MÔ HÌNH TỰ ĐỘNG PHÂN LOẠI DỮ LIỆU LỚP PHỦ BỀ MẶT PHỤC VỤ KIỂM KÊ KHÍ NHÀ KÍNH BẰNG ẢNH VIỄN THÁM

NÔNG THỊ OANH[(1)], TRẦN XUÂN TRƯỜNG[(1)]
TẠ HOÀNG TRUNG[(2)], TRỊNH VIỆT NGA[(3)]
[(1)]Trường Đại học Mỏ - Địa chất
[(2)]Cục Đo đạc, Bản đồ và Thông tin địa lý Việt Nam
[(3)]Cục Viễn thám Quốc gia

Tóm tắt:

Gần đây, việc giám sát phát thải khí nhà kính đang nhận được sự quan tâm của các nhà khoa học, các nhà hoạch định chính sách và các cơ quan. Nhằm tính toán mức độ phát thải khí nhà kính, một bản đồ phủ bề mặt chính xác khu vực cần đánh giá là rất cần thiết. Nghiên cứu này sử dụng thuật toán Random Forest, ảnh vệ tinh Landsat 8, Sentinel 1,2, dữ liệu mô hình số địa hình xây dựng mô hình phân loại ảnh viễn thám thành lập bản đồ phủ bề mặt theo khuyến cáo của AFOLU phục vụ kiểm kê khí nhà kính. Kết quả thử nghiệm phân loại ảnh viễn thám trên khu vực Tây Bắc và Tây Nam Bộ cho kết quả độ chính xác tổng thể đạt lần lượt 83% và 80.5%, phù hợp cho xây dựng bản đồ phủ bề mặt phục vụ kiểm kê khí nhà kính.

Keywords: Ảnh viễn thám, phủ bề mặt, Random Forest, AFOLU.

1. Đặt vấn đề

Biến đổi khí hậu và nước biển dâng đã và đang có các ảnh hưởng tiêu cực trực tiếp tới cuộc sống của con người bao gồm xói lở, trượt đất, sụt lún bề mặt, sa mạc hóa hoặc ngập lụt (IPCC, 2006). Nếu không có các biện pháp thích ứng và giảm thiểu phù hợp, ước tính biến đổi khí hậu sẽ khiến Việt Nam mất khoảng 12% đến 14,5% GDP mỗi năm vào năm 2050 và có thể khiến tới một triệu người vào tình trạng nghèo cùng cực vào năm 2030 [1]. Các hoạt động của còn người là nguyên nhân chính phát thải khí nhà kính, như hoạt động đi lại, sản xuất công nghiệp, nông nghiệp và chặt phá rừng [2], các hoạt động này hầu hết đều có tác động đến lớp phủ bề mặt của trái đất. Do đó, giám sát lớp phủ bề mặt trái đất liên tục là một yêu cầu quan trọng nhằm giám sát lượng phát thải khí nhà kính. Bản đồ phủ bề mặt cung cấp dữ liệu về diện tích của từng loại phủ bề mặt, từ đó làm căn cứ để tính toán lượng phát thải, hoặc hấp thụ khí nhà kính của mỗi khu vực cụ thể. Vì vậy, trong tính toán, giám sát phát thải khí nhà kính, việc thành lập bản đồ phủ bề mặt chính xác, theo đúng các lớp đã được IPCC khuyến cáo rất cần thiết.

Để phục vụ cho mục đích giám sát phát thải khí nhà kính, nhiều tác giả đã sử dụng các thuật toán học máy để phân loại ảnh viễn thám, từ đó tạo ra các bản đồ phủ bề mặt phục vụ giám sát việc phát thải khí nhà kính. Tác giả Doãn Hà Phong và nkk đã nghiên cứu sử

Ngày nhận bài: 1/7/2023, ngày chuyển phản biện: 5/7/2023, ngày chấp nhận phản biện: 9/7/2023, ngày chấp nhận đăng: 28/7/2023

