So sánh hiệu quả các thuật toán Random Forest, SVM và Naive Bayes trong phân loại lớp phủ bề mặt sử dụng dữ liệu Sentinel-2 trên Google Earth Engine: Trường hợp nghiên cứu tại khu vực Thái Nguyên, Việt Nam

Lê Duy Thành[1], Trịnh Thị Hoài Thu[1]*, Bùi Thị Hồng Thắm[1], Trịnh Thị Loan[2]

[1]*Trường Đại học Tài nguyên và Môi trường Hà Nội, Hà Nội, Việt Nam*

[2]*Công ty cổ phần khảo sát đo đạc và thiết kế Hưng Bình, Hà Nội, Việt Nam*

Email tác giả liên hệ: tththu@hunre.edu.vn

https://doi.org/10.5281/zenodo.xxxxxxxx

Tóm tắt:

Nghiên cứu này đánh giá hiệu quả của ba thuật toán học máy gồm Random Forest (RF), Support Vector Machine (SVM) và Naive Bayes (NB) trong phân loại lớp phủ bề mặt từ dữ liệu ảnh vệ tinh Sentinel-2 tại tỉnh Thái Nguyên. Dữ liệu ảnh được xử lý trên nền tảng Google Earth Engine (GEE), kết hợp các kênh phổ, chỉ số phổ, yếu tố địa hình và đặc trưng kết cấu ảnh để xây dựng tập dữ liệu đầu vào cho mô hình phân loại. Tổng cộng 18.524 pixel mẫu được sử dụng, trong đó 70% số mẫu dùng để huấn luyện và 30% dùng để đánh giá độ chính xác mô hình. Kết quả cho thấy sự khác biệt rõ rệt về hiệu quả phân loại giữa các thuật toán. RF đạt độ chính xác cao nhất với độ chính xác tổng thể (OA) 90,27% và hệ số Kappa 0,880; SVM cũng cho kết quả tốt với OA đạt 88,78% và Kappa 0,862. Trong khi đó, NB cho độ chính xác thấp hơn đáng kể với OA 37,41% và Kappa 0,238. Điều này cho thấy RF và SVM có khả năng mô hình hóa tốt các mối quan hệ phi tuyến giữa các biến phổ và lớp phủ bề mặt, trong khi giả định độc lập giữa các biến đầu vào của NB chưa phù hợp với đặc trưng phổ phức tạp của dữ liệu viễn thám đa phổ. Nghiên cứu khẳng định tiềm năng của việc kết hợp dữ liệu Sentinel-2 với các thuật toán học máy trên nền tảng GEE trong phân loại và xây dựng bản đồ lớp phủ bề mặt, đồng thời cung cấp cơ sở cho việc lựa chọn thuật toán phù hợp trong các nghiên cứu viễn thám.

Từ khóa: *Sentinel-2, phân loại lớp phủ bề mặt, Random Forest, Support Vector Machine, GEE*

Ngày nhận bài: 16/03/2026 Ngày sửa lại: 06/04/2026 Ngày chấp nhận đăng: 10/04/2026 Ngày xuất bản: 30/04/2026

Comparative evaluation of Random Forest, SVM, and Naive Bayes algorithms for land cover classification using Sentinel-2 data on Google Earth Engine: A case study in Thai Nguyen, Vietnam

Le Duy Thanh[1], Trinh Thi Hoai Thu[1]*, Bui Thi Hong Tham[1], Trinh Thi Loan[2]

[1]*Hanoi University of Natural Resources and Environment, Hanoi, Vietnam*

[2]*Hung Binh Surveying and Design Joint Stock Company, Hanoi, Vietnam*

Corresponding Author Email: tththu@hunre.edu.vn

Abstract:

This study evaluates the performance of three machine learning algorithms, namely Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB), for land cover classification using Sentinel-2 satellite imagery in Thai Nguyen province, Vietnam. The imagery was processed on the Google Earth Engine (GEE) platform, integrating spectral bands, spectral indices, topographic variables, and image texture features to construct the input dataset for the classification models. A total of 18,524 sample pixels were used, of which 70% were allocated for model training and 30% for accuracy assessment. The results indicate a clear difference in classification performance among the algorithms. RF achieved the highest accuracy with an overall accuracy (OA) of 90.27% and a Kappa coefficient of 0.880. SVM also produced good classification results, with an OA of 88.78% and a Kappa value of 0.862. In contrast, NB showed

significantly lower performance, with an OA of 37.41% and a Kappa coefficient of 0.238. These findings suggest that RF and SVM are capable of effectively modeling nonlinear relationships between spectral variables and land cover classes, whereas the independence assumption of NB is not well suited to the complex spectral characteristics of multispectral remote sensing data. The study highlights the potential of integrating Sentinel-2 data with machine learning algorithms on the GEE platform for land cover classification and mapping, and provides a scientific basis for selecting appropriate classification algorithms in remote sensing studies.

Keywords: *Sentinel-2, land cover classification, Random Forest, Support Vector Machine, GEE*

Submission received: 16/03/2026 Revised: 06/04/2026 Accepted:10/04/2026 Published:30/04/2026

1. Mở đầu

Lớp phủ bề mặt (land cover) phản ánh trạng thái vật lý của bề mặt Trái Đất và là một trong những thông tin quan trọng phục vụ nghiên cứu môi trường, quản lý tài nguyên và quy hoạch phát triển kinh tế – xã hội. Việc xác định và theo dõi sự phân bố của các loại lớp phủ cho phép đánh giá những biến động của hệ sinh thái và tác động của hoạt động con người lên môi trường tự nhiên [1,2]. Trong bối cảnh gia tăng dân số, mở rộng đô thị và biến đổi khí hậu toàn cầu, thông tin lớp phủ có vai trò quan trọng trong giám sát tài nguyên đất, đánh giá phát thải khí nhà kính và xây dựng các chiến lược phát triển bền vững [3-5]. Nhiều nghiên cứu gần đây cho thấy sự thay đổi lớp phủ đang diễn ra nhanh chóng ở nhiều khu vực trên thế giới, đặc biệt tại các quốc gia đang phát triển, làm gia tăng nhu cầu xây dựng các cơ sở dữ liệu lớp phủ có độ chính xác cao và cập nhật thường xuyên [1,3, 6-8].

Trong bối cảnh đó, công nghệ viễn thám đã khẳng định vai trò then chốt nhờ khả năng cung cấp dữ liệu quan sát Trái Đất một cách liên tục, đồng nhất trên phạm vi rộng [9, 10]. Đáng chú ý, dữ liệu vệ tinh Sentinel-2 với độ phân giải không gian 10–20m và chu kỳ lặp ngắn đã trở thành nguồn tài nguyên mang tính cách mạng trong phân loại lớp phủ chi tiết [11,12]. Hệ thống kênh phổ đa dạng, trải dài từ vùng nhìn thấy (visible) đến hồng ngoại sóng ngắn (SWIR), không chỉ hỗ trợ phân biệt rõ nét các đối tượng thảm thực vật, đất nông nghiệp, mặt nước và khu vực xây dựng mà còn nâng cao đáng kể độ chính xác cho các mô hình phân loại phức tạp [13-15].

