PHÂN CẤP BỆNH ĐỐM LÁ MẮT ẾCH HẠI ĐẬU TƯƠNG BẰNG CÁCH SỬ DỤNG PHẢN XẠ SIÊU QUANG PHỔ CỦA LÁ

Shuang Liu, Haiye Yu, Yuanyuan Sui, Haigen Zhou, Junhe Zhang, Lijuan Kong, Jingmin Dang, Lei Zhang
Võ Như Cầm biên dịch

Cao đẳng Kỹ thuật Sinh học và Nông nghiệp, Đại học Cát Lâm, Trường Xuân, Cát Lâm, Trung Quốc

TÓM TẮT

Trong nghiên cứu này, nghiên cứu tính khả thi của việc phân cấp bệnh đốm lá mắt ếch đậu tương (FLS). Hình ảnh lá và dữ liệu phản xạ siêu quang phổ của lá đậu tương khỏe mạnh và bị bệnh FLS đã được thu thập. Đầu tiên, xử lý hình ảnh được sử dụng để phân cấp FLS nhằm tạo ra một tham chiếu cho việc phân tích dữ liệu siêu quang phổ cận sau đó. Sau đó, các phương pháp giảm kích thước của dữ liệu siêu quang phổ được sử dụng để thu được thông tin liên quan đến FLS. Ba phương pháp đơn lẻ, cụ thể là chỉ số quang phổ (SI), phân tích thành phần chính (PCA) và lấy mẫu trọng số thích ứng cạnh tranh (CARS), cùng với phương pháp kết hợp PCA và SI, đã được đưa vào. PCA được sử dụng để lựa chọn các thành phần chính hiệu quả (PC) và đánh giá SI. Các bước sóng đặc trưng (CW) được chọn bằng cách sử dụng CARS. Cuối cùng, bước sóng đầy đủ, CW, PC hiệu dụng, SI và SI quan trọng được chia thành 14 bộ dữ liệu (DS1 – DS14) và được sử dụng làm đầu vào để xây dựng mô hình phân cấp. Màn trình diễn của các mẫu được đánh giá dựa trên độ chính xác của phân cấp cho cả cấp tổng thể và cấp riêng lẻ. Kết quả của chúng tôi cho thấy FLS bao gồm năm cấp dựa trên tỷ lệ tổng bề mặt lá được phủ FLS. Trong mô hình kết hợp PCA và SI, 5 PC và 20 SI có hệ số trọng lượng của mỗi PC cao hơn được trích xuất. Đối với dữ liệu siêu quang phổ, 20 CW và 26 PC hiệu quả cũng được chọn. Trong số 14 bộ dữ liệu, các biến đầu vào mô hình được cung cấp bởi năm bộ dữ liệu (DS2, DS3, DS4, DS10 và DS11) vượt trội hơn so với các biến có bước sóng đầy đủ (DS1) cả trong máy vectơ hỗ trợ (SVM) và vectơ hỗ trợ bình phương nhỏ nhất bộ máy phân cấp (LS-SVM). Các mô hình được phát triển bằng cách sử dụng năm bộ dữ liệu này đạt được độ chính xác tổng thể nằm trong khoảng từ 91,8–94,5% trong SVM và 94,5–97,3% trong LS-SVM. Ngoài ra, độ chính xác phân cấp đã được cải thiện từ 0,9–3,6% (SVM) và 0,9–3,7% (LS-SVM).

GIỚI THIỆU

Đậu tương, một loại cây họ đậu, là nguồn cung cấp protein và axit béo quan trọng [1], nguồn protein thức ăn sẵn có lớn nhất và là nguồn cung cấp dầu ăn lớn thứ hai [2]. Tổng sản lượng đậu tương là 7,6 triệu tấn trong năm 2018 trên toàn thế giới và khoảng 40% tổng sản lượng là ở Trung Quốc [3]. Do dân số đông và sở thích sử dụng dầu đậu nành từ lâu, Trung Quốc có mức tiêu thụ đậu nành cao nhất trên toàn cầu. Ngoài ra, đậu tương là nguồn cung cấp protein chính cho thức ăn cho lợn, giúp tăng tốc độ tiêu thụ của nó [4]. Tuy nhiên, một số dịch bệnh đã đe dọa nghiêm trọng đến năng suất và chất lượng đậu tương. Ví dụ, bệnh đốm lá mắt ếch (FLS) do nấm *Cercospora sojina* Hara (CSH) gây ra, là một loại bệnh hại lá trên cây đậu tương gây giảm năng suất và hư hỏng hạt giống, do đó thiệt hại về kinh tế. Dịch bệnh FLS có thể gây thiệt hại đến 60% năng suất. FLS là một bệnh đa vòng, trong đó nhiễm trùng, phát triển triệu chứng và sinh sản có thể lặp lại

nhiều lần trong một mùa [5]. Do đó, việc phát hiện và đánh giá mức độ bệnh để ước tính tác động kinh tế và áp dụng các chiến lược kiểm soát là điều cần thiết.

Thông thường, việc phát hiện các bệnh trên lá và mức độ nghiêm trọng của chúng chủ yếu dựa vào đánh giá bằng mắt thường hoặc các phương pháp hóa học. Các nhà nông học kiểm tra thủ công các kiểu màu lá, kích thước của vùng vết bệnh và cấu trúc ngọn, chẳng hạn như mật độ cây trồng, số lượng, hình dạng và sự phân bố của lá, số lượng thân và cành của cây trồng, … [6]. Tuy nhiên, phương pháp đánh giá trực quan là chủ quan và có thể bị ảnh hưởng bởi kiến thức thực nghiệm của người quan sát [7]. Khi các loại bệnh hại cây trồng gia tăng, các loại bệnh có thể phát hiện được sẽ hạn chế. Bên cạnh đó, khi nhiều bệnh có biểu hiện hình thái giống nhau, người quan sát sẽ khó phân biệt chính xác. Ngược lại, các phương pháp hóa học, bao gồm phản ứng chuỗi polymerase (PCR) và xét nghiệm chất hấp thụ miễn dịch liên kết với enzym (ELISA), có độ nhạy cao và có thể phát hiện chính xác bệnh. Mặc dù chúng có độ chính xác cao, nhưng các phương pháp hóa học tốn nhiều thời gian, công sức và tính hủy diệt [8]. Hơn nữa, hình ảnh siêu quang phổ đã được sử dụng để phát hiện mức độ nghiêm trọng của bệnh cây trồng [7, 9–12], nhưng đòi hỏi tính toán và đào tạo rộng rãi và tốn thời gian [13]. Do đó, một phương pháp thích hợp, không phá hủy, nhanh chóng và hiệu quả cao vẫn được đảm bảo để đánh giá mức độ nghiêm trọng của bệnh hại cây trồng.

Được phát triển cho các ứng dụng nông nghiệp chính xác, kỹ thuật siêu quang phổ đã thu hút nhiều sự chú ý trong việc phát hiện bệnh cây trồng và ước tính mức độ nghiêm trọng của chúng [14, 15]. Phương pháp này mang tính khách quan, không phá hoại, có thể phát hiện những biến đổi sinh lý bên trong của lá trong các điều kiện căng thẳng khác nhau với hiệu quả cao. Nó cũng tiết kiệm thời gian, giúp tạo điều kiện thuận lợi cho các chiến lược quản lý và cải thiện năng suất [16, 17]. Kỹ thuật siêu quang phổ cũng có thể được sử dụng để tiến hành kiểm tra chuyên sâu các đặc tính của cây trồng, chẳng hạn như cấu trúc tế bào cây trồng, hàm lượng chất diệp lục, độ ẩm, hàm lượng nguyên tố vi lượng, đặc điểm phản xạ và hấp thụ ánh sáng. Điều này có thể không đạt được bằng cách sử dụng dữ liệu đa quang, có băng thông thô tương đối [18]. Nhiều nghiên cứu đã chỉ ra tính khả thi và tiềm năng của dữ liệu phản xạ siêu quang phổ để đánh giá mức độ bệnh hại cây trồng như bệnh bạc lá hại lúa [19], bệnh khảm vàng đậu đen [20] và bệnh vàng lá gân xanh trên cây có múi [21]. Gần đây, các phương pháp học máy đã được sử dụng để phát triển các mô hình phân cấp bệnh cây trồng [22, 23], chẳng hạn như mạng nơron [24–26], máy vectơ hỗ trợ (SVM) [11] và máy vectơ hỗ trợ bình phương nhỏ nhất (LS-SVM) [22, 27, 28]. SVM là một thuật toán học máy có giám sát được sử dụng để phân cấp và hồi quy, trong khi LS-SVM là một phần mở rộng của SVM thường chuyển dữ liệu phi tuyến có chiều thấp thành dữ liệu tuyến tính có chiều cao để giải quyết các thách thức liên quan đến mô hình dữ liệu phi tuyến [29]. Để tránh trang bị quá nhiều, LS-SVM ủng hộ nguyên tắc giảm thiểu rủi ro cấu trúc hơn nguyên tắc giảm thiểu rủi ro theo kinh nghiệm truyền thống được sử dụng trong mạng nơ-ron thông thường [30]. LS-SVM nhanh hơn và chính xác hơn SVM, và có thể được sử dụng trong cả phân tích đa biến tuyến tính và phi tuyến [31]. Dữ liệu siêu ảnh có chứa thông tin quá mức, dư thừa và có độ tương quan cao trên một số lượng lớn các bước sóng, điều này có thể làm tăng số lượng phép tính. Do đó, việc giảm kích thước dữ liệu, tối ưu hóa và đơn giản hóa toàn bộ bước sóng, là điều cần thiết. Một số phương pháp đã được sử dụng hiệu quả để giảm kích thước của dữ liệu siêu quang phổ, bao gồm các thuật toán khai thác bước sóng hiệu quả như lấy mẫu trọng số thích ứng cạnh tranh (CARS) [32, 33], tính toán các chỉ số quang phổ (SI) [34–36] và phân tích thành phần (PCA) [37, 38]. Thuật toán CARS tuân theo thuyết tiến hóa của

