

ĐÁNH GIÁ ĐỘ CHÍNH XÁC TRONG PHÂN LOẠI LỚP PHỦ DỰA TRÊN THUẬT TOÁN HỌC MÁY VÀ DỮ LIỆU VIỄN THÁM THÔNG QUA GOOGLE EARTH ENGINE: ÁP DỤNG TẠI TỈNH ĐẮK LẮK

Giang Thị Phương Thảo¹, Phạm Thị Thu Hương²

Phạm Việt Hòa¹, Nguyễn An Bình¹

¹Viện Địa lý tài nguyên TP. HCM, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

²Trường Đại học Tài nguyên và Môi trường Hà Nội

Tóm tắt

Các thuật toán học máy cho phép tăng cường độ chính xác của các mô hình phân loại lớp phủ trên dữ liệu ảnh vệ tinh quan sát trái đất. Kết hợp với các nền tảng điện toán đám mây, quy trình này được thực hiện tự động nhằm đảm bảo hiệu suất xử lý một tệp dữ liệu lớn chứa nhiều thông tin không gian có giá trị được triết tách từ ảnh viễn thám. Nghiên cứu này đánh giá độ chính xác của các mô hình phân loại lớp phủ từ ảnh vệ tinh quang học Landsat 8 (LS8) và Sentinel 2 (S2) trên Google Earth Engine (GEE). Giá trị phổ phản xạ mặt đất được sử dụng làm đầu vào cho mô hình học máy Classification And Regression Tree (CART) và Random Forest (RF) phục vụ phân loại 7 loại hình hiện trạng lớp phủ tỉnh Đắk Lắk năm 2021. Kết quả đánh giá độ chính xác sau phân loại đã chỉ ra rằng, với diện tích khoảng 13.000 km² trên quy mô cấp tỉnh Đắk Lắk, ảnh LS8 cùng thuật toán CART đạt được độ chính xác cao nhất với hệ số Kappa là 0,85. Quy trình hoàn thiện cung cấp giải pháp tự động hóa trong phục vụ thành lập bản đồ hiện trạng lớp phủ chính xác và tin cậy, hỗ trợ cho các công tác trong lĩnh vực tài nguyên và môi trường.

Từ khóa: Đắk Lắk; Google Earth Engine; Học máy; Landsat 8; Sentinel 2; Viễn thám.

Abstract

Accuracy assessment of landcover classification based on various machine learning algorithms and remotely sensed data through Google Earth Engine: Application in Dak Lak province

Machine learning algorithms allow for increasing the accuracy of land cover classification models from earth observation satellite images. Combined with cloud computing platforms, the procedure is performed automatically to ensure efficient processing of a big dataset containing a plethora of valuable spatial information extracted from remotely sensed images. This study evaluates the performance of various classification models from Landsat 8 (LS8) and Sentinel 2 (S2) optical satellite images through Google Earth Engine (GEE). Spectral surface reflectance values are used as input to the Classification And Regression Tree (CART) and Random Forest (RF) machine learning models to classify 7 land cover classes in Dak Lak province in 2021. Statistical error revealed that, with an area of about 13,000 km² on the scale of Dak Lak province, the LS8 image combined with the CART algorithm achieved

the highest accuracy with Kappa of 0.85. The all-inclusive procedure provides an automatic solution of accurate and reliable land cover mapping, as well as supporting natural resources management and the environmental monitoring.

Keywords: Dak Lak; Google Earth Engine; Machine learning; Landsat 8; Sentinel 2; Remote sensing.

1. Mở đầu

Khái niệm hiện trạng lớp phủ được dùng để chỉ các đối tượng vật lý (sinh học) có thể quan sát được trên bề mặt trái đất. Nhu cầu chi tiết hóa để giám sát và định lượng chính xác lớp phủ trên quy mô không gian - thời gian đang là những đòi hỏi cấp bách, chính vì thế vấn đề này luôn được quan tâm đặc biệt trong các nghiên cứu về quản lý tài nguyên thiên nhiên và môi trường, phòng chống thiên tai, quy hoạch, xây dựng các chính sách của Chính phủ. Kể từ khi bức ảnh viễn thám đầu tiên được chụp từ vệ tinh America's Earth Explore 6 năm 1959 [1], ứng dụng của công nghệ viễn thám trong thành lập bản đồ hiện trạng sử dụng đất và lớp phủ (Land Use Land Cover - LULC) cho đến nay đã trở thành một lĩnh vực nghiên cứu điển hình và lớn mạnh. Mặc dù vậy, công tác này đã, đang và sẽ vẫn là một nhiệm vụ luôn luôn cấp thiết phục vụ cho thống kê thường kỳ biến động sử dụng đất và lớp phủ cũng như cung cấp nguồn dữ liệu đầu vào quan trọng cho các nghiên cứu chuyên sâu trong lĩnh vực khoa học trái đất.

Chuyển đổi từ việc áp dụng các phương pháp phân loại truyền thống sang các mô hình trí tuệ nhân tạo đang là xu thế mới hiện nay. Việc cập nhật liên tục các mô hình học máy phục vụ phân loại không giám sát (Unsupervised classification) và có giám sát (Supervised classification) dựa trên pixel [2], hoặc tích hợp mô hình

học sâu cùng với phương pháp hướng đối tượng (Object-oriented) với đa dạng các thuật toán mở ra nhiều nghiên cứu chuyên sâu cho phép tăng cường kiến thức và hiểu biết xung quanh lĩnh vực quan tâm [3]. Trên thế giới cũng như tại Việt Nam hiện nay, chủ đề nghiên cứu này đang được thực hiện với nhiều công trình công bố trong những năm gần đây.

