Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.385

DỰ ĐOÁN ĐỘ NGỌT CỦA XOÀI TRÊN CƠ SỞ DỮ LIỆU PHỔ THU THẬP TỪ CẢM BIẾN ĐA PHỔ GIÁ THÀNH THẤP

Nguyễn Phước Lộc[1,2], Dương Văn Sử[1], Trần Nhựt Thanh[1], Nguyễn Chí Ngôn[1] và Nguyễn Chánh Nghiệm[1*]

[1]*Khoa Tự động hóa, Trường Bách Khoa, Trường Đại học Cần Thơ*
[2]*Khoa Điện – Điện tử Máy tính, Trường Cao đẳng nghề Kiên Giang*
**Tác giả liên hệ (Corresponding author): ncnghiem@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 22/05/2024
Sửa bài (Revised): 29/07/2024
Duyệt đăng (Accepted): 25/08/2024

Title: *Prediction of mangoes' sweetness based on spectral data acquired from a low-cost multispectral sensor*

Author(s): *Nguyen Phuoc Loc[1,2], Duong Van Su[1], Tran Nhut Thanh[1], Nguyen Chi Ngon[1] and Nguyen Chanh Nghiem[1*]*

Affiliation(s): *[1]Can Tho University, [2]Kien Giang Vocational College*

TÓM TẮT

Nhiều nghiên cứu gần đây cho thấy cảm biến đa phổ giá thành thấp được quan tâm nhiều trong việc phát triển các ứng dụng trong nông nghiệp. Nghiên cứu này đánh giá tiềm năng sử dụng cảm biến đa phổ giá thành thấp trong việc dự đoán độ ngọt của xoài, loại trái cây có giá trị xuất khẩu cao. Để phát triển được mô hình dự đoán chính xác, một số giải thuật tiền xử lý và lựa chọn bước sóng đã được áp dụng. Kết quả cho thấy dữ liệu phổ không qua tiền xử lý trích xuất từ mười bốn bước sóng được chọn bởi giải thuật "hệ số hồi quy" là phù hợp để xây dựng mô hình hồi quy bình phương tối thiểu từng phần có hệ số tương quan bằng 0,703 và sai số RMSE là 1,439 °Brix. Kết quả này có thể so sánh được với các nghiên cứu gần đây sử dụng cùng loại cảm biến đa phổ vì thế khẳng định tiềm năng sử dụng cảm biến đa phổ giá thành thấp trong việc phát triển ứng dụng và thiết bị cầm tay để đánh giá chất lượng trái cây.

Từ khoá: *Đánh giá không phá hủy, cảm biến đa phổ, giá thấp*

ABSTRACT

Recent studies have shown that low-cost multispectral sensors have attracted much interest in developing agricultural applications. This study evaluated the potential of using a low-cost multispectral sensor to predict the sweetness of mango fruit with high export values. A few spectral data preprocessing and wavelength selection algorithms were applied to develop an accurate prediction model. Experimental results showed that unprocessed spectral data of fourteen wavelengths selected by the "regression coefficients" algorithm were suitable for developing a partial least square regression model with a correlation coefficient of 0.703 and a root mean square error of 1.439 °Brix. These results were comparable to recent studies using the same multispectral sensor, confirming the potential use of low-cost multispectral sensors in developing applications and portable devices for fruit quality assessment.

Keywords: *Low-cost, nondestructive assessment, multispectral sensor*

1. GIỚI THIỆU

Để đánh giá chất lượng các loại nông sản, đặc biệt là trái cây tươi, các giải pháp không làm hư mẫu và có thể ứng dụng trong các hệ thống tự động thường được quan tâm. Kết quả khảo sát trong thời gian gần đây cho thấy giải pháp phân tích quang phổ và thị giác máy được đặc biệt quan tâm (Nghiệm và ctv., 2021). Ưu điểm của thị giác máy tính là dễ thực hiện nhưng chủ yếu phù hợp để đánh giá chất lượng của đối tượng dựa vào đặc trưng bên ngoài. Nếu không có tương quan đủ lớn giữa đặc trưng bên ngoài và chất lượng bên trong, giải pháp thị giác máy tính có thể gặp một số hạn chế và kém chính xác. Khi cần thiết phải đánh giá chất lượng trái cây tươi như độ trưởng thành và giai đoạn chín thông qua độ ngọt, độ chua, hàm lượng chất khô,… (Mishra et al., 2020), giải pháp phân tích quang phổ có nhiều ưu thế hơn vì các chỉ tiêu đánh giá này phụ thuộc vào hàm lượng các thành phần hóa học của phần thịt/vỏ trái cây với khả năng hấp thụ ánh sáng khác nhau tại các bước sóng khác nhau. Vì thế, các đánh giá chất lượng trái cây dựa trên quang phổ dần trở thành giải pháp thay thế khi xem xét yếu tố tốc độ và bản chất không phá hủy (Lu et al., 2020).

Theo thống kê bởi Rogers et al. (2023), các nghiên cứu áp dụng phương pháp quang phổ khảo sát các vùng phổ với mức độ quan tâm giảm dần từ vùng phổ khả kiến và cận hồng ngoại có bước sóng ở khoảng 400–1000 nm, vùng phổ cận hồng ngoại có bước sóng khoảng 900–1700 nm và vùng hồng ngoại bước sóng ngắn 1700–2500 nm. Các vùng phổ được chọn phụ thuộc vào ứng dụng cần quan tâm. Ví dụ, vùng hồng ngoại phù hợp hơn cho các ứng dụng liên quan đến độ ẩm vì nước hấp thụ nhiều ánh sáng trong vùng này. Khi màu sắc có sự tương quan đến đặc tính quan tâm, cảm biến trong vùng ánh sáng khả kiến sẽ phù hợp hơn.