Darwin về "sự sống sót của những cá thể khỏe mạnh nhất" [39]. Trong thuật toán này, biến của mỗi bước sóng được coi là một cá thể riêng lẻ. Trong quá trình chọn lọc, chỉ những cá thể có khả năng thích nghi mạnh mới được giữ lại. Trong khi đó, trong quá trình chọn bước sóng, nhiều tập con của các biến bước sóng có giá trị tuyệt đối lớn của hệ số hồi quy thu được bằng cách loại bỏ bước sóng có hệ số hồi quy nhỏ. Sau khi thu được nhiều tập con của các biến, phương pháp xác nhận chéo sẽ được sử dụng để thu được tập con tối ưu của các biến có sai số bình phương trung bình căn nhỏ nhất của phép xác nhận chéo (RMSECV), được định nghĩa là tập con bước sóng tối ưu [40, 41]. CARS là một cách tiếp cận thực tế được sử dụng để chọn các dải đặc trưng [42–44]. Các chỉ số quang phổ (SI), thu được bằng cách kết hợp các giá trị phản xạ ở hai hoặc nhiều bước sóng hoặc dải bước sóng, được sử dụng để loại bỏ thông tin không liên quan, do đó nâng cao đặc tính của vật thể [45–47]. PCA là một phương pháp thống kê đa biến có thể giảm kích thước của dữ liệu một cách hiệu quả trong khi vẫn giữ lại thông tin hữu ích từ dữ liệu gốc. Tất cả các thành phần chính (PC) đều độc lập với nhau, điều này giúp loại bỏ ảnh hưởng của thông tin dư thừa trong bộ dữ liệu thứ nguyên cao. PCA cũng có thể giảm sự phụ thuộc nhiều vào các dải sóng liền kề. Do đó, một số nghiên cứu đã báo cáo sử dụng PCA để nén dữ liệu của dữ liệu siêu quang phổ [18, 48, 49]. Giảm dư thừa và tương quan cải thiện độ chính xác và độ tin cậy của kết quả phân tích dữ liệu siêu quang phổ. Tuy nhiên, các nghiên cứu nói trên đã sử dụng một phương pháp giảm kích thước dữ liệu đơn lẻ và bỏ qua tầm quan trọng của việc kết hợp các phương pháp trích xuất đặc trưng khác nhau. Cho đến nay, chưa có nghiên cứu nào phân cấp toàn diện các loại bệnh hại cây trồng khác nhau bằng cách sử dụng cả phương pháp chiết xuất đơn tính và kết hợp. Các ứng dụng khác của sự kết hợp khác nhau của các phương pháp chiết xuất đặc trưng bao gồm phân cấp hình ảnh quang phổ [50, 51] và các nguyên tố dinh dưỡng cây trồng và phát hiện thông tin sinh lý, chẳng hạn như dự đoán trạng thái nitơ của lúa [52] và ước tính hàm lượng diệp lục trong lúa [53] và lúa mì [54] thực vật. Các nghiên cứu này đã chứng minh rằng các phương pháp trích xuất tính năng kết hợp ưu việt hơn so với chỉ sử dụng một phương pháp.

Trong nghiên cứu này, các phương pháp đơn lẻ của SI, CARS và PCA, và sự kết hợp của PCA và SI được sử dụng để trích xuất thông tin hiệu quả về FLS. Hơn nữa, hiệu suất của các bước sóng và SI hiệu quả để phát hiện FLS đã được so sánh. Công nghệ siêu quang phổ cận đã được sử dụng để đánh giá bệnh cây trồng, tuy nhiên, việc áp dụng độ phản xạ siêu quang phổ của lá kết hợp với các phương pháp giảm kích thước dữ liệu khác nhau để thiết lập các mô hình phân cấp FLS đậu tương vẫn chưa được báo cáo. Do đó, các mục tiêu chính của nghiên cứu này là (i) ước tính loại bệnh FLS bằng cách sử dụng phân tích phản xạ siêu quang phổ trên lá, (ii) đánh giá hiệu suất của các phương pháp chiết xuất đặc điểm và mô hình hóa để phát hiện loại bệnh FLS, và (iii) xác định tính khả thi của việc phân cấp FLS thông qua các phương pháp học máy sử dụng 14 bộ dữ liệu.

VẬT LIỆU VÀ PHƯƠNG PHÁP

Nuôi cấy và cấy mẫu

Cây đậu tương (Hushan 60) được trồng trong chậu nhựa (ø 260mm) trong nhà kính năng lượng mặt trời được kiểm soát môi trường tại Đại học Cát Lâm, Trung Quốc, ở 25/20^{0}C (ngày/đêm), độ ẩm tương đối 60–80% (RH) và khoảng thời gian chụp ảnh là 15 giờ mỗi ngày (Hình 1). Mỗi chậu chứa một cây và tổng số 125 hạt đã được trồng vào ngày 10 tháng 7 năm 2019 (Hình 1A). Trong số đó, 100 cây được sử dụng để chống nhiễm bệnh,

trong khi 25 cây còn lại dùng để kiểm soát. Tất cả các cây đều được tưới nước và bón phân bình thường theo cùng một cách trước khi bị nhiễm bệnh.

Chất cấy FLS của đậu tương được lấy từ Học viện Khoa học Nông nghiệp Cát Lâm, Trung Quốc. Bào tử CSH được thu hoạch từ môi trường nuôi cấy thạch dextrose khoai tây bằng cách làm ngập trong nước cất vô trùng và cạo nhẹ bề mặt bằng que thủy tinh vô trùng. Sau đó, các bào tử thu hoạch được lọc bằng gạc, và 3% sucrose được thêm vào để tạo ra huyền phù cấy với nồng độ bào tử 1 × 105/ml. Hỗn dịch cấy này đã được sử dụng cho nhiễm trùng. Tổng số 100 cây đã được cấy các mầm bệnh. Cây đậu tương được cấy trong giai đoạn ra hoa đầu tiên vào ngày 15 tháng 8 năm 2019. Trong nghiên cứu này, một phương pháp cấy nhân tạo đã được áp dụng, trong đó toàn bộ lá cây được phun đồng nhất bằng một vòi phun nhỏ vào những ngày nhiều mây (Hình 1B). Hình 1C cho thấy trạng thái của lá sau khi bị nhiễm bệnh. Sau khi cấy, cây được bao phủ bằng túi nhựa trong 48 giờ để duy trì độ ẩm cao. Quy trình cấy được lặp lại sau một tuần để đảm bảo tỷ lệ mắc bệnh FLS. Cuối cùng, các cây bị nhiễm bệnh và nhóm đối chứng được tách ra ở các vị trí khác nhau trong hai lán nhựa được duy trì ở cùng nhiệt độ 28/18^{0}C (ngày/đêm) và độ ẩm (75 – 80%).

Hình 1. Hình ảnh đại diện của quá trình trồng và cấy giống đậu tương.
(a) Môi trường trồng đậu tương, (b) cấy nhân tạo, và (c) lá sau khi cấy.

Thu thập dữ liệu

Dữ liệu thí nghiệm được thu thập trong khoảng thời gian từ 10 giờ đến 14 giờ ngày 20 tháng 9 năm 2019. Trong quá trình đo không có gió, ít mây và đủ ánh sáng mặt trời. Trong nghiên cứu này, hai loại dữ liệu là hình ảnh lá và độ phản xạ siêu quang phổ của lá đã được thu thập. Tuy nhiên, không phải tất cả các lá bị nhiễm vi khuẩn đều phát triển các triệu chứng, điều này thường gặp ở bệnh FLS. Các lá có FLS có thể nhìn thấy được chọn từ nhóm xử lý để thu thập dữ liệu, trong khi các lá khỏe mạnh được thu thập từ nhóm đối chứng. Một lá bị bệnh hoặc khỏe mạnh được coi là một mẫu. Tổng cộng, 440 mẫu, bao gồm 340 lá bệnh và 100 lá khỏe mạnh, đã được thu thập và nghiên cứu.

Trước khi đo lường dữ liệu siêu quang phổ của lá, một hình ảnh được chụp cho tất cả 440 mẫu. Hình ảnh chiếc lá thu được bằng điện thoại di động thông thường (OPPO R9s) với độ phân giải 1600 triệu pixel. Tất cả các bức ảnh được chụp với nền trắng.

Dữ liệu về độ phản xạ siêu quang phổ của lá được thu thập bằng cách sử dụng máy đo phổ Field Spec HandHeld-2 (Thiết bị quang phổ phân tích, Boulder, Colorado, Mỹ), có phụ kiện kẹp lá (của cùng một công ty) được kết nối với nó. Phần trung tâm của mỗi lá đã được đo. Vùng siêu quang phổ dao động từ 325–1.075nm, với độ phân giải 3nm. Số kênh siêu âm là 512. Máy đo phổ được làm ấm trong 30 phút trước khi sử dụng để loại bỏ ảnh hưởng của phông nền lên thông tin phổ và đèn được bật trong 5 phút để duy trì độ ổn