Nhiều bộ cảm vệ tinh độ phân giải từ trung bình đến cao và miễn phí đã được thử nghiệm tại các khu vực có cảnh quan lớp phủ đồng nhất và phức tạp tại Việt Nam. Ảnh Landsat 7, 8 có độ phân giải trung bình 15 m với kênh toàn sắc, 60 m với kênh nhiệt và 30 m với kênh phổ đã được kết luận là phù hợp cho việc xây dựng bản đồ lớp phủ khu vực Tây Nguyên [4]. Ảnh Sentinel 2 cũng đã được kết luận có độ phân giải cao phục vụ cho thành lập bản đồ hiện trạng rừng [5]. Mô hình phân loại tự động từ thuật toán học máy Random Forest trên bộ dữ liệu Landsat 5, 7, 8 giai đoạn 30 năm cho phép đánh giá chi tiết sự thay đổi của các hoạt động sinh kế nông nghiệp vùng Đồng Tháp Mười thuộc vùng đồng bằng sông Cửu Long [2]. Cùng với việc chuyển đổi mô hình phân loại lên các nền tảng điện toán đám mây, một loạt các ưu thế khi so sánh với các giải pháp thực hiện trên máy tính để bàn đã được kết luận bao gồm (1) Tiết kiệm thời gian và công sức khi thực hiện tiền xử lý dữ liệu; (2) Quy trình được thực hiện tự động cùng với hiệu năng xử lý dữ liệu lớn; (3) Khả năng chia sẻ dữ liệu và

sản phẩm đến người dùng cuối và (4) Cho phép tạo ra một cơ sở dữ liệu không gian vô cùng lớn phục vụ các phân tích tiếp theo như chuỗi dữ liệu, mô hình hóa [6].

Nghiên cứu này đánh giá khả năng của 2 thuật toán học máy phổ biến hiện nay trong phân loại hiện trạng lớp phủ bề mặt từ dữ liệu ảnh quang học. Lựa chọn khu vực nghiên cứu điển hình với điều kiện cảnh quan không đồng nhất tại tỉnh Đắk Lắk, giá trị phổ phản xạ mặt đất (Surface reflectance) sau hiệu chỉnh khí quyển triết tách từ ảnh vệ tinh Landsat 8 (LS8) và Sentinel 2 (S2) được sử dụng để xây dựng mô hình phân loại dựa trên thuật toán Random Forest (RF) và Classification And Regression Tree (CART). Đánh giá độ chính xác sau phân loại được thực hiện dựa trên độ chính xác tổng thể, độ chính xác người dùng, độ chính xác nhà sản xuất và hệ số Kappa nhằm so sánh và lựa chọn thuật toán cũng như loại hình dữ liệu ảnh vệ tinh phù hợp cho quy mô khu vực nghiên cứu. Quy trình các bước xử lý được thiết lập trên GEE minh chứng cho khả năng tự động hóa, tiết kiệm thời gian xử lý, đồng thời dễ dàng chuyển đổi để áp dụng cho quy mô không gian và thời gian khác nhau một cách linh hoạt.

2. Khu vực nghiên cứu và dữ liệu sử dụng

2.1. Khu vực nghiên cứu

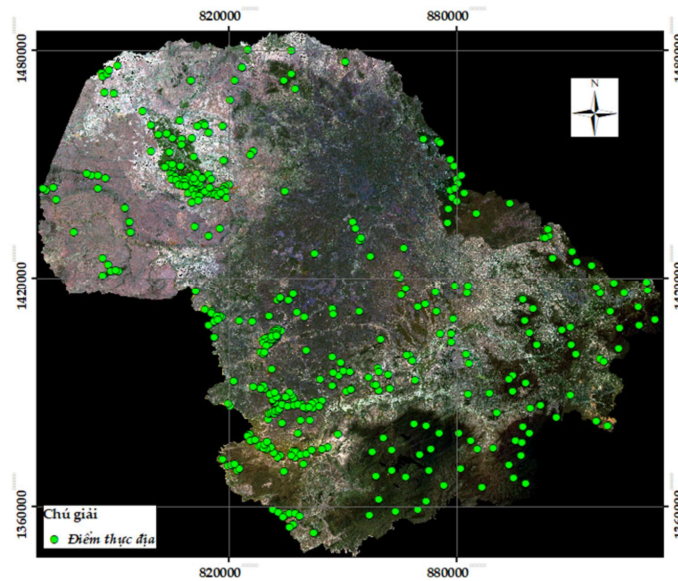
Đắk Lắk thuộc khu vực Tây Nguyên, có diện tích 13.125,4 km² chiếm 27,6 % diện tích vùng Tây Nguyên và 3,9 % diện tích tự nhiên của cả nước. Khu vực nghiên cứu nằm trong khoảng tọa độ địa lý từ 107°28'57" - 108°59'37" độ kinh Đông và từ 12°9'45" - 13°25'06" độ vĩ

Bắc. Khí hậu toàn tỉnh được chia thành 2 vùng tiểu khí hậu. Vùng phía Tây Bắc có khí hậu nắng nóng, khô hanh về mùa khô, vùng phía Đông có khí hậu mát mẻ, ôn hòa. Do đặc điểm vị trí địa lý, địa hình nên khí hậu ở Đắk Lắk vừa chịu sự chi phối của khí hậu nhiệt đới gió mùa, vừa mang tính chất của khí hậu cao nguyên mát dịu. Điều kiện khí hậu mang lại các ưu thế cho việc phát triển nông nghiệp trên địa bàn toàn tỉnh.

