TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Check for updates

Bài báo khoa học

Đánh giá độ chính xác kết quả phân loại sử dụng đất trên ảnh viễn thámSentinel 2 bằngcác thuật toán học máy

Nguyễn Thị Phương Hoa[1]*, Nguyễn Thị Thanh Bình[1], Lê Thành An[2], Nguyễn Duy Phương [2]

[1] Cục Viễn thám quốc gia, Bộ Tài nguyên và Môi trường; ntphoa19@gmail.com; binhntt1191@gmail.com
[2] Học Viện quốc tế, Bộ Công an; minhtn122019@gmail.com; nanh22022022@gmail.com

*Tác giả liên hệ: ntphoa19@gmail.com; Tel.: +84–386635353

Ban Biên tập nhận bài: 8/10/2024; Ngày phản biện xong: 27/11/2024; Ngày đăng bài: 25/3/2025

Tóm tắt: Nghiên cứu sử dụng 4 thuật toán phân loại học máy tiêu biểu như cây quyết định (CART), thuật toán rừng ngẫu nhiên (RF), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGBoost) để phân loại sử dụng đất tỉnh Nam Định trên ảnh viễn thám quang học Sentinel 2. Bằng cách sử dụng cùng một bộ mẫu huấn luyện và tập dữ liệu đánh giá độ chính xác, nghiên cứu này đánh giá độ chính xác của các mô hình phân loại từ ảnh viễn thám quang học Sentinel 2 cho 7 loại sử dụng đất tỉnh Nam Định năm 2024. Kết quả đánh giá độ chính xác sau phân loại đã chỉ ra rằng với diện tích khoảng 1.700 km^2 trên quy mô cấp tỉnh thuật toán XGBoost cho độ chính xác cao nhất với hệ số Kappa là 0,7 và độ chính xác toàn cục đạt 78,8%, mô hình CART cho độ chính xác thấp nhất với Kappa là 0,58. Các đối tượng sử dụng đất có diện tích nhỏ như đất làm muối không nhận diện được bằng mô hình CART và SVM. Tuy nhiên, đối tượng mặt nước hai mô hình này cho độ chính xác cao hơn khi sử dụng thuật toán RF và XGBoost.

Từ khóa: Phân loại ảnh; Phân loại sử dụng đất; Viễn thám; Cây quyết định CART; Random Forest (RF); Support Vector Machines (SVM); Extreme Gradient (XCBoost).

1. Đặt vấn đề

Hiện nay, ứng dụng viễn thám đang bước vào một kỉ nguyên mới với sự bùng nổ của các nguồn dữ liệu lớn được thu thập từ không gian (*space-borne, air-borne*), các thiết bị bay không người lái (UAV) hay từ đám mây điểm (*point cloud*), nguồn dữ liệu này có dung lượng lớn, kiểu dữ liệu đa dạng và tốc độ thu thập dữ liệu. Việc khai thác các dữ liệu này nhằm giải quyết các bài toán lớn trong việc giám sát quá trình biến đổi của tự nhiên và nhân tạo ở qui mô toàn cầu. Sự ra đời của trí tuệ nhân tạo (*Artificial Intelligence - AI*) diễn ra vào những năm 1950 dựa trên sự khẳng định rằng mọi khía cạnh của việc học hoặc bất kỳ đặc điểm nào khác của trí tuệ con người đều có thể được mô tả chính xác đến mức có thể tạo ra một cỗ máy để mô phỏng nó [1]. Trí tuệ nhân tạo là lĩnh vực nghiên cứu nhằm phát triển các chương trình máy tính để giải quyết các vấn đề theo cách bắt chước trí tuệ của con người, trong đó phương pháp tiếp cận học máy (*Machine Learning - ML*) là một phần của trí tuệ nhân tạo với các thuật toán máy tính được sử dụng để tự học từ dữ liệu [2]. Các thuật toán này đã và đang được ứng dụng rất hiệu quả trong việc phân loại lớp phủ bề mặt đất/sử dụng đất từ ảnh viễn thám như thuật toán vector hỗ trợ (SVM), mạng thần kinh nhân tạo (NNET), cây quyết định (CART), rừng ngẫu nhiên (RF) và cũng đánh giá được độ tin cậy của từng thuật toán trên [3–5]. Kết quả phân loại lớp phủ mặt đất từ ảnh viễn thám không chỉ phụ thuộc vào các thuật toán mà còn phụ thuộc vào các yếu tố như vị trí địa lý, điều kiện tự nhiên, thời điểm chụp ảnh, chất lượng ảnh, độ phân giải ảnh, mẫu khóa ảnh...[3].

Tuy nhiên, việc đánh giá các thuật toán học máy sử dụng trong phân loại sử dụng đất trên ảnh viễn thám nhằm giám sát sự thay đổi hiện trạng sử dụng đất so với quy hoạch sử dụng đất cho đến nay hầu như chưa thực hiện. Do đó, nghiên cứu này được thực hiện để xác định thuật toán phân loại ảnh viễn thám bằng học máy nào cho kết quả có độ chính xác nhất phù hợp với mục tiêu phân loại 7 loại sử dụng đất khu vực tỉnh Nam Định năm 2024. Thuật toán cho kết quả tốt nhất sẽ được lựa chọn để xây dựng bản đồ hiện trạng sử dụng đất tỷ lệ 1: 50.000 cũng như giám sát việc chuyển đổi sử dụng đất có đúng với quy hoạch sử dụng đất tỉnh Nam Định đến năm 2030.

2. Phương pháp nghiên cứu và dữ liệu thu thập

2.1. Giới thiệu khu vực nghiên cứu

Địa bàn nghiên cứu là tỉnh Nam Định với đường bờ biển dài 72 km. Nhiệt độ trung bình năm của khu vực này là 23,7^{o}C, lượng mưa trung bình năm đạt khoảng 1.699 mm [6]. Do sự gia tăng dân số, phát triển kinh tế - xã hội, sự khai thác tài nguyên quá mức và những tác động tiêu cực của con người đang gây ra những biến đổi xấu đến môi trường, nguồn nước, không khí, môi trường sinh thái của tỉnh. Sự xâm nhập mặn tiến sâu trong nội đồng, diện tích rừng phòng hộ suy giảm, mất đất khu vực bãi bồi vườn quốc gia Xuân Thủy… khiến thu hẹp diện tích đất canh tác, đặc biệt đất trồng lúa [7]. Theo số liệu thống kê [8] cho thấy năm 2015 diện tích đất lúa của tỉnh là 154.434 ha nhưng đến năm 2020 diện tích đất này chỉ còn 145.376 ha. Như vậy sau 5 năm diện tích đất nông nghiệp giảm trên 9000 ha. Thực trạng này đặt ra một áp lực lớn cho việc phát triển kinh tế - xã hội của tỉnh và những khó khăn trong việc giám sát biến động sử dụng đất phục vụ quy hoạch sử dụng đất của tỉnh.

Hình 1. Khu vực thử nghiệm tỉnh Nam Định (màu trắng: tập dữ liệu đánh giá độ chính xác, màu đen: tập dữ liệu huấn luyện mô hình).

