TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN



*Bài báo khoa học*

# 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

**Phạm Thị Thanh Hòa[1,2]*, Vũ Ngọc Quang[3], Lê Thanh Nghị[1,2], Đoàn Thị Nam Phương[1,2], Nguyễn Minh Hải[1]**

[1] Khoa Trắc địa - Bản đồ và Quản lý đất đai, Trường Đại học Mỏ - Địa chất; phamthithanhhoa@humg.edu.vn; lethanhnghi@humg.edu.vn; doanthinamphuong@humg.edu.vn; nguyenminhhai@humg.edu.vn

[2] Nhóm nghiên cứu Công nghệ Địa tin học trong Khoa học Trái đất (GES), Trường Đại học Mỏ - Địa chất; phamthithanhhoa@humg.edu.vn; lethanhnghi@humg.edu.vn; doanthinamphuong@humg.edu.vn

[3] Khoa Công trình, Trường Đại học Công nghệ giao thông vận tải; quangvn@utt.edu.vn

***Tác giả liên hệ:** phamthithanhhoa@humg.edu.vn; Tel.: +84–977732505

Ban Biên tập nhận bài: 8/9/2023; Ngày phản biện xong: 18/10/2023; Ngày đăng bài: 25/12/2023

**Tóm tắt:** Trong kỷ nguyên công nghệ mới, phương pháp học máy (*Machine learning*) dần thay thế các phương pháp truyền thống trong lĩnh vực viễn thám. Một trong những thuật toán có độ chính xác cao trong phân loại là *Random Forest* (Rừng ngẫu nhiên - RF). Cùng với đó, thay vì phân loại ảnh trên các phần mềm thương mại, nền tảng đám mây Google Colab giúp tối ưu hóa thời gian xử lý với nguồn thư viện phong phú và đặc biệt phù hợp với phương pháp học máy. Vì vậy, nghiên cứu đã tiến hành phân loại lớp phủ mặt đất sử dụng thuật toán Random Forest trên nền tảng Google Colab, thực nghiệm tại tỉnh Quảng Bình với thời gian là tháng 8 năm 2022. Ảnh vệ tinh Sentinel-2 được lựa chọn do độ phân giải không gian cao hơn so với các ảnh miễn phí khác. Đồng thời, nghiên cứu cũng so sánh kết quả phân loại RF trong hai trường hợp: (1) sử dụng bốn kênh ảnh có độ phân giải 10m của ảnh Sentinel-2, (2) kết hợp 4 kênh ảnh trên và các ảnh chỉ số NDVI, NDWI, NDBI. Cả hai trường hợp đều đạt độ chính xác tổng thể trên 90% và Kappa trên 0,9, cho thấy tính khả thi của thuật toán RF. Trong đó, trường hợp (2) đạt độ chính xác cao hơn, khẳng định rằng việc sử dụng các chỉ số quang phổ giúp làm tăng thông tin và cải thiện kết quả phân loại.

**Từ khóa:** Random Forest; Sentinel-2; Lớp phủ bề mặt; Google Colab.

## 1. Mở đầu

Trong điều kiện tự nhiên, lớp phủ mặt đất (*land cover*) tích hợp và phản ánh khí hậu, địa chất, đất đai và hệ sinh vật sẵn có của một khu vực tại một thời điểm, theo tháng hoặc năm, có thể hàng thập kỷ hoặc lâu hơn. Nó được xem là nguồn thông tin đầu vào quan trọng trong các nghiên cứu lũ lụt, hạn hán, xói mòn, cũng như cần thiết trong quản lý, giám sát đối tượng lớp phủ mặt đất [1–2]. Nhiều nghiên cứu và nhiều phương pháp được lựa chọn để theo dõi lớp đối tượng này. Cho đến nay, việc chiết tách lớp phủ sử dụng phương pháp chủ yếu là viễn thám (*Remote sensing*) [3]. Việc sử dụng tư liệu viễn thám trong thành lập bản đồ lớp phủ tương đối đơn giản và khá nhanh chóng, được đánh giá là mang lại hiệu quả tốt, vừa có thể tiết kiệm được chi phí và công sức. Công nghệ viễn thám đang ở giai

đoạn phát triển vượt trội, với số lượng ảnh viễn thám và các ứng dụng không ngừng phát triển qua các năm. Số lượng lớn ảnh miễn phí với nhiều độ phân giải khác nhau được dùng trong thành lập bản đồ lớp phủ [4–7].

