NGHIÊN CỨU CÁC GIẢI THUẬT PHÂN LOẠI TRONG GOOGLE EARTH ENGINE ĐỂ GIÁM SÁT LỚP PHỦ TỈNH NINH BÌNH

Đinh Bảo Ngọc[1], Lê Thị Kim Thoa[2], Lê Đức Hoàng[2], Đào Thị Lưu[2], Dương Thị Hồng Yến[2], Phí Thị Thu Hoàng[2], Ngô Thị Bích Hồng[2], Trịnh Xuân Quang[3]*

TÓM TẮT

Trong bối cảnh nhu cầu giám sát lớp phủ đất ngày càng tăng nhằm hỗ trợ quản lý tài nguyên thiên nhiên và quy hoạch đô thị, nghiên cứu này tập trung vào việc ứng dụng các giải thuật học máy Random Forest (RF), Support Vector Machine (SVM) và Classification and Regression Tree (CART) trên nền tảng Google Earth Engine (GEE) để phân loại và giám sát lớp phủ tại tỉnh Ninh Bình. Dữ liệu ảnh vệ tinh Sentinel-2 được sử dụng làm nguồn dữ liệu chính nhờ độ phân giải không gian cao và khả năng cập nhật thường xuyên. Kết quả nghiên cứu cho thấy hiệu suất của các giải thuật trong việc phân loại các loại lớp phủ như rừng, đất nông nghiệp, đô thị và mặt nước, đồng thời đánh giá độ chính xác của từng phương pháp. Nghiên cứu cung cấp cơ sở khoa học cho việc lựa chọn giải thuật phù hợp trong giám sát lớp phủ tại khu vực địa phương, góp phần vào quản lý bền vững tài nguyên đất đai.

Từ khóa: Google Earth Engine, Giải thuật phân loại, Random Forest, Support Vector Machine, Classigication and Regression Tree, Ninh Bình, Lớp phủ

[1] *Đại học Mỏ Địa chất*
[2] *Viện Các Khoa học Trái Đất, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[3] *Phân hiệu trường ĐH Tài nguyên và Môi trường Hà Nội tại tỉnh Thanh Hóa*
[*] *Email: dinhbaongoc1612@gmail.com*

I. GIỚI THIỆU

Tỉnh Ninh Bình là một khu vực điển hình của miền Bắc Việt Nam, nơi các hoạt động kinh tế - xã hội như nông nghiệp, du lịch và đô thị hóa đang gây áp lực lớn lên lớp phủ đất [Nguyen et al., 2020]. Sự thay đổi lớp phủ không chỉ phản ánh quá trình phát triển mà còn liên quan đến các vấn đề môi trường như xói mòn đất, ngập lụt và suy giảm đa dạng sinh học. Do đó, việc giám sát lớp phủ đất đai một cách chính xác và kịp thời là yếu tố quan trọng để hỗ trợ các nhà quản lý trong quy hoạch và bảo vệ tài nguyên thiên nhiên.

Công nghệ viễn thám đã trở thành công cụ không thể thiếu trong giám sát lớp phủ nhờ khả năng thu thập dữ liệu trên diện rộng và liên tục [Lillesand et al., 2015]. Kết hợp với các giải thuật học máy, việc phân loại lớp phủ từ dữ liệu vệ tinh đạt được độ chính xác cao hơn so với các phương pháp truyền thống [Maxwell et al., 2018]. Google Earth Engine (GEE), một nền tảng xử lý dữ liệu đám mây, cho phép truy cập và phân tích nhanh chóng các bộ dữ liệu vệ tinh lớn như Sentinel-2, Landsat, và MODIS [Gorelick et al., 2017]. Nghiên cứu này tập trung vào việc áp dụng ba giải thuật học máy phổ biến - RF [Breiman, 2001], SVM [Vapnik, 1995] và CART [Breiman et al., 1984] - để phân loại lớp phủ tại Ninh Bình, nhằm đánh giá hiệu suất của từng phương pháp trong bối cảnh địa lý và sinh thái đặc thù của khu vực.

Mục tiêu chính của nghiên cứu bao gồm: (1) Xây dựng bản đồ lớp phủ hiện trạng dựa trên dữ liệu Sentinel-2; (2) So sánh hiệu quả của RF, SVM và CART trong phân loại lớp phủ; (3) Đề xuất giải pháp tối ưu cho giám sát lớp phủ dài hạn tại Ninh Bình.