2.2. Dữ liệu thu thập

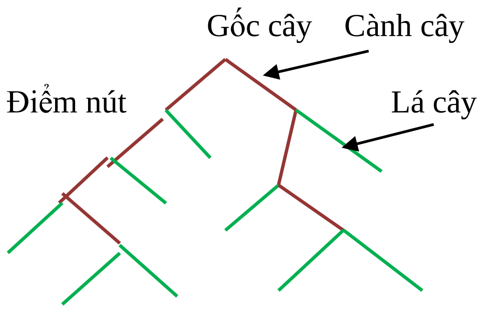
Vật liệu chính bao gồm ảnh Sentinel 2 chụp khu vực tỉnh Nam Định năm 2024. Nguồn ảnh này được được cung cấp miễn phí từ cổng thông tin của hệ thống Cơ quan môi trường Châu Âu [9] (https://browser.dataspace.copernicus.eu).

Nghiên cứu này sử dụng nền tảng Google Earth Engine (GEE) để thử nghiệm và phân loại tự động 7 loại hình sử dụng đất bao gồm: Lúa, Cây hàng năm, Cây lâu năm, Mặt nước và Nuôi trồng thủy sản, Đất làm muối, Đất rừng và Đất phi nông nghiệp.

2.3. Cơ sở khoa học của các thuật toán học máy ứng dụng phân loại ảnh viễn thám

2.3.1. Thuật toán (*Classification and Regression Tree - CART*)

Thuật toán CART là một thuật toán học máy có giám sát trong hệ thống phân loại dựa trên cây quyết định (Decision tree) và sử dụng các mẫu huấn luyện để xác định, nhận dạng, phân loại đối tượng trên ảnh viễn thám [10]. Cây quyết định là một phương pháp có thể học quy nạp bằng cách đào tạo các mẫu và xây dựng cây quyết định, sau đó sử dụng cây quyết định để phân loại dữ liệu. Cây quyết định được cấu tạo bởi một nút gốc, một loạt các nút bên trong và các nút lá. Mỗi nút chỉ có thể có một nút chính và hai hoặc nhiều nút phụ. Các nút được kết nối với nhau bằng các nhánh [11]. Mỗi nút bên trong tương ứng với một thuộc tính hoặc nhóm thuộc tính thử nghiệm và mọi bên tương ứng với mọi giá trị có thể có của thuộc tính. Nút tương ứng với một giá trị thuộc tính của lớp và nút khác nhau có thể tương ứng với cùng một giá trị thuộc tính của lớp [12]. Mỗi đường từ gốc đến lá tương ứng với một quy tắc và điều kiện của quy tắc là tùy chọn tất cả các giá trị thuộc tính của các nút, kết quả của quy tắc là thuộc tính lớp của nút lá trên đường [11].



Hình 2. Mô hình cây quyết định.

Quá trình xây dựng cây quyết định là: xuất phát từ gốc cây với đầu vào như trên, cần tạo ra một cây có đầy đủ cành và các lá, tập hợp T được phân chia dần về các nút cây, ngoài ra, cành cuối cùng của cây chỉ còn các lá (đại diện cho các lớp phân loại).

Thuật toán CART xây dựng cây quyết định như sau: Giả sử cho tập hợp T gồm N đối tượng của m lớp $\omega_1, \omega_2, \ldots, \omega_m$. Mỗi đối tượng chỉ thuộc một lớp và có n dấu hiệu. Mỗi lớp ω_i có N_i đối tượng ($i = 1 \div m$).

Bước 1: Tính entropy S_0 của tập hợp gốc T theo công thức:

$$S = -\sum(\rho_i \times \log_2 \rho_i) \tag{1}$$

Tỷ số $\rho_i = N_i / N$ là xác suất của đối tượng thuộc lớp ω_i ($i = 1 \div m$).

Trường hợp $S_0 = 0$, chỉ xảy ra khi tất cả các đối tượng của tập hợp gốc đều thuộc về một lớp phân loại. Lưu giữ lại lớp này như là lá của cây.

Trường hợp $S_0 \neq 0$, lấy $\Delta S = 0$, cho $j = 0$.

Bước 2: Tính $j = j + 1$, nếu $j > N$, chuyển đến bước 4.

Xác định đối tượng $X_j \in T$, véc tơ dấu hiệu là $(x_{1j}, x_{2j}, \ldots, x_{nj})$. Lấy $k = 0$.

Bước 3: Tính $k = k + 1$, nếu $k > n$, quay lại bước 2.

Xác định dấu hiệu $x_{kj} \in X_j$, phân chia tập hợp gốc thành hai tập hợp con theo dấu hiệu này.

Tính entropy cho từng tập hợp con, xác định giá trị trung bình và tính gia số ΔS_{kj} theo công thức $\Delta S = S_0 - S_{TB}$

Nếu $\Delta S_{kj} < \Delta S$, quay lại bước 3.

Nếu $\Delta S_{kj} \geq \Delta S$, lấy $\Delta S = \Delta S_{kj}$, $p = k$ và $q = j$.

Quay lại bước 3.

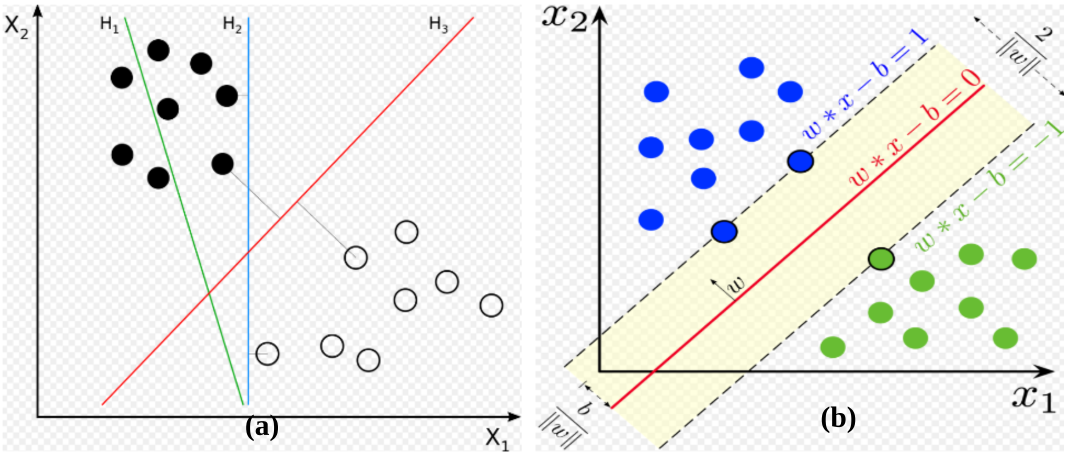
Bước 4: Phân chia tập hợp gốc thành hai tập hợp con theo dấu hiệu x_{pq}.

Nếu tồn tại tập hợp con có số đối tượng không thuộc một lớp, quá trình lặp lại từ bước 1 cho tất cả các nút cây, trừ các nút đã sinh ra toàn lá. Tập hợp gốc tại các nút là tập hợp được phân cành trước đó.

Nếu mỗi tập hợp con chỉ gồm một hoặc nhiều hơn các đối tượng ở cùng một lớp, ta xác định được chúng thuộc lớp nào (các đối tượng của tập hợp mẫu đã được phân loại), lưu giữ chúng như là lá cây. Quá trình xây dựng cây quyết định kết thúc khi tất cả các cành cây cuối cùng đều ra toàn lá.