Trong kỷ nguyên công nghệ mới, chuyển đổi số mang lại nhiều sự đột phá trong nhiều lĩnh vực với việc xuất hiện của Trí tuệ nhân tạo (*Artificial Intelligence - AI*) và Học máy (*Machine Learning*). Khi đó, khoa học công nghệ gắn liền với việc xử lý nguồn dữ liệu lớn (Big data) và phương tiện hiện đại. Hiện nay, việc kết hợp mô phỏng chủ đề vào các thuật toán Machine learning trở thành một hướng nghiên cứu mới mà nhiều nhà khoa học quan tâm [8–9], trong đó đặc biệt nhấn mạnh sự kết hợp của học máy và lĩnh vực viễn thám. Các kỹ thuật phân loại dựa trên học máy xuất hiện và trở thành hướng tiếp cận mới trong nghiên cứu lớp phủ mặt đất [10]. Một trong những thuật toán học máy có giám sát mang tính khả thi là rừng ngẫu nhiên (*Random Forest - RF*). Các nhà khoa học đã sử dụng Random Forest trong thành lập bản đồ lớp phủ mặt đất với độ chính xác cao. Nghiên cứu [11] cho thấy thuật toán RF mang lại sự phân loại lớp phủ mặt đất ở phía nam Tây Ban Nha với độ chính xác tổng thể là 92% và chỉ số Kappa là 0,92. Trong khi kết quả này ở nghiên cứu của [12] tương ứng là 84,6% và Kappa 0,808. Các nghiên cứu [13–14] đánh giá RF có độ chính xác cao hơn một số phương pháp phân loại khác như Maximum Likelihood, khoảng cách tối thiểu, cây quyết định, mạng Nơ Ron nhân tạo và Máy vectơ hỗ trợ (*Support Vector Machine*). Nhìn chung, các nghiên cứu đã chứng minh tính hiệu quả của thuật toán rừng ngẫu nhiên trong nghiên cứu lớp phủ [11–14].

Một vấn đề khác cần chú ý là việc phân loại lớp phủ mặt đất truyền thống thường yêu cầu khối lượng tính toán khổng lồ, đôi khi gây ra áp lực trong quá trình phân tích và xử lý ảnh viễn thám. Do đó, cần lựa chọn một nền tảng xử lý cho phép giảm bớt sự phụ thuộc vào tài nguyên cơ sở hạ tầng máy tính, cũng như giảm bớt gánh nặng về dung lượng ổ cứng máy tính. Một trong những nền tảng cho phép thực hiện trong lĩnh vực viễn thám là Google Colab. Sự xuất hiện của Google Colab phép các nhà nghiên cứu thực thi mã xử lý ảnh thông qua kết nối Internet và đặc biệt phù hợp với phương pháp học sâu [15] và học máy [16].

Hầu hết các nghiên cứu trước đây về lớp phủ mặt đất chưa tiếp cận sử dụng nền tảng Google Colab. Đồng thời, muốn khẳng định tính hiệu quả về độ chính xác của kỹ thuật học máy, nhóm nghiên cứu đã tích hợp thuật toán Random Forest trên nền tảng Google Colab để phân loại lớp phủ mặt đất ở tỉnh Quảng Bình, trên cơ sở sử dụng ảnh vệ tinh Sentinel-2. Như vậy, thay vì phân loại ảnh trên các phần mềm thương mại, nghiên cứu tiến hành lập trình Python trong môi trường Google Colab, giúp tối ưu hóa thời gian xử lý ảnh, tận dụng ưu điểm đơn giản và nguồn thư viện phong phú của ngôn ngữ Python. Ảnh Sentinel-2 được lựa chọn do được tích hợp sẵn trên nền tảng điện toán đám mây, miễn phí và có độ phân giải không gian cao hơn so với ảnh khác như Landsat, Modis. Đây chính là hướng tiếp cận mới cho khu vực tỉnh Quảng Bình khi mà trước đây có rất ít các nghiên cứu về lớp phủ.