2.2. Dữ liệu sử dụng

Dữ liệu thực địa bao gồm 390 điểm thực địa đại diện cho 7 loại hình lớp phủ chủ yếu được điều tra trên toàn tỉnh Đắk Lắk bao gồm (1) Mặt nước; (2) Dân cư và công trình; (3) Lúa; (4) Cây nông nghiệp khác bao gồm cây ăn trái và cây công nghiệp như cao su, cà phê, hồ tiêu; (5) Rừng dày; (6) Rừng thưa và (7) Đất trống. Trong số điểm dữ liệu thực địa này, 70 % sẽ được dùng để thực hiện phân loại và 30 % còn lại được dùng để đánh giá độ chính xác sau phân loại (Hình 1).

Dữ liệu viễn thám được sử dụng trong nghiên cứu bao gồm dữ liệu ảnh vệ tinh LS8 và S2. Bộ cảm Operational Land Imager (OLI) trên vệ tinh LS8 của Hoa Kỳ gồm 9 kênh phổ (30 m), 1 kênh toàn sắc (15 m) và 2 kênh hồng ngoại nhiệt (100 m). Đối với thể hệ vệ tinh S2 của châu Âu, bộ cảm MultiSpectral Instrument (MSI) gồm 12 kênh phổ có độ phân giải từ 10 - 60 m. Lưu ý rằng giá trị phổ phản xạ mặt đất (Surface reflectance) sau hiệu chỉnh khí quyển và lọc mây được sử dụng làm đầu vào cho mô hình phân loại nhằm giảm thiểu các sai số.

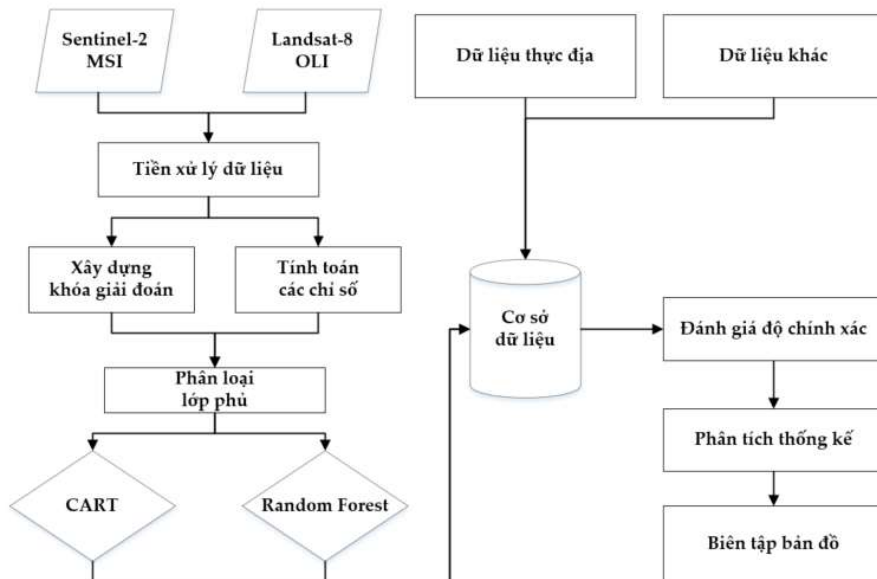


Hình 1: Sơ đồ các điểm thực địa khu vực nghiên cứu

3. Phương pháp nghiên cứu

Quy trình nghiên cứu tổng thể bao gồm các bước tiền xử lý dữ liệu, xây dựng

khóa giải đoán, tính toán các chỉ số hỗ trợ cho việc phân loại, thực hiện phân loại theo các phương pháp khác nhau (Hình 2).



Hình 2: Sơ đồ quy trình nghiên cứu

Kết quả các quá trình phân loại kết hợp với dữ liệu thực địa và dữ liệu khác tạo thành bộ cơ sở dữ liệu địa không gian tổng hợp phục vụ cho việc đánh giá độ chính xác sau phân loại, phân tích thống kê và biên tập bản đồ. Trong nghiên cứu này, quy trình được thực hiện trên nền tảng điện toán đám mây GEE nhằm mục đích tận dụng ưu thế trong

việc sử dụng một cơ sở dữ liệu viễn thám lớn sẵn có, cùng với các phép phân tích xử lý và mô hình phân loại học máy hiệu năng cao nhằm cung cấp một giải pháp toàn diện và tự động hóa trong thành lập bản đồ lớp phủ. Các sản phẩm bản đồ sau phân loại và đánh giá độ chính xác được tải về nhằm mục đích phân tích thống kê và trình bày hiển thị.

3.1. Thuật toán phân loại

Nghiên cứu lựa chọn và thử nghiệm 2 thuật toán phân loại học máy (1) Cây phân loại và hồi quy CART và (2) RF. Trong đó, thuật toán CART dựa trên cây quyết định có ưu điểm là rất đơn giản và trực quan, ít yêu cầu chuẩn bị dữ liệu cả dữ liệu số và dữ liệu phân loại, tuy nhiên lại có nhược điểm là có thể tạo ra các cây không có khả năng khái quát hóa tốt, các biến thể nhỏ trong dữ liệu có thể tạo ra một cây hoàn toàn khác. Thuật toán rừng ngẫu nhiên cũng được xây dựng dựa trên CART, tuy nhiên bao gồm nhiều cây quyết định CART khác nhau cấu thành bởi yếu tố ngẫu nhiên (Random). Khác với CART, kết quả dự đoán từ mô hình RF sẽ được tổng hợp từ các quyết định, do đó sẽ phức tạp hơn so với cây quyết định đơn lẻ CART.

