Vietnam Journal of Food Control

A Scientific Journal of National Institute for Food Control

Journal homepage: https://vjfc.nifc.gov.vn

Research Article

# Machine learning and deep learning models applied to identification and classification of mango

**Nguyen Duc Phong[1], Nguyen Manh Son[1], Nguyen Manh Ha[1], Bui Xuan Thanh[1], Ta Thi Thao[1], Nguyen Thi Van Anh[2], Le Thi Hong Hao[1,3], Nguyen Duc Thanh[1,4*]**

[1]*University of Science, Vietnam National University - Hanoi, Hanoi, Vietnam*

[2]*Viet Nam University of Traditional Medicine and Pharmacy, Hanoi, Vietnam*

[3]*National Institute for Food Control, Hanoi, Vietnam*

[4]*Vietnam Military Medical Academy, Hanoi, Vietnam*

*(Received: 19 Jul 2024; Revised: 07 Sep 2024; Accepted: 09 Sep 2024)*

## *Abstract*

This study utilizes the data published on the website https://data.mendeley.com/datasets/46htwnp833/2, which includes visible-near-infrared (Vis-NIR) spectral data at wavelengths ranging from 309 nm to 1149 nm for 11691 mangoes in Australia, collected from 10 mango varieties across 2 different growing regions. The research developed machine learning models with open-source programming language Python such as: principal component analysis (PCA) combined with support vector machines (SVM), decision trees (DT), random forests (RF), and artificial neural networks (ANN); partial least squares model combined with discriminant analysis (PLS-DA); and a deep learning model 1-dimensional convolutional neural network (1D-CNN). The preprocessing steps were carried out based on the full spectral data with second derivative, smoothing using the Savitzky-Golay algorithm, and data balancing via a new Synthetic Minority Oversampling Technique (SMOTE). The results demonstrated that applying the SMOTE data preprocessing technique before running the machine learning models significantly enhanced classification accuracy. Furthermore, using a 1D-CNN model with a complex structure provided higher classification efficiency than conventional machine learning models. The accuracy of the 1D-CNN model in classifying mango ripeness, mango variety, and growing location was 99.40%, 94.35%, and 96.92%, respectively. The 1D-CNN deep learning model is well-suited for sample classification when dealing with large datasets containing tens of thousands of samples based on spectral data.

***Keywords:*** *mango classification, machine learning, deep learning, 1D-CNN, Vis-NIR spectra.*

* *Corresponding author: Nguyen Duc Thanh (E-mail: nguyenducthanh@vmmu.edu.vn )*

*Doi: https://doi.org/10.47866/2615-9252/vjfc.4370*

# Ứng dụng mô hình học máy và học sâu trong nhận dạng và phân loại quả xoài

**Nguyễn Đức Phong[1], Nguyễn Mạnh Sơn[1], Nguyễn Mạnh Hà[1], Bùi Xuân Thành[1], Tạ Thị Thảo[1], Nguyễn Thị Vân Anh[2], Lê Thị Hồng Hảo[1,3], Nguyễn Đức Thanh[1,4*]**