Sự kết hợp giữa nguồn dữ liệu chất lượng cao và các nền tảng điện toán đám mây như Google Earth Engine (GEE) đã tạo ra một bước tiến mới trong xử lý dữ liệu địa không gian quy mô lớn (Big Data). GEE cho phép truy cập trực tiếp kho dữ liệu vệ tinh toàn cầu và tích hợp các công cụ phân tích mạnh mẽ, giúp tối ưu hóa thời gian xử lý và giảm thiểu rào cản về hạ tầng kỹ thuật đối với các nghiên cứu giám sát môi trường [16, 17]. Trên nền tảng này, các thuật toán AI đã chứng minh ưu thế vượt trội so với phương pháp truyền thống nhờ khả năng xử lý hiệu quả các mối quan hệ phi tuyến giữa các đối tượng địa lý [18, 19]. Các mô hình sử dụng trí tuệ nhân tạo như Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, Gradient Boost (GBoost), Vision Transformer (ViT) đã được sử dụng rộng rãi trong phân loại ảnh viễn thám và thành lập bản đồ lớp phủ bề mặt tại Việt Nam [20-24].

Tuy nhiên, kết quả phân loại lớp phủ bề mặt thường phụ thuộc mạnh vào đặc điểm khu vực nghiên cứu, đặc trưng của dữ liệu ảnh vệ tinh và thuật toán được sử dụng trong quá trình xử lý và phân loại ảnh. Do đó, việc so sánh và đánh giá hiệu quả của các thuật toán học máy trên cùng một nguồn dữ liệu và trong cùng một môi

trường xử lý là cần thiết nhằm xác định phương pháp phù hợp cho từng điều kiện nghiên cứu cụ thể. Việc sử dụng dữ liệu Sentinel-2 kết hợp với hệ sinh thái xử lý trên nền tảng Google Earth Engine cho phép thực hiện các thí nghiệm so sánh thuật toán một cách nhất quán và hiệu quả, qua đó làm rõ sự khác biệt về độ chính xác phân loại, khả năng tổng quát hóa và mức độ phù hợp của từng thuật toán đối với các loại lớp phủ bề mặt.

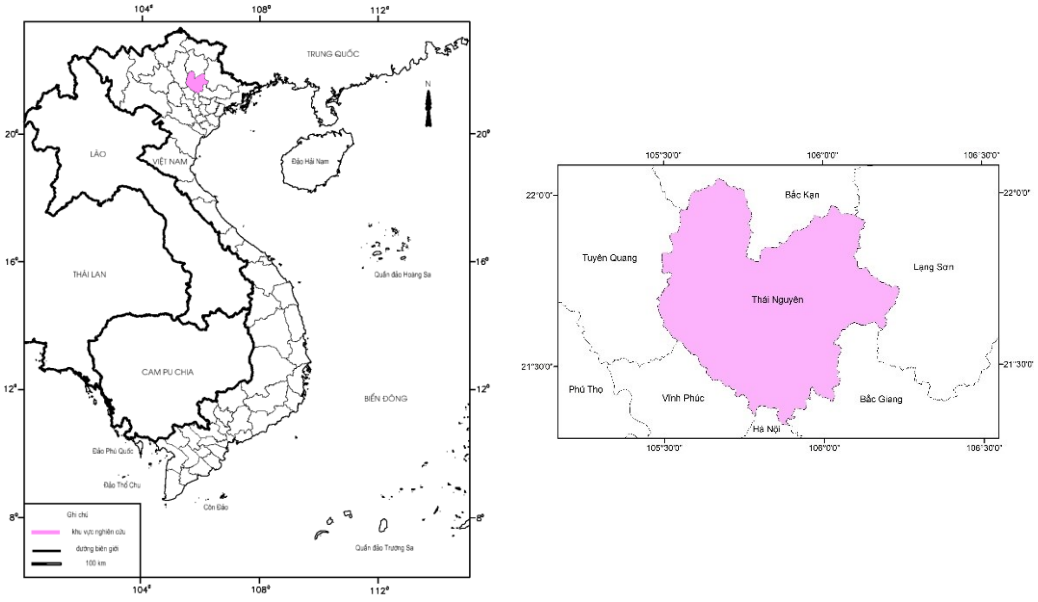
Trong nghiên cứu này, ba thuật toán học máy phổ biến được lựa chọn để phân loại lớp phủ bề mặt bao gồm Random Forest (RF), Support Vector Machine (SVM) và Naive Bayes (NB). Trong đó, Random Forest (RF) có khả năng hạn chế hiện tượng quá khớp (overfitting) nhờ cơ chế tổ hợp nhiều cây quyết định được xây dựng từ các tập mẫu và biến ngẫu nhiên [25]. Support Vector Machine (SVM) được đánh giá cao trong các bài toán viễn thám nhờ khả năng tổng quát hóa tốt và vẫn đạt độ chính xác cao ngay cả khi số lượng mẫu huấn luyện hạn chế [26], Naive Bayes (NB) là thuật toán có cấu trúc đơn giản và tốc độ tính toán nhanh, phù hợp với các bài toán phân loại dữ liệu lớn [2]. Các thuật toán này được áp dụng để phân loại lớp phủ bề mặt từ dữ liệu ảnh Sentinel-2 trên nền tảng Google Earth Engine tại khu vực tỉnh Thái Nguyên. Kết quả nghiên cứu không chỉ góp phần đánh giá hiệu quả của các thuật toán học máy trong điều kiện dữ liệu viễn thám tại Việt Nam mà còn cung cấp cơ sở khoa học cho việc lựa chọn phương pháp phân loại phù hợp phục vụ công tác quản lý tài nguyên và giám sát môi trường tại địa phương.

2. Khu vực và dữ liệu nghiên cứu

Khu vực nghiên cứu được lựa chọn là tỉnh Thái Nguyên (ranh giới hành chính trước khi sáp nhập), nằm ở vùng trung du và miền núi phía Bắc Việt Nam. Thái Nguyên có tọa độ địa lý khoảng từ 21°20′ đến 22°25′ vĩ độ Bắc và từ 105°25′ đến 106°16′ kinh độ Đông, tiếp giáp với các tỉnh Bắc Kạn ở phía Bắc, Vĩnh Phúc và Tuyên Quang ở phía Tây, Lạng Sơn và Bắc Giang ở phía Đông, và thủ đô Hà Nội ở phía Nam. Với diện tích tự nhiên khoảng 3.562 km², khu vực này có địa hình đa dạng bao gồm đồi núi thấp, thung lũng và các khu vực đồng bằng xen kẽ, tạo nên sự phân hóa rõ rệt về các loại lớp phủ bề mặt.

Các dữ liệu được sử dụng cho nghiên cứu:

- Dữ liệu ảnh vệ tinh Sentinel-2 Surface Reflectance (COPERNICUS/S2_SR_HARMONIZED) được truy cập và xử lý trên nền tảng Google Earth Engine (GEE). Bộ dữ liệu này thuộc chương trình Copernicus và được cung cấp bởi Cơ quan Vũ trụ châu Âu (ESA). Các ảnh Sentinel-2 được lựa chọn trong khoảng thời gian từ tháng 01 đến tháng 12 năm 2020. Điều kiện lọc ảnh được áp dụng với tỷ lệ che phủ mây nhỏ hơn 30% cho từng cảnh ảnh nhằm đảm bảo chất lượng dữ liệu đầu vào. Sau đó, các pixel bị ảnh hưởng bởi mây và bóng mây tiếp tục được loại bỏ bằng lớp Scene Classification Layer (SCL) trước khi thực hiện các bước xử lý tiếp theo. Các kênh phổ được sử dụng trong nghiên cứu bao gồm B2 (Blue), B3 (Green), B4 (Red), B8 (Near Infrared), B11 và B12 (Shortwave Infrared) với độ phân giải không gian từ 10–20 m, cung cấp thông tin phản xạ phổ quan trọng phục vụ cho việc tính toán các chỉ số phổ và phân loại lớp phủ bề mặt.