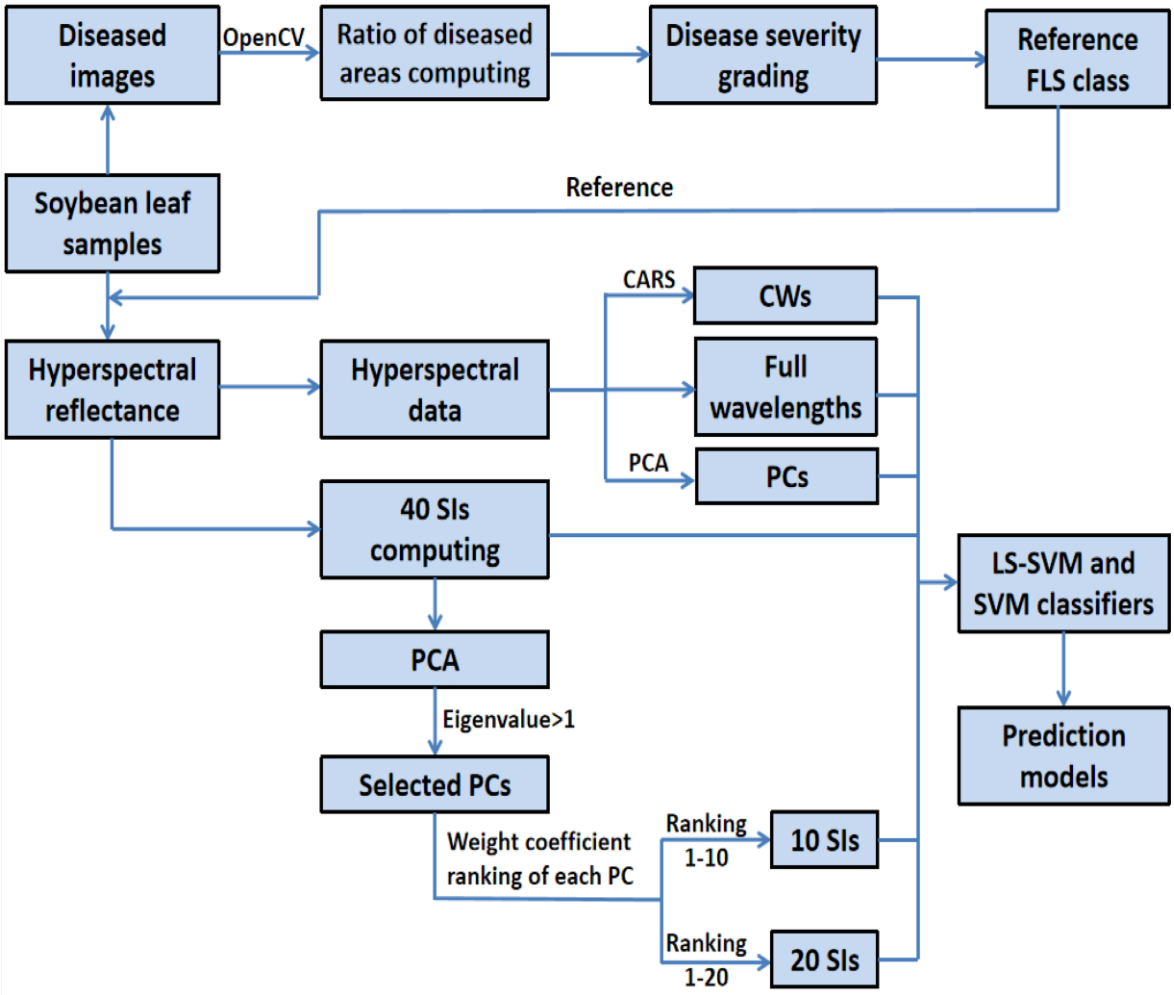
định của quang phổ. Máy đo phổ đã được hiệu chuẩn để thu được hệ số phản xạ siêu quang phổ tương đối của mẫu theo công thức sau:

$$R = \frac{R_s - R_d}{R_w - R_d} \qquad (1)$$

Trong đó, R là hệ số phản xạ tương đối, Rs là phổ mẫu, Rw là tham chiếu màu trắng và Rd là dòng điện tối. Tham chiếu màu trắng thu được bằng một tấm tham chiếu màu trắng hình tròn có đường kính 3 inch, và dòng điện tối thu được bằng cách che thấu kính bằng một tấm bảng mờ. Mười đường cong phản xạ trên mỗi mẫu và giá trị trung bình của chúng đã được tính toán.

Xử lí dữ liệu

Lưu đồ trong Hình 2 trình bày các phương pháp xử lý dữ liệu được thực hiện trong nghiên cứu này. Lưu đồ cho thấy một phương pháp xử lý hình ảnh và một phương pháp xử lý dữ liệu phản xạ siêu quang phổ để giảm kích thước siêu quang phổ.



Hình 2. Sơ đồ các phương pháp xử lý và phân tích dữ liệu được sử dụng trong nghiên cứu này.

Mức độ nặng nhẹ của bệnh ở vảy lá phụ thuộc vào vùng vết bệnh trên phiến lá. Phương pháp xử lý ảnh xác nhận loại bệnh bằng cách tính toán tỷ lệ của tổng bề mặt lá được bao phủ bởi các vùng đốm trong ảnh của lá đậu tương. Phương pháp này chủ yếu dựa trên thư viện hàm OpenCV, sử dụng công nghệ nhận dạng thị giác máy tính để trích xuất vùng bị bệnh của FLS dựa trên không gian màu sắc, độ bão hòa và giá trị (HSV) của hình ảnh. Mỗi hình ảnh bị bệnh được chia thành hai khu vực khác nhau dựa trên đặc điểm màu sắc.

Trong không gian màu HSV, các pixel có giá trị ngưỡng lớn hơn (1, 1, 1) và nhỏ hơn (37, 255, 255) được phân cấp là pixel FLS, trong khi những pixel có giá trị lớn hơn (37, 255, 255) và nhỏ hơn (255, 255, 255) được phân cấp là pixel sức khỏe. Cuối cùng, tỷ lệ phần trăm của tổng số diện tích bị bệnh, đối với diện tích của toàn bộ lá, được tính toán. Các giá trị thu được được sử dụng như một tham chiếu chân lý cơ bản trong nghiên cứu này, sau đây sẽ được gọi là cấp FLS tham chiếu. Tỷ lệ tổn thương được chia thành sáu loại dựa trên các thông số kỹ thuật để đánh giá bệnh đốm lá mắt ếch đậu tương (Số kỹ thuật: NY/T3114.2 – 2017) ở Trung Quốc: 0–1% (cấp 1), 1–3 % (cấp 2), 3–6% (cấp 3), 6–20% (cấp 4), 20–50% (cấp 5), > 50% (cấp 6). Các lá khỏe mạnh được gọi là loại 0.

Bảng 1. Danh sách các chỉ số quang phổ được trích dẫn trong các tài liệu trước đây và được phát triển trong nghiên cứu này.

Chỉ số quang phổ	Công thức	Tài liệu tham khảo
Chỉ số thực vật khác nhau bình thường	$NDVI_1 = (R_{900}-R_{680})/(R_{900}+R_{680})$	(55)
	$NDVI_2 = (R_{800}-R_{670})/(R_{800}+R_{670})$	(13)
	$NDVI_3 = (R_{870}-R_{530})/(R_{870}+R_{530})$	Nghiên cứu này
	$NDVI_4 = (R_{850}-R_{680})/(R_{850}+R_{680})$	Nghiên cứu này
	$NDVI_5 = (R_{855}-R_{650})/(R_{855}+R_{650})$	Nghiên cứu này
	$NDVI_6 = (R_{750}-R_{500})/(R_{750}+R_{500})$	Nghiên cứu này
	$NDVI_7 = (R_{815}-R_{715})/(R_{815}+R_{715})$	Nghiên cứu này
Chỉ số thực vật điều chỉnh hình tam giác	$MTVI = 1,2 \times \{1,2 \times (R_{800}-R_{550}) - 2,5 \times (R_{670}-R_{550})\}$	(13)
Chỉ số thực vật khác nhau chuẩn	$RDVI = (R_{800}-R_{670})/\sqrt{(R_{800}-R_{670})}$	(13)
Chỉ số sức khỏe	$HI = \{(R_{534}-R_{698})/(R_{534}+R_{698})\}-0,5 \times R_{704}$	(44)
Tỷ lệ đơn giản	$SR_1 = R_{800}/R_{670}$	(13)
	$SR_2 = R_{800}/R_{550}$	(13)
	$SR_3 = R_{750}/R_{550}$	(13)
	$SR_4 = R_{850}/R_{615}$	Nghiên cứu này
	$SR_5 = R_{800}/R_{680}$	Nghiên cứu này
	$SR_6 = R_{780}/R_{500}$	Nghiên cứu này
Chỉ số xanh lá cây	$GI = R_{570}/R_{670}$	(56)
Chỉ số stress bệnh nước	$DSWI = R_{550}/R_{680}$	(13)
Chỉ số phản xạ Anth	$ARI = 1/R_{550} - 1/R_{700}$	(13)
Chỉ số xanh đậm	$BI = R_{450}/R_{490}$	(13)
Chỉ số thực vật khác nhau bình thường cạnh đỏ	Red-edge $NDVI = (R_{750}-R_{705})/(R_{750}+R_{705})$	(57)
Chỉ số thực vật điều chỉnh đất tối ưu	$OSAVI = (1+0,16) \times (R_{800}-R_{670})/(R_{800}+R_{670}+ 0,16)$	(58)
Tỷ lệ đơn giản sắc tố cụ thể diệp lục b	$PSSR_b = R_{800}/R_{635}$	(13)
Tỷ lệ đơn giản sắc tố cụ thể carotenoid	$PSSR_c = R_{800}/R_{550}$	(13)
Chỉ số giảm đỏ	$RI = R_{700}/R_{550}$	(13)
Chỉ số khá nhau bình thường	$NDI_1 = (R_{800}-R_{680})/(R_{800}+R_{680})$	(13)
	$NDI_2 = (R_{750}-R_{660})/(R_{750}+R_{660})$	(13)
	$NDI_3 = (R_{750}-R_{705})/(R_{750}+R_{705})$	(13)
	$NDI_4 = (R_{755}-R_{680})/(R_{755}+R_{680})$	Nghiên cứu này
	$NDI_5 = (R_{800}-R_{550})/(R_{800}+R_{550})$	Nghiên cứu này
Chỉ số phản xạ quang hóa	$PRI_1 = (R_{531}-R_{570})/R_{531}+R_{570})$	(59)
	$PRI_2 = (R_{530}-R_{570})/R_{530}+R_{570})$	(13)
Chỉ số phản xạ trong điều chỉnh diệp lục	$MCARI = \{(R_{700}-R_{670}) - 0,2 \times (R_{700}-R_{550})\}/(R_{700}/R_{670})$	(60)
Chỉ số thực vật khác nhau bình thường xanh lá	$GNDVI = (R_{801}-R_{550})/(R_{801}+R_{550})$	(60)

cây		
Chỉ số phản xạ trong biến đổi diệp lục	$TCARI = 3 \times \{(R_{700}-R_{670}) - 0,2 \times (R_{700}-R_{550})/(R_{700}/R_{670})\}$	(13)
Chỉ số thực vật hình tam giác	$TVI = 0,5 \times \{(120 \times (R_{750}-R_{550}) - 200 \times (R_{670}-R_{550})\}$	(13)
Chỉ số màu xanh lá cây xanh đậm	$BGI_1 = R_{400}/R_{550}$	(58)
	$BGI_2 = R_{450}/R_{550}$	(58)
Chỉ số màu đỏ xanh đậm	$BRI_1 = R_{400}/R_{690}$	(58)
	$BRI_1 = R_{450}/R_{690}$	(58)

