Tạp chí Khoa học Nông nghiệp Việt Nam 2026, 24(6): 809-820 https://doi.org/10.31817/tckhnnvn.2026.24.06.08
vie.vjas.vn

PHÂN LOẠI MẪU GẠO LỨT BẰNG PHƯƠNG PHÁP HỌC MÁY

Lương Minh Quân

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Tác giả liên hệ: lmquan@vnua.edu.vn

Ngày nhận bài: 26.09.2025 Ngày chấp nhận đăng: 05.05.2026

TÓM TẮT

Gạo lứt là một loại ngũ cốc nguyên hạt có nhiều chất xơ, chất chống oxy hóa, nhiều vitamin và khoáng chất quan trọng. Sản phẩm này phù hợp với những người ăn kiêng hay người có bệnh lý như tiểu đường tuýp 2, cao huyết áp, béo phì, có nguy cơ đột quỵ hay cholesterol cao. Nghiên cứu này nhằm mục đích phân loại hạt gạo lứt Đen và lứt Huyết Rồng dựa trên các đặc trưng hình thái học. Ảnh chụp các mẫu hạt gạo lứt được tiền xử lý để tách tời từng hạt riêng biệt kèm theo ảnh mặt nạ tương ứng bằng mô hình SAM. Mỗi hạt này tiếp tục được phân tích để trích xuất các thông tin quan trọng về hình thái học liên quan tới cấu trúc hình học bao gồm 19 đặc trưng độc lập. Các đặc trưng này được phân tích bằng các mô hình học máy: cây quyết định (DT), rừng ngẫu nhiên (RF) và máy vector hỗ trợ (SVM). Kết quả cho thấy cả ba mô hình học máy đều cho kết quả với độ chính xác cao, đặc biệt là SVM, với độ chính xác đạt 76%. Mô hình SVM cũng vượt trội hơn các mô hình DT và RF với các chỉ số Recall, F1-Score đạt lần lượt là 83% và 80,2% cho mẫu gạo lứt Đen. Tuy nhiên, các chỉ số này cho mẫu Huyết Rồng chỉ đạt dưới 70% do tính chất chồng lấn của các thuộc tính trong không gian đặc trưng.

Từ khóa: gạo lứt Đen, gạo Huyết Rồng, học máy.

Classification of Brown Rice Samples using Machine Learning Method

ABSTRACT

Brown Rice is whole grain high in fiber, antioxidants, and many important vitamins and minerals. This product is suitable for people who on a diet or those with medical conditions such as diabetes type 2, high blood pressure, obesity, risk of heart stroke or high cholesterol. This study aimed to classify black brown and red brown rice based on the morphological characteristics. Images of brown rice samples were pre-processed to separate each separately, accompanied by corresponding image mask using the SAM. Each of these seeds was further analyzed to extract important morphological information related to the geometric structure including 19 independent features. These features were analyzed using machine learning models: decision tree (DT), random forest (RF) and support véc tơ machine (SVM). The results show that all three algorithms resulted highly accurate results, especially SVM, with an accuracy of 76%. The SVM model also outperformed the DT and RF models with Recall and F1-Score indexes reaching 83% and 80.2%, respectively, for the black brown rice sample. However, these indexes for the red brown sample were only below 70% due to the overlapping nature of the attributes in the feature space.

Keywords: Black brown rice, red brown rice, machine learning.

1. ĐẶT VẤN ĐỀ

Phương pháp phân tích dựa trên đặc trưng hình thái có nhiều ưu điểm vượt trội so với các phương pháp phân tích phổ cận hồng ngoại (NIR) và phân tích hóa - lý đòi hỏi phải có thiết bị chuyên dụng, chi phí cao và nguồn nhân lực có chuyên môn. Tuy nhiên, phương pháp này vẫn gặp phải thách thức lớn khi các mẫu phân tích có các đặc trưng hình thái tương tự nhau đặc biệt là đối với mẫu gạo lứt Đen và Huyết Rồng trong nghiên cứu.

Hiện nay, có nhiều tiêu chí để đánh giá chất lượng của hạt gạo như: cảm quan, mầu sắc; các đặc trưng sau khi chế biến như: mùi thơm, hương vị, độ dẻo,… Từ góc nhìn của người tiêu

dùng, hình thức và tiêu chí liên quan tới dinh dưỡng là những yếu tố quan trọng trong việc lựa chọn sản phẩm trên thị trường. Vũ Mạnh Ẩn & cs. (2023) sử dụng các tiêu chí về chiều dài hạt, chiều rộng hạt, tỷ lệ dài/rộng và hàm lượng protein để đánh giá chất lượng của 171 giống lúa địa phương nhằm tuyển chọn giống tiềm năng cho sản xuất và công tác tạo giống lúa chất lượng. Theo Nguyễn Thị Lang & cs. (2021), nghiên cứu đánh giá chất lượng của giống lúa mùa AG3 tại An Giang sử dụng phương pháp phản ứng gạo với KOH và phương pháp PCR với hai chỉ thị RM223 và FMU1-2 để đánh giá mùi thơm. Dùng máy đo Baker E-02 của Nhật đo hình dạng, kích thước hạt để phân tích phẩm chất gạo. Kết quả đã chọn ra dòng lúa có mùi thơm tốt nhất và tính đồng nhất cao của hình dạng hạt gạo.

Trong nghiên cứu phục chế giống lúa thơm VD20 phục vụ nhu cầu xuất khẩu tại đồng bằng sông Cửu Long, Trần Thị Thanh Thúy & cs. (2021), đã tiến hành phân tích các tiêu chí về ngoại hình hạt như kích thước hạt gạo, tỷ lệ dài/rộng của hạt gạo, tỷ lệ gạo lứt, tỷ lệ gạo nguyên, độ bạc bụng kết hợp các tiêu chí về lượng khoáng chất: hàm lượng amylose và protein, độ bền thể gel, nhiệt độ trở hồ và mùi thơm theo phương pháp của IRRI (1996).

Ngày nay, sự đa dạng về chủng loại và sản lượng sản xuất ngày càng lớn, nhu cầu truy xuất thông tin nguồn gốc sản phẩm ngày càng trở nên quan trọng, đòi hỏi nhu cầu áp dụng các phương pháp nhận dạng để nhận biết và phân biệt chủng loại cũng như chất lượng của các sản phẩm lúa gạo chủ lực đặc biệt là dòng sản phẩm xuất khẩu. Tuy nhiên, nguồn cơ sở dữ liệu mô tả đặc điểm hình thái của hạt gạo nói chung và gạo lứt nói riêng đáp ứng yêu cầu về học máy còn rất hạn chế. Nghiên cứu của Cinar & Koklu (2019) đã ứng dụng các phương pháp học máy và mạng nơ ron nhân tạo để phân loại hạt gạo Osmancik và Kameo, những sản phẩm lúa gạo có nguồn gốc ở Thổ Nhĩ Kỳ. Phân tích học máy dựa trên các thuộc tính về hình dạng của hạt gạo bao gồm: diện tích, chu vi, bán trục nhỏ, bán trục lớn, độ phủ, tỷ lệ dài/ rộng,… Kết quả phân tích với các mô hình học máy cho kết quả phân lớp tới 93% với mô hình hồi quy tuyến tính.