Tùy vào độ phân giải phổ cao hay thấp, cảm biến có thể được phân loại là siêu phổ hay đa phổ. Ưu điểm của cảm biến siêu phổ là độ phân giải cao hay khả năng đo được tín hiệu ở dải tần hẹp nhờ đó thu được nhiều thông tin phổ đặc trưng (Noguera et al., 2022). Với độ phân giải phổ thấp của cảm biến đa phổ, các ứng dụng phát triển dựa trên các cảm biến này có thể đạt độ chính xác không quá cao. Vì thế, cảm biến đa phổ thường có giá thành thấp hơn so với cảm biến siêu phổ. Tuy nhiên, nhiều cảm biến đa phổ giá thành thấp lại được phát triển cho vùng phổ ánh sáng khả kiến, rất có thể vì đây là vùng phổ phù hợp cho nhiều ứng dụng thực tế. Phát triển các ứng dụng trong nông nghiệp dựa trên các cảm biến này cũng trở thành chủ đề thu hút nhiều quan tâm (Walsh et al., 2020). Một số nghiên cứu cũng cho thấy cảm biến đa phổ có thể được sử dụng để phát triển giải pháp hay thiết bị cầm tay với độ chính xác phù hợp dùng cho việc đánh giá chất lượng trái cây (Nguyen et al., 2020; Tran & Fukuzawa, 2020; Noguera et al., 2022; Mohammed et al., 2023; Srinivasagan et al., 2023).

Để ứng dụng phương pháp phân tích quang phổ, một số vấn đề đã được xem xét giải quyết như loại bỏ bớt nhiễu do hiện tượng tán xạ ánh sáng bị ảnh hưởng bởi cấu trúc mô của phần thịt trái cây (Lu et al., 2020) thông qua tiền xử lý dữ liệu phổ (Mishra et al., 2020; Luka et al., 2024), sự ảnh hưởng của nhiệt độ (Golic et al., 2003; Mishra et al., 2020). Việc chọn lựa bước sóng quan trọng để loại bỏ các bước sóng không đóng góp đáng kể cho độ chính xác của mô hình cũng cần thiết để giảm thời gian xử lý và tăng tính bền vững của giải pháp khi xây dựng mô hình hồi quy đa biến (Rogers et al., 2023).

Nhằm đánh giá khả năng sử dụng cảm biến đa phổ giá thành thấp để phát triển các ứng dụng hay thiết bị đánh giá chất lượng trái cây dựa trên phương pháp phân tích quang phổ, nghiên cứu này thực hiện dự đoán độ ngọt của trái cây trên cơ sở dữ liệu phổ thu thập từ cảm biến đa phổ giá thành thấp với đối tượng nghiên cứu là trái xoài tươi vì tiềm năng xuất khẩu và giá trị kinh tế của xoài khá cao. Một số giải thuật tiền xử lý dữ liệu phổ và lựa chọn bước sóng được khảo sát để tìm ra giải thuật phù hợp nhằm phát triển mô hình đánh giá độ ngọt của xoài (thông qua độ Brix) đạt được độ chính xác cao. Kết quả nghiên cứu sẽ là cơ sở cho việc sử dụng cảm biến đa phổ giá thành thấp để phát triển các ứng dụng và thiết bị cầm tay trong việc đánh giá chất lượng trái cây.