Ghi chú: R đại diện cho sự phản xạ siêu quang phổ và hình vẽ đại diện cho bước sóng tương ứng

Trong nghiên cứu này, dữ liệu phản xạ siêu quang phổ được xử lý bằng ba phương pháp đơn lẻ, đó là SI, CARS, PCA và phương pháp kết hợp PCA và SI. Bốn phương pháp này được sử dụng để giảm kích thước của dữ liệu siêu quang phổ. Đầu tiên, chúng tôi tính toán 40 SI bằng cách sử dụng hai hoặc nhiều tổ hợp phản xạ bước sóng, sau đó tiến hành lựa chọn SI bằng PCA dựa trên xếp hạng của các hệ số trọng số. PC và CW được trích xuất từ dữ liệu siêu quang phổ trên toàn bộ dải bước sóng bằng cách sử dụng PCA và CARS tương ứng cùng một lúc. Cuối cùng, SI, PC và CW đã chọn được sử dụng để xây dựng mô hình phân cấp với bộ phân cấp SVM và LS-SVM, các mô hình này được so sánh với mô hình được thiết lập bởi đầy đủ các bước sóng để đánh giá tính khả thi và tính ưu việt của phương pháp phân biệt của cấp FLS được sử dụng trong nghiên cứu này.

Chỉ số quang phổ (SI)

Nghiên cứu này liệt kê 30 SI thường được sử dụng từ các nghiên cứu trước và 10 SI được phát triển trong nghiên cứu này (Bảng 1). Các SI này được lấy từ các bước sóng trong vùng khả kiến và vùng cận hồng ngoại được tính toán từ dữ liệu siêu quang phổ âm thô sau khi hiệu chỉnh đường cơ sở, sau đó, ngoại lệ được loại bỏ.

Phân tích thành phần chính (PCA)

Trong nghiên cứu này, phương pháp PCA được sử dụng để trích xuất các PC hiệu quả và chọn các SI hiệu quả.

Dữ liệu siêu quang phổ cận được nén bởi PCA, công cụ này trích xuất các PC hiệu quả làm biến đầu vào. Giá trị riêng là một chỉ số cho biết mức độ ảnh hưởng của PC. Nói cách khác, giá trị riêng đại diện cho lượng thông tin của các biến ban đầu có thể được giải thích trung bình sau khi PC được đưa vào. Giá trị riêng nhỏ hơn 1 cho bất kỳ PC nào cho biết rằng tác động của PC đó ít quan trọng hơn tác dụng của một biến đơn lẻ. Do đó, chỉ những PC có giá trị riêng lớn hơn 1 mới được chọn để phân tích tiếp theo trong nghiên cứu này.

40 SI nói trên đã phải chịu PCA. Trong mỗi PC được chọn, 40 SI được sắp xếp theo hệ số trọng lượng của chúng. Mỗi PC bao gồm 40 SI, trong đó hệ số trọng lượng của mỗi SI riêng lẻ thể hiện tầm quan trọng của nó. Xét theo nguyên tắc, "số mẫu lớn hơn 5–10 lần số SI tối ưu", số SI tối thiểu và tối đa được coi là giới hạn dưới và giới hạn trên, các thành phần của chúng được cho điểm trong một cột được liệt kê cho mỗi PC. Do đó, ảnh hưởng của số lượng SI có thể được đánh giá dựa trên kết quả phân cấp của chúng.

Lấy mẫu trọng số thích ứng cạnh tranh (CARS).

Lựa chọn CW nhằm mục đích chỉ chọn một số bước sóng mang thông tin hữu ích nhất để giảm tính toán dữ liệu siêu quang phổ [61]. Trong nghiên cứu này, tất cả dữ liệu siêu

quang phổ được xử lý bằng thuật toán CARS; các bước sóng được chọn là CWs, được coi là đầu vào để phát triển các mô hình phân cấp.

Phương pháp mô hình hóa

Hai phương pháp mô hình hóa, máy vectơ hỗ trợ (SVM) và máy vectơ hỗ trợ bình phương nhỏ nhất (LS-SVM), được phát triển tương đối để xác định và phân cấp FLS của đậu tương trong nghiên cứu này. Chức năng quyết định của SVM là:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} y_i a_i K(x, x_i) + b) \qquad (2)$$

Trong đó a_i là cấp số nhân Lagrangian, b là giá trị độ lệch, (x_i, y_i) là vectơ hỗ trợ.

Chức năng quyết định của LS-SVM có thể được tính như sau:

$$f(x) = \sum_{i=1}^{n} a_i K(x, x_i) + b \qquad (3)$$

Trong đó a_i là một cấp số nhân Lagrangian, K (x, x_i) là một hàm nhân cơ sở bán kính và b là độ lệch thống kê.

Trong cả hai mô hình SVM và LS-SVM, một yếu tố chung, chức năng hạt nhân, được mong đợi để chọn. Hàm nhân cơ sở xuyên tâm (RBF) được khuyến nghị làm chức năng nhân của SVM và LS-SVM, vì RBF có thể xử lý các mối quan hệ phi tuyến giữa các thuộc tính phổ và mục tiêu và cung cấp một hiệu suất tốt trong các giả định chung về độ mượt mà. Do đó, RBF được sử dụng làm chức năng nhân của SVM và LS-SVM trong nghiên cứu này. Hai tham số quan trọng trong mô hình SVM là hệ số phạt (C) và tham số hàm nhân gamma (g). C là hệ số phạt, là khả năng chịu lỗi, g là một tham số xác định tính phi tuyến của nhân RBF. Giá trị này xác định một cách ngầm định sự phân bố của dữ liệu sau khi nó được lập bản đồ không gian đặc trưng mới. Trong mô hình LS-SVM, cũng có hai tham số quan trọng cần được xác định: tham số chính quy gam (γ) và tham số sig^2 (σ^2) của hàm nhân RBF. γ xác định sự cân bằng giữa việc giảm thiểu độ phức tạp của mô hình và giảm thiểu lỗi huấn luyện, và σ^2 là băng thông, xác định ngầm định lập bản đồ phi tuyến từ không gian đầu vào đến không gian đặc trưng chiều cao.

Trong nghiên cứu này, phương pháp xác nhận chéo k-lần được sử dụng để tối ưu hóa các kết hợp tham số (C, g) và (γ, σ^2). Bộ mẫu huấn luyện được chia đều thành k nhóm. Mỗi lần nhóm (k-1) được sử dụng để huấn luyện, nhóm còn lại được sử dụng để xác minh. Mỗi nhóm dữ liệu được xoay vòng làm dữ liệu xác minh để xác minh tỷ lệ nhận dạng của mô hình. K = 10 đã được sử dụng trong nghiên cứu này. Trong trường hợp của mỗi bộ (C, g) và (γ, σ^2), mỗi nhóm dữ liệu được luân chuyển và xác minh. Nói chung, sử dụng nhiều nhóm hơn, làm tăng độ chính xác của các phép tính, nhưng cũng làm tăng thời gian tính toán. Vì vậy, cần phải đánh đổi giữa hiệu quả và độ chính xác. Ngoài ra, để đạt được sự kết hợp tối ưu của (C, g) và (γ, σ^2) và tránh những thách thức quá mức, một phương pháp tìm kiếm lưới đã được sử dụng.

Trong SVM và LS-SVM, các giá trị giả, chẳng hạn như 0, 1, 2, 3, 4, được sử dụng làm giá trị y. Các giá trị giả của 0, 1, 2, 3 và 4 được sử dụng để đại diện cho năm cấp, tương ứng với các cấp đậu nành FLS là “cấp 0”, “cấp 1”, “cấp 2”, “cấp 3’ và “cấp 4”.

Đánh giá hiệu suất trên các mô hình

Hiệu suất của cả hai mô hình SVM và LS-SVM được xác định với tỷ lệ nhận dạng của thử nghiệm được đặt là độ chính xác phân cấp của các cấp FLS đối với đậu tương. Độ chính xác của phân cấp đưa ra ước tính về mức độ hoạt động của các biến đầu vào nhất định. Hai loại độ chính xác phân cấp, độ chính xác phân cấp trong mỗi loại, được gọi là

độ chính xác riêng lẻ và cấp còn lại là độ chính xác tổng thể, cả hai đều được sử dụng để đánh giá mô hình. Độ chính xác phân cấp cung cấp ước tính về mức độ chính xác của mẫu được phân cấp, giá trị càng cao, hiệu suất của mô hình càng tốt. Hai độ chính xác được xác định bằng cách sử dụng Công thức (4) và (5) như sau:

$$\text{individual accuracy (\%)} = \frac{\text{correctly classified samples in A}}{\text{all samples in A}} \times 100 \quad (4)$$

$$\text{overall accuracy (\%)} = \frac{\text{the total number of correctly classified samples in each class}}{\text{all samples}} \times 100 \quad (5)$$

Trong đó A đại diện cho các cấp tương ứng: "cấp 0", "cấp 1", "cấp 2", "cấp 3' và "cấp 4".