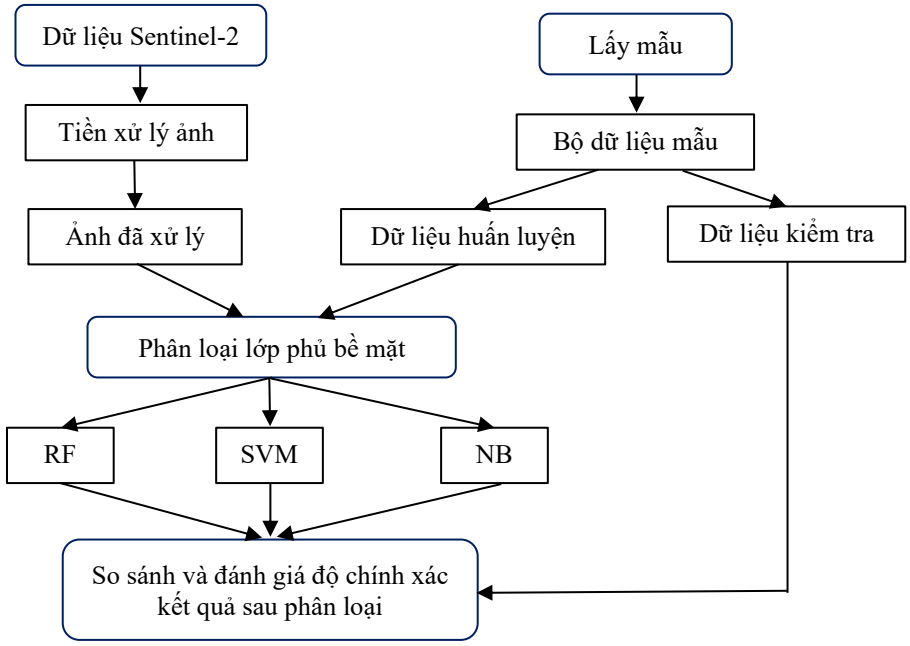
Hình 1. Khu vực thực nghiệm

- Dữ liệu mô hình số độ cao (DEM) từ Shuttle Radar Topography Mission (SRTM) với độ phân giải không gian 30 m do NASA cung cấp trên nền tảng Google Earth Engine được sử dụng để trích xuất các biến địa hình gồm độ cao (elevation) và độ dốc (slope). Trong quá trình xử lý, dữ liệu DEM được nội suy (resample) về độ phân giải 10 m để phù hợp với dữ liệu Sentinel-2. Các biến địa hình này được sử dụng như dữ liệu bổ trợ nhằm cải thiện khả năng phân biệt giữa các lớp phủ có đặc trưng phổ tương tự nhưng khác nhau về điều kiện địa hình.

- Dữ liệu ảnh độ phân giải rất cao được khai thác từ nền tảng Google Earth Pro, bao gồm các nguồn ảnh từ Maxar Technologies và CNES/Airbus Defence and Space, tại thời điểm năm 2020. Các ảnh này có độ phân giải không gian rất cao (nhỏ hơn 1 m), cho phép nhận dạng trực quan các đối tượng bề mặt như công trình xây dựng, thảm thực vật, mặt nước và khu vực đất trống. Bộ dữ liệu được sử dụng làm nguồn tham chiếu để số hóa các mẫu huấn luyện và mẫu kiểm tra phục vụ cho quá trình phân loại lớp phủ bề mặt.

3. Phương pháp nghiên cứu

Quy trình nghiên cứu so sánh hiệu quả các thuật toán Random Forest, SVM và Naive Bayes trong phân loại lớp phủ bề mặt sử dụng dữ liệu Sentinel-2 được thể hiện ở sơ đồ Hình 2.



Hình 2. Quy trình phân loại lớp phủ sử dụng dữ liệu Sentinel-2 trên Google Earth Engine

3.1. Tiền xử lý dữ liệu ảnh Sentinel-2

Dữ liệu ảnh vệ tinh Sentinel-2 Surface Reflectance được thu thập và xử lý trên nền tảng Google Earth Engine trong phạm vi khu vực nghiên cứu cho năm 2020. Quá trình tiền xử lý nhằm tạo ra bộ dữ liệu đầu vào có chất lượng tốt trước khi tiến tính toán chỉ số phổ và đưa vào phân loại. Các bước tiền xử lý bao gồm:

- Lọc theo thời gian: Lựa chọn các ảnh được thu nhận trong khoảng thời gian nghiên cứu nhằm đảm bảo tính đồng nhất của dữ liệu.

- Lọc mây và bóng mây: Sử dụng lớp Scene Classification Layer (SCL) của Sentinel-2 để loại bỏ các pixel bị che phủ bởi mây, bóng mây và các đối tượng không mong muốn.

- Tạo ảnh tổng hợp: Sau khi loại bỏ các pixel bị ảnh hưởng bởi mây và bóng mây, các ảnh hợp lệ được tổng hợp bằng phương pháp tính giá trị trung vị (median composite) nhằm tạo ra một ảnh đại diện cho toàn bộ thời gian nghiên cứu và giảm ảnh hưởng của nhiễu hoặc các giá trị ngoại lai trong chuỗi ảnh Sentinel-2. Phương pháp median composite giúp giảm ảnh hưởng của các pixel nhiễu còn sót lại và đảm bảo giữ được đặc trưng phổ ổn định của các lớp phủ, từ đó nâng cao độ tin cậy của dữ liệu đầu vào phục vụ cho quá trình phân loại.

- Cắt theo ranh giới khu vực nghiên cứu: Ảnh tổng hợp được cắt theo ranh giới khu vực nghiên cứu nhằm giới hạn phạm vi phân tích và giảm dung lượng dữ liệu cho các bước xử lý tiếp theo.

3.2. Tính toán các chỉ số phổ và dữ liệu bổ trợ

Từ các kênh phổ của Sentinel-2, các chỉ số phổ và dữ liệu bổ trợ được tính toán nhằm tăng khả năng phân biệt giữa các loại lớp phủ. Ngoài ra, các biến địa

hình bao gồm độ cao (Elevation) và độ dốc (Slope) được trích xuất từ dữ liệu DEM và tích hợp cùng với các kênh phổ, chỉ số phổ và đặc trưng kết cấu để tạo thành tập biến đầu vào cho các mô hình phân loại học máy, thể hiện ở Bảng 1. Trong quá trình phân loại, các biến địa hình đóng vai trò là dữ liệu phụ trợ (auxiliary variables) giúp cải thiện khả năng phân biệt giữa các lớp phủ có đặc trưng phổ tương tự nhưng phân bố khác nhau theo điều kiện địa hình, đặc biệt giữa các loại rừng, đất nông nghiệp và khu vực xây dựng.

Bảng 1. Các chỉ số phổ và dữ liệu bổ trợ

Nhóm dữ liệu	Dữ liệu đầu vào phân loại
Chỉ số phổ (Spectral indices)	NDVI, NDBI, MNDWI, BSI, NDVI_SD (độ lệch chuẩn NDVI)
Yếu tố địa hình	Độ cao (Elevation), Độ dốc (Slope)
Đặc trưng kết cấu ảnh (Texture)	Contrast (Độ tương phản), Entropy (Độ hỗn loạn), ASM (Angular Second Moment – mô men góc bậc hai)

3.3. Thu thập mẫu: huấn luyện – kiểm tra

Dữ liệu mẫu đại diện cho các loại lớp phủ bề mặt được thu thập để phục vụ huấn luyện và kiểm định mô hình. Các điểm mẫu (point) và vùng mẫu (polygon) đại diện cho từng loại lớp phủ được lựa chọn dựa trên ảnh vệ tinh độ phân giải rất cao từ Google Earth và thông tin thực địa. Trong nghiên cứu này, các polygon mẫu được xây dựng nhằm bao phủ đầy đủ sự biến thiên phổ của từng loại lớp phủ trong khu vực nghiên cứu. Từ các polygon này, các pixel đại diện được trích xuất trên nền tảng Google Earth Engine để xây dựng tập dữ liệu huấn luyện và kiểm tra cho các mô hình phân loại. Việc sử dụng polygon thay vì chỉ sử dụng các điểm mẫu đơn lẻ giúp giảm sai số do pixel hỗn hợp và phản ánh tốt hơn đặc trưng phổ của từng lớp phủ trong dữ liệu ảnh vệ tinh.