Mục tiêu trọng tâm của nghiên cứu này là xác lập cơ sở khoa học cho việc phân loại các mẫu gạo lứt thông qua các chỉ số hình thái đặc trưng, hướng tới việc tự động hóa quy trình kiểm định và quản lý chất lượng nông sản.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Vật liệu

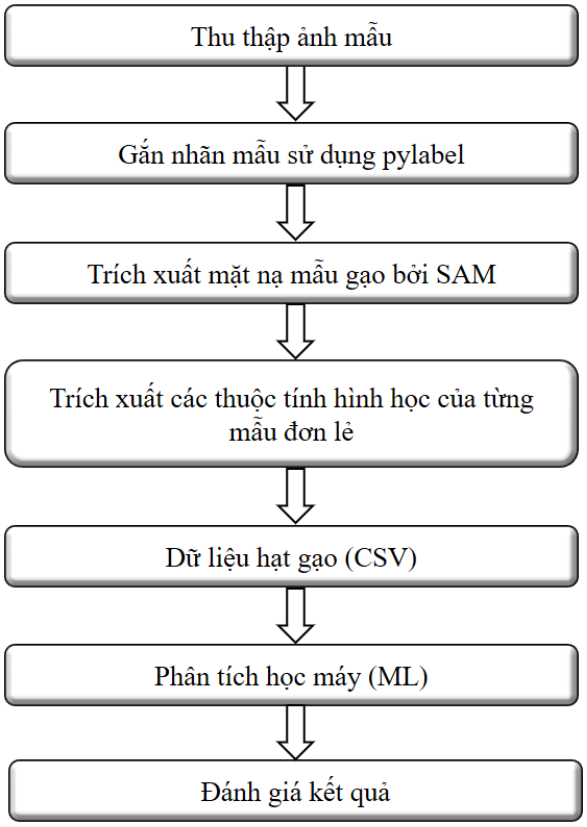
Nghiên cứu này kết hợp kỹ thuật xử lý hình ảnh kỹ thuật số và các thuật toán học máy bao gồm: cây quyết định (DT), rừng ngẫu nhiên (RF) và máy vector hỗ trợ (SVM) để thiết lập hệ thống phân loại tự động cho các mẫu gạo lứt dựa trên đặc trưng hình thái.

Trong quá trình thu thập mẫu gạo lứt, nhóm nghiên cứu nhận thấy ngoại trừ mầu sắc thì hai loại gạo lứt Đen và lứt Huyết Rồng được sản xuất bởi Tập đoàn Giống cây trồng Việt Nam Vinaseed có cấu trúc hình thái tương đồng nhau, điều này thúc đẩy nhóm nghiên cứu tìm ra phương pháp phù hợp để nhận diện các mẫu hạt này dựa trên các đặc trưng hình thái.

Bảng 1 giới thiệu các mẫu hạt được lựa chọn thỏa mãn điều kiện nguyên vẹn về cấu trúc hình học, không bị gãy hay biến dạng bởi các tác nhân vật lý. Quy trình thu thập ảnh mẫu và tiền xử lý làm tiền đề cho việc áp dụng các phương pháp học máy để phân loại hạt gạo dựa trên các đặc tính hình học được thực hiện theo quy trình được mô tả bởi biểu đồ trên hình 1.

Bảng 1. Thống kê các mẫu gạo lứt trong nghiên cứu

Tên mẫu	Khối lượng	Số lượng hạt mẫu thu thập	Nơi cung cấp mẫu
Gạo lứt Đen	37 gram	1.602	VINASEED
Gạo lứt Huyết Rồng	35 gram	1.161	VINASEED
Tổng số mẫu		2.763	



Hình 1. Quy trình thu thập và tiền xử lý ảnh mẫu gạo

2.2. Phương pháp nghiên cứu

2.2.1. Thu thập mẫu

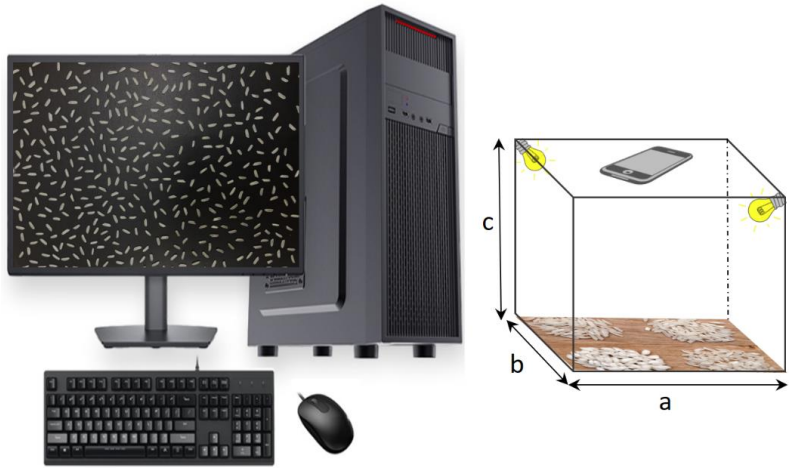
Nhóm nghiên cứu đã thiết kế mô hình thu thập chụp ảnh được mô tả trong hình 2. Một buồng mẫu hình hộp chữ nhật có đáy là nơi đặt mẫu với kích thước các cạnh lần lượt là a × b × c = 20 × 25 × 20cm. Buồng mẫu được chiếu sáng bởi hệ thống bóng đèn LED có ánh sáng trắng để đáp ứng yêu cầu về cường độ sáng cho mẫu ảnh cần phân tích, vị trí đèn được chiều chỉnh sao cho giảm thiểu bóng của mẫu hạt trên nền ảnh. Các mẫu hạt được sắp xếp một cách cẩn thận sao cho không có sự xếp chồng lên nhau giữa các hạt, ảnh mẫu được chụp bằng điện thoại Iphone8-plus với dung lượng trung bình là 5MB cho mỗi bức ảnh.