2. PHƯƠNG PHÁP NGHIÊN CỨU

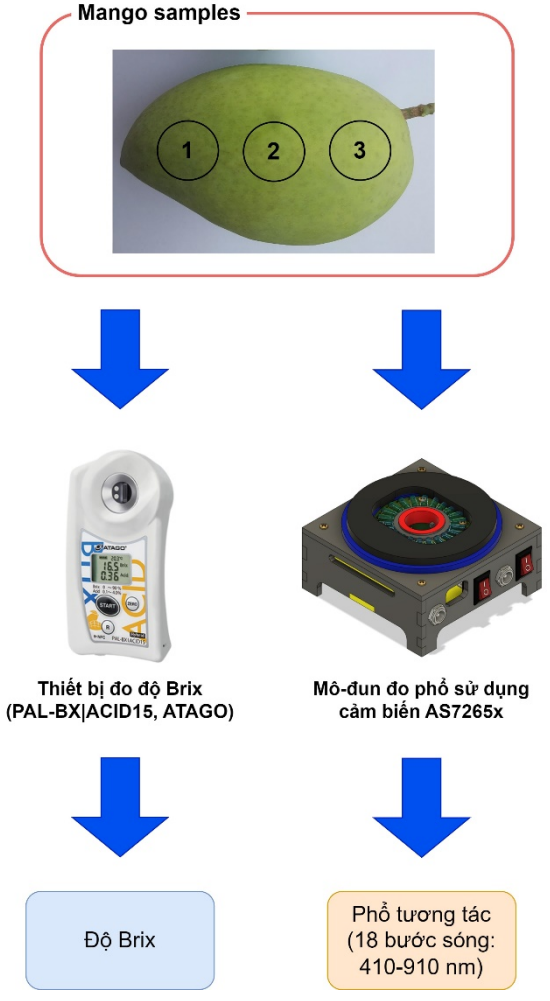
2.1. Bố trí thu mẫu

Để đảm bảo tính tổng quát của nghiên cứu, xoài sử dụng trong nghiên cứu được thu hoạch ở nhiều thời điểm khác nhau, từ 70, 75, 80, 85, 90 và 95 ngày sau khi đậu trái (Days after fruit set – DAFS) (Bảng 1). Các mẫu xoài đều được thu hoạch tại Nông trường Sông Hậu, huyện Cờ Đỏ, thành phố Cần Thơ để đảm bảo chất lượng xoài không chịu ảnh hưởng bởi vùng trồng và tập quán canh tác. Đối với mỗi mẫu xoài, dữ liệu độ Brix và phổ tương tác được đo từ ba vị trí trên mặt bên của xoài, là mặt có diện tích lớn nhất của xoài như mô tả ở Hình 1. Một mô-đun đo phổ sử dụng cảm biến phổ AS7265x (ams-OSRAM AG) để thu phổ tại 18 bước sóng trong vùng ánh sáng khả kiến và cận hồng ngoại từ 410 đến 940 nm (Hình 2). Mô-đun này được thiết kế để

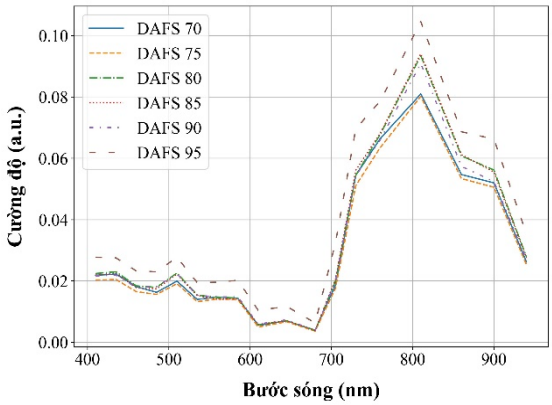
thu phổ ở chế độ tương tác (interactance mode), nghĩa là thu tín hiệu ánh sáng từ bên trong mẫu xoài truyền ra khi chặn hoàn toàn sự ảnh hưởng của ánh sáng bên ngoài, dựa trên một số kết quả mang tính khả thi và thiết kế trước đó (Nguyen et al., 2020; Tran & Fukuzawa, 2020; Tran et al., 2021).

Bảng 1. Thông tin số lượng và thời điểm thu hoạch xoài được sử dụng trong nghiên cứu

Thời điểm thu hoạch	Ngày sau khi đậu trái						Tổng
	70	75	80	85	90	95	
05/2021	11	12	12	12	12	11	70
03/2022	24	19	07	00	00	00	50
04/2022	08	19	30	37	24	12	130
03/2023	05	04	00	00	00	00	09
04/2023	00	00	13	16	07	03	39
Tổng	**48**	**54**	**62**	**65**	**43**	**26**	**298**

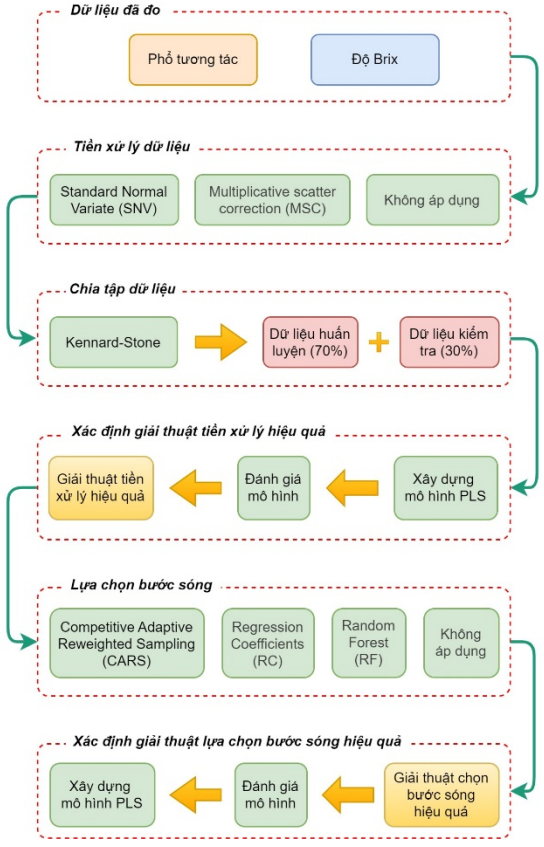


Hình 1. Mô tả quá trình đo mẫu



Hình 2. Phổ trung bình theo ngày sau khi đậu trái (DAFS)

2.2. Xây dựng mô hình



Hình 3. Lưu đồ xây dựng mô hình dự đoán độ ngọt của xoài

Trong nghiên cứu này, mô hình hồi quy được xây dựng để dự đoán độ ngọt của xoài dựa trên dữ liệu phổ. Phương pháp hồi quy bình phương tối thiểu từng phần (Partial Least Square – PLS) được áp dụng vì PLS là một giải thuật học máy hiệu quả khi xây dựng các mô hình hồi quy sử dụng dữ liệu phổ

(Flynn et al., 2023) và được ứng dụng phổ biến để đánh giá chất lượng trái cây dựa trên quang phổ (Zhang & Yang, 2024). Các giải thuật tiền xử lý dữ liệu và lựa chọn bước sóng phổ biến được áp dụng để phát triển mô hình hồi quy có hiệu suất cao nhất. Giải thuật tiền xử lý dữ liệu và lựa chọn bước sóng phù hợp được lựa chọn dựa trên cơ sở so sánh hiệu suất của mô hình PLS khi áp dụng các giải thuật tiền xử lý dữ liệu và lựa chọn bước sóng khác nhau như được tóm tắt ở Hình 3.