Số lượng mẫu của từng lớp phủ được lựa chọn nhằm đảm bảo tính đại diện cho sự biến thiên phổ của các đối tượng trong khu vực nghiên cứu. Các lớp có diện tích lớn và mức độ biến thiên phổ cao như rừng tự nhiên và rừng sản xuất được bố trí số lượng mẫu nhiều hơn nhằm nâng cao độ ổn định và khả năng tổng quát hóa của mô hình phân loại. Số lượng mẫu cho từng lớp phủ được xác định dựa trên diện tích phân bố của lớp phủ trong khu vực nghiên cứu và mức độ biến thiên phổ của đối tượng. Các lớp bao gồm: nước, đất ở, đất công nghiệp, lúa, đất trống, rừng tự nhiên và rừng sản xuất.

Tập mẫu được chia thành hai phần: 70% dữ liệu huấn luyện (Train) dùng để xây dựng mô hình; 30% dữ liệu kiểm định (Test): dùng để đánh giá độ chính xác của mô hình. Việc tách dữ liệu nhằm đảm bảo quá trình phân loại và đánh giá được thực hiện phân loại một cách độc lập và khách quan.

3.4. Huấn luyện các thuật toán học máy và thực hiện phân loại ảnh

Ba thuật toán học máy được sử dụng để xây dựng mô hình phân loại:

Random Forest (RF): thuật toán dựa trên tập hợp nhiều cây quyết định, có khả năng xử lý dữ liệu đa chiều và giảm hiện tượng quá khớp.

Support Vector Machine (SVM): tìm siêu phẳng tối ưu để phân tách các lớp dữ liệu trong không gian đặc trưng.

Naive Bayes: mô hình xác suất dựa trên định lý Bayes với giả định độc lập giữa các biến.

3.5 Đánh giá, so sánh kết quả phân loại theo các phương pháp

Kết quả phân loại được đánh giá bằng tập mẫu kiểm tra thông qua các chỉ tiêu:

- Confusion Matrix: Ma trận sai số thể hiện mức độ nhầm lẫn giữa các lớp.

- Overall Accuracy: Độ chính xác tổng thể của mô hình.

- Kappa: Hệ số đánh giá mức độ phù hợp giữa kết quả phân loại và dữ liệu tham chiếu.

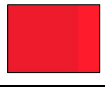
Các chỉ tiêu này được sử dụng để so sánh hiệu quả của các thuật toán phân loại và lựa chọn phương pháp có độ chính xác cao nhất trong việc xây dựng bản đồ lớp phủ bề mặt.

4. Kết quả nghiên cứu và thảo luận

4.1. Kết quả phân loại lớp phủ

Việc phân loại lớp phủ bề mặt được thực hiện theo sơ đồ quy trình trình bày tại Hình 1, sử dụng ba thuật toán học máy giám sát gồm Random Forest (RF), Support Vector Machine (SVM) và Naive Bayes (NB). Quá trình giải đoán và xây dựng mẫu huấn luyện được thực hiện cho 7 lớp đối tượng bao gồm: nước, đất ở, đất khu công nghiệp, lúa, đất trống, rừng tự nhiên và rừng sản xuất; chi tiết các lớp đối tượng được trình bày trong Bảng 2.

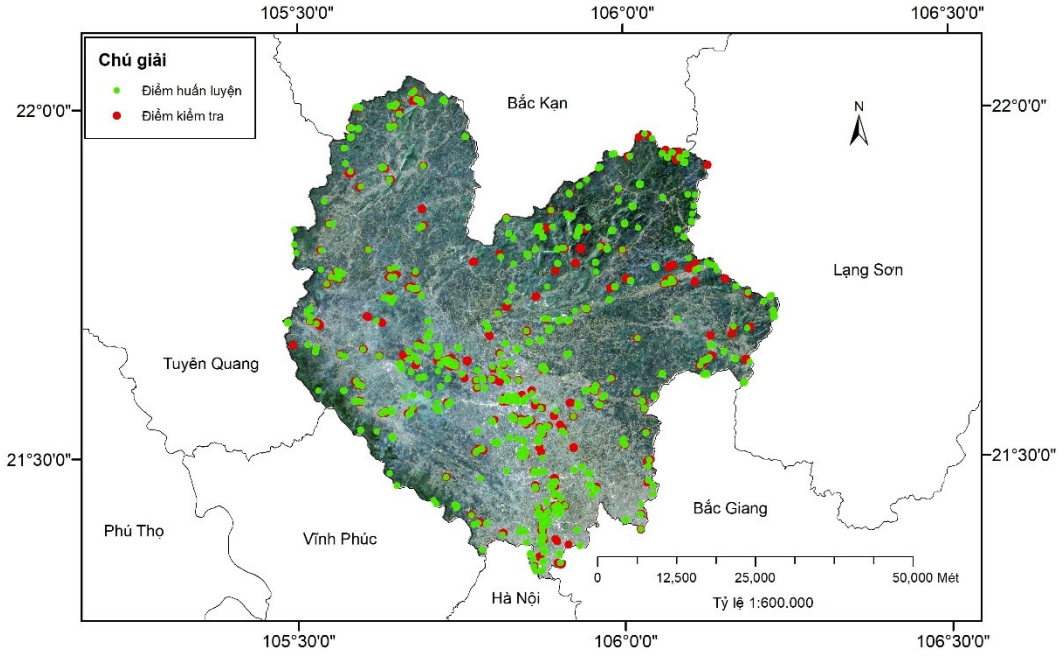
Bảng 2. Các mẫu giải đoán ảnh

TT	Lớp đối tượng	Mô tả cấu trúc chi tiết	Mẫu trên ảnh tổ hợp màu giả chuẩn
1	Nước	Sông, suối, hồ, ao và các vùng mặt nước tự nhiên hoặc nhân tạo. Bề mặt tương đối đồng nhất, ít kết cấu, ranh giới thường rõ theo hình dạng thủy hệ.	
2	Đất ở	Khu dân cư, nhà ở, công trình xây dựng và hệ thống giao thông trong khu dân cư. Cấu trúc không gian phức tạp, xen kẽ giữa nhà ở, đường giao thông và cây xanh.	
3	Đất khu công nghiệp	Khu nhà xưởng, nhà máy, kho bãi và các công trình hạ tầng công nghiệp. Cấu trúc dạng khối lớn, bố trí theo quy hoạch tương đối đồng đều.	
4	Lúa	Đất trồng lúa, thường phân bố thành các thửa ruộng nhỏ có bờ bao rõ ràng. Đặc trưng phổ thay đổi theo giai đoạn sinh trưởng.	
5	Đất trống	Các loại đất bỏ hoang, đá trơ hoặc đất chưa có thảm phủ thực vật. Bề mặt thường không đồng nhất và ít che phủ thực vật.	
6	Rừng tự nhiên	Khu rừng có cấu trúc sinh thái tự nhiên, đa dạng loài và nhiều tầng tán. Tán cây dày, độ che phủ lớn.	
7	Rừng sản xuất	Khu rừng trồng phục vụ mục đích sản xuất lâm nghiệp như keo, bạch đàn. Cấu trúc tương đối đồng đều theo lô trồng.	