3.2. Sử dụng chỉ số hỗ trợ phân loại

Một trong những ưu điểm của việc thực hiện phân loại trên các thuật toán phân loại học máy là có thể bổ sung các thông tin khác ngoài thông tin có sẵn trên các kênh ảnh của dữ liệu vệ tinh. Việc này nhằm làm tăng độ chính xác của kỹ thuật học máy, cụ thể trong trường hợp này là cây quyết định CART và RF. Nghiên cứu đã bổ sung chỉ số thực vật chuẩn hóa (Normalized Difference Vegetation Index - NDVI) và chỉ số khác biệt nước chuẩn hóa (Normalised Difference Water Index - NDWI).

Chỉ số thực vật chuẩn hóa NDVI được phân tách từ các kênh cận hồng ngoại (Near Infrared - NIR) và kênh sóng nhìn thấy đỏ (Red) để thấy được sự khác nhau của các từng loại hình thảm thực vật như rừng dày, rừng thưa, cây nông nghiệp. Có nhiều các chỉ số thực vật khác nhau, nhưng chỉ số thực vật chuẩn hóa NDVI đã được chứng minh là chỉ số phổ biến nhất

để giám sát sự thay đổi trạng thái thực vật. Chỉ số NDVI được tính theo công thức:

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} \quad (1)$$

Trong đó: NIR, RED là phổ phản xạ ở dải sóng cận hồng ngoại và dải đỏ. Giá trị NDVI nằm trong khoảng từ -1 đến +1. Giá trị NDVI từ 0 - 1 cho thấy khu vực có độ phủ thực vật biến thiên từ thấp tới cao. NDVI có giá trị âm cho thấy tại khu vực đó không có thực vật, có thể là những đối tượng mặt nước hay công trình xây dựng

Chỉ số NDWI được đề xuất xây dựng bởi McFeeters [7] nhằm tối đa hóa độ phản xạ của vùng nước mặt trong dải sóng màu xanh lục (Green) và giảm thiểu độ phản xạ của vùng nước mặt trong dải hồng ngoại gần NIR, với công thức như sau:

$$NDWI = \frac{\rho_{Green} - \rho_{NIR}}{\rho_{Green} + \rho_{NIR}} \quad (2)$$

Trong đó NIR, Green là phổ phản xạ ở dải sóng cận hồng ngoại và dải xanh lục. Giá trị NDWI thông thường được xác định với ngưỡng 0 với các giá trị dương là nước và các giá trị âm là đối tượng không nước.

3.3. Đánh giá độ chính xác

Để đánh giá độ chính xác của phương pháp, 30 % dữ liệu thực địa được sử dụng để tính toán các chỉ số bao gồm độ chính xác toàn cục (Overall accuracy), độ chính xác nhà sản xuất (Producer's accuracy), độ chính xác người dùng (User's accuracy) và hệ số Kappa (K).

$$\text{Độ chính xác toàn cục} = \frac{\sum_{i=1}^k n_{ij}}{n} \quad (3)$$

$$\text{Độ chính xác nhà sản xuất được tính theo công thức} \quad \frac{n_{ij}}{n_{+j}} \quad (4)$$

$$\text{Độ chính xác người dùng được tính theo công thức} \quad \frac{n_{ij}}{n_{i+}} \quad (5)$$

$$\hat{K} = \frac{n \sum_{i=1}^k n_{ii} - \sum_{i=1}^k n_{i+} n_{+i}}{n^2 - \sum_{i=1}^k n_{i+} n_{+i}} \quad (6)$$

Trong đó:

k: Số lớp.

n: Tổng số điểm thực địa được dùng để đánh giá độ chính xác.

n_{ij} : Tổng số điểm thuộc lớp i ($i = 1, 2, \dots, k$) trên ảnh được phân loại vào lớp j ($j = 1, 2, \dots, k$) ngoài thực tế, đây là những điểm được phân loại đúng.

$n_{i+} = \sum_{j=1}^k n_{ij}$ là tổng số mẫu từ kết quả phân loại ảnh vệ tinh.

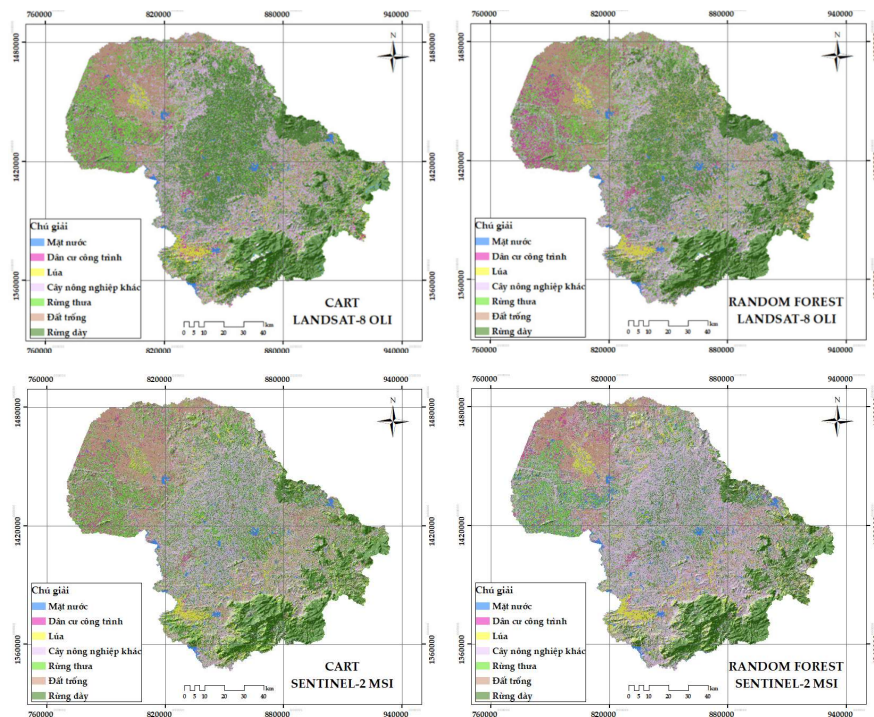
$n_{+j} = \sum_{i=1}^k n_{ij}$ là tổng số mẫu có được từ dữ liệu tham khảo.