dụng thuật toán Random Forest và ảnh Sentinel-1 phân loại rừng khu vực tỉnh Quảng Bình, kết quả phân loại đạt độ chính xác khoảng 90% [3]. Mặc dù đạt độ chính xác cao, nhưng nghiên cứu mới được thực hiện trên phạm vi khá nhỏ, đối tượng phủ bề mặt của khu vực phân loại chủ yếu là rừng, nên khó đánh giá được độ chính xác của mô hình khi áp dụng trên phạm vi rộng hơn, ở các khu vực có nhiều đối tượng phủ bề mặt hơn. Thêm vào đó, nghiên cứu mới chỉ phân loại 2 lớp là rừng và không phải là rừng nên khả năng đạt được độ chính xác cao cũng lớn hơn. Đối với các nghiên cứu có yêu cầu phân loại nhiều lớp phủ, tác giá Phạm Thị Làn và nnk sử dụng ảnh VNREDSat -1 và phương pháp phân loại dựa trên đối tượng để thành lập bản đồ phủ bề mặt khu vực thành phố Uông Bí, tỉnh Quảng Ninh. Sản phẩm phân loại gồm 14 lớp phủ với độ chính xác 81% [4], hoặc tác giả Vu Thi Thu và nnk sử dụng ảnh Landsat và thuật toán Maximum Likelihood thành lập bản đồ phủ bề mặt khu vực thành phố Đông Triều, tỉnh Quảng Ninh [5]. Trên phạm vi rộng hơn, tác giả Trần Tuấn Ngọc và nnk sử dụng 12 cảnh ảnh Landsat 7+ ETM và thuật toán Maximum Likelihood để thành lập bản đồ phủ bề măt khu vực bắc Lào, kết quả đạt độ chính xác tổng thể 75%, hệ số Kappa đạt 0.79 [6]. Độ chính xác của sản phẩm không được cao như các nghiên cứu khác có thể do độ chính xác của thuật toán sử dụng không được cao khi so với các thuật toán khác như SVM hoặc Random Forest [7]. Nhìn chung Các kết quả nghiên cứu đã ứng dụng thuật toán máy học cho công tác phân loại ảnh viễn thám tuy nhiên, các nghiên cứu có kết quả phân loại đạt độ chính xác cao trên 80% thường chỉ được thử nghiệm trên một phạm vi nghiên cứu tương đối nhỏ, hoặc có ít lớp cần phân loại. Các khu vực được lựa chọn thử nghiệm đều tương đối đồng nhất về địa hình, ví dụ như khu vực đô thị, hoặc khu vực có phần lớn diện tích là địa hình núi cao. Việc thực nghiệm ở các khu vực tương đối đồng nhất về địa hình như trên dẫn đến vấn đề kiểm nghiệm độ chính xác của mô hình trên các khu vực có điều kiện địa hình khác nhau gặp khó khăn, ví dụ mô hình có thể cho độ chính xác tốt trên khu vực đồng bằng, nhưng trên khu vực miền núi lại cho độ chính xác không cao. Thêm vào đó, các nghiên cứu chỉ sử dụng một loại ảnh vệ tinh duy nhất, mà không sử dụng thêm dữ liệu độ cao, địa hình làm dữ liệu đầu vào cho mô hình phân loại, có thể ảnh hưởng đến độ chính xác của mô hình.

Để phục vụ kiểm kê khí nhà kính tại Việt Nam được nhanh chóng, kịp thời, đáp ứng các nhu cầu về ứng phó với biến đổi khí hậu, nhóm nghiên cứu đã tiến hành thử nghiệm mô hình phân loại sử dụng kết hợp ảnh vệ tinh Landsat 8, ảnh vệ tinh Sentinel 1, 2 và dữ liệu mô hình số địa hình thành lập bản đồ phủ bề mặt phục vụ kiểm kê khí nhà kính theo khuyến nghị của AFOLU, gồm 10 lớp phân loại. Mô hình phải có khả năng áp dụng trên các khu vực có điều kiện địa hình khác nhau để có thể phục vụ thành lập bản đồ phủ bề mặt phục vụ kiểm kê khí nhà kính trên phạm vi cả nước, và có độ chính xác trên 80%.

2. Dữ liệu và phương pháp nghiên cứu

2.1. Cơ sở lý luận

Hiện nay có rất nhiều thuật toán được sử dụng để phân loại ảnh viễn thám như Random forest (RF), Support Vector Machine (SVM), và k-Nearest Neighbor (k-NN). Mặc dù đã có nhiều nghiên cứu, so sánh độ chính xác của các thuật toán trên để tìm ra một thuật toán tối ưu, nhưng kết quả của các nghiên cứu này lại không thống nhất với nhau [8]. Do mỗi thuật toán có các ưu điểm khác nhau, việc lựa chọn

thuật toán phân loại cần phải được thử nghiệm kỹ càng.

Phương pháp rừng cây phân loại (Random Forest) là một bộ phân loại học máy tập hợp sử dụng nhiều cây quyết định, mỗi cây trong đó được cung cấp các tập con ngẫu nhiên của dữ liệu huấn luyện để hình thành các cây phân loại. Các cây quyết định trong tập hợp được tạo ra một cách độc lập. Sau khi được huấn luyện, mỗi điểm ảnh chưa biết được phân loại dựa trên đa số phiếu bầu của cây phân loại [9]. RF được sử dụng phổ biến trong phân loại ảnh viễn thám bởi kết quả phân loại độ chính xác phân loại tốt hơn của nó so với các thuật toán chỉ sử dụng một cây phân loại duy nhất, dễ tham số hóa và ít bị ảnh hưởng bởi nhiễu, phù hợp cho các khu vực khó lấy được bộ mẫu phân loại tốt [7], [10], [11]. SVM là một thuật toán học máy có giám sát, phi tham số nhằm tìm kiếm ranh giới siêu phẳng để phân tách các lớp dữ liệu. Dữ liệu được phân tách dựa trên đường thẳng tuyến tính, phân chia 2 lớp dữ liệu [12], [13]. K-Nearest Neighbors (k-NN) là một bộ phân loại không tham số, gán nhãn phân loại cho dữ liệu đầu vào dựa trên độ gần của chúng với k dữ liệu mẫu đã được gắn nhãn trước gần nhất trong không gian véc tơ thuộc tính. k-NN thường được mô tả là một thuật toán phân loại lười học vì nó không được đào tạo; dữ liệu chưa biết được so sánh trực tiếp với dữ liệu huấn luyện.