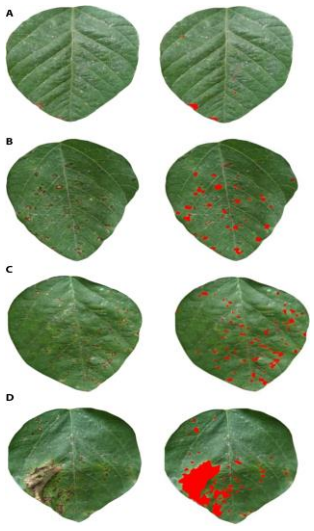
Phần mềm xử lý dữ liệu

Xử lý trước dữ liệu, tính toán thống kê và phân tích dữ liệu được thực hiện bằng ViewSpec Pro (ASD Inc., Boulder, Colorado, Mỹ), MATLAB R2018a (Mathworks Inc., Natick, Mỹ), SPSS 24.0 (IBM Inc., Chicago , IL, MỸ) và Origin 19.0 (OriginLab, Hampton, Mỹ). Các đường cong siêu quang phổ được tính trung bình bằng ViewSpec Pro. Nhận dạng và tính toán FLS đã được hoàn thành bằng cách sử dụng MATLAB R2018a. Các mô hình SVM và LS-SVM được thiết lập bằng cách sử dụng hộp công cụ libsvm và hộp công cụ LS-SVM v1.8 chạy trên MATLAB R2018a, tương ứng. SI đã được chọn thông qua PCA sử dụng SPSS 24.0 và tất cả các biểu đồ được vẽ bằng Origin 19.0. Tất cả các hoạt động được thực hiện bằng cách sử dụng nền tảng Microsoft Windows 10 (64bit).

KẾT QUẢ VÀ THẢO LUẬN

Khu vực FLS khai thác và phân cấp

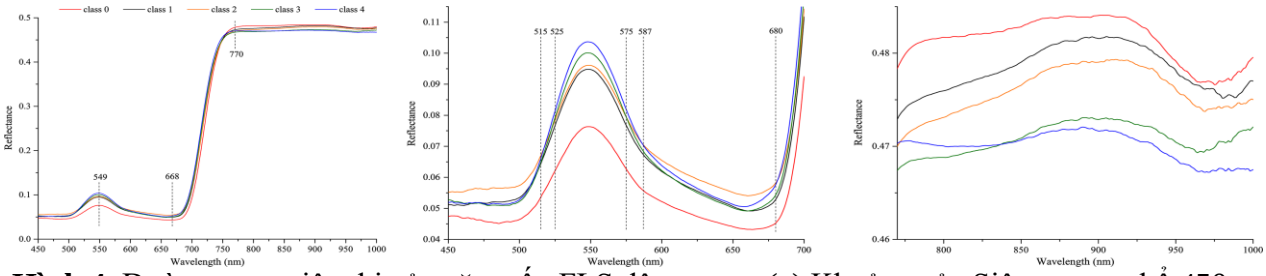
Hình 3 cho thấy các ảnh FLS gốc và đã xử lý, tỷ lệ các vùng FLS và kết quả phân cấp dựa trên công nghệ thị giác máy vi tính. Tất cả 440 lá đậu tương được chia thành độ 0–4 theo tỷ lệ vùng FLS. Số lá ở các cấp 0, 1, 2, 3 và 4 lần lượt là 100, 100, 100, 80 và 60. Sau đó, phân tích tiếp theo về dữ liệu siêu quang phổ cận được thực hiện dựa trên kết quả phân cấp như cấp FLS tham chiếu.



Hình 3. Hình ảnh gốc và đã xử lý, tỷ lệ vùng FLS và kết quả phân cấp.
(a) 0,43%, cấp 1;
(b) 2,62%, cấp 2;
(c) 3,41%, cấp 3;
(d) 10,87%, cấp 4.
Các khu vực FLS được sơn màu đỏ trên lá.
Mỗi cấp được thể hiện bằng một hình

Dấu hiệu của phản xạ siêu quang phổ

Hình 4 minh họa các đường cong siêu thị của năm cáp FLS đậu tương sau khi thực hiện hiệu chỉnh đường cơ sở (Phương trình (1)). Chỉ những bước sóng siêu quang phổ âm trong khoảng 450–1.000nm mới được hiển thị vì các vùng ở đầu và cuối của dải bước sóng hoàn chỉnh có tín hiệu nhiễu. Năm đường cong siêu âm này cho thấy các cấu hình và xu hướng tương tự nhau với đỉnh ở khoảng 549nm và vùng trũng ở khoảng 668nm (Hình 4A). Độ phản xạ tăng mạnh từ 668nm và đạt điểm cao nhất ở 770nm, và độ phản xạ tương đối cao được duy trì lên đến 1.000nm.



Hình 4. Đường cong siêu thị của năm cấp FLS đậu tương. (a) Khoảng của Siêu quang phổ 450–1.000nm, (b) dải siêu quang phổ 450–700nm, và (c) dải siêu quang phổ 770–1.000nm.

Sự phản xạ trong dải nhìn thấy (450–700nm) và cận hồng ngoại (770–1.000nm) được phóng to để quan sát chi tiết các đường cong siêu quang phổ (Hình 4B và 4C). Như thể hiện trong Hình 4B, độ bội nhiễm của các lá bị bệnh FLS (loại 1–4) đều cao hơn đáng kể so với các lá khỏe mạnh (loại 0). Hệ số phản xạ trong phạm vi nhìn thấy chủ yếu bị ảnh hưởng bởi hàm lượng chất diệp lục [46]. Các lá bị nhiễm FLS chứa ít chất diệp lục hơn và hấp thụ ít ánh sáng xanh hơn, do đó, có độ phản xạ trong phạm vi nhìn thấy cao hơn các lá khỏe mạnh. Bốn đường cong của các lá bị nhiễm (cấp 1–4) giao nhau trong khoảng 512–525nm, 575–587nm và 680–1.000nm. Hơn nữa, các đường cong phản xạ của các cấp 1, 3 và 4 giao nhau trong khoảng 450–525nm và 575–700nm. Như thể hiện trong Hình 4C, độ phản xạ của lá khỏe mạnh (cấp 0) cao hơn so với lá bị nhiễm bệnh (cấp 1–4) trong vùng cận hồng ngoại. Nguyên nhân chủ yếu là do cấu trúc tế bào lá đậu tương bị hư hỏng hoặc sụp đổ với sự lây lan của FLS. Các tế bào trên lá bị bệnh bị tổn thương và độ phẳng của bề mặt lá bị giảm đi rất nhiều. Do đó, ánh sáng tới sẽ có phản xạ không đều hoặc khuếch tán từ bề mặt lá, làm suy yếu tín hiệu quang phổ mà máy quang phổ nhận được, do đó làm giảm độ phản xạ trong vùng cận hồng ngoại. Những kết quả này tương tự như những phát hiện trước đó [62, 63]. Tuy nhiên, rất khó để phân biệt và phân cấp bằng mắt thường các cấp FLS trong lá đậu tương dựa trên hiện tượng phản xạ siêu quang phổ bao phủ toàn bộ vùng bước sóng 450-1.000nm.

Lựa chọn PC sử dụng PCA

PCA đã được áp dụng cho các đường cong siêu quang phổ để giảm khối lượng dữ liệu và trích xuất thông tin tính năng. Tổng cộng 26 PC hiệu quả đã được trích xuất. Giá trị đặc trưng và tỷ lệ đóng góp tổng thể của chúng được liệt kê trong Bảng 2. Tỷ lệ đóng góp tổng thể của các PC hiệu quả (PC1 – PC26) lớn hơn 94%, cho thấy rằng thông tin chứa trong các bước sóng có thể được giải thích bằng cách nghiên cứu các PC hiệu quả.

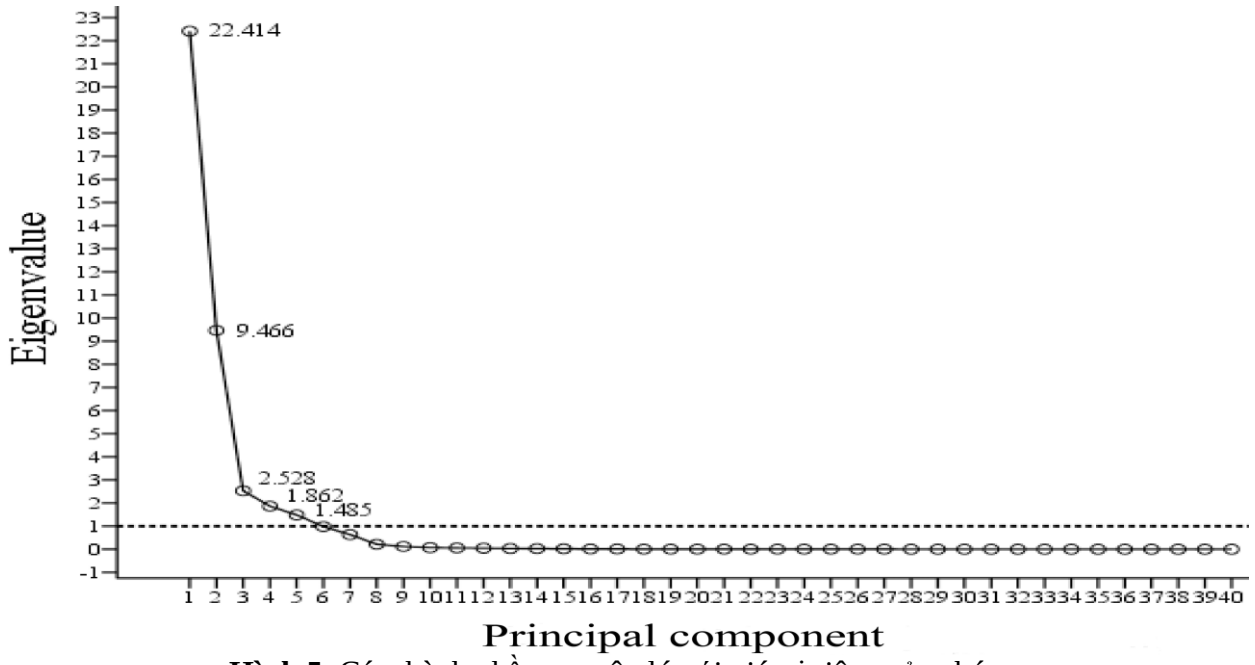
Bảng 2. Các giá trị riêng và tỷ lệ đóng góp tổng thể của các PC hiệu quả.