*[1]Trường Đại học Khoa học Tự Nhiên - ĐHQGHN, Hà Nội, Việt Nam*

*[2]Học viện Y dược học Cổ truyền Việt Nam, Hà Nội, Việt Nam*

*[3]Viện kiểm nghiệm vệ sinh an toàn thực phẩm quốc gia, Hà Nội, Việt Nam*

*[4]Học viện Quân Y, Hà Nội, Việt Nam*

## *Tóm tắt*

Nghiên cứu này kế thừa các số liệu đã công bố trên trang https://data.mendeley.com/datasets/46htwnp833/2 về dữ liệu phổ vùng khả kiến - hồng ngoại gần (Vis-NIR) ở khoảng bước sóng từ 309 nm đến 1149 nm của 11691 quả xoài tại Úc, lấy từ 10 giống xoài, thuộc 2 vùng trồng. Nghiên cứu đã phát triển các mô hình học máy với mã nguồn mở Python như: phân tích thành phần chính (PCA) kết hợp với máy vector hỗ trợ (SVM), cây quyết định (DT), rừng ngẫu nhiên (RF) và mạng thần kinh nhân tạo (ANN); mô hình bình phương tối thiểu từng phần kết hợp với phân tích biệt thức (PLS-DA), cùng với mô hình học sâu mạng thần kinh tích chập 1 chiều (1D-CNN) với các bước tiền xử lý dữ liệu phổ toàn phần bao gồm đạo hàm bậc hai và làm mịn bằng thuật toán Savitzky-Golay, cân bằng dữ liệu thông qua kỹ thuật tạo mẫu tổng hợp mới cho mẫu thiểu số (SMOTE). Kết quả cho thấy sử dụng thêm kĩ thuật tiền xử lý số liệu SMOTE trước khi chạy các mô hình học máy đã làm tăng đáng kể khả năng phân loại. Ngoài ra, mô hình 1D-CNN cho hiệu quả phân loại cao hơn so với các mô hình học máy thông thường với độ chính xác (qua phần trăm số mẫu nhận dạng đúng) của mô hình 1D-CNN trong phân loại độ chín của xoài, giống xoài, và địa điểm trồng lần lượt là 99,40%, 94,35% và 96,92%. Mô hình học sâu 1D-CNN thích hợp cho việc phân loại đối tượng khi có lượng lớn hàng chục nghìn mẫu dựa trên dữ liệu phổ.

***Từ khóa**: phân loại xoài, học máy, học sâu, 1D-CNN, phổ Vis- NIR.*

## 1. ĐẶT VẤN ĐỀ

Quả xoài (*Mangifera indica* L.), loại trái cây nhiệt đới phổ biến và được yêu thích, có nguồn gốc từ Nam Á, đặc biệt là từ khu vực Ấn Độ và Myanma, nhưng ngày nay được trồng rộng rãi ở các khu vực nhiệt đới và cận nhiệt đới trên toàn cầu như Mỹ Latinh, Châu Phi, và Đông Nam Á. Nhu cầu sử dụng xoài ngày càng tăng không chỉ do hương vị thơm ngon mà còn vì những lợi ích sức khỏe mà nó mang lại. Quả xoài có thể được tiêu thụ tươi như một món tráng miệng ngon, hay chế biến thành nhiều sản phẩm khác nhau như nước ép, sinh tố, kem, salad, mứt… và các món ăn chín khác. Về thành phần hóa học của xoài, có thể quan tâm đến ba nhóm chất chính gồm: i) chất dinh dưỡng đa lượng (carbohydrate, protein, lipid, chất béo và acid hữu cơ); ii) vi chất dinh dưỡng (vitamin và khoáng chất) và iii) chất thực

vật (phytochemicals) như các hợp chất phenolic, polyphenol, sắc tố và các thành phần dễ bay hơi [2]. Quả xoài chứa carbohydrate tạo cấu trúc như pectin và cellulose, các acid amine chủ yếu gồm lysine, leucine, cysteine, valine, arginine, phenylalanine và methionine. Hàm lượng lipid tăng lên trong quá trình chín, đặc biệt là acid béo omega-3 và omega-6. Các sắc tố quan trọng nhất của quả xoài gồm diệp lục (a và b) và carotenoid. Các acid hữu cơ như malic acid và citric acid tạo nên tính acid đặc trưng của loại trái cây này. Các hợp chất dễ bay hơi góp phần tạo nên hương thơm đặc trưng của xoài [3]. Hàm lượng các chất hóa học trong quả xoài có thể thay đổi do nhiều yếu tố như giống [4], điều kiện canh tác [5], giai đoạn chín của quả [6], nhiệt độ tại thời điểm thu hoạch [7] và điều kiện bảo quản [8]. Do đó, thông tin về các thành phần hóa học trong xoài này không chỉ giúp tối ưu hóa việc sử dụng dinh dưỡng quả xoài mà qua đó cho phép nhận dạng, phân loại từ đó lựa chọn được các quả xoài có chất lượng.

Trong phân loại quả xoài, nếu như con người thường sử dụng cảm quan thông qua các đặc trưng vật lý của quả xoài như màu sắc, khối lượng, kích thước, hình dáng thì với sự phát triển của công nghệ thông tin, việc ứng dụng thị giác máy tính (computer vision) kết hợp với các thuật toán trí tuệ nhân tạo (Artificial Inteligent- AI) nhờ cơ sở dữ liệu ảnh chụp quả xoài với camera CCD trong phòng thí nghiệm theo thời gian thực [9] hoặc hệ đo cầm tay kết hợp với phần mềm [10] là thông tin quan trọng giúp dự đoán được độ ngọt và độ chín của xoài. Trong số dữ liệu thu thập từ các thiết bị phân tích thì phổ hồng ngoại gần (NIR) hoặc vùng trung với kỹ thuật đo truyền qua hoặc phản xạ khuếch tán kết hợp với các thuật toán học máy đã được sử dụng để xác thực và phân loại giống xoài [11] hoặc phân loại chất lượng quả xoài theo mùi vị (như ngọt, và chua, chua và nhạt) phục vụ cho kiểm soát chất lượng quả trước khi được bán ra thị trường [12, 13].

