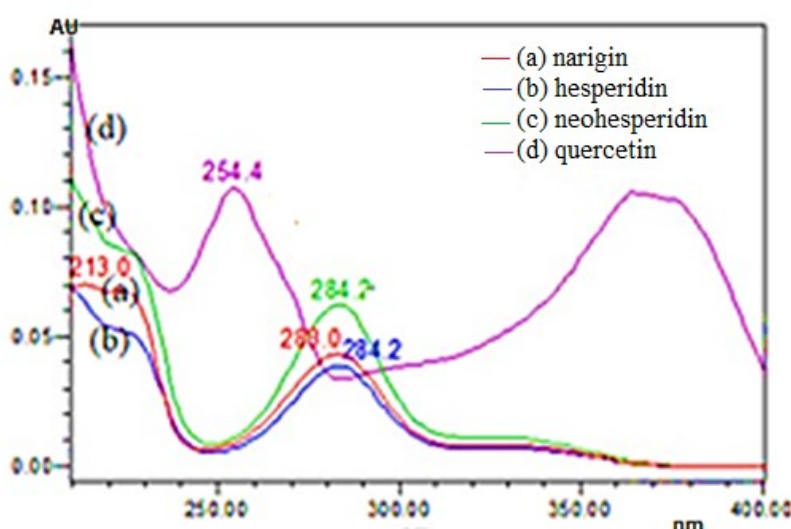
Theo AOAC, yêu cầu độ thu hồi tại mức hàm lượng 100 ppb - 10 ppm phải đạt từ 80 - 110 % và độ lệch chuẩn tương đối không vượt quá 2%. Kết quả thẩm định của cả 3 chất đều có hiệu suất thu hồi và RSD nằm trong giới hạn cho phép. Điều này chứng minh phương pháp đáp ứng được yêu cầu theo tiêu chuẩn của AOAC về độ thu hồi.

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

a. Lựa chọn bước sóng phát hiện

Để lựa chọn bước sóng thích hợp ghi sắc ký đồ khi phân tích đồng thời naringin, hesperidin, neohesperidin và quercetin trên detector PDA, tiến hành quét phổ UV-Vis trong khoảng từ 200 - 800nm, xác định cực đại hấp thụ của mỗi chất. Kết quả thể hiện trên hình 3.17.

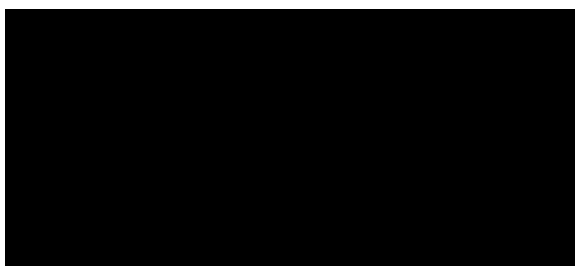


Hình 3.17. Phổ hấp thụ UV-Vis của các flavonoid

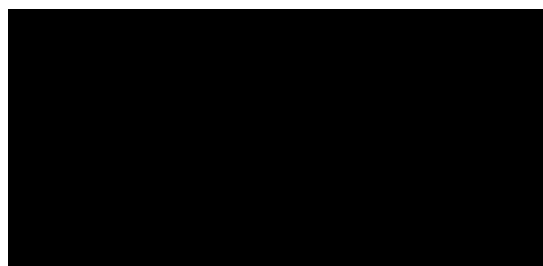
Naringin, hesperidin, neohesperidin có cực đại hấp thụ lần lượt ở 283,0 nm; 284,2 nm; 284,2 nm và quercetin có 2 cực đại tại 254,4 nm và 364,2 nm. Để giảm ảnh hưởng của dung môi và tạp chất những vẫn đảm bảo được độ nhạy của phương pháp khi phân tích đồng thời bốn chất trong nước cam nghiên cứu đã lựa chọn bước sóng phát hiện tại 284 nm đối với naringin, hesperidin, neohesperidin; bước sóng phát hiện tại 254 nm đối với quercetin.

b. Lựa chọn cột pha tĩnh

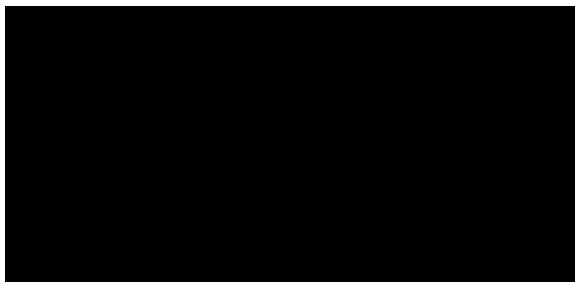
Tính chất của pha tĩnh và thành phần pha động có mối liên quan chặt chẽ với nhau, có ảnh hưởng rất nhiều đến quá trình tách sắc ký do độ phân cực tương đối của pha tĩnh và pha động. Kết quả khảo sát với 3 loại cột được thể hiện trên Hình 3.18.



Cột Symmetry® C18



Cột Sunfire™ C18



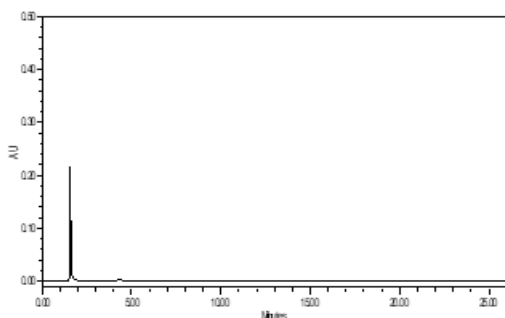
Cột Waters XBridge®

Hình 3.18. Sắc ký đồ hỗn hợp chuẩn của 3 cột khảo sát

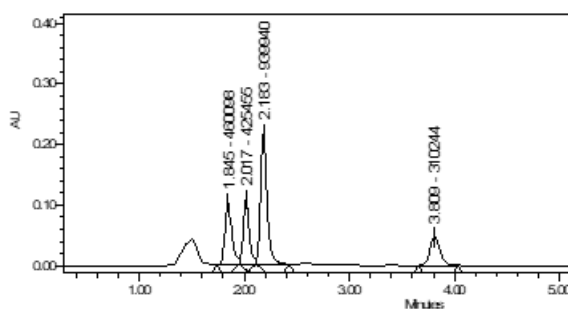
Kết quả cho thấy, cột Cột Symmetry® C18 và cột Sunfire™ C18 cho độ tách các chất không tốt, peak sắc ký không đối xứng, chân peak bị giãn, cột Sunfire™ C18 còn gặp tình trạng chẻ peak. Cột Waters XBridge® peak sắc ký của các chất được tách nhau hoàn toàn, peak gọn, đối xứng. Từ kết quả trên, cột Waters XBridge® C18 (4,6 mm x 150 mm; 5µm) được lựa chọn cho nghiên cứu này.

c. Lựa chọn pha động

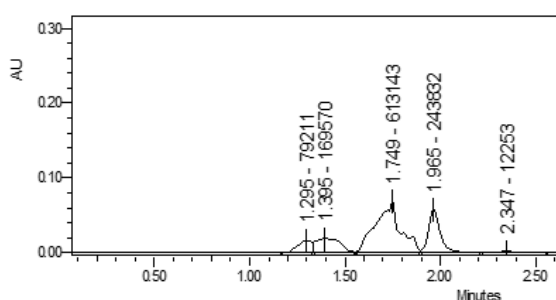
Hiệu quả tách sắc ký phụ thuộc phần lớn vào pha động. Pha động có thể ảnh hưởng đến: độ chọn lọc của hệ pha, thời gian lưu giữ của chất phân tích, độ rộng chân peak sắc ký... Kết quả khảo sát đẳng dòng được thể hiện trên sắc ký đồ Hình 3.19.



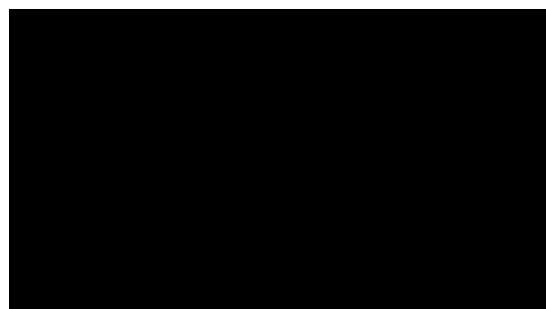
ACN 100%



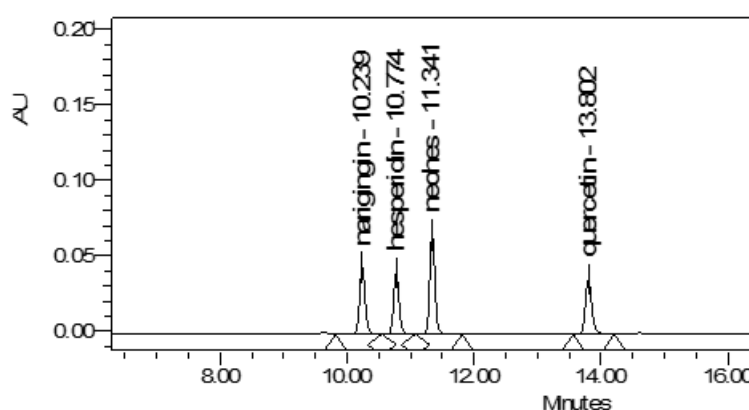
Acid phosphoric 0,1% : ACN (70 : 30)



Acid phosphoric 0,1% : ACN (50 : 50)



Gradient 1



Gradient 2

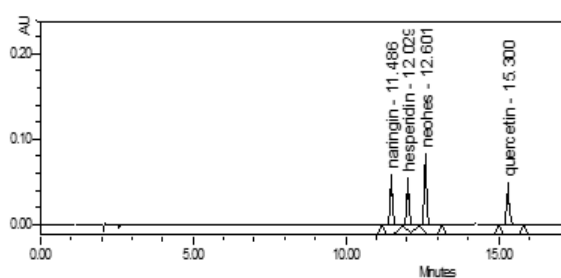
Hình 3.19. Sắc ký đồ khảo sát pha động

Kết quả cho thấy chương trình gradient tách tốt hơn nhiều so với chương trình đẳng dòng. Chương trình đẳng dòng cho khả năng tách các peak phân tích rất kém, thời gian lưu ngắn từ 1,8 - 3,8 phút khó tách khỏi các tạp chất khi phân tích mẫu thực, hình dáng peak xấu. Chương trình gradient thu được các peak cân đối, nhọn, tín hiệu đường nền thấp, các chất tách tốt ra khỏi nhau.

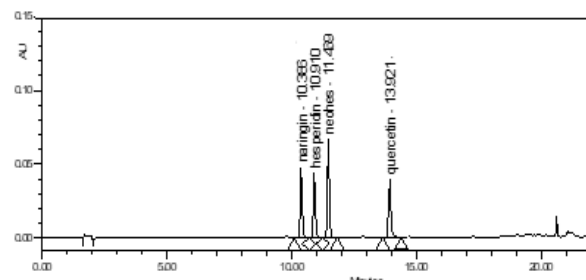
Hai chương trình gradient đều cho khả năng tách tốt các chất, chương trình gradient 2 cho độ rộng chân peak nhỏ hơn và thời gian lưu ngắn hơn so với chương trình gradient 1 giúp tiết kiệm thời gian và chi phí. Vì vậy, pha động gồm 2 thành phần acid phosphoric 0,1% và ACN với chương trình gradient 2 được lựa chọn để phân tích đồng thời naringin, hesperidin, neohesperidin và quercetin.

d. Lựa chọn tốc độ dòng pha động

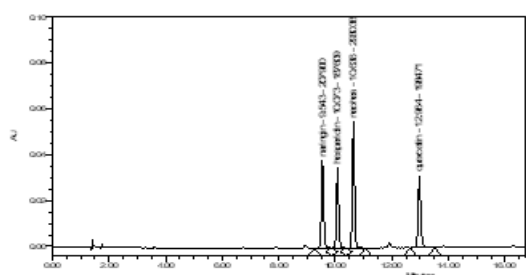
Tiến hành khảo sát 3 tốc độ dòng 0,8 mL/phút; 1,0 mL/phút và 1,2 mL/phút với các điều kiện pha động và pha tính đã tối ưu. Kết quả được thể hiện trên Hình 3.20.



Tốc độ dòng 0,8 mL/phút



Tốc độ dòng 1,0 mL/phút



Tốc độ dòng 1,2 mL/phút

Hình 3.20. Sắc ký đồ khảo sát tốc độ dòng tách 4 flavonoid

Bảng 3.13. Các thông số khảo sát tách 4 flavonoid với các tốc độ dòng

Tốc độ dòng (mL/phút)	Chất phân tích	Thời gian lưu (phút)	S/N	Hệ số đối xứng	Độ rộng chân peak
0,8	Naringin	11,4	490,6451	1,1688	0,098
	Hesperidin	12,0	453,2829	1,1714	0,097
	Neohesperidin	12,6	730,9057	1,1732	0,094
	Quercetin	15,3	390,2619	1,2609	0,117
1,0	Naringin	10,2	1215,8197	0,8988	0,1141
	Hesperidin	10,8	910,2566	0,9324	0,1110
	Neohesperidin	11,3	1661,3051	0,9485	0,1047
	Quercetin	13,8	1079,9414	1,1571	0,1138
1,2	Naringin	9,5	122,4529	1,0791	0,0879
	Hesperidin	10,0	111,6097	1,0992	0,0876
	Neohesperidin	10,6	180,0869	1,0883	0,0847
	Quercetin	12,9	99,9388	1,1951	0,1009

Kết quả cho thấy, với tốc độ dòng 1,0 mL/phút tỉ lệ tín hiệu nhiễu trên nền (S/N) có giá trị lớn nhất, hệ số đối xứng gần 1 nhất (đối với hesperidin, neohesperidin và quercetin). Với tốc độ dòng 1,2 mL/phút độ rộng chân peak nhỏ nhất nhưng tỷ lệ nhiễu trên nền rất nhỏ, thực tế diện tích peak cũng nhỏ nhất làm giảm độ nhạy của phương pháp. Do vậy, để thu được sắc ký đồ với các peak gọn, tách tốt, thời gian phân tích vừa phải nghiên cứu đã lựa chọn tốc độ dòng 1,0 mL/phút trong phương pháp xác định đồng thời naringin, hesperidin, neohesperidin và quercetin trong nước cam.

Như vậy, sau khi tiến hành khảo sát điều kiện phân tích đồng thời naringin, hesperidin, neohesperidin và quercetin trong nước cam bằng HPLC cụ thể như sau:

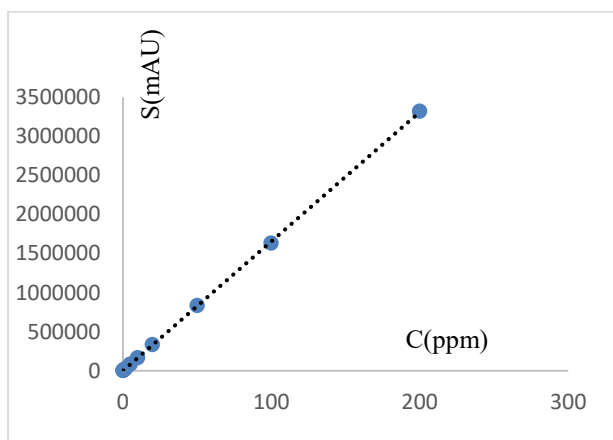
- Cột sắc ký: Cột XBridge® Water C18 (4,6mm x 150mm; 5 μ m).
- Pha động: acid phosphoric 0,1% và ACN với chương trình gradient 2 ở Bảng 2.1.
- Tốc độ dòng 1,0 mL/phút.
- Thể tích tiêm mẫu: 10 μ L.
- Detector PDA
- Bước sóng phát hiện tại 284 nm đối với naringin, hesperidin, neohesperidin; tại 254 nm đối với quercetin.

e. Xây dựng đường chuẩn

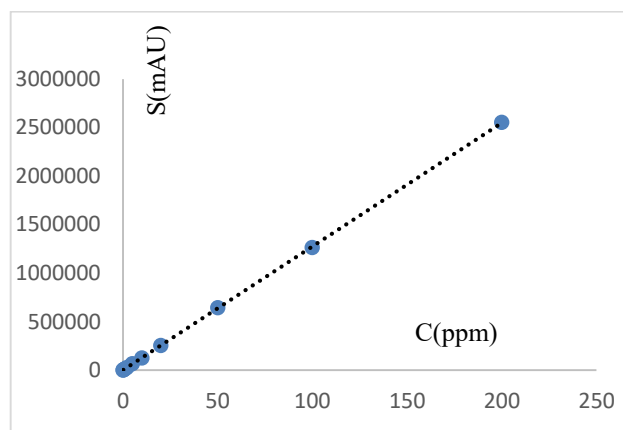
Xây dựng đường chuẩn bằng cách phân tích các mẫu chuẩn hỗn hợp với nồng độ thay đổi từ 0,1 μ g/mL - 200 μ g/mL với cả 4 chất theo quy trình đã xây dựng. Xác định sự tương quan giữa nồng độ chất phân tích và diện tích peak tương ứng thu được trên sắc ký đồ bằng phương pháp hồi quy tuyến tính

Bảng 3.14. Đường chuẩn của các chất phân tích

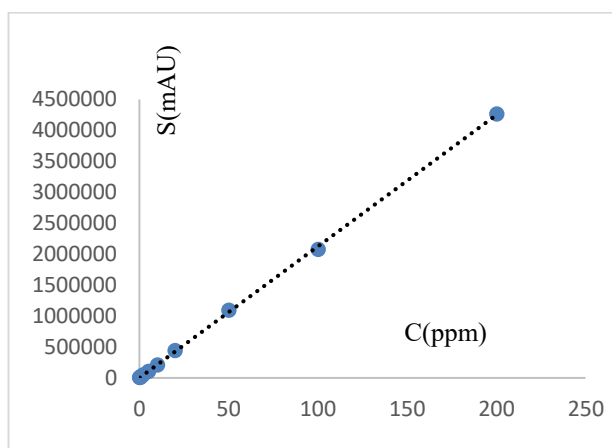
Chất phân tích	Khoảng tuyến tính	Đường chuẩn	R ²
Naringin	0,1 – 200 μ g/mL	$y = 16581x - 266,93$	0,9999
Hesperidin	0,1 – 200 μ g/mL	$y = 12741x + 1318,9$	1,0
Neohesperidin	0,1 – 200 μ g/mL	$y = 21239x - 622,24$	0,9998
Quercetin	0,1 – 200 μ g/mL	$y = 16311x - 902,92$	0,9998



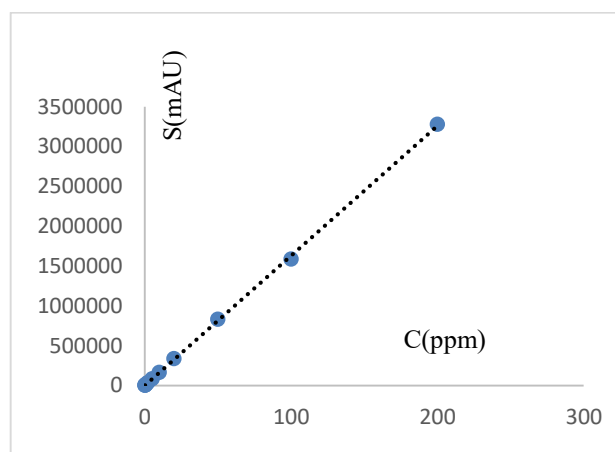
Naringin



Hesperidin



Neohesperidin



Quercetin

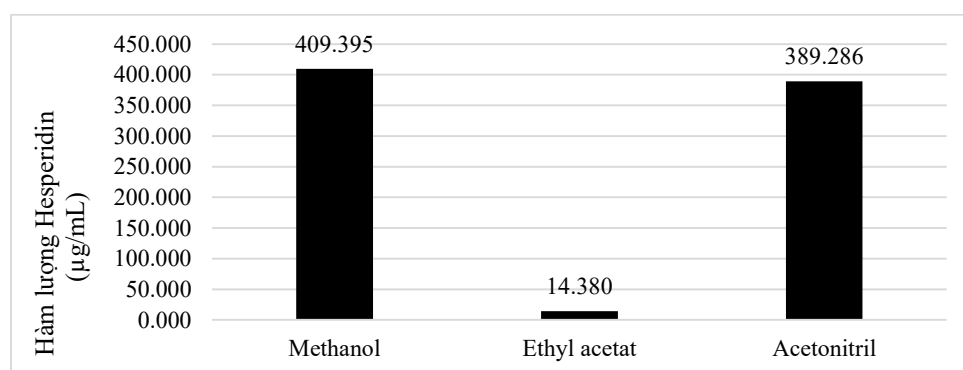
Hình 3.21. Đường chuẩn của 4 flavonoid trong nước cam

Kết quả cho thấy khoảng tuyến tính của 4 flavonoid trong khoảng nồng độ từ 0,1 – 200 $\mu\text{g/mL}$, các phương trình tuyến tính đều có hệ số tương quan $R^2 > 0,99$, đạt yêu cầu của AOAC.

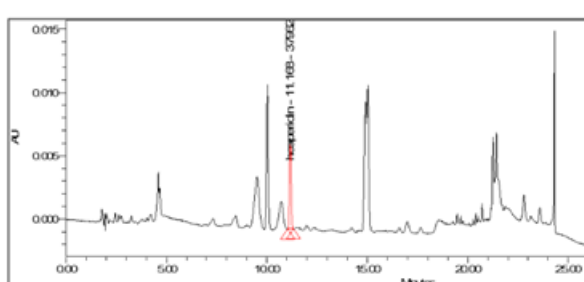
3.1.3.2. Xây dựng quy trình xử lý mẫu

a. Khảo sát lựa chọn dung môi hòa tan

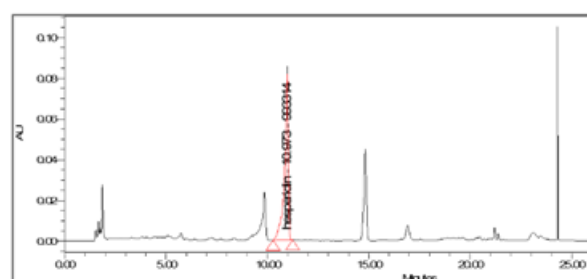
Cả 4 chất đều khó tan trong nước, tan tốt trong dung môi hữu cơ. Vì vậy, tiến hành khảo sát 3 dung môi hòa tan: methanol, ethyl acetat, ACN trên nền mẫu cam Vinh.



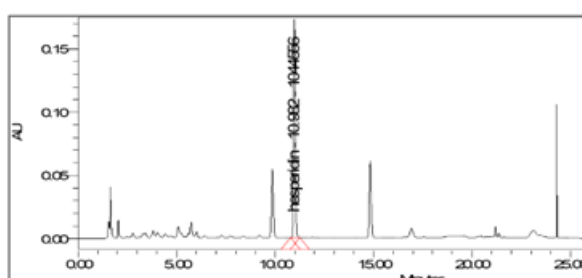
Hình 3.22. Kết quả khảo sát dung môi hòa tan



Ethyl acetat



ACN



Methanol

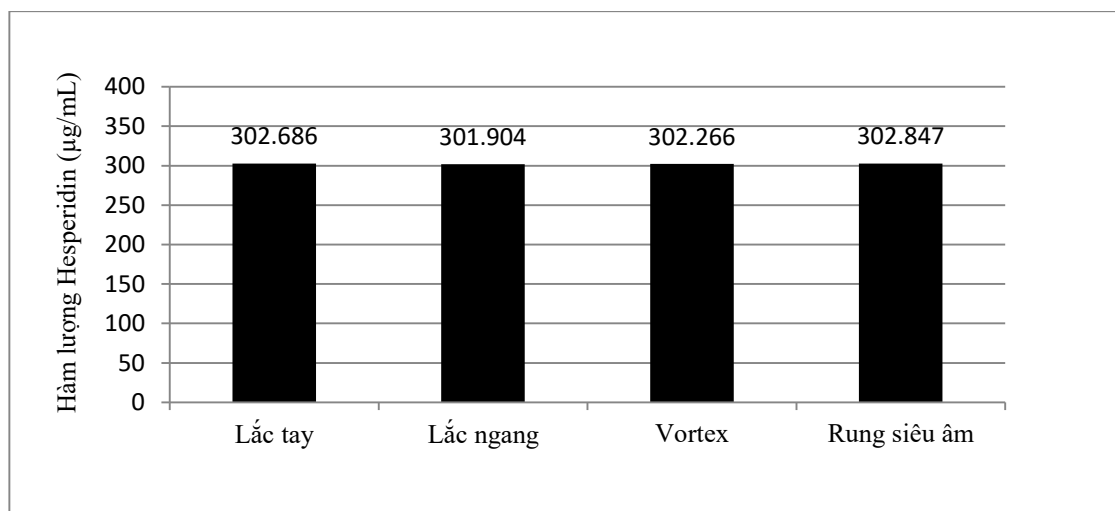
Hình 3.23. Sắc ký đồ mẫu nước cam được hòa tan trong 3 dung môi

Kết quả sắc ký đồ cho thấy, trên mẫu thực chỉ phát hiện thấy hesperidin, khi hòa tan bằng ethyl acetat không loại bỏ được hết tạp chất nên peak lẫn rất nhiều tạp, hàm lượng hesperidin thấp hơn rất nhiều so với 2 dung môi còn lại. Với dung môi methanol cho peak sắc ký đối xứng và độ rộng chân peak nhỏ hơn, hàm lượng hesperidin cao hơn so với ACN.

Vì vậy, nghiên cứu đã lựa chọn methanol là dung môi hòa tan cho các mẫu nước cam.

b. Khảo sát lựa chọn kỹ thuật hòa tan mẫu

Kỹ thuật hòa tan mẫu ảnh hưởng nhiều đến khả năng phân tán và tách hoạt chất cần phân tích ra khỏi nền mẫu. Dựa vào điều kiện phòng thí nghiệm, tiến hành khảo sát 4 cách hòa tan: lắc tay, lắc ngang, lắc vortex và rung siêu âm.

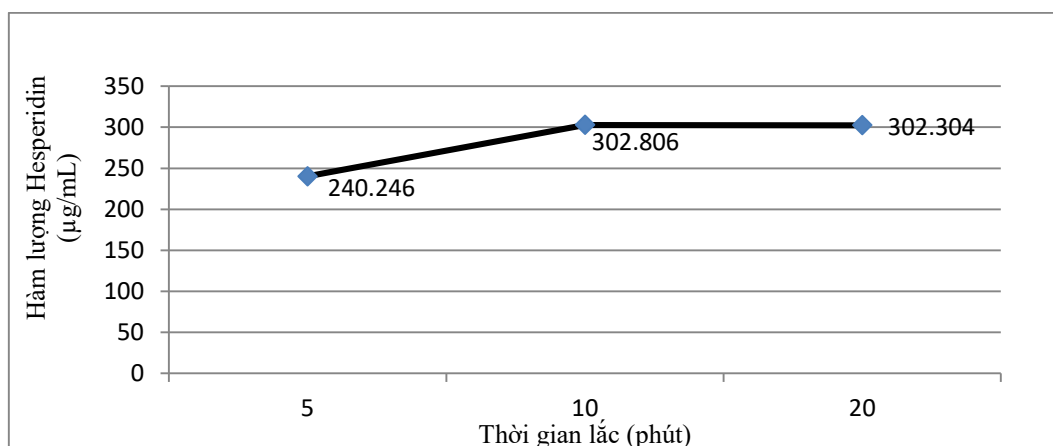


Hình 3.24. Kết quả khảo sát kỹ thuật hòa tan

Với cả 4 kỹ thuật, hiệu suất thu được khác nhau không đáng kể. Tuy nhiên, lắc tay cho năng suất làm việc kém, vortex mỗi lần chỉ được tối đa 2 mẫu. Do vậy, để thuận tiện và phù hợp với điều kiện phòng thí nghiệm nghiên cứu lựa chọn kỹ thuật lắc ngang để tối ưu khả năng xử lý mẫu, tăng năng suất làm việc, ít tốn kém.

c. Khảo sát lựa chọn thời gian hòa tan mẫu

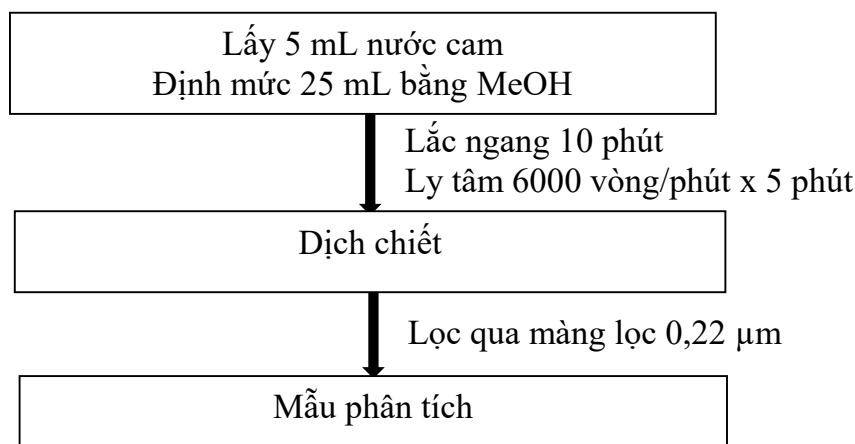
Trên cùng một nền mẫu nước cam Vinh, tiến hành lắc ngang bằng methanol ở các mức thời gian khác nhau: 5 phút, 10 phút, 20 phút.