2.2.1. Tiền xử lý dữ liệu

Dữ liệu phổ thông thường là kết quả của việc hấp thụ và tán xạ ánh sáng (Lu et al., 2020), lần lượt do các thành phần hóa học có trong phần thịt trái cây và cấu trúc vật lý của phần thịt và vỏ của trái cây quyết định (Mishra et al., 2020). Vì thế, hiện tượng tán xạ ánh sáng ảnh hưởng đến kết quả dự đoán đặc tính của trái cây (như độ ngọt, độ chua) chủ yếu phụ thuộc vào các thành phần hóa học của thịt quả (như lượng đường, lượng axit). Trong nghiên cứu này, hai giải thuật tiền xử lý là Biến chuẩn hóa (standard normal variate – SNV) và hiệu chỉnh phân tán nhân (multiplicative scatter correction – MSC) được chọn vì hai giải thuật này được áp dụng phổ biến để loại bỏ các sai lệch trong dữ liệu phổ do ảnh hưởng của hiện tượng tán xạ ánh sáng.

Giải thuật SNV là một giải thuật phổ biến để chuẩn hóa dữ liệu phổ nhằm loại bỏ các hiệu ứng số nhân (multiplicative effects) do tán xạ ánh sáng và hiệu ứng cộng (additive effects) do các khác biệt về cường độ tín hiệu toàn cục (global signal intensities) (Barnes et al., 1989; Mishra et al., 2020). SNV sửa sai phổ bằng cách trừ cho phổ trung bình sau đó chia cho độ lệch chuẩn của phổ tín hiệu như sau (Luka et al., 2024):

$$x_{ij}(\text{SNV}) = \frac{x_{ij} - \overline{x}_i}{\sqrt{\frac{\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_i\right)^2}{m-1}}}, \quad (1)$$

trong đó x_{ij}, $x_{ij}(\text{SNV})$ lần lượt là biên độ của tín hiệu tại bước sóng thứ *j* của mẫu phổ thứ *i* trước và sau khi điều chỉnh bởi SNV, *m* là số bước sóng và giá trị trung bình của mẫu phổ thứ *i* được tính bởi

$$\overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij} \ . \quad (2)$$

Giải thuật MSC xem phổ như kết quả tổng hợp của tán xạ và hấp thu ánh sáng. Giải thuật sử dụng một phổ tham chiếu thông thường là trị trung bình và dịch chuyển các mẫu phổ gần nhất đến phổ tham chiếu thông qua việc chia tỉ lệ và dời phổ. MSC xem tán xạ khuếch tán là như nhau cho tất cả mẫu phổ và tại tất cả các bước sóng từ đó áp dụng hồi quy bình phương cực tiểu để ước lượng độ dốc và độ lệch của một hàm tuyến tính dùng để điều chỉnh từng mẫu phổ theo phương trình

$$x_{corr} = \frac{1}{b}(x - a), \quad (3)$$

với x_{corr} là mẫu phổ được điều chỉnh, *x* là phổ chưa điều chỉnh, *a* là tham số độ lệch và *b* là độ dốc tìm được bởi MSC (Mishra et al., 2020).

2.2.2. Phân chia dữ liệu

Để xây dựng và đánh giá hiệu suất của mô hình được xây dựng, dữ liệu phổ và độ Brix được chia thành tập dữ liệu huấn luyện và tập dữ liệu kiểm tra với tỉ lệ 7:3. Giải thuật Kennard-Stone (KS), một giải thuật phổ biến để phân chia dữ liệu phổ, được sử dụng trong nghiên cứu. Giải thuật này chia dữ liệu sao cho tối đa hóa khoảng cách Euclidean giữa các mẫu dữ liệu. Giải thuật KS được trình bày chi tiết bởi Luka et al. (2024).

2.2.3. Lựa chọn bước sóng

Các đặc trưng phổ có thể chứa thông tin phổ dư thừa, không liên quan đến biến đáp ứng của mô hình. Các đặc trưng này cần được loại bỏ để tăng độ chính xác của mô hình, giảm bớt độ phức tạp của mô hình toán và tránh hiện tượng quá khớp đối khi xây dựng mô hình hồi quy đa biến. Trong nghiên cứu này, giải thuật "Lấy mẫu có trọng số thích ứng cạnh tranh" (competitive adaptive reweighted sampling – CARS), "Hệ số hồi quy" (regression coefficients – RC) và "Mô hình rừng cây" (random forest – RF) được áp dụng.

Giải thuật CARS được ứng dụng rất phổ biến trong việc lựa chọn bước sóng (Rogers et al., 2023). CARS được phát triển trên cơ sở thuyết tiến hóa của Darwin và ứng dụng phương pháp chọn mẫu Monte Carlo. CARS chọn tập con các biến thông qua các chu kỳ tiến hóa và cạnh tranh dựa trên hàm mũ suy biến và việc lấy mẫu có trọng số thích ứng (Luka et al., 2024). Thông tin chi tiết hơn về giải thuật được trình bày bởi Li et al. (2009).

Cũng giống như CARS, giải thuật RC cũng là một trong những giải thuật phổ biến nhất để lựa chọn bước sóng (Rogers et al., 2023). So với giải thuật CARS, giải thuật RC có độ phức tạp ít hơn khi sử dụng các hệ số hồi quy để phát hiện các bước sóng có đóng góp nhiều cho sự thay đổi của biến đáp ứng.

Hệ số hồi quy được tính khi xây dựng mô hồi quy PLS. Các bước sóng có hệ số hồi quy lớn được xem như chứa đựng nhiều thông tin hơn và các bước sóng có hệ số hồi quy nhỏ sẽ được loại bỏ (Luka et al., 2024). Vì thế, dữ liệu tương ứng với các bước sóng quan trọng được sử dụng để xây dựng mô hình PLS. Kỹ thuật tìm kiếm tham số grid search với việc kiểm chứng chéo trên 10 tập con (10-fold cross validation) được áp dụng để xây dựng mô hình PLS tốt nhất.