Tập số liệu sử dụng trong nghiên cứu này được trích từ bài báo của tác giả N.T. Anderson và cộng sự về đánh giá tính ổn định của hàm lượng chất khô của quả xoài còn nguyên vẹn theo các yếu tố như mùa thu hoạch, địa điểm và giống xoài trên cơ sở thu thập dữ liệu độ hấp thụ quang trong phổ NIR của từng đối tượng (từng quả) [14] được công bố trên trang web https://data.mendeley.com/datasets/46htwnp833/2. Trong công trình nghiên cứu của các tác giả, hàm lượng chất khô đã được dự đoán với mô hình học máy có giám sát với thuật toán hồi qui tuyến tính PLS và phi tuyến tính ANN. Trong nghiên cứu này, thay vì dự đoán hàm lượng chất khô trong quả xoài như bài báo trích dẫn, mục tiêu của nghiên cứu nhằm xây dựng các mô hình học máy và học sâu, sử dụng kỹ thuật SMOTE [15] để làm cân bằng dữ liệu trên ngôn ngữ lập trình mã nguồn mở Python để trực tiếp phân loại quả xoài với các chỉ tiêu gồm độ chín, giống xoài, và vùng trồng (các thông số này đã được đề cập trong tập số liệu của tác giả) dựa trên toàn bộ dữ liệu phổ vùng Vis-NIR. Kết quả phân loại với ngôn ngữ lập trình mã nguồn mở hướng tới có thể tích hợp thuật toán vào các thiết bị đo cầm tay, phục vụ cho việc kiểm định, kiểm tra và xác thực trước khi quả xoài được đưa ra thị trường.

## 2. VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

### 2.1 Phương pháp lấy mẫu xoài và thu thập số liệu phổ

Tổng số mẫu (quả xoài thu thập tại Úc) các tác giả công bố là 11691 (mỗi quả là 1 hàng dữ liệu). Mỗi quả đi kèm thông tin gồm: i) 4 vụ (năm) thu hoạch (gán nhãn 1 đến 4 với các năm từ 2015-2019); ii) 2 vùng lấy mẫu (gán nhãn NT (Northern Territory ) và QLD (Central Queensland) (thuộc vùng nhiệt đới và cận nhiệt đới); iii) 10 giống *Calypso*™(Caly), *Honey Gold* (HG), *Keitt*, *Kensington Pride* (KP), *Lady Grace* (LadyG), *Lady Jane* (LadyJ), R2E2 và ba giống xoài thuộc Chương trình nhân giống Quốc gia Úc bao gồm NMBP1201, NMBP1243 và NMBP4069 và iv) 2 loại xoài (xanh, chín) và các thông tin khác về nhiệt độ khi phân tích mẫu[14].

Mỗi quả xoài được chụp phổ Vis-NIR trong vùng bước sóng từ 309-1149 nm, khoảng cách các điểm đo 3 nm, trên thiết bị cầm tay F750 (Felix Instruments, Camas, USA) với độ phân giải pixel khoảng 3,3 nm, độ phân giải quang học 10 nm, và độ lặp lại khoảng 1 đơn vị mAbs. Mỗi quả được đo hai lần ở phần rộng nhất của mỗi mặt (khoảng giữa quả), vuông góc với mặt phẳng hạt, tại ba mức nhiệt độ thấp-trung bình-cao là ~ 15, 25, 35°C. Cài đặt mặc định trên thiết bị đo bốn lần quét phổ trên mỗi mặt quả rồi tính trung bình cho mỗi lần đo mẫu.