2.3.2. Thuật toán (*Support Vector Machine - SVM*)

Đây là một thuật toán phân loại nhị phân. Một bộ các mẫu huấn luyện thuộc hai thể loại cho trước, thuật toán huấn luyện SVM xây dựng một mô hình SVM để phân loại các dữ liệu khác vào hai thể loại đó. Mô hình SVM là cách biểu diễn các điểm giữa hai thể loại sao cho khoảng cách từ các mẫu huấn luyện tới ranh giới là xa nhất có thể. Các dữ liệu mới cũng được biểu diễn trong cùng một không gian và được thuật toán dự đoán thuộc một trong hai thể loại tùy vào dữ liệu đó nằm ở phía nào của ranh giới [4]. Bản chất của phương pháp SVM là chuyển không gian dữ liệu ban đầu thành một không gian mới hữu hạn chiều mà ở đó cho khả năng phân lớp dễ dàng hơn. Điểm làm SVM hiệu quả hơn các phương pháp khác chính là việc SVM không bị giới hạn việc phân lớp một cách tuyến tính.



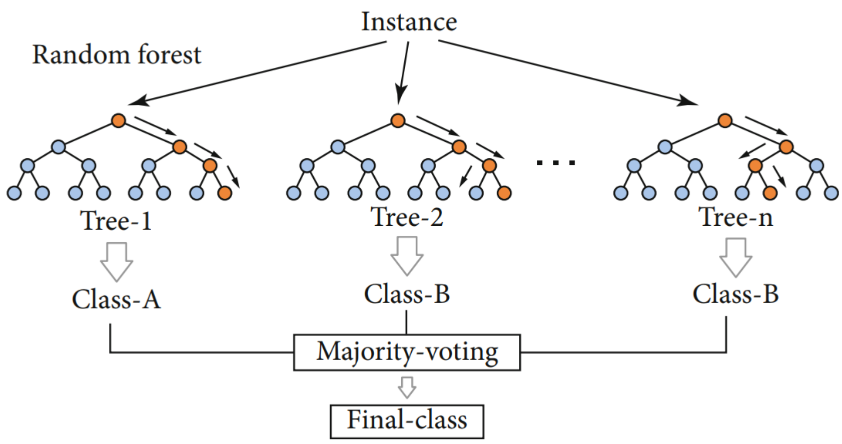
Hình 3. (a) H_1 không phân tách các lớp, H_2 phân tách các lớp nhưng chỉ với mép lề nhỏ, H_3 phân tách các lớp với mép lề tối đa, (b) xác định khoảng cách lề giữa hai lớp dữ liệu.

Nếu xem xét dữ liệu đầu vào như là hai tập các véc-tơ trong không gian đa chiều (*n-dimentional space*) thì thuật toán SVM sẽ xây dựng một ranh giới tuyến tính giữa hai lớp, nghĩa là một mặt siêu phẳng phân tách (*separating hyperplane*), cho phép tối đa hóa khoảng cách lề giữa hai lớp dữ liệu (Hình 3). Để tính toán khoảng cách lề, hai mặt siêu phẳng song song được dựng về hai phía của mặt siêu phẳng phân tách và được "đẩy hết cỡ" về phía hai tập dữ liệu. Sự phân tách được cho là tốt nếu như khoảng cách từ mặt siêu phẳng đến các điểm dữ liệu của cả hai lớp là xa nhất, khi đó, khoảng cách lề càng lớn thì sai số tổng quát hóa của thuật toán phân loại càng tốt.

2.3.3. Thuật toán (*Random Forest - RF*)

Thuật toán Random Forest được ứng dụng cho phân loại, tính hồi quy bằng cách xây dựng nhiều cây quyết định (*Decision tree*). Một cây quyết định là một cách đơn giản để biểu diễn một giao thức (*Protocol*) hay cây quyết định biểu diễn một kế hoạch, trả lời câu hỏi phải làm gì trong một hoàn cảnh nhất định. Mỗi Node của cây sẽ là các thuộc tính, và các nhánh là giá trị lựa chọn của thuộc tính đó. Bằng cách đi theo các giá trị thuộc tính trên cây sẽ cho biết giá trị dự đoán [13] (Hình 4).

Để thiết lập mô hình RF, cần có hai tham số mà được coi là cơ sở của phương pháp đó là số lượng cây phân loại "n-tree" và số lượng các lớp đối tượng "m-try" trong một phân nhánh. "Lá phiếu" đa số của tất cả các cây được sử dụng để gán một lớp chung cuộc cho mỗi lớp chưa biết. Điều này trực tiếp giải quyết được vấn đề rằng mỗi một cây bất kỳ nào đó có thể không phải là tối ưu, nhưng bằng cách kết hợp nhiều cây thì sẽ đạt được một giải pháp tối ưu tổng thể. Ý tưởng này được mở rộng hơn nữa bằng cách huấn luyện mỗi cây với tập hợp con



Hình 4. Phân loại RF với 2 lớp đối tượng A và B.

dữ liệu huấn luyện được tạo ra ngẫu nhiên của riêng nó và cũng chỉ sử dụng một tập hợp con các biến cho cây đó. Sự kết hợp giữa việc giảm bớt các mẫu dữ liệu huấn luyện giảm và giảm bớt số lượng biến có nghĩa là các cây riêng lẻ sẽ kém chính xác hơn và cũng sẽ ít tương quan hơn, nhưng chúng làm cho tổng thể trở nên đáng tin cậy hơn. Dữ liệu không được sử dụng làm dữ liệu huấn luyện có thể được dùng để đánh giá độc lập độ chính xác tổng thể của thuật toán phân loại RF.

2.3.4. Thuật toán (*Extreme Gradient Boosting - XGBoost*)

Thuật toán XGBoost nhằm mục đích xử lý bài toán học máy có giám sát cho độ tin cậy cao. Với phương pháp học máy chuyên sâu thông thường chỉ nhận nguồn vào là dạng thô, khi đó phải quy đổi sang n-vector trong khoảng trống số thực thì XGBoost nhận nguồn dữ liệu đầu vào là dạng bảng với mọi kích cỡ dữ liệu và dạng tài liệu gồm có cả phân loại. XGBoost có tốc độ huấn luyện nhanh do có thể tính toán song song khi sử dụng tất cả các lõi CPU trong quá trình đào tạo. XGBoost là sự mở rộng của thuật toán GTB (Gradient Tree Boosting được đề xuất bởi Friedman [14]), tuy nhiên kèm theo đó là những cải tiến để tối ưu thuật toán, bộ nhớ đệm của cấu trúc dữ liệu, sự kết hợp tối ưu giữa phần mềm và phần cứng nên có khả năng ứng dụng với bộ dữ liệu lớn [15].

Về mặt toán học, XGBoost là một phương pháp học tập tổng hợp kết hợp các dự đoán của nhiều mô hình yếu để tạo ra một dự đoán mạnh. Các mô hình yếu trong XGBoost là các cây quyết định, được đào tạo bằng cách sử dụng gradient enhancement. Điều này có nghĩa là tại mỗi lần lặp, thuật toán sẽ điều chỉnh một cây quyết định cho phù hợp với phần dư của lần lặp trước đó.

Cây quyết định trong XGBoost được đào tạo bằng cách sử dụng hàm mục tiêu sau:

$$\min_{\theta}\left(\sum_{i}^{n} l\left(y_i, y_i\right)\right) + \sum_{i}^{k} \Omega\left(f_k\right) \tag{2}$$

Trong đó l là hàm mất mát, y_i là giá trị thực của ví dụ đào tạo thứ i, y_i là nhãn dự đoán của ví dụ đào tạo thứ i, f_k là cây quyết định thứ k và Ω là một thuật ngữ chính quy hóa phạt độ phức tạp của cây. Hàm mục tiêu này được tối ưu hóa bằng cách sử dụng phương pháp giảm dần độ dốc.