2.2. Dữ liệu sử dụng

Các dữ liệu được sử dụng để xây dựng mô hình gồm có ảnh vệ tinh Sentinel- 2, độ phân giải 10 m có độ che phủ mây nhỏ hơn 20% được thu thập trong năm 2022, sau đó lấy giá trị median trong của từng pixel để loại bỏ vùng bị che phủ bởi mây, ảnh vệ tinh Sentinel-1, phân cực HH, HV (H: horizontal - phân cực ngang; V: vertical - phân cực đứng), dữ liệu mô hình số độ cao cung cấp bởi ESA, và ảnh Landsat 8 OLI TOA.

Để nâng cao độ chính xác phân loại, các chỉ số thực vật được tính toán từ các kênh của ảnh vệ tinh quang học, sau đó các chỉ số này được gộp chung với dữ liệu ảnh ban đầu, tạo thành bộ dữ liệu đầu vào cho mô hình phân loại. Ví dụ, mô hình phân loại sử dụng 9 kênh ảnh Landsat, gộp thêm 4 chỉ số được tính toán thêm, tổng cộng mô hình phân loại sử dụng 13 kênh của ảnh Landsat làm dữ liệu đầu vào.

Bảng 1: Các sản phẩm ảnh sử dụng xây dựng mô hình

STT	Loại ảnh	Tên sản phẩm	Mô tả
1	Sentinel-1	Level 1 SLC	Loại bỏ sai số do hiệu ứng Doppler, và nắn chỉnh hình học
2	Sentinel-2	Level 1C	Hiệu chỉnh sai số khí quyển (Top of Atmosphere), và nắn chỉnh hình học
3	Landsat-8	Level 1	Hiệu chỉnh sai số do địa hình, khí quyển
4	DEM	Copernicus DSM	Mô hình số mặt đất

Các chỉ số NDWI (chỉ số nước khác biệt chuẩn hóa), và NDPI (chỉ số ao hồ khác biệt chuẩn hóa) được tính toán nhằm nâng cao độ chính xác khi phân loại các đối tượng là nước mặt, ao hồ [14]. Trong khi chỉ ố NDVI (chỉ số thực vật khác biệt chuẩn hóa) được tính toán để nâng cao khả năng phân biệt các đối tượng là thực vật [15]. Chỉ số công trình khác biệt

chuẩn hóa (NDBI) được tính toán cho việc phân biệt các khu vực được xây dựng bởi con người [16]. Bên cạnh đó, ngoài các kênh ảnh trong dải nhìn thấy được sử dụng, các kênh hồng ngoại nhiệt cũng được thử nghiệm sử dụng trong mô hình do chúng có khả năng phân biệt các đối tượng thay đổi theo mùa như đất sử dụng để trồng trọt, hoặc thực vật thay lá [17], [18].

Bảng 2: Công thức tính các chỉ số chuẩn hóa

STT	Tên chỉ số	Công thức	Landsat 8	Sentinel-2
1	NDVI	$NDVI = \frac{NIR-Red}{NIR+Red}$	$NDVI = \frac{B5 - B4}{B5 + B4}$	$NDVI = \frac{B7 - B4}{B7 + B4}$
2	NDWI	$NDWI = \frac{Green-NIR}{Green+NIR}$	$NDWI = \frac{B3 - B5}{B3 + B5}$	$NDWI = \frac{B3 - B7}{B3 + B7}$
3	NDPI	$NDPI = \frac{NIR - (0.74\,x\,Red + 0.26\,x\,SWIR1)}{NIR + (0.74\,x\,Red + 0.26\,x\,SWIR1)}$	$NDPI = \frac{B5 - (0.74\,x\,B4 + 0.26\,x\,B6)}{B5 + (0.74\,x\,B4 + 0.26\,x\,B6)}$	$NDPI = \frac{B7 - (0.74\,x\,B4 + 0.26\,x\,B11)}{B7 + (0.74\,x\,B4 + 0.26\,x\,B11)}$
4	NDBI	$NDBI = \frac{NIR-SWIR}{NIR+SWIR}$	$NDBI = \frac{B5-B6}{B5+B6}$	$NDBI = \frac{B7-B11}{B7+B11}$