Hình 3.25. Kết quả khảo sát thời gian hòa tan

Hiệu suất khi lắc 5 phút thấp hơn nhiều so với 2 mức còn lại. Ở mức thời gian 10 phút và 20 phút hiệu suất tương đương nhau. Vì vậy để tiết kiệm thời gian xử lý mẫu, nghiên cứu đã lựa chọn thời gian lắc là 10 phút.

Dựa trên các khảo sát, quy trình xử lý mẫu được trình bày trong sơ đồ Hình 3.26.



Hình 3.26. Quy trình xử lý mẫu

3.1.3.3. Thẩm định phương pháp

a. Độ ổn định hệ thống

Thực hiện tiêm lặp lại 6 lần vào nền mẫu hỗn hợp chuẩn có nồng độ 200 ppm. Kết quả thu được như Bảng 3.15.

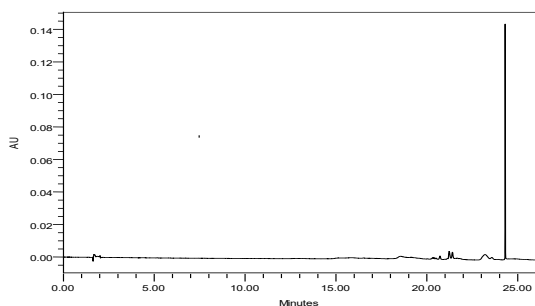
Bảng 3.15. Kết quả độ ổn định hệ thống

Chất phân tích	RSD _{IR} (%)	RSD _{Speak} (%)
Naringin	0,09	0,49
Hesperidin	0,09	0,94
Neohesperidin	0,07	0,44
Quercetin	0,07	0,53

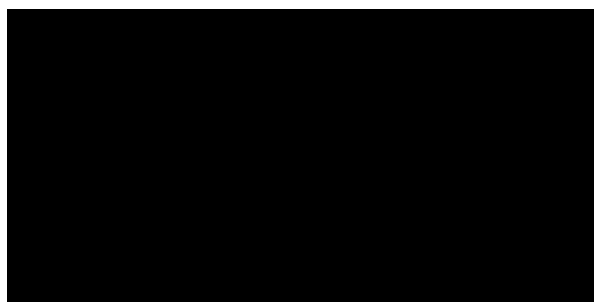
Giá trị RSD theo thời gian lưu và theo diện tích pic của cả 4 chất chất đều nhỏ hơn 2%, đáp ứng theo yêu cầu của AOAC.

b. Độ đặc hiệu

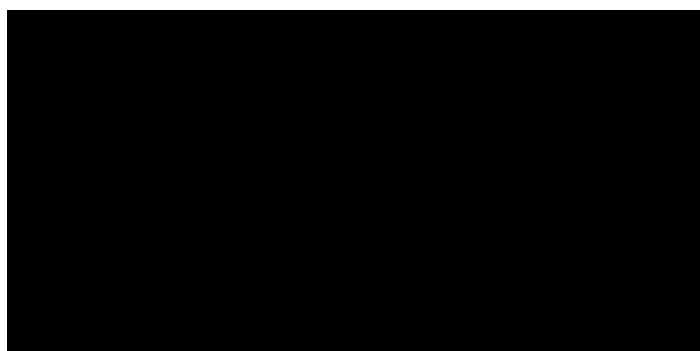
Độ đặc hiệu của phương pháp được đánh giá thông qua phân tích mẫu trắng, mẫu dung dịch hỗn hợp chuẩn và mẫu thực thêm chuẩn theo phương pháp đã xây dựng.



Mẫu trắng



Mẫu chuẩn



Mẫu thử thêm chuẩn

Hình 3.27. Sắc ký đồ khảo sát độ đặc hiệu phân tích flavonoid

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. Như vậy, phương pháp đã xây dựng có độ đặc hiệu cao.

d. Giới hạn phát hiện (MDL), giới hạn định lượng (MQL) của phương pháp

Các kết quả xác định MDL và MQL của các flavonoid được trình bày ở phụ lục 2.3. Kết quả tổng hợp được trình bày ở bảng 3.16.

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, do đó giá trị MDL, MQL thu được là đáng tin cậy.

e. Độ lặp lại và độ thu hồi

Độ lặp lại được thực hiện trên mẫu cam Vinh và tiến hành làm thí nghiệm 6 lần lặp lại. Do naringin, neohesperidin và quercetin không có trong mẫu thực, vì vậy nghiên cứu thêm một lượng chuẩn (0,6 ppm) của naringin, neohesperidin, quercetin vào mẫu thực để đánh giá độ thu hồi ở mức lặp lại của ba chất. Độ lặp lại của phương pháp được đánh giá thông qua độ lệch chuẩn (SD) và độ lệch chuẩn tương đối (RSD (%)).

Bảng 3.17. Kết quả độ lặp lại của 4 flavonoid

Chất phân tích	Naringin	Hesperidin	Neohesperidin	Quercetin
Hàm lượng trung bình ($\mu\text{g/mL}$)	3,20	376,26	3,32	3,07
SD	0,07	7,61	0,08	0,05
RSD (%)	2,07	2,02	2,48	1,35

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%, đáp ứng yêu cầu của AOAC khi phân tích đồng thời naringin, hesperidin, neohesperidin và quercetin trên nền mẫu nước cam.

Độ thu hồi được thực hiện bằng cách thêm 1 lượng chuẩn xác định ở 3 mức nồng độ thấp, trung bình và cao, mỗi mức thực hiện thêm chuẩn 3 lần.

Bảng 3.18. Kết quả độ thu hồi của 4 flavonoid

Chất phân tích		Naringin	Hesperidin	Neohesperidin	Quercetin
Thêm chuẩn mức thấp	Hiệu suất (%)	105,0 ÷ 110,5	97,0 ÷ 99,0	102,0 ÷ 107,0	100,0 ÷ 101,4
	RSD (%)	2,17	0,34	2,42	0,70

Chất phân tích		Naringin	Hesperidin	Neohesperidin	Quercetin
Thêm chuẩn mức trung bình	Hiệu suất (%)	98,9 ÷ 99,6	104,1 ÷ 104,3	95,7 ÷ 96,8	85,3 ÷ 90,0
	RSD (%)	0,37	0,09	0,60	3,01
Thêm chuẩn mức cao	Hiệu suất (%)	101,2 ÷ 101,5	92,5 ÷ 93,0	92,2 ÷ 92,5	85,5 ÷ 87,9
	RSD (%)	0,14	0,20	0,17	1,58

Theo AOAC độ thu hồi phải đạt từ 80 - 110% tại mức nồng độ từ 100 ppb - 10ppm; từ 90 - 107% tại mức nồng độ 100ppm và RSD không vượt quá 11%. Kết quả thẩm định của cả 4 chất đều có hiệu suất thu hồi và RSD nằm trong giới hạn cho phép. Điều này chứng minh phương pháp đáp ứng được yêu cầu theo tiêu chuẩn của AOAC về độ thu hồi.

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

3.2.1. Phân tích thống kê

Tiến hành định lượng các thành phần hóa học chính theo các quy trình đã xây dựng 82 mẫu cam thuộc 4 giống, từ 10 tỉnh của Việt Nam (cụ thể đã trình bày ở mục 2.3.2.1). Bảng kết quả phân tích các thành phần hóa học chính của các mẫu nước cam tự nhiên được trình bày trong Bảng 3.19.

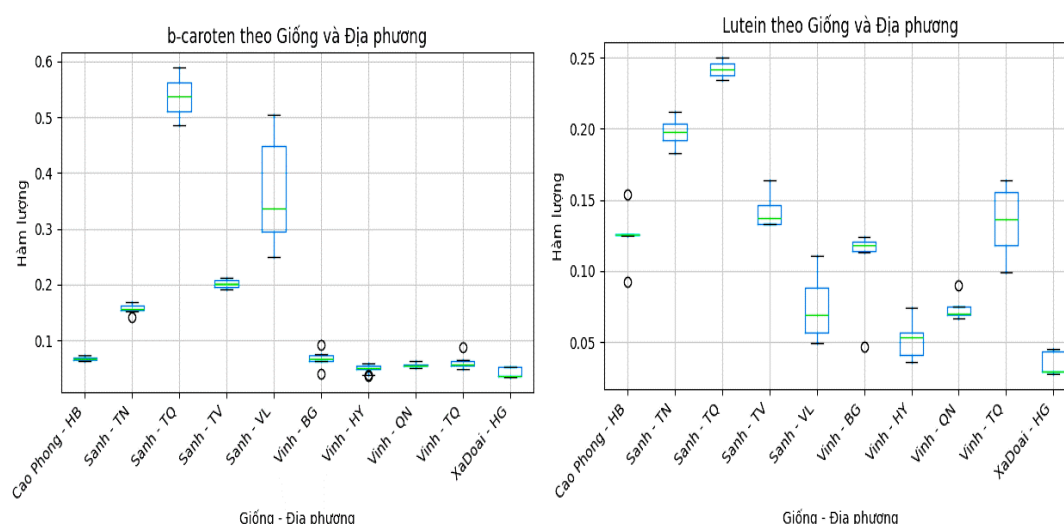
Bảng 3.19. Hàm lượng các hợp chất chính trong nước cam tươi ở các giống và được trồng ở các tỉnh khác nhau

<i>Giống</i>	<i>Tỉnh</i>	<i>Đường (g/100mL)</i>			<i>Carotenoid (mg/L)</i>		<i>Acid (g/L)</i>			<i>Flavonoid (mg/L)</i>	
		<i>fructose</i>	<i>glucose</i>	<i>sucrose</i>	<i>lutein</i>	<i>β-caroten</i>	<i>citric</i>	<i>ascorbic</i>	<i>fumaric</i>	<i>hesperidin</i>	<i>quercetin</i>
Vinh	BG	2,367 ±0,238	2,148 ±0,091	2,720 ±0,179	0,107 ±0,030	0,067 ±0,017	187,002 ±6,987	1,457 ±0,533	0,147 ±0,018	281,831 ±128,859	39,737 ±3,894
Vinh	HY	2,509 ±0,423	2,226 ±0,390	4,184 ±0,837	0,052 ±0,011	0,049 ±0,007	205,120 ±38,511	1,951 ±0,474	0,058 ±0,027	385,315 ±152,662	21,040 ±4,621
Vinh	QN	3,185 ±0,566	3,458 ±0,286	3,888 ±0,840	0,074 ±0,009	0,056 ±0,005	264,383 ±9,171	3,786 ±0,094	0,163 ±0,003	425,561 ±206,081	92,040 ±0,198
Vinh	TQ	2,198 ±0,446	1,966 ±0,414	3,042 ±0,655	0,136 ±0,022	0,059 ±0,011	317,082 ±16,412	14,224 ±8,366	0,755 ±0,781	363,231 ±168,668	71,207 ±23,816
Xã Đoài	HG	2,480 ±0,541	2,385 ±0,589	2,445 ±0,545	0,033 ±0,008	0,041 ±0,009	192,286 ±6,149	2,773 ±0,173	0,066 ±0,003	624,549 ±113,299	20,615 ±0,876
Xã Đoài	NA	2,081 ±0,185	1,878 ±0,232	3,744 ±0,431	0,070 ±0,012	0,052 ±0,005	188,552 ±66,515	3,560 ±4,378	0,071 ±0,023	288,836 ±79,365	23,523 ±8,130
Sành	TN	1,424 ±0,236	1,427 ±0,273	3,662 ±0,663	0,197 ±0,010	0,157 ±0,008	176,225 ±10,347	1,837 ±1,582	0,245 ±0,084	259,500 ±142,291	69,624 ±33,347
Sành	TQ	1,770 ±0,269	1,345 ±0,163	3,085 ±0,573	0,242 ±0,011	0,537 ±0,074	238,320 ±11,355	0,612 ±0,004	0,026 ±0,006	360,126 ±152,941	85,685 ±4,831
Sành	TV	1,705 ±0,941	1,663 ±0,985	3,408 ±0,540	0,143 ±0,015	0,202 ±0,010	149,529 ±32,019	1,455 ±0,411	0,160 ±0,063	350,763 ±206,175	40,418 ±2,850
Sành	VL	2,062 ±0,470	1,939 ±0,468	4,236 ±0,792	0,074 ±0,022	0,362 ±0,093	205,128 ±46,589	0,440 ±0,152	0,210± 0,176	197,343 ±60,182	26,380 ±17,340
Cao Phong	HB	3,038 ±1,021	2,647 ±0,982	4,327 ±1,501	0,125 ±0,020	0,068 ±0,004	184,309 ±1,937	1,737 ±0,022	0,118 ±0,020	514,923 ±155,664	44,040 ±2,151

Ghi chú: BG: Bắc Giang, HY: Hưng Yên, QN: Quảng Ninh, TQ: Tuyên Quang, HG: Hà Giang, NA: Nghệ An, TN: Tây Ninh, TV: Trà Vinh, VL: Vĩnh Long, HB: Hòa Bình, giá trị hàm lượng = hàm lượng trung bình ± SD.

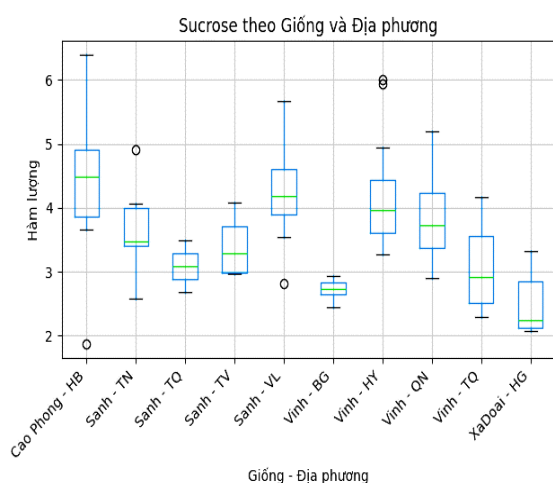
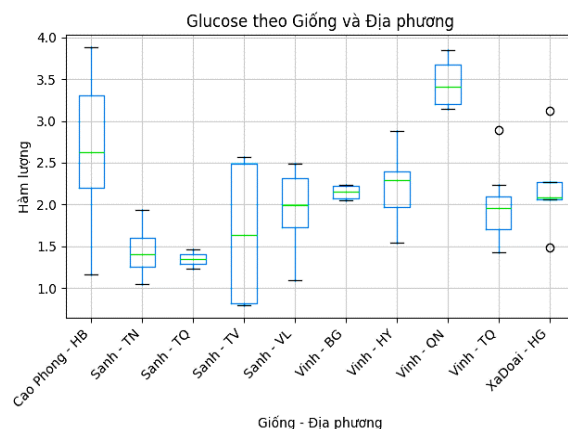
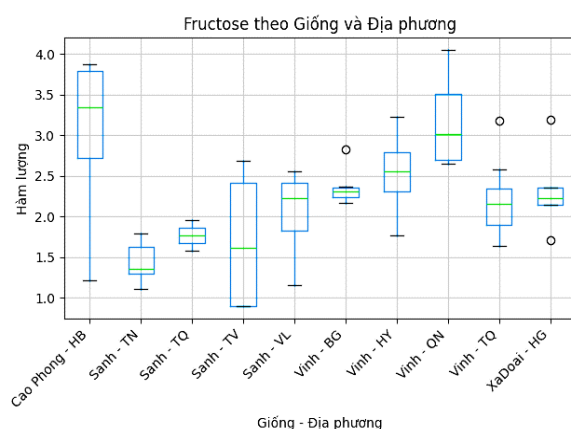
Kết quả phân tích hàm lượng các hợp chất chính cho thấy nồng độ các chất này biến thiên đáng kể theo giống cam và vùng địa lý trồng, với sự khác biệt có ý nghĩa thống kê.

Đối với nhóm carotenoid, giống cam Sành thể hiện hàm lượng β -caroten và lutein cao hơn rõ rệt so với các giống còn lại trong nghiên cứu. Hàm lượng lutein dao động từ $0,033 \pm 0,008$ đến $0,242 \pm 0,011$ mg/L, trong đó giá trị cao nhất được ghi nhận ở cam trồng tại Tuyên Quang. Hàm lượng β -carotene nằm trong khoảng $0,041 \pm 0,009$ đến $0,537 \pm 0,074$ mg/L, đồng thời cũng đạt giá trị cực đại ở cam Sành trồng tại Tuyên Quang.



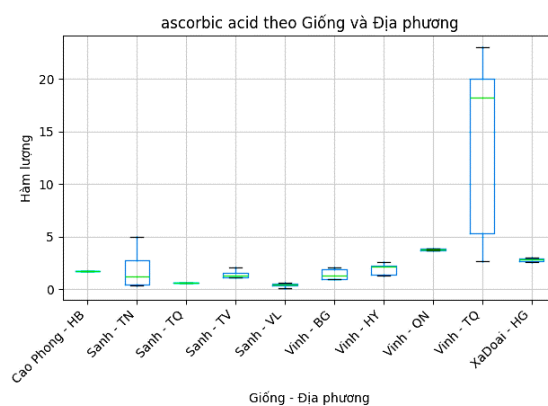
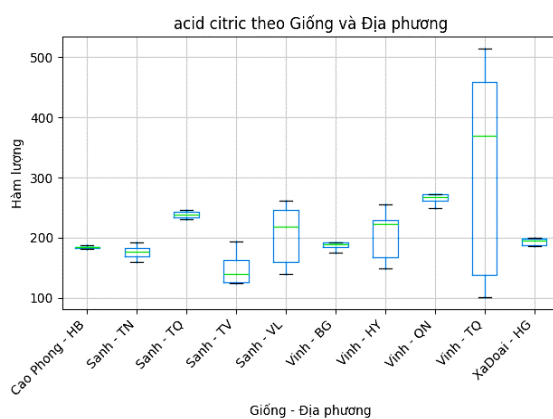
Hình 3.28. Sự phân bố hàm lượng carotenoid trong nước cam

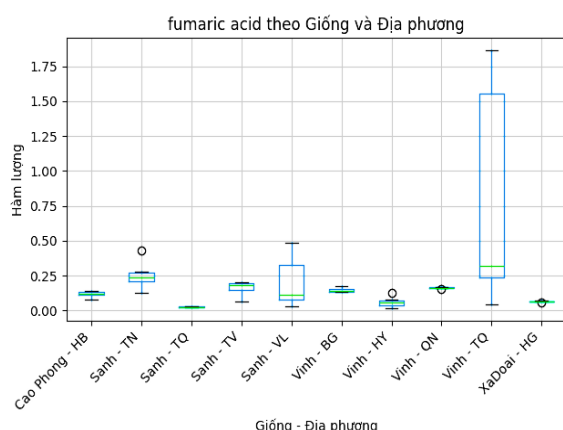
Đối với nhóm đường, hàm lượng fructose dao động từ $1,424 \pm 0,236$ đến $3,185 \pm 0,566$ g/100mL, với giá trị cao nhất quan sát được ở cam Vinh trồng tại Quảng Ninh. Tương tự, glucose có hàm lượng trong khoảng $1,427 \pm 0,273$ đến $3,458 \pm 0,286$ g/100mL, cũng đạt cực đại ở cam Vinh của Quảng Ninh. Đáng chú ý, sucrose có hàm lượng cao hơn đáng kể so với hai loại đường còn lại, dao động từ $2,445 \pm 0,545$ đến $4,327 \pm 1,501$ mg/L, với giá trị lớn nhất được ghi nhận ở cam Cao Phong trồng tại Hòa Bình.



Hình 3.29. Sự phân bố hàm lượng đường trong nước cam

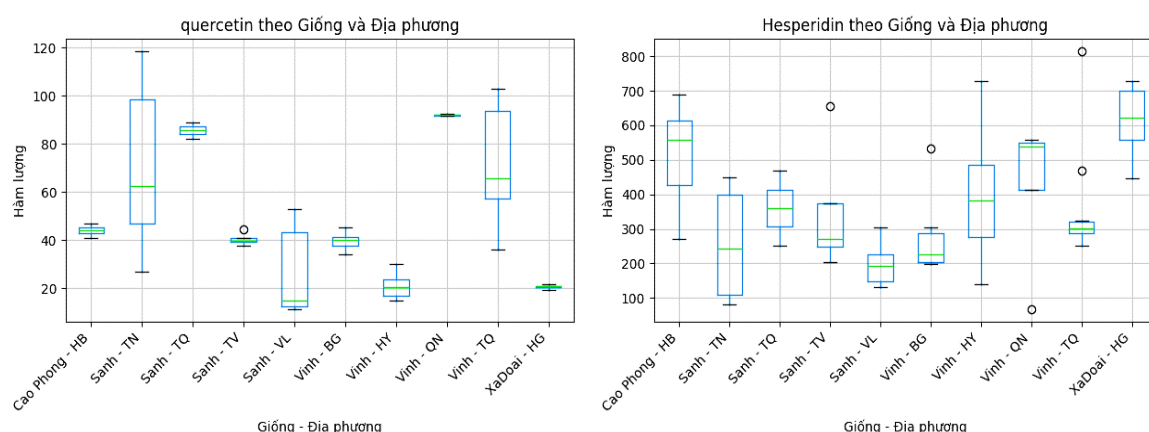
Đối với nhóm acid hữu cơ, hàm lượng acid citric biến thiên trong khoảng $149,529 \pm 32,019$ đến $317,082 \pm 167,412$ g/L, với giá trị cao nhất ở cam Vinh trồng tại Tuyên Quang. Hàm lượng acid ascorbic dao động từ $0,440 \pm 0,152$ đến $14,224 \pm 8,366$ g/L, trong khi acid fumaric nằm trong khoảng $0,026 \pm 0,006$ đến $0,755 \pm 0,781$ g/L; cả hai acid này đều đạt giá trị cao nhất ở cam Vinh của Tuyên Quang.





Hình 3.30. Sự phân bố hàm lượng acid trong nước cam

Đối với nhóm flavonoid, hàm lượng quercetin dao động từ $20,615 \pm 0,876$ đến $92,040 \pm 0,198$ mg/L, với giá trị cao nhất được ghi nhận ở cam Vinh trồng tại Quảng Ninh. Hàm lượng hesperidin nằm trong khoảng $197,343 \pm 60,182$ đến $624,549 \pm 113,299$ mg/L, trong đó cam Xã Đoài trồng tại Hà Giang cho giá trị cao nhất.



Hình 3.31. Sự phân bố hàm lượng flavonoid trong nước cam

Những kết quả này phù hợp với các báo cáo trước đây trên thế giới [36, 44, 50, 53, 75, 83], trong đó chỉ ra rằng thành phần và hàm lượng các hợp chất hóa học trong nước cam phụ thuộc mạnh vào giống và điều kiện trồng trọt. Đáng lưu ý, đối với một số hợp chất như β -carotene, acid citric và quercetin, một số mẫu thể hiện giá trị đặc biệt cao so với phần lớn các mẫu còn lại. Các điểm dữ liệu này được xem là ngoại lai (outliers) và cần được loại bỏ trước khi áp dụng các mô hình học máy, nhằm tránh ảnh hưởng tiêu cực đến quá trình huấn luyện mô hình và đạt được độ chính xác dự đoán tối ưu.

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

Phân tích phương sai một chiều (one-way ANOVA) được sử dụng để kiểm định sự khác biệt về hàm lượng các hợp chất hóa học chính giữa các giống cam và các địa phương Việt Nam với mức ý nghĩa thống kê $\alpha = 0,05$. Bảng kết quả ANOVA một chiều được thể hiện ở bảng 3.20

Bảng 3.20. Kết quả ANOVA thành phần hóa học

Hợp chất	Theo giống		Theo địa phương	
	p	Ý nghĩa	p	Ý nghĩa
Lutein	$5,654.10^{-07}$	***	$4,329.10^{-22}$	***
Quercetin	$1,860.10^{-02}$	*	$1,208.10^{-13}$	***
β -caroten	$5,645.10^{-18}$	***	$6,942.10^{-11}$	***
Acid ascorbic	$9,249.10^{-03}$	**	$5,460.10^{-08}$	***
Glucose	$8,825.10^{-05}$	***	$8,838.10^{-07}$	***
Fructose	$3,722.10^{-07}$	***	$1,073.10^{-06}$	***
Sucrose	$1,211.10^{-01}$	ns	$4,240.10^{-05}$	***
Hesperidin	$1,492.10^{-03}$	**	$7,566.10^{-05}$	***
Acid citric	$3,283.10^{-02}$	*	$9,975.10^{-04}$	***
Acid fumaric	$2,296.10^{-01}$	ns	$4,374.10^{-03}$	**

Ghi chú: ns (không ý nghĩa), * ($p < 0,05$), ** ($p < 0,01$), *** ($p < 0,001$).

Kết quả ANOVA một chiều cho thấy các thành phần hóa học chính của nước cam Việt Nam biến thiên theo cả hai tiêu chí phân loại là giống và địa phương trồng.

Xét theo giống, đa số hợp chất có sự khác biệt có ý nghĩa thống kê ($p < 0,05$), đặc biệt nhóm carotenoid gồm lutein ($p = 5,654.10^{-7}$) và β -caroten ($p = 5,645.10^{-18}$) thể hiện mức khác biệt lớn giữa các giống, gợi ý sự khác biệt di truyền trong con đường sinh tổng hợp carotenoid. Nhóm đường cũng tạo sự phân tách mạnh, fructose ($p = 3,722.10^{-7}$) và glucose ($p = 8,825.10^{-5}$) thay đổi đáng kể theo giống. Ngược lại, sucrose ($p = 0,1211$) và acid fumaric ($p = 0,2296$) sự khác biệt không có ý nghĩa giữa các giống. Điều này gợi ý rằng hai chỉ tiêu này ít chịu chi phối bởi yếu tố di truyền theo giống.

Xét theo địa phương trồng, tất cả hợp chất khảo sát đều có sự khác biệt có ý nghĩa thống kê ($p < 0,05$), trong đó lutein ($p = 4,329.10^{-22}$), quercetin ($p = 1,208.10^{-13}$) và β -caroten ($p = 6,942.10^{-11}$) cho khả năng phân tách mạnh do có giá trị p rất nhỏ. Kết quả này cho thấy vai trò của điều kiện sinh thái (nhiệt độ, lượng mưa, thổ nhưỡng, dinh dưỡng ...) đối với tích lũy chất chuyển hóa thứ cấp (carotenoid và flavonoid).