2.2.2. Gắn nhãn mẫu hạt gạo sử dụng pylabel

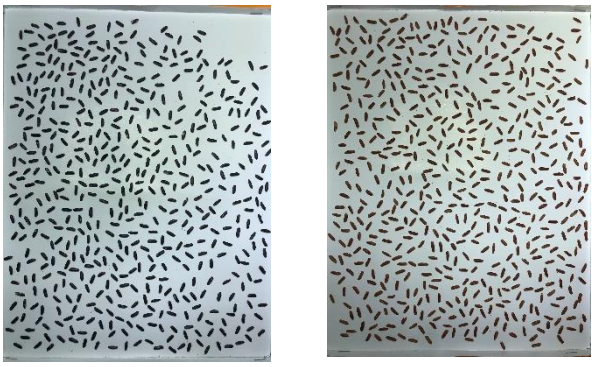
Jeremy & cs. (2021) đã phát triển công cụ pylabel để gắn nhãn các đối tượng trong ảnh chụp. Đây là một gói chương trình ứng dụng được phát triển trên nền tảng ngôn ngữ lập trình Python. Quá trình gắn nhãn cho từng mẫu hạt gạo được thực hiện trên nền tảng Web. Công cụ này cho phép người dùng gắn nhãn một cách linh hoạt, tiện lợi và dễ dàng ngay cả khi trong ảnh mẫu có nhiều loại nhãn cần phân tích. Dữ liệu thu thập được bao gồm nhãn, tọa độ hộp giới hạn (vị trí, chiều cao, chiều rộng) và được chuyển tới bước tiền xử lý tiếp theo với mô hình phân mảnh.

2.2.3. Mạng học sâu phân mảnh mọi thứ (Segment Anything Model - SAM)

Hộp giới hạn bao quanh mẫu hạt không đủ thông tin để phản ánh đặc trưng hình thái học của mẫu đó vì nó phụ thuộc vào định hướng ngẫu trong không gian của mẫu. Với mẫu hạt có bán trục lớn nằm theo phương ngang, diện tích hộp giới hạn là nhỏ nhất; với mẫu hạt nằm theo phương tạo một góc 45° so với phương ngang, diện tích hộp giới hạn là lớn nhất. Do đó, việc phân mảng để xác định diện tích mà mẫu hạt chiếm chỗ trong hộp giới hạn đó là cần thiết.



Hình 2. Thiết kế hệ thống chụp ảnh mẫu hạt gạo



a. Gạo lứt Đen b. Gạo lứt Huyết Rồng

Hình 3. Mẫu gạo lứt

Trong nghiên cứu này, chúng tôi sử dụng mô hình học sâu có giám sát được đề xuất bởi Alexander & cs. (2023) để phân vùng nhằm tách các mẫu hạt gạo lứt trong ảnh chụp dựa trên thông tin gắn nhãn thu được với pylabel để trích xuất mặt nạ của các mẫu hạt này. Mặt nạ thu được từ kết quả phân mảnh sẽ được sử dụng để trích xuất các dữ liệu quan trọng về đặc trưng hình thái như: diện tích, chu vi, bán trục lớn, bán trục nhỏ, độ phủ, các tính chất bất biến với các phép biến đổi không gian. Các đặc trưng này là dữ liệu đầu vào của các thuật toán nhận dạng và phân loại bằng các mô hình máy học. Với hai mẫu gạo lứt, nhóm nghiên cứu đã thu được bộ dữ liệu bao gồm: (1) ảnh mặt nạ của các mẫu hạt (điểm sáng có giá trị pixel bằng 255) trên nền tối (có giá trị pixel bằng 0) và (2) ảnh các hạt mẫu được phân tách một cách độc lập, được sử dụng để phân tích các thuộc tính về mặt cấu trúc hình học, phân bố khối lượng.

2.2.4. Trích xuất các đặc trưng hình học của mẫu hạt với opencv-python

Opencv-python là một thư viện thị giác máy tính mã nguồn mở được sử dụng rộng rãi trong các bài toán xử lý ảnh. Các đặc trưng quan trọng của mặt nạ mô tả hình dạng của mẫu có thể được trích xuất từ các hàm chức năng của opencv-python: hàm mô ment (cv2.moments) cho các thông tin về khối lượng, khối tâm, diện tích, chu vi của hình khối cần phân tích; các tính chất biến đổi về không gian với hàm cv2.HuMoments của từng mô ment tương ứng với mỗi mặt nạ sẽ bổ sung tính bất biến đối với: phép tịnh tiến, phép quay và phép co giãn trong không.

Với mỗi mặt nạ, các thuộc tính về cấu trúc hình học và tính bất biến với các phép biến đổi hình học được phân tích. Trong bảng 2 giới thiệu một số mẫu hạt đã được xử lý với các thông tin mô tả như đường biên phân tách mẫu với phông nền, hình chữ nhật thẳng đứng bao quanh mẫu, hình chữ nhật có diện tích nhỏ nhất giới hạn bởi mẫu, bán trục nhỏ và bán trục lớn.

Đặc trưng về mặt cấu trúc hình thái học có vai trò quan trọng cho bài toán phân loại bằng phương pháp học máy được mô tả trong bảng 3. Các thông số thống kê về mẫu được mô tả trong bảng 4a và 4b, trong đó, μ là giá trị trung bình, σ là độ lệch chuẩn.

2.2.5. Đánh giá hiệu suất mô hình phân lớp

Với bài toán phân lớp, sau khi tiến hành tiền xử lý dữ liệu và đưa vào mô hình hoc máy, đầu ra của mô hình là một vector xác suất tương ứng của từng lớp. Độ chính xác của mô hình được đánh giá thông qua chỉ số Accuracy: là phần trăm các lớp đã phân loại đúng trên tổng số dự đoán. Tuy nhiên, với chỉ số các lớp đã phân loại sai đặc biệt trong trường hợp dữ liệu ban đầu mất cân bằng giữa các lớp thì việc bổ xung các đại lượng nhằm mô tả cả những dự đoán không chính xác hoặc dự đoán nhầm lẫn giữa các lớp sẽ cung cấp nhiều thông tin có ý nghĩa hơn. Để giải quyết bài toán này ma trận phức hợp được sử dụng để đánh giá với tập dữ liệu kiểm tra.