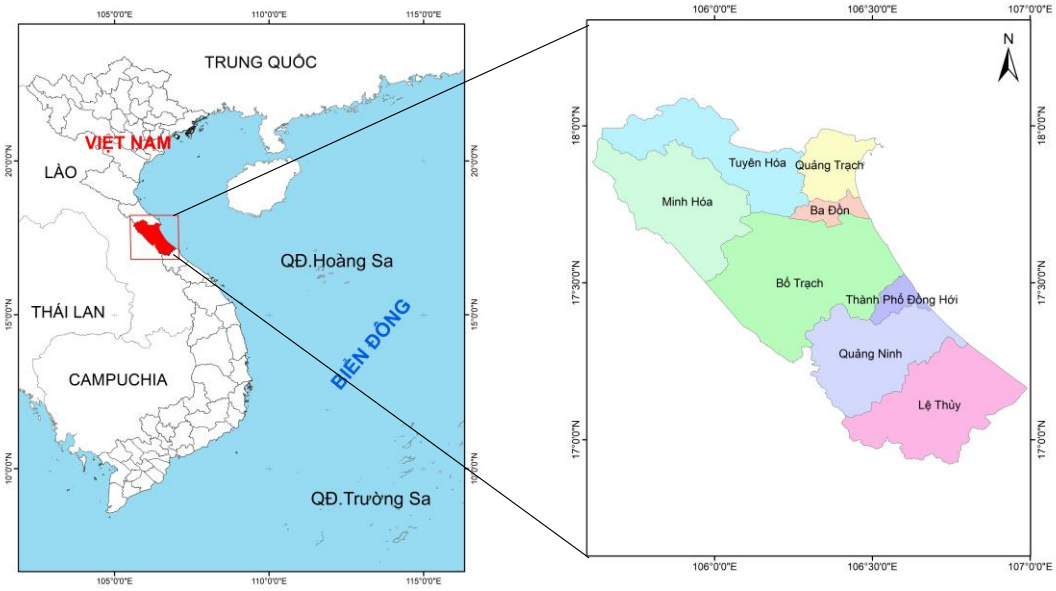
Nghiên cứu sử dụng thuật toán Random Forest được tiến hành theo hai hướng: (1) sử dụng bốn kênh ảnh có độ phân giải 10 m (kênh 2, kênh 3, kênh 4, kênh 8) của Sentinel-2, (2) sử dụng bốn kênh ảnh có độ phân giải 10m trên và bổ sung thêm các ảnh chỉ số phổ NDVI (*Normalized Difference Vegetation Index* - chỉ số thực vật khác biệt chuẩn hóa), NDWI (*Normalized Difference Water Index* - chỉ số nước khác biệt chuẩn hóa), NDBI (*Normalized Difference Built-up Index* - chỉ số xây dựng khác biệt chuẩn hóa). Mục tiêu cụ thể của nghiên cứu là: (1) Phân loại đối tượng lớp phủ mặt đất ở tỉnh Quảng Bình; (2) Đánh giá tiềm năng của thuật toán Random Forest thông qua kết quả đánh giá độ chính xác; (3) So sánh hai hướng tiếp cận; từ đó lựa chọn hướng tối ưu trong phân loại lớp phủ ở tỉnh Quảng Bình khi sử dụng thêm các chỉ số phổ làm tăng lượng thông tin và khả năng nhận biết từng đối tượng đặc trưng; (4) Đánh giá tiềm năng của nền tảng Google Colab.