### 2.2. Phương pháp xử lý tín hiệu phổ và xây dựng mô hình phân loại

Dữ liệu phổ toàn phầnVis-NIR từ 309 nm đến 1149 nm được xử lý với ngôn ngữ lập trình Python phiên bản 3.11. Tập dữ liệu được chia ngẫu nhiên theo tỷ lệ 80:20, tức là số lượng mẫu trong tập huấn luyện mô hình chiếm 80% còn số lượng mẫu trong tập kiểm tra chiếm 20% toàn bộ dữ liệu tương đương với 9352 mẫu luyện và 2339 mẫu kiểm tra. Toàn bộ dữ liệu phổ được tiền xử lý bằng đạo hàm bậc hai toàn dải phổ và làm mượt với thuật toán Savitzky-Golay [16], chuẩn hóa bằng cách lấy giá trị dữ liệu của một số đỉnh đặc trưng trừ đi giá trị trung bình toàn khoảng phổ và chia cho độ lệch chuẩn của các đỉnh đặc trưng đó.

***Bảng 1.*** *Kích thước ma trận dữ liệu phổ của dữ liệu ban đầu (None) và dữ liệu cân bằng theo từng tiêu chí phân loại (SMOTE)*

| Cân bằng dữ liệu | Độ chín | Giống xoài | Vùng trồng |
|---|---|---|---|
| None | | 11691 × 281 | |
| SMOTE | 15536×281 | 27680×281 | 13024×281 |

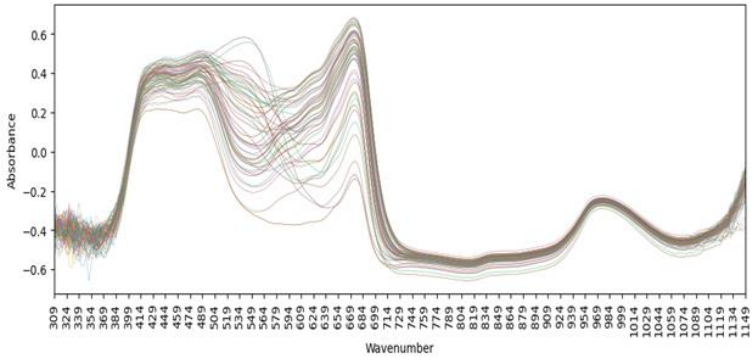
Xây dựng các thuật toán học máy để phân loại đối tượng sau khi đã tiến hành các phép giảm chiều dữ liệu (nhằm giảm số biến cần học cho mô hình) bằng PCA bao gồm PCA-SVC, PLS-DA, PCA-DT, PCA-RF, PCA-ANN và thuật toán học sâu 1D-CNN. Các tham số cần thiết của mô hình học máy được khảo sát và tối ưu bằng phương pháp đánh giá chéo sử dụng kỹ thuật GridSearchCV [17]. Độ chính xác của các mô hình được tính bằng tỷ lệ giữa số mẫu đoán đúng và số mẫu dự đoán.

Sau khi khảo sát có được các tham số tối ưu nhất của từng mô hình, tiến hành xây dựng mô hình có các tham số đã tối ưu, từ đó tiến hành cho mô hình huấn luyện trên toàn bộ dữ liệu huấn luyện rồi phân loại trên dữ liệu kiểm tra.

## 3. KẾT QUẢ VÀ BÀN LUẬN

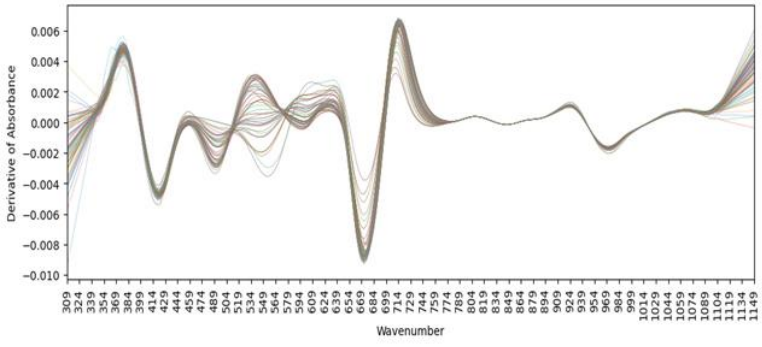
### 3.1 Tiền xử lý dữ liệu phổ

Dữ liệu đầu vào là tín hiệu Vis- NIR trong vùng từ 309 - 1149 nm phổ dạng bảng (kích thước 11691 mẫu × 281 cột giá trị Abs) từ file dữ liệu dạng Excel; các dữ liệu thông tin đầu ra là các biến rời rạc, bao gồm nguồn gốc địa lý (2 nhãn cho vùng trồng), độ chín (2 nhãn cho quả xanh, chín), giống (10 nhãn cho 10 giống) để xây dựng mô hình và kiểm tra độ chính xác của mô hình phân loại. Phổ Vis-NIR của 100 mẫu đầu tiên được hiển thị trong Hình 1.