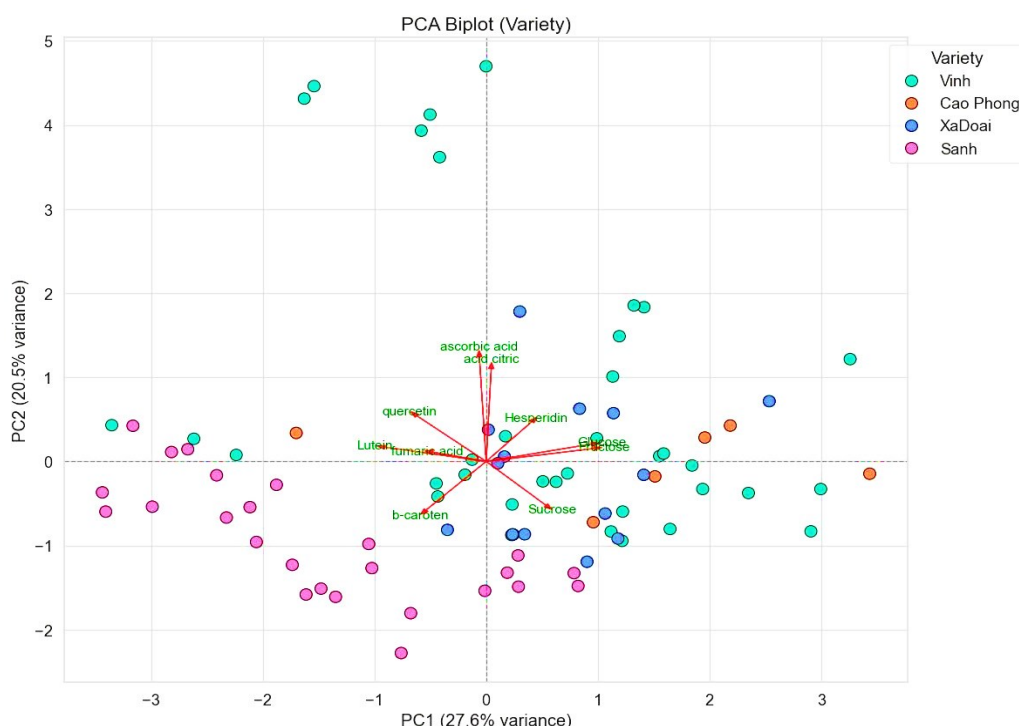
Vavoura và cộng sự cũng đã chứng minh rằng việc kết hợp chỉ tiêu hóa lý - flavonoid - hợp chất bay hơi với chemometric có thể phân biệt nước cam theo giống với độ chính xác cao; trong đó flavonoid là nhóm biến đóng góp quan trọng [145]. Dasenaki và Thomaidis kết luận rằng dấu vân tay hóa học của nước quả thường thể hiện khác biệt theo nguồn gốc địa lý, đặc biệt đối với nhóm chất chuyển hóa thứ cấp (polyphenol/flavonoid) và một số acid hữu cơ [87].

Như vậy, hợp chất có p -value rất nhỏ theo cả giống và địa phương trồng (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. Kết quả này cho thấy, 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

Phân tích thành phần chính (PCA) được thực hiện nhằm khảo sát mối quan hệ giữa thành phần hóa học của các mẫu nước cam và các giống cam tương ứng. Kết quả trên Hình 3.32 chỉ ra rằng hai thành phần chính đầu tiên, PC1 và PC2, lần lượt giải thích 27,6% và 20,5% tổng phương sai của bộ dữ liệu, tương ứng với 48,1% tổng biến thiên. Mặc dù tỷ lệ phương sai giải thích ở mức trung bình, biểu đồ PCA biplot vẫn cho thấy sự phân tách tương đối rõ ràng giữa bốn giống cam gồm Vinh, Cao Phong, Xã Đoài và Sành, phản ánh sự khác biệt đáng kể về hồ sơ chuyển hóa giữa các giống.



Hình 3.32. Kết quả biểu đồ PCA biplot dựa trên tiêu chí phân loại theo giống

Cụ thể, giống cam Sành phân bố chủ yếu về phía âm của PC1, gắn liền với hàm lượng cao của β -carotene, lutein, quercetin và acid fumaric. Điều này cho thấy cam Sành có xu hướng giàu hợp chất chống oxy hóa và sắc tố, phản ánh đặc trưng sinh hóa nổi bật của giống này. Ngược lại, các giống Xã Đoài và Cao Phong nằm về phía dương của PC1, đặc trưng bởi hàm lượng cao của hesperidin, glucose và sucrose, qua đó phản ánh độ ngọt cao hơn và hàm lượng flavonoid đáng kể. Trong khi đó, giống cam Vinh chủ yếu phân bố dọc theo trục dương của PC2, tương ứng với hàm lượng cao của acid ascorbic và acid citric. Các hợp chất này đóng vai trò quan trọng trong việc tạo độ chua cũng như tăng cường tiềm năng chống oxy hóa, cho thấy cam Vinh có đặc trưng sinh hóa khác biệt so với các giống còn lại.

Ngoài ra, hướng vector của glucose, sucrose và hesperidin gần như đồng nhất, cho thấy mối tương quan dương chặt chẽ giữa các hợp chất này. Ngược lại, vector của β -carotene có xu hướng đối nghịch với acid ascorbic, gợi ý mối tương quan nghịch giữa hai nhóm hợp chất này trong các mẫu cam khảo sát.

Tổng thể, các 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. Đây là cơ sở quan trọng cho việc lựa chọn biến đầu vào và phát triển các mô hình hóa học đa biến và học máy nhằm phân loại giống cam ở các bước nghiên cứu tiếp theo.

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 minh họa một hiện tượng kinh điển trong học máy khi phân bố dữ liệu huấn luyện bị mất cân bằng nghiêm trọng giữa các lớp. Trong trường hợp này, các mô hình có xu hướng tối thiểu hóa hàm mất mát chủ yếu theo lớp chiếm ưu thế, dẫn đến độ chính xác tổng thể cao nhưng độ thu hồi (recall) thấp đối với lớp thiểu số. Việc áp dụng SMOTE (Synthetic Minority Over-sampling Technique) đã giảm thiểu một phần hạn chế này thông qua cơ chế nội suy trong không gian đặc trưng, tạo ra các mẫu tổng hợp nằm trong lân cận của các điểm dữ liệu thuộc lớp thiểu số. Nhờ đó, phân bố dữ liệu trở nên đẳng hướng hơn trong không gian học, góp phần hình thành ranh giới quyết định ổn định và công bằng hơn giữa các lớp.

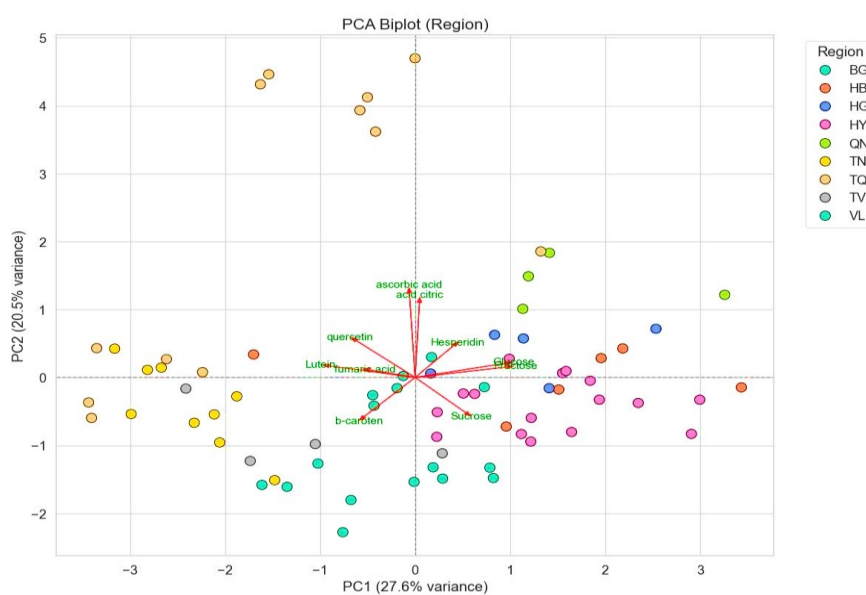
Sự cải thiện nhất quán về hiệu năng của các mô hình SVC, Decision Tree, Random Forest và XGBoost từ khoảng 0,6 - 0,7 lên gần 0,9 cho thấy SMOTE không chỉ làm giảm thiên lệch mà còn làm trơn ranh giới quyết định, đặc biệt có lợi đối với các mô hình dựa trên phân hoạch như Decision Tree và Random Forest. Điều này xuất phát từ việc dữ liệu được bổ sung có cấu trúc giúp các thuật toán phân chia không gian đặc trưng hiệu quả hơn, tránh thiên lệch về lớp chiếm ưu thế.

Tuy nhiên, cần lưu ý rằng SMOTE dựa trên nội suy tuyến tính; do đó, nếu không gian đặc trưng có cấu trúc đa tạp phi tuyến mạnh, các mẫu tổng hợp sinh ra có thể không nằm trên đa tạp dữ liệu thực, từ đó tiềm ẩn nguy cơ quá khớp nhẹ. Dẫu vậy, 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. Kết quả này cho thấy khi phân bố dữ liệu đã được cân bằng, các bộ xấp xỉ phi tuyến toàn cục như MLP có thể khai thác hiệu quả toàn bộ dữ liệu theo cách liên tục, từ đó tạo ra ranh giới quyết định mượt mà và ổn định hơn.

3.2.3.2. Phân loại cam theo nguồn gốc địa lý

Biểu đồ PCA biplot (Hình 3.33) với PC1 giải thích 27,6% và PC2 giải thích 20,5% tổng phương sai cho thấy sự phân hóa hóa học rõ rệt giữa các mẫu cam thu thập từ các vùng địa lý khác nhau. Các mẫu từ Bắc Giang, Tuyên Quang và Quảng Ninh có xu hướng phân cụm tách biệt, phản ánh hồ sơ chuyển hóa đặc trưng theo vùng, nhiều khả năng chịu ảnh hưởng của điều kiện địa lý và môi trường sinh thái. Các vector đại diện cho acid ascorbic, acid citric và hesperidin định hướng gần nhau, cho thấy mối tương quan dương mạnh giữa các acid hữu cơ và nhóm hợp chất liên quan đến vị ngọt. Ngược lại, β -carotene, lutein và quercetin hình thành một nhóm vector riêng biệt, đại diện cho sự biến thiên của các hợp chất sắc tố và flavonoid.



Hình 3.33. Kết quả biểu đồ PCA biplot dựa trên tiêu chí phân loại theo nguồn gốc địa lý

Mặc dù PC1 và PC2 chỉ giải thích khoảng 48% tổng phương sai, hiện tượng phân cụm theo vùng địa lý quan sát được vẫn làm nổi bật đa dạng sinh hóa tiềm ẩn giữa các khu vực trồng cam. Kết quả này khẳng định PCA có khả năng nắm bắt cấu trúc

chính của sự biến thiên chuyển hóa, qua đó hỗ trợ hiệu quả cho việc phân loại vùng địa lý dựa trên dấu vân tay thành phần hóa học của cam.

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, qua đó khẳng định lợi ích của việc cân bằng lớp đối với các bộ dữ liệu chất chuyển hóa có mức độ mất cân bằng trung bình. Trong trường hợp chưa áp dụng SMOTE, Random Forest (0,882) và SVC (0,823) thể hiện hiệu năng vượt trội so với các phương pháp học máy khác, trong khi MLP (0,765) đạt độ chính xác tương đối nhưng thấp hơn một chút, do độ nhạy cao với số lượng mẫu hạn chế ở mỗi lớp.

Sau khi áp dụng SMOTE, tất cả các mô hình đều được cải thiện rõ rệt, trong đó SVC và Random Forest đạt độ chính xác tuyệt đối (1,0), còn XGBoost đạt 0,969. Kết quả này cho thấy quá trình lấy mẫu tăng cường đã nâng cao khả năng biểu diễn đặc trưng, đồng thời giảm thiên lệch về các lớp chiếm ưu thế, giúp các mô hình học được ranh giới phân lớp công bằng và ổn định hơn. Những phát hiện trên củng cố vai trò của SMOTE như một bước tiền xử lý quan trọng trong các bài toán phân loại dựa trên dữ liệu thành phần hóa học, đặc biệt khi mục tiêu là tối ưu hóa hiệu năng mô hình và đảm bảo tính công bằng giữa các lớp trong bối cảnh dữ liệu không cân bằng.

Kết quả của nghiên cứu này cho thấy 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. Trong số các mô hình được đánh giá, mạng nơ-ron nhiều lớp (Multi-Layer Perceptron – MLP) đạt độ chính xác dự đoán cao nhất trong nhận diện giống cam (93,1%), trong khi thuật toán XGBoost cho khả năng phân loại nguồn gốc địa lý tốt nhất (96,9%). Những kết quả này khẳng định tiềm năng mạnh mẽ của các phương pháp phân tích dựa trên dữ liệu trong việc hỗ trợ truy xuất nguồn gốc và xác thực tính xác thực của sản phẩm nông nghiệp.

Nghiên cứu này cũng cung cấp bộ dữ liệu toàn diện đầu tiên về thành phần hóa học của bốn giống cam được trồng tại mười tỉnh khác nhau trên lãnh thổ Việt Nam, qua đó tạo nền tảng khoa học quan trọng cho việc xây dựng chỉ dẫn địa lý và chứng nhận xuất xứ cho cam Việt Nam. Đây là một đóng góp có ý nghĩa không chỉ về mặt học thuật mà còn về mặt quản lý và phát triển thương hiệu nông sản. Bên cạnh giá trị khoa học, các kết quả thu được còn mang ý nghĩa thực tiễn rõ rệt đối với ngành công nghiệp chế biến nước cam trong nước, giúp các nhà sản xuất xác nhận tính xác thực của nguồn nguyên liệu đầu vào, đồng thời cung cấp cho người tiêu dùng sản phẩm nước cam chất lượng cao, có nguồn gốc trong nước, thay thế cho các sản phẩm dựa trên dịch cô đặc nhập khẩu.

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

Các phổ hấp thụ của 111 mẫu cam thu thập từ ba tỉnh của Việt Nam, trải dài từ vùng UV đến NIR, được trình bày trong Hình 3.34 (a: Hòa Bình, b: Quảng Ninh, c: Tuyên Quang). Việc mở rộng miền khảo sát, bao gồm UV-Vis (250 - 780 nm) và NIR (780 - 1000 nm), cho phép đánh giá các đỉnh hấp thụ tương ứng với hỗn hợp đa dạng các hợp chất hoạt tính sinh học (flavonoid, carotenoid...) có trong nước cam.

Từ Hình 3.34, có thể nhận thấy các dải hấp thụ đặc trưng tập trung chủ yếu tại hai bước sóng 280 nm và 330 nm. Dải hấp thụ thứ nhất trong vùng tử ngoại được quy cho các chuyển mức điện tử $\pi \rightarrow \pi^*$ của các hệ vòng thơm kiểu benzoyl và các hợp chất phenolic [105]. Đỉnh này liên quan đến sự hiện diện của các flavonoid như quercetin, hesperidin và naringin [116]. Ngoài ra, vitamin C (acid ascorbic) cũng từng được gán cho dải hấp thụ 260 - 280 nm do hệ enediol với các nhóm hydroxyl liên hợp [88]. Tuy nhiên, việc không quan sát thấy đỉnh này trong Hình 3.34b, cho thấy khả năng tín hiệu tại vùng này chủ yếu bắt nguồn từ các hợp chất phenolic. Trong khi đó, một đỉnh hấp thụ UV đặc trưng khác tại 330 nm được quy cho các chuyển mức cục bộ của các hệ vòng cinnamoyl, chẳng hạn như quercetin [14, 105]. Hệ vòng thơm này bao gồm một vòng benzen liên kết với nhóm carbonyl α, β -không no ba carbon ($-\text{CH}=\text{CH}-\text{CO}-$), với các chuyển mức $\pi \rightarrow \pi^*$ của vòng benzen và $n \rightarrow \pi^*$ của nhóm

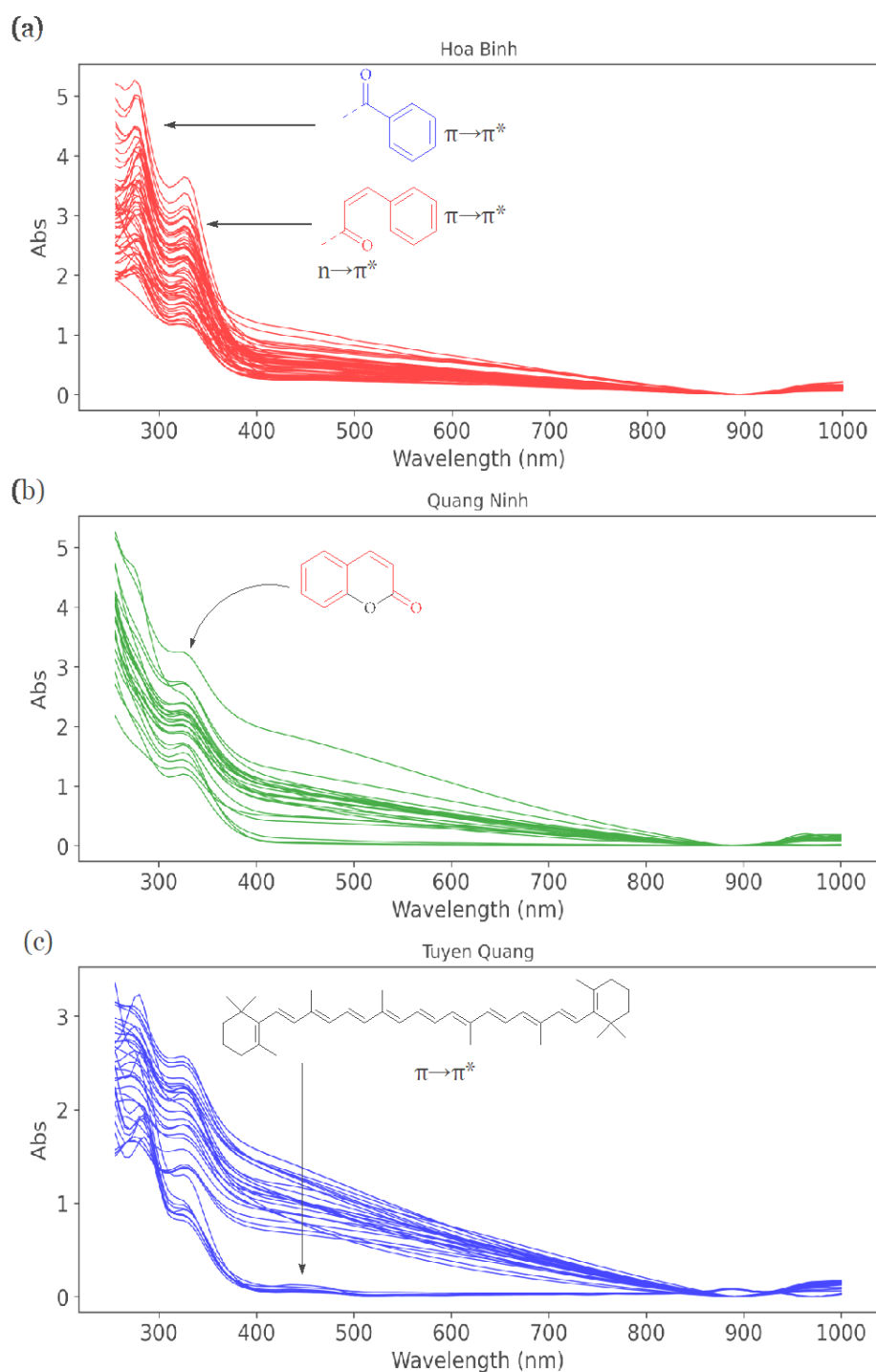
carbonyl (C=O). Tuy nhiên, phần lớn các phổ UV-Vis của mẫu cam từ tỉnh Quảng Ninh (Hình 3.34b) không cho thấy sự xuất hiện của dải hấp thụ thấp hơn, cho thấy rằng đỉnh hấp thụ xấp xỉ 330 nm có thể liên quan đến các nhóm cinnamoyl trong những hợp chất tương tự như các lacton liên hợp (coumarin) [88]. Nhìn chung, nước cam tự nhiên chỉ chứa một lượng rất nhỏ coumarin (auraptene) [73, 126]. Bên cạnh đó, một số dấu hiệu yếu của dải hấp thụ tại 450 nm được ghi nhận trong các mẫu cam Tuyên Quang (Hình 3.34c). Hiện tượng này được cho là do sự hiện diện của carotenoid (ví dụ: β -carotene), là các sắc tố tự nhiên có chuỗi π liên hợp dài [88]. Các hợp chất này hấp thụ trong vùng khả kiến thông qua các chuyển mức $\pi \rightarrow \pi^*$ và là nguyên nhân cơ bản tạo nên màu cam đặc trưng của quả. Gebishu và cộng sự [88] cũng như Zeb và cộng sự [15] đã báo cáo các đỉnh hấp thụ trội tại khoảng 390 nm và 680 nm, lần lượt liên quan đến vitamin A và chlorophyll trong nước cam tươi; tuy nhiên, các dải hấp thụ đặc trưng này không được phát hiện trong các mẫu của nghiên cứu hiện tại.

Nhìn chung, dữ liệu UV-Vis của các mẫu cam từ ba tỉnh khác nhau, đặc biệt là Hòa Bình và Tuyên Quang, cho thấy kiểu phổ tương tự nhau, nhưng chỉ khác biệt về giá trị độ hấp thụ tại từng bước sóng. Quan sát này cho thấy sự tương đồng về thành phần hóa học chính của cam ở các vùng địa lý khác nhau. Cụ thể, các phổ thu được phản ánh thông tin phức tạp về thành phần hóa học của cam thông qua nhiều dải hấp thụ đặc trưng của các chất chống oxy hóa. Trong đó, sự hấp thụ mạnh trong vùng UV (< 400 nm) bắt nguồn từ các chuyển mức $\pi \rightarrow \pi^*$ và $n \rightarrow \pi^*$ của các hợp chất phenolic, flavonoid, vitamin và coumarin, trong khi các hệ liên hợp như carotenoid (tiền chất của vitamin A) hấp thụ trong vùng khả kiến khoảng 400–500 nm do các chuyển mức $\pi \rightarrow \pi^*$.

Một số nghiên cứu đã công bố trước đây [117] cũng ghi nhận các đỉnh hấp thụ tương tự trong phổ của các sản phẩm từ cam, tuy nhiên chưa có sự gán cụ thể cho các dải hấp thụ UV này. Do đó, nghiên cứu hiện tại có thể đóng vai trò như một nguồn tham chiếu nhằm xác định các đỉnh hấp thụ đặc trưng của cam cho các nghiên cứu tiếp theo. Để củng cố và kiểm chứng các kết luận trên, phân tích định lượng bằng

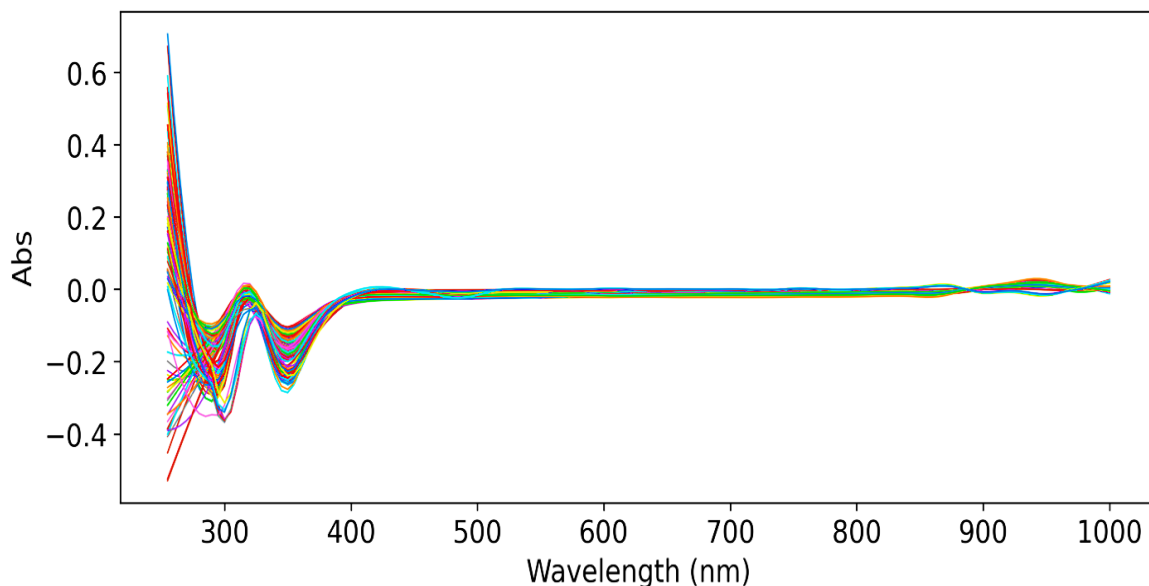
HPLC đã được thực hiện. Ngoài ra, phân tích UV - Vis cho thấy tín hiệu phổ hầu như không bị ảnh hưởng trong vùng bước sóng trên 400 nm. Do đó, việc sử dụng toàn bộ dải bước sóng hay chỉ một phần phổ có thực sự cần thiết cho mục đích nhận diện thành phần hóa học và xác định nguồn gốc địa lý của cam hay không sẽ được đánh giá thông qua các chỉ số độ chính xác của mô hình nghiên cứu.

Mặc dù 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 trong Hình 3.34. Như đã đề cập trước đó, ảnh hưởng của tính phi tuyến, tính đa cộng tuyến và nhiễu là không đáng kể do sự khác biệt phổ giữa các mẫu là rất nhỏ khi sử dụng dữ liệu UV-Vis. Vì vậy, các phương pháp xử lý dữ liệu và đánh giá dựa trên dữ liệu được đề xuất như những công cụ hỗ trợ thiết yếu nhằm trích xuất và làm rõ sự phân biệt giữa các lớp mẫu.



Hình 3.34. Phổ UV-Vis của các mẫu cam ở 3 tỉnh Việt Nam

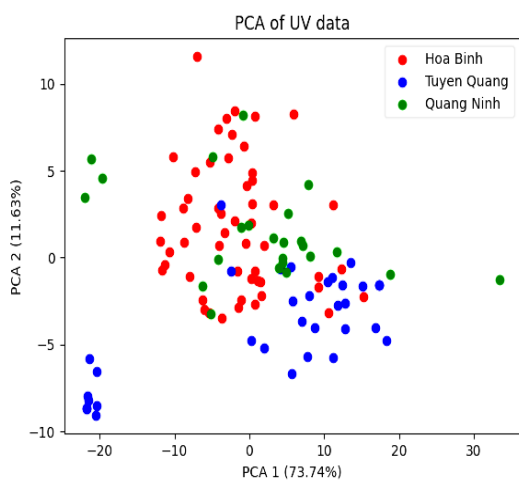
Do đó, một bước tiền xử lý dữ liệu sử dụng tổ hợp SNV + SG + FD đã được áp dụng nhằm chuẩn hóa cường độ phổ và nâng cao độ tin cậy của phân tích phổ. Kết quả hiệu chỉnh phổ bằng các kỹ thuật tiền xử lý này (Hình 3.35) đã làm nổi bật dạng “dấu vân tay” của các nhóm cinnamoyl với đỉnh đặc trưng tại vùng khoảng 330 nm.



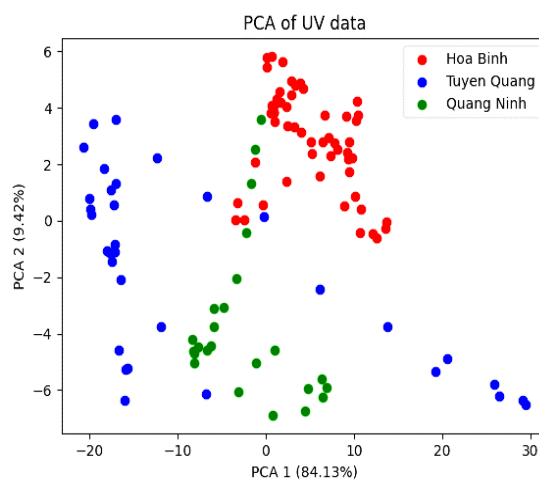
Hình 3.35. Phổ UV - Vis của các mẫu nghiên cứu sau khi xử lý bằng $SNV+SG+FD$.

3.3.2. Phân tích thăm dò dữ liệu bằng PCA

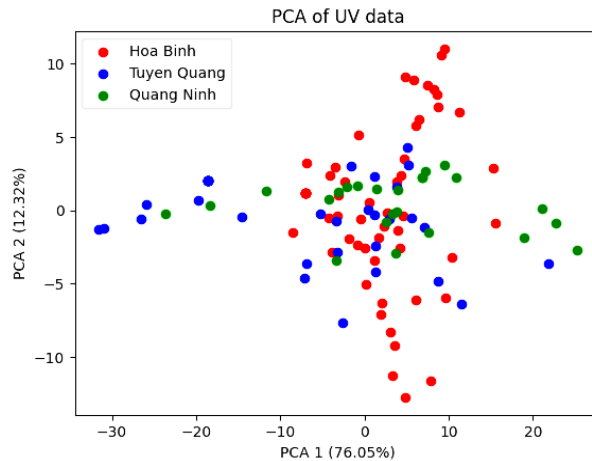
Để hiểu rõ hơn cấu trúc và đặc điểm của bộ dữ liệu, phân tích thành phần chính (PCA) đã được áp dụng cho cả dữ liệu phổ thô và dữ liệu sau tiền xử lý (SNV + SG và SNV + SG + FD), như minh họa trong Hình 3.36. Bên cạnh mục đích trực quan hóa dữ liệu và giảm chiều, việc sử dụng PCA trong nghiên cứu này còn nhằm đánh giá khả năng cải thiện xu hướng phân cụm của các mẫu sau tiền xử lý so với dữ liệu ban đầu.