Bảng 2. Mẫu hạt gạo thu được sau khi tiền xử lý và gắn nhãn

Mẫu gạo Huyết Rồng	
Mẫu gạo lứt Đen	

Bảng 3. Các đặc trưng hình học của mẫu hạt được trích xuất từ mỗi hạt gạo

Tên thuộc tính	Ý nghĩa
Segmented_area	Số lượng các điểm ảnh được giới hạn bởi đường biên bao quanh mẫu hạt
Perimeter	Độ dài đường biên bao quanh mẫu hạt
MajorAxisLength	Độ dài lớn nhất của trục đi qua khối tâm mẫu hạt giao cắt với đường biên của mẫu
MinorAxisLength	Độ dài nhỏ nhất của trục đi qua khối tâm mẫu hạt giao cắt với đường biên của mẫu, trục này vuông góc với bán trục lớn
Eccentricity	Giá trị thuộc [0,1], cho biết độ tròn của mẫu tương ứng với các bán trục. Giá trị càng nhỏ, mẫu hạt càng có hình dạng tròn.
HullArea	Số lượng các điểm ảnh chứa trong đa giác lồi, nhỏ nhất bao quanh mẫu
Solidity	Tỉ số giữa HullArea và Area
Extent_to_min_area	Tỉ số giữa Area và diện tích nhỏ nhất bao quanh mẫu hạt
Extent_to_bbox	Tỉ số giữa Area và diện tích hình chữ nhật bao quanh mẫu hạt
Angle_orientation	Góc lệch của mẫu hạt so với phương ngang
Skew	Độ xoắn, mô tả sự phân bố không đồng đều về cấu trúc hình học của mẫu hạt
Roundness	Mô tả độ tròn của mẫu khi so sánh mẫu hạt thu được với hình tròn tương ứng với cùng diện tích
Hu_Moments ($\varphi_1 \rightarrow \varphi_7$)	7 thuộc tính humentt tương ứng với các tính chất bất biến của mẫu với các phép biến đổi hình học cơ bản: phép tịnh tiến, phép quay và phép co giãn không gian
Classes	Thông tin gắn nhãn với từng mẫu hạt: lứt Đen và Huyết Rồng

Bảng 4a. Mô tả dữ liệu thống kê các thuộc tính bộ cơ sở dữ liệu gạo lứt Đen

Thuộc tính	Số mẫu	μ	σ	min	25%	50%	75%	max
Segmented_area	1.602	3.895,59	230,73	3.500,50	3.715,75	3.883,75	4.050,63	5.028,50
Periphery	1.602	274,12	9,52	249,30	267,68	273,79	280,52	305,20
HullArea	1.602	3972,71	233,23	3551,50	3790,13	3962,50	4128,63	5111,50
Solidity	1.602	0,98	0,00	0,95	0,98	0,98	0,98	0,99
Extent_to_bbox	1.602	0,63	0,10	0,45	0,54	0,60	0,72	0,87
Extent_to_min_area	1.602	0,85	0,02	0,79	0,84	0,86	0,87	0,91
Angle_orientation	1.602	88,25	53,10	0,04	40,34	88,26	133,73	179,81
Major_ax	1.602	116,78	5,49	100,73	113,07	116,92	120,34	138,95
Minor_ax	1.602	43,41	1,98	37,91	42,04	43,31	44,70	52,90
Ecentricity	1.602	0,93	0,01	0,89	0,92	0,93	0,94	0,95
Skew	1.602	0,01	0,69	-1,31	-0,55	-0,01	0,60	1,41
Roundness	1.602	1,54	0,06	1,36	1,50	1,54	1,58	1,77
Hu_Moment_1	1.602	1,42	0,05	1,26	1,39	1,42	1,46	1,57
Hu_Moment_2	1.602	3,43	0,18	2,91	3,31	3,42	3,55	4,05
Hu_Moment_3	1.602	9,90	1,12	7,68	9,16	9,72	10,41	16,15
Hu_Moment_4	1.602	12,00	1,09	9,16	11,29	11,87	12,55	19,64
Hu_Moment_5	1.602	-0,67	23,83	-37,59	-23,51	-20,78	23,33	35,13
Hu_Moment_6	1.602	-3,71	14,08	-21,40	-14,55	-13,51	13,73	20,91
Hu_Moment_7	1.602	0,71	23,71	-38,57	-23,07	20,04	23,40	34,52

Bảng 4b. Mô tả dữ liệu thống kê các thuộc tính bộ cơ sở dữ liệu gạo lứt Huyết Rồng

Thuộc tính	Số mẫu	μ	σ	min	25%	50%	75%	max
Segmented_area	1.161	3870,66	253,54	3500,00	3670,50	3831,50	4024,00	4945,00
Periphery	1.161	279,05	11,77	241,66	271,54	279,10	286,99	320,88
HullArea	1.161	3953,23	258,05	3552,00	3748,00	3921,50	4109,00	5037,00
Solidity	1.161	0,98	0,00	0,96	0,98	0,98	0,98	0,99
Extent_to_bbox	1.161	0,62	0,11	0,43	0,52	0,59	0,71	0,87
Extent_to_min_area	1.161	0,85	0,02	0,79	0,84	0,85	0,87	0,92
Angle_orientation	1.161	89,30	51,06	0,23	48,65	89,07	132,52	179,97
Major_ax	1.161	120,86	7,38	93,76	116,78	121,61	125,92	141,59
Minor_ax	1.161	41,82	2,54	35,44	40,01	41,59	43,35	52,18
Ecentricity	1.161	0,94	0,02	0,85	0,93	0,94	0,95	0,97
Skew	1.161	-0,02	0,76	-1,44	-0,68	-0,01	0,62	1,58
Roundness	1.161	1,60	0,09	1,28	1,55	1,61	1,66	1,93
Hu_Moment_1	1.161	1,37	0,07	1,16	1,32	1,36	1,41	1,65
Hu_Moment_2	1.161	3,25	0,26	2,62	3,08	3,22	3,37	4,47
Hu_Moment_3	1.161	9,81	1,20	7,27	8,95	9,62	10,47	18,18
Hu_Moment_4	1.161	11,76	1,19	8,90	10,95	11,61	12,42	20,24
Hu_Moment_5	1.161	3,01	23,16	-39,72	-22,46	20,22	23,27	31,94
Hu_Moment_6	1.161	-1,34	14,16	-20,97	-14,05	-12,61	13,67	21,85
Hu_Moment_7	1.161	-1,77	23,38	-39,88	-23,20	-20,19	22,55	37,33

Bảng 5. Ma trận nhầm lẫn với bài toán nhị phân

		Giá trị thực	
		Lứt Đen (0)	Huyết Rồng (1)
Dự đoán bởi mô hình	Lứt Đen (0)	TP	FP
	Huyết Rồng (1)	FN	TN

Bài toán phân loại hạt gạo là bài toán nhị phân: gạo lứt Đen tương ứng với phân lớp 0, gạo Huyết Rồng tương ứng với phân lớp 1.

Trong đó:

TP (True Positive): Số lượng dự đoán chính xác. Khi mô hình dự đoán đúng mẫu gạo đó thuộc phân lớp 0.