Quá trình xây dựng và huấn luyện mô hình được thực hiện trên nền tảng

Google Earth Engine, sử dụng tập dữ liệu gồm khoảng 18.524 pixel mẫu đại diện cho 7 lớp phủ bề mặt. Ba thuật toán Random Forest (RF), Support Vector Machine (SVM) và Naive Bayes (NB) được cấu hình và huấn luyện song song nhằm đánh giá và so sánh hiệu quả phân loại.

Phân bố không gian của các điểm huấn luyện và kiểm tra được thể hiện trong Hình 3. Số lượng pixel mẫu của từng lớp phủ cho hai tập dữ liệu được thống kê trong Bảng 3. Trong đó, 70% số mẫu được sử dụng để huấn luyện mô hình, trong khi 30% còn lại được dùng làm tập kiểm tra nhằm đánh giá độ chính xác phân loại.



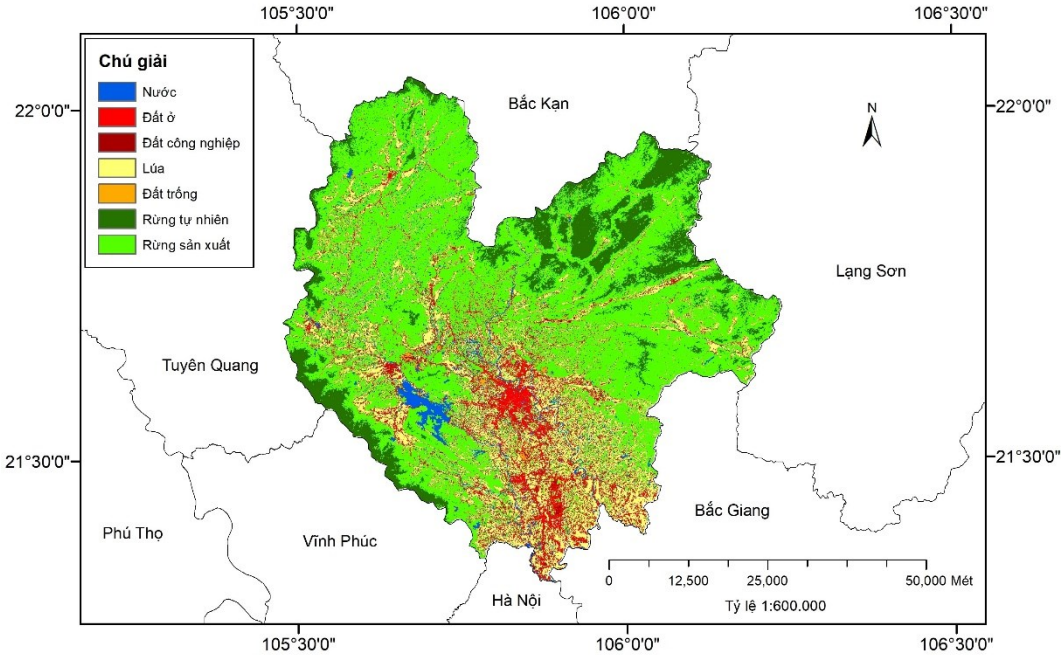
Hình 3. Sơ đồ vị trí các điểm huấn luyện và điểm kiểm tra

Bảng 3. Thống kê số lượng mẫu thực nghiệm cho 7 lớp phủ

Mã lớp	Tên lớp phủ	Số mẫu huấn luyện	Số mẫu kiểm tra	Tổng số lượng mẫu
1	Nước	1821	900	2721
2	Đất ở	1527	350	1877
3	Đất công nghiệp	1731	715	2446
4	Lúa	1927	1440	3367
5	Đất trống	956	218	1174
6	Rừng tự nhiên	5186	1826	7012
7	Rừng sản xuất	5376	903	6279

Sau khi các mô hình được huấn luyện, ba thuật toán RF, SVM và NB được triển khai để thực hiện phân loại lớp phủ bề mặt từ dữ liệu Sentinel-2. Kết quả phân loại được biểu diễn dưới dạng các bản đồ lớp phủ bề mặt tương ứng với từng thuật toán, làm cơ sở cho việc so sánh và đánh giá hiệu quả phân loại. Các bản đồ kết quả cũng như độ chính xác phân loại của từng phương pháp được trình bày trong các phần tiếp theo.

- Đối với thuật toán RF, bản đồ kết quả phân loại được thể hiện tại Hình 4.



Hình 4. Kết quả phân loại lớp phủ bề mặt bằng thuật toán RF

Thuật toán RF với thiết lập 80 cây quyết định cho thấy hiệu suất phân loại cao và ổn định nhất trong ba thuật toán được thử nghiệm. Kết quả đánh giá từ ma trận nhầm lẫn tính toán trên Google Earth Engine cho thấy mô hình RF đạt độ chính xác tổng thể (OA) là 90,27% và hệ số Kappa đạt 0,880, cho thấy mức độ phù hợp cao giữa kết quả phân loại và dữ liệu tham chiếu. Ma trận nhầm lẫn của RF được trình bày trong Bảng 4, trong đó các giá trị trên đường chéo chính biểu thị số lượng điểm ảnh được phân loại đúng.

Bảng 4. Ma trận nhầm lẫn của mô hình RF

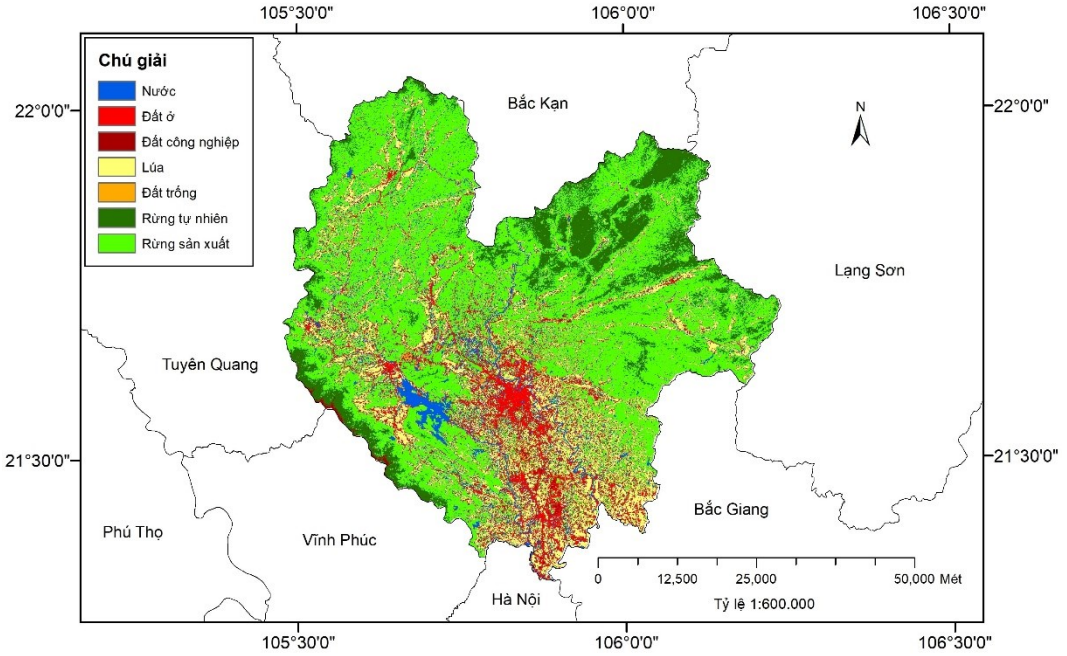
Lớp	Nước	Đất ở	Đất công nghiệp	Lúa	Đất trống	Rừng tự nhiên	Rừng sản xuất
Nước	885	0	1	11	0	0	3
Đất ở	0	325	21	1	2	0	1
Đất công nghiệp	1	72	663	5	4	0	0
Lúa	1	57	2	1376	0	0	4
Đất trống	2	3	5	4	204	0	0
Rừng tự nhiên	0	0	0	0	0	1568	258
Rừng sản xuất	0	0	0	4	0	156	743