***Hình 1.*** *Phổ Vis- NIR của 100 mẫu xoài đầu tiên trong tập dữ liệu*

Hiện tượng tín hiệu nhiễu xuất hiện ở các khoảng bước sóng 350 - 390 nm và 1065 - 1155 nm, các peak trong vùng 390 - 630 nm bị chồng chéo lên nhau, đồng thời do sử dụng thiết bị cầm tay để lấy tín hiệu của mẫu nên các ảnh hưởng của thiết bị cũng thể hiện trên phổ dữ liệu. Các yếu tố này chính là các nguy cơ dẫn đến việc gây khó khăn cho khả năng phân loại mô hình, khiến mô hình khó tối ưu và làm giảm tốc độ hội tụ của mô hình. Để khắc phục điều này, tiến hành đạo hàm bậc hai và làm mượt phổ sử dụng thuật toán Savitzky-Golay thu được phổ trên Hình 2.



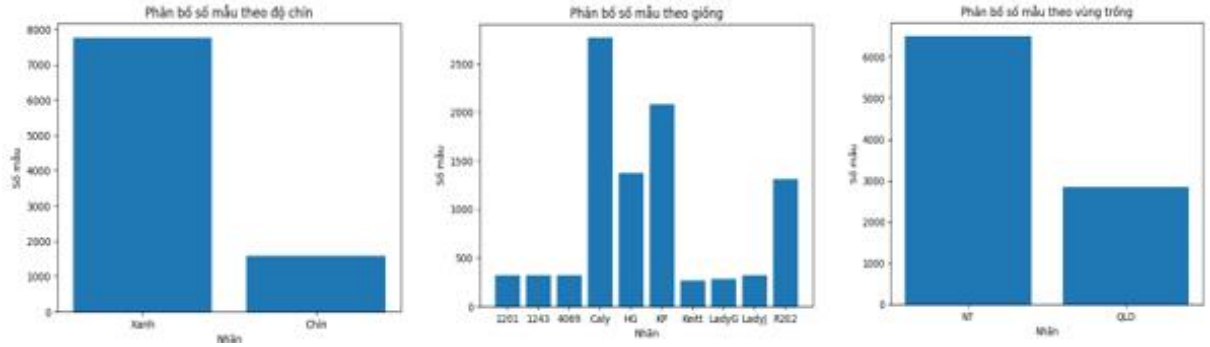

***Hình 2.*** *Phổ sau khi đạo hàm bậc 2 và làm mượt (smoothing) của 100 mẫu xoài đầu tiên trong tập dữ liệu*

Sau khi đạo hàm và làm mượt, các hiện tượng nhiễu phổ đã biến mất, đồng thời các đặc trưng của mẫu sẽ được thể hiện rõ hơn bằng đạo hàm bậc 2, điều này làm cho mô hình dễ nhận dạng các mẫu theo nhãn hơn. Đồng thời các ảnh hưởng của thiết bị và môi trường đã được loại bỏ để khiến việc phân loại của mô hình trở nên ổn định hơn. Sau đó tiếp tục chuẩn hóa dữ liệu phổ bằng StandardScaler để tối đa hóa hiệu suất của các mô hình. Trên các tập dữ liệu huấn luyện mô hình và kiểm tra mô hình với phổ đạo hàm bậc 2, sử dụng tín

hiệu toàn phổ, sử dụng câu lệnh *train_test_split* của thư viện *scikit-learn* với tham số *stratify=y* để phân bố các lớp trong tập kiểm tra sẽ gần giống với phân bố các lớp trong tập dữ liệu ban đầu, từ đó có thể đánh giá mô hình một cách khách quan nhất.

### 3.2. Phân bố số lượng mẫu xoài theo các tiêu chí dán nhãn

Sự phân bố số lượng mẫu xoài được thể hiện trên Hình 3 theo các chỉ tiêu phân loại của xoài theo độ chín (xoài xanh, xoài chín), theo giống (1201, 1243, 4069, Caly, HG, KP, Keitt, LadyG, Ladyj, R2E2) và theo vị trí vùng trồng (NT, QLD).