Sau khi cây quyết định đã được đào tạo, XGBoost đưa ra dự đoán bằng cách kết hợp các dự đoán của tất cả các cây bằng cách sử dụng một giá trị trung bình có trọng số. Trọng số cho mỗi cây được học trong quá trình đào tạo bằng cách sử dụng cùng một hàm mục tiêu. Điều này cho phép thuật toán tự động học cây nào quan trọng hơn và nên được đưa ra nhiều trọng số hơn trong dự đoán cuối cùng.

XGBoost là một gói tăng cường độ dốc đa nền tảng triển khai một khuôn khổ tăng cường độ dốc. Thuật toán có thể mở rộng cho tính toán song song. Ngoài Python, nó có sẵn trong C++, Java, R, Julia và các ngôn ngữ tính toán khác. XGBoost đã thu hút sự chú ý trong các cuộc thi học máy như một thuật toán được lựa chọn để phân loại và hồi quy.

2.3.5. Các phương pháp

Ảnh Sentinel 2 được phân loại theo các thuật toán cây quyết định (CART), thuật toán rừng ngẫu nhiên(RF), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGBoost). Đối tượng phân loại được xác định dựa theo bộ khóa giải đoán của 7 loại đất nông nghiệp chính (tổng hợp từ Phân loại nhóm đất nông nghiệp theo Luật đất đai (2024) và đất phi nông nghiệp [16].

- Phương pháp đánh giá độ chính xác: được đánh giá thông qua chỉ số Kappa [17].

- Phương pháp thu thập mẫu giải đoán: Các mẫu thực địa (groundtruth) được thu thập trên địa bàn nghiên cứu thông qua phương pháp khảo sát thực tế. Điểm thu mẫu được chọn theo phương pháp ngẫu nhiên phân nhóm [18] trên nền ảnh viễn thám có độ phân giải cao CNES image (MapsGoogle).

2.3.6. Sơ đồ quy trình công nghệ thực hiện nghiên cứu

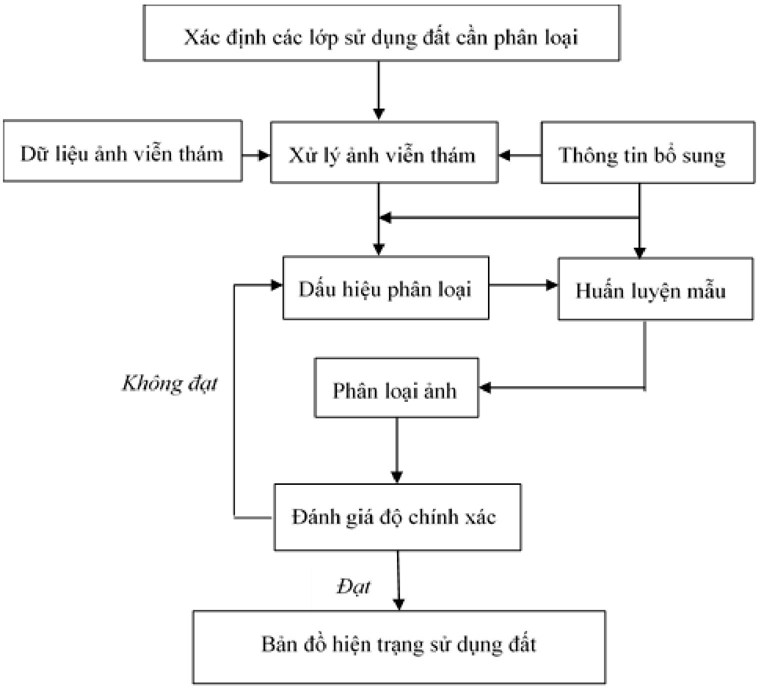
Sơ đồ quy trình nghiên cứu được thể hiện trên hình 5.

a) Tiền xử lý ảnh viễn thám

Hiệu chỉnh ảnh hưởng của đầu thu ảnh vệ tinh, khí quyển, góc cao mặt trời. Phản xạ viễn thám xác định từ tư liệu ảnh viễn thám tuân theo quy luật sự tồn tại khách quan khi có sự tương tác của ánh sáng đối với các đối tượng trong tự nhiên.

b) Dấu hiệu phân loại

Dấu hiệu phân loại là các dấu hiệu đặc trưng cho các lớp thông tin và lớp phổ thể hiện trên ảnh viễn thám đa phổ. Bộ dấu hiệu cấu trúc ảnh được đặc trưng bằng sự lặp lại cấu trúc ảnh của các vùng đồng nhất có kích thước tương đối giống nhau. Ví dụ như, cấu trúc ảnh các khu nhà trong thành phố, các khu nuôi trồng thủy sản ven biển,… Bộ dấu hiệu này được xác định bằng đặc trưng vô hướng của ma trận cấu trúc, thể hiện sự tương đồng của các cặp mức độ sáng ở các pixel lân cận.



Hình 5. Sơ đồ quy trình nghiên cứu.

c) Huấn luyện mẫu

Huấn luyện mẫu nhằm mục đích tạo ra mẫu huấn luyện là các bộ pixel được sử dụng trong bộ phân loại để nhận biết các lớp và xác định ranh giới phân khúc trong không gian dấu hiệu tương ứng với các tính chất của các pixel được huấn luyện.

d) Phân loại ảnh

Ảnh được phân thành 7 phân lớp (trồng lúa, trồng cây lâu năm, cây hàng năm khác, nuôi trồng thủy sản, làm muối, lâm nghiệp và đất khác) dựa trên dấu hiệu phân loại đề cập ở trên. Thuật toán phân lớp thực hiện theo "phân loại định hướng đối tượng". Phương pháp này thuộc kiểu phân loại có giám sát, dựa trên nguyên lý về sự khác biệt phổ phản xạ giữa các đối tượng lớp phủ [19–21].

e) Điều tra bổ sung ngoại nghiệp

Điều tra bổ sung ngoại nghiệp nhằm kiểm tra kết quả phân loại ảnh và tách chiết thông tin ở nội nghiệp; Xác minh các đối tượng còn nghi vấn ở công tác nội nghiệp; Bổ sung các đối tượng không thể phân loại được trong nội nghiệp; Thu thập các tài liệu về địa danh, địa giới, định tính và định lượng của đối tượng và các thông tin khác không có trên ảnh hoặc không thể xác định được.

2.3.7. Phương pháp đánh giá độ chính xác

Để đánh giá độ chính xác của các 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*).

Đánh giá độ chính xác được đánh giá thông qua chỉ số Kappa [17] như sau:

$$K = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)}{N^2 - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)} \quad (3)$$

Trong đó N là tổng số pixel lấy mẫu; r là số lớp đối tượng phân loại; x_{ii} là số pixel đúng trong lớp thứ i; x_{i+} là tổng pixel lớp thứ i của mẫu (groundtruth); x_{+i} là tổng pixel lớp thứ i sau phân loại (kết quả giải đoán ảnh). Ý nghĩa của hệ số K được diễn giải theo mức độ giống nhau giữa số liệu cần kiểm chứng và số liệu tham chiếu như sau [22]: < 0: không có sự giống nhau; 0,0 - 0,20: Giống nhau rất ít; 0,2 - 0,4: Giống nhau trung bình; 0,61 - 0,80: Giống nhau nhiều; 0,81 - 1,00: Giống nhau gần hoàn toàn.