TN (True Negative): Số lương dự đoán chính xác một cách gián tiếp. Khi mô hình dự đoán đúng mẫu gạo đó thuộc phân lớp 1.

FP (False Positive): Số lượng các dự đoán sai lệch. Khi mô hình dự đoán mẫu hạt đó thuộc phân lớp 0 nhưng thực tế mẫu này thuộc về phân lớp 1.

FN (False Negative): Số lượng các dự đoán sai lệch một cách gián tiếp. Khi mô hình dự đoán mẫu hạt gạo thuộc phân lớp 1 nhưng thực tế mẫu hạt đó thuộc phân lớp 0.

Từ bảng ma trận phức hợp trên, các đại lượng sau đây có thể được sử dụng để đánh giá hiệu suất của mô hình .

Accuracy: Số lượng mẫu được dự đoán chính xác trong tất cả các mẫu cần phân tích

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (1)$$

Precision: Trong tất cả các dự đoán mẫu hạt thuộc phân lớp 1 được đưa ra, thì bao nhiêu dự đoán là chính xác:

$$\text{Precison} = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

Recall: Trong tất cả các mẫu thuộc phân lớp 1, thì bao nhiêu mẫu được dự đoán chính xác:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

F1-Score: Đánh giá độ tin cậy chung của mô hình bằng cách kết hợp 2 chỉ số Precision và Recall thành một chỉ số duy nhất.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \times 100\% \quad (4)$$

2.2.6. Đánh giá chéo

Khi xây dựng các mô hình học máy, việc tối ưu hóa các tham số của mô hình trong quá trình huấn luyện với dữ liệu đào tạo thường dẫn tới hiệu suất mô hình rất cao với dữ liệu huấn luyện nhưng lại không tốt đối với bộ dữ liệu kiểm tra. Để khắc phục vấn đề trên, có thể sử dụng phương pháp đánh giá chéo nhiều lớp được đề xuất bởi Berrar (2018).

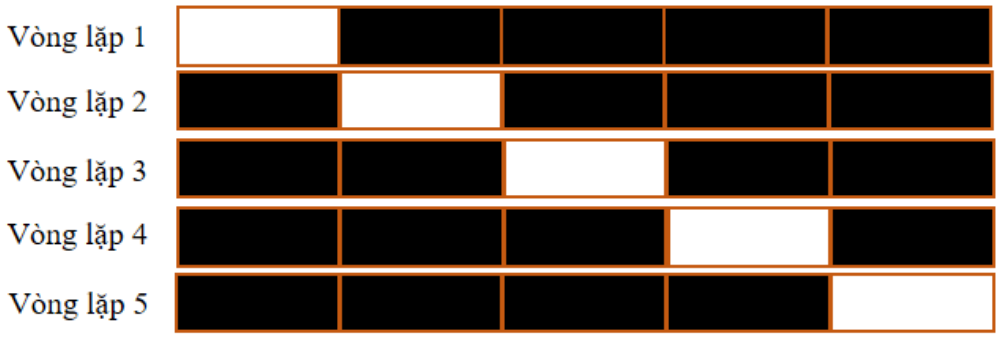
Hình 4 cho thấy bộ dữ liệu huấn luyện được chia thành 5 phần bằng nhau, ở mỗi vòng lặp 80% dữ liệu được sử dụng cho việc huấn luyện (ô màu đen), 20% dữ liệu còn lại được dùng để đánh giá kết quả (ô mầu trắng). Dữ liệu đánh giá ở mỗi vòng lặp độc lập nhau. Hiệu suất chung của mô hình được lấy trung bình theo số vòng lặp.

2.2.7. Lựa chọn biến đặc trưng dựa trên ma trận tương quan

Ma trận tương quan được đưa ra bởi Benesty & cs. (2009), là một công cụ thống kê mạnh mẽ cho phép trực quan hóa và định lượng mối quan hệ tuyến tính giữa tất cả các cặp biến trong tập dữ liệu.

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \quad (5)$$

Trong đó: r_{xv} là hệ số tương qua giữa biến x và y; n là số lượng mẫu; x_i, y_i là giá trị tương ứng của các biên; $\bar{x}, \bar{y}$ giá trị trung bình của các biến.



Hình 4. Mô hình đánh giá chéo 5 lớp với bộ dữ liệu huấn luyện

Hệ số tương quan có giá trị nằm trong khoảng [-1, 1]. Nếu $r_{xv} > 0$ mối tương quan là đồng biến, ngược lại với $r_{xv} < 0$ mối tương quan là nghịch biến. Khi $r_{xv} \sim 0$, thì các biến không có mối tương quan tuyến tính với nhau.

3. CÁC MÔ HÌNH HỌC MÁY

Hiện tại có rất nhiều mô hình học máy có thể được lựa chọn để đánh giá hiệu quả của bài toán phân lớp. Tuy nhiên, với bài toán này nhóm tác giả sử dụng các mô hình phân lớp dựa trên cơ sở của lý thuyết tập hợp. Từ đây có thể so sánh kết quả của phân lớp của một cây quyết định (DT) và một tập hợp cây quyết định theo các thuật toán khác nhau như như rừng ngẫu nhiên (RF). Ngoài ra SVM cũng là một trong số các thuật toán máy học được sử dụng hiệu quả trong bài toán phân lớp dựa trên việc tìm siêu phẳng trong không gian đa chiều.

3.1. Mô hình cây quyết định (DT)

Cây quyết định là một mô hình học dựa trên tri thức được đề xuất bởi Quinlan (1985) và được sử dụng rộng rãi trong các lĩnh vực khác nhau như máy học, khai phá dữ liệu và thống kê. Mô hình này cho phép biểu diễn một cách rõ ràng và trực quan quá trình đưa ra quyết định dựa trên dữ liệu bằng cách mô hình hóa mối liên hệ giữa các thuộc tính khác nhau. Về mặt cấu trúc, mô hình cây quyết định bao gồm các nút đại diện cho các quyết định hoặc kiểm tra các thuộc tính, các nhánh đại diện cho kết quả của các quyết định này và các nút lá đại diện cho kết quả hoặc dự đoán cuối cùng.