4. Kết quả và thảo luận

4.1. Kết quả phân loại hiện trạng lớp phủ tỉnh Đắk Lắk

Hiện trạng lớp phủ tỉnh Đắk Lắk tại thời điểm tháng 3 năm 2021 được phân loại

trên ảnh LS8 và S2 bằng thuật toán CART và RF được trình bày tại Hình 3. Đối với các phương pháp phân loại dựa trên ảnh LS8, kết quả cho thấy là rừng dày phần lớn tập trung tại trung tâm phía Đông nam của tỉnh. Ngược lại, rừng thưa chủ yếu phân bố tại phía Tây, Tây bắc và rải rác tại trung tâm của khu vực nghiên cứu. Mặt nước của tỉnh chủ yếu là các con sông và hồ chứa nước phân bố đều khắp khu vực nghiên cứu. Cả tỉnh có hai khu vực có lớp phủ lúa nổi bật phân bố tại phía Tây của trung tâm và phía Tây nam của khu vực nghiên cứu. Đối với dữ liệu lớp phủ được thành lập từ ảnh có độ phân giải không gian cao hơn S2 cho thấy độ chi tiết hơn so với LS8 khi áp dụng thuật toán CART và RF. Tuy nhiên, sự phân bố của các đối tượng tương tự như kết quả phân loại trên ảnh LS8. Điểm khác biệt lớn nhất đến từ việc phân loại các vùng canh tác nông nghiệp bao gồm cả cây ăn trái và cây công nghiệp như cà phê, hồ tiêu phần lớn phân bố xung quanh khu vực trung tâm với diện tích tương đối lớn.



Hình 3: Kết quả phân loại hiện trạng lớp phủ khi áp dụng thuật toán CART và RF trên dữ liệu ảnh viễn thám LS8 và S2

Thông kê tại Bảng 1 cho thấy diện tích và cơ cấu các loại hình lớp phủ theo phần trăm tại tỉnh Đắk Lắk khác nhau khi áp dụng các thuật toán phân loại khác nhau trên các dữ liệu ảnh khác nhau. Tổng diện tích các loại hình lớp phủ trên toàn khu vực nghiên cứu là 1.306.927,8 ha. Lớp phủ chiếm diện tích lớn nhất là rừng dày, với cơ cấu 36,8 % khi sử dụng thuật toán CART và 33,9 % khi áp dụng thuật toán RF cùng trên ảnh LS8. Đối với ảnh S2, cơ cấu phân bố lớn nhất thuộc về đối tượng cây nông nghiệp khác, với 29,9 % trên CART và 43 % trên RF. Ngược lại, đối tượng lớp phủ có độ bao phủ thấp nhất trên cả hai loại ảnh và thuật toán phân loại là mặt nước. Trên ảnh LS8 với cả 2 thuật toán, mặt nước bao phủ lần lượt là

1,4 % và 1,3 %. Trên ảnh S2, giá trị này là 1,7 % và 3,3 %. Rừng thưa cũng có diện tích phân bố đáng kể. Trên ảnh LS8, đối tượng này có phần trăm phân bố là 17,3 % bằng thuật toán CART và 15,5 % bằng thuật toán RF. Trên ảnh S2, đối tượng này có giá trị phân bố là 18,6 % trên CART và 11,9 % trên Random Forest. Trong các loại hình thực vật, lúa là loại hình có diện tích bao phủ thấp nhất. Trên ảnh LS8, loại hình này chiếm 3,2 % tổng diện tích với thuật toán CART và 3,5 % với thuật toán RF. Với ảnh S2, giá trị này là 4,1 % trên CART và 4,8 % trên RF. Đất trống có diện tích phân bố đáng kể tại khu vực nghiên cứu, cụ thể là 10,7 % (CART), 11,1 % (FF) trên ảnh LS8 và 15,9 % (CART), 10 % (RF) trên ảnh S2.

Bảng 1. Thống kê diện tích (ha) và cơ cấu (%) các loại hình lớp phủ tỉnh Đắk Lắk theo các phương pháp phân loại và dữ liệu ảnh sử dụng

	CART - LS8		RF - LS8		CART - S2		RF - S2	
	Diện tích	%	Diện tích	%	Diện tích	%	Diện tích	%
Mặt nước	17.992,7	1,4	17.186,8	1,3	22.425,2	1,7	43.332,8	3,3
Dân cư công trình	56.101,5	4,3	82.390,9	6,3	71.129,4	5,4	55.417,1	4,2
Lúa	42.163,5	3,2	45.893,1	3,5	53.611,3	4,1	62.151,1	4,8
Cây nông nghiệp	343.127,1	26,3	370.717,8	28,4	390.118,3	29,9	562.200,6	43,0
Rừng thưa	226.338,5	17,3	203.107,8	15,5	243.177,0	18,6	155.321,5	11,9
Đất trống	139.891,1	10,7	145.016,4	11,1	207.927,5	15,9	131.048,3	10,0
Rừng dày	481.313,4	36,8	442.615,2	33,9	318.539,1	24,4	297.456,5	22,8
Tổng	1.306.927,8							