Mức độ giống nhau ở trên thể hiện cấp độ chính xác của giải đoán ảnh. Nếu bản đồ được giải đoán ra giống với số liệu thực địa hoặc bản đồ tham chiếu có độ tin cậy cao thì được hiểu là sản phẩm giải đoán chính xác.

3. Kết quả và thảo luận

3.1. Kết quả thu thập và hiệu chỉnh ảnh

Ảnh viễn thám sử dụng là ảnh Sentinel 2 được cung cấp miễn phí từ cổng thông tin của hệ thống Cơ quan môi trường Châu Âu [9] (Bảng 1).

Bảng 1. Thông tin tư liệu viễn thám thu thập.

STT	Mã ảnh (ID)	Ngày chụp	Đặc điểm chung
1	20240926T031519_20240926T032821_T48QXG	26/09/2024	Độ phân giải: 10m, 20m Kiểu dữ liệu: Float Khuôn dạng dữ liệu: BSQ Lưới chiếu: UTM, Zone 48 North Hệ tọa độ: WGS-84
2	20241006T031519_20241006T032825_T48QXG	06/10/2024	
3	20241006T031519_20241006T032825_T48QXH	06/10/2024	
4	20241009T032539_20241009T033232_T48QXG	09/10/2024	
5	20241011T031641_20241011T033004_T48QXH	11/10/2024	

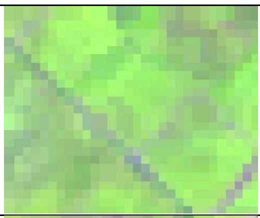
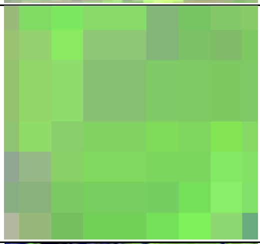
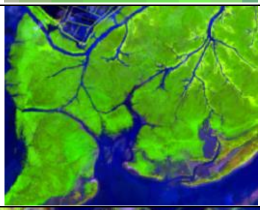
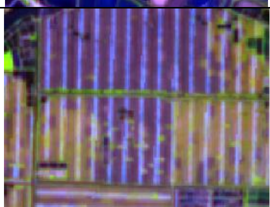
Hình 6. Ảnh viễn thám khu vực tỉnh Nam Định năm 2024 (Trước và sau khi xử lý lọc mây).

Nghiên cứu này sử dụng các thông số được cung cấp cùng với tập ảnh về điểm cực “Zenith” và góc phương vị “Azimuth” kết hợp với việc sử dụng kỹ thuật vùng đệm để tìm kiếm khu vực bóng mây. Sau khi xác định được vị trí của mây và bóng mây, tiến hành tạo mặt nạ cho các vị trí này, ở đó các điểm ảnh có giá trị bằng 0. Giá trị điểm ảnh tại những vị trí này sẽ được gán bởi giá trị điểm ảnh từ các ảnh khác không mây bằng thuật toán trung vị (Hình 6).

3.2. Xây dựng khóa giải đoán

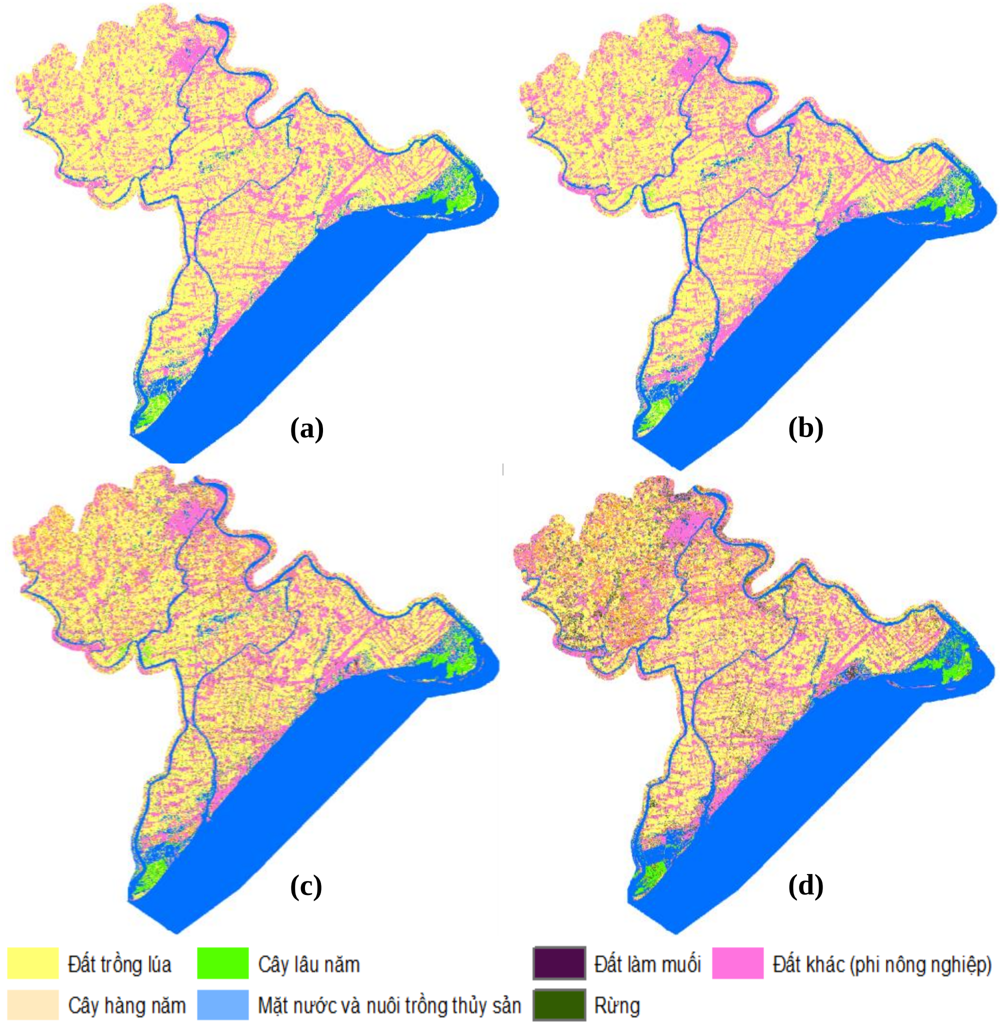
Khóa giải đoán ảnh của 7 loại đất nông nghiệp chính bao gồm: đất trồng lúa, đất trồng cây hằng năm, đất trồng cây lâu năm, đất lâm nghiệp, đất nuôi trồng thủy sản và đất làm muối và các loại đất phi nông nghiệp còn lại. Đặc điểm của các khóa giải đoán được xây dựng như trong bảng 2.

Bảng 2. Hệ thống khóa giải đoán phân loại ảnh viễn thám.