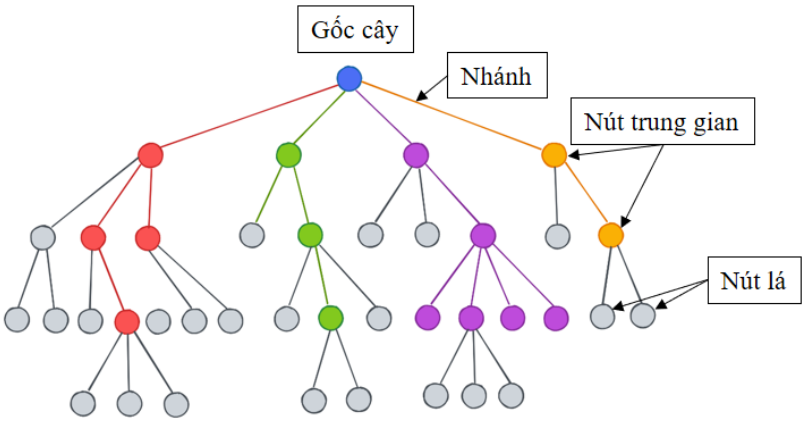
3.2. Thuật toán phân lớp máy véc tơ hỗ trợ (SVM)

Cortes & cs. (1995) đã đề xuất thuật toán mạng lưới vector hỗ trợ để giải quyết bài toán phân lớp nhị phân. Ngày nay, thuật toán này được sử dụng trong cả bài toán phân lớp lẫn hồi quy với mục đích tìm ra một siêu phẳng trong không gian N chiều chia dữ liệu thành lớp tương ứng. Trên hình 6 biểu diễn phân lớp đối với hai nhóm dữ liệu tuyến tính trong không gian hai chiều. Siêu phẳng tối ưu tìm được có lề lớn nhất tức là khoảng cách tới các điểm gần nhất của hai lớp là lớn nhất.

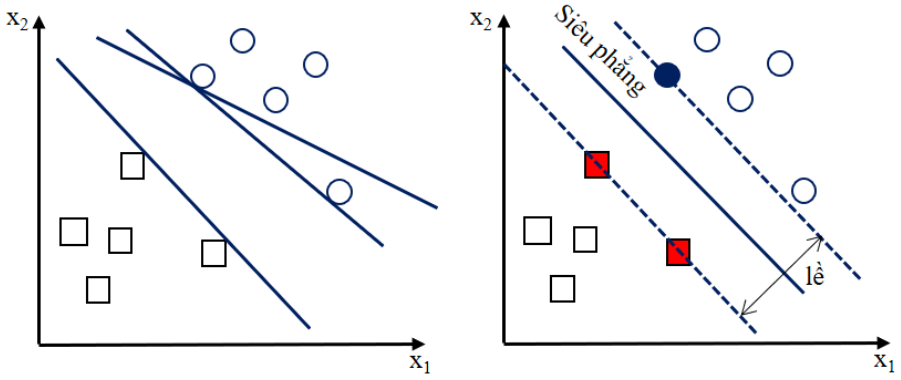
Các điểm dữ liệu nằm trên đường đứt đoạn (----) của đồ thị được gọi là các vector hỗ trợ, chúng cách đều siêu phẳng và ảnh hưởng tới hướng và vị trí của siêu phẳng.

Với dữ liệu phi tuyến tính, siêu phẳng không còn là đường thẳng trong không gian hai chiều nữa. Để giải quyết bài toán này, nhóm nghiên cứu tiếp cận theo hai phương pháp sau:

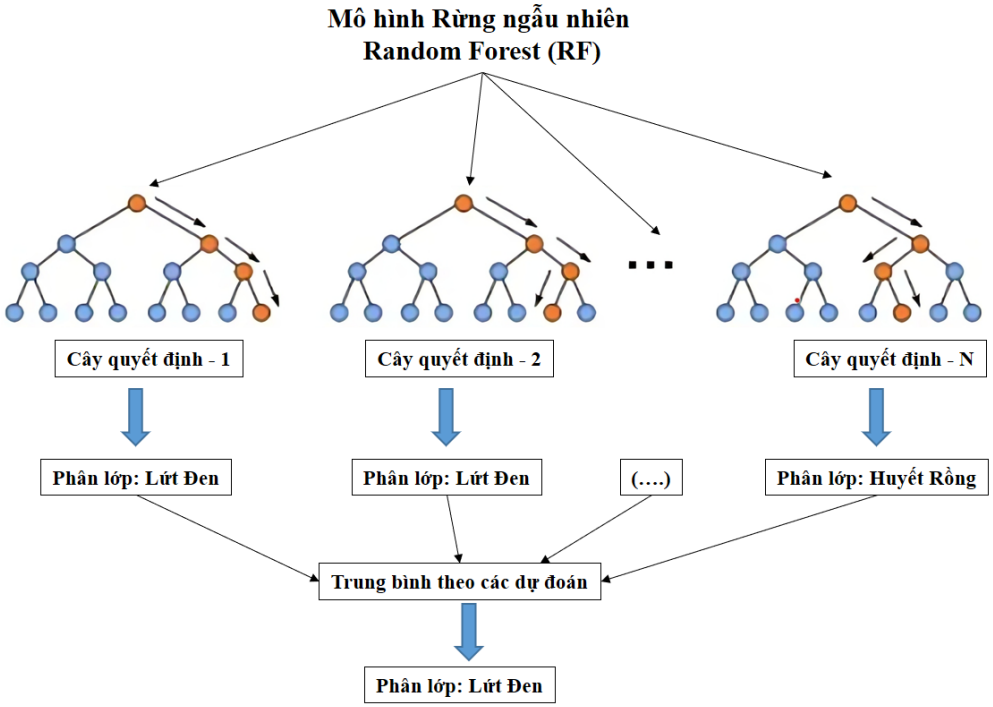
Lề mềm (soft margin): Thuật toán này cho phép SVM mắc một số lỗi nhất định và giữ cho lề càng rộng càng tốt để các điểm khác vẫn có thể được phân loại chính xác. Nói một cách khác, việc điều chỉnh cân bằng giữa phân loại sai và tối đa hóa lề có thể được thực hiện bằng cách điều chỉnh siêu tham số C. C càng lớn, thuật toán SVM càng bị phạt nặng khi tiến hành phân loại sai, do đó khoảng cách từ các véc tơ hỗ trợ tới siêu phẳng càng nhỏ (lề cứng). C càng nhỏ thì hình thức phạt càng ít, do đó khoảng cách này càng lớn (lề mềm).



Hình 5. Mô hình cây quyết định với bài toán phân lớp



Hình 6. Siêu phẳng trong không gian 2 chiều



Hình 7. Mô hình rừng ngẫu nhiên với bài toán phân lớp - Random Forest

Thủ thuật kernel: Kernel là một hàm ánh xạ dữ liệu từ không gian ít chiều sang không gian nhiều chiều hơn, từ đó tìm được siêu phẳng phân tách dữ liệu. Một số kernel phổ biến trong các thuật toán phân loại dữ liệu phi tuyến: tuyến tính (linear); đa thức (poly); RBF; sigmoid.