## 2. Tài liệu thu thập và phương pháp nghiên cứu

### *2.1. Khu vực nghiên cứu*

Quảng Bình là tỉnh thuộc vùng duyên hải Bắc Trung Bộ, Việt Nam với giới hạn tọa độ địa lý từ 16°55’ đến 18°05’ vĩ độ Bắc và từ 105°37’ đến 107°00’ kinh độ Đông. Tỉnh tiếp giáp với tỉnh Hà Tĩnh ở phía Bắc, phía Nam giáp tỉnh Quảng Trị, phía Đông giáp biển với chiều dài trên 116,04 km và đường biên giới phía Tây giáp Lào có tổng chiều dài 222,118 km.

Tỉnh Quảng Bình nằm ở sườn Đông dãy Trường Sơn, với địa hình đồi núi cao hiểm trở, bề ngang hẹp và dốc, nghiêng từ Tây sang Đông và chia cắt ở các khu vực phía Tây của tỉnh. Địa hình phân chia thành các tiểu vùng: Vùng núi cao tập trung ở phía Tây của Tỉnh và nằm dọc theo sườn Đông dãy Trường Sơn; Vùng gò đồi và trung du; Vùng đồng bằng và vùng cát ven biển. Phần lớn diện tích tỉnh là đồi núi (chiếm trên 85% diện tích tự nhiên), các dạng địa hình khác là trung du, đồng bằng duyên hải và bãi cát ven biển chỉ chiếm gần 15%. Do vậy, diện tích đất canh tác lúa tương đối hạn chế và thường xuyên gánh chịu lũ lụt bất thường vào mùa mưa do hệ thống sông suối ngắn, dốc, chảy từ Tây sang Đông (theo Cổng thông tin điện tử tỉnh Quảng Bình).