TT	Loại sử dụng	Ảnh viễn thám	Chìa khóa giải đoán	Đặc điểm
1	Đất trồng lúa			Màu đồng nhất, cấu trúc mịn, có thể thấy các dải thửa
2	Cây hàng năm			Màu không đồng nhất, cấu trúc dải thửa. Thường ở gần khu vực dân cư.
3	Cây lâu năm			Màu không đồng nhất, cấu trúc không mịn. Thường ở các khu vực ven đường, ven sông hoặc vườn trong khu dân cư.
4	Rừng			Màu đồng nhất, cấu trúc mịn, tập hợp thành vùng lớn. Thấy nhiều ở vùng ven biển.
5	Đất nuôi trồng thủy sản			Màu sẫm, cấu trúc không mịn, có thể thấy rõ các ô thửa
6	Đất làm muối			Màu không đồng nhất, cấu trúc không min, thấy rõ các dải đều nhau.
7	Đất phi nông nghiệp			Trên ảnh vệ tinh màu hồng, trắng đan xen, cấu trúc không mịn

3.3. Kết quả phân loại ảnh

Hiện trạng lớp phủ sử dụng đất tỉnh Nam Định tại thời điểm tháng 9, tháng 10 năm 2024 được phân loại trên ảnh Sentinel 2 bằng thuật toán CART, SMV, RF, XGBoost được trình bày ở hình 7.



Hình 7. Kết quả phân loại sử dụng đất tỉnh Nam Định theo các thuật toán CART(a), SMV (b), RF (c), XGBoost (d).

3.4. Kết quả diện tích hiện trạng lớp phủ sử dụng đất

Kết quả thống kê từ bản đồ hiện trạng sử dụng phân loại trên nền ảnh viễn thám được thể hiện ở bảng 3 và hình 7.

Bảng 3. Diện tích đất nông nghiệp phân tích từ ảnh viễn thám qua các giai đoạn khác nhau (ha).

TT	Loại đất	CART		SMV		RF		XGBoost		Số liệu bản đồ hiện trạng 2020	
		Diện tích	Tỷ lệ	Diện tích	Tỷ lệ	Diện tích	Tỷ lệ	Diện tích	Tỷ lệ	Diện tích	Tỷ lệ
1	Lúa	106897,4	61,3	96835,7	55,5	84898,6	48,7	74759,1	42,9	86251,4	49,4
2	Cây hàng năm	335,2	0,2	1322,7	0,8	13166,8	7,5	16873,7	9,7	6193,3	3,5
3	Cây lâu năm	0,0	0,0	121,2	0,1	5370,1	3,1	8130,1	4,7	184,4	0,1
4	Mặt nước và NTS	21321,3	12,2	23091,4	13,2	23634,5	13,5	25411,9	14,6	29304,7	16,8
5	Đất làm muối	0,0	0,0	0,0	0,0	297,7	0,2	1003,7	0,6	800,3	0,5

TT	Loại đất	CART		SMV		RF		XGBoost		Số liệu bản đồ hiện trạng 2020	
		Diện tích	Tỷ lệ	Diện tích	Tỷ lệ	Diện tích	Tỷ lệ	Diện tích	Tỷ lệ	Diện tích	Tỷ lệ
6	Đất rừng	2859,6	1,6	2870,4	1,6	3764,5	2,2	3316,7	1,9	3097,3	1,8
7	Đất phi nông nghiệp	43049,9	24,7	50221,9	28,8	43331,2	24,8	44968,0	25,8	48631,8	27,9
	Tổng	174463.3	100	174463,3	100,0	174463,3	100	174463,3	100	174463,3	100

Theo bảng trên, đất trồng lúa chiếm tỷ trọng cao nhất trong sử dụng đất tại khu vực ven biển tỉnh Nam Định (trên 40% ở tất cả các thời kì).

3.5. Đánh giá độ chính xác kết quả phân loại ảnh viễn thám bằng các thuật toán học máy

Bảng 4 thể hiện chi tiết độ chính xác của các kết quả phân loại theo các thuật toán lựa chọn. Độ chính xác tổng thể (OA) của SVM là 71,6%, hệ số Kappa là 0,58 trong khi phương pháp XGGboost đạt 78,9% với Kappa là 0,70 và RF có OA là 75,0%, Kappa đạt 0,65 và CART có OA thấp nhất là 71,3%, Kappa đạt 0,58.

Bảng 4. So sánh độ chính xác các kết quả phân loại.

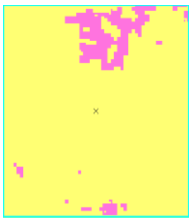
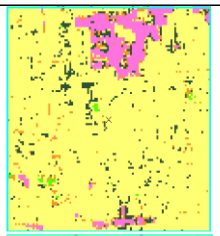
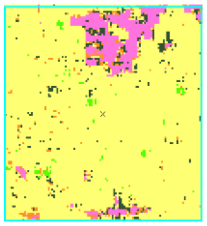
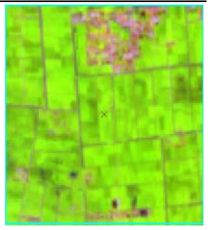
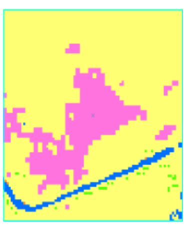
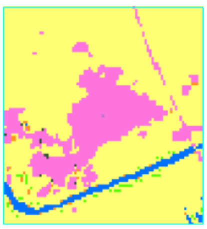
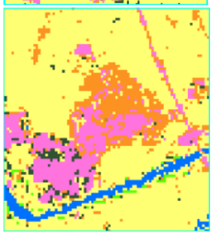
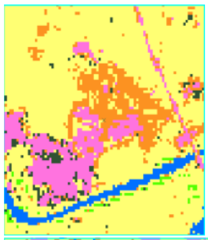
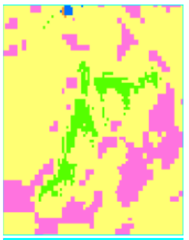
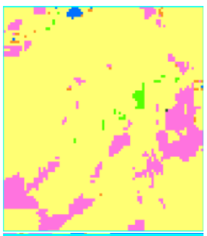
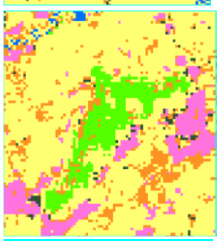
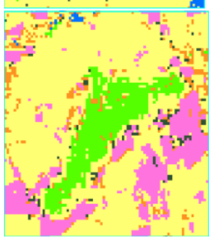
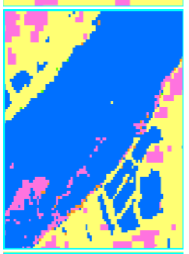
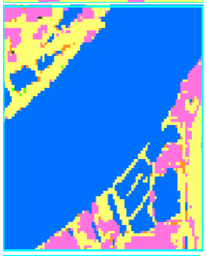
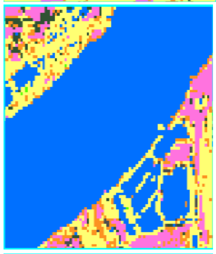
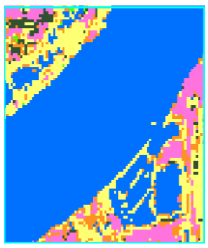
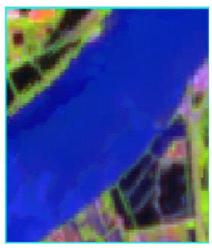
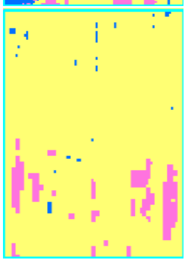
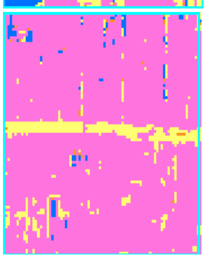
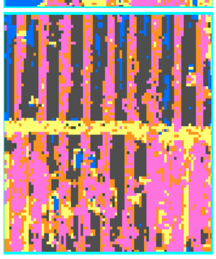
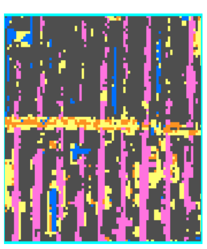
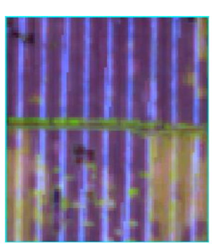
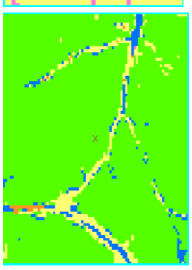
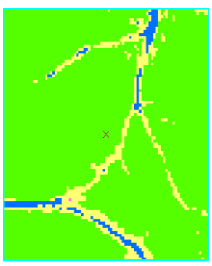
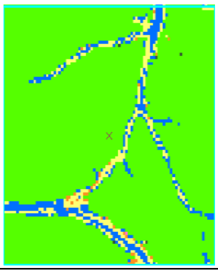
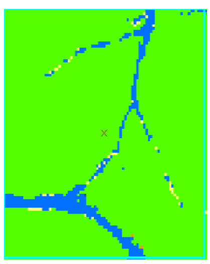
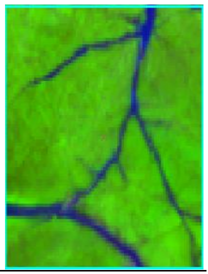
Thuật toán	Độ chính xác	Loại hiện trạng sử dụng đất							Độ chính xác tổng thể(OA)	Kappa
		Lúa	Cây hàng năm	Cây lâu năm	Mặt nước và Nuôi trồng thủy sản	Đất làm muối	Đất rừng	Đất phi nông nghiệp		
CART	Nhà sản xuất	61,4	0	61,0	80,0	0,0	83,3	76,3	71,3	0,58
	Người dùng	83,0	23,3	30,1	97,0	0,0	59,5	84,9		
SMV	Nhà sản xuất	63,5	0	42,2	77,1	0,0	86,2	76,6	71,6	0,58
	Người dùng	82,7	30,1	24,7	97,0	0,0	49,0	84,9		
RF	Nhà sản xuất	74,0	53,9	48,3	78,8	83,3	78,0	83,3	75,0	0,65
	Người dùng	77,3	46,1	37,7	95,5	33,3	76,2	86,8		
XGBoost	Nhà sản xuất	75,5	60,3	61,8	76,2	93,3	91,1	86,8	78,9	0,70
	Người dùng	77.5	54.1	54.5	97.0	56.0	78.8	89.1		