II. PHƯƠNG PHÁP NGHIÊN CỨU

1. Khu vực nghiên cứu

Tỉnh Ninh Bình nằm ở tọa độ địa lý từ 19°56′ đến 20°31′ vĩ độ Bắc và từ 105°40′ đến 106°08′ kinh độ Đông, với tổng diện tích tự nhiên khoảng 1.391 km² [General Statistics Office of Vietnam, 2023]. Khu vực này bao gồm các dạng địa hình đa dạng: Đồng bằng phù sa (chiếm khoảng 70% diện tích), đồi núi thấp (chủ yếu ở phía Tây Bắc, nơi có Vườn Quốc gia Cúc Phương), và vùng ven biển ngập mặn ở phía Đông Nam. Các lớp phủ chính tại đây bao gồm rừng tự nhiên, đất nông nghiệp (lúa nước, hoa màu), khu vực đô thị (thành phố Ninh Bình, Tam Điệp), và mặt nước (sông, hồ, đầm phá). Sự phân bố lớp phủ chịu ảnh hưởng lớn từ hoạt động con người và điều kiện tự nhiên như lũ lụt mùa mưa [Nguyen et al., 2020]. Dữ liệu ranh giới hành chính được lấy từ cơ sở dữ liệu GEE.

2. Dữ liệu

Dữ liệu chính được sử dụng trong nghiên cứu là ảnh vệ tinh Sentinel-2 Level-2A (đã được hiệu chỉnh khí quyển) từ GEE, thu thập trong khoảng thời gian từ tháng 1 đến tháng 12 năm 2024 [ESA, 2015]. Sentinel-2 cung cấp độ phân giải không gian 10m cho các băng phổ khả kiến (Blue, Green, Red) và cận hồng ngoại (Near-Infrared), phù hợp cho phân loại lớp phủ chi tiết [Drusch et al., 2012]. Ngoài ra, chỉ số thực vật chuẩn hóa (NDVI) được tính toán từ băng Red và Near-Infrared để tăng cường khả năng phân biệt giữa các loại thực vật và phi thực vật [Rouse et al., 1974].

Dữ liệu lấy mẫu gồm 80 điểm được lấy mẫu trên ảnh vệ tinh sentinel 2A vào năm 2024

3. Các giải thuật phân loại: Ba giải thuật phân loại được thử nghiệm:

Random Forest (RF): RF là một giải thuật học tập tổng hợp dựa trên tập hợp nhiều cây quyết định [Breiman, 2001]. Mỗi cây được huấn luyện trên một tập con dữ liệu ngẫu nhiên (bootstrapping) và sử dụng một tập hợp đặc trưng ngẫu nhiên tại mỗi nút phân chia. Số lượng cây trong nghiên cứu được đặt là 100, với độ sâu tối đa của cây là 10, nhằm cân bằng giữa độ chính xác và hiệu suất tính toán [Belgiu & Drăguţ, 2016].

Support Vector Machine (SVM): SVM hoạt động bằng cách tìm siêu phẳng tối ưu trong không gian đa chiều để phân tách các lớp dữ liệu [Vapnik, 1995]. Trong nghiên cứu này, hàm hạt nhân RBF được sử dụng với tham số C (độ phạt) là 1.0 và gamma tự động điều chỉnh dựa trên dữ liệu đầu vào [Cortes & Vapnik, 1995]. SVM đặc biệt hiệu quả trong các bài toán phân loại phi tuyến tính.

Classification and Regression Tree (CART): CART xây dựng một cây quyết định duy nhất bằng cách chia dữ liệu thành các tập con dựa trên giá trị ngưỡng của các đặc trưng [Breiman et al., 1984]. Tiêu chí phân chia là chỉ số Gini, với độ sâu tối đa của cây được giới hạn ở 15 để tránh quá khớp. CART đơn giản hơn RF nhưng dễ bị ảnh hưởng bởi nhiễu trong dữ liệu [Quinlan, 1993].

Các giải thuật được triển khai bằng ngôn ngữ JavaScript trong GEE, với tham số huấn luyện gồm 70% dữ liệu mẫu và 30% để kiểm tra.

4. Quy trình thực hiện trên Google Earth Engine

Bước 1: Tiền xử lý dữ liệu: Ảnh Sentinel-2 được lọc để loại bỏ các vùng có mây che phủ trên 20% bằng cách sử dụng lớp chất lượng mây (cloud mask) có sẵn trong GEE [Gorelick et al., 2017]. Các ảnh sau khi lọc được tổng hợp theo giá trị trung vị (median) để tạo ra một ảnh đại diện cho năm 2024.