Ma trận nhầm lẫn của mô hình RF ở Bảng 4 cho thấy các lớp phủ được phân tách tương đối tốt với phần lớn các giá trị nằm trên đường chéo chính. Tuy nhiên, vẫn tồn tại nhầm lẫn giữa các lớp có đặc trưng phổ tương đồng, đặc biệt giữa đất ở và đất công nghiệp, cũng như giữa rừng tự nhiên và rừng sản xuất. Ngoài ra, lớp lúa cũng xuất hiện một số nhầm lẫn với lớp đất ở do ảnh hưởng của các pixel hỗn hợp tại vùng chuyển tiếp giữa khu dân cư và khu vực canh tác. Kết quả đánh giá theo từng lớp cho thấy lớp nước và lúa đạt độ chính xác cao nhất, trong khi lớp rừng sản xuất có độ chính xác thấp hơn do sự tương đồng phổ với các lớp thực vật khác.

- Đối với thuật toán SVM, bản đồ kết quả phân loại được thể hiện tại Hình 5.

Thuật toán SVM được thiết lập với hạt nhân RBF (Radial Basis Function), tham số Gamma = 0,1 và Cost = 5 để thực hiện phân loại lớp phủ bề mặt từ dữ liệu Sentinel-2. Kết quả đánh giá trên tập kiểm tra cho thấy mô hình đạt độ chính xác tổng thể (OA) 88,78% và hệ số Kappa 0,862, phản ánh mức độ phù hợp cao giữa kết quả phân loại và dữ liệu tham chiếu. Ma trận nhầm lẫn của SVM được trình bày trong Bảng 5.

Ma trận nhầm lẫn của thuật toán SVM cho thấy các lớp phủ bề mặt được phân tách tương đối tốt với phần lớn các giá trị tập trung trên đường chéo chính. Tuy nhiên, vẫn tồn tại một số nhầm lẫn giữa các lớp có đặc trưng phổ tương đồng. Cụ thể, lớp đất ở và đất công nghiệp có sự nhầm lẫn đáng kể với 22 mẫu đất ở bị phân loại nhầm sang đất công nghiệp và 85 mẫu theo chiều ngược lại. Lớp lúa cũng xuất hiện một số nhầm lẫn với đất ở (55 mẫu) và đất công nghiệp (106 mẫu), phản ánh sự tương đồng phổ giữa các khu vực nông nghiệp và bề mặt xây dựng. Ngoài ra, sự giao thoa giữa các lớp thực vật vẫn xảy ra, đặc biệt giữa rừng tự nhiên và rừng sản xuất với 222 và 136 mẫu bị phân loại nhầm lẫn giữa hai lớp này.



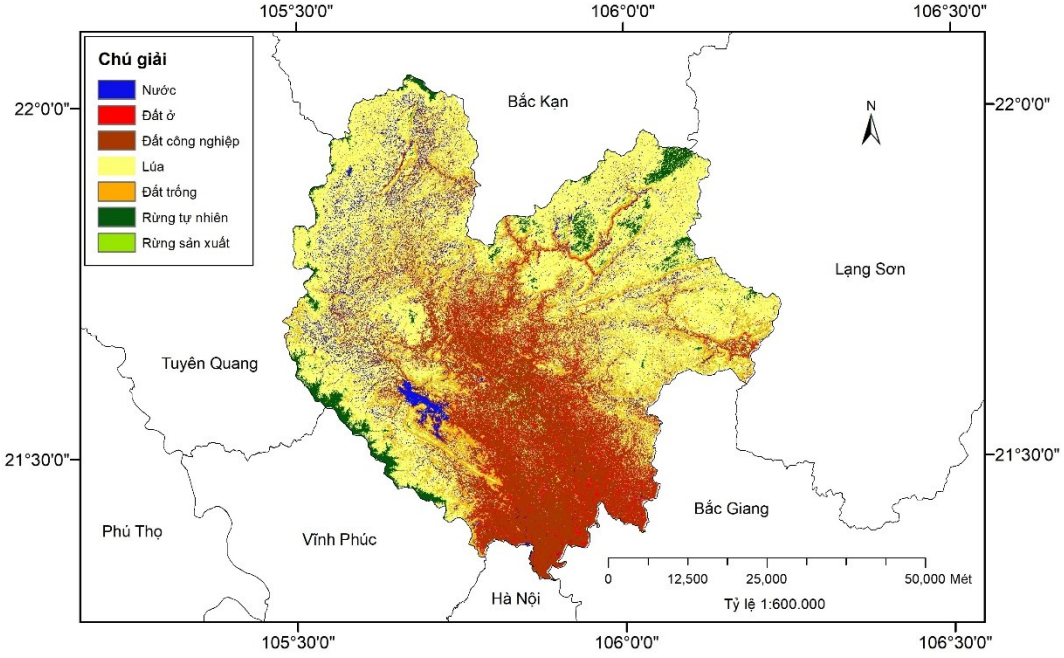
Hình 5. Kết quả phân loại lớp phủ bề mặt bằng thuật toán SVM

Bảng 5. Ma trận nhầm lẫn của mô hình SVM

Lớp	Nước	Đất ở	Đất công nghiệp	Lúa	Đất trống	Rừng tự nhiên	Rừng sản xuất
Nước	881	0	1	17	0	0	1
Đất ở	1	323	22	1	2	0	1
Đất công nghiệp	9	85	618	2	1	0	0
Lúa	5	55	106	1267	0	0	7
Đất trống	9	0	27	0	182	0	0
Rừng tự nhiên	0	0	0	0	0	1604	222
Rừng sản xuất	0	0	0	3	0	136	764

- Đối với thuật toán NB, bản đồ kết quả phân loại được thể hiện tại Hình 6.

Kết quả phân loại bằng thuật toán Naive Bayes (NB) cho thấy sự phân bố lớp phủ chưa phù hợp với thực tế. Lớp lúa chiếm tỷ lệ diện tích rất lớn (khoảng 53,2% diện tích toàn tỉnh), do mô hình có xu hướng gán nhãn lớp này cho nhiều khu vực rừng trồng, đồi thấp và các dạng đất nông nghiệp khác. Đồng thời, diện tích của một số lớp phủ khác bị giảm đáng kể như đất ở và rừng sản xuất. Điều này phản ánh khả năng phân biệt lớp phủ của mô hình NB còn hạn chế. Ma trận nhầm lẫn của NB được trình bày trong Bảng 6.



Hình 6. Kết quả phân loại lớp phủ bề mặt bằng thuật toán NB

Bảng 6. Ma trận nhầm lẫn của mô hình NB

Lớp	Nước	Đất ở	Đất công nghiệp	Lúa	Đất trống	Rừng tự nhiên	Rừng sản xuất
Nước	493	18	80	86	19	204	0
Đất ở	17	2	213	116	2	0	0
Đất công nghiệp	1	4	627	76	7	0	0
Lúa	69	432	182	735	22	0	0
Đất trống	46	0	0	119	53	0	0
Rừng tự nhiên	7	0	0	1306	47	436	3
Rừng sản xuất	12	13	4	557	302	12	3

Ma trận nhầm lẫn của mô hình NB cho thấy sự phân tách giữa các lớp phủ chưa rõ ràng, với nhiều điểm ảnh phân bố ngoài đường chéo chính. Lớp lúa có mức độ nhầm lẫn cao khi nhiều mẫu của các lớp khác bị gán nhãn sang lớp này, đặc biệt từ rừng tự nhiên (1306 mẫu) và rừng sản xuất (557 mẫu). Bên cạnh đó, lớp đất ở cũng có sự nhầm lẫn đáng kể với đất công nghiệp và lúa. Điều này cho thấy khả năng phân biệt giữa các lớp phủ của mô hình NB còn hạn chế khi áp dụng cho dữ liệu Sentinel-2.