PCs	Giá trị riêng	Tỷ lệ đóng góp tổng thể	PCs	Giá trị riêng	Tỷ lệ đóng góp tổng thể
PC1	318,79	42,45	PC14	1,53	92,81
PC2	228,32	72,85	PC15	1,46	93,00
PC3	60,23	80,87	PC16	1,41	93,19

PC4	30,29	84,90	PC17	1,39	93,38
PC5	24,00	88,10	PC18	1,35	93,56
PC6	11,34	89,61	PC19	1,28	93,73
PC7	8,26	90,71	PC20	1,23	93,89
PC8	5,35	91,42	PC21	1,20	94,05
PC9	2,12	91,70	PC22	1,17	94,20
PC10	1,83	91,95	PC23	1,15	94,36
PC11	1,75	92,18	PC24	1,13	94,51
PC12	1,63	92,40	PC25	1,03	94,64
PC13	1,56	92,61	PC26	1,00	94,78

Vì số mẫu tối đa trong một cấp là 100 nên số SI tối ưu của cực đại và nhỏ nhất lần lượt là 20 và 10. Các máy tính cá nhân có giá trị riêng lớn hơn 1 và xếp hạng SI thành phần của chúng từ 1 đến 20 được thể hiện trong Hình 5 và Bảng 3.

Các giá trị riêng của năm PC đầu tiên lớn hơn 1, và do đó được sử dụng cho bước sau (Hình 5). 20 SI đầu tiên được liệt kê cho mỗi PC và được chia thành hai nhóm: 10 SI đầu tiên và 10 SI thứ hai theo hệ số trọng lượng của chúng. Trong số 100 SI được chọn trong tất cả các tổ hợp PC khác nhau, SI có tần số cao nhất là SR_5, $NDVI_1$, $NDVI_4$, SR_1, NDI_1 và NDI_4 (Bảng 3). Các SI tần số cao này được liên kết với các bước sóng trong vùng đỏ và đỏ xa (670, 680 và 755nm) và vùng hồng ngoại gần (800, 850 và 900nm) (Bảng 1). Các SI chỉ có một tần số là $NDVI_3$, HI, ARI, $PSSR_b$, PRI_1, PRI_2 và BGI_2 (Bảng 3). Các SI tần số thấp này có một hoặc hai bước sóng trong vùng xanh lục 450–577nm ($NDVI_3$, HI, ARI, PRI_1, PRI_2 và BGI_2) hoặc ở 635nm ($PSSR_b$) (Bảng 1). Chỉ các SI xếp hạng từ 1 đến 20 trong mỗi PC được chọn làm đầu vào trong bộ phân cấp để phân tích dữ liệu sâu hơn vì các SI khác chỉ tương quan một chút với phân cấp tham chiếu.



Hình 5. Các thành phần nguyên lý với giá trị riêng của chúng.

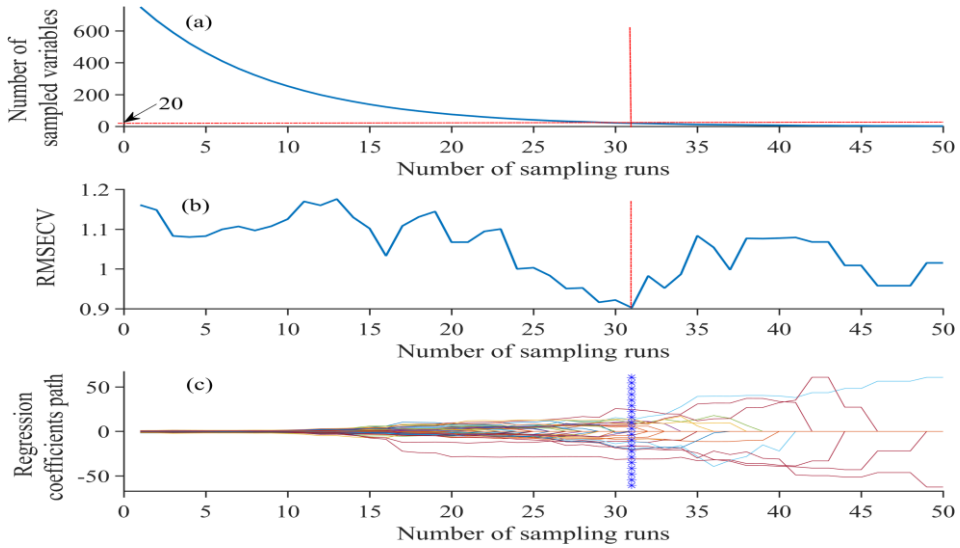
Bảng 3. Năm PC và các SI thành phần của chúng được xếp hạng từ 1 đến 20.

PCs	Xếp hạng SI 1-10	Xếp hạng SI 11-20
PC1	$PSSR^b$, SR_4, $NDVI_5$, NDI_4, OSAVI, NDI_1,	$NDVI_4$, $NDVI_1$, SR_3, RDVI, NDI_5, GNDVI,

	SR_1, $NDVI_2$, NDI_2	NDI_3, Rededge NDVI, SR_2, $PSSR_c$
PC2	RI, GI, DSWI, MCARI, TCARI, $NDVI_6$, MTV_1, SR_6, TVI, NDI_2	$NDVI_2$, $NDVI_1$, $NDVI_4$, NDI_4, NDI_1, SR_1, $NDVI_5$, OSAVI, SR_5, RDVI
PC3	ARI, BGI_1, BRI_1, SR_6, $NDVI_7$, $NDVI_6$, $PSSR_c$, SR_2, GNDVI, SR_1	NDI_5, RI, SR_3, $NDVI_4$, SR_5, SR_1, NDI_1, SR_4, $NDVI_2$, $NDVI_5$
PC4	BRI_1, BGI_1, PRI_1, PRI_2, HI, DSWI, SR_6, $NDVI_6$, BRI_2, GI	TVI, NDI_3, Red-edge NDVI, SR_5, NDI_4, SR_4, NDI_2, BI, MTV_1, SR_1
PC5	BI, BRI_2, MTV_1, BGI_2, TVI, RDVI, BRI_1, BGI_1, OSAVI, $PSSR_b$	TCARI, GI, $NDVI_1$, $NDVI_4$, NDI_1, DSWI, $NDVI_3$, SR_5, $NDVI_7$, NDI_4

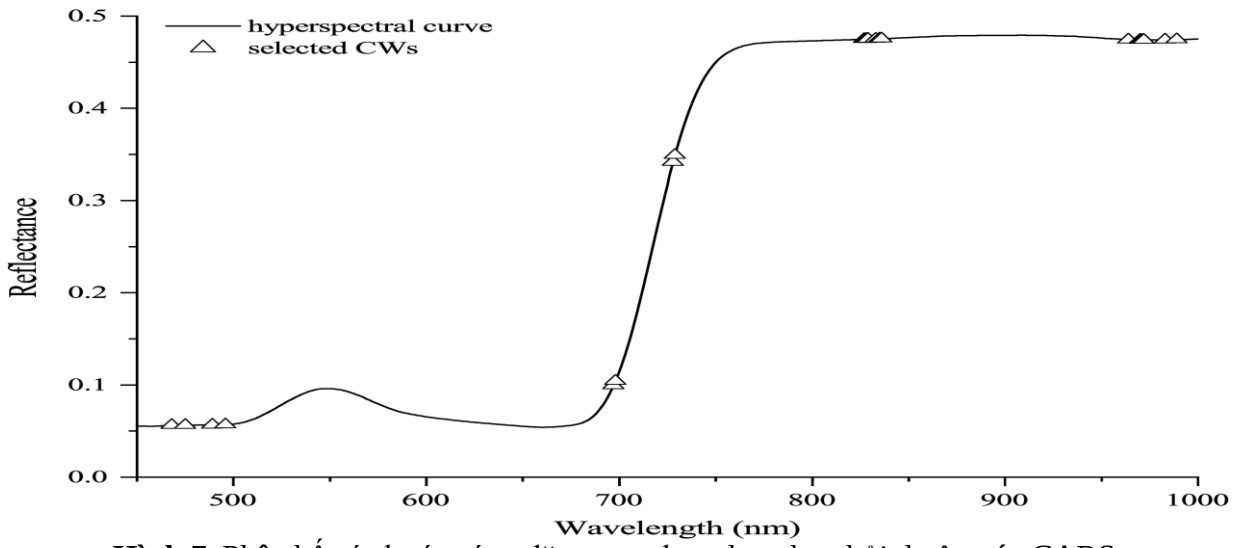
Lựa chọn CW theo CARS

Trong thuật toán CARS, có hai tham số quan trọng: số lượng mẫu Monte Carlo (MCS) và các biến tiềm ẩn để xác nhận chéo. Tham số của số MCS nằm trong khoảng từ 10 đến 100, trong khi tham số của các biến tiềm ẩn dao động từ 1 đến 10 trong nghiên cứu này. Sau một số thử nghiệm, các tham số của CARS được đặt như sau: số MCS là năm mươi, số biến tiềm ẩn tối đa để xác nhận chéo là sáu và "trung tâm" được sử dụng làm phương pháp tiền xử lý. Sau 50 lần chạy, giá trị của sai số trung bình bình phương tối thiểu của xác nhận chéo (RMSECV) đã được trích xuất (Hình 6). Hình 6A trình bày mối quan hệ giữa số lần chạy lấy mẫu và các biến được lấy mẫu dự trữ. Với sự gia tăng số lần chạy lấy mẫu, số bước sóng được chọn giảm và cuối cùng ổn định. Như thể hiện trong Hình 6B và 6C, RMSECV lần đầu tiên được giảm xuống khi các bước sóng không liên quan bị loại bỏ. Tại 31 lần chạy lấy mẫu, RMSECV đạt được giá trị tối thiểu là 0,9026.