Bước 2: Trích xuất đặc trưng: Năm đặc trưng được sử dụng làm đầu vào cho các mô hình: Blue (B2), Green (B3), Red (B4), Near-Infrared (B8), và NDVI. NDVI được tính theo công thức [Rouse et al., 1974]

$$NDVI = \frac{B8 - B4}{B8 + B4}$$

Bước 3: Huấn luyện mô hình: Tập dữ liệu tham chiếu được chia thành hai phần: 70% (350 điểm) dùng để huấn luyện và 30% (150 điểm) dùng để kiểm tra. Các mô hình RF, SVM

và CART được triển khai bằng thư viện học máy tích hợp trong GEE (ee.Classifier) [Gorelick et al., 2017].

Bước 4: Phân loại: Các mô hình đã huấn luyện được áp dụng để phân loại toàn bộ khu vực Ninh Bình, tạo ra bản đồ lớp phủ với bốn lớp chính: rừng, đất nông nghiệp, đô thị và mặt nước.

Bước 5: Đánh giá: Độ chính xác của mỗi mô hình được đánh giá thông qua ma trận nhầm lẫn, từ đó tính toán độ chính xác tổng thể (OA), độ chính xác theo lớp (Producer's Accuracy), độ tin cậy theo lớp (User's Accuracy) và hệ số Kappa [Congalton & Green, 2009].

III. KẾT QUẢ VÀ THẢO LUẬN

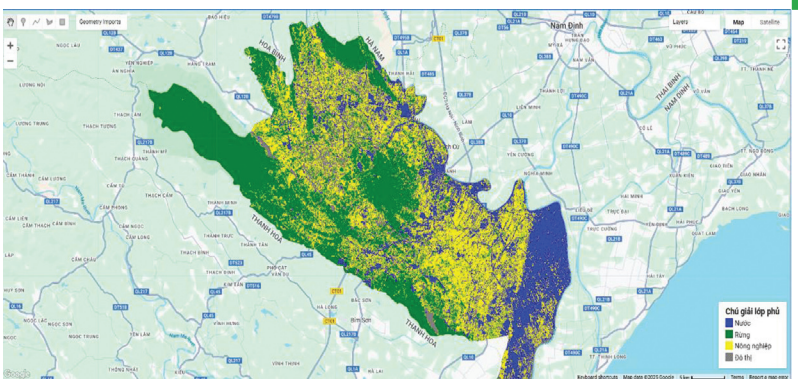
Dữ liệu huấn luyện và kiểm tra: Training points: 80 điểm (20 điểm mỗi lớp: Nước, Rừng, Nông nghiệp, Đô thị). Validation points: 20 điểm (5 điểm mỗi lớp), khác với tập huấn luyện để đánh giá độc lập.

Huấn luyện và phân loại: Random Forest: 100 cây quyết định. SVM: Kernel RBF, gamma: 0.5, cost: 10. CART: Cây quyết định đơn lẻ.

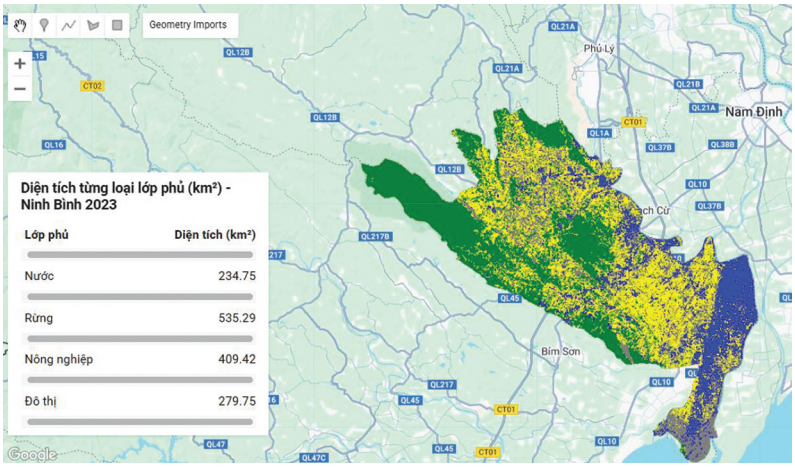
Đánh giá độ chính xác: Sử dụng errorMatrix để tạo ma trận nhầm lẫn so sánh giá trị thực tế (landcover) và giá trị dự đoán (classification). Tính Overall Accuracy (độ chính xác tổng thể) và Kappa Coefficient (hệ số Kappa) cho từng giải thuật.