(a) - phổ thô



(b) phổ sau xử lý SNV + SG



(c) phổ sau xử lý SNV + SG + FD.

Hình 3.36. Biểu đồ PCA của phổ UV - Vis các mẫu nước cam với 2 PC đầu tiên

Trước hết, phân tích đồ thị điểm số PCA hai chiều đối với dữ liệu phổ chưa được làm trơn (Hình 3.36a) cho thấy hai thành phần chính đầu tiên (PC1 và PC2) giải thích 85,4% tổng phương sai tích lũy của bộ dữ liệu. Tuy nhiên, trong phân tích không giám sát này, không thể phân tách rõ ràng ba nhóm mẫu (Hòa Bình, Quảng Ninh và Tuyên Quang), và không có thành phần chính riêng lẻ nào đủ khả năng phân biệt các quần thể cam theo nguồn gốc địa lý. Mặc dù tồn tại xu hướng hình thành cụm trong từng nhóm mẫu, sự phân bố vẫn chồng lấp và phân tán mạnh, dẫn đến độ rõ ràng trực quan thấp. Hiện tượng chồng lấp và phân tán này có thể bắt nguồn từ nhiều nền và sự khác biệt phổ không đáng kể, làm hạn chế khả năng phân biệt nguồn gốc địa lý của nước cam chỉ dựa trên dữ liệu phổ thô.

Nhằm cải thiện mức độ phân tán của các điểm dữ liệu và xu hướng phân cụm, PCA tiếp tục được thực hiện trên các tập dữ liệu đã qua tiền xử lý (Hình 3.36b và 3.36c). Đối với dữ liệu xử lý bằng SNV + SG, hai thành phần chính đầu tiên giải thích 93,5% tổng phương sai, trong khi tổ hợp SNV + SG + FD cho giá trị phương sai tích lũy đạt 88,4%. Mặc dù các chiến lược tiền xử lý này giúp làm tăng tỷ lệ phương sai được giải thích và cải thiện phần nào cấu trúc dữ liệu, khả năng phân tách hoàn toàn ba vùng địa lý vẫn còn hạn chế, cho thấy PCA không đủ mạnh để phân loại rõ ràng nguồn gốc cam trong điều kiện không giám sát.

Tuy nhiên, cần lưu ý rằng các mẫu cam từ những vùng địa lý khác nhau vẫn thể hiện một số đặc điểm phân cụm nhất định trên các đồ thị điểm số PCA. Cụ thể, trong Hình 3.36b, các mẫu cam Hòa Bình chủ yếu phân bố ở giá trị dương của PC1, trong khi các mẫu cam Quảng Ninh tập trung chủ yếu ở vùng giá trị âm của PC1. Ngược lại, các mẫu cam Tuyên Quang có xu hướng phân tán trên nhiều phần tư khác nhau của không gian PCA, cho thấy sự tương đồng nhất định về thành phần hóa học với cả hai nhóm Hòa Bình và Quảng Ninh.

Hiện tượng này có thể được giải thích dựa trên yếu tố địa lý, khi tỉnh Tuyên Quang là khu vực miền núi phía Bắc, nằm giữa hai tỉnh Hòa Bình và Quảng Ninh. Do đó, sự tương đồng về điều kiện sinh thái và khí hậu có thể dẫn đến sự khác biệt hóa học ở mức độ nhỏ, đặc biệt thể hiện qua một số dải hấp thụ nhất định (chẳng hạn vùng 450 nm) trong phổ UV-Vis của các mẫu cam Tuyên Quang.

Tóm lại, 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

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

Để phân biệt nguồn gốc địa lý của cam, bộ dữ liệu gồm 111 mẫu cam sau khi tiền xử lý đã được chia ngẫu nhiên thành hai tập con: tập huấn luyện (chiếm 80% số mẫu, $n = 88$) dùng để xây dựng và huấn luyện mô hình, và tập kiểm tra (chiếm 20% số mẫu, $n = 23$) dùng để đánh giá hiệu năng của các mô hình đã huấn luyện. Nhãn lớp của dữ liệu được mã hóa thành ba nhóm tương ứng với nguồn gốc địa lý: Hòa Bình, Quảng Ninh và Tuyên Quang.

Mặc dù các phân tích PCA trên dữ liệu đã tiền xử lý đã được thực hiện nhằm làm rõ khả năng phân tách theo nguồn gốc địa lý, không có kết quả PCA nào cho thấy sự cải thiện đáng kể trong phân loại địa lý. Do đó, cần nhấn mạnh vai trò của các kỹ

thuật học giám sát, bao gồm các phương pháp phân biệt như LDA, các thuật toán học máy như Random Forest (RF) và các phương pháp học sâu như Mạng nơ-ron nhân tạo (ANN), trong việc phân biệt nguồn gốc địa lý của cam. Xuất phát từ thực tế rằng các phân tích phân loại không ổn định và sai số lớn khi không áp dụng tiền xử lý dữ liệu, nghiên cứu này chỉ triển khai các kỹ thuật học giám sát trên dữ liệu đã được tiền xử lý.

Bên cạnh đó, nhằm đánh giá ảnh hưởng của các miền phổ đến hiệu năng mô hình, phân tích không chỉ giới hạn trong dải UV-Vis truyền thống mà còn mở rộng sang vùng NIR. Cụ thể, hiệu năng của các mô hình được đánh giá trong dải hẹp 250 - 400 nm nhằm kiểm chứng tính cần thiết của việc sử dụng toàn bộ dải bước sóng 250 - 1000 nm đối với hiệu quả phân loại. Kết quả tổng hợp và so sánh hiệu năng của các thuật toán phân loại được trình bày trong Bảng 3.23 (trong đó A_{training} và A_{test} lần lượt là độ chính xác phân loại của tập huấn luyện và tập kiểm tra; các chỉ số còn lại được tính trên tập kiểm tra, trừ khi có ghi chú khác). 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. Cụ thể, tỷ lệ nhận dạng tổng thể của các mô hình học giám sát trên tập dự đoán khi sử dụng ma trận phổ toàn dải đều không thấp hơn 90% với tiền xử lý SNV + SG, và xấp xỉ 95% khi áp dụng quy trình SNV + SG + FD.

Bảng 3.23. Hiệu suất phân loại của các mô hình (%) LDA, RF, ANN

Bước sóng (nm)	Tiền xử lý	Hiệu suất (%)	Mô hình phân loại		
			LDA	RF	ANN
250-400	SNV + SG (order = 0)	A_{test}	100	86,9	86,9
		A_{training}	98,8	100	95,4
	SNV + SG + FD (order = 1)	A_{test}	100	91,3	95,6
		A_{training}	98,8	100	97,7

Bước sóng (nm)	Tiền xử lý	Hiệu suất (%)	Mô hình phân loại		
			LDA	RF	ANN
250-1000	SNV + SG (order = 0)	A_{test}	95,6	91,3	91,3
		$A_{training}$	96,5	100	98,8
		E	4,40	8,70	8,70
		P	96,5	90,9	90,9
		R	97,7	90,9	90,9
		S	98,8	90,9	95,6
		FI	96,5	90,9	90,9
	SNV + SG + FD (order = 1)	A_{test}	95,6	95,6	100
		$A_{training}$	96,5	100	100
		E	4,40	4,40	0
		P	96,5	94,4	100
		R	97,7	97,7	100
		S	98,8	98,8	100
		FI	96,5	95,6	100

Ngược lại, hiệu năng phân loại thấp hơn khi chỉ sử dụng tập con các dải phổ chủ yếu xuất phát từ các mô hình học máy như RF và ANN. Hiện tượng này có thể được giải thích dựa trên đặc điểm của dữ liệu phổ và hành vi của các mô hình trong bài toán nhận dạng. Nhìn chung, RF và ANN là các bộ phân loại phi tuyến phức tạp, đòi hỏi kích thước tập dữ liệu đủ lớn và quá trình tinh chỉnh siêu tham số cẩn thận (ví dụ: số cây, độ sâu cây đối với RF; tốc độ học, số lớp nơ-ron ẩn, hàm kích hoạt đối với ANN) [31]. Do đó, khi được triển khai trên các ma trận dữ liệu có kích thước [m

mẫu x n bước sóng] tương đối nhỏ hoặc trung bình, RF dễ gặp hiện tượng quá khớp, trong khi ANN nhạy cảm hơn với việc lựa chọn siêu tham số, dẫn đến suy giảm hiệu năng phân loại.

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

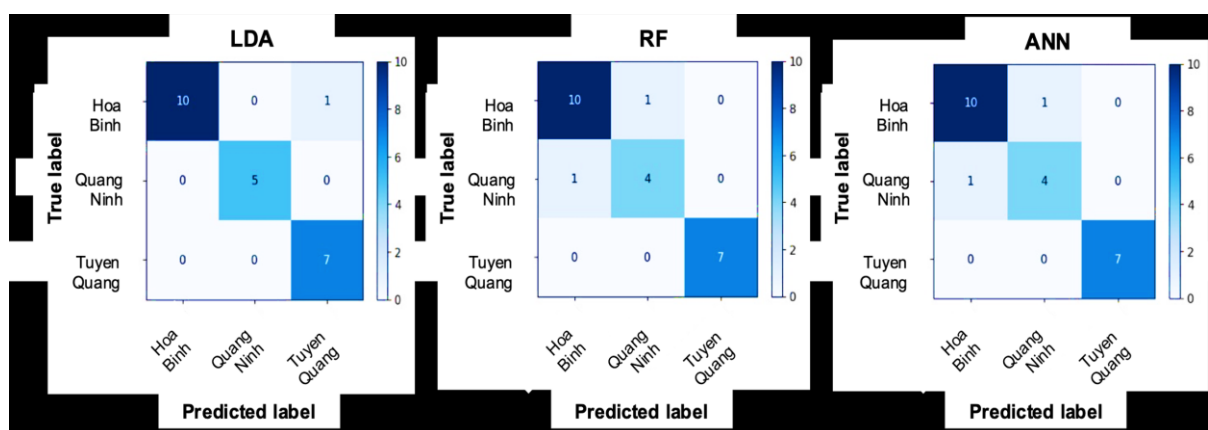
Dựa trên hiệu năng dự đoán trên tập kiểm tra (Bảng 3.23), thuật toán phân tích biệt thức tuyến tính (LDA) đạt độ chính xác phân loại cao nhất, lên tới 100%, đối với cả hai phương pháp tiền xử lý khi dữ liệu chỉ được lựa chọn trong dải UV hẹp (250 - 400 nm). Số lượng thành phần chính (PCs) được lựa chọn để xây dựng mô hình LDA nhằm tối thiểu hóa sai số phân loại là 10 PC. So với LDA, hiệu năng của hai thuật toán học máy RF và ANN thấp hơn khoảng 13,1% khi dữ liệu chỉ được tiền xử lý bằng chuẩn hóa SNV kết hợp làm trơn Savitzky-Golay (SNV + SG). Tuy nhiên, độ chính xác của hai mô hình này được cải thiện đáng kể khi dữ liệu tiếp tục được xử lý bằng đạo hàm bậc nhất (FD), với giá trị đạt 91,3% đối với RF và 95,6% đối với ANN.

Hiệu năng vượt trội của LDA so với các mô hình học máy trong trường hợp dữ liệu có kích thước nhỏ có thể được lý giải bởi việc kết hợp PCA với LDA giúp tăng cường khả năng phân biệt trong không gian có chiều thấp, đồng thời giảm thiểu hiện tượng quá khớp (overfitting) [81, 84]. Điều này cho thấy LDA đặc biệt phù hợp với các tập dữ liệu có số lượng mẫu hạn chế nhưng số biến lớn, như dữ liệu phổ UV-Vis.

Xuất phát từ hiệu năng dự đoán thỏa đáng của các mô hình học giám sát khi sử dụng toàn bộ dải UV-Vis, các chỉ số đánh giá khác bao gồm tỷ lệ lỗi (E), độ chính xác dự đoán (P), độ nhạy (R), độ đặc hiệu (S) và điểm F1 đã được xác định trong Bảng 3.23, dựa trên các ma trận nhầm lẫn tương ứng (Hình 3.37). Khi sử dụng dữ liệu tiền xử lý bằng SNV + SG, kết quả phân loại cho thấy cả ba thuật toán LDA, RF và ANN đều đạt độ chính xác, độ đặc hiệu và điểm F1 lớn hơn 90%. Trên tập kiểm tra ($n = 23$), LDA tiếp tục thể hiện hiệu năng tốt nhất với độ chính xác 95,6%, trong khi RF và ANN đạt độ chính xác thấp hơn, lần lượt là 91,3%.

Cụ thể, mô hình LDA dự đoán sai một mẫu cam Hòa Bình thành nguồn gốc Tuyên Quang; trong khi đó, cả RF và ANN đều nhầm lẫn chéo một mẫu cam giữa

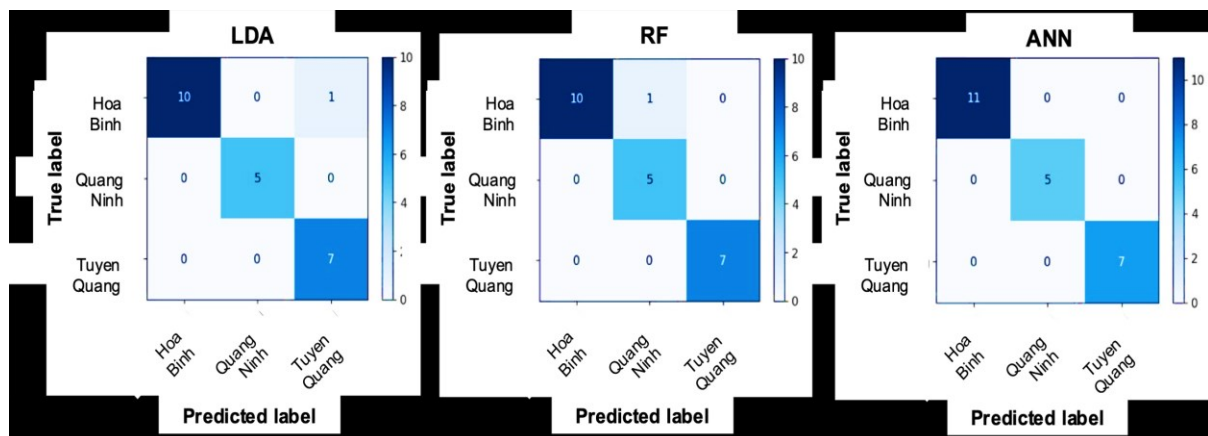
hai tỉnh Hòa Bình và Quảng Ninh (Hình 3.37a - 3.37c). Kết quả này cho thấy các mẫu cam từ ba tỉnh Hòa Bình, Quảng Ninh và Tuyên Quang có thành phần hóa học tương đối tương đồng, dẫn đến việc có nhiều đặc trưng phổ UV-Vis chung. Đồng thời, điều này cũng phản ánh ảnh hưởng của kích thước tập dữ liệu đến hiệu năng dự đoán: thuật toán LDA với số lượng tham số thấp tỏ ra hiệu quả hơn trong không gian dữ liệu có nhiều chiều mà số mẫu không lớn [84]. Ngược lại, độ phức tạp của các mô hình RF và ANN có thể dẫn đến việc xây dựng biên quyết định quá phức tạp, từ đó làm giảm hiệu năng so với bộ phân loại tuyến tính như LDA [23].



(a)- SNV + SG + LDA

(b)- SNV + SG + RF

(c)- SNV + SG + ANN



(d)- SNV + SG + FD + LDA

(e)- SNV + SG + FD + RF

(f)- SNV + SG + FD + ANN

Hình 3.37. Ma trận nhầm lẫn của các mô hình

Khi xét trên tập huấn luyện ($n = 88$), mô hình RF phân loại chính xác 100% các mẫu cam, trong khi ANN và LDA lần lượt dự đoán sai 1,20% và 3,50% số mẫu. Điều này cho thấy các mô hình học máy vượt trội hơn LDA trên tập huấn luyện, thể hiện

xu hướng ngược lại so với tập kiểm tra. Sự đảo chiều này có thể là do kích thước ma trận dữ liệu lớn hơn trong tập huấn luyện, vốn phù hợp hơn với nguyên lý hoạt động của các mô hình phức tạp như RF và ANN.

Để tiếp tục đánh giá ảnh hưởng của tiền xử lý đạo hàm, FD đã được áp dụng sau bước làm mượt và chuẩn hóa nhằm kiểm chứng ưu thế của phương pháp này đối với khả năng truy xuất nguồn gốc. Các ma trận nhầm lẫn thu được từ LDA, RF và ANN trên tập kiểm tra của dữ liệu phổ toàn dải sau tiền xử lý SNV + SG + FD được trình bày trong Hình 3.37d - 3.37f. Kết quả cho thấy khả năng nhận dạng của LDA hầu như không thay đổi trên cả tập huấn luyện và tập kiểm tra khi áp dụng FD (Bảng 3.23). Ngược lại, hiệu năng dự đoán của các mô hình học máy được cải thiện rõ rệt, đặc biệt đối với ANN. Cụ thể, RF đạt độ chính xác 95,6% trên tập kiểm tra, trong khi ANN phân loại hoàn toàn chính xác toàn bộ 11 mẫu Hòa Bình, 5 mẫu Quảng Ninh và 7 mẫu Tuyên Quang. Trên tập huấn luyện, ANN cũng đạt tỷ lệ lỗi bằng 0 trong việc phân biệt các mẫu cam từ ba tỉnh. Việc đạt 100% độ chính xác, độ đặc hiệu và điểm F1 của mô hình ANN cho thấy không có mẫu cam Việt Nam nào bị nhận dạng sai, qua đó khẳng định hiệu năng phân biệt vượt trội của ANN và RF khi kết hợp với tiền xử lý SNV + SG + FD, so với bộ phân loại tuyến tính LDA, ngay cả khi kích thước tập dữ liệu không thay đổi. Sự cải thiện này có thể được giải thích do tác động của phép biến đổi đạo hàm trong tiền xử lý dữ liệu. Khi áp dụng FD, dữ liệu phổ được chuyển sang dạng biểu diễn giàu thông tin hơn và làm nổi bật các cấu trúc phi tuyến, tạo điều kiện cho các mô hình phi tuyến như RF và ANN khai thác hiệu quả hơn các biên phân lớp phức tạp [21, 124]. Trong khi đó, LDA dựa trên giả định về tính phân tách tuyến tính tối ưu, nên không tận dụng được đầy đủ lợi thế của dữ liệu đã được biến đổi đạo hàm. So sánh giữa hai mô hình học máy, ANN thể hiện ưu thế nhẹ so với RF trong việc xử lý tính phi tuyến nhờ khả năng điều chỉnh linh hoạt và học các mẫu liên tục trong dữ liệu [13, 63].

Như vậy, mặc dù 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 sự kết hợp SNV + SG + FD đã nâng cao đáng kể hiệu năng phân loại của hai

mô hình học máy RF và ANN. Kết quả thực nghiệm cho thấy dữ liệu phổ UV-Vis chứa đựng đủ thông tin hóa học để phát triển các mô hình học giám sát phục vụ phân biệt nguồn gốc địa lý của cam Việt Nam. 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 [141]. 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 [49, 109, 140, 147, 148], phương pháp trong nghiên cứu này cho thấy kết quả phân loại nguồn gốc cam tương đương hoặc vượt trội, đồng thời phương pháp này có ưu điểm hơn do chi phí thấp, dễ thực hiện và tốc độ phân tích nhanh khi kết hợp quang phổ UV-Vis với học giám sát.

Tuy nhiên, bên cạnh những ưu điểm rõ rệt, nghiên cứu này vẫn tồn tại một số hạn chế. Cụ thể, kích thước tập dữ liệu còn hạn chế và chỉ tập trung vào một giống cam duy nhất (cam Vinh) có thể ảnh hưởng đến khả năng khái quát hóa của mô hình. Do đó, trong các nghiên cứu tiếp theo, cần mở rộng số lượng mẫu với nhiều giống cam khác nhau, cũng như xem xét các yếu tố môi trường và canh tác trong quá trình lấy mẫu nhằm đánh giá toàn diện hơn độ bền vững và tính ổn định của các mô hình phân loại.

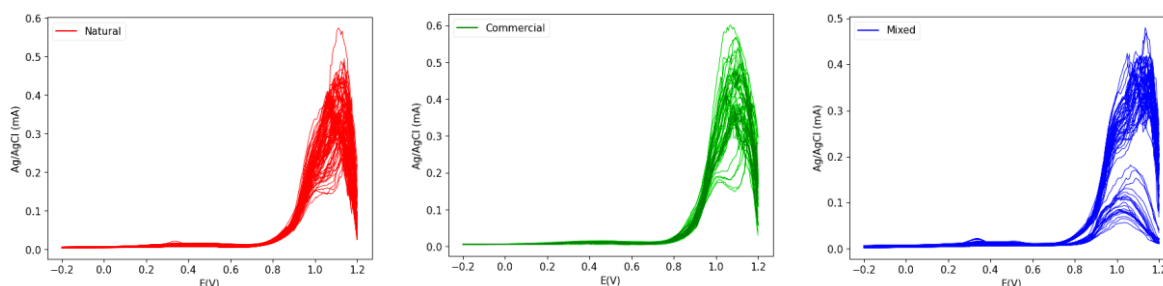
Tóm lại, nghiên cứu đã đề xuất kết hợp giữa quang phổ UV-Vis và các mô hình học giám sát như một giải pháp tốt và có tính thực tiễn cao trong phân loại nguồn gốc địa lý của cam Việt Nam. Những kết quả thu được cho thấy phương pháp này hoàn toàn có thể ứng dụng trong thực tiễn sản xuất và công nghiệp, góp phần nâng cao công tác đảm bảo chất lượng, kiểm soát chất lượng trong lĩnh vực thực phẩm.

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

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

Kết quả đo volt-ampere của 198 mẫu nước cam được trình bày trong Hình 3.38. Các tín hiệu điện hóa thể hiện rõ trong khoảng thế 0,8 - 1,2 V (so với điện cực Ag/AgCl), với dòng cực đại dao động trong khoảng 0,04 - 0,60 mA. Đối với một số mẫu nước cam pha trộn giữa nước cam tự nhiên và nước cam nhân tạo, cường độ dòng cực đại thấp hơn đáng kể, chỉ nằm trong khoảng 0,02 - 0,04 mA.

Tuy nhiên, phần lớn các tín hiệu volt-ampere chỉ thể hiện sự khác biệt nhỏ giữa các mẫu theo nguồn gốc địa lý, đồng thời có hình dạng đỉnh khá tương đồng với các mẫu nước cam tự nhiên và thương mại. Do sự tương đồng về hình dạng đỉnh cùng với biến thiên nhỏ về chiều cao đỉnh nên không thể phân loại bằng trực quan. Hơn nữa, ngay cả trong cùng một nhóm mẫu (tự nhiên hoặc nhân tạo), giá trị dòng cực đại vẫn dao động đáng kể, cho thấy mức độ biến thiên nội nhóm cao. Sự biến thiên này có thể gây ra nhiều đáng kể trong quá trình xây dựng mô hình phân loại, làm giảm khả năng nhận dạng dựa trên tín hiệu thô. Do đó, việc tiền xử lý tín hiệu điện hóa là bước bắt buộc trước khi tiến hành các phân tích phân biệt và phân loại bằng các phương pháp thống kê và học máy, nhằm chuẩn hóa dữ liệu, giảm nhiễu và làm nổi bật các đặc trưng thông tin có ý nghĩa phục vụ cho quá trình nhận diện và phát hiện pha trộn nước cam.



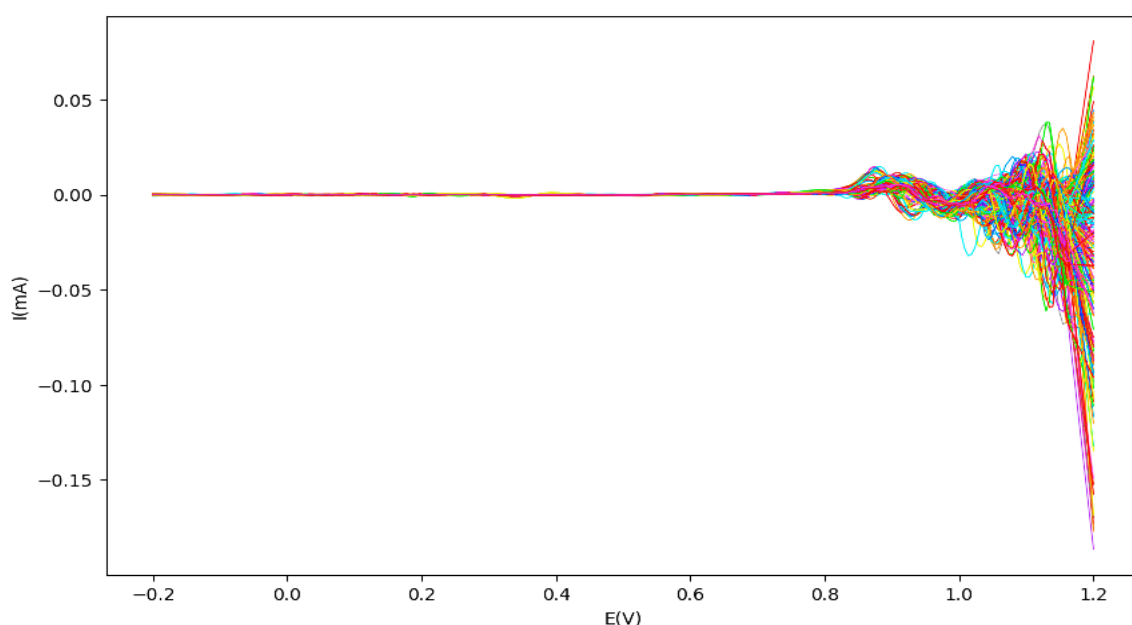
a- nước cam tự nhiên (b)- nước cam thương mại (c)- nước cam pha trộn

Hình 3.38. Volt-ampere đồ của 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. Các đường volt-ampere đồ sau tiền xử lý được trình bày trong Hình 3.39. Sau khi sử dụng các thuật toán tiền xử lý này đã giảm đáng kể nhiễu nền, bao gồm nhiễu xuất phát từ bản thân mẫu, cũng như nhiễu do thiết bị và điều kiện môi trường thí nghiệm gây ra. Đồng thời, đạo hàm bậc hai giúp làm nổi bật sự khác biệt tín hiệu giữa các mẫu, qua đó tăng cường khả năng phân biệt. Những cải thiện này góp phần nâng cao tính ổn định và tính khách quan của các mô hình phân loại, từ đó cải thiện hiệu năng phân loại tổng thể. Sau khi tiền xử lý, toàn bộ tập dữ liệu volt-ampere được chia

ngẫu nhiên thành 158 mẫu (80%) cho tập huấn luyện và 40 mẫu (20%) cho tập kiểm tra, phục vụ việc xây dựng và đánh giá các mô hình học máy.

Cụ thể hơn, dữ liệu ban đầu được tiền xử lý bằng SNV nhằm hiệu chỉnh các hiệu ứng tán xạ và sự biến thiên do độ dày mẫu. Tiếp theo, các phổ sau khi hiệu chỉnh bằng SNV được làm trơn và lấy đạo hàm bằng bộ lọc Savitzky-Golay với cửa sổ 15 điểm, đa thức bậc ba và tính đạo hàm bậc hai, nhằm tăng cường các đặc trưng tín hiệu quan trọng trong khi vẫn giảm thiểu nhiễu. Quy trình tiền xử lý này tạo ra bộ dữ liệu có chất lượng cao hơn, làm cơ sở đáng tin cậy cho việc huấn luyện và đánh giá các mô hình học máy.