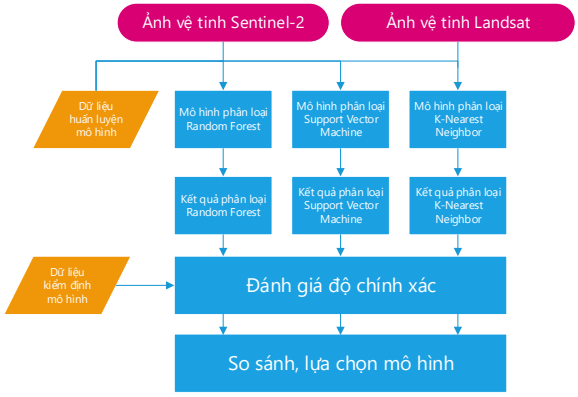
2.3. Phương pháp nghiên cứu

2.3.1. Khu vực nghiên cứu

Các khu vực được lựa chọn để thử nghiệm phân loại gồm 2 khu vực có đặc điểm địa hình, khí hậu rất khác nhau là khu vực Tây Bắc bao gồm tỉnh Điện Biên, Lai Châu và Sơn La và khu vực đồng bằng sông Cửu Long. Khu vực Tây Bắc có địa hình phức tạp, chủ yếu là đồi núi dốc, hiểm trở và chia cắt mạnh, được cấu tạo bởi những dãy núi chạy dài theo hướng Tây Bắc - Đông Nam với độ cao biến đổi từ 200 m đến hơn 1.800 m. Địa hình thấp dần từ Bắc xuống Nam và nghiêng dần từ Tây sang Đông. Vùng đồng bằng sông Cửu Long (còn được gọi là, Vùng Tây Nam Bộ, Cửu Long, hay Miền Tây) là vùng cực nam của Việt Nam. Khu vực này có 1 thành phố trực thuộc Trung ương là TP. Cần Thơ và 12 tỉnh: Long An, Tiền Giang, Bến Tre, Vĩnh Long, Trà Vinh, Hậu Giang, Sóc Trăng, Đồng Tháp, An Giang, Kiên Giang, Bạc Liêu và Cà Mau. Đồng bằng sông Cửu Long có tổng diện tích 40.577,6 km² và có tổng dân số là 17.744.947 người (2022). Vùng chiếm 12,8% diện tích cả nước nhưng có 17,9% dân số cả nước. Đồng bằng sông Cửu Long gồm ba tiểu vùng. Vùng cao ở phía tây, đây là vùng thường bị ngập vào mùa mưa bởi nước sông Cửu Long dâng lên, vùng thấp ở duyên hải phía đông, đây là vùng thường bị mặn xâm nhập vào mùa khô.

2.3.2. Sơ đồ tổng quát

Trước tiên, mẫu huấn luyện mô hinh cho khu vực 3 tỉnh Tây Bắc (Điện Biên, Lai Châu, Sơn La) và khu vực Tây Nam Bộ (đồng bằng Sông Cửu Long) được thành lập. Mẫu huấn luyên phục vụ thử nghiệm mô hình được chọn bằng cách đoán đọc trên ảnh vệ tinh độ phân giải cao. Sau đó mẫu này được sử dụng để thử nghiệm trên các mô hình phân loại Random Forest, Support Vector machine và k-Nearest neighbour. Kết quả phân loại sau đó được so sánh độ chính xác. Mô hình phân loại nào cho ra kết quả độ chính xác tổng hợp cao nhất sẽ được chọn để làm mô hình phân loại ảnh phục vụ kiểm kê khí nhà kính. Sơ đồ tổng thể công tác chọn mô hình phân loại được trình bày trong hình 1.



Hình 1: Sơ đồ quy trình hiệu chỉnh mô hình phân loại

Các thử nghiệm mô hình phân loại lần lượt sử dụng ảnh quang học, sau đó sử dụng ảnh quang học kết hợp với tính thêm các chỉ số thực vật, và cuối cùng sử dụng kết hợp ảnh quang học, ảnh radar và dữ liệu trích xuất từ mô hình số địa hình đưa vào mô hình phân loại. Trong thử nghiệm cuối cùng, ngoài 14 kênh ảnh của ảnh Landsat và Sentinel-2, nhóm nghiên cứu sử dụng thêm kênh ảnh VV, và VH của ảnh Sentinel-1. Để tăng khả năng phân loại các đối tượng thay đổi theo mùa, ví dụ như lúa, hoặc các cây trồng ngắn ngày, cũng như tận dụng lợi thế không bị ảnh hưởng bởi mây của ảnh radar, đối với dữ liệu ảnh Sentinel-1, dữ liệu ảnh Setninel-1 trong 1 năm được chia thành 3 nhóm, mỗi nhóm 4 tháng (từ tháng 1-4, từ tháng 5-8 và từ tháng 8-12) nhằm ghi nhận sự thay đổi của các loại phủ bề mặt thay đổi theo mùa. Từ tập hợp ảnh của mỗi nhóm này, tổng hợp thành một ảnh duy nhất, đại diện cho mỗi nhóm. Chi tiết các trường hợp thử nghiệm được trình bày trong bảng 2.