Hình 6. Xu hướng thay đổi của thuật toán CARS với sự gia tăng của các lần chạy lấy mẫu. (a) Số lượng biến được lấy mẫu, (b) giá trị RMSECV, (c) hệ số hồi quy của mỗi bước sóng

20 CW (468, 475, 489, 496, 697, 698, 728, 729, 827, 828, 829, 833, 835, 836, 964, 970, 971, 972, 983 và 989nm) đã được chọn để xác định cấp FLS; sự phân bố của chúng được thể hiện trong Hình 7. Số bước sóng giảm sau khi lựa chọn CW sử dụng CARS, điều này làm giảm đáng kể độ phức tạp tính toán.



Hình 7. Phân bố các bước sóng đặc trưng được lựa chọn bởi thuật toán CARS.

Trong số các CW được chọn, các bước sóng 468, 475, 489 và 496nm có thể liên quan đến sự hấp thụ anthocyanin trong mô đậu tương [64]. Ngoài ra, chất diệp lục b (620nm) và chất diệp lục a (675nm) cho thấy các đỉnh đặc trưng, có thể giải thích việc chọn 697, 698, 728 và 729nm làm CW. Trong khi đó, các bước sóng ở 827, 828, 829, 833, 835 và 836nm có thể liên quan đến sự hấp thụ các nguyên tố dinh dưỡng như nitơ và kẽm, trong khi các bước sóng ở 964, 970, 971, 972, 983 và 989nm, tương ứng đến âm bội kéo dài O – H thứ nhất và thứ hai [65], có liên quan đến hàm lượng nước. Những phát hiện này cho thấy CARS rất hữu ích cho việc lựa chọn các bước sóng có liên quan.

Ảnh hưởng của việc giảm kích thước dữ liệu đến độ chính xác của phân cấp

Bảng 4 trình bày 14 bộ dữ liệu (DS1 – DS14) được đặt trực tiếp vào bộ phân cấp SVM và LS-SVM để xây dựng mô hình phân cấp cho FLS. Bước sóng siêu âm đầy đủ ban đầu được gọi là "Thô" trong Bảng 4. Mỗi trong số năm PC chứa hai kết hợp SI (10SI và 20SI). Các mẫu từ mỗi cấp được chia thành bộ huấn luyện và bộ thử nghiệm với tỷ lệ 3: 1. Do đó, có tổng cộng 330 mẫu được chọn làm bộ huấn luyện và 110 mẫu còn lại được sử dụng làm bộ thử nghiệm (Bảng 5).

Bảng 4. 14 bộ dữ liệu được sử dụng để xây dựng mô hình phân cấp SVM và LS-SVM của FLS.

Raw	CWs	PCs	SIs	10SIs					20SIs				
				PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	DS10	DS11	DS12	DS13	DS14

Bảng 5. Tổng số mẫu, bộ huấn luyện và bộ thử nghiệm trong năm cấp.

Phân cấp	Số lượng mẫu	Số lượng bộ huấn luyện	Số lượng bộ thử nghiệm
Cấp 0	100	75	25
Cấp 1	100	75	25
Cấp 2	100	75	25
Cấp 3	80	60	20
Cấp 4	60	45	15
Tổng	**440**	**330**	**110**

Độ chính xác phân cấp của các mô hình SVM và LS-SVM cho cấp tổng thể và cấp riêng lẻ (cấp 0–4) với tổng số 14 bộ dữ liệu, được minh họa và so sánh tương ứng trong Bảng 6

và 7. Các tham số tối ưu trong mô hình SVM và LS-SVM thu được bằng cách sử dụng quy trình tìm kiếm theo lưới (tương ứng Bảng 6 và 7). Các thông số tối ưu được sử dụng để đạt được mô hình huấn luyện tối ưu dự đoán các mẫu thử nghiệm. Phạm vi của γ và σ^2 được đặt trong khoảng 10^{-2}–10^5 trong LS-SVM. Các thông số C và g trong SVM được đặt trong khoảng 2^{-8}–2^8. Tổng cộng có 14 kết quả được đưa vào một cấp. Hiệu suất của các bộ dữ liệu khác nhau thay đổi với nhau.

Bảng 6. Độ chính xác phân cấp chung và phân cấp riêng của các mô hình SVM với các bộ dữ liệu khác nhau.

Dữ liệu (DS)	Tham số (C,g)	Độ chính xác bộ tập huấn (%)						Độ chính xác bộ thử nghiệm (%)					
		Cấp0	Cấp1	Cấp2	Cấp3	Cấp4	Tổng thể	Cấp0	Cấp1	Cấp2	Cấp3	Cấp4	Tổng thể
DS1	(16,21; 0,06)	100,0	100,0	100,0	100,0	100,0	100,0	88,0	100,0	92,0	70,0	86,7	90,9
DS2	(14,35; 3,28)	100,0	100,0	100,0	100,0	100,0	100,0	96,0	100,0	92,0	80,0	93,3	92,7
DS3	(4,05; 2,67)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	92,0	75,0	86,7	91,8
DS4	(4,22; 8,73)	100,0	100,0	100,0	100,0	100,0	100,0	92,0	100,0	96,0	85,0	86,7	91,8
DS5	(8,21; 8,06)	100,0	100,0	100,0	100,0	98,3	99,7	76,0	88,0	88,0	70,0	86,7	81,8
DS6	(64,55; 4,43)	100,0	100,0	98,7	100,0	100,0	99,7	52,0	88,0	72,0	75,0	80,0	72,7
DS7	(16,14; 16,62)	100,0	100,0	98,7	98,3	97,8	99,1	72,0	72,0	68,0	70,0	66,7	72,7
DS8	(64,13; 16,02)	100,0	100,0	98,7	97,5	100,0	99,1	76,0	84,0	64,0	60,0	66,7	70,9
DS9	(4,38; 8,07)	100,0	100,0	100,0	100,0	100,0	100,0	64,0	88,0	80,0	70,0,	73,3	76,4
DS10	(15,51; 2,33)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	92,0	85,0	93,3	94,5
DS11	(4,65; 12,44)	100,0	100,0	100,0	100,0	100,0	100,0	96,0	100,0	96,0	85,0	86,7	93,6
DS12	(32,45; 16,06)	100,0	100,0	100,0	100,0	100,0	100,0	84,0	100,0	92,0	80,0	86,7	89,1
DS13	(8,54; 250,37)	100,0	100,0	100,0	100,0	100,0	100,0	76,0	100,0	92,0	80,0	80,0	86,4
DS14	(16,65; 8,77)	100,0	100,0	100,0	100,0	100,0	100,0	88,0	96,0	72,0	65,0	86,7	81,8

Trong các mô hình SVM (Bảng 6), độ chính xác phân cấp tổng thể của các bộ huấn luyện và bộ thử nghiệm cho các bộ dữ liệu khác nhau thay đổi tương ứng giữa 99,1 – 100% và 70,9 – 94,5%. Nhìn chung, hầu hết các mô hình hoạt động tốt, nhưng các mô hình được xây dựng trên DS10 vượt trội hơn so với các mô hình khác, với độ chính xác phân cấp của bộ huấn luyện và bộ thử nghiệm lần lượt đạt 100% và 94,5%. Mô hình được xây dựng bằng DS8 thu được hiệu suất tương đối kém hơn với độ chính xác tổng thể trong cả bộ huấn luyện và thử nghiệm (tương ứng là 99,1% và 70,9%).

Trong các mô hình LS-SVM (Bảng 7), các cấp 0, 1, 3 và 4 có độ chính xác từng cấp cao nhất (ở mức 100%) trong bộ thử nghiệm, trong khi độ chính xác cao nhất cho cấp 2 là 96%. Trong số tất cả các bộ dữ liệu, độ chính xác thấp nhất của các cấp 0–4 lần lượt là 80%, 76%, 64%, 70% và 73,3%. Khó khăn trong việc phân cấp cấp 2 xảy ra trong PC4 khi sử dụng 10 SI (DS8) và độ chính xác là 64%. Ngược lại, độ chính xác phân cấp của các cấp khác cao hơn 70%.

Bảng 7. Độ chính xác phân cấp chung và phân cấp riêng của các mô hình LS-SVM với các bộ dữ liệu khác nhau.

Dữ liệu (DS)	Tham số (γ, σ^2)	Độ chính xác bộ tập huấn (%)						Độ chính xác bộ thử nghiệm (%)					
		Cấp0	Cấp1	Cấp2	Cấp3	Cấp4	Tổng thể	Cấp0	Cấp1	Cấp2	Cấp3	Cấp4	Tổng thể
DS1	(16,21; 0,06)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	92,0	85,0	86,7	93,6
DS2	(14,35; 3,28)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	96,0	90,0	93,3	96,4
DS3	(4,05; 2,67)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	92,0	85,0	93,3	94,5
DS4	(4,22; 8,73)	100,0	100,0	100,0	100,0	100,0	100,0	96,0	100,0	96,0	90,0	86,7	94,5
DS5	(8,21; 8,06)	100,0	100,0	100,0	100,0	100,0	100,0	84,0	100,0	84,0	75,0	73,3	84,5
DS6	(64,55; 4,43)	100,0	100,0	98,7	100,0	100,0	99,7	84,0	100,0	80,0	80,0	80,0	85,5
DS7	(16,14; 16,62)	100,0	100,0	100,0	100,0	100,0	100,0	80,0	88,0	76,0	80,0	73,3	80,0
DS8	(64,13; 16,02)	100,0	100,0	97,3	100,0	98,3	99,1	90,0	95,0	64,0	75,0	73,3	79,1
DS9	(4,38; 8,07)	100,0	97,3	98,7	100,0	100,0	99,1	92,0	76,0	76,0	75,0	80,0	80,0

DS10	(15,51; 2,33)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	96,0	90,0	100,0	97,3
DS11	(4,65; 12,44)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	96,0	95,0	93,3	97,3
DS12	(32,45; 16,06)	100,0	100,0	100,0	100,0	100,0	100,0	92,0	100,0	88,0	80,0	73,3	88,2
DS13	(8,54; 250,37)	100,0	100,0	100,0	100,0	100,0	100,0	96,0	92,0	88,0	70,0	86,7	87,3
DS14	(16,65; 8,77)	100,0	100,0	98,7	100,0	100,0	99,7	92,0	100,0	84,0	80,0	93,3	90,0

Như thể hiện trong Bảng 7, độ chính xác phân cấp tổng thể của bộ thử nghiệm đối với các bộ dữ liệu khác nhau thay đổi từ 79,1–97,3%. Hầu hết các mô hình đều cho kết quả khả quan khi độ chính xác phân cấp tổng thể lớn hơn 80%. Các mô hình phân cấp LS-SVM được thiết lập bởi DS8 và DS9 có hiệu quả dự đoán tốt nhất, với độ chính xác là 97,3%. So với mô hình được phát triển bằng cách sử dụng đầu vào của bước sóng đầy đủ (DS1), có độ chính xác tổng thể là 93,6%, năm mô hình cho thấy hiệu suất phân cấp tốt hơn; các mô hình này dựa trên DS2, DS3, DS4, DS10 và DS11, và đạt được độ chính xác tổng thể lần lượt là 96,4%, 94,5%, 94,5%, 97,3% và 97,3%. Độ chính xác phân cấp cũng được cải thiện lần lượt là 2,8%, 0,9%, 0,9%, 3,7% và 3,7%.