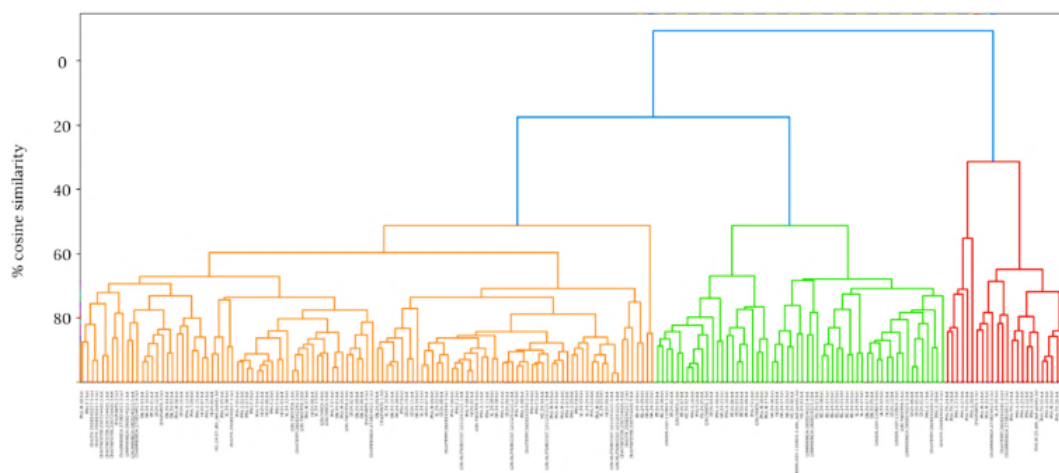
Hình 3.39. Volt-ampere đồ của nước cam sau khi chuẩn hóa bằng SNV, đạo hàm bậc hai và làm mượt với Savitzky-Golay

3.4.2. Nhận diện và xác thực nước cam

Hình 3.40 trình bày kết quả phân tích phân cụm thứ bậc (Hierarchical Clustering Analysis - HCA) được thực hiện trên toàn bộ tập dữ liệu sau tiền xử lý bằng chuẩn hóa SNV, sử dụng độ tương đồng cosine làm thước đo khoảng cách. Độ tương đồng cosine được lựa chọn cho phân cụm thứ bậc do đặc tính nhấn mạnh sự tương đồng về hình dạng tín hiệu hơn là giá trị cường độ tuyệt đối. Trong dữ liệu volt-ampere, cường độ dòng điện có thể thay đổi do quá trình chuẩn bị mẫu, mức độ pha loãng hoặc nhiễu

từ thiết bị, trong khi hình dạng tổng thể của volt-ampere đồ vẫn mang tính đặc trưng cho từng loại mẫu.

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 (Bảng 3.24). Điều này chứng tỏ rằng dấu vân tay volt-ampe, sau khi được tiền xử lý thích hợp, đã chứa đựng đủ thông tin đặc trưng để phục vụ cho việc nhận diện và xác thực nước cam. Đồng thời, kết quả phân cụm này cũng tạo tiền đề thuận lợi cho việc triển khai các mô hình học giám sát nhằm nâng cao hơn nữa độ chính xác và độ tin cậy trong phân loại và phát hiện pha trộn.



Hình. 3.40. Phân cụm phân cấp sử dụng tương đồng cosin

Kết quả phân tích cho thấy 80% các mẫu nước cam pha trộn được phân vào cụm thứ nhất, trong khi 62% các mẫu nước cam tự nhiên tập trung ở cụm thứ hai. Tuy nhiên, tại cụm thứ ba, sự phân bố của các mẫu thuộc ba nhóm không thể hiện sự phân biệt rõ ràng. Do đó, các thuật toán không giám sát chỉ cho thấy sự khác biệt tương đối rõ giữa nhóm nước cam pha trộn và nước cam tự nhiên, nhưng chưa đủ khả năng phân tách hoàn toàn tất cả các nhóm mẫu. Như vậy, dữ liệu volt-ampere, dù đã được tiền xử lý, vẫn tồn tại sự chồng lấp đáng kể giữa các nhóm khi sử dụng các phương pháp không giám sát. Tuy vậy, việc quan sát được sự khác biệt ban đầu giữa các nhóm mẫu cũng tạo cơ sở quan trọng cho việc triển khai các mô hình học máy có

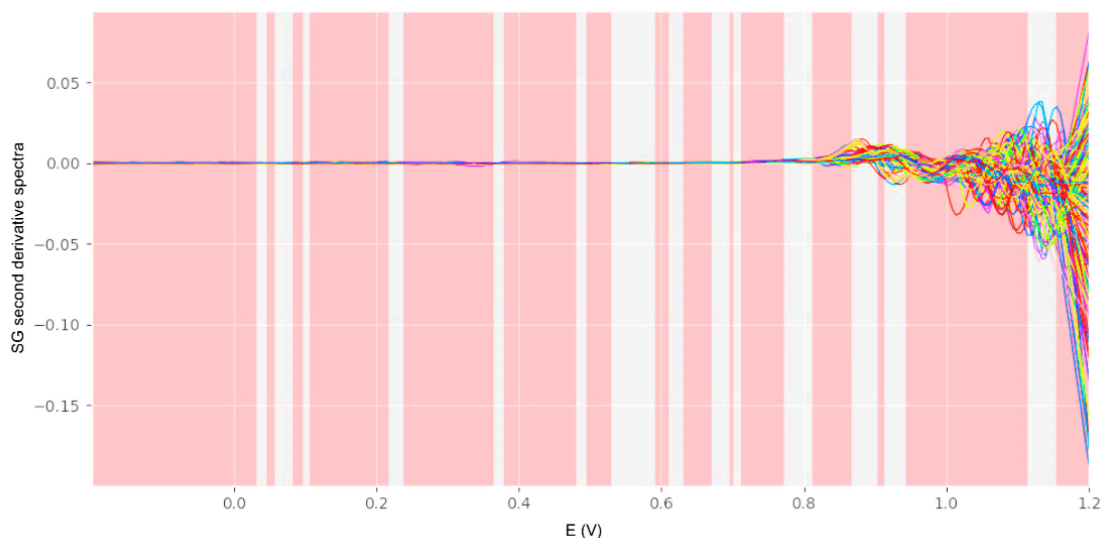
giám sát, nhằm khai thác hiệu quả hơn các đặc trưng tiềm ẩn trong dữ liệu và nâng cao độ chính xác trong bài toán phân loại và phát hiện pha trộn nước cam.

Bảng 3.24. Phần trăm mỗi loại mẫu trong nhóm phân loại

Cụm	Pha trộn	Thương mại	Tự nhiên
1	80%	16%	4%
2	16%	22%	62%
3	33%	27%	40%

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

Sau khi áp dụng đạo hàm bậc hai, bộ dữ liệu thu được có số lượng đặc trưng tương đối lớn (tổng cộng 284 đặc trưng), đặt ra yêu cầu cần giảm chiều dữ liệu trước khi xây dựng các mô hình học máy. Do đó, thuật toán iPLS đã được áp dụng trên tập huấn luyện nhằm lựa chọn tập hợp đặc trưng tối ưu cho quá trình xây dựng mô hình phân loại. Thuật toán iPLS cho phép lựa chọn các vùng phổ có khả năng phân biệt cao nhất, đồng thời loại bỏ các vùng phổ nhiễu hoặc chứa ít thông tin, qua đó giúp giảm nguy cơ quá khớp (overfitting) và nâng cao độ chính xác của mô hình. Kết quả lựa chọn đặc trưng bằng iPLS được trình bày trong Hình 3.41. Trong số 284 đặc trưng ban đầu, 205 đặc trưng có tiềm năng phân loại thấp (được tô màu hồng trong hình) đã bị loại bỏ, chỉ giữ lại 79 đặc trưng có giá trị phân biệt tốt nhất. Tập 79 đặc trưng được chọn lọc này tiếp tục được sử dụng cho cả tập huấn luyện và tập kiểm tra, nhằm đảm bảo tính nhất quán trong quá trình đánh giá độ chính xác và hiệu năng của các mô hình học giám sát trong các phân tích tiếp theo. Việc giảm đáng kể số chiều dữ liệu không chỉ giúp đơn giản hóa mô hình, mà còn góp phần tăng tính ổn định và khả năng khái quát hóa của các thuật toán phân loại trong bài toán phát hiện pha trộn nước cam.



Hình 3.41. Lựa chọn các đặc trưng bằng iPLS

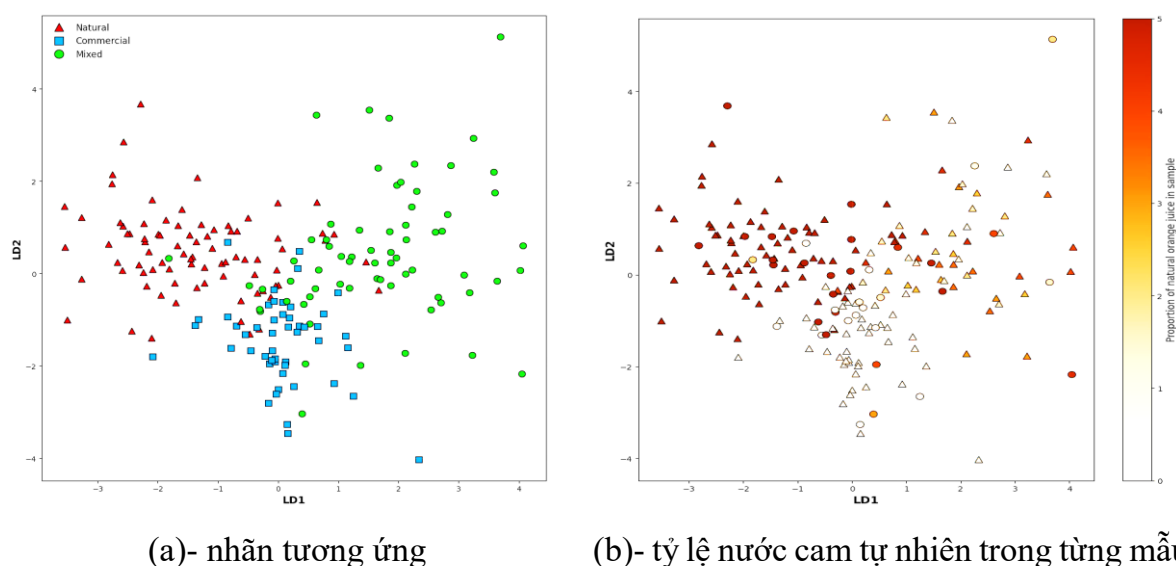
Trong quá trình lựa chọn mô hình, các tiêu chí như tính tuyến tính, độ chính xác, khả năng phân biệt nhãn (đặc biệt giữa nước cam tự nhiên và nước cam thương mại), cũng như hiệu năng đối với tập dữ liệu kích thước nhỏ đã được xem xét. Thông thường, các mô hình PLS-DA, RF, SVM và ANN được ưu tiên cho các bài toán phân loại. Tuy nhiên, trong nghiên cứu này, ANN và SVM chỉ đạt độ chính xác dự đoán khoảng 75%, do đó không đáp ứng yêu cầu về hiệu năng. Bên cạnh đó, kích thước tập dữ liệu hạn chế làm suy giảm khả năng học của ANN, trong khi SVM thể hiện sự nhầm lẫn đáng kể giữa các mẫu nước cam tự nhiên và nước cam thương mại, làm giảm tính ứng dụng của mô hình. Vì vậy, nghiên cứu này chỉ tập trung vào hai mô hình phù hợp hơn với tập dữ liệu, bao gồm iPLS-LDA và RF.

3.4.3.1. Mô hình iPLS-LDA

Thuật toán iPLS được áp dụng để giảm chiều dữ liệu trên tập huấn luyện kích thước 158×79 , bao gồm 158 mẫu được phân loại thành ba nhóm (nước cam tự nhiên, nước cam thương mại và nước cam pha trộn), với 79 đặc trưng được trích xuất từ dữ liệu thô. Trên cơ sở đó, mô hình LDA được xây dựng nhằm phân loại các nhóm mẫu này.

Kết quả cho thấy một số mẫu nước cam pha trộn bị phân loại nhầm sang nhóm nước cam tự nhiên hoặc nước cam thương mại. Sự phân tách tuyến tính hoàn toàn

không đạt được do sự chồng lấp giữa các nhóm tự nhiên và pha trộn. Để đánh giá sâu hơn hiện tượng này, tỷ lệ phần trăm nước cam tự nhiên trong các mẫu được biểu diễn chi tiết trong Hình 3.42, trong đó màu sắc mẫu chuyển dần từ sáng sang đậm tương ứng với tỷ lệ nước cam tự nhiên tăng dần. Các ký hiệu hình tam giác biểu thị các mẫu thuộc tập huấn luyện, hình tròn đại diện cho các mẫu thuộc tập kiểm tra. Các kết quả cho thấy mô hình LDA gặp khó khăn trong việc phân biệt giữa nước cam tự nhiên và nước cam pha trộn. Điều này cho thấy các hợp chất có mặt trong nước cam tự nhiên có thể che lấp hoặc làm suy yếu các tín hiệu đặc trưng, từ đó làm tăng độ phức tạp của bài toán phân loại khi sử dụng bộ phân loại tuyến tính. Nhận định này cũng giải thích vì sao LDA, dù có ưu thế về tính đơn giản và ổn định với tập dữ liệu nhỏ, vẫn bị hạn chế trong trường hợp dữ liệu có ranh giới lớp phi tuyến.



Hình 3.42. Biểu đồ LDA của tập huấn luyện và tập kiểm tra được chiếu lên LDA

Dựa trên Hình 3.42a, có thể nhận thấy rằng khi tập huấn luyện và tập kiểm tra được chiếu vào không gian LDA theo các nhãn phân loại tương ứng, xuất hiện sự chồng lấp đáng kể giữa các cụm mang nhãn “Pha trộn (Mixed)” và “Thương mại (Commercial)”. Hiện tượng này giải thích vì sao một số mẫu bị phân loại nhầm vào nhóm “Pha trộn”. Sự chồng lấp phản ánh mức độ tương đồng đặc trưng giữa các mẫu nước cam thương mại và nước cam pha trộn, đặc biệt trong trường hợp tỷ lệ nước cam tự nhiên trong sản phẩm biến thiên, khiến mô hình gặp khó khăn trong việc phân biệt rạch ròi giữa các nhóm. Hình 3.42b cung cấp thêm thông tin khi các mẫu được

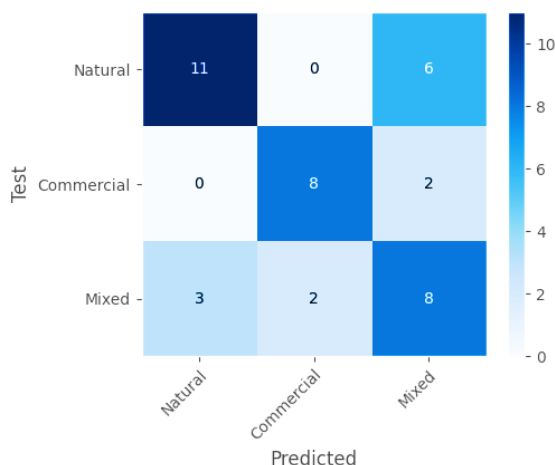
chiều vào không gian LDA theo tỷ lệ phần trăm nước cam tự nhiên. Kết quả cho thấy các mẫu có tỷ lệ nước cam tự nhiên nằm gần ngưỡng chuyển tiếp giữa các nhóm “Pha trộn” - “Tự nhiên” hoặc “Pha trộn” - “Thương mại” có xu hướng bị phân loại sai cao hơn. Điều này cho thấy ranh giới phân lớp trong không gian LDA chưa đủ mạnh, không phản ánh đầy đủ sự khác biệt liên tục về hàm lượng nước cam tự nhiên giữa các nhóm mẫu.

Nhìn chung, các kết quả trên chỉ ra rằng giả định phân tách tuyến tính của LDA bị 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.

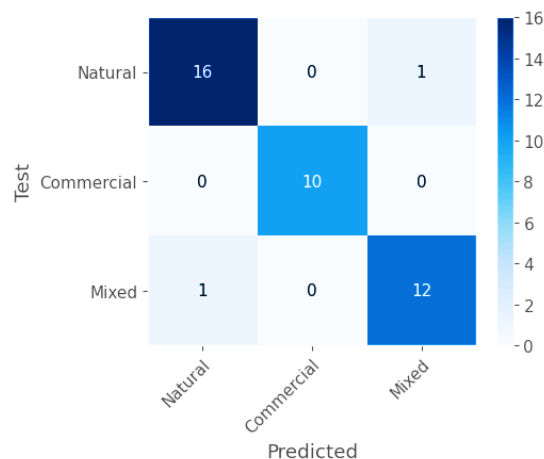
3.4.3.2. Mô hình *iPLS-RF*

Khi sử dụng tập dữ liệu kiểm tra, mô hình *iPLS-RF* đạt độ chính xác phân loại 95%, cho thấy sự cải thiện rõ rệt so với thuật toán LDA (độ chính xác 67,5%) như minh họa trong Hình 3.43. Dựa trên kết quả của mô hình *iPLS-RF*, chỉ có một mẫu nước cam tự nhiên bị phân loại nhầm sang nhóm nước cam pha trộn, và một mẫu nước cam pha trộn bị dự đoán nhầm là nước cam tự nhiên. Đáng chú ý, không xảy ra hiện tượng nhầm lẫn giữa hai nhóm nước cam pha trộn và nước cam thương mại, cho thấy khả năng phân biệt vượt trội của mô hình.

Kết quả này làm nổi bật tiềm năng ưu việt của thuật toán RF, cũng như các phương pháp học tập tổ hợp (ensemble learning) nói chung, trong việc phân loại đối tượng dựa trên các đặc trưng phổ đã được chọn lọc. Khả năng xử lý tính phi tuyến, tương tác phức tạp giữa các biến, cùng với cơ chế tổng hợp quyết định từ nhiều cây phân loại, giúp RF khắc phục hiệu quả các hạn chế của mô hình tuyến tính như LDA. Do đó, *iPLS-RF* là một giải pháp tốt, đáng tin cậy cho bài toán phát hiện và phân loại pha trộn nước cam dựa trên dữ liệu phổ điện hóa.



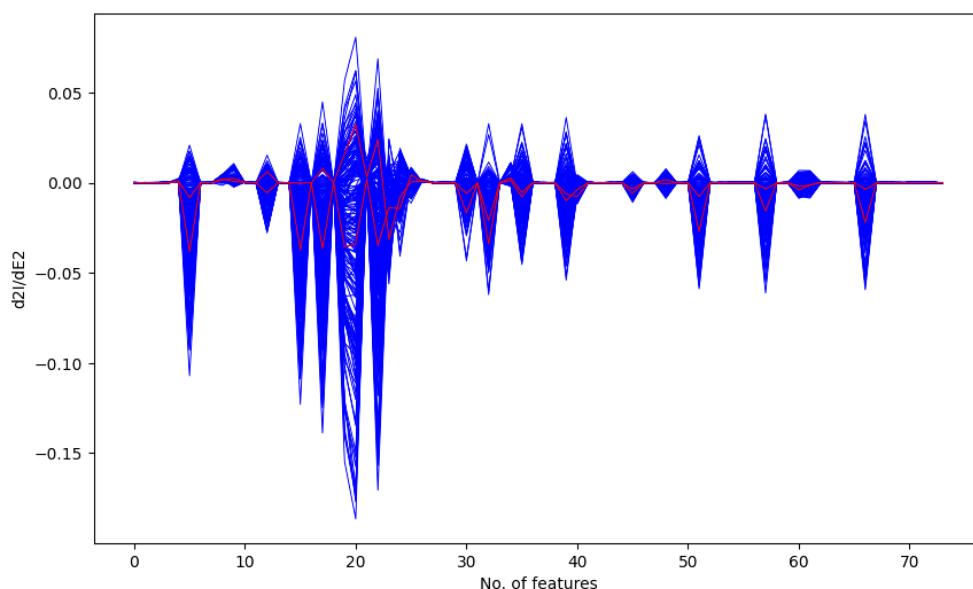
(a)- LDA trên tập kiểm tra



(b)- RF trên tập kiểm tra

Hình 3.43. Ma trận nhầm lẫn mô hình iPLS

Mặc dù đạt hiệu năng phân loại cao, mô hình RF vẫn gặp khó khăn trong việc phân biệt chính xác giữa hai nhóm nước cam tự nhiên (natural) và nước cam pha trộn (mixed), đặc biệt đối với các mẫu có tỷ lệ nước cam tự nhiên trung bình, khoảng 45 - 50%. Hiện tượng này có thể bắt nguồn từ việc các đặc trưng dữ liệu hiện có chưa đủ khả năng tách biệt hiệu quả hai nhóm này trong không gian đặc trưng. Cụ thể, các mẫu có hàm lượng nước cam tự nhiên nằm trong vùng trung gian tạo thành dải chuyển tiếp liên tục về thành phần hóa học và đặc trưng phổ, dẫn đến ranh giới phân lớp không rõ ràng. Điều này cho thấy rằng, ngay cả đối với các mô hình phi tuyến mạnh như RF, chất lượng của đặc trưng đầu vào vẫn đóng vai trò quyết định trong việc nâng cao khả năng phát hiện pha trộn. Nhận định này cũng gợi ý rằng việc bổ sung thêm đặc trưng hóa học đặc hiệu hoặc kết hợp đa kỹ thuật phân tích có thể giúp cải thiện hơn nữa hiệu năng phân loại trong các nghiên cứu tiếp theo.



Hình 3.44. Mối tương quan giữa tín hiệu đạo hàm bậc hai và các đặc trưng của mẫu trong mô hình iPLS-LDA (đỏ: mẫu phân loại sai; xanh: mẫu phân loại đúng)

Để làm rõ nguyên nhân khiến mô hình iPLS-DA cho độ chính xác dự đoán thấp, các tín hiệu đạo hàm bậc hai của các mẫu bị phân loại sai (đường màu đỏ) đã được so sánh với các mẫu được phân loại đúng (đường màu xanh), như trình bày trong Hình 3.44. Kết quả cho thấy hình dạng của dữ liệu đạo hàm bậc hai theo các đặc trưng có sự khác biệt rõ rệt giữa hai nhóm mẫu. Cụ thể, các mẫu bị phân loại sai thường xuất hiện những bất thường như thiếu đỉnh đặc trưng, tín hiệu bị làm phẳng hoặc không có sự biến thiên đáng kể. Những bất thường này đồng thời liên quan đến các điểm ngoại lai tại các vị trí đỉnh hoặc vai đỉnh trên đường volt-ampere đồ. Điều này cho thấy rằng các mẫu có hình dạng tín hiệu bất thường so với tổng thể dữ liệu trong cùng một nhóm cần được nhận diện và loại bỏ trước khi tiến hành các bước xử lý và xây dựng mô hình. Việc kiểm soát chất lượng dữ liệu đầu vào như vậy là cần thiết nhằm nâng cao độ ổn định và độ tin cậy của các mô hình phân loại dựa trên dữ liệu volt-ampere.

Tóm lại, sử dụng dữ liệu volt-ampere kết hợp với các mô hình học máy là phương pháp sàng lọc hiệu quả để xác định tính xác thực của nước cam. Phân tích phân cụm cho thấy sự khác biệt rõ rệt giữa nước cam tự nhiên, nước cam có nguồn gốc hóa học và các mẫu pha trộn giữa nước cam tự nhiên và nước cam nhân tạo. Trên

cơ sở đó, các mô hình học giám sát, bao gồm iPLS-LDA và RF, đã được xây dựng để phục vụ mục tiêu phân loại, trong đó mô hình RF thể hiện kết quả tốt về khả năng phân loại và dự đoán. Các kết quả thu được cũng nhấn mạnh tính hữu ích của phương pháp volt-ampere sử dụng điện cực vàng trong việc đánh giá tính xác thực của nước cam. Những hạn chế của phương pháp nhận dạng mẫu đã được xem xét thông qua việc mở rộng kích thước tập dữ liệu, qua đó góp phần xây dựng các mô hình mạnh và bền vững hơn. Bên cạnh đó, phương pháp điện hóa cho thấy nhiều ưu điểm như thời gian phân tích nhanh, chi phí thấp và khả năng triển khai tại hiện trường, phù hợp cho các ứng dụng kiểm tra nhanh.

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: H_2O : $\text{CH}_3\text{COOC}_2\text{H}_5$ = 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 Hảo, **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*, trang 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 Hảo, 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), trang 8 - 17.

4. **Nguyễn Đức Thanh**, Nguyễn Thị Quyên, Nguyễn Thị Vân Anh, Lê Việt Ngân, Lê Thị Hồng Hảo, 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), trang 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.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

1. Đỗ Huy Bích và cộng sự (2004), *Cây thuốc và động vật làm thuốc ở Việt Nam*, NXB Khoa học Kỹ thuật, Hà Nội.
2. Bộ Y tế (2007), *Dược lý học*, NXB Y học, Hà Nội.
3. Lê Doãn Diên (2004), *Công nghệ sau thu hoạch thuộc ngành nông nghiệp Việt Nam trong xu thế hội nhập và toàn cầu hóa*, NXB Nông nghiệp, Hà Nội.
4. Lê Văn Hòa, Phan Thị Xuân Thủy (2010), "Cải thiện màu sắc và phẩm chất trái cam soàn (*citrus sinensis* (L.) Cv. Soan) bằng biện pháp xử lý hóa chất trước thu hoạch", *Tạp chí Khoa học Trường Đại học Cần Thơ* Tập (16a), tr. 178 - 187.
5. Nguyễn Văn Lợi, Nguyễn Thị Minh Tú, Hoàng Đình Hòa (2012), "Nghiên cứu tách chiết và xác định hoạt tính sinh học của các thành phần tạo hương trong tinh dầu vỏ Bưởi và vỏ cam của Việt Nam", *Tạp chí khoa học và công nghệ* Tập 51(2), tr. 153 - 162.
6. Nguyễn Thiện Luân, Lê Doãn Diên, Phan Quốc Kinh (1999), *Các loại thực phẩm thuốc, thực phẩm chức năng ở Việt Nam*, NXB Nông nghiệp, Hà Nội.
7. Trần Thị Huyền Nga (2010), *Nghiên cứu chiết xuất, tinh chế và xác định hoạt tính sinh học của một vài carotenoid từ cây cỏ Việt Nam dùng để sản xuất thực phẩm chức năng*, Luận án tiến sĩ sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQGHN.
8. Nguyễn Bá Phú, Nguyễn Quốc Sĩ, Nguyễn Bảo Vệ (2021), "Đặc tính hình thái thực vật của cây cam Soàn (*Citrus sinensis* (L.) cv. Soan) không hạt được phát hiện tại tỉnh Hậu Giang", *Tạp chí khoa học và nông nghiệp Việt Nam*, tr. 1567-1575.

9. Mai Thành Thái, Tô Nguyễn Phước Mai, Nguyễn Văn Mười (2022), “Tính chất hóa lý và các hợp chất sinh học của quả cam sành được trồng tại Vĩnh Long”, *Tạp chí công thương* Số 22, tr 388 - 393.
10. Nguyễn Đức Thanh, Nguyễn Thị Thanh Hương (2020), *Phân tích công cụ*, NXB Quân đội nhân dân, Hà nội.
11. Trần Văn Thanh, Trần Đức Hiệp (2022), "Nghiên cứu những cây thuốc chi Citrus ở Việt Nam có flavonoid để sản xuất chế phẩm Citroflavonoid", *Tạp chí Khoa học và công nghệ Trường đại học Hòa Bình* Số 6, tr. 130-144.

Tiếng Anh

12. A. El Orche, M. El Karbane, O. Ait El Alia, H. Bouchafra, L. Zarayby, M. Bouatia (2025), “Rapid Authentication of Moringa oleifera Powder: Detection and Quantification of Adulteration Using ATR-FTIR Spectroscopy and Chemometrics”, *Food Analytical Methods* Vol.18, pp. 898 - 910.
13. A. Mustapha, I. Ishak, N.N.M. Zaki, M.R. Ismail-Fitry, S. Arshad, A.Q. Sazili (2024), “Application of machine learning approach on halal meat authentication principle, challenges, and prospects: A review”, *Heliyon* Vol.10, e32189.
14. A.M. Pupin, M.J. Dennis, M.C.F Toledo (1999). “HPLC analysis of carotenoids in orange juice”, *Food Chemistry* Vol. 64(2), pp. 269 - 275.
15. A. Zeb, W.S. Qureshi, A. Ghafoor, A. Malik, M. Imran, A. Mirza, M.I. Tiwana, E. Alanazi (2023), “Towards sweetness classification of orange cultivars using short-wave NIR spectroscopy”, *Scientific Reports* Vol.3, e325.
16. A. Zeb, W.S. Qureshi, A. Ghafoor, D.O. Sullivan (2022), “Learning Fruit Class from Short Wave Near Infrar”, *8th International Conference on Mechatronics and Robotics Engineering (ICMRE)*, pp. 193-196.