Bảng 3: Chi tiết các trường hợp thử nghiệm

STT	Danh sách kênh ảnh tham gia vào quá trình phân loại	Thuật toán		
		RF	SVM	k-NN
I	**Sử dụng ảnh Landsat**			
1	Landsat 8: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11	L08B-RF	L08B-SVM	L08B-kNN
2	Landsat 8: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11 Chỉ số: NDVI, NDBI, NDPI, NDWI	L12B-RF	L12B-SVM	L12B-kNN
3	Landsat 8: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11 Chỉ số: NDVI, NDBI, NDPI, NDWI Thông tin độ dốc, độ cao	L15B-RF	L15B-SVM	L15B-kNN
II	**Sử dụng ảnh Sentinel-2**			
5	Sentinel-2: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11	S08B-RF	S08B-SVM	S08B-kNN
6	Sentinel-2: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11 Chỉ số: NDVI, NDBI, NDPI, NDWI	S12B-RF	S12B-SVM	S12B-kNN
7	Sentinel-2: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11 Chỉ số:NDVI, NDBI, NDPI, NDWI Thông tin độ dốc, độ cao	S15B-RF	S15B-SVM	S15B-kNN
III	**Sử dụng kết hợp ảnh Landsat 8, Sentinel-2 và Sentinel-1**			
	Landsat 8: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11 Chỉ số: NDVI, NDBI, NDPI, NDWI	LS-RF	LS-SVM	LS-kNN

STT	Danh sách kênh ảnh tham gia vào quá trình phân loại	Thuật toán		
		RF	SVM	k-NN
	Sentinel-2: Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, Kênh 10, Kênh 11 Chỉ số: NDVI, NDBI, NDPI, NDWI Sentinel-1: VH, VV Thông tin độ dốc, độ cao			

2.3.3. Đồ hình phân loại

Theo khuyến nghị của AFOLU, 10 lớp dữ liệu sau được chọn để tiến hành phân loại ảnh, xây dựng dữ liệu phủ bề mặt. Đồ hình phân loại này được áp dụng chung cho cả 2 khu vực thử nghiệm nhằm kiểm tra tính thích ứng của mô hình phân loại đối với các khu vực có điều kiện địa hình khác nhau và có các loại phủ bề mặt khác nhau.

Bảng 4: Đồ hinh phân loại

STT	Tên lớp
1	Rừng ngập mặn
2	Rừng thường xanh
3	Rừng trồng
4	Lúa
5	Các loại cây trồng ngắn ngày gồm cả Cỏ dùng để chăn nuôi
6	Cây trồng lâu năm
7	Có mọc hoang dã, không có sự chăm sóc và thu hoạch
8	Gồm tất cả các loại đất cho mục đích đât ở và CSHT, cả những khu khai thác khoáng sản
9	Nuôi trồng thủy sản, sông hồ, ao …..
10	Bao gồm các khu vực, nơi lớp phủ mặt đất là đất trống, đồi trọc, đá sỏi, cát.., kể cả khu vực băng tuyết

2.3.4. Tạo dữ liệu huấn luyện và kiểm định mô hình

Trong nghiên cứu này, mẫu huấn luyện và kiểm định mô hình được đoán đọc từ ảnh vệ tinh độ phân giải cao, cung cấp bởi Google Earth. Lần lượt 10, 20, 50 khu vực trên mỗi lớp được lấy mẫu để đưa vào mô hình phân loại và đánh giá độ chính xác. Trong bộ dữ liệu này 70% số mẫu dùng để phân loại, và 30% số mẫu dùng để kiểm định mô hình.