4.2. Đánh giá độ chính xác sau phân loại

Bảng 2 thể hiện chi tiết độ chính xác của các kết quả phân loại theo các cách thức kết hợp ảnh vệ tinh và mô hình phân loại khác nhau. Bằng cả hai thuật toán CART và RF, ảnh LS8 đạt độ chính xác toàn cục lần lượt là 85,25 % và 79,51 %, với hệ số Kappa lần lượt là 0,85 và 0,8. Trên ảnh S2, độ chính xác toàn cục là 75,83 % và 73,53 %, Kappa lần lượt là 0,76 và 0,73. Có thể thấy, Kappa của các loại ảnh trên từng thuật

toán đều từ mức trung bình đến cao, minh chứng cho độ tin cậy của quy trình đề xuất. So sánh độ chính xác khi áp dụng riêng các thuật toán phân loại CART và RF cho từng nguồn ảnh vệ tinh khác nhau cho thấy độ chính xác tổng thể cao nhất thuộc về thuật toán CART áp dụng phân loại trên ảnh LS8 với độ chính xác tổng thể đạt 85,25 % và Kappa đạt 0,85 thuộc loại tốt. Loại ảnh và phương pháp phân loại có độ chính xác tổng thể thấp nhất là ảnh S2 và thuật toán RF với độ chính xác tổng thể đạt 72,95 %

Nghiên cứu

và hệ số Kappa là 0,73 thuộc loại trung bình. Bên cạnh đó, qua độ chính xác người dùng và độ chính xác nhà sản xuất, có thể thấy đối tượng lớp phủ mặt nước luôn có độ chính xác cao và ít bị nhầm lẫn với các đối tượng khác (>85 %). Nhìn chung,

phương pháp RF thường có độ chính xác của các loại lớp phủ thấp hơn CART ở cả 2 loại ảnh. Thêm vào đó, trong nghiên cứu này, ảnh LS8 cũng thể hiện là tư liệu ưu thế hơn khi sử dụng để phân loại lớp phủ tỉnh Đắk Lắk.

Bảng 2. So sánh độ chính xác các kết quả phân loại

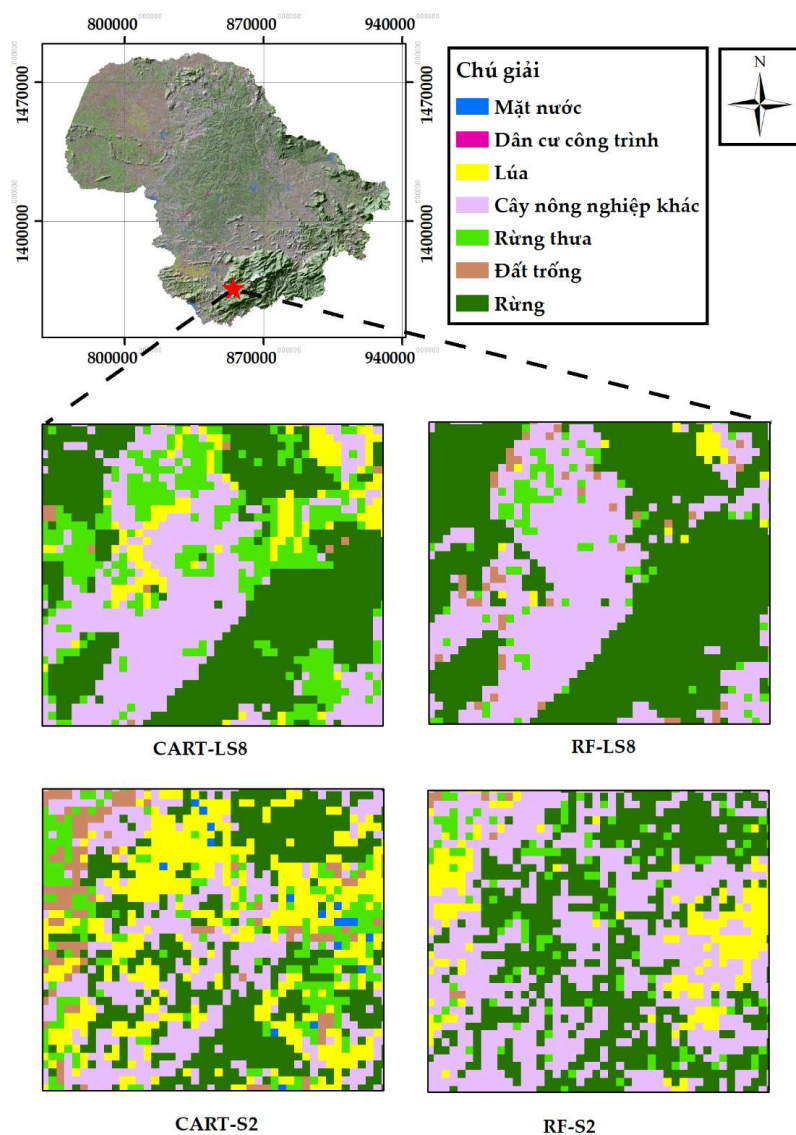
	Độ chính xác	Loại hình							Độ chính xác tổng thể	Kappa
		Mặt nước	Dân cư công trình	Lúa	Cây nông nghiệp	Rừng thưa	Đất trống	Rừng dày		
CART	Nhà sản xuất	100,00	76,92	100,00	76,92	60,00	79,17	88,89	85,25	0,85
LS8	Người dùng	100,00	100,00	80,00	55,56	66,67	79,17	100,00		
RF	Nhà sản xuất	100,00	83,33	87,50	57,89	35,71	90,00	85,71	79,51	0,8
LS8	Người dùng	100,00	50,00	70,00	61,11	55,56	75,00	100,00		
CART	Nhà sản xuất	100,00	91,67	66,67	53,33	36,36	55,56	87,10	76,23	0,76
S2	Người dùng	96,67	78,57	66,67	44,44	57,14	62,50	87,10		
RF	Nhà sản xuất	88,46	90,00	57,89	66,67	69,23	68,18	70,00	72,95	0,73
S2	Người dùng	85,19	64,29	91,67	57,14	45,00	88,24	77,78		