17. Abdel-Salam OM, Youness ER, Mohammed NA, Morsy SM, Omara EA, Sleem AA. (2014), “Acid citric effects on brain and liver oxidative stress in lipopolysaccharide-treated mice”, *J Med Food* Vol. 17(5):588-98.
18. AIJN – European Fruit Juice Association (2024), *Reference Guideline for Orange Juice*, AIJN, Brussels.
19. Altmeyer PJ, Matthes U, Pawlak F, et al. (1994), “Antipsoriatic effect of acid fumaric derivatives. Results of a multicenter double-blind study in 100 patients”, *Journal of the American Academy of Dermatology* Vol.30(6), pp. 977-981.
20. AOAC Research Institute (2016), *Guidelines for Standard Method Performance Requirements*, AOAC Official Methods of Analysis, AOAC international, United States of America, Appendix F: 1-18.
21. B. Ozen, C. Cavdaroglu, F. Tokatli (2024), “Trends in authentication of edible oils using vibrational spectroscopic techniques”, *Analytical Methods* Vol.16, pp. 4216-4233.
22. Belajova E., Suhaj M. (2004), “Determination of phenolic constituents in citrus juices: Method of high performance liquid chromatography”, *Food Chemistry* Vol. 86(3), pp. 339 - 343.
23. Barnes R.J. Dhanoa M.S. Lister S.J. (1989), “Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra”, *Applied Spectroscopy* Vol.43 (5), pp.772-777.
24. Bi S. Sun S. Lao F. Liao X. Wu J. (2020), “Gas chromatography–mass spectrometry combined with multivariate data analysis as a tool for differentiating between processed orange juice samples on the basis of their volatile markers”, *Food Chemistry* Vol.311, e125913.
25. Bishop C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York.

26. Boulesteix A.L. et al. (2012), “Overview of random forest methodology and practical guidance”, *WIREs Data Mining and Knowledge Discovery* Vol.2, pp . 493–507.
27. Breiman L. (2001), “Random forests”, *Machine Learning* 45, pp.5–32.
28. Brendel R., Schwolow S., Rohn S., Weller P. (2021), “Volatilomic profiling of citrus juices by HS-GC-MS-IMS and machine learning”, *Journal of Agricultural and Food Chemistry* Vol.69 (5), pp.1727-1738.
29. C. Martinez-Castillo, G. Astray, J.C. Mejuto, J. Simal-Gandara (2020), “Random Forest, Artificial Neural Network, and Support Vector Machine Models for Honey Classification”, *eFood* Vol.1, pp.69-76.
30. Chen T. Guestrin C. (2016), “XGBoost: A scalable tree Boosting system”, *Proceedings of the 22nd ACM SIGKDD Conference*, pp.785–794.
31. Cortes C. Vapnik V. (1995), “Support-vector networks”, *Machine Learning* 20, pp.273–297.
32. Cozzolino D. (2020), “Use of infrared spectroscopy and chemometrics for food authenticity and traceability”, *Food Research International* Vol. 137(1), pp.109541-109548.
33. D. Dimakopoulou-Papazoglou, N. Ploskas, S. Serrano, C.S. Silva, V. Valdramidis, K. Koutsoumanis, E. Katsanidis (2023), “Application of UV–Vis spectroscopy for the detection of adulteration in Mediterranean honeys”, *European Food Research and Technology* Vol.249, pp.3043-3053.
34. D. Pullancheri, G. Vaidyanathan, N. Gayathree (2013), “Qualitative and Quantitative Analyses of Water Soluble Vitamins and Flavonoids in Pomegranate Aril Juice, Skin, and Commercially Available Fruit Juice Using the ACQUITY UPLC H-Class with PDA Detector”, *Waters Corporation*.

35. D. Suhandy, M. Yulia (2021), "The Use of UV Spectroscopy and SIMCA for the Authentication of Indonesian Honeys According to Botanical, Entomological and Geographical Origins", *Molecules* Vol.26, e915.
36. Dhuique-Mayer, C., et al. (2005), "Varietal and interspecific influence on micronutrient contents in citrus from the Mediterranean area", *Journal of Agricultural and Food Chemistry* Vol. 53(6), pp. 2140 - 2145.
37. Donaldson S., (2004). "Nutrition and cancer: A review of the evidence for an anti-cancer diet", *Nutr J*, Vol.3(1), pp.19-20.
38. E. Ryckebosch, S.P.C. Bermúdez, R. Termote-Verhalle, C. Bruneel, K. Muylaert, R. Parra-Saldivar, I. Foubert (2014), "Influence of extraction solvent system on the extractability of lipid components from the biomass of *Nannochloropsis gaditana*", *Journal of Applied Phycology* Vol.26, pp.1501-1510.
39. E.d.B. Santos, N.V. Madalossi, F.A. Sigoli, I.O. Mazali (2015), "Silver nanoparticles: green synthesis, self-assembled nanostructures and their application as SERS substrates", *New Journal of Chemistry* Vol.39, pp.2839-2846.
40. Ehsani S., Homayouni A., Ahbhashlo M. (2023), "Portable near-infrared spectroscopy coupled with chemometrics for rapid authentication of orange juice", *Food Control* Vol. 148 (1), pp. 109648-109657.
41. Ehsani S. Zubiaur, J.A. García-González, F.J. Callejón R.M. (2023), "NIR spectroscopy combined with chemometrics for detection of pulp wash addition in orange juice", *Microchemical Journal* Vol.194, e109304.
42. Ellis D.I., Muhamadali H., Allen D.P. (2022), "Fingerprinting food: Current technologies and future trends", *TrAC Trends in Analytical Chemistry* Vol. 152, pp. 116635-116646.
43. Erdman J.W.(2005), "How do nutritional and hormonal status modify the Bioavailability, uptake and distribution of different isomers of lycopene", *J. Nutr.* Vol.135, pp. 2046S-2047S.

44. Escamilla-García, B., et al. (2014), "Determination of Flavanones in Orange Juices Obtained from Different Sources by HPLC/DAD", *Journal of Analytical Methods in Chemistry*, Vol. 2014, art. 296838.
45. Eva Bauersachs Andreas Dunkel Veronika Mall Klaas Reglitz Martin Steinhaus (2025), "Using a combined volatilomics and sensomics approach to identify candidate markers for the differentiation of variously preserved not from concentrate (NFC) orange juices", *Food Chemistry* 480, e143966.
46. Evans, J.R.; Lawrenson, J.G. (2017). "Antioxidant vitamin and mineral supplements for slowing the progression of age-related macular degeneration". *Cochrane Database Syst* 9(9), eCD000254.
47. F. Ammari, L. Redjdal, D.N. Rutledge (2015), "Detection of orange juice frauds using front-face fluorescence spectroscopy and Independent Components Analysis", *Food Chemistry* Vol.168, pp.211-217.
48. F.J. Cuevas, G. Pereira-Caro, J.M. Moreno-Rojas, J.M. Muñoz-Redondo, M.J. Ruiz-Moreno (2017), "Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-MS combining data fusion and chemometrics", *Food Control* Vol.82, pp. 203-211.
49. F.J. Cuevas, J.M. Moreno-Rojas, M.J. Ruiz-Moreno (2017), "Assessing a traceability technique in fresh oranges (*Citrus sinensis* L. Osbeck) with an HS-SPME-GC-MS method. Towards a volatile characterisation of organic oranges", *Food Chemistry* Vol.221, pp.1930-1938.
50. Fanciullino, A. L., et al. (2006), "Carotenoid diversity in cultivated citrus is highly influenced by genetic factors", *Journal of Agricultural and Food Chemistry*, Vol. 54(12), pp. 4397- 4406.
51. Fisher R.A. (1936), "The use of multiple measurements in taxonomic problems", *Annals of Eugenics* Vol.7, pp.179–188.

52. G. Xu, D. Liu, J. Chen, X. Ye, Y. Ma, J. Shi (2008), "Juice components and antioxidant capacity of citrus varieties cultivated in China", *Food Chemistry* Vol.106, pp.545-551.
53. Gattuso G., Barreca D., Gargiulli C., Leuzzi U., Caristi C. (2007). "Flavonoid composition of Citrus juices", *Molecules* Vol. 12(8), pp. 1641-1673.
54. Gulsun Akdemir Evrendilek Birsen Hitit Özkan (2024), "Pulsed electric field processing of fruit juices with inactivation of enzymes with new inactivation kinetic model and determination of changes in quality parameters", *Innovative Food Science & Emerging Technologies* Vol.94, e103678.
55. H. Cen, Y. He, M. Huang (2006), "Measurement of Soluble Solids Contents and pH in Orange Juice Using Chemometrics and Vis-NIRS", *Journal of Agricultural and Food Chemistry* Vol.54, pp.7437-7443.
56. H. Lin, C. He, H. Liu, G. Shen, F. Xia, J. Feng (2021), "NMR-based quantitative component analysis and geographical origin identification of China's sweet orange", *Food Control* Vol.130, e108292.
57. H.R. Keller, D.L. Massart (1991), "Evolving factor analysis", *Chemometrics and Intelligent Laboratory Systems* Vol.12, pp.209-224.
58. Haifeng Lin, Chenxi He, Huili Liu, Guiping Shen, Feng Xia, Jianghua Feng (2021), "NMR-Based quantitative component analysis and geographical origin identification of China's sweet orange", *Food Control* Vol.130, e108292.
59. Harri V., Franca B., (2006), "Evaluation of Cancer Preventive Agents", *Carotenoids* Vol.2, IARC, pp.170-180.
60. Hasim Kelebek, Serkan Selli, Ahmet Canbas, Turgut Cabaroğlu (2019), "HPLC determination of organic acids, sugars, phenolic compositions and antioxidant capacity of orange juice and orange wine made from a Turkish cv. Kozan". *Microchemical Journal* Vol.91(2), 178-192

61. Hemachandra P., Pathirana A., Kamalasekara R. (2025), “Application of optical spectroscopy techniques in fruit juice authentication”, *Trends in Food Science & Technology* Vol. 146(2), pp. 102351-102370
62. Hinek A, Kim HJ, Wang Y, Wang A, Mitts TF. (2014), “Sodium L-ascorbate enhances elastic fibers deposition by fibroblasts from normal and pathologic human skin”, *J Dermatol Sci* Vol.75(3), pp.173-82.
63. I. Goodfellow, Y. Bengio, A. Courville (2016), *Deep Learning*, The MIT Press.
64. İ.Y. Genç, R. Gürfidan, E. Açikgözoğlu (2025), “Quality Determination of Frozen-Thawed Shrimp Using Machine Learning Algorithms Powered by Explainable Artificial Intelligence”, *Food Analytical Methods* Vol.18, pp.935-945.
65. J.J.T. Gama, C.M. Sylos (2005). “Major carotenoid composition of Brazilian Valencia orange juice: Identification and quantification by HPLC”, *Food Research International* Vol. 38(8-9), pp. 899 - 903.
66. J. Queffelec, W. Beraud, M. Dolores Torres, H. Domínguez, “Advances in obtaining ready to use extracts with natural solvents”, *Sustainable Chemistry and Pharmacy* Vol.38, e101478.
67. J. F. Robyt (1998), *Essentials of Carbohydrate Chemistry*. New York: Springer.
68. J. L. Jameson et al. (2018), *Harrison’s Principles of Internal Medicine*, 20th ed. New York: McGraw-Hill.
69. Jahani R., Nasiri A., Eslamizad S., Mokhtari S., Bayanati M., Haratian A., Yazdanpanah H. (2024), “Detection and verification of citrus juice adulteration: A comprehensive review of targeted and untargeted analytical methods”, *Journal of Food Composition and Analysis* Vol.127, e.107127.

70. Jolliffe I.T., Cadima J. (2016), "Principal component analysis: a review and recent developments", *Philosophical Transactions of the Royal Society A* Vol.374, e20150202.
71. Jungen M., Gilcher C., Patz C.D., Steingass C.B., Schweiggert R. (2025), "Phenolic compounds and quantitative ¹H-NMR spectroscopy for authentication of lemon, lime, orange and grapefruit juices", *Food Chemistry* Vol.485, e144379.
72. Justesen U., Knuthsen P., Leth, T. (1998), "Quantitative analysis of flavonols, flavones, and flavanones in fruits, vegetables and beverages by HPLC with photo-diode array and MS detection", *Journal of Chromatography A* Vol. 799(1-2), pp. 101 - 110.
73. K. Ogawa, A. Kawasaki, T. Yoshida, H. Nesumi, M. Nakano, Y. Ikoma, M. Yano (2000), "Evaluation of Auraptene Content in Citrus Fruits and Their Products", *Journal of Agricultural and Food Chemistry* Vol.48, pp.1763 - 1769.
74. K. Wang and Z. Xu (2022), "Comparison of freshly squeezed and processed orange juice using LC-HRMS," *Food Chemistry* Vol. 373, art. 131430.
75. Karadeniz F. (2004), "Main organic acid distribution of authentic citrus juices in Turkey", *Turkish Journal of Agriculture and Forestry* Vol. 28(4), pp. 267-271.
76. Katie Weikle. (2012). "Determination of acid citric in fruit juices using HPLC", *Concordia College Journal of Analytical Chemistry* 3, pp.57-62.
77. Kaufman L., Rousseeuw P.J. (2005), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, Hoboken.
78. Kewen Wang, Zhenzhen Xu (2022), "Comparison of freshly squeezed, non-thermally and thermally processed orange juice based on traditional quality characters, untargeted metabolomics, and volatile overview", *Food Chemistry* Vol.373 (Part B), art. 131430.

79. L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblá (2016), “Quality performance metrics in multivariate classification methods for qualitative analysis”, *TrAC Trends in Analytical Chemistry* Vol.80, pp.612-624.
80. L. Cuadros-Rodríguez, L. Valverde-Som, A.M. Jiménez-Carvelo, M. Delgado-Aguilar (2020), “Validation requirements of screening analytical methods based on scenario-specified applicability indicators”, *TrAC Trends in Analytical Chemistry* Vol.122, art.115705.
81. L. Xu, Z. Xu, S. Kelly, X. Liao (2020), “Integrating untargeted metabolomics and targeted analysis for not from concentrate and from concentrate orange juices discrimination and authentication”, *Food Chemistry* Vol.329, e127130.
82. Lee H. S. (2001), “Characterization of carotenoids in juice of red navel orange (Cara Cara)”, *Journal of Agricultural and Food Chemistry* Vol. 49(5), pp. 2563-2568.
83. Lee HS, Coates GA (2000), “Quantitative study of free sugars and myo-inositol in citrus juices by HPLC and a literature compilation”, *J Liq Chromatogr Relat Technol.* Vol. 23(14), pp. 2123 - 2141.
84. Lei Xu, Zhenzhen Xu, Simon Kelly, Xiaojun Liao (2020), “Integrating untargeted metabolomics and targeted analysis for not from concentrate and from concentrate orange juices discrimination and authentication”, *Food Chemistry* Vol.329, art. 127130.
85. Lin H., He C., Liu H., Shen G., Xia F., Feng J. (2021), “NMR-Based quantitative component analysis and geographical origin identification of sweet orange”, *Food Control* Vol.130, art.108292.,
86. Li Y., Fang X., Wang Z. (2022), “Machine learning strategies for food authentication based on chemical composition data”, *Analytica Chimica Acta* Vol.1203, pp. 339707-339717.

87. M. E. Dasenaki and N. S. Thomaidis, "Quality and Authenticity Control of Fruit Juices-A Review," *Molecules* Vol. 24(6), art. 1014.
88. M. Gebishu, B. Fikadu, B. Bekele, L. Tesfaye Jule, N. N, K. Ramaswamy (2022), "Fluorescence and UV/visible spectroscopic investigation of orange and mango fruit juice quality in case of Adama Town", *Scientific Reports* Vol.12, art. 7345.
89. M. Singh, K. Domijan (2019), "Comparison of Machine Learning Models in Food Authentication Studies", *30th Irish Signals and Systems Conference*, pp. 1-6.
90. M. V. Vavoura, I. K. Karabagias, I. S. Kosma, A. V. Badeka, and M. G. Kontominas, "Characterization and Differentiation of Fresh Orange Juice Variety Based on Conventional Physicochemical Parameters, Flavonoids, and Volatile Compounds Using Chemometrics", *Molecules* Vol. 27(19), art.6166.
91. M. Vavoura et al. (2022), "Differentiation of orange juices according to botanical origin using flavonoids", *Molecules* Vol. 27, No. 19, art. 6166.
92. M. Wang, A. Perera, R. Gutierrez-Osuna (2004), "Principal discriminants analysis for small-sample-size problems: application to chemical sensing", *Sensors* Vol.592, pp. 591-594.
93. M. Yulia, D. Suhandy (2019), "Authentication of organic Lampung robusta ground roasted coffee by UV-visible spectroscopy and PLS-DA method", *Journal of Physics: Conference Series* Vol.1341, art.022006.
94. M.A. Farag, M. Sheashea, C. Zhao, A.A. Maamoun (2022), "UV Fingerprinting Approaches for Quality Control Analyses of Food and Functional Food Coupled to Chemometrics: A Comprehensive Analysis of Novel Trends and Applications", *Foods* Vol.11, art. 2867.
95. M.E. Dasenaki, N.S. Thomaidis (2019), "Quality and Authenticity Control of Fruit Juices-A Review", *Molecules* Vol.24, art.1014.

96. M.J. Hidalgo, M. Pérez-Rodríguez, J.E. Gaiad, H.C. Goicoechea, A. Mendoza, R.G. Pellerano (2024), "Elemental tracer determination and modeling for geographical origin designation of sweet oranges", *Journal of Agriculture and Food Research* Vol.18, art.101386.
97. M.O. Salazar, P.L. Pisano, M. González Sierra, R.L.E. Furlan (2018), "NMR and multivariate data analysis to assess traceability of argentine citrus", *Microchemical Journal* Vol.141, pp.264-270.
98. M. Ladaniya (2008), *Citrus Fruit: Biology, Technology and Evaluation*. Academic Press.
99. Mac H.X., Sun R., Liu Y. (2023), "Current techniques for fruit juice adulteration detection and authenticity evaluation", *Journal of Food Composition and Analysis* Vol.115(4), pp. 104870-104884.
100. Maria V. Vavoura, Ioannis K. Karabagias, Ioanna S. Kosma, Anastasia V. Badeka, Michael G. Kontominas (2022), "Characterization and Differentiation of Fresh Orange Juice Variety Based on Conventional Physicochemical Parameters, Flavonoids, and Volatile Compounds Using Chemometrics", *Molecules* Vol.27 (19), pp. 6166.
101. Martin D., Ruck C., Schmidt M., Sell S., Beutner S., Mayer B., Walsh R.,(2005). *Chemistry of carotenoid oxidation and free radical reaction*, Institute of Organic Chemistry and Macromolecular Chemistry, UK, pp.205-211.
102. Mohammad Ehsani José A., Zubiaur M., Dolores Torres-Cardona Francisco J., García-González Rosa M. Callejón (2023), "NIR spectroscopy combined with chemometrics for detection of pulp wash addition in orange juice", *Microchemical Journal* Vol.194, art. 109304.
103. Mohan Lal Ladaniya (2008), *Citrus Fruit: Biology, Technology and Evaluation*, Academic Press.

104. Monago-Maraña O., Zapardiel A., Crevillén A.G. (2024), “Untargeted authentication of fruit juices based on electrochemical fingerprints combined with chemometrics”, *LWT* Vol.209, art.116797.
105. N. Jayaprakash, R. Suresh, S. Rajalakshmi, S. Raja, E. Sundaravadivel, M. Gayathri, M. Sridharan (2020), “One-step synthesis, characterisation, photocatalytic and bio-medical applications of ZnO nanoplates”, *Materials Technology* Vol.35, pp.112-124.
106. N. Yindeethirathana, N. Sricharoen, N. Panyain, T. Vilaivan, K. Wongravee, T. Praneenararat (2025), “Classifications of cultivars and geographical origins of durians by chemometric analyses of fatty acid compositions”, *Microchemical Journal* Vol.213, art. 113812.
107. Nørgaard L. et al. (2000), “Interval partial least-squares regression (iPLS): a comparative chemometric study”, *Applied Spectroscopy* Vol.54(3), pp.413–419.
108. Nour V., Trandafir I., Ionica M. E. (2010). “HPLC Organic Acid Analysis in Different Citrus Juices under Reversed Phase Conditions”. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* Vol38(1), pp.44–48.
109. O. Monago-Maraña, A. Zapardiel Palenzuela, A.G. Crevillén (2024), “Untargeted authentication of fruit juices based on electrochemical fingerprints combined with chemometrics. Adulteration of orange juice as case of study”, *LWT* Vol.209, art. 116797.
110. O. Rodionova, S. Kucheryavskiy, A. Pomerantsev (2021), “Efficient tools for principal component analysis of complex data— a tutorial”, *Chemometrics and Intelligent Laboratory Systems* Vol.213, art. 104304.
111. Ooghe W. C., Ooghe S. J., Detavernier, C. M., Huyghbaert A. (1994). “Differentiation of Citrus juices by multivariate analyses. Part II. Orange and mandarin juices”, *Journal of Agricultural and Food Chemistry* Vol. 42(10), pp. 2191-2195.

112. Pantić M., Kovačević D., Pantić S. (2021), “Chemometric-assisted UV–Vis spectroscopy for fruit juice authenticity assessment”, *Spectrochimica Acta Part A* Vol. 246, pp. 119017 - 119025.
113. Peng-jun Lu et al. (2017), “Cytological and molecular characterization of carotenoid accumulation in normal and high-lycopene mutant oranges”, *Scientific Reports*, art.761 (2017).
114. Q. Yang, S. Tian, H. Xu (2022), “Identification of the geographic origin of peaches by VIS-NIR spectroscopy, fluorescence spectroscopy and image processing technology”, *Journal of Food Composition and Analysis* Vol.114, art. 104843.
115. R. Díaz et al. (2014), “Metabolomic approaches for orange origin discrimination by UHPLC-QTOF-MS,” *Food Chemistry* Vol. 157, pp. 84 - 93.
116. R. Hendrickson, J.W. Kesterson, G.J. Edwards (1959), “Hesperidin in orange juice and peel extracts determined by UV absorption”, *Florida State Horticultural Society*, pp.258-263.
117. R. Mare, R. Pujia, S. Maurotti, S. Greco, A. Cardamone, A.R. Coppoletta, S. Bonacci, A. Procopio, A. Pujia (2023), “Assessment of Mediterranean Citrus Peel Flavonoids and Their Antioxidant Capacity Using an Innovative UV-Vis Spectrophotometric Approach”, *Plants* Vol.12, art. 4046.
118. R. Ríos-Reina, S.M. Azcarate (2023), “How Chemometrics Revives the UV-Vis Spectroscopy Applications as an Analytical Sensor for Spectralprint (Nontargeted) Analysis”, *Chemosensors* Vol.11, art. 8.
119. R. Sun et al. (2023), “UPLC-QTOF-MS coupled with machine learning to discriminate NFC and FC orange juice,” *Food Control* Vol.145, art. 109487.

120. R. Sun, R. Xing, J. Zhang, L. Wei, Y. Ge, T. Deng, W. Zhang, Y. Chen (2022), "Authentication and quality evaluation of not from concentrate and from concentrate orange juice by HS-SPME-GC-MS coupled with chemometrics", *LWT* Vol.162, art. 113504.
121. R. Yusun (2020), "Optimization for deep learning: An overview", *Journal of the Operations Research Society of China* Vol. 8, pp. 249-294.
122. R. J. Braddock (1999), *Handbook of Citrus By-Products and Processing Technology*, New York: Wiley.
123. Rodrigo Scherer, Ana Cecília Poloni Rybka, et.al (2012). "Validation of a HPLC method for simultaneous determination of main organic acids in fruits and juices", *Food Chemistry* Vol.135 (1), pp.150-154.
124. S. Grassi, C. Alamprese (2024), "Spectroscopic Non-targeted Techniques in Combination with Linear Discriminant Analysis for Wine Vinegar Authentication", *Food and Bioprocess Technology* Vol.17, pp.479-488.
125. S. Liang, G. Chen, C. Ma, J. Gu, C. Zhu, L. Li, H. Gao, Z. Yang, J. Cao, Z. Chen (2025), "Qualitative Identification and Adulteration Quantification of Extra Virgin Olive Oil Based on Raman Spectroscopy Combined with Multi-task Deep Learning Model", *Food Analytical Methods* Vol.18, pp. 385-397.
126. S.C. Heghes, O. Vostinaru, C. Mogosan, D. Miere, C.A. Iuga, L. Filip (2022), "Safety Profile of Nutraceuticals Rich in Coumarins: An Update", *Frontiers in Pharmacology* Vol.13, e2022.
127. S.M. Afifi, R. Gök, I. Eikenberg, D. Krygier, E. Rottmann, A.-S. Stübler, K. Aganovic, S. Hillebrand, T. Esatbeyoglu (2023), "Comparative flavonoid profile of orange (*Citrus sinensis*) flavedo and albedo extracted by conventional and emerging techniques using UPLC-IMS-MS, chemometrics and antioxidant effects", *Frontiers in Nutrition* Vol.10, pp. 1-16.

128. S. Ranganna (2000), *Handbook of Analysis and Quality Control for Fruit and Vegetable Products*, 2nd ed. New Delhi: Tata McGraw-Hill
129. Samaneh Ehsani, Hassan Yazdanpanah, Hadi Parastar (2023), “An innovative screening approach for orange juice authentication using dual portable/handheld NIR spectrometers and chemometrics”, *Microchemical Journal* Vol.194, art.109304.
130. Silva L. C., David J. M., Borges R. D. S., Ferreira S. L., David J. P., Reis P. S. D., Bruns R. E. (2014), “Determination of flavanones in orange juices obtained from different sources by HPLC/DAD”, *Journal of analytical methods in chemistry* Vol. 2014 Aug 7, art. 296838.
131. Shuang Bi, Sicheng Sun, Fei Lao, Xiaojun Liao, Jihong Wu (2020), “Gas chromatography–mass spectrometry combined with multivariate data analysis as a tool for differentiating between processed orange juice samples on the basis of their volatile markers”, *Food Chemistry* Vol.311, art. 125913.
132. Shuang Bi, Sicheng Sun, Fei Lao, Xiaojun Liao, Jihong Wu (2020), “Volatile markers and multivariate discrimination of HHP vs pasteurized orange juice during storage”, *Food Chemistry* Vol.311, pp. 125913.
133. Sun R. et al. (2023), “Quality changes of high hydrostatic pressure processed orange juice during storage based on metabolomics”, *Food Chemistry* Vol.404, art.134612.
134. Sun R., Xing R., Zhang J., Wei L., Ge Y., Deng T., Zhang W., Chen Y. (2022), “Authentication and quality evaluation of not from concentrate and from concentrate orange juice by HS-SPME-GC-MS coupled with chemometrics”, *LWT* Vol.162, art.113504.
135. Sun R. , Xing R., Zhang J., Yu N., Ge Y., Zhang W., Chen Y. (2023), “UPLC-QTOF-MS coupled with machine learning to discriminate between NFC and FC orange juice”, *Food Control* Vol.145, art.109487.

136. T.T. Ta, D.L. Nguyen, T.H.L. Nguyen, T.T. Hoang, T.H. Do, L.H. Nguyen, T.H. Tran, D.T. Nguyen, G.B. Pham (2024), "Application of machine learning models for green and simultaneous determination of asiaticoside and madecassoside in *Centella asiatica*", *Journal of Science and Technology* Vol.7, pp.1-8.
137. Tyagi G., Jangir D.K., Singh P. et al (2014), "Rapid determination of main constituents of packed juices by reverse phase-high performance liquid chromatography: an insight in to commercial fruit drinks", *J Food Sci Technol* Vol. 51, pp. 476 - 484.
138. Tappy L, Lê KA (2010), "Metabolic effects of fructose and the worldwide increase in obesity," *Physiological Reviews* Vol. 90, pp. 23-46.
139. V. Bhaumik (2025), "Disease-resistant sweet oranges", *Nature Ecology & Evolution* Vol.9, art.753.
140. V. Centonze, V. Lippolis, S. Cervellieri, A. Damascelli, G. Casiello, M. Pascale, A.F. Logrieco, F. Longobardi (2019), "Discrimination of geographical origin of oranges (*Citrus sinensis* L. Osbeck) by mass spectrometry-based electronic nose and characterization of volatile compounds", *Food Chemistry* Vol.277, pp.25-30.
141. V.F. de Moraes-Neto, A.C. Costa-Santos, J.A.L. Pallone (2025), "Portable Near-Infrared Spectrometer in Tandem with Chemometrics as an Option for the Authenticating Commercial A2 Bovine Milk", *Food Analytical Methods* Vol.18, pp.1009-1018.
142. Vachiramon V., Vargas M. A., Pashley D. H., Tay F. R., Geraldeli S., Qian F., Armstrong S. R. (2008). "Effects of oxalate on dentin bond after 3-month simulated pulpal pressure". *Journal of Dentistry* Vol.36(3), pp.178–185.
143. Varnasserri A., Xu Y., Goodacre R. (2022), "Handheld Raman spectroscopy and chemometrics for detecting adulteration in orange juice", *Analytical Methods* Vol.14, pp.1663–1670.