3. Kết quả nghiên cứu

3.1. So sánh kết quả các trường hợp thử nghiệm

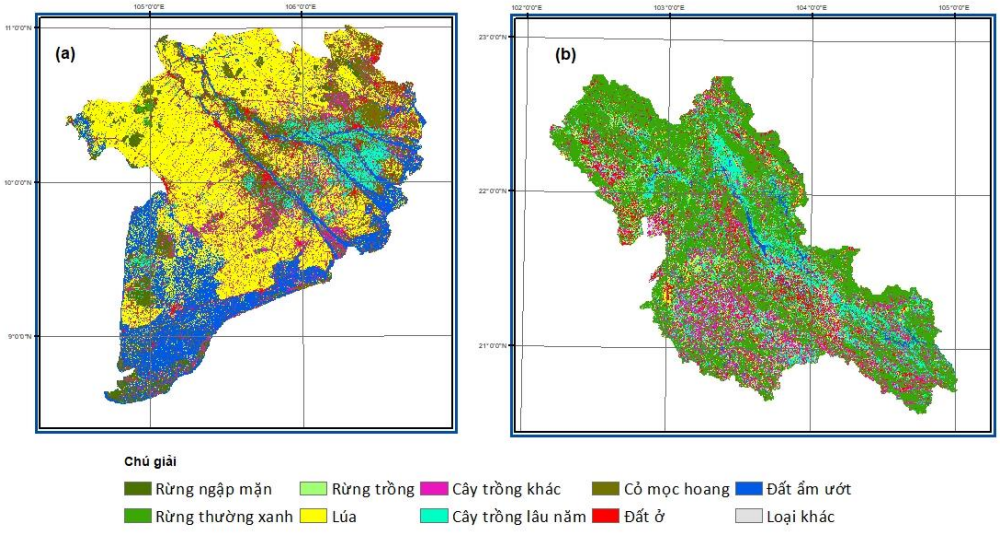
Khi sử dụng nhiều kênh ảnh, độ chính xác của mô hình có xu hướng tăng lên đáng kể, xu hướng này phù hợp trong cả 3 thuật toán thử nghiệm là RF, SVM và k-NN. Khi sử dụng 8 kênh ảnh của ảnh Landsat và Sentinel-2, bao gồm Kênh 2, Kênh 3, Kênh 4, Kênh 5, Kênh 6, Kênh 7, và Kênh 11 độ chính xác tổng thể của mô hình đạt khoảng 50(±2)%. Sau đó, khi bổ sung thêm các chỉ số thực vật bao gồm NDVI, NDPI, NDWI và NDBI, tương ứng với tổng số kênh sử dụng là 13 kênh (9 kênh của ảnh gốc, và 4 kênh chứa thông tin các chỉ số thực vật), chỉ số độ chính xác tổng thể của ảnh Landsat giảm đi không đáng kể (1% ,từ 50% xuống 49%), trong khi ngược lại, độ chính xác tổng thể của ảnh Sentinel-2 lại có xu hướng tăng mạnh (tăng gần 5%). Cuối cùng, khi sử dụng thêm dữ liệu độ dốc, và độ cao, tương

ứng với việc đưa 19 kênh ảnh vào mô hình, độ chính xác cho sự cải thiện đáng kể, đặc biệt đối với ảnh Landsat khi độ chính xác tổng thể mô hình đạt được tăng tới gần 10%, đạt 80%.

Đối với các thuật toán, khi so sánh độ chính xác của các thuật toán, cả 3 thuật toán có độ chính xác tương đương nhau khi sử dụng số lượng mẫu lớn nhất (50 vùng), và đều đạt độ chính xác trên 0.8. Tuy nhiên có một sự khác biệt khi số lượng mẫu sử dụng nhỏ (20 vùng), khi đó RF cho ra kết quả độ chính xác tổng thể tốt nhất (0.62) trong khi k-NN cho ra kết quả thấp nhất (0.46) .

3.2. Bản đồ phủ bề mặt khu vực Tây Bắc và Tây Nam Bộ

Sau khi thử nghiệm các mô hình phân loại, mô hình phân loại sử dụng thuật toán Random Forest, sử dụng 19 kênh ảnh đã được lựa chọn để thành lập bản đồ phủ bề mặt khu vực Tây Bắc và Tây Nam Bộ. Mô hình phân loại cho ra kết quả tương đối chính xác ở cả 2 khu vực phân loại có đặc điểm địa hình khác nhau. Trong khi tại khu vực Tây Nam Bộ, rừng ngập mặn được phân bố nhiều, chủ yếu tại khu vực Cà Mau, Kiên Giang, và hầu như không có rừng thường xanh, thì tại khu vực Tây Bắc, kết quả chủ yếu là rừng thường xanh, không có rừng ngập mặn. Kết quả phân loại này phù hợp với thực địa, và kết quả giải đoán ảnh bằng mắt, do khu vực Tây Nam Bộ là khu vực ven biển, nơi phân bố rừng ngập mặn chủ yếu của Việt Nam.



Hình 2: Kết quả phân loại khu vực Tây Nam Bộ (a) và Tây Bắc (b)

Kết quả đánh giá độ chính xác cho thấy, độ chính xác phân loại tổng thể đạt 80.5%, trong đó một số loại như đất âm ướt, cỏ mọc hoang có độ chính xác nhà sản xuất (producer accuracy) cao. Độ chính xác nhà sản xuất thể hiện tỷ lệ phần trăm số pixel được phán ảnh đúng trên bản đồ, ví dụ trên thực địa là đất ẩm ướt, trên bản đồ phân loại đúng là đất ẩm ướt. Tỷ lệ chính xác cao của đất ẩm ướt cho thấy các khu vực là đất ẩm ướt trên thực địa được thể hiện đúng trên bản đồ với tỷ lệ cao.