4.2. Đánh giá các thuật toán phân loại

Kết quả đánh giá độ chính xác của ba thuật toán phân loại được thể hiện trong Bảng 7.

Bảng 7. Độ chính xác của các thuật toán phân loại

Thuật toán	Overall Accuracy	Kappa
Random Forest	90,27%	0,880
Support Vector Machine	88,78%	0,862
Naive Bayes	37,41%	0,238

Bảng 7 cho thấy sự khác biệt rõ rệt về hiệu quả giữa ba thuật toán phân loại. Trong đó, RF đạt độ chính xác cao nhất với OA đạt 90,27% và hệ số Kappa đạt 0,880, cho thấy khả năng mô hình hóa tốt mối quan hệ phi tuyến giữa các biến phổ và lớp phủ bề mặt. Thuật toán này thể hiện khả năng phân tách các lớp phủ tương đối ổn định và hạn chế sự nhầm lẫn giữa các đối tượng có đặc trưng phổ gần nhau. SVM cũng đạt hiệu quả phân loại cao với OA đạt 88,78% và Kappa đạt 0,862. Mặc dù thấp hơn RF một mức nhỏ, SVM vẫn cho thấy khả năng phân loại tốt đối với dữ liệu ảnh Sentinel-2 đa phổ và đảm bảo độ tin cậy trong thành lập bản đồ lớp phủ bề mặt. Ngược lại, NB cho độ chính xác thấp hơn đáng kể với OA chỉ đạt 37,41% và Kappa đạt 0,238, cho thấy khả năng mô hình hóa của phương pháp này còn hạn chế đối với dữ liệu viễn thám đa chiều. Điều này cho thấy giả định độc lập giữa các biến đầu vào của NB chưa phù hợp với đặc trưng phổ phức tạp của ảnh Sentinel-2, dẫn đến sự suy giảm đáng kể về độ chính xác phân loại.

4.3. Thảo luận

Trong nghiên cứu này, RF đạt độ chính xác phân loại cao nhất, tiếp theo là SVM, trong khi NB cho kết quả thấp hơn đáng kể. Xu hướng này phù hợp với nhiều nghiên cứu trước đây về phân loại lớp phủ bề mặt sử dụng dữ liệu viễn thám. RF thường cho hiệu quả cao trong phân loại ảnh viễn thám nhờ khả năng xử lý dữ liệu đa chiều và mô hình hóa các mối quan hệ phi tuyến giữa các biến phổ. Cơ chế tổ hợp nhiều cây quyết định giúp tăng tính ổn định của mô hình và giảm hiện tượng quá khớp [25]. Trong khi đó, SVM là một thuật toán phân loại mạnh, thường được áp dụng trong xử lý dữ liệu viễn thám. Phương pháp này hoạt động bằng cách xác định ranh giới phân tách tối ưu giữa các lớp dữ liệu trong không gian đặc trưng, giúp cải thiện khả năng phân biệt giữa các loại lớp phủ bề mặt.Tuy nhiên, hiệu quả của SVM có thể phụ thuộc đáng kể vào việc lựa chọn hàm kernel và tham số mô hình [26]. Trong khi đó, NB thường cho độ chính xác thấp hơn do giả định các biến đầu vào độc lập, trong khi các kênh phổ của ảnh viễn thám thường có mối tương quan nhất định.

Xu hướng này cũng phù hợp với nhiều nghiên cứu trước đây sử dụng dữ liệu Sentinel-2 trong phân loại lớp phủ bề mặt. Một số công trình tại Việt Nam cho thấy RF và SVM đều đạt độ chính xác cao trong thành lập bản đồ lớp phủ hoặc sử dụng đất từ dữ liệu viễn thám đa phổ [12, 18], [20, 21], [24]. Ngoài ra, việc khai thác dữ liệu viễn thám kết hợp với các thuật toán học máy trên nền tảng Google Earth Engine đã được chứng minh là phương pháp hiệu quả trong giám sát lớp phủ và biến động sử dụng đất [11], [15], [22]. Nhìn chung, kết quả của nghiên cứu này phù hợp với xu hướng chung của các công trình đã công bố về ứng dụng các thuật toán

học máy trong phân loại lớp phủ bề mặt từ dữ liệu viễn thám [2], [10].

5. Kết luận

Nghiên cứu đã thực hiện phân loại lớp phủ bề mặt tỉnh Thái Nguyên từ dữ liệu ảnh vệ tinh Sentinel-2 trên nền tảng Google Earth Engine bằng ba thuật toán học máy gồm RF, SVM và NB. Tập dữ liệu đầu vào được xây dựng từ các kênh ảnh Sentinel-2, các chỉ số phổ, yếu tố địa hình và đặc trưng kết cấu ảnh, với tổng số 18.524 pixel mẫu được sử dụng cho quá trình huấn luyện và kiểm tra mô hình.

Kết quả đánh giá độ chính xác cho thấy thuật toán RF đạt hiệu quả phân loại cao nhất với độ chính xác tổng thể (OA) 90,27% và hệ số Kappa 0,880. Thuật toán SVM cũng cho kết quả phân loại tốt với OA đạt 88,78% và Kappa 0,862, trong khi NB có độ chính xác thấp hơn đáng kể (OA 37,41%, Kappa 0,238). Kết quả này cho thấy RF và SVM có khả năng mô hình hóa tốt các mối quan hệ phi tuyến giữa các biến phổ và lớp phủ bề mặt, trong khi giả định độc lập giữa các biến của NB chưa phù hợp với đặc trưng phổ phức tạp của dữ liệu viễn thám.

Nhìn chung, việc kết hợp dữ liệu Sentinel-2 với các thuật toán học máy trên nền tảng Google Earth Engine cho thấy tiềm năng lớn trong thành lập bản đồ lớp phủ bề mặt và giám sát tài nguyên môi trường. Kết quả nghiên cứu cung cấp cơ sở khoa học cho việc lựa chọn thuật toán phân loại phù hợp trong các nghiên cứu viễn thám và hỗ trợ công tác quản lý tài nguyên, quy hoạch sử dụng đất tại địa phương.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Bộ Nông nghiệp và Môi trường thông qua đề tài khoa học và công nghệ "Nghiên cứu công nghệ viễn thám, xử lý dữ liệu lớn kết hợp trí tuệ nhân tạo phục vụ xác định lượng bù đắp các-bon từ rừng", mã số TNMT.ĐL.2025.08.01.

Cam kết của các tác giả

Các tác giả cam kết không có xung đột lợi ích trong quá trình thực hiện và công bố nghiên cứu này.

Tài liệu tham khảo

[1] M. Herold, P. Mayaux, C. Woodcock, A. Baccini, and C. Schmullius, "Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets," *Remote Sensing of Environment,* vol. 112, no. 5, pp. 2538-2556, 2008, doi: https://doi.org/10.1016/j.rse.2007.11.013.

[2] J. A. Richards and X. Jia, *Remote sensing digital image analysis: an introduction*. Springer, 2006.

[3] FAO, "FAOSTAT Analytical Brief 88 – Land Statistics 2001–2022. Global, regional and country trends," 2023.