**Hình 1.** Vị trí khu vực nghiên cứu.

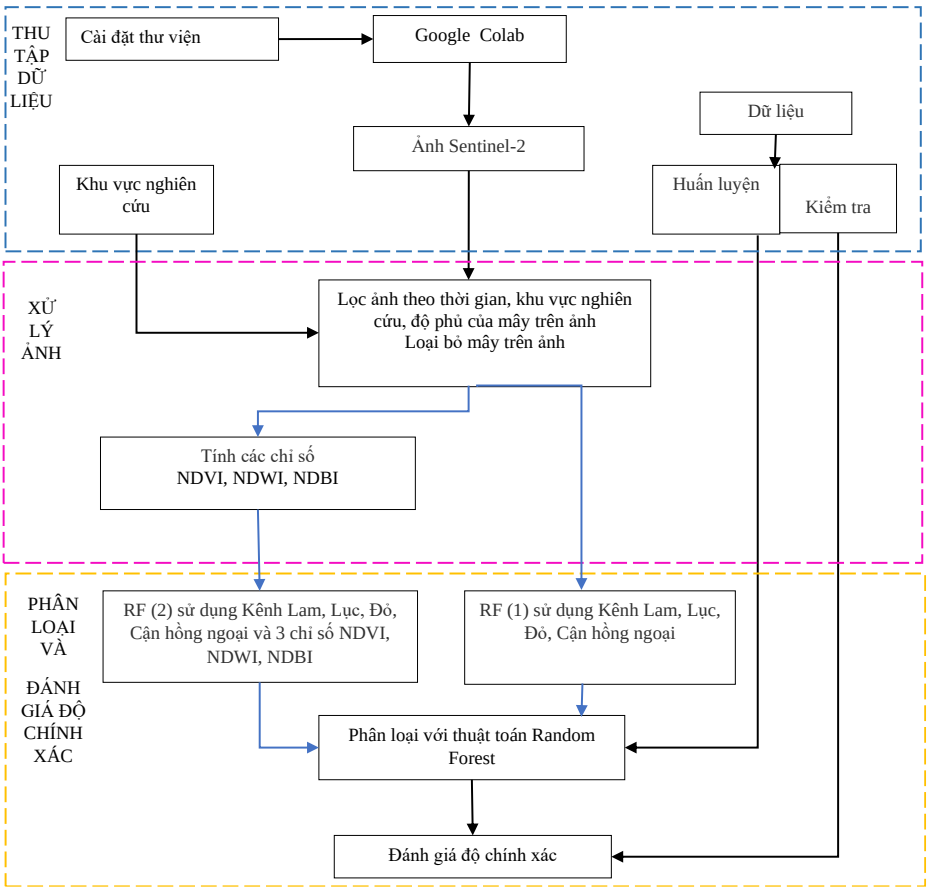
### *2.2. Dữ liệu ảnh vệ tinh*

Nghiên cứu sử dụng ảnh Sentinel-2 với độ che phủ mây dưới 30%, phủ trùm khu vực nghiên cứu vào tháng 8 năm 2022. Sentinel-2 là vệ tinh được nghiên cứu và phát triển bởi ESA - cơ quan hàng không vũ trụ châu Âu, có nhiệm vụ quan sát Trái đất, hỗ trợ các nghiên cứu giám sát thảm thực vật, lớp phủ mặt đất cũng như các tai biến thiên nhiên.

Hệ thống này bao gồm hai vệ tinh quay quanh cực trên cùng một quỹ đạo nhưng lệch pha nhau 180°, được thiết kế với chu kỳ lặp 5 ngày (kết hợp cả hai vệ tinh), độ rộng dải chụp 290 km. Ảnh Sentinel-2 đa phổ (*MultiSpectral Instrument*) gồm 13 kênh phổ: bốn kênh ở độ phân giải không gian10 m, sáu kênh ở 20 m và ba kênh ở độ phân giải 60 m (theo ESA).

### *2.3. Phương pháp nghiên cứu*

Hình 2 thể hiện sơ đồ quy trình ứng dụng thuật toán Random Forest và ảnh Sentinel-2 trên nền tảng Google Colab để phân loại lớp phủ mặt đất.

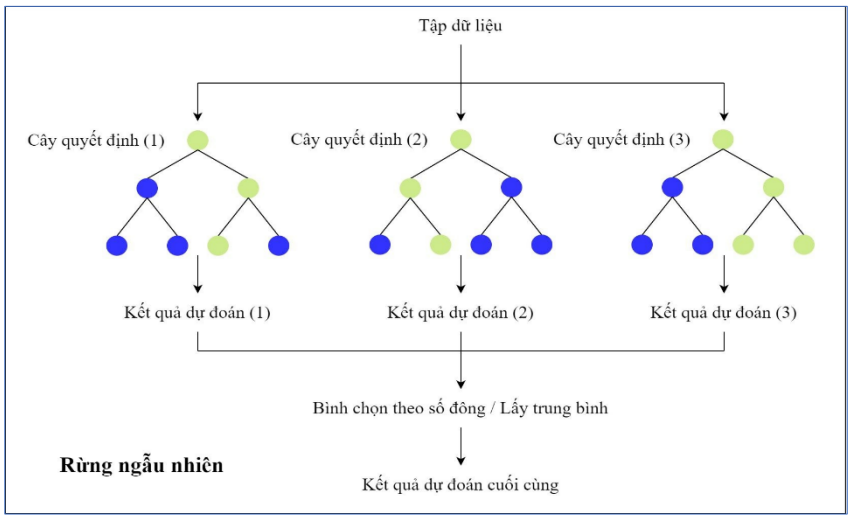



**Hình 2.** Sơ đồ quy trình.

### 2.3.1. Thuật toán Random Forest (Rừng ngẫu nhiên - RF)

Random Forest được đề xuất bởi Breiman [17] vào năm 2001. Đây là một thuật toán học máy có giám sát dễ sử dụng và linh hoạt, đồng thời cũng được sử dụng phổ biến để giải quyết nhiệm vụ phân loại và hồi quy.

Random Forest được hiểu là "Rừng ngẫu nhiên", bắt nguồn từ thuật toán Decision tree (Cây quyết định), nó phát triển nhiều cây quyết định và kết hợp chúng với nhau. Với phân loại Random Forest, cây quyết định được tạo bằng cách sử dụng các tập hợp con ngẫu nhiên khác nhau của dữ liệu và tính năng nhất định. Mỗi cây sẽ cung cấp dự đoán của nó để phân loại. Random Forest dựa trên đa số kết quả dự đoán và lấy kết quả phổ biến nhất làm đầu ra cuối cùng.