Thời gian gần đây, giải thuật RF được ứng dụng hiệu quả để lựa chọn bước sóng và xây dựng mô hình phân loại xoài dựa vào độ ngọt (Nguyen et al., 2020). Vì thế, RF cũng được ứng dụng trong nghiên cứu này. Với mỗi một mô hình RF, thông số Mean decrease in impurity – MDI đều có thể được tính và xem là thước đo cho mức độ quan trọng của từng đặc trưng (hoặc biến độc lập) của mô hình (Li et al., 2019). Một đặc trưng có thể xem như quan trọng nếu thông số MDI của đặc trưng đó vượt ngưỡng được tính bởi giá trị nghịch đảo của số lượng đặc trưng. Bằng cách thực hiện giải thuật tìm kiếm tham số grid search nhiều lần, nhiều mô hình RF tối ưu được tạo ra. Một đặc trưng hay biến độc lập được xem là quan trọng và được lựa chọn nếu hơn một nữa số mô hình RF được tạo ra đều cho thấy đặc trưng đó là quan trọng dựa trên thông số MDI của đặc trưng đó. Cách chọn lựa đặc trưng được trình bày chi tiết bởi Nguyen et al. (2020).

2.3. Tiêu chí đánh giá mô hình

Hiệu suất mô hình ước lượng độ ngọt của xoài được xác định thông qua hệ số tương quan *R*, căn bậc hai của trung bình bình phương sai số (Root Mean Square Error – RMSE), trung bình sai số tuyệt đối (Mean Absolute Error – MAE) và tỉ lệ dự đoán đối với độ lệch (Ratio of Prediction to Deviation – RPD).

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (5)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (6)$$

$$RPD = \frac{SD}{SEP} \quad (7)$$

với

$$SEP = \sqrt{\frac{\sum_{i=1}^{n_p}(\hat{y}_i - y_i - b)^2}{n_p}} \quad (8)$$

$$b = \frac{1}{n_p}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (9)$$

trong đó,

- $\hat{y}_i$, y_i và x_i lần lượt là giá trị dự đoán, giá trị đo và phổ của mẫu đo thứ *i*;
- $\bar{x}$ và $\bar{y}$ là giá trị trung bình của *n* mẫu phổ và giá trị đo.
- *SD* là độ lệch chuẩn của giá trị tham chiếu của tập dữ liệu kiểm tra và *SEP* là độ lệch chuẩn của sai số dự đoán thực hiện trên tập dữ liệu kiểm tra (Cayuela & García, 2017).
- n_p là số mẫu trong tập dữ liệu kiểm tra.

Là một tiêu chí được sử dụng phổ biến để đánh giá các mô hình ước lượng từ dữ liệu phổ, giá trị *RPD* cũng được sử dụng trong nghiên cứu này. Giá trị *RPD* có thể được chia thành ba vùng giá trị [1,5; 2,0], (2,0; 2,5] và (2,5; ∞) lần lượt thể hiện khả năng của mô hình có thể phân biệt được các giá trị cao và thấp, đưa ra dự đoán định lượng mang tính chất thô (coarse quantitative prediction) và dự đoán xuất sắc giá trị quan tâm (Nicolaï et al., 2007).

3. KẾT QUẢ VÀ THẢO LUẬN

Hiệu suất mô hình PLS được xây dựng với các giải thuật tiền xử lý khác nhau được trình bày ở Bảng 2. Mô hình PLS xây dựng dựa trên bộ dữ liệu huấn luyện đã qua bước tiền xử lý bằng giải thuật SNV và MSC không cho thấy sự khác biệt đáng kể trong việc cải thiện sai số *RMSE* và *MAE* so với mô hình được xây dựng từ dữ liệu thô. Ngoài ra, kết quả dự đoán của mô hình PLS xây dựng từ phổ thô đều cho kết quả tốt hơn. Cụ thể, khi không tiền xử lý dữ liệu phổ, sai số $RMSE_p$ chỉ khoảng 1,4 so với 2,2 khi thực hiện tiền xử lý. Trung bình sai số tuyệt đối MAE_p khi không áp dụng giải thuật tiền xử lý chỉ là 0,873 °Brix, nhỏ hơn khoảng 28% so với áp dụng giải thuật MSC. Nhiều nghiên cứu cũng chỉ ra rằng mô hình hồi quy xây dựng dựa trên dữ liệu thô có hiệu suất tốt hơn (Nordey et al., 2017; Malvandi et al., 2022). Ngoài ra, việc tìm giải thuật tiền xử lý dữ

liệu phù hợp để nâng cao hiệu suất mô hình cần thực nghiệm thử-và-sai nhiều lần (Engel et al., 2013). Dựa trên kết quả so sánh hiệu suất các mô hình được trình bày ở Bảng 2, phổ thô được sử dụng để xây dựng các mô hình PLS nhằm đánh giá giải thuật lựa chọn bước sóng hiệu quả. Hiệu suất các mô hình được xây dựng dựa trên dữ liệu phổ đã được áp dụng giải thuật chọn lựa các bước sóng khác nhau được tóm tắt ở Bảng 3. Kết quả cho thấy giải thuật RF không hiệu quả khi mô hình tương ứng có sai số lớn hơn và hệ số tương quan, giá trị *RPD* đều nhỏ hơn so với mô hình có áp dụng giải thuật RC hay CARS.