Bảng 6: Ma trận đánh giá độ chính xác khu vực Tây Nam Bộ

LSR31B		Dữ liệu tham chiếu											
		Rừng ngập mặn	Rừng thường xanh	Rừng trồng	Lúa	Cây trồng khác	Cây trồng lâu năm	Cỏ mọc hoang	Đất ở	Đất ẩm ướt	Khác	Tổng	*User accuracy*
Dữ liệu phân loại	Rừng ngập mặn	**111**	0	0	0	8	8	0	0	0	0	127	87.4
	Rừng thường xanh	**7**	**22**	0	0	6	0	0	0	0	0	35	62.86
	Rừng trồng	**0**	0	**63**	0	0	0	0	7	0	0	70	90
	Lúa	**0**	0	0	**129**	0	0	1	0	0	0	130	99.23
	Cây trồng khác	**1**	0	0	0	**76**	1	0	0	0	0	78	97.44
	Cây trồng lâu năm	**1**	0	0	0	0	**120**	1	0	0	0	122	98.36
	Cỏ mọc hoang	**0**	2	0	0	0	0	**153**	0	0	3	158	96.84
	Đất ở	**0**	0	6	0	0	0	0	**60**	4	0	70	85.71
	Đất ẩm ướt	**0**	0	0	4	2	0	0	0	**80**	0	86	93.02
	Khác	**0**	0	0	1	0	0	0	0	1	**20**	22	90.91
	Tổng	120	24	69	134	92	129	155	67	85	23	898	
	Producer accuracy	92.5	91.67	91.3	96.27	82.61	93.02	98.71	89.55	94.12	86.96		**80.5**

Bảng 7: Ma trận đánh giá độ chính xác khu vực Tây Bắc

LSR31B		Dữ liệu tham chiếu										
		Rừng thường xanh	Rừng trồng	Lúa	Cây trồng khác	Cây trồng lâu năm	Cỏ mọc hoang	Đất ở	Đất ẩm ướt	Khác	Tổng	*User accuracy*
Dữ liệu phân loại	Rừng thường xanh	**30**	2	0	0	3	0	0	0	0	35	85.71
	Rừng trồng	2	**55**	1	3	1	0	2	0	0	64	85.94
	Lúa	1	1	**58**	12	0	0	3	6	3	84	69.05
	Cây trồng khác	0	1	2	**70**	9	0	6	1	8	97	72.16
	Cây trồng lâu năm	3	5	4	4	**89**	0	3	0	0	108	82.41
	Cỏ mọc hoang	0	0	3	0	1	**30**	0	0	0	34	88.24
	Đất ở	1	0	3	2	5	0	**225**	1	4	241	93.36
	Đất ẩm ướt	0	0	3	2	1	0	11	**54**	2	73	73.97
	Khác	0	1	3	2	2	1	1	1	**60**	71	84.51
	Tổng	37	65	77	95	111	31	251	63	77	807	
	Producer accuracy	81.08	84.62	75.32	73.68	80.18	96.77	89.64	85.71	77.92		**83.1**

4. Kết luận và kiến nghị

Thuật toán Random Forest cho ra kết quả phân loại chính xác nhất trong các thuật toán toán được thử nghiệm phân loại. Random Forest không chỉ chính xác hơn khi sử dụng tập mẫu huấn luyện kích thước lớn, mà cũng cho thấy sự chính xác khi sử dụng tập mẫu huấn luyện có kích thước nhỏ. Việc đạt độ chính xác tốt ngay khi sử dụng tập huấn luyện nhỏ rất có ý nghĩa khi một số lớp phân loại không phổ biến ngoài thực địa, chỉ có thể cung cấp được tập mẫu kích thước không lớn của các lớp phân loại này cho mô hình. Một mô hình phân loại có thể hiện tốt đối với tập mẫu kích thước nhỏ sẽ có thể cải thiện kết quả phân loại của các lớp phân loại có tỷ lệ diện tích nhỏ trên khu vực nghiên cứu.

Độ chính xác của kết quả phân loại đạt được cao khi sử dụng kết hợp các kênh ảnh của ảnh radar, ảnh quang học và dữ liệu địa hình. Tuy nhiên từ kết quả thực nghiệm cho thấy, việc tính toán thêm các chỉ số thực vật không phải lúc nào cũng có thể cải thiện độ chính xác. Trong một số trường hợp, tính toán thêm các chỉ số còn làm giảm đi độ chính xác tổng thể. Vì vậy, việc lựa chọn sử dụng các chỉ số nào cần phải được thử nghiệm căn cứ vào các bài toán phân loại cụ thể.❍

Bài báo là kết quả nghiên cứu của Đề tài "Nghiên cứu sử dụng Dữ liệu lớn (bigdata) và Học máy (Machine Learning) để xây dựng phương pháp tự động phân loại lớp phủ mặt đất phục vụ kiểm kê phát thải khí nhà kính quốc gia", Mã số: TNMT.2022.04.08

Tài liệu tham khảo

[1]. World Bank, "World Bank Group. 2022. Vietnam Country Climate and Development Report. CCDR Series," Washington, DC: World Bank. [Online]. Available: http://hdl.handle.net/10986/37618

[2]. United Nations in Vietnam, "Nguyên nhân và ảnh hưởng của biến đổi khí hậu." [Online]. Available: https://vietnam.un.org/

[3]. Phong D. H. and Huệ N., "Giám sát và kiểm kê phát thải khí nhà kính (CO2 tương đương) trên cơ sở phân loại lớp phủ bằng ảnh Sentinel 1 tỉnh Quảng Bình," no. 2022, 2022.