[4] S. Hu *et al.*, "Converging trend of global urban land expansion sheds new light on sustainable development," *arXiv preprint arXiv:2310.02293,* 2023, doi: https://doi.org/10.48550/arXiv.2310.02293.

[5] H. A. Nguyễn, "Ứng dụng giải thuật trí tuệ nhân tạo phân loại và dự báo sự phân bố lớp phủ thực vật sử dụng ảnh Landsat – vùng nghiên cứu tại đới ven bờ của tỉnh Bà Rịa Vũng Tàu,"

Tạp chí Khoa học Đại học Cần Thơ, vol. 61, no. 2, pp. 67-79, 2025, doi: https://doi.org/10.22144/ctujos.2025.032.

[6] G. A. Afuye *et al.*, "Global trend assessment of land use and land cover changes: A systematic approach to future research development and planning," *Journal of King Saud University-Science,* vol. 36, no. 7, p. 103262, 2024, doi: https://doi.org/10.1016/j.jksus.2024.103262.

[7] J. Chen *et al.*, "Global land cover mapping at 30 m resolution: A POK-based operational approach," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 103, pp. 7-27, 2015, doi: https://doi.org/10.1016/j.isprsjprs.2014.09.002.

[8] M. A. Wulder, J. G. Masek, W. B. Cohen, T. R. Loveland, and C. E. Woodcock, "Opening the archive: How free data has enabled the science and monitoring promise of Landsat," *Remote Sensing of Environment,* vol. 122, pp. 2-10, 2012, doi: https://doi.org/10.1016/j.rse.2012.01.010.

[9] T. S. Unger Holtz, "Introductory digital image processing: A remote sensing perspective," ed: Association of Environmental & Engineering Geologists, 2007.

[10] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

[11] S. Gadal and G. Mozgeris, "Advances of remote sensing in land cover and cand use mapping," vol. 17, ed: MDPI, 2025, p. 1980.

[12] T. P. T. Giang, T. T. H. Phạm, V. H. Phạm, and A. B. Nguyễn, "Đánh giá độ chính xác trong phân loại lớp phủ dự trên thuật toán học máy và dữ liệu viễn thám thông qua Google Earth Engine: Ápdụng tại tỉnh Đắk Lắk," *Journal of Science on Natural Resources Environment,* no. 46, pp. 55-65, 2021.

[13] T. T. H. Lê and T. P. T. Giang, "Sử dụng ảnh vệ tinh Sentinel-2 trong giám sát sự phát triển của cây lúa tại tỉnh Đồng Tháp, Việt Nam," *Tạp chí Khí tượng thuỷ văn,* vol. 764, pp. 93-108, 2024, doi: https://doi.org/10.36335/VNJHM.2024(764).93-108.

[14] V. T. Nguyễn, T. N. P. Đoàn, and T. D. Bùi, "Sử dụng các chỉ số phổ của dữ liệu ảnh vệ tinh Sentinel-2 và Landsat-8 thành lập bản đồ mức độ cháy rừng ở xã Na Ngoi, Kỳ Sơn, Nghệ An," *Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất,* vol. 59, no. 5, pp. 44-54, 2018.

[15] T. N. T. N. Nguyễn, K. D. Nguyễn, and K. D. Phan, "Xây dựng bản đồ phân bố không gian hiện trạng sử dụng đất nông nghiệp huyện Tân Hưng, tỉnh Long An sử dụng kết hợp chuỗi ảnh Sentinel 2 và Sentinel 1," *Tạp chí Khoa học Đại học Cần Thơ,* vol. 61, pp. 144-154, 2025, doi: https://doi.org/10.22144/ctujos.2025.065.

[16] H. S. Nguyen, "Ứng dụng viễn thám và Google Earth Engine thành lập bản đồ hiện trạng sử dụng đất nông nghiệp năm 2023 ở huyện Hòa Vang, thành phố Đà Nẵng," *Hue University Journal of Science: Agriculture and Rural Development,* vol. 133, no. 3B, pp. 17–33-17–33, 2024, doi: https://doi.org/10.26459/hueunijard.v133i3B.7443.

[17] Đ. C. Nguyễn, Đ. C. Phạm, and V. B. Nguyễn, "Sử dụng ảnh Sentinel-2 và Google Earth Engine để đánh giá biến động diện tích rừng phòng hộ và đặc dụng tại huyện Võ Nhai, tỉnh Thái Nguyên," *Tạp chí Khoa học Lâm nghiệp,* vol. 1, pp. 106-114, 2022.

[18] T. T. H. Phạm, N. Q. Vũ, T. N. Lê, T. N. P. Đoàn, and M. H. H. Nguyễn, "Nghiên cứu khả năng ứng dụng thuật toán Random Forest và ảnh vệ tinh Sentinel-2 trong phân loại lớp phủ mặt đất tỉnh Quảng Bình trên nền tảng Google Colab," *Tạp chí Khí tượng thuỷ văn,* vol. 756, pp. 29-41, 2023, doi: https://doi.org/10.36335/VNJHM.2023(756).29-41.

[19] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing,* vol. 152, pp. 166-177, 2019, doi: https://doi.org/10.1016/j.isprsjprs.2019.04.015.

[20] V. A. Trần *et al.*, "Nghiên cứu một số phương pháp học máy trong thành lập bản đồ lớp phủ bề mặt tỉnh Cà Mau trên nền tảng Google Earth Engine," *Tạp chí Khoa học Đo đạc và Bản đồ,* no. 55, pp. 18-26, 2023.

[21] T. C. Nguyễn, Q. B. Trần, T. Đ. Trương, T. H. Nguyễn, V. D. Phạm, and H. H. Nguyễn, "Kết hợp trí tuệ nhân tạo (AI) và Google Earth Engine (GEE) để phân loại các lớp phủ từ ảnh Sentinel-2: trường hợp nghiên cứu tại xã Quảng Sơn và xã Tà Đùng, tỉnh Lâm Đồng," *Tạp chí Khoa học và công nghệ Lâm nghiệp,* vol. 14, no. 7, pp. 60-70, 2025, doi: https://doi.org/10.55250/jo.vnuf.14.7.2025.060-070.

[22] C. H. Phạm and T. N. Nguyễn, "Ứng dụng Google Earth Engine giám sát biến động không gian xanh tại thành phố Thủ Đức bằng ảnh Sentinel 2 giai đoạn 2019-2024," *Tạp chí Trắc địa - Bản đồ,* vol. 11, no. 3, pp. 25-39, 2025, doi: 10.5281/zenodo.15795229.

[23] T. O. Nông, X. T. Trần, H. T. Tạ, and V. N. Trịnh, "Mô hình tự động phân loại dữ liệu lớp phủ bề mặt phục vụ kiểm kê khí nhà kính bằng ảnh viễn thám," *Tạp chí Khoa học Đo đạc và Bản đồ,* no. 57, pp. 55-64, 2023.

[24] T. P. T. Đỗ, M. H. Lê, N. N. Nguyễn, T. T. H. Vũ, and K. V. Nguyễn, "Giám sát lớp phủ bề mặt khu dự trữ sinh quyển Cần Giờ sử dụng thuật toán Random Forest trên nền tảng điện toán đám mây," *Tạp chí Khí tượng thuỷ văn,* vol. 770, pp. 58-67, 2025, doi: https://doi.org/10.36335/VNJHM.2025(770).58-67.

[25] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing,* vol. 114, pp. 24-31, 2016, doi: https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[26] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS journal of photogrammetry and remote sensing,* vol. 66, no. 3, pp. 247-259, 2011, doi: https://doi.org/10.1016/j.isprsjprs.2010.11.001.