Cả hai giải thuật RC và CARS đều hiệu quả đối với dữ liệu phổ trong nghiên cứu này vì hiệu suất mô hình có áp dụng CARS và RC đều tương đồng với mô hình sử dụng dữ liệu phổ của tất cả 18 bước sóng. Dữ liệu phổ thể hiện mối tương quan tốt với đáp ứng của mô hình với R_p = 0,703. Trong đó, mô hình áp dụng giải thuật RC tốt hơn CARS khi xem xét các tiêu chí R_c, $RMSE_c$ và *RPD*.

Kết quả lựa chọn bước sóng được trình bày ở Bảng 4 cũng cho thấy sự tương đồng cao giữa giải thuật RC và CARS. Trong 14 bước sóng quan trọng để xây dựng mô hình có hiệu suất cao, chỉ có sự khác biệt ở 1 bước sóng. Trong khi CARS chọn bước sóng 680 nm, có liên hệ với sự hấp thụ Chlorophyll ở bước sóng trong lân cận 670 nm (Posom et al., 2020), giải thuật RC chọn bước sóng 810 nm với sự liên hệ gần với màu sắc của thịt xoài do thành phần β-carotene của thịt xoài (Rungpichayapichet et al., 2015; Nordey et al., 2017). Cả RC và CARS đều chọn các bước sóng có liên hệ với thành phần của thịt xoài. Các bước sóng 435 và 460 nm đều rất gần với vùng hấp thụ hẹp của chiết xuất Chlorophyll a (gần bước sóng 428 nm) và Chlorophyll b (gần bước sóng 453 nm) (Alós et al., 2019) là các thành phần có liên hệ chặt chẽ với thành phần đường trong thịt xoài ở các giai đoạn chín khác nhau (Gill et al., 2017). Các bước sóng tại 730, 900 và 940 nm đều liên quan chặt với vùng hấp thụ của đường do các dao động của các gốc C–H and O–H (Golic et al., 2003; Omar et al., 2012b). Bước sóng 860 nm gần vùng chịu ảnh hưởng bởi thành phần axit citric (tại bước sóng 850 nm (Omar et al., 2012a)) có hàm lượng giảm dần khi lượng đường trong xoài tăng do quá trình chín của xoài (Maldonado-Celis et al., 2019).

Bảng 2. Hiệu suất mô hình với các giải thuật tiền xử lý dữ liệu khác nhau

Giải thuật tiền xử lý	Kết quả xây dựng mô hình			Kết quả dự đoán			
	R_c	$RMSE_c$	MAE_c	R_p	$RMSE_p$	MAE_p	*RPD*
Không áp dụng (sử dụng phổ thô)	0,713	2,534	1,591	0,703	1,434	0,873	1,405
SNV	0,704	2,391	1,572	0,677	2,188	1,232	1,358
MSC	0,711	2,428	1,569	0,582	2,165	1,208	1,230

Bảng 3. Hiệu suất mô hình với các giải thuật lựa chọn bước sóng khác nhau

Giải thuật lựa chọn bước sóng	Số bước sóng được chọn	Kết quả xây dựng mô hình			Kết quả dự đoán			
		R_c	$RMSE_c$	MAE_c	R_p	$RMSE_p$	MAE_p	*RPD*
Không áp dụng [a]	18	0,713	2,534	1,591	0,703	1,434	0,873	1,405
RF	3	0,611	2,860	1,915	0,538	1,665	1,023	1,186
RC	14	*0,720* [b]	*2,508*	1,596	0,703	1,439	0,890	*1,407*
CARS	14	0,701	2,577	1,594	0,703	1,458	*0,870*	*1,407*

[a] *Mô hình không áp dụng giải thuật lựa chọn bước sóng là mô hình xây dựng với dữ liệu phổ thô được mô tả ở Bảng 2*

[b] *Các giá trị in nghiêng thể hiện kết quả tốt hơn so với mô hình không áp dụng giải thuật lựa chọn bước sóng*

Bảng 4. Kết quả lựa chọn bước sóng

Giải thuật	Bước sóng																	
	410	**435**	**460**	**485**	**510**	**535**	**560**	**585**	**610**	**645**	**680**	**705**	**730**	**760**	**810**	**860**	**900**	**940**
RF	-	-	-	-	-	-	-	-	-	X	X	X	-	-	-	-	-	-
RC	X	X	X	X	X	-	X	X	X	X	-	X	X	-	X	X	-	X
CARS	X	X	X	X	X	-	X	X	X	X	X	X	X	-	-	X	-	X

"X" chỉ trường hợp bước sóng được chọn

Bảng 5 trình bày hiệu suất của các mô hình được xây dựng để dự đoán độ Brix của một số loại trái cây dựa trên dữ liệu phổ thu thập bởi cảm biến AS7265x trong các nghiên cứu liên quan. Kết quả cho thấy mô hình dự đoán độ ngọt của xoài có sai số $RMSE_p$ lớn hơn so với các nghiên cứu liên quan (từ 0,224 đến 1,031 °Brix). Tuy nhiên, kết quả này có thể xem như so sánh được với các nghiên cứu cùng sử dụng cảm biến AS7265x vì vỏ xoài tương đối dầy hơn so với vỏ táo và nho. Ngoài ra, mô hình đã xây dựng đạt RPD = 1,407, bằng với kết quả công bố của một nghiên cứu dự đoán độ ngọt của cà chua khi sử dụng cảm biến siêu phổ với độ phân giải cao trong vùng 400–1100 nm (Huang et al., 2018). Như vậy, các kết quả nghiên cứu cho thấy tiềm năng trong việc sử dụng cảm biến giá thành thấp, cụ thể là cảm biến đa phổ AS7265x được sử dụng trong nghiên cứu này.