[4]. Pham L. T. *et al.*, "Establishment of land cover map using object-oriented classification method for VNREDSat-1 data," *J. Min. Earth Sci.*, vol. 61, no. 2, pp. 134–144, Apr. 2020, doi: 10.46326/JMES.2020.61(2).15.

[5]. T.-T. Vu and Y. Shen, "Land-Use and Land-Cover Changes in Dong Trieu District, Vietnam, during Past Two Decades and Their Driving Forces," *Land*, vol. 10, no. 8, p. 798, Jul. 2021, doi: 10.3390/land10080798.

[6]. T. N. Trần, T. T. Vũ, T. N. Nguyễn, and T. O. Nông, "Ứng dụng công nghệ viễn thám trong thành lập bản đồ lớp phủ mặt đất theo hướng dẫn của IPCC phục vụ công tác giám sát tài nguyên môi trường và biến đổi khí hậu," *Tạp Chí Khoa Học Đo Đạc Và Bản Đồ*, no. 27, pp. 39–45, Mar. 2016, doi: 10.54491/jgac.2016.27.169.

[7]. A. Balha and C. K. Singh, "Comparison of Maximum Likelihood, Neural Networks, and Random Forests Algorithms in Classifying Urban Landscape," in *Application of Remote Sensing and GIS in Natural Resources and Built Infrastructure Management*, vol. 105, V. P. Singh, S. Yadav, K. K. Yadav, G. A. Corzo Perez, F. Muñoz-Arriola, and R. N. Yadava, Eds., in Water Science and Technology Library, vol. 105. , Cham: Springer International Publishing, 2022, pp. 29–38. doi: 10.1007/978-3-031-14096-9_2.

[8]. P. Thanh Noi and M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery," *Sensors*, vol. 18, no. 2, p. 18, Dec. 2017, doi: 10.3390/s18010018.

[9]. P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random Forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, Mar. 2006, doi: 10.1016/j.patrec.2005.08.011.

[10]. M.-J. Jun, "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area," *Int. J. Geogr. Inf. Sci.*, vol. 35, no. 11, pp. 2149–2167, Nov. 2021, doi: 10.1080/13658816.2021.1887490.

[11]. E. Adam, O. Mutanga, J. Odindi, and E. M. Abdel-Rahman, "Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers," *Int. J. Remote Sens.*, vol. 35, no. 10, pp. 3440–3458, May 2014, doi: 10.1080/01431161.2014.903435.

[12]. G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, May 2011, doi: 10.1016/j.isprsjprs.2010.11.001.

[13]. M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 1007–1011, Mar. 2005, doi: 10.1080/01431160512331314083.

[14]. S. Szabó, Z. Gácsi, and B. Balázs, "Specific features of NDVI, NDWI and

MNDWI as reflected in land cover categories," *Landsc. Environ.*, vol. 10, no. 3–4, pp. 194–202, Sep. 2016, doi: 10.21120/LE/10/3-4/13.

[15]. S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao, "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing," *J. For. Res.*, vol. 32, no. 1, pp. 1–6, Feb. 2021, doi: 10.1007/s11676-020-01155-1.

[16]. R. Kaur and P. Pandey, "A review on spectral indices for built-up area extraction using remote sensing technology," *Arab. J. Geosci.*, vol. 15, no. 5, p. 391, Mar. 2022, doi: 10.1007/s12517-022-09688-x.

[17]. V. Eisavi, S. Homayouni, A. M. Yazdi, and A. Alimohammadi, "Land cover mapping based on random forest classification of multitemporal spectral and thermal images," *Environ. Monit. Assess.*, vol. 187, no. 5, p. 291, May 2015, doi: 10.1007/s10661-015-4489-3.

[18]. L. Sun and K. Schulz, "The Improvement of Land Cover Classification by Thermal Remote Sensing," *Remote Sens.*, vol. 7, no. 7, pp. 8368–8390, Jun. 2015, doi: 10.3390/rs70708368.❍

Summary

Automatic classification model for greenhouse gas inventory using remote sensing data

Nong Thi Oanh, Tran Xuan Truong

Hanoi University of Mining and Geology

Trinh Viet Nga

Department of National Remote Sensing

Ta Hoang Trung

Department of Survey, Mapping and Geographic Information Viet Nam

Monitoring greenhouse gas emissions has recently received attention from scientists, policy-makers and agencies. In order to estimate greenhouse gas emissions, an accurate land cover map of the area to be evaluated is essential. This study uses the Random Forest algorithm to classify Landsat 8, Sentinel 1,2 satellite images to build land cover maps according to AFOLU classification schema recommendations, serving greenhouse gas inventories. The remote sensing image classification results in the Northwest and Southwest regions showed an overall accuracy of 80.5%, suitable for building surface coverage maps for greenhouse gas inventory.❍

Keywords: Remote sensing, Random Forest, Landcover, AFOLU