4.3. So sánh giữa các kết quả phân loại

Cả hai phương pháp phân loại CART và RF khi sử dụng cho ảnh S2 đều cho các kết quả có độ chi tiết cao hơn so với ảnh LS8. Điều này có thể lý giải là bởi vì do ảnh S2 có độ phân giải không gian cao hơn LS8. Trong khi S2 với các kênh phổ chính đỏ, lam, lục, cận hồng ngoại có độ phân giải không gian 10 m, LS8 bao gồm các kênh trong dải sóng nhìn thấy, cận hồng ngoại và hồng ngoại sóng ngắn đều có chung độ phân giải 30 m. Mặc dù có độ chi tiết cao hơn khi tiến hành phân loại, song quá trình phân loại cũng cho thấy xuất hiện các pixel phân loại sai (nhiều) trên các khu vực có cùng một loại hình lớp phủ, ví dụ như rừng dày với rừng thưa, đất trồng lúa với các loại cây nông nghiệp khác (Hình 4).

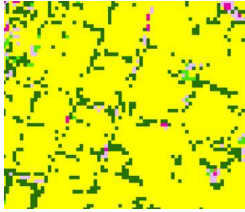
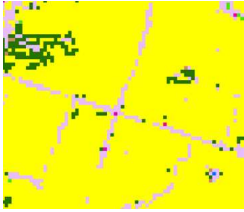
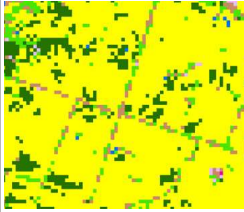
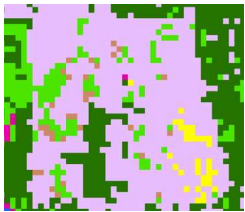
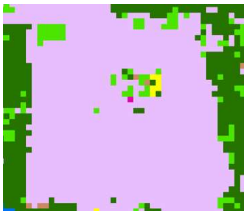
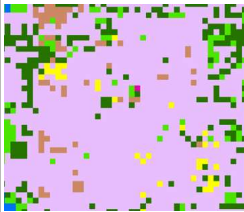
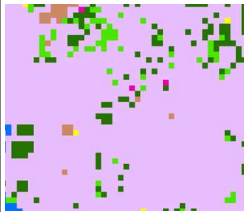
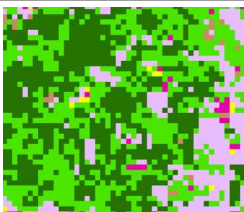
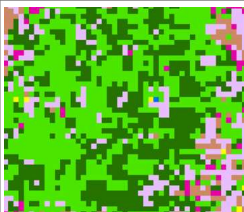
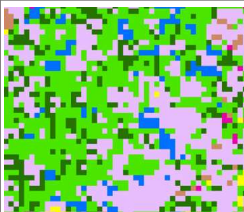
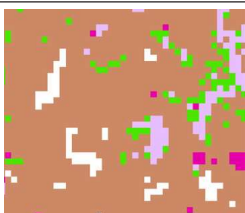
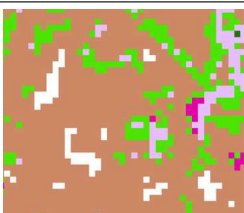
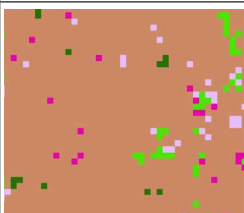
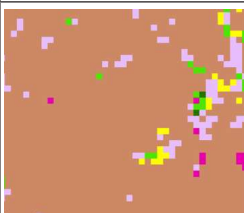
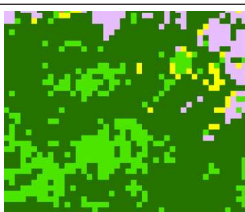
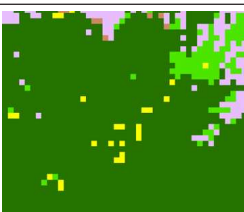
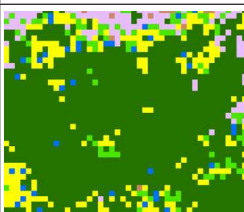
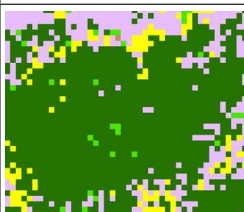
Đối với từng loại hình lớp phủ, các phương pháp phân loại dựa trên thuật toán CART và RF khi áp dụng lên dữ liệu ảnh LS8 và S2 cũng cho các kết quả phân loại