144. Varnasser M., Xu Y., Goodacre R. (2022), “Rapid detection and quantification of the adulteration of orange juice with grapefruit juice using handheld Raman spectroscopy and multivariate analysis”, *Analytical Methods* Vol.14, pp. 1587–1595.
145. Vavoura M.V., KaraBagias I.K., Kosma I.S., Badeka A.V., Kontominas M.G. (2022), “Characterization and differentiation of fresh orange juice variety based on physicochemical parameters, flavonoids and volatile compounds using chemometrics”, *Molecules* Vol.27 (19), art.6166.
146. W.M. Reuter (2015), *Analysis of Organic Acids in Fruit Juices by HPLC and UV Detection, Liquid Chromatography*, PerkinElmer, Inc, Shelton, CT.
147. X. Shu, M. Xie, X. Zhang, N. Wang, W. Zhang, J. Lin, J. Yang, X. Yang, Y. Li (2025), “Untargeted Metabolomics Comparison and Nutrition Evaluation of Geographical Indication Newhall Navel Oranges in China”, *Foods* Vol.14, art.355.
148. Y. Li, Z. Ren, C. Zhao, G. Liang (2025), “Geographical Origin Traceability of Navel Oranges Based on Near-Infrared Spectroscopy Combined with Deep Learning”, *Foods* Vol.14(3), art.484.
149. Z. Jandrić et al. (2014), “Assessment of fruit juice authenticity using UPLC–QToF MS: A metabolomics approach,” *Food Chemistry* Vol. 148, pp.7 - 17.
150. Zhang Z., Chen J., Li H. (2024), “Rapid food authenticity assessment using chemometrics-assisted spectroscopy: Current status and future perspectives”, *Comprehensive Reviews in Food Science and Food Safety* Vol. 23(3), pp. 1682-1705.
151. Zhou Q., Li G., Ou-Yang Z., Yi X., Huang L., Wang H. (2020), “Volatile organic compounds profiles to determine authenticity of sweet orange juice using headspace gas chromatography coupled with multivariate analysis”, *Foods* Vol.9 (4), art.505.

PHỤ LỤC

Phụ lục 1

VĂN BẢN XÁC NHẬN ỨNG DỤNG QUY TRÌNH PHÂN TÍCH

BỘ Y TẾ
VIỆN KIỂM NGHIỆM
AN TOÀN VỆ SINH
THỰC PHẨM QUỐC GIA

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 30 tháng 12 năm 2025

VĂN BẢN XÁC NHẬN

Căn cứ chức năng, nhiệm vụ của Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia và hệ thống quản lý chất lượng phòng thử nghiệm theo tiêu chuẩn ISO/IEC 17025:2017 đang áp dụng tại Viện, chúng tôi xác nhận:

NCS Nguyễn Đức Thanh, là người tham gia nghiên cứu xây dựng, cập nhật quy trình:

1. Xác định hàm lượng acid hữu cơ và dạng muối tương ứng trong thực phẩm (bao gồm nước cam) bằng phương pháp HPLC, mã hiệu phương pháp thử NIFC.02.M.32.
2. Xác định hàm lượng carotenoid trong thực phẩm (bao gồm nước cam) bằng phương pháp UV-Vis, mã hiệu phương pháp thử NIFC.02.M.57.
3. Xác định hàm lượng flavonoid trong thực phẩm (bao gồm nước cam), mã hiệu phương pháp thử NIFC.05.M.139.

Hiện nay, các quy trình nêu trên đang được triển khai thường quy tại phòng thử nghiệm của Khoa Dinh dưỡng và phụ gia thực phẩm, Khoa Nghiên cứu và phát triển. Ngoài ra, phương pháp NIFC.02.M.32 và NIFC.02.M.57 đã được công nhận phù hợp theo tiêu chuẩn ISO/IEC 17025:2017 trên nền thực phẩm theo Quyết định số 2269/QĐ-VPCNCL ngày 14/12/2023 và Quyết định số 3452/QĐ-VPCNCL ngày 31/12/2024 của Văn phòng Công nhận chất lượng quốc gia, còn hiệu lực tại thời điểm xác nhận. Số hiệu phòng thí nghiệm: VILAS 203.

Văn bản này được lập nhằm xác nhận việc ứng dụng kết quả nghiên cứu của NCS Nguyễn Đức Thanh tại Viện Kiểm nghiệm an toàn vệ sinh thực phẩm Quốc gia, phục vụ cho việc hoàn thiện hồ sơ và bảo vệ luận án tiến sĩ của NCS Nguyễn Đức Thanh.

Trân trọng.

VIỆN TRƯỞNG
VIỆN KIỂM NGHIỆM
AN TOÀN VỆ SINH
THỰC PHẨM
QUỐC GIA
Lê Thị Hồng Hào

Phụ lục 2

KẾT QUẢ THẨM ĐỊNH CÁC PHƯƠNG PHÁP ĐỊNH LƯỢNG

Phụ lục 2.1. Kết quả thẩm định phương pháp định lượng các acid trong nước cam

Kết quả thẩm định MDL, MQL của acid oxalic

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
1	5829	2,34	2,6	0,2	0,60	2,0	4,3
2	6932	2,7					
3	6316	2,5					
4	6419	2,53					
5	5988	2,39					
6	7023	2,73					
7	7074	2,75					
8	6923	2,7					
9	6820	2,66					
10	7375	2,84					

Kết quả thẩm định MDL, MQL của acid malic

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
1	6249	7,59	7,4	0,3	0,9	3,0	8,2
2	6403	7,8					
3	6237	7,57					
4	5985	7,22					

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
5	5806	6,97					
6	6279	7,63					
7	5993	7,23					
8	5674	6,78					
9	6122	7,41					
10	5742	6,88					

Kết quả thẩm định MDL, MQL của acid ascorbic

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
1	33427	16,46	16,2	1,3	3,9	13,0	4,2
2	34167	16,8					
3	32206	15,9					
4	32558	16,06					
5	30125	14,94					
6	35355	17,35					
7	35738	17,53					
8	26656	13,34					
9	35052	17,21					
10	33179	16,35					

Kết quả thẩm định MDL, MQL của acid citric

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
1	67540	63,38	61,5	4,8	14,4	48,0	4,3
2	66356	62,27					
3	64487	60,52					
4	67237	63,1					
5	62016	58,2					
6	70111	65,8					
7	67888	63,71					
8	52966	49,71					
9	66499	62,41					
10	70747	66,39					

Kết quả thẩm định MDL, MQL của acid fumaric

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
1	2183	1,28	1,3	0,05	0,15	0,5	8,7
2	3343	1,39					
3	3412	1,4					
4	2652	1,32					
5	3232	1,38					
6	2699	1,33					
7	2796	1,34					

Lần PT	Diện tích peak (mAU.s)	Hàm lượng $\mu\text{g/g}$	Trung bình	SD	LOD $\mu\text{g/g}$	LOQ $\mu\text{g/g}$	R
8	2614	1,32					
9	3620	1,42					
10	2420	1,3					

Kết quả độ lặp lại của 5 acid

STT	Hàm lượng acid oxalic ($\mu\text{g/g}$)	Hàm lượng acid malic ($\mu\text{g/g}$)	Hàm lượng acid ascorbic ($\mu\text{g/g}$)	Hàm lượng acid citric ($\mu\text{g/g}$)	Hàm lượng acid fumaric ($\mu\text{g/g}$)
1	122,9	698,8	1400,3	15905,0	31,9
2	128,6	709,0	1373,2	16065,1	32,2
3	121,7	736,7	1348,4	16048,9	31,7
4	126,1	719,6	1359,1	16091,6	32,6
5	129,1	698,4	1365,1	16413,2	31,7
6	116,6	726,0	1236,7	14602,6	31,6
7	123,0	714,9	1385,7	16605,0	32,3
8	121,5	708,4	1358,7	16165,9	31,5
9	121,7	714,7	1391,2	16522,9	31,7
10	123,4	701,8	1398,0	16114,4	31,6
TB	123,5	712,8	1361,6	16054	31,9
RSD	3,0	1,7	3,5	3,5	1,3

Kết quả độ thu hồi của acid oxalic

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
1	123,5	50	172,3	97,6	99,1	2,3
2			174,7	102,4		
3			173,0	99,0		
4			172,2	97,4		
5		100	222,6	99,1	98,2	0,9
6			220,8	97,3		
7			221,0	97,5		
8			222,4	98,9		
9		200	319,5	98,0	99,2	2,1
10			328,0	102,3		
11			320,5	98,5		
12			319,0	97,8		

Kết quả độ thu hồi của acid malic

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
1	712,8	300	1017,4	101,5	101,8	0,6
2			1017,5	101,6		
3			1016,9	101,4		
4			1020,5	102,6		
5		600	1310,4	99,6	98,8	0,8
6			1300,2	97,9		
7			1302,6	98,3		

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
8		1200	1308,8	99,3	98,0	0,8
9			1880,5	97,3		
10			1901,5	99,1		
11			1883,0	97,5		
12			1888,5	98,0		

Kết quả độ thu hồi của acid ascorbic

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
1	1361,6	550	1931,1	103,6	103,2	1,0
2			1921,5	101,8		
3			1933,6	104,0		
4			1930,7	103,5		
5		1100	2490,4	102,6	103,0	0,7
6			2492,6	102,8		
7			2506,4	104,1		
8			2489,0	102,5		
9		2200	3549,5	99,5	102,2	2,1
10			3594,0	101,5		
11			3649,5	104,0		
12			3644,5	103,8		

Kết quả độ thu hồi của acid citric

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
1	16054	6500	22632,8	101,2	102,1	1,1
2			22732,6	102,8		
3			22788,4	103,6		
4			22751,5	103,1		
5		13000	29495,6	103,4	103,6	0,3
6			29572,6	104,0		
7			29484,4	103,3		
8			29505,0	103,5		
9		26000	43030,5	103,8	101,4	2,9
10			42527,5	101,8		
11			41357,0	97,3		
12			42784,5	102,8		

Kết quả độ thu hồi của acid fumaric

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
1	0	100	98,4	98,4	99,4	0,8
2			100,1	100,1		
3			99,3	99,3		
4			99,9	99,9		
5		200	205,6	102,8	99,6	2,2
6			199,1	99,6		
7			196,6	98,3		

STT mẫu	Hàm lượng ban đầu (µg/g)	Lượng chuẩn thêm (µg/g)	Hàm lượng thu hồi (µg/g)	Hiệu suất (%)	Trung bình hiệu suất	RSD
8		400	195,5	97,8	98,5	0,9
9			393,7	98,4		
10			398,9	99,7		
11			390,6	97,7		
12			391,8	98,0		

Phụ lục 2.2. Kết quả thẩm định phương pháp định lượng các carotenoid trong nước cam

Kết quả thẩm định MDL, MQL của lutein

Lần phân tích	Diện tích peak (mAU.s)	Hàm lượng (µg/mL)	Trung bình (µg/mL)	SD	MDL (mg/L)	MQL (mg/L)	R
1	6981	0,016	0,016	0,0006	0,002	0,006	8,59
2	6717	0,016					
3	6511	0,015					
4	7439	0,017					
5	6678	0,016					
6	6586	0,015					
7	6532	0,015					
8	7197	0,016					
9	7173	0,016					
10	7297	0,017					

Kết quả thẩm định MDL, MQL của lycopene

Lần phân tích	Diện tích peak (mAU.s)	Hàm lượng (µg/mL)	Trung bình (µg/mL)	SD	LOD (mg/L)	LOQ (mg/L)	R
1	3333	0,012	0,012	0,0010	0,003	0,010	4,11
2	3028	0,011					
3	2697	0,010					
4	3179	0,012					
5	3408	0,013					
6	3116	0,011					
7	3260	0,012					
8	3620	0,013					
9	3346	0,012					
10	3510	0,013					

Kết quả thẩm định MDL, MQL của β-caroten

Lần phân tích	Diện tích peak (mAU.s)	Hàm lượng (µg/mL)	Trung bình (µg/mL)	SD	LOD (mg/L)	LOQ (mg/L)	R
1	5298	0,027	0,024	0,002	0,006	0,019	4,17
2	5081	0,026					
3	4550	0,024					
4	4506	0,023					
5	4246	0,022					
6	4827	0,025					
7	4765	0,025					
8	4262	0,022					
9	4168	0,022					
10	4762	0,025					

Kết quả độ lặp lại của lutein

Lần phân tích	Thể tích mẫu (mL)	Thể tích hòa cần (mL)	Hệ số pha loãng	Diện tích peak (mAU.s)	Nồng độ (µg/mL)	Hàm lượng trong mẫu (µg/mL)
1	20	5	1	25048	0,193	0,048
2	20	5	1	24823	0,191	0,048
3	20	5	1	24663	0,190	0,047
4	20	5	1	23369	0,181	0,045
5	20	5	1	23923	0,185	0,046
6	20	5	1	24605	0,190	0,047
Hàm lượng trung bình (µg/mL)						0,047
SD						0,001
RSD (%)						2,4

Kết quả độ lặp lại của lycopene

[illegible]

Kết quả độ lặp lại của β -caroten

Lần phân tích	Thể tích mẫu (mL)	Thể tích hòa cần (mL)	Hệ số pha loãng	Diện tích peak (mAU.s)	Nồng độ ($\mu\text{g/mL}$)	Hàm lượng trong mẫu ($\mu\text{g/mL}$)
1	20	5	1	49504	0,590	0,148
2	20	5	1	49330	0,588	0,147
3	20	5	1	49258	0,588	0,147
4	20	5	1	49145	0,586	0,147
5	20	5	1	49073	0,585	0,146
6	20	5	1	49081	0,585	0,146
Hàm lượng trung bình ($\mu\text{g/mL}$)						0,147
SD						0,00049
RSD (%)						0,3

Độ thu hồi của lutein

STT	Thể tích mẫu (mL)	Lượng thêm chuẩn ($\mu\text{g/mL}$)	Hàm lượng tìm lại ($\mu\text{g/mL}$)	Hiệu suất thu hồi R (%)	SD	RSD (%)
1	5,00	0,50	0,454	90,8	0,0004	0,35
2			0,468	93,6		
3			0,458	91,6		
4		1,00	0,804	80,4	0,0003	0,33
5			0,807	80,7		
6			0,815	81,5		
7		2,00	2,056	102,8	0,0017	1,12
8			2,116	105,8		
9			2,110	105,5		

Độ thu hồi của lycopene

STT	Thể tích mẫu (mL)	Lượng thêm chuẩn (µg/mL)	Hàm lượng tìm lại (µg/mL)	Hiệu suất thu hồi R (%)	SD	RSD (%)
1	5,00	0,25	0,202	81,0	0,0004	0,73
2			0,212	84,8		
3			0,218	87,1		
4		0,50	0,047	88,0	0,0004	0,78
5			0,048	90,5		
6			0,047	87,9		
7		1,00	0,075	99,9	0,0012	1,62
8			0,076	102,9		
9			0,077	104,8		

Độ thu hồi của β -caroten

STT	Thể tích mẫu (mL)	Lượng thêm chuẩn (µg/mL)	Hàm lượng tìm lại (µg/mL)	Hiệu suất thu hồi R (%)	SD	RSD (%)
1	5,00	1,5	1,452	96,8	0,0005	0,18
2			1,438	95,9		
3			1,431	95,4		
4		3,00	2,871	95,7	0,0010	0,36
5			2,883	96,1		
6			2,913	97,1		
7		6,00	5,022	83,7	0,0048	1,20
8			5,130	85,5		
9			5,214	86,9		

Phụ lục 2.3. Kết quả thẩm định phương pháp định lượng các flavonoid trong nước cam

Kết quả thẩm định MDL, MQL của narigin

Lần PT	Tín hiệu	Hàm lượng (mg/L)	Trung bình (mg/L)	SD	LOD (mg/L)	LOQ (mg/L)	R
1	2869	0,95	0,97	0,04	0,12	0,40	8,07
2	2964	0,97					
3	3082	1,01					
4	3121	1,02					
5	2855	0,94					
6	3005	0,99					
7	2692	0,89					
8	2832	0,93					
9	3051	1,00					
10	3001	0,99					

Kết quả thẩm định MDL, MQL của hesperidin

Lần PT	Tín hiệu	Hàm lượng (mg/L)	Trung bình (mg/L)	SD	LOD (mg/L)	LOQ (mg/L)	R
1	4264	1,16	1,17	0,06	0,17	0,57	6,87
2	4165	1,12					
3	4116	1,10					
4	4240	1,15					
5	4356	1,19					
6	4463	1,23					

Lần PT	Tín hiệu	Hàm lượng (mg/L)	Trung bình (mg/L)	SD	LOD (mg/L)	LOQ (mg/L)	R
7	4567	1,27					
8	4240	1,15					
9	4249	1.15					
10	4465	1.23					

Kết quả thẩm định MDL, MQL của neohesperidin

Lần PT	Tín hiệu	Hàm lượng (mg/L)	Trung bình (mg/L)	SD	LOD (mg/L)	LOQ (mg/L)	R
1	3398	0,82	0,85	0,04	0,13	0,44	6,48
2	3605	0,90					
3	3499	0,86					
4	3560	0,88					
5	3400	0,82					
6	3584	0,89					
7	3677	0,93					
8	3364	0,80					
9	3484	0,85					
10	3356	0,80					

Kết quả thẩm định MDL, MQL của neohesperidin

Lần PT	Tín hiệu	Hàm lượng (mg/L)	Trung bình (mg/L)	SD	LOD (mg/L)	LOQ (mg/L)	R
1	3212	0,74	0,77	0,06	0,18	0,61	4,18
2	3347	0,80					
3	3034	0,67					
4	3312	0,78					
5	3097	0,70					
6	3127	0,71					
7	3293	0,77					
8	3394	0,81					
9	3494	0,85					
10	3463	0,84					

Độ lặp lại của hesperidin

Lần phân tích	Thể tích mẫu (mL)	Thể tích dung môi hòa tan (mL)	Hệ số pha loãng	Diện tích peak (mAU.s)	Nồng độ (ppm)	Hàm lượng (µg/mL)
1	5,00	25	1	929515	72,85	364,25
2	5,00	25	1	966615	75,76	378,81
3	5,00	25	1	976454	76,53	382,67
4	5,00	25	1	976358	76,53	382,63
5	5,00	25	1	942925	73,90	369,51
6	5,00	25	1	968784	75,93	379,66
Hàm lượng trung bình (µg/mL)						376,26
SD						7,61
RSD (%)						2,02

Độ lặp lại của naringin

<i>Lần phân tích</i>	<i>Thể tích mẫu (mL)</i>	<i>Thể tích dung môi hòa tan (mL)</i>	<i>Hệ số pha loãng</i>	<i>Diện tích Peak (mAU.s)</i>	<i>Nồng độ (ppm)</i>	<i>Hàm lượng (µg/mL)</i>
1	5,00	25	1	10317	0,64	3,19
2	5,00	25	1	10538	0,65	3,26
3	5,00	25	1	10112	0,63	3,13
4	5,00	25	1	10576	0,65	3,27
5	5,00	25	1	10474	0,65	3,24
6	5,00	25	1	10065	0,62	3,12
Hàm lượng trung bình (µg/mL)						3,20
SD						0,07
RSD (%)						2,07

Độ lặp lại của neohesperidin

<i>Lần phân tích</i>	<i>Thể tích mẫu (mL)</i>	<i>Thể tích dung môi hòa tan (mL)</i>	<i>Hệ số pha loãng</i>	<i>Diện tích Peak (mAU.s)</i>	<i>Nồng độ (ppm)</i>	<i>Hàm lượng (µg/mL)</i>
1	5,00	25	1	12952	0,64	3,20
2	5,00	25	1	13286	0,65	3,27
3	5,00	25	1	12469	0,62	3,08
4	5,00	25	1	13244	0,65	3,26
5	5,00	25	1	13420	0,66	3,31
6	5,00	25	1	13130	0,65	3,24
Hàm lượng trung bình (µg/mL)						3,32
SD						0,08
RSD (%)						2,48

Độ lặp lại của quercetin

<i>Lần phân tích</i>	<i>Thể tích mẫu (mL)</i>	<i>Thể tích dung môi hòa tan (mL)</i>	<i>Hệ số pha loãng</i>	<i>Diện tích Peak (mAU.s)</i>	<i>Nồng độ (ppm)</i>	<i>Hàm lượng (µg/mL)</i>
1	5,00	25	1	12952	0,63	3,14
2	5,00	25	1	13286	0,61	3,07
3	5,00	25	1	12469	0,61	3,05
4	5,00	25	1	13244	0,60	3,02
5	5,00	25	1	13420	0,62	3,09
6	5,00	25	1	13130	0,60	3,02
Hàm lượng trung bình (µg/mL)						3,07
SD						0,05
RSD (%)						1,35

Độ thu hồi của naringin

<i>Nồng độ chuẩn thêm vào (µg/mL)</i>	<i>Lượng thể tích mẫu (mL)</i>	<i>Nồng độ chuẩn tìm lại được (µg/mL)</i>	<i>Độ thu hồi (%)</i>	<i>RSD (%)</i>
0,60	5,00	0,63	105,0	2,71
		0,66	110,5	
		0,64	106,3	
6,00		5,95	99,1	0,37
		5,98	99,6	
		5,94	98,9	
60,00		60,90	101,5	0,14
		60,74	101,2	
		60,75	101,3	

Độ thu hồi của hesperidin

<i>Nồng độ chuẩn thêm vào (μg/mL)</i>	<i>Lượng thể tích mẫu (mL)</i>	<i>Nồng độ chuẩn tìm lại được (μg/mL)</i>	<i>Độ thu hồi (%)</i>	<i>RSD (%)</i>
35,00	5,00	109,89	99,0	0,34
		109,20	97,0	
		109,80	98,7	
140,00		220,95	104,1	0,09
		221,09	104,2	
		221,33	104,3	
280,00		335,55	93,0	0,20
		334,26	92,5	
		334,62	92,6	

Độ thu hồi của neohesperidin

<i>Nồng độ chuẩn thêm vào (μg/mL)</i>	<i>Lượng thể tích mẫu (mL)</i>	<i>Nồng độ chuẩn tìm lại được (μg/mL)</i>	<i>Độ thu hồi (%)</i>	<i>RSD (%)</i>
0,60	5,00	0,61	102,0	2,42
		0,62	103,9	
		0,64	107,0	
6,00		5,79	96,6	0,60
		5,81	96,8	
		5,74	95,7	
60,00		55,49	92,5	0,17
		55,43	92,4	
		55,30	92,2	

Độ thu hồi của quercetin

<i>Nồng độ chuẩn thêm vào (μg/mL)</i>	<i>Lượng thể tích mẫu (mL)</i>	<i>Nồng độ chuẩn tìm lại được (μg/mL)</i>	<i>Độ thu hồi (%)</i>	<i>RSD (%)</i>
0,60	5,00	0,60	100,5	0,07
		0,60	100,0	
		0,61	101,4	
6,00		5,40	90,0	3,01
		5,14	85,6	
		5,12	85,3	
60,00		52,66	87,8	1,58
		52,77	87,9	
		51,29	85,5	

Phụ lục 3

THÔNG TIN CÁC MẪU CAM

Phụ lục 3.1. Các mẫu cam xác định thành phần hóa học

STT	Kí hiệu mẫu	Địa điểm	Giống	Thời gian lấy mẫu
1	BG.V.01	Bắc giang	Vinh	2022
2	BG.V.02	Bắc giang	Vinh	2022
3	BG.V.03	Bắc giang	Vinh	2023
4	BG.V.04	Bắc giang	Vinh	2023
5	BG.V.05	Bắc giang	Vinh	2023
6	BG.V.06	Bắc giang	Vinh	2023
7	HB.CP.01	Hòa Bình	Cao Phong	2022
8	HB.CP.02	Hòa Bình	Cao Phong	2022
9	HB.CP.03	Hòa Bình	Cao Phong	2023
10	HB.CP.04	Hòa Bình	Cao Phong	2023
11	HB.CP.05	Hòa Bình	Cao Phong	2023
12	HB.CP.06	Hòa Bình	Cao Phong	2023
13	HG.XD.01	Hà giang	Xã đoàn	2022
14	HG.XD.02	Hà giang	Xã đoàn	2022
15	HG.XD.03	Hà giang	Xã đoàn	2023
16	HG.XD.04	Hà giang	Xã đoàn	2023
17	HG.XD.05	Hà giang	Xã đoàn	2023
18	HY.V.01	Hưng yên	Vinh	2022
19	HY.V.02	Hưng yên	Vinh	2022
20	HY.V.03	Hưng yên	Vinh	2022
21	HY.V.04	Hưng yên	Vinh	2023
22	HY.V.05	Hưng yên	Vinh	2023

STT	Kí hiệu mẫu	Địa điểm	Giống	Thời gian lấy mẫu
23	HY.V.06	Hung yên	Vinh	2023
24	HY.V.07	Hung yên	Vinh	2023
25	HY.V.08	Hung yên	Vinh	2024
26	HY.V.09	Hung yên	Vinh	2024
27	HY.V.10	Hung yên	Vinh	2024
28	HY.V.11	Hung yên	Vinh	2024
29	HY.V.12	Hung yên	Vinh	2024
30	HY.V.13	Hung yên	Vinh	2024
31	HY.V.14	Hung yên	Vinh	2024
32	HY.V.15	Hung yên	Vinh	2024
33	HY.V.16	Hung yên	Vinh	2024
34	NA.XD.01	Nghệ an	Xã đoài	2023
35	NA.XD.02	Nghệ an	Xã đoài	2023
36	NA.XD.03	Nghệ an	Xã đoài	2023
37	NA.XD.04	Nghệ an	Xã đoài	2024
38	NA.XD.05	Nghệ an	Xã đoài	2024
39	NA.XD.06	Nghệ an	Xã đoài	2024
40	NA.XD.07	Nghệ an	Xã đoài	2024
41	NA.XD.08	Nghệ an	Xã đoài	2024
42	NA.XD.09	Nghệ an	Xã đoài	2024
43	NA.XD.10	Nghệ an	Xã đoài	2024
44	QN.V.01	Quảng ninh	Vinh	2022
45	QN.V.02	Quảng ninh	Vinh	2022
46	QN.V.03	Quảng ninh	Vinh	2023

STT	Kí hiệu mẫu	Địa điểm	Giống	Thời gian lấy mẫu
47	QN.V.04	Quảng ninh	Vinh	2023
48	TN.S.01	Tây ninh	Sành	2022
49	TN.S.02	Tây ninh	Sành	2022
50	TN.S.03	Tây ninh	Sành	2022
51	TN.S.04	Tây ninh	Sành	2023
52	TN.S.05	Tây ninh	Sành	2023
53	TN.S.06	Tây ninh	Sành	2024
54	TN.S.07	Tây ninh	Sành	2024
55	TN.S.08	Tây ninh	Sành	2024
56	TN.S.09	Tây ninh	Sành	2024
57	TQ.S.01	Tuyên quang	Sành	2023
58	TQ.S.02	Tuyên quang	Sành	2023
59	TQ.V.01	Tuyên quang	Vinh	2022
60	TQ.V.02	Tuyên quang	Vinh	2022
61	TQ.V.03	Tuyên quang	Vinh	2022
62	TQ.V.04	Tuyên quang	Vinh	2023
63	TQ.V.05	Tuyên quang	Vinh	2023
64	TQ.V.06	Tuyên quang	Vinh	2023
65	TQ.V.07	Tuyên quang	Vinh	2023
66	TQ.V.08	Tuyên quang	Vinh	2024
67	TQ.V.09	Tuyên quang	Vinh	2024
68	TQ.V.10	Tuyên quang	Vinh	2024
69	TV.S.01	Trà vinh	Sành	2023
70	TV.S.02	Trà vinh	Sành	2023

STT	Kí hiệu mẫu	Địa điểm	Giống	Thời gian lấy mẫu
71	TV.S.03	Trà vinh	Sành	2024
72	TV.S.04	Trà vinh	Sành	2024
73	VL.S.01	Vĩnh long	Sành	2022
74	VL.S.02	Vĩnh long	Sành	2022
75	VL.S.03	Vĩnh long	Sành	2023
76	VL.S.04	Vĩnh long	Sành	2023
77	VL.S.05	Vĩnh long	Sành	2023
78	VL.S.06	Vĩnh long	Sành	2023
79	VL.S.07	Vĩnh long	Sành	2024
80	VL.S.08	Vĩnh long	Sành	2024
81	VL.S.09	Vĩnh long	Sành	2024
82	VL.S.10	Vĩnh long	Sành	2024