3.3. Thuật toán rừng ngẫu nhiên (Random Forest)

Mô hình phân lớp rừng ngẫu nhiên (RF), Breiman (2001), được xây dựng dựa trên một tập hợp gồm nhiều cây quyết định. Mỗi cây quyết định (DT) sẽ đưa ra dự đoán riêng với cùng một dữ liệu đầu vào. Kết quả dự đoán được đưa ra bởi mô hình này dựa trên nguyên tắc lấy trung bình (hay bỏ phiếu) để lựa chọn ra dự đoán nào chiếm ưu thế nhất. Trong hình 7 mô tả trường hợp dự đoán của mô hình là mẫu hạt gạo lứt Đen.

Rừng ngẫu nhiên có thể được sử dụng một cách hiệu quả với bộ dữ liệu với số thuộc tính lớn, dữ liệu không đầy đủ. Tuy nhiên, nhược điểm lớn nhất của mô hình này là tính không tường minh khi giải thích kết quả dự báo của từng cây quyết định cũng như kết quả bỏ phiếu cuối sau cùng của mô hình.

4. KẾT QUẢ VÀ THẢO LUẬN

Phân tích mối tương quan tuyến tính giữa các biến đặc trưng với nhãn được biểu diễn trên hình 8. Kết quả cho thấy sự khác biệt tương đối rõ ràng về mặt hình thái: lớp Huyết Rồng (mầu xanh) đặc trưng bởi độ lệch tâm cao (Eccentricity) và bán trục (Mninor_ax) bé hơn tương ứng với hình dáng dài và mảng của hạt gạo. Trong khi đó, lớp lứt Đen (mầu đỏ) có diện tích và chiều rộng lớn hơn, thể hiện hình dáng to và bè. Ngoài ra, hệ số tương quan của các đặc trưng: ['Skew', 'Periphery', 'Roundness', 'Angle_orientation', 'Solidity'] có giá trị nhỏ nhất nên bị loại bỏ.

Bước kế tiếp là loại bỏ các mẫu ngoại lai (outliers) được tiến hành với các biến đặc trưng có hệ số tương quan lớn nhất bao gồm ['Minor_ax', 'Eccentricity']. Các mẫu có giá trị nằm ngoài khoảng (25%, 75%) bị loại bỏ. Bộ cơ sở dữ liệu hạt gạo sau tiền xử lý được phân chia theo tỷ lệ 75/25 cho dữ liệu huyến luyện và dữ liệu kiểm tra như trong bảng 6.

Kỹ thuật tăng cường mẫu với các lớp thiểu số (SMOTE) được sử dụng để làm giảm ảnh hưởng mất cân bằng dữ liệu, tránh thiên vị với các lớp có nhiều mẫu (lứt Đen) và có thể học được nhiều đặc trưng hơn của lớp thiểu số (Huyết Rồng).

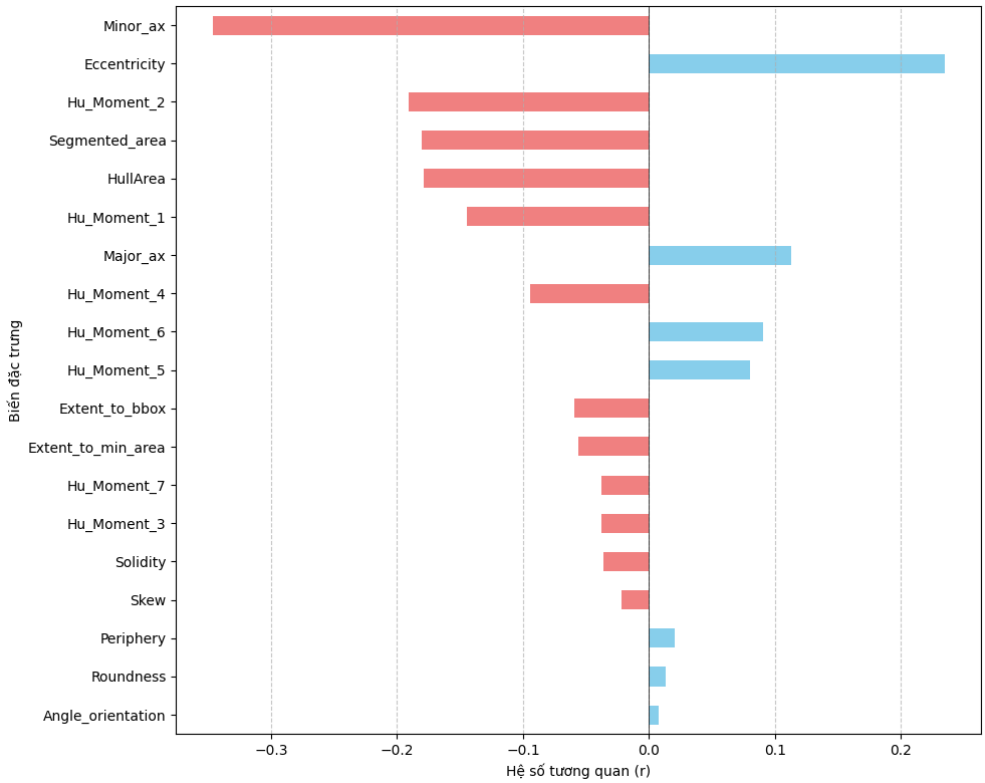
Với mỗi mô hình học máy, các tham số của được tối ưu hóa với kỹ thuật tìm tham số theo lưới chia (grid search). Với bộ tham số tối ưu thu được (Bảng 7), việc phân tích hiệu suất phân lớp được tiến hành dựa trên thuật toán đánh giá chéo với số vòng lặp k = 5.

Bảng 8 phân tích kết quả đánh giá các chỉ số Precision, Recall, F1-Score và Accuracy của ba mô hình máy học: Cây quyết định (DT), Rừng ngẫu nhiên (RF) và máy véc tơ hỗ trợ (SVM) đối với hai mẫu gạo trong bộ cơ sở dữ liệu gạo lứt Đen (GĐ) và gạo Huyết Rồng (HR).

Mô hình SVM có hiệu suất phân lớp cao nhất với độ chính xác vượt trội 76%, cho thấy khả năng phân loại của SVM với siêu phẳng tối ưu, tối đa khoảng cách giữa các lớp trong không gian đặc trưng. Mô hình rừng ngẫu nhiên có độ chính xác nhỉnh hơn cây quyết định nhờ vào cơ chế học tổng hợp, làm giảm phương sai và tăng tính ổn định.