***Hình 3.*** *Phân bố mẫu đo theo lần lượt độ chín, giống xoài, vùng trồng*

Ở các chỉ tiêu phân loại về độ chín, giống, vùng trồng phân bố số lượng xoài theo từng chỉ tiêu xảy ra hiện tượng mất cân bằng một cách trầm trọng, nếu áp dụng các mô hình phân loại được thiết kế hiện hành có thể xảy ra sự thiên vị phân loại đối với các nhãn có số lượng mẫu lớn hơn. Vì vậy việc làm cân bằng dữ liệu (sử dụng thuật toán SMOTE) là cần thiết để khiến cho số lượng mẫu ở các nhãn phân loại trở nên đồng đều nhau.

### 3.3. Kết quả phân loại xoài

Các kết quả phân loại đánh giá trên các tiêu chí về độ chính xác thông qua phần trăm các mẫu phân loại đúng trong tập số liệu kiểm tra được trình bày ở Bảng 2.

***Bảng 2****. Kết quả phân loại của các thuật toán học máy và mô hình học sâu 1D-CNN ở trường hợp sử dụng dữ liệu ban đầu (None) và cân bằng dữ liệu SMOTE*

| Tiêu chí phân loại | Cân bằng dữ liệu | PCA - SVC | PLS- DA | PCA- DT | PCA -RF | PCA- ANN | 1D- CNN |
|---|---|---|---|---|---|---|---|
| Độ chín | None | 97,90% | 98,11% | 95,68% | 97,30% | 98,60% | 99,02% |
| | SMOTE | 98,18% | 98,29% | 96,32% | 98,10% | 98,50% | 99,40% |
| Giống | None | 73,92% | 3,42% | 69,4% | 85,12% | 81,74% | 92,00% |
| | SMOTE | 73,94% | 3,84% | 71,01% | 85,10% | 82,3% | 94,35% |
| Vùng trồng | None | 87,60% | 87,17% | 82,60% | 90,17% | 91,10% | 96,37% |
| | SMOTE | 87,08% | 85,53% | 83,53% | 91,36% | 89,52% | 96,92% |

Khi phân loại giữa xoài chín và xoài xanh, các mô hình học máy và học sâu đều đạt độ chính xác cao, đặc biệt với mô hình 1D-CNN độ chính xác cao hơn 99% gần như tuyệt đối, từ đó chứng tỏ rằng việc phân loại nhị phân độ chín của xoài có thể đạt được trực tiếp bằng các thuật toán học máy và học sâu mà không cần thông qua việc dự đoán hàm lượng chất khô như trong bài báo gốc.

Với các tiêu chí phân loại theo giống, khi bài toán trở thành phân loại đa lớp thì các mô hình học máy thể hiện sự khó khăn khi giá trị độ chính xác đều giảm xuống dưới 90%, đặc biệt mô hình PLS-DA có độ chính xác thấp đến bất ngờ khi chỉ đạt tới 3,42% và 3,84%, điều này có thể giải thích do khi tăng số lượng lớp phân loại lên sẽ làm cho các điểm dữ liệu trong không gian nằm ở các vị trí tương đối phức tạp, đồng nghĩa với việc mô hình PLS-DA gặp càng nhiều khó khăn hơn khi tìm ra siêu phẳng tối ưu. Các mô hình đáng tin cậy nhất vẫn là mô hình PCA-RandomForest với độ chính xác cao nhất với giá trị là 85,10% và 85,12%, theo sau đó là mô hình PCA-ANN với độ chính xác lần lượt là 81,74% và 82,30%. Sự có mặt của thuật toán SMOTE nhìn chung vẫn giúp cho các mô hình học máy tăng độ chính xác, từ đó có thể khẳng định rằng kỹ thuật SMOTE vẫn có ích trong trường hợp này. Mô hình 1D-CNN có độ chính xác cao nhất 94,35% và 92,00% chứng minh sự phù hợp khi sử dụng mô hình có cấu trúc phức tạp cho bài toán phân loại đa lớp.

Đối với trường hợp phân loại dựa theo tiêu chí vùng trồng, việc áp dụng kỹ thuật SMOTE so với khi không áp dụng cho ra hai kết quả tương đối giống nhau. Thuật toán PLS -DA và PCA -ANN sau khi được sử dụng dữ liệu được làm cân bằng với SMOTE lại cho ra kết quả độ chính xác giảm. Điều đó có thể do trong quá trình làm cân bằng dữ liệu, thuật toán đã vô tình giả định một số mẫu mới bị nhầm lẫn vào phần dữ liệu của nhãn khác và trở thành dữ liệu nhiễu, dẫn tới mô hình học sai. Nhìn chung, ba thuật toán PCA-RF, PCA-ANN và 1D-CNN cho độ chính xác tốt nhất, với thuật toán 1D-CNN có độ chính xác là hơn 96% cho thấy sự phù hợp của mô hình trong phân loại dựa trên tiêu chí vùng trồng.