Phụ lục 3.2. Các mẫu cam quét phổ UV - Vis

STT	Kí hiệu mẫu	Địa điểm	Giống	Thời gian lấy mẫu
1	QN.V.23.01	Quảng ninh	Vinh	2023
2	QN.V.23.02	Quảng ninh	Vinh	2023
3	QN.V.23.03	Quảng ninh	Vinh	2023
4	QN.V.23.04	Quảng ninh	Vinh	2023
5	QN.V.23.05	Quảng ninh	Vinh	2023
6	QN.V.23.06	Quảng ninh	Vinh	2023
7	QN.V.23.07	Quảng ninh	Vinh	2023
8	QN.V.23.08	Quảng ninh	Vinh	2023
9	QN.V.23.09	Quảng ninh	Vinh	2023

10	QN.V.23.10	Quảng ninh	Vinh	2023
11	QN.V.23.11	Quảng ninh	Vinh	2023
12	QN.V.23.12	Quảng ninh	Vinh	2023
13	QN.V.23.13	Quảng ninh	Vinh	2023
14	QN.V.23.14	Quảng ninh	Vinh	2023
15	QN.V.24.01	Quảng ninh	Vinh	2024
16	QN.V.24.02	Quảng ninh	Vinh	2024
17	QN.V.24.03	Quảng ninh	Vinh	2024
18	QN.V.24.04	Quảng ninh	Vinh	2024
19	QN.V.24.05	Quảng ninh	Vinh	2024
20	QN.V.24.06	Quảng ninh	Vinh	2024
21	QN.V.24.07	Quảng ninh	Vinh	2024
22	QN.V.24.08	Quảng ninh	Vinh	2024
23	QN.V.24.09	Quảng ninh	Vinh	2024
24	QN.V.24.10	Quảng ninh	Vinh	2024
25	HB.V.23.01	Hòa Bình	Vinh	2023
26	HB.V.23.02	Hòa Bình	Vinh	2023
27	HB.V.23.03	Hòa Bình	Vinh	2023
28	HB.V.23.04	Hòa Bình	Vinh	2023
29	HB.V.23.05	Hòa Bình	Vinh	2023
30	HB.V.23.06	Hòa Bình	Vinh	2023
31	HB.V.23.07	Hòa Bình	Vinh	2023
32	HB.V.23.08	Hòa Bình	Vinh	2023
33	HB.V.23.09	Hòa Bình	Vinh	2023
34	HB.V.23.10	Hòa Bình	Vinh	2023
35	HB.V.23.11	Hòa Bình	Vinh	2023

36	HB.V.23.12	Hòa Bình	Vinh	2023
37	HB.V.23.13	Hòa Bình	Vinh	2023
38	HB.V.23.14	Hòa Bình	Vinh	2023
39	HB.V.23.15	Hòa Bình	Vinh	2023
40	HB.V.23.16	Hòa Bình	Vinh	2023
41	HB.V.23.17	Hòa Bình	Vinh	2023
42	HB.V.23.18	Hòa Bình	Vinh	2023
43	HB.V.23.19	Hòa Bình	Vinh	2023
44	HB.V.23.20	Hòa Bình	Vinh	2023
45	HB.V.23.21	Hòa Bình	Vinh	2023
46	HB.V.23.22	Hòa Bình	Vinh	2023
47	HB.V.23.23	Hòa Bình	Vinh	2023
48	HB.V.23.24	Hòa Bình	Vinh	2023
49	HB.V.23.25	Hòa Bình	Vinh	2023
50	HB.V.24. 01	Hòa Bình	Vinh	2024
51	HB.V.24.02	Hòa Bình	Vinh	2024
52	HB.V.24.03	Hòa Bình	Vinh	2024
53	HB.V.24.04	Hòa Bình	Vinh	2024
54	HB.V.24.05	Hòa Bình	Vinh	2024
55	HB.V.24.06	Hòa Bình	Vinh	2024
56	HB.V.24.07	Hòa Bình	Vinh	2024
57	HB.V.24.08	Hòa Bình	Vinh	2024
58	HB.V.24.09	Hòa Bình	Vinh	2024
59	HB.V.24.10	Hòa Bình	Vinh	2024
60	HB.V.24.11	Hòa Bình	Vinh	2024
61	HB.V.24.12	Hòa Bình	Vinh	2024

62	HB.V.24.13	Hòa Bình	Vinh	2024
63	HB.V.24.14	Hòa Bình	Vinh	2024
64	HB.V.24.15	Hòa Bình	Vinh	2024
65	HB.V.24.16	Hòa Bình	Vinh	2024
66	HB.V.24.17	Hòa Bình	Vinh	2024
67	HB.V.24.18	Hòa Bình	Vinh	2024
68	HB.V.24.19	Hòa Bình	Vinh	2024
69	HB.V.24.20	Hòa Bình	Vinh	2024
70	HB.V.24.21	Hòa Bình	Vinh	2024
71	HB.V.24.22	Hòa Bình	Vinh	2024
72	HB.V.24.23	Hòa Bình	Vinh	2024
73	HB.V.24.24	Hòa Bình	Vinh	2024
74	HB.V.24.25	Hòa Bình	Vinh	2024
75	HB.V.24.26	Hòa Bình	Vinh	2024
76	HB.V.24.27	Hòa Bình	Vinh	2024
77	HB.V.24.28	Hòa Bình	Vinh	2024
78	HB.V.24.29	Hòa Bình	Vinh	2024
79	TQ.V.23.01	Tuyên Quang	Vinh	2023
80	TQ.V.23.02	Tuyên Quang	Vinh	2023
81	TQ.V.23.03	Tuyên Quang	Vinh	2023
82	TQ.V.23.04	Tuyên Quang	Vinh	2023
83	TQ.V.23.05	Tuyên Quang	Vinh	2023
84	TQ.V.23.06	Tuyên Quang	Vinh	2023
85	TQ.V.23.07	Tuyên Quang	Vinh	2023
86	TQ.V.23.08	Tuyên Quang	Vinh	2023
87	TQ.V.23.09	Tuyên Quang	Vinh	2023

88	TQ.V.23.10	Tuyên Quang	Vinh	2023
89	TQ.V.23.11	Tuyên Quang	Vinh	2023
90	TQ.V.23.12	Tuyên Quang	Vinh	2023
91	TQ.V.23.13	Tuyên Quang	Vinh	2023
92	TQ.V.23.14	Tuyên Quang	Vinh	2023
93	TQ.V.23.15	Tuyên Quang	Vinh	2023
94	TQ.V.24.01	Tuyên Quang	Vinh	2024
95	TQ.V.24.02	Tuyên Quang	Vinh	2024
96	TQ.V.24.03	Tuyên Quang	Vinh	2024
97	TQ.V.24.04	Tuyên Quang	Vinh	2024
98	TQ.V.24.05	Tuyên Quang	Vinh	2024
99	TQ.V.24.06	Tuyên Quang	Vinh	2024
100	TQ.V.24.07	Tuyên Quang	Vinh	2024
101	TQ.V.24.08	Tuyên Quang	Vinh	2024
102	TQ.V.24.09	Tuyên Quang	Vinh	2024
103	TQ.V.24.10	Tuyên Quang	Vinh	2024
104	TQ.V.24.11	Tuyên Quang	Vinh	2024
105	TQ.V.24.12	Tuyên Quang	Vinh	2024
106	TQ.V.24.13	Tuyên Quang	Vinh	2024
107	TQ.V.24.14	Tuyên Quang	Vinh	2024
108	TQ.V.24.15	Tuyên Quang	Vinh	2024
109	TQ.V.24.16	Tuyên Quang	Vinh	2024
110	TQ.V.24.17	Tuyên Quang	Vinh	2024
111	TQ.V.24.18	Tuyên Quang	Vinh	2024

Phụ lục 3.3. Mã hóa cam đo điện hóa

Cam tự nhiên: Tên tỉnh.năm.số thứ tự

Cam Cao phong: CP.24.01 – CP.24.02.....- CP.24.14

Cam Hà Giang: HG.24.01 – HG.24.02.....- HG.24.07

Cam Bến tre: BT.24.01 – BT.24.02.....- BT.24.07

Cam Vĩnh long: VL.24.01 – VL.24.02.....- VL.24.13

Cam Nghệ an: NA.24.01 – NA.24.02.....- NA.24.10

Cam Bắc Giang: BG.24.01 – BG.24.02.....- BG.24.10

Cam Hưng Yên: HY.24.01 – HY.24.02.....- HY.24.10

Cam Quảng Ninh: QN.24.01 – QN.24.02.....- QN.24.10

Mẫu nước cam đóng chai: TM.Tên thương hiệu.năm.số thứ tự

Mirinda: TM.M.24.01 - – TM.M.24.10

Fanta: TM.F.24.01 - – TM.M.24.10

Twister: TM.T.24.01 - – TM.T.24.10

Teppy: TM.Te.24.01 - ... – TM.Te.24.10

Nutri: TM.N.24.01 -.....-TM.N.24.05

TH: TM.TH.24.01 - ... – TM.TH.24.05

Các mẫu pha trộn: PHA.Tên thương hiệu viết tắt. Số thứ tự

Số thứ tự đánh từ 01-09 tương ứng từ 10% TM đến 90% TM

Ví dụ: PHA.F.01: mẫu nước cam tự nhiên pha 10% nước cam Fanta.

Phụ lục 4

KẾT QUẢ HÀM LƯỢNG CÁC CHẤT HÓA HỌC CHÍNH TRONG NƯỚC CAM

Địa phương	Giống	Đường (g/100 mL)			Carotenoid (mg/L)		Acid (g/L)			Flavonoid (mg/L)	
		Fructose	Glucose	Sucrose	Lutein	β- caroten	Acid citric	Acid ascorbic	Acid fumaric	Hesperidin	Quercetin
Bắc giang	Vinh	2.37	2.24	2.44	0.12	0.062	175.136	2.09	0.135	201.992	34.169
Bắc giang	Vinh	2.17	2.06	2.94	0.121	0.091	187.115	0.998	0.175	212.426	36.941
Bắc giang	Vinh	2.22	2.05	2.83	0.124	0.066	192.99	1.018	0.149	198.592	40.711
Bắc giang	Vinh	2.83	2.09	2.65	0.047	0.04	191.557	0.942	0.132	242.291	41.678
Bắc giang	Vinh	2.33	2.23	2.82	0.116	0.075	192.4	1.691	0.16	303.079	45.415
Bắc giang	Vinh	2.28	2.22	2.64	0.113	0.07	182.812	2.002	0.131	532.606	39.508
Hòa Bình	Cao Phong	3.85	2.18	4.51	0.126	0.066	183.375	1.734	0.115	391.381	46.953
Hòa Bình	Cao Phong	1.21	1.16	1.88	0.126	0.065	185.272	1.748	0.082	575.88	44.849
Hòa Bình	Cao Phong	3.09	3.02	4.47	0.125	0.069	187.22	1.775	0.127	688.231	43.206
Hòa Bình	Cao Phong	2.6	2.24	5.05	0.092	0.063	183.086	1.726	0.139	271.17	42.79
Hòa Bình	Cao Phong	3.87	3.88	6.39	0.154	0.073	181.788	1.723	0.132	625.471	40.942
Hòa Bình	Cao Phong	3.61	3.4	3.66	0.126	0.069	185.11	1.715	0.11	537.404	45.499

Địa phương	Giống	Đường (g/100 mL)			Carotenoid (mg/L)		Acid (g/L)			Flavonoid (mg/L)	
		Fructose	Glucose	Sucrose	Lutein	β- caroten	Acid citric	Acid ascorbic	Acid fumaric	Hesperidin	Quercetin
Hà giang	Xã đoài	1.71	1.49	2.85	0.043	0.052	197.808	3.045	0.067	556.772	21.042
Hà giang	Xã đoài	3.19	3.12	2.24	0.029	0.037	195.77	2.714	0.066	622.528	20.356
Hà giang	Xã đoài	2.23	2.09	2.13	0.029	0.035	186.458	2.843	0.07	727.207	21.062
Hà giang	Xã đoài	2.36	2.27	3.33	0.028	0.037	187.213	2.606	0.065	447.611	19.372
Hà giang	Xã đoài	2.14	2.06	2.08	0.045	0.053	199.703	2.928	0.061	700.849	21.669
Hưng yên	Vinh	2.46	2.29	4.81	0.053	0.05	169.372	1.845	0.13	726.778	23.761
Hưng yên	Vinh	2.67	2.39	3.84	0.055	0.054	256.035	2.635	0.02	335.668	17.88
Hưng yên	Vinh	3.00	2.79	5.93	0.041	0.051	195.011	2.325	0.025	401.683	16.022
Hưng yên	Vinh	3.23	2.88	6.01	0.04	0.049	253.428	2.208	0.039	366.087	29.37
Hưng yên	Vinh	2.46	2.2	3.27	0.055	0.055	254.177	1.838	0.072	397.715	30.165
Hưng yên	Vinh	2.96	2.69	4.95	0.048	0.048	221.595	2.485	0.028	347.585	21.098
Hưng yên	Vinh	2.00	1.73	3.47	0.058	0.056	225.986	1.284	0.079	229.003	19.87
Hưng yên	Vinh	1.77	1.54	4.32	0.053	0.039	228.841	1.36	0.063	568.36	23.565
Hưng yên	Vinh	2.61	2.37	4.00	0.036	0.038	225.271	1.318	0.078	513.552	23.216

Địa phương	Giống	Đường (g/100 mL)			Carotenoid (mg/L)		Acid (g/L)			Flavonoid (mg/L)	
		Fructose	Glucose	Sucrose	Lutein	β- caroten	Acid citric	Acid ascorbic	Acid fumaric	Hesperidin	Quercetin
Hung yên	Vinh	1.85	1.8	4.09	0.074	0.053	225.729	1.334	0.07	498.255	24.572
Hung yên	Vinh	2.61	2.3	3.43	0.037	0.036	229.843	1.354	0.079	481.296	22.664
Hung yên	Vinh	2.78	2.16	3.48	0.041	0.047	148.912	2.221	0.038	139.548	17.205
Hung yên	Vinh	2.02	1.7	3.65	0.058	0.059	172.992	2.35	0.056	193.134	19.442
Hung yên	Vinh	2.5	2.32	3.97	0.073	0.05	162.46	2.142	0.054	235.113	16.471
Hung yên	Vinh	2.4	2.03	3.77	0.049	0.048	158.533	2.246	0.047	440.753	16.589
Hung yên	Vinh	2.83	2.43	3.96	0.056	0.058	153.739	2.268	0.056	290.513	14.757
Nghệ an	Xã đoài	2.41	2.23	4.3	0.073	0.051	146.316	1.464	0.058	366.937	21.3
Nghệ an	Xã đoài	2.17	2.04	3.92	0.075	0.05	349.555	11.614	0.105	221.439	41.262
Nghệ an	Xã đoài	2.1	1.98	3.58	0.089	0.062	155.913	1.619	0.054	243.038	22.302
Nghệ an	Xã đoài	1.8	1.6	3.23	0.09	0.061	180.528	1.439	0.063	209.067	16.58
Nghệ an	Xã đoài	1.94	1.59	3.47	0.061	0.044	155.795	1.453	0.104	341.51	15.161
Nghệ an	Xã đoài	2.23	2.11	4.31	0.062	0.053	164.607	1.637	0.06	459.486	29.787
Nghệ an	Xã đoài	2.03	1.93	3.48	0.063	0.051	164.517	1.224	0.099	240.246	21.166

Địa phương	Giống	Đường (g/100 mL)			Carotenoid (mg/L)		Acid (g/L)			Flavonoid (mg/L)	
		Fructose	Glucose	Sucrose	Lutein	β- caroten	Acid citric	Acid ascorbic	Acid fumaric	Hesperidin	Quercetin
Nghệ an	Xã đoài	2.23	1.98	4.27	0.055	0.049	150.523	1.421	0.044	236.939	17.312
Nghệ an	Xã đoài	2.02	1.63	3.68	0.073	0.051	265.828	1.618	0.065	302.23	30.959
Nghệ an	Xã đoài	1.88	1.69	3.2	0.059	0.052	151.939	12.108	0.056	267.471	19.396
Quảng ninh	Vinh	4.05	3.85	5.2	0.067	0.05	249.3	3.728	0.165	529.582	92.059
Quảng ninh	Vinh	2.71	3.61	2.9	0.07	0.055	264.718	3.887	0.165	557.125	92.356
Quảng ninh	Vinh	2.65	3.22	3.91	0.09	0.063	271.555	3.662	0.157	546.513	91.863
Quảng ninh	Vinh	3.33	3.15	3.54	0.07	0.056	271.957	3.865	0.164	69.023	91.88
Tây ninh	Sành	1.53	1.38	3.96	0.212	0.162	182.109	4.985	0.432	80.374	31.718
Tây ninh	Sành	1.63	1.6	4.91	0.204	0.152	168.759	0.817	0.241	108.536	26.732
Tây ninh	Sành	1.34	1.46	4.07	0.196	0.155	184.677	2.917	0.273	332.563	109.229
Tây ninh	Sành	1.79	1.94	2.58	0.198	0.168	159.519	2.595	0.282	243.898	47.038
Tây ninh	Sành	1.11	1.05	3.4	0.183	0.156	177.188	1.208	0.236	208.519	76.815
Tây ninh	Sành	1.64	1.62	3.99	0.198	0.154	168.325	0.454	0.208	109.363	62.611
Tây ninh	Sành	1.35	1.26	3.44	0.185	0.141	170.215	0.454	0.182	398.991	55.276

Địa phương	Giống	Đường (g/100 mL)			Carotenoid (mg/L)		Acid (g/L)			Flavonoid (mg/L)	
		Fructose	Glucose	Sucrose	Lutein	β- caroten	Acid citric	Acid ascorbic	Acid fumaric	Hesperidin	Quercetin
Tây ninh	Sành	1.13	1.12	3.47	0.192	0.162	182.304	2.767	0.219	449.599	118.497
Tây ninh	Sành	1.3	1.41	3.14	0.209	0.166	192.933	0.34	0.128	403.655	98.704
Tuyên quang	Sành	1.58	1.46	3.49	0.25	0.485	246.349	0.614	0.03	251.98	82.269
Tuyên quang	Sành	1.96	1.23	2.68	0.234	0.589	230.291	0.609	0.022	468.271	89.101
Tuyên quang	Vinh	3.18	2.89	4.16	0.119	0.054	245.428	13.063	0.646	468.798	35.967
Tuyên quang	Vinh	2.39	2.06	2.73	0.118	0.056	365.543	21.774	0.272	311.673	66.379
Tuyên quang	Vinh	2.16	1.84	3.78	0.134	0.065	102.145	2.715	1.864	323.524	103.049
Tuyên quang	Vinh	2.21	2.11	3.09	0.118	0.059	372.364	22.978	0.331	813.699	64.804
Tuyên quang	Vinh	2.58	2.23	3.7	0.164	0.087	385.302	20.138	0.222	305.211	64.815
Tuyên quang	Vinh	1.78	1.53	2.32	0.156	0.048	102.571	2.708	1.864	296.226	102.829
Tuyên quang	Vinh	2.07	1.9	3.15	0.155	0.051	102.801	2.718	1.861	289.627	102.274
Tuyên quang	Vinh	1.64	1.43	2.76	0.138	0.056	513.812	18.84	0.31	288.536	54.935
Tuyên quang	Vinh	1.83	1.66	2.43	0.156	0.054	482.672	19.654	0.045	284.258	68.303
Tuyên quang	Vinh	2.14	2.01	2.3	0.099	0.064	498.179	17.648	0.13	250.76	48.718

Địa phương	Giống	Đường (g/100 mL)			Carotenoid (mg/L)		Acid (g/L)			Flavonoid (mg/L)	
		Fructose	Glucose	Sucrose	Lutein	β- caroten	Acid citric	Acid ascorbic	Acid fumaric	Hesperidin	Quercetin
Trà vinh	Sành	0.9	2.57	3.59	0.133	0.212	153.038	2.043	0.068	262.381	37.656
Trà vinh	Sành	2.33	0.8	2.97	0.141	0.207	126.513	1.175	0.197	204.658	39.968
Trà vinh	Sành	2.69	2.46	4.09	0.133	0.191	125.051	1.169	0.203	279.76	39.638
Trà vinh	Sành	0.9	0.82	2.98	0.164	0.196	193.512	1.434	0.172	656.254	44.408
Vĩnh long	Sành	1.94	1.89	4.61	0.055	0.249	213.462	0.646	0.087	218.499	15.034
Vĩnh long	Sành	2.17	1.96	5.66	0.111	0.476	153.536	0.162	0.484	227.449	46.154
Vĩnh long	Sành	2.42	2.25	3.55	0.066	0.473	248.462	0.376	0.03	272.247	11.721
Vĩnh long	Sành	1.15	1.09	3.89	0.057	0.331	247.296	0.625	0.075	132.057	12.146
Vĩnh long	Sành	2.29	2.03	5.01	0.049	0.258	261.698	0.555	0.067	168.96	13.451
Vĩnh long	Sành	2.46	2.49	4.12	0.09	0.505	160.867	0.42	0.481	132.325	52.791
Vĩnh long	Sành	1.46	1.29	4.24	0.105	0.289	241.662	0.33	0.112	304.812	15.036
Vĩnh long	Sành	2.55	2.37	3.89	0.072	0.342	159.6	0.304	0.313	146.804	41.99
Vĩnh long	Sành	1.78	1.68	2.81	0.083	0.378	140.258	0.488	0.331	154.035	43.988
Vĩnh long	Sành	2.4	2.34	4.58	0.056	0.317	224.437	0.496	0.118	216.246	11.491

Phụ lục 5

MỘT SỐ CÂU LỆNH HỌC MÁY

SMOOTE

```
sm = SMOTE(k_neighbors=1)

X, y = sm.fit_resample(X.astype(str), y.astype(str))


le = LabelEncoder()
y = le.fit_transform(y)


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y, random_state=42)
X_train.shape, X_test.shape, y_train.shape, y_test.shape

((128, 10), (32, 10), (128,), (32,))


scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

PCA-SVC

PCA SVC

```
pca = PCA()
svc = SVC()
pipeline = Pipeline(steps=[('pca', pca), ('svc', svc)])
param_grid = {
    'svc__kernel': ['rbf', 'poly', 'sigmoid', 'linear'],
}
grid = GridSearchCV(pipeline, param_grid, refit = True, verbose = 3)
grid.fit(X_train, y_train)
print("\n\n")
ic(grid.best_params_)
model = grid.best_estimator_
model.fit(X_train, y_train)
print("Accuracy score:", accuracy_score(y_test, model.predict(X_test)))
```

PCA-DT

PCA-DT

Generate Code Markdown

```
pca = PCA()
dt = DecisionTreeClassifier()
pipeline = Pipeline(steps=[('pca', pca), ('dt', dt)])
param_grid = {

}
grid = GridSearchCV(pipeline, param_grid, refit = True, verbose = 3)
grid.fit(X_train, y_train)
print("\n\n")
ic(grid.best_params_)
model = grid.best_estimator_
model.fit(X_train, y_train)
print("Accuracy score:", accuracy_score(y_test, model.predict(X_test)))
```

PCA-RF

PCA - RF

Generate Code Markdown

```
pca = PCA()
rf = RandomForestClassifier()
pipeline = Pipeline(steps=[('pca', pca), ('rf', rf)])
param_grid = {

}
grid = GridSearchCV(pipeline, param_grid, refit = True, verbose = 3)
grid.fit(X_train, y_train)
print("\n\n")
ic(grid.best_params_)
model = grid.best_estimator_
model.fit(X_train, y_train)
print("Accuracy score:", accuracy_score(y_test, model.predict(X_test)))
```

PCA - ANN

PCA-ANN

Generate Code Markdown

```
pca = PCA()
mlp = MLPClassifier()
pipeline = Pipeline(steps=[('pca', pca), ('mlp', mlp)])
param_grid = {
    'mlp__solver': ['adam', 'sgd'],
    'mlp__hidden_layer_sizes': [(2,2), (3,3), (4,4), (5,5), (6,6), (7,7), (8,8), (9,9), (10, 10), (12,12), (20, 20)],
    'mlp__batch_size': [8],
    'mlp__early_stopping': [True],
}
grid = GridSearchCV(pipeline, param_grid, refit = True, verbose = 3)
grid.fit(X_train, y_train)
print("\n\n")
ic(grid.best_params_)
model = grid.best_estimator_
model.fit(X_train, y_train)
print("Accuracy score:", accuracy_score(y_test, model.predict(X_test)))
```

XGBoost

XGBoost

```
from xgboost import XGBClassifier
```

```
pca = PCA()
xgb = XGBClassifier()
pipeline = Pipeline(steps=[('pca', pca), ('xgb', xgb)])
param_grid = {
    'xgb__n_estimators': [50, 100, 150, 200],
}
grid = GridSearchCV(pipeline, param_grid, refit = True, verbose = 3)
grid.fit(X_train, y_train)
print("\n\n")
ic(grid.best_params_)
model = grid.best_estimator_
model.fit(X_train, y_train)
print("Accuracy score:", accuracy_score(y_test, model.predict(X_test)))
```

SNV

```
def snv(input_data):
    # Define a new array and populate it with the corrected data
    output_data = np.zeros_like(input_data)
    for i in range(input_data.shape[0]):
        # Apply correction
        output_data[i,:] = (input_data[i,:] - np.mean(input_data[i,:])) / np.std(input_data[i,:])
    return output_data
```

```
X = snv(X)
X = savgol_filter(X, window_length=15, polyorder=3, deriv=2)
```

```
plt.plot(wl, X.T, linewidth = 0.7)
plt.xlabel("E(V)")
plt.ylabel("I(mA)")
plt.show()
```

SPA

```
import numpy as np
from joblib import Parallel, delayed
class SPA():
    def __init__(self,
                 n_features_to_select: int = None,
                 n_cv_folds: int = 5,
                 pls: PLSRegression = None,
                 n_jobs: int = -1):
        self.n_features_to_select = n_features_to_select
        self.pls = PLSRegression()
        self.n_cv_folds = n_cv_folds
        self.n_jobs = n_jobs

    def fit_spa(self, X, y, seed):
        wavelength_map = np.arange(X.shape[1])
        selected_wave_length = [seed]
        selected_vector = X[:, seed:seed+1]
        remain_wavelength = np.delete(wavelength_map, seed)
        remain_vectors = np.delete(X, seed, 1)

        for i in range(self.n_features_to_select - 1):
            selected_vector = selected_vector / np.linalg.norm(selected_vector, ord = 2)
            orthogonal_vectors = remain_vectors - selected_vector @ np.transpose(np.transpose(remain_vectors) @ selected_vector)
            orthogonal_norms = np.linalg.norm(orthogonal_vectors, ord = 2, axis=0)
            max_length_vector_idx = np.argmax(orthogonal_norms)

            selected_wave_length.append(remain_wavelength[max_length_vector_idx])
            wavelength_map = np.delete(wavelength_map, max_length_vector_idx, )

            selected_vector = orthogonal_vectors[:, max_length_vector_idx:max_length_vector_idx+1]
            remain_wavelength = np.delete(remain_wavelength, max_length_vector_idx)
            remain_vectors = np.delete(remain_vectors, max_length_vector_idx, 1)
```

CA

```
sample_names = label
# Create a table showing which sample belongs to which cluster
# Create the cluster table
cluster_table = pd.DataFrame({
    "Sample": sample_names, # Replace 'data["Type"]' with your sample labels
    "Cluster": clusters
})

# Sort the table by Cluster for clarity
cluster_table = cluster_table.sort_values("Cluster")

# Print the cluster table
counts = cluster_table.groupby(["Cluster", "Sample"]).size().unstack(fill_value=0)

# Step 2: Plot the stacked bar chart
counts.plot(kind="bar", stacked=True, color=["skyblue", "lightcoral", "orange"], figsize=(8, 6))

# Step 3: Customize the plot
# plt.title("Count of 'T' and 'D' in Each Cluster")
plt.xlabel("Cluster")
plt.ylabel("Count")
plt.legend(title="Sample Type", labels=["TN", "TM", "HH"])
plt.xticks(rotation=0)
plt.show()
print(counts)
proba_table = counts
for i in range(len(proba_table)):
    proba_table.iloc[i, :] = proba_table.iloc[i, :] / proba_table.iloc[i, :].sum()
print(proba_table)
```