Cả hai mô hình phân cấp SVM và LS-SVM được phát triển sau khi trích xuất tính năng đều thể hiện hiệu suất tốt hơn với độ chính xác phân cấp cao hơn cho cả cấp tổng thể và cấp riêng lẻ (Bảng 6 và 7). Tất cả ba phương pháp giảm khối lượng dữ liệu đơn lẻ (SI, CARS và PCA) đã cải thiện hiệu suất phân cấp, điều này chứng tỏ tính khả thi của việc sử dụng các phương pháp này trong nghiên cứu này. So sánh sâu hơn đã được thực hiện với phương pháp trích xuất tính năng kết hợp sử dụng PCA và SI. Mười tổ hợp PC và SI (DS5 – DS14) được sử dụng làm đầu vào. Hai đầu vào dựa trên DS10 và DS11 có độ chính xác cao hơn cho cả cấp tổng thể và cấp riêng lẻ. 40 SI tích lũy của DS8 và DS9 được phân tích thêm, cho thấy có tới 10 SI ($NDVI_1$, $NDVI_2$, $NDVI_4$, $NDVI_5$, RDVI, SR_1, SR_5, OSAVI, NDI_2 và NDI_4) là trùng hợp (Bảng 3). Có 5 SI tần số cao và không có tần số thấp được bao gồm trong số đó. Ngoài ra, bước sóng của 10 SI này chủ yếu nằm trong vùng đỏ và vùng cận hồng ngoại 660–855nm với 670, 680 và 800nm là các bước sóng phổ biến nhất (Bảng 1). Các bước sóng này cũng liên quan đến sức sống của thực vật [66]. Kết quả cho thấy không chỉ các bước sóng đỏ và cận hồng ngoại chứa nhiều thông tin về tính năng hơn mà còn chịu trách nhiệm về hiệu quả của SI tần số cao đối với việc phân cấp FLS đậu tương trong nghiên cứu này.

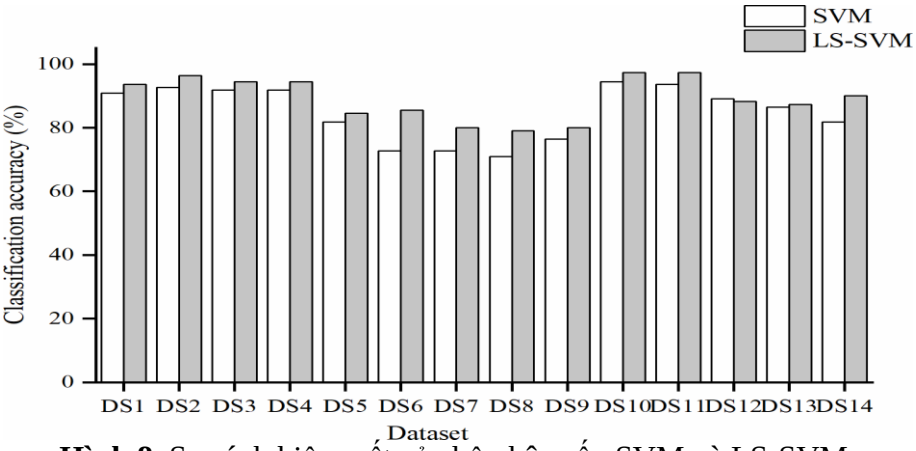
Cả hai mô hình phân cấp SVM và LS-SVM có hiệu suất tổng thể kém nhất đều được xây dựng bằng cách sử dụng cùng một bộ dữ liệu (DS8) và có độ chính xác tổng thể lần lượt chỉ là 70,9% và 79,1%. Tám SI (BGI_1, PRI_1, PRI_2, HI, DSWI, SR_6, $NDVI_6$ và GI) có ít nhất một bước sóng trong vùng xanh lục (500–570nm) trong số 10 SI được sử dụng làm đầu vào (Bảng 1 và 3). Tuy nhiên, những bước sóng này có liên quan đến các sắc tố, chẳng hạn như carotenoid và chất diệp lục [36]. Một nghiên cứu cho thấy không có sự cải thiện nào trong các mô hình khi sử dụng các vùng quang phổ xanh lam và xanh lục cho thông tin đầu vào [18]. Trong nghiên cứu này, chúng tôi nhận thấy rằng hệ số phản xạ siêu quang phổ của lá trong vùng xanh tương quan kém với phân cấp FLS của đậu tương. Điều này càng được khẳng định khi các bước sóng trong vùng màu xanh lá cây không được CARS chọn làm CW (Hình 7). Mối tương quan kém này cũng có thể là lý do khiến kết quả phân cấp chính xác tổng thể là thấp nhất đối với DS8.

Các kết quả của nghiên cứu này chứng minh rằng sự kết hợp của các phương pháp trích xuất tính năng (kết hợp PCA và SI) cải thiện hiệu suất phân cấp cho cả hai mô hình SVM và LS-SVM và cho phép phân cấp chính xác cao nhất cho cả cấp tổng thể và riêng lẻ. Phương pháp kết hợp này tận dụng lợi thế của cả PC và SI hiệu quả hơn là chỉ một loại phương pháp giảm kích thước dữ liệu, chẳng hạn như chỉ CW, PC hoặc SI. Do đó, sự kết

hợp giữa PCA và SI là một phương pháp hiệu quả để phân cấp FLS đậu tương và là một phương pháp đầy hứa hẹn cho các nghiên cứu trong tương lai.

So sánh mô hình phân cấp SVM và LS-SVM

Kết quả của các mô hình SVM và LS-SVM cùng với các phương pháp trích xuất tính năng khác nhau được minh họa trong Hình 8. So sánh khả năng phân cấp FLS, chúng tôi nhận thấy rằng các mô hình LS-SVM thường cho độ chính xác phân cấp cao hơn so với các mô hình SVM và luôn vượt trội hơn SVM trong phân cấp FLS đậu tương ở nghiên cứu này. LS-SVM sử dụng tốt hơn thông tin phi tuyến tiềm ẩn của dữ liệu siêu quang phổ, có thể đã góp phần vào hiệu suất dự đoán tốt hơn của nó.



Hình 8. So sánh hiệu suất của bộ phân cấp SVM và LS-SVM.

KẾT LUẬN

Tóm lại, nghiên cứu này đã chứng minh tiềm năng và tính khả thi của việc ước lượng cấp FLS đậu tương bằng kỹ thuật siêu quang phổ. Phương pháp CARS, SI, PCA và phương pháp kết hợp PCA và SI được sử dụng để chọn 20 băng tần quan trọng (CW), 26 PC hiệu quả và SI hiệu quả hơn để phân biệt các cấp FLS khác nhau. Các phương pháp này trích xuất thông tin nhạy cảm liên quan đến các cấp FLS. Hơn nữa, sự kết hợp giữa PCA và SI có thể trích xuất thông tin hiệu quả hơn từ FLS so với các phương pháp đơn lẻ. Các mô hình phân cấp LS-SVM tạo ra hiệu suất khả quan hơn sau khi giảm kích thước của dữ liệu siêu quang phổ so với các mô hình SVM. Độ chính xác của phân cấp (cho cả cấp tổng thể và cấp riêng lẻ) sử dụng DS2, DS3, DS4, DS10 và DS11 là hơn 90% và việc sử dụng các bộ dữ liệu này có lợi hơn so với việc sử dụng bộ dữ liệu siêu quang phổ hoàn chỉnh (DS1). Các mô hình kết hợp PC1-20SIs-LS-SVM và PC2-20SIs-LS-SVM, cả hai đều có độ chính xác phân cấp tổng thể là 97,3%, thể hiện hiệu suất tốt nhất trong số tất cả các mô hình được xây dựng bởi SVM và LS-SVM. Hơn nữa, dữ liệu từ vùng đỏ và vùng cận hồng ngoại có hiệu quả trong việc phân cấp bệnh FLS. Kết quả của chúng tôi cung cấp một tài liệu tham khảo lý thuyết để cải thiện hệ thống giám sát bệnh.

THÔNG TIN BỔ SUNG

S1 Data. Data of Fig 1.

https://doi.org/10.1371/journal.pone.0257008.s001 (XLSX)

S2 Data. Data of Fig 5.

https://doi.org/10.1371/journal.pone.0257008.s002 (XLSX)

S3 Data. Values of 40SIs in classes 0–4.

https://doi.org/10.1371/journal.pone.0257008.s003 (XLSX)

S4 Data. Siêu quang phổ data used in this study.

https://doi.org/10.1371/journal.pone.0257008.s004 (XLSX)