Trong hầu hết các trường hợp phân loại, việc sử dụng kỹ thuật tiền xử lý dữ liệu SMOTE với các mô hình học máy và học sâu có độ chính xác cao hơn so với việc không sử dụng kỹ thuật này. Trong các thuật toán học máy, với mô hình có cấu trúc phức tạp như PCA -RF và PCA -ANN thể hiện khả năng phân loại tốt hơn so với các thuật toán học máy đơn giản. Mô hình 1D-CNN với cấu trúc phức tạp có các lớp tích chập, lớp kéo và các lớp kết nối đầy đủ cho kết quả phân loại tốt, độ ổn định cao ở các tiêu chí phân loại. Độ chính xác của mô hình 1D-CNN ở các tiêu chí phân loại đều lớn hơn 92%, thể hiện tính ứng dụng cao trong thực tế khi mô hình có thể nhìn được toàn bộ phổ dữ liệu mà không cần tiến hành giảm chiều, từ đó tăng tính tự động hóa và dễ dàng tích hợp vào các thiết bị đo cầm tay phục vụ phân tích hiện trường.

### 3.4. Bàn luận

Nghiên cứu đã chỉ ra sự tiến bộ đáng kể trong việc nhận dạng và phân loại xoài dựa trên dữ liệu quang phổ kết hợp với các mô hình học máy và học sâu. Các kết quả cho thấy rằng việc sử dụng dữ liệu quang phổ vùng Vis-NIR kết hợp với các phương pháp phân tích dữ liệu tiên tiến có thể mang lại hiệu quả cao trong việc phân loại xoài dựa trên các yếu tố như độ chín, giống, vùng trồng. So với các công trình nghiên cứu được đề cập [4, 5, 9],

nghiên cứu này đã xây dựng mô hình đơn giản và hiệu quả hơn, không cần tiến hành khảo sát khoảng phổ tối ưu mà sử dụng toàn bộ dải phổ đo được mà vẫn đáp ứng được mục đích phân loại. Đồng thời, kết quả phân loại đạt được (đặc biệt là các bài toán phân loại đa lớp) sử dụng mô hình học sâu 1D-CNN đều có độ chính xác trên 95%, trong nghiên cứu [4] độ chính xác chỉ đạt 91,49%, nghiên cứu [5] độ chính xác rất cao tuy nhiên bắt buộc phải phân loại theo cặp ở các khoảng bước sóng khác nhau và nghiên cứu [9] chỉ dự đoán hàm lượng chất khô DMC với $R^2$ đạt 0,89. Như vậy, việc sử dụng mô hình học sâu kết hợp với đạo hàm bậc 2 và làm mượt, đồng thời làm giàu dữ liệu sử dụng thuật toán SMOTE khiến cho có thể loại bỏ đi bước giảm chiều dữ liệu, từ đó mở ra tiềm năng nghiên cứu đa dạng và linh hoạt với dữ liệu phổ nhiều chiều.

Mặc dù nghiên cứu đã đạt được kết quả đáng khích lệ, vẫn còn một số hạn chế cần được xem xét. Đầu tiên, bộ dữ liệu sử dụng trong nghiên cứu là khá lớn và đa dạng, nhưng việc mở rộng bộ dữ liệu để bao gồm nhiều giống xoài và nhiều vùng địa lý hơn và các điều kiện môi trường khác nhau có thể giúp làm tăng độ tin cậy và khả năng tổng quát của các mô hình. Thứ hai, việc tối ưu hóa cấu trúc và tham số của mô hình học sâu 1D-CNN có thể mang lại những cải thiện thêm về hiệu suất. Cuối cùng, nghiên cứu có thể mở rộng ứng dụng của các kỹ thuật phân tích này cho các loại trái cây khác và các sản phẩm nông nghiệp khác để đánh giá hiệu quả và khả năng ứng dụng rộng rãi hơn.