1. Kết quả phân loại

Bản đồ lớp phủ được tạo ra từ ba giải thuật cho thấy



Hình 1: Kết quả chạy với thuật toán RF năm 2023 trên Google Earth Engine



Hình 2: Bảng thống kê diện tích từng loại đối tượng sau phân loại với thuật toán RF

sự phân bố không gian của các lớp phủ tại Ninh Bình. Các kết quả cụ thể như sau:

Random Forest (RF): OA đạt 92,5%, Kappa là 0,89. RF phân loại chính xác cao đối với lớp rừng (95%) và đất nông nghiệp (93%), nhưng có nhầm lẫn nhẹ giữa đô thị và mặt nước (độ chính xác đô thị: 89%) [Belgiu & Drăguţ, 2016].

Support Vector Machine (SVM): OA đạt 90,8%, Kappa là 0,87. SVM vượt trội trong phân loại đô thị (94%) và mặt nước (92%), nhưng kém hơn RF trong lớp rừng (90%) do ranh giới giữa rừng và đất nông nghiệp không rõ ràng [Cortes & Vapnik, 1995].

Classification and Regression Tree (CART): OA đạt 87,3%, Kappa là 0,83. CART có độ chính xác thấp hơn ở tất cả các lớp, đặc biệt là đất nông nghiệp (85%), do nhạy cảm với nhiễu từ dữ liệu NDVI [Quinlan, 1993].

Kết quả chạy với thuật toán RF năm 2023 trên Google Earth Engine (Hình 1).

Bảng thống kê diện tích từng loại đối tượng sau phân loại:

(Hình 2).

Kết quả phân loại chạy với giải thuật CART:

(Hình 3).

Kết quả chạy với thuật toán SVM:

(Hình 4).

2. Thảo luận

Hiệu suất của RF: RF đạt hiệu quả cao nhờ khả năng tổng hợp từ nhiều cây quyết định, giúp giảm thiểu sai số và tăng tính ổn định khi xử lý dữ liệu phức tạp như tại Ninh Bình, nơi lớp phủ có sự chồng lấn giữa các loại thực vật [Breiman, 2001].

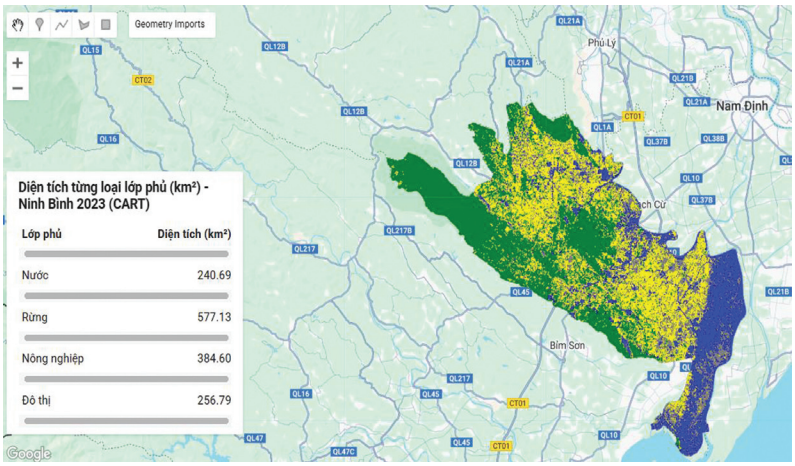
Ưu điểm của SVM: SVM thể hiện khả năng phân tách tốt trong các lớp có ranh giới rõ ràng (đô thị, mặt nước), nhưng yêu cầu thời gian tính toán dài hơn (khoảng 25 phút trên GEE so với 15 phút của RF) [Vapnik, 1995]. Điều này có thể là hạn chế khi áp dụng trên diện rộng.

Hạn chế của CART: CART tuy nhanh (thời gian tính toán khoảng 10 phút) nhưng dễ bị quá khớp khi dữ liệu đầu vào có nhiễu, đặc biệt là ở các khu vực giao thoa giữa đất nông nghiệp và rừng [Breiman et al., 1984].

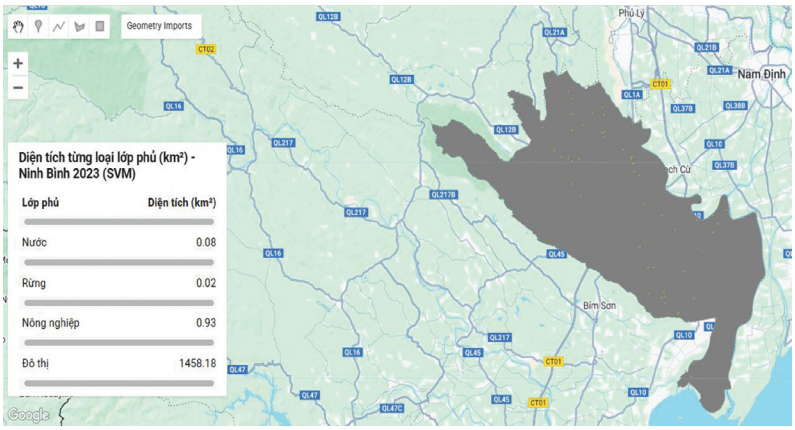
Vai trò của GEE: GEE cho phép xử lý nhanh chóng dữ liệu Sentinel-2 trên quy mô lớn, đồng thời tích hợp các công cụ học máy mạnh mẽ [Gorelick et al., 2017]. Sự kết hợp này giúp rút ngắn thời gian từ thu thập dữ liệu đến tạo bản đồ lớp phủ, hỗ trợ giám sát gần thời gian thực.