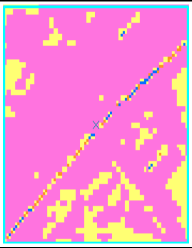
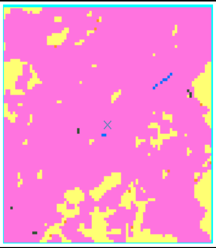
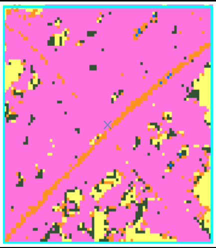
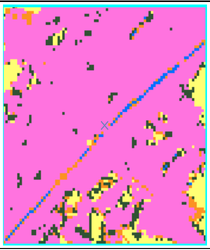
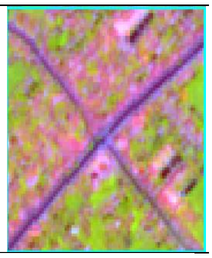
Mặt nước và nuôi trồng thủy sản cho độ chính xác cao ở tất cả các thuật toán sử dụng. Tuy nhiên đối với các loại hiện trạng sử dụng đất có diện tích nhỏ cũng như không có dấu hiện nhận biết rõ ràng trên ảnh viễn thám như đất làm muối và đất trồng cây hàng năm thi sai số phân loại rất lớn. Hai thuật toán là CART và RF không nhận diện được đối tượng đất làm muối. Đất rừng cũng có độ chính xác cao khi sử dụng thuật toán XGBoost (độ chính xác nhà sản xuất 91,1%). Đất phi nông nghiệp cũng có độ chính xác cao khi sử dụng 4 thuật toán học máy. Quy hoạch sử dụng đất tỉnh Nam Định đến năm 2030 chủ yếu là sự chuyển đổi từ đất nông nghiệp sang đất phi nông nghiệp. Do đó, sai số đối với các loại đất nông nghiệp có diện tích nhỏ, manh mún là không đáng kể và ảnh hưởng nhiều đến mục tiêu giám giám sát thực hiện quy hoạch sử dụng đất cấp tỉnh của nghiên cứu.

4. Kết luận

Nghiên cứu này sử dụng nguồn dữ liệu ảnh viễn thám Sentinel 2 kết hợp với các thuật toán phân loại cây quyết định (CART), thuật toán rừng ngẫu nhiên (RF), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGBoost) bằng cách xử dụng cùng một bộ mẫu huấn luyện và tập dữ liệu đánh giá độ chính xác để phân loại cho 7 loại hình hiện trạng sử dụng đất tỉnh Nam Định năm 2024. Kết quả đánh giá độ chính xác sau phân loại đã chỉ ra rằng với diện tích khoảng 1.700 km trên quy mô cấp tỉnh thuật toán XGBoost cho độ chính xác cao nhất với hệ số Kappa là 0,7 và độ chính xác toàn cục đạt 78,8%, mô hình CART cho độ chính xác thấp nhất với Kappa là 0,58. Quy hoạch sử dụng đất tỉnh Nam Định đến năm 2030 chủ yếu là sự chuyển đổi từ đất nông nghiệp sang đất phi nông nghiệp [23]. Do đó sai số đối với các loại đất nông nghiệp có diện tích nhỏ, manh mún là không đáng kể và ảnh hưởng nhiều đến mục tiêu giám giám sát thực hiện quy hoạch sử dụng đất cấp tỉnh của nghiên cứu.

Bảng 5. Khác biệt kết quả phân loại giữa hiện trạng lớp phủ sử dụng đất tỉnh Nam Định qua các thuật toán CART, SMV, RF, XGBoost.

Loại sử dụng	CART	SVM	RF	XGBOOST	Ảnh viễn thám
Lúa					
Cây hàng năm					
Cây lâu năm					
Mặt nước và NTS					
Đất làm muối					
Rừng					

Loại sử dụng	CART	SVM	RF	XGBOOST	Ảnh viễn thám
Đất phi nông nghiệp					

Đất trồng lúa · Cây lâu năm · Đất làm muối · Đất khác (phi nông nghiệp)
Cây hàng năm · Mặt nước và nuôi trồng thủy sản · Rừng

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.T.P.H.; Thu thập dữ liệu: L.T.A., N.D.P.; Phân tích và xử lí số liệu, chạy mô hình: N.T.T.B., L.T.A., N.D.P.; Phân tích kết quả nghiên cứu: N.T.P.H., N.T.T.B.; Viết bản thảo bài báo: N.T.P.H., N.T.T.B.