## 4. KẾT LUẬN

So với các thuật toán học máy thông thường có kết hợp với giảm chiều dữ liệu, mô hình 1D-CNN đã chứng tỏ là mô hình có cấu trúc tối ưu nhất để phân loại xoài dựa trên dữ liệu phổ Vis-NIR với độ chính xác cao trên 95%. Điều đó chứng tỏ khả năng ứng dụng cao của mô hình 1D-CNN cho việc phân loại các đặc điểm của xoài, cung cấp thông tin một cách tổng quát về việc ứng dụng dữ liệu phổ toàn phần và các phương pháp học máy, học sâu trong phân loại xoài, và mở ra hướng đi mới cho việc cải thiện quy trình thu hoạch và quản lý chất lượng sản phẩm nông nghiệp.

## TÀI LIỆU THAM KHẢO

[1]. K. A. Shah, M. B. Patel, R.J. Patel, P.K. Parmar, "Mangifera indica (mango)," *Phar. Rev. Jan*, 4(7), pp. 42-48, 2010.

[2]. M. E. Maldonado-Celis, E. M. Yahia, R. Bedoya, et al., "Chemical composition of mango (Mangifera indica L.) fruit: Nutritional and phytochemical compounds," *Front Plant Sci.*, 10:450160, 2019.

[3]. V. Bennett, A.K.Inengite, "Comparative Study of the Chemical Composition of Three Mango Stem Bark," *Journal of Diseases and Medicinal Plants*, 8(3), pp. 55-60, 2022.

[4]. A. K. Kouassi, T. Alabi, G. Purcaro, C. Blecker, S. Danthine, "Assessment of the Impact of Annual Growing Conditions on the Physicochemical Properties of Mango Kernel Fat," *Horticulturae*, 10:814, 2024.

[5]. Tran Van Hau, Nguyen Chi Linh, Nguyen Long, "Determining the harvest time of Hoa Loc mango (mangifera indica l.) in Hoa Hung commune, Cai Be district, Tien Giang province," *CTU Journal of Innovation and Sustainable Development*, 37(2), pp.111-119, 2015.

[6]. M. D. K. Vithana, Z. Singh, & S. K. Johnson, "Cold storage temperatures and durations affect the concentrations of lupeol, mangiferin, phenolic acids and other health-promoting compounds in the pulp and peel of ripe mango fruit," *Postharvest Biology and Technology*, 139, pp. 91-98, 2018.

[7]. G. Gizachew, G. Gezahegn, & F. Seifu, "Chemical Composition of Mango (Mangifera Indica L) Fruit as Influence by Postharvest Treatments in Arba Minch, Southern Ethiopia," IOSR Journal of Environmental Science Toxicology and Food Technology, 10(11), pp.70-77, 2016.

[8]. Nguyen Truong Thịnh, Nguyen Duc Thong, Huynh Thanh Cong, Nguyen Tran Thanh Phong, "Mango classification system based on machine vision and artificial intelligence," 7th International Conference on Control, Mechatronics and Automation, pp. 475-482, 2019.

[9]. D. G. A. Al-Sanabani, M. I. Solihin, L. P. Pui, W. Astuti, C.K. Ang and L. W. Hong, "Development of non-destructive mango assessment using Handheld Spectroscopy and Machine Learning Regression," Journal of Physics: Conference Series, pp. 12- 30, 2019.

[10]. S. N. Jha, P. Jaiswal, K. Narsaiah, et al., "Authentication of mango varieties using near-infrared spectroscopy," *Agricultural Research*, vol. 2, pp. 229-235, 2013.

[11]. I. W. Budiastra, & H. K. Punvadaria, "Classification of mango by artificial neural network based on near infrared diffuse reflectance," *IFAC Proceedings*, 33(29), pp. 157-161, 2000.

[12]. R. Pronprasit, & J. Natwichai, "Prediction of mango fruit quality from Nir spectroscopy using an ensemble classification," *International Journal of Computer Applications*, 83(14), 2013.

[13]. N. T. Anderson, K. B. Walsh, P. P. Subedi and C. H. Hayes, "Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content," *Postharvest Biology and Technology*, 168, 2020.

[14]. A. Fernández, S. Garcia, F. Herrera and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, 61, pp. 863-905, 2018.

[15]. R. W. Schafer, "What is a Savitzky-Golay filter?" *IEEE Signal processing magazine*, 28(4), pp. 111-117, 2011.

[16]. I. Syarif, A. Prugel-Bennett, G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telecommunication Computing Electronics and Control*, 14(4), 2016.