Bảng 5. Các nghiên cứu sử dụng cảm biến AS7265x để dự đoán độ Brix của trái cây

Nghiên cứu	Đối tượng	Mô hình	Hiệu suất dự đoán
Tran & Fukuzawa (2020)	Táo	MLR [a]	$R_p^2 = 0,861$ [b]; $RMSE_p = 0,403$
Noguera et al. (2022)	Nho	ANN [c]	$R_p^2 = 0,7$; $RMSE_p = 1,210$
Zhao et al. (2023)	Táo	PLS	$R_p = 0,8568$; $RMSE_p = 0,7753$
Nghiên cứu này	Xoài	PLS	$R_p = 0,703$ $R =$ $RMSE_p = 1,439$

[a] *Mô hình hồi quy đa biến (multiple linear regression)*

[b] *Hệ số xác định (coefficient of determination)*

[c] *Mô hình mạng nơ-ron nhân tạo (artificial neural network)*

Mô hình PLS đã được đề xuất để bước đầu đánh giá tiềm năng ứng dụng cảm biến giá thành thấp. Vì thế, các mô hình học máy khác cần được xem xét trong các nghiên cứu tiếp theo để có thể phát triển được mô hình có độ chính xác cao hơn (Zhang & Yang, 2024) nhằm ứng dụng các cảm biến giá thành thấp hiệu quả hơn.

4. KẾT LUẬN

Nghiên cứu đã xây dựng mô hình PLS để dự đoán độ ngọt của xoài trên cơ sở dữ liệu phổ thu thập từ cảm biến đa phổ Vis-NIR giá thành thấp. Kết quả cho thấy phổ thu thập tương quan tốt với độ Brix và có thể dùng thông tin phổ của 14 bước sóng quan trọng để dự đoán độ ngọt của xoài với sai số dự đoán $RMSE_p$ = 1,439 °Brix. Kết quả trên có thể so sánh được với một số nghiên cứu liên quan vì thế cho thấy tiềm năng sử dụng các loại cảm biến đa phổ giá thành thấp trong việc phát triển các giải pháp hay thiết bị cầm tay để đánh giá định lượng một số chỉ tiêu chất lượng của trái cây tươi.

TÀI LIỆU THAM KHẢO

Alós, E., Rodrigo, M. J., & Zacarias, L. (2019). Ripening and Senescence. In *Postharvest Physiology and Biochemistry of Fruits and Vegetables* (pp. 131–155). Elsevier. https://doi.org/10.1016/B978-0-12-813278-4.00007-5

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*, *43*(5), 772–777. https://doi.org/10.1366/0003702894202201

Cayuela, J. A., & García, J. F. (2017). Sorting olive oil based on alpha-tocopherol and total tocopherol content using near-infra-red spectroscopy (NIRS) analysis. *Journal of Food Engineering*, *202*, 79–88. https://doi.org/10.1016/j.jfoodeng.2017.01.015

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, *50*, 96–106. https://doi.org/10.1016/j.trac.2013.04.015

Flynn, K. C., Baath, G., Lee, T. O., Gowda, P., & Northup, B. (2023). Hyperspectral reflectance and machine learning to monitor legume biomass and nitrogen accumulation. *Computers and Electronics in Agriculture*, *211*, 107991. https://doi.org/10.1016/j.compag.2023.107991

Gill, P. P. S., Jawandha, S. K., & Kaur, N. (2017). Transitions in mesocarp colour of mango fruits kept under variable temperatures. *Journal of Food Science and Technology*, *54*(13), 4251–4256. https://doi.org/10.1007/s13197-017-2894-z

Golic, M., Walsh, K., & Lawson, P. (2003). Short-Wavelength Near-Infrared Spectra of Sucrose, Glucose, and Fructose with Respect to Sugar Concentration and Temperature. *Applied*

Spectroscopy, *57*(2), 139–145. https://doi.org/10.1366/000370203321535033

Huang, Y., Lu, R., & Chen, K. (2018). Assessment of tomato soluble solids content and pH by spatially-resolved and conventional Vis/NIR spectroscopy. *Journal of Food Engineering*, *236*(May), 19–28. https://doi.org/10.1016/j.jfoodeng.2018.05.008

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, *648*(1), 77–84. https://doi.org/10.1016/j.aca.2009.06.046

Li, X., Wang, Y., Basu, S., Kumbier, K., & Yu, B. (2019). A debiased MDI feature importance measure for random forests. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Lu, R., Van Beers, R., Saeys, W., Li, C., & Cen, H. (2020). Measurement of optical properties of fruits and vegetables: A review. *Postharvest Biology and Technology*, *159*, 111003. https://doi.org/10.1016/j.postharvbio.2019.111003

Maldonado-Celis, M. E., Yahia, E. M., Bedoya, R., Landázuri, P., Loango, N., Aguillón, J., Restrepo, B., & Guerrero Ospina, J. C. (2019). Chemical Composition of Mango (Mangifera indica L.) Fruit: Nutritional and Phytochemical Compounds. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01073

Malvandi, A., Kapoor, R., Feng, H., & Kamruzzaman, M. (2022). Non-destructive measurement and real-time monitoring of apple hardness during ultrasonic contact drying via portable NIR spectroscopy and machine learning. *Infrared Physics & Technology*, *122*(February), 104077. https://doi.org/10.1016/j.infrared.2022.104077