Lời cảm ơn: Nghiên cứu này được hỗ trợ bởi Đề tài: "Nghiên cứu ứng dụng trí tuệ nhân tạo và viễn thám phục vụ giám sát thực hiện quy hoạch sử dụng đất" Mã số: TNMT.2023.04.05 do Bộ Tài nguyên và Môi trường cấp kinh phí.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Đại học Kỹ thuật - Hậu cần CAND. Trực tuyến: https://dhkthc.bocongan.gov.vn/TrangChu/tin-tuc/99-tri-tue-nhan-tao-la-gi-lich-su-phat-trien-tri-tue-nhan-tao-ai.html.
2. Chen, Y.N.; Fan, K.C.; Chang, Y.L.; Moriyama, T. Special Issue Review: Artificial intelligence and machine learning applications in remote sensing. *Remote Sens.* **2023**, *15*, 569. https://doi.org/10.3390/rs15030569.
3. Tùng, Đ.T.; Ngọc, T.M. So sánh thuật toán học máy về phân loại lớp phủ bề mặt từ ảnh vệ tinh Sentinel 2 trên nền tảng Google Earth Engine. *J. Educ. Equip. Appl. Res.* **2023**, *2(301)*, 64–66.
4. Oanh, N.T.; Tám, P.M. Khả năng ứng dụng phân loại bề mặt lớp phủ bằng thuật toán Support Vector Machine. *Tạp chí Khoa học và Công nghệ lâm nghiệp* **2015**, *1-2015*, 140–147.
5. Tuấn, N.T.; Tuấn, P.V.; Quý, N.V.; Nhung, H.T.P. Sử dụng ảnh vệ tinh Sentinel 2 và thuật toán học máy thành lập bản đồ hiện trạng rừng tại huyện Bù Đăng, tỉnh Bình Phước. *Tạp chí Khoa học Trường Đại học Cần Thơ* **2022**, *58(6B)*, 150–163. Doi:10.22144/ctu.jvn.2022.254.
6. Trực tuyến: https://namdinh.gov.vn/portal/Pages/dieu_kien_tu_nhien.aspx.
7. Ân, N.T.; Bằng, T.N. Tác động của biến đổi khí hậu tới sử dụng đất nông nghiệp và sự thích ứng của người dân huyện Giao Thủy. *Tạp chí Nông nghiệp và Phát triển nông thôn* **2014**, tr. 110–116.
8. Cục Thống kê tỉnh Nam Định. 2021. Trực tuyến: https://www.namdinh.gso.gov.vn/.
9. Copernicus Data Space Ecosystem. Online available: https://browser.dataspace.copernicus.eu.
10. Hải, V.N. Ứng dụng thuật toán học máy theo dõi lớp phủ mặt nước phục vụ đào tạo, nghiên cứu trong lĩnh vực quản lý đất đai. *Equip. New General Educ. Program* **2023**, *1(304)*, 333–334.
11. Ngọc, N.T.A.; Hùng, T.Đ.; Hà, L.P. Ứng dụng phương pháp học máy - cây quyết định trong đánh giá biến động rừng ngập mặn khu vực xã đất mũi. *Tạp chí Khoa học Biến đổi khí hậu* **2021**, *20*, 28–34.
12. Li, F.; Li, M. Remote sensing image auto classify study based on combination of artificial neural networks and decision tree. *Remote Sens. Inf.* **2003**, *3*, 3–25.

13. Hải, P.M.; Long, V.K. Nghiên cứu ứng dụng phương pháp học máy (machine learning) trong phân loại rừng ngập mặn trên ảnh viễn thám spot6 với khu vực thử nghiệm tại tỉnh Cà Mau. *Tạp chí Khoa học và Bản đồ* **2019**, *40,* 17–21.
14. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29(5)*, 1189–1232. https://doi.org/10.1214/aos/1013203451.
15. Hà, B.M.; Thủy, N.B.; Ngọc, P.K.; Tiến, P.V. Kết quả bước đầu thử nghiệm thuật toán XGBoots dự báo nước dâng do bão tại trạm Hòn Dáu. *Tạp chí Khí tượng Thủy văn* **2023**, *756*, 83–94. doi: 10.36335/VNJHM.2023(756).83-94.
16. Luật Đất đai. 2024.
17. Jensen, J.R.; Lulla, K. Introductory digital image processing: A remote sensing perspective. *Geocarto Int.* **1987**, *2(1)*, 65–65. Doi:10.1080/10106048709354084.
18. Trực tuyến: https://naue.edu.vn/.
19. Oanh, N.T.; Trường, T.X.; Tuyết, V.T.; Trung, T.H. Nâng cao độ chính xác phân loại ảnh viễn thám bằng phương pháp kết hợp phân loại dựa trên đối tượng và phân loại dựa trên điểm ảnh. *Tạp chí Khoa học Đo đạc và Bản đồ* **2022**, *53*, 48–56.
20. Hải, L.V.H.; Hoài, Đ.T.; Long, V.K. Nghiên cứu phương pháp phân loại hướng đối tượng trên tư liệu ảnh máy bay không người lái. *Tạp chí Khoa học Đo đạc và Bản đồ* **2018**, *35*, 38–43.
21. Điệp, N.T.H.; Nghĩa, N.M.; Phát, C.N.; Hiếu, D.C.; Nguyễn, N.T.; Diễm, P.K. Phân tích tổn thương trên hiện trạng sử dụng đất nông nghiệp do tác động xâm nhập mặn tại tỉnh Tiền Giang năm 2020 ứng dụng ảnh viễn thám. *Tạp chí Khoa học Trường Đại học Cần Thơ* **2023**, *59*, 185–192.
22. Roberts, C. Modelling patterns of agreement for nominal scales. *Stat. Med.* **2008**, *27(6)*, 810–830. Doi: 10.1002/sim.2945.
23. Quyết định số 1396/QĐ-UBND ngày 02/7/2021 của Ủy ban Nhân dân tỉnh Nam Định về việc Quy hoạch sử dụng đất đến năm 2030 và kế hoạch sử dụng đất năm đầu của quy hoạch sử dụng đất thành phố Nam Định, tỉnh Nam Định.

Evaluation the accuracy results of land use classification based on machine learning algorithms and remote sensing data

Nguyen Thi Phuong Hoa[1]*, Nguyen Thi Thanh Binh[1], Le Thanh An[2], Nguyen Duy Phuong[2]

[1] National Remote Sensing Department, Ministry of Natural Resources and Environment, ntphoa19@gmail.com; binhntt1191@gmail.com

[2] Instterof International Studies, Ministry of Public Security, minhtn122019@gmail.com; nanh22022022@gmail.com

Abstract: This research used four machine learning algorithms such as decision trees (CART), random forest algorithms (RF), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGBoost) for land use classification of Sentinel 2 into land use categories in Nam Dinh province. By using the same training set and accuracy evaluation dataset, this study will evaluate the accuracy of classification models from Sentinel 2 optical satellite images for 7 types of land use in Nam Dinh province in 2024. The results of the accuracy evaluation after classification showed that with an area of about 1,700 km^2 on a provincial scale, the XGBoost algorithm gave the highest accuracy with a Kappa coefficient of 0.7 and a global accuracy of 78.8%, the CART model gave the lowest accuracy with a Kappa of 0.58. Small land use objects such as salt land were not recognized by the CART and SVM models. However, the water surface objectwas more accurate when using the RF and XGBoost algorithms.

Key words: Nam Dinh, Sentinel 2; Classification And Regression Tree (CART); Random Forest (RF); Support Vector Machines (SVM); Extreme Gradient (XCBoost).