có sự khác biệt. Mặt nước là đối tượng mà các phương pháp phân loại khác nhau dễ nhận biết nhất và các kết quả gần như đồng nhất. Trong trường hợp này, chỉ số NDWI cũng có tác dụng hỗ trợ cho việc phân biệt các đối tượng nước - không nước tốt hơn. Đối với loại hình lớp phủ khu dân cư và công trình xây dựng, lớp phủ cây nông nghiệp khác và rừng thưa thường bị nhầm lẫn, trong đó thể hiện chi tiết hơn khi sử dụng dữ liệu ảnh S2. Lúa cho thấy sự phân biệt khá tốt với các loại hình lớp phủ còn lại. Loại hình lớp phủ cây nông nghiệp khác thường bị nhầm lẫn với rừng. Rừng thưa và rừng dày thường bị nhầm lẫn với nhau, ngoại trừ thuật toán phân loại RF trên ảnh S2 cho kết quả tương đối khác với ba cách thức kết hợp còn lại. Khu vực đất trống có thể nhận biết rõ ràng trên ảnh LS8 và S2, cũng như khi áp dụng thuật toán phân loại CART và RF. Ngoài ra, ảnh S2 cũng ghi nhận sự nhầm lẫn giữa đối tượng rừng dày và đất trồng lúa, trong khi ảnh LS8 có thể phân biệt rõ hai loại hình này (Hình 5).



Hình 4: Khác biệt kết quả phân loại giữa hiện trạng lớp phủ tỉnh Đắk Lắk giữa ảnh LS8 và S2, thuật toán CART và RF

	CART - LS8	RF - LS8	CART - S2	RF - S2
Mặt nước				
Dân cư và công trình				

	CART - LS8	RF - LS8	CART - S2	RF - S2
Lúa				
Cây nông nghiệp khác				
Rừng thưa				
Đất trống				
Rừng				

Hình 5: Sai khác giữa các cách thức kết hợp phân loại trên các loại hình lớp phủ

5. Kết luận

Trong nghiên cứu này sử dụng nguồn dữ liệu ảnh vệ tinh quang học LS8 và S2, kết hợp với các thuật toán phân loại CART và RF để phân loại hiện trạng lớp phủ tỉnh Đắk Lắk tháng 3 năm 2021. Nhận định chung là thuật toán CART và RF đều cho ra kết quả khá tương đồng và đáng tin cậy với độ chính xác cao. Với quy mô khu vực nghiên cứu, khi phân loại trên ảnh vệ tinh LS8 cho độ chính xác cao hơn và ít nhiễu hơn so với ảnh S2. Dữ liệu ảnh S2 với độ phân giải không gian 10 m cần có chỉnh

sửa sau đánh giá độ chính xác để giảm bớt nhiễu. Cả hai loại ảnh LS8 và S2 đều cần có những chỉnh sửa sau khi phân loại, phần lớn dựa vào kinh nghiệm của người phân loại để cho ra kết quả chính xác nhất. Quy trình áp dụng trên nền tảng điện toán đám mây GEE cho phép thực hiện nhanh chóng và chính xác, tự động hóa trong việc áp dụng mô hình theo quy mô không gian và thời gian khác nhau phục vụ các phân tích chuyên sâu tiếp theo.

Lời cảm ơn: Nghiên cứu này được hỗ trợ bởi Đề tài: “Nghiên cứu tác động

của yếu tố khí hậu tới các loại hình sử dụng đất nông nghiệp tại tỉnh Đắk Lắk”, do Sở Khoa học và Công nghệ tỉnh Đắk Lắk cấp kinh phí.

TÀI LIỆU THAM KHẢO

[1]. European Space Agency (2022). *50 years of Earth Observation*. https://www.esa.int/About_Us/ESA_history/50_years_of_Earth_Observation.

[2]. N. A. Binh et al (2021). *Thirty-year dynamics of LULC at the Dong Thap Muoi area, Southern Vietnam, using Google Earth Engine*. ISPRS Int J Geoinf, vol. 10, no. 4. Doi: 10.3390/ijgi10040226.

[3]. M. Papadomanolaki, M. Vakalopoulou and K. Karantzas (2019). *A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks*. Remote Sens (Basel), vol. 11, no. 6. Doi: 10.3390/RS11060684.

[4]. Hồ Thị Lệ Thu, Nguyễn Thanh Hoàn, Lê Minh Hằng, Đặng Xuân Phong

(2020). *Xây dựng bản đồ lớp phủ khu vực Tây Nguyên sử dụng dữ liệu ảnh Landsat đa thời gian*. Tạp chí Khoa học và Công nghệ Việt Nam, số 3, tr. 43 - 45.

[5]. Trần Quang Bảo, Cao Lê Quốc Việt, Võ Minh Hoàn, Nguyễn Thị Hoa (2022). *Xây dựng bản đồ hiện trạng rừng của huyện Đồng Phú, tỉnh Bình Phước bằng ảnh vệ tinh Sentinel 2A*. Tạp chí Khoa học và Công nghệ Lâm nghiệp, số 3, tr. 53 - 63.

[6]. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau and R. Moore (2017). *Google Earth Engine: Planetary-scale geospatial analysis for everyone*. Remote Sens Environ, vol. 202, p. 18 - 27. Doi: <https://doi.org/10.1016/j.rse.2017.06.031>.

[7]. S. K. McFeeters (1996). *The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features*. Int J Remote Sens, vol. 17, no. 7. Doi: 10.1080/01431169608948714.

BBT nhận bài: 16/5/2023; Phản biện xong: 30/5/2023; Chấp nhận đăng: 29/6/2023