Bảng 6. Thống kê số lượng mẫu hạt trong tập huyến luyện và kiểm tra bởi mô hình học máy

Tên mẫu	Tổng số hạt	Khối lượng	Dữ liệu huyến luyện	Dữ liệu kiểm tra
Gạo lứt Đen	1.598	37 gram	1.198	400
Gạo Huyết Rồng	1.137	35 gram	853	284
Tổng số mẫu	2.735	72 gram	2.051	684



Hình 8. Đánh giá mối tương quan giữa các biến đặc trưng với nhãn [lứt Đen (màu đỏ), Huyết Rồng (màu xanh)]

Bảng 7. Bảng siêu tham số với các mô hình học máy

Tên mô hình	Siêu tham số
Cây quyết định (DT)	criterion = ' gini '; max_depth = 4; min_samples_split = 4; min_samples_leaf = 30; max_leaf_nodes = 16
Máy véc tơ hỗ trợ (SVM)	C = 2; kernel = 'poly'; gamma = 1; degree = 3
Rừng ngẫu nhiên (RF)	max_depth = 20; max_features = 'sqrt'; min_samples_leaf = 1; min_sample_split = 2; n_estimators = 173.

Bảng 8. Kết quả phân loại với các mô hình học máy

Mô hình	Mẫu hạt	Precision	Recall	F1-Score	Accuracy
Cây quyết định (DT)	Lứt Đen	74,0	81,3	77,5	72,9
	Huyết Rồng	69,4	59,9	64,3	
Rừng ngẫu nhiên (Random Forest)	Lứt Đen	76,5	79,0	77,8	
	Huyết Rồng	69,0	65,9	67,4	73,5
Máy véc tơ hỗ trợ (SVM)	Lứt Đen	77,6	83,0	80,2	76,0
	Huyết Rồng	73,4	66,2	69,6	

Hiệu suất phân lớp được đánh giá bằng các chỉ số Recall và F1-Score cho thấy cả ba mô hình đều cho hiệu suất phân lớp rất tốt với mẫu lứt Đen, trên 77%, vượt trội hơn cả là SVM với giá trị F1-Score tới 80,2%. Tuy nhiên, với mẫu Huyết Rồng thì các chỉ số này đạt dưới 70%. Điều này cho thấy lớp 'thiểu số' có sự chồng lấn của các thuộc tính trong không gian đặc trưng, dẫn tới sự khó khăn khi phân loại dựa vào đặc trưng hình thái học.

5. KẾT LUẬN

Nghiên cứu này sử dụng các đặc trưng hình thái làm cơ sở để phân loại hai loại gạo lứt Đen và Huyết Rồng. Kết quả phân lớp có độ chính xác đến 76% với mô hình SVM mặc dù chưa cao nhưng hoàn toàn có ý nghĩa về mặt khoa học, vì ưu điểm của phương pháp phân loại này là tính khả thi trong triển khai mô hình một cách dễ dàng, với chi phí thấp. Việc bổ sung thêm nguồn dữ liệu cho việc học của mô hình là rất quan trọng cũng như sự kết hợp giữa các đặc trưng hình thái với các thuộc tính về mầu sắc sẽ cải thiện hiệu suất phân lớp và sẽ tiếp tục được thực hiện trong các nghiên cứu tiếp theo.

Phương pháp nghiên cứu trong bài báo này là cơ sở cho việc phân loại mẫu nông sản và là tiền để để xây dựng bộ cơ sở dữ liệu cho các sản phẩm nông sản Việt Nam nói chung và lúa, gạo nói riêng nhằm đáp ứng các yêu cầu về phân loại giống và việc phân biệt vùng trồng khác nhau trở nên thuận lợi và dễ dàng hơn. Với mô hình trên, một hệ thống tự động có thể được thiết kế để phân loại hạt giống cũng như các tạp chất không mong muốn.

TÀI LIỆU THAM KHẢO

Alexander K., Eric M., Nikhila R., Hanzi M., Chloe R., Laura G., Tete X., Spencer W., Alexander C. B., Wan Y.L., Piotr D. & Ross G. (2023). Segment anything. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3992-4003.

Berrar D. (2018). Cross-Validation. Reference Module in Life Sciences. doi:10.1016/b978-0-12-809633-8.20349-x

Benesty J., Chen J., Huang Y. & Cohen I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing. Springer. pp. 37-40.

Breiman L. (2001). Random Forests. Machine Learning. 45: 5-32. https://doi.org/10.1023/A:1010933404324

Cinar I. & Koklu M. (2019). Classification of rice varieties using artificial intelligence methods. Internaional Jounal of Intelligent Systems and Applications in Engineering. ISSN: 2147-6799.

Cortes C. & Vapnik V. (1995). Support-Vector Networks. Machine Learning. 20: 273-297. http://dx.doi.org/10.1007/BF00994018.

Jeremy F., Alex H. & Derek T. (2021). PyLabel. Retrieved from https://github.com/pylabel-project/pylabel on Jun 15, 2024.

Kim T.H., Kim E.K., Lee M.S., Lee H.K., Hwang W.S., Choe S.J., Kim T.Y., Han S.J., Kim H.J., Kim D.J. & Lee K.W. (2011). Intake of brown rice lees reduces waist circumference and improves metabolic parameters in type 2 diabetes. Nutr. Res. 31(2): 131-138. https:// doi.org/10.1016/ j.nutres.2011.01.010

Nguyễn Thị Lang, Lê Hoàng Phương, Bùi Chí Hiếu, Nguyễn Trọng Phước & Bùi Chí Bửu (2021). Phân tích chất lượng của giống lúa mùa AG3 tại An Giang. Tạp chí Nông nghiệp và Phát triển nông thôn. 11: 3-9.

Quinlan J.R. (1985). Induciton of Decision Trees. Machine Learning. 1: 81-106.

Tập đoàn Giống cây trồng Việt Nam (Vinaseed). Gạo lứt Phúc Thọ. Truy cập từ https://vinaseed.com.vn/vi/ product/m7/gao-lut-phuc-tho-den-63.htm ngày 02/07/2024.

Tianqi C. & Carlos G. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-794. https://doi.org/10.1145/2939672.293978

Trần Thị Thanh Thúy, Nguyễn Tấn & Võ Công Thành (2021). Nghiên cứu phục tráng giống lúa thơm đặc sản VD20 phục vụ cho xuất khẩu tại đồng bằng sông Cửu Long. Tạp chí Khoa học và Công nghệ Nông nghiệp Việt Nam. 4(125).

Vũ Mạnh Ấn, Hoàng Ngọc Đỉnh, Trần Hiền Linh, Phạm Xuân Hội & Hoàng Thị Giang (2023). Đánh giá một số chỉ tiêu chất lượng gạo của các giống lúa địa phương. Tạp chí Khoa học và Công nghệ Nông nghiệp Việt Nam (ISSN1859-1558). 2(144).