Mishra, P., Roger, J. M., Rutledge, D. N., & Woltering, E. (2020). SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials. *Postharvest Biology and Technology*, *168*, 111271. https://doi.org/10.1016/j.postharvbio.2020.111271

Mohammed, M., Srinivasagan, R., Alzahrani, A., & Alqahtani, N. K. (2023). Machine-Learning-Based Spectroscopic Technique for Non-Destructive Estimation of Shelf Life and Quality of Fresh Fruits Packaged under Modified Atmospheres. *Sustainability (Switzerland)*, *15*(17). https://doi.org/10.3390/su151712871

Nghiệm, N. C., Lộc, N. P., Dũng, N. H. & Ngôn, N. C. (2021). Tổng quan về đánh giá chất lượng trái cây bằng phương pháp không phá hủy. *Tạp chí Khoa học và Công nghệ Đại học Thái Nguyên*, *226*(11), 158–167. https://doi.org/10.34238/tnu-jst.4673

Nguyen, C.-N., Phan, Q.-T., Tran, N.-T., Fukuzawa, M., Nguyen, P.-L., & Nguyen, C.-N. (2020). Precise Sweetness Grading of Mangoes (Mangifera indica L.) Based on Random Forest Technique with Low-Cost Multispectral Sensors. *IEEE Access*, *8*, 212371–212382. https://doi.org/10.1109/ACCESS.2020.3040062

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, *46*(2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Noguera, M., Millan, B., & Andújar, J. M. (2022). New, Low-Cost, Hand-Held Multispectral Device for In-Field Fruit-Ripening Assessment. *Agriculture*, *13*(1), 4. https://doi.org/10.3390/agriculture13010004

Nordey, T., Joas, J., Davrieux, F., Chillet, M., & Léchaudel, M. (2017). Robust NIRS models for non-destructive prediction of mango internal quality. *Scientia Horticulturae*, *216*, 51–57. https://doi.org/10.1016/j.scienta.2016.12.023

Omar, A. F., Atan, H., & MatJafri, M. Z. (2012a). NIR Spectroscopic Properties of Aqueous Acids Solutions. *Molecules*, *17*(6), 7440–7450. https://doi.org/10.3390/molecules17067440

Omar, A. F., Atan, H., & MatJafri, M. Z. (2012b). Peak Response Identification through Near-Infrared Spectroscopy Analysis on Aqueous Sucrose, Glucose, and Fructose Solution. *Spectroscopy Letters*, *45*(3), 190–201. https://doi.org/10.1080/00387010.2011.604065

Posom, J., Klaprachan, J., Rattanasopa, K., Sirisomboon, P., Saengprachatanarug, K., & Wongpichet, S. (2020). Predicting Marian Plum Fruit Quality without Environmental Condition Impact by Handheld Visible–Near-Infrared Spectroscopy. *ACS Omega*, *5*(43), 27909–27921. https://doi.org/10.1021/acsomega.0c03203

Rogers, M., Blanc-Talon, J., Urschler, M., & Delmas, P. (2023). Wavelength and texture feature selection for hyperspectral imaging: a systematic literature review. *Journal of Food Measurement and Characterization*, *17*(6), 6039–6064. https://doi.org/10.1007/s11694-023-02044-x

Rungpichayapichet, P., Mahayothee, B., Khuwijitjaru, P., Nagle, M., & Müller, J. (2015). Non-destructive determination of β-carotene content in mango by near-infrared spectroscopy compared with colorimetric measurements. *Journal of Food Composition and Analysis*, *38*,

32–41. https://doi.org/10.1016/j.jfca.2014.10.013

Luka, S. B., Mohammed Yunusa, B., Msurshima Vihikwagh, Q., Fanan Kuhwa, K., Hannah Oluwasegun, T., Ogalagu, R., Kenneth Yuguda, T., & Adnouni, M. (2024). Hyperspectral imaging systems for rapid assessment of moisture and chromaticity of foods undergoing drying: Principles, applications, challenges, and future trends. *Computers and Electronics in Agriculture*, *224*(June), 109101. https://doi.org/10.1016/j.compag.2024.109101

Srinivasagan, R., Mohammed, M., & Alzahrani, A. (2023). TinyML-Sensor for Shelf Life Estimation of Fresh Date Fruits. *Sensors*, *23*(16), 7081. https://doi.org/10.3390/s23167081

Tran, N.-T., & Fukuzawa, M. (2020). A Portable Spectrometric System for Quantitative Prediction of the Soluble Solids Content of Apples with a Pre-calibrated Multispectral Sensor Chipset. *Sensors*, *20*(20), 5883. https://doi.org/10.3390/s20205883

Tran, N.-T., Phan, Q.-T., Nguyen, C.-N., & Fukuzawa, M. (2021). Machine Learning-Based Classification of Apple Sweetness with Multispectral Sensor. *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, 23–27. https://doi.org/10.1109/SNPDWinter52325.2021.00014

Walsh, K. B., Blasco, J., Zude-Sasse, M., & Sun, X. (2020). Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biology and Technology*, *168*, 111246. https://doi.org/10.1016/j.postharvbio.2020.111246

Zhang, X., & Yang, J. (2024). Advanced chemometrics toward robust spectral analysis for fruit quality evaluation. *Trends in Food Science & Technology*, *150*, 104612. https://doi.org/10.1016/j.tifs.2024.104612

Zhao, X., Peng, Y., Li, Y., Wang, Y., Li, Y., & Chen, Y. (2023). Intelligent micro flight sensing system for detecting the internal and external quality of apples on the tree. *Computers and Electronics in Agriculture*, *204*(17), 107571. https://doi.org/10.1016/j.compag.2022.107571