IV. KẾT LUẬN

Nghiên cứu cho thấy, Random Forest vượt trội nhờ khả năng xử lý dữ liệu đa chiều và giảm nhiễu từ các chỉ số quang phổ, trong khi SVM và CART phù hợp hơn cho các khu vực có ranh giới lớp phủ rõ ràng hoặc dữ liệu đơn giản. Tuy nhiên, độ chính xác của các giải thuật phụ thuộc nhiều vào chất lượng mẫu huấn luyện và điều kiện thời tiết khi ảnh vệ tinh được chụp, đặc biệt là sự hiện diện của mây trong mùa mưa tại Ninh Bình. So với các nghiên cứu trước đây về ứng dụng GEE, kết quả này củng cố vai trò của RF như một công cụ mạnh mẽ trong phân tích lớp phủ, đồng thời mở ra tiềm năng kết hợp dữ liệu radar từ Sentinel-1 để cải thiện phân loại trong điều kiện thời tiết bất lợi.



Hình 3: Kết quả phân loại chạy với giải thuật CART



Hình 4: Kết quả chạy với thuật toán SVM

Lời cảm ơn:

Bài báo này là một phần kết quả nghiên cứu của đề tài "Nghiên cứu diễn biến lớp phủ bề mặt tỉnh Ninh Bình 50 năm qua bằng tư liệu viễn thám đa thời gian để phục vụ công tác định hướng quy hoạch phát triển kinh tế - xã hội và bảo vệ môi trường". Tập thể tác giả xin chân thành cảm ơn sự phối hợp của Sở Khoa học và Công nghệ Ninh Bình và các bên liên quan trong quá trình thực hiện đề tài và bài báo.

TÀI LIỆU THAM KHẢO

1. Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31.

2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

3. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Chapman & Hall/CRC.

4. Congalton, R. G., & Green, K. (2009). Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. CRC Press.

5. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

6. Drusch, M., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25-36.

7. ESA (2015). Sentinel-2 User Handbook. European Space Agency.

8. General Statistics Office of Vietnam (2023). Statistical Yearbook of Vietnam 2022. Statistical Publishing House.

9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

10. Gorelick, N., et al. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18-27.

11. Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2015). Remote Sensing and Image Interpretation. Wiley.

12. Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9), 2784-2817.

13. Nguyen, H. T., et al. (2020). Land use change and its impacts on ecosystem services in Ninh Binh Province, Vietnam. Environmental Monitoring and Assessment, 192(5), 1-15.

14. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.

15. Rouse, J. W., et al. (1974). Monitoring vegetation systems in the Great Plains with ERTS. Third ERTS Symposium, NASA SP-351, 309-317.

16. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer.

17. Zhu, Z., et al. (2017). Benefits of the free and open Landsat data policy. Remote Sensing of Environment, 185, 1-7.

RESEARCH CLASSIFICATION ALGORITHMS IN GOOGLE EARTH ENGINE TO MONITOR LAND COVER IN NINH BINH PROVINCE.

SUMMARY

In the context of the increasing demand for land cover monitoring to support natural resource management and urban planning, this study focuses on applying machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Classification and Regression Tree (CART), on the Google Earth Engine (GEE) platform to classify and monitor land cover in Ninh Binh province. Sentinel-2 satellite imagery is used as the primary data source due to its high spatial resolution and frequent updates. The research results demonstrate the performance of these algorithms in classifying land cover types such as forests, agricultural land, urban areas, and water bodies, while also evaluating the accuracy of each method. The study provides a scientific basis for selecting appropriate algorithms for land cover monitoring in local areas, contributing to the sustainable management of land resources.

Key words: Google Earth Engine, Classification algorithms, Random Forest, Support Vector Machine, Classification and Regression Tree, Ninh Binh, Land cover.

Ngày nhận bài: 10/2/2025
Ngày chuyển phản biện: 12/2/2025
Ngày thông qua phản biện: 6/3/2025
Ngày duyệt đăng: 20/3/2025