**Hình 3.** Thuật toán Random Forest (*https://interactivechaos.com/en/wiki/random-forest*)

Sử dụng thuật toán Random Forest, cần đặc biệt chú ý tham số "number of trees"- số lượng cây quyết định

bởi nó ảnh hưởng lớn đến kết quả và độ chính xác dự báo [17]. Ngoài ra, còn có các tham số khác liên quan đến số lượng thuộc tính dùng để xây dựng cây.

Random Forest hoạt động như sau:

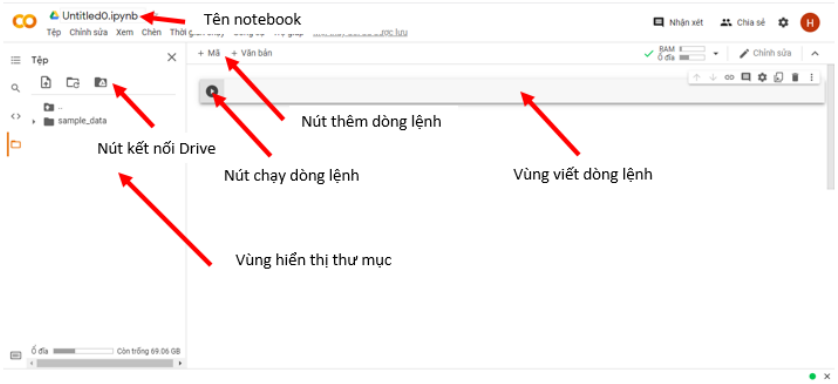
Từ tập dữ liệu D ban đầu, tạo dữ liệu ngẫu nhiên (mẫu bootstrap) Dj có kích thước N sử dụng phương pháp "*bagging*". Những dữ liệu còn lại không được lựa chọn tham gia vào quá trình huấn luyện gọi là dữ liệu "*out -of -bag*".

Tiếp đó, sử dụng mẫu Dj làm dữ liệu huấn luyện: Các tập con dữ liệu lấy mẫu ngẫu nhiên $D_1$, $D_2$,…, $D_k$ để thiết lập nên các cây quyết định $T_1$, $T_2$,…,$T_k$. Khi đó, sử dụng các cây đã tạo để nhận kết quả dự đoán. Cuối cùng, tổng hợp kết quả từ các cây quyết định theo hình thức: với bài toán phân loại, lựa chọn theo đa số (tức là số phiếu bầu cao nhất), với bài toán hồi quy, lấy trung bình các giá trị dự đoán.

### 2.3.2. Nền tảng Google Colab

Bên cạnh sự xuất hiện của *Google Earth Engine (GEE)* - một nền tảng lưu trữ và xử lý dữ liệu lớn (*Big data*), *Google Research* tạo ra *Google Colaboratory* (còn gọi là *Google Colab*) cho phép thực thi trên nền tảng đám mây. Google Colab thay thế cho Google Earth Engine khi dễ dàng xử lý dữ liệu trong tài nguyên máy tính [18]. Google Colab phù hợp để thực thi các bài toán đòi hỏi khối lượng tính toán khổng lồ, đi kèm với các thư viện hỗ trợ cho dự án deep learning/machine learning và khoa học dữ liệu.

Google Colab có ưu điểm là được thiết kế chạy mã Python thông qua trình duyệt, cho phép thay thế các phần mềm xử lý ảnh, cũng như không cần nâng cấp phần cứng máy tính. Khi sử dụng Google Colab, các cơ sở hạ tầng như bộ nhớ, khả năng xử lý, đơn vị xử lý đồ họa (GPU) và đơn vị xử lý tensor (TPU) được cung cấp quyền truy cập miễn phí. Và đặc biệt Google Colab giải quyết được các bài toán trong nhiều lĩnh vực với chi phí thấp và thời gian ngắn.




**Hình 4.** Giao diện Google Colab (*https://interactivechaos.com/en/wiki/random-forest*).

### 2.3.3. Công thức xác định các chỉ số phổ trong nghiên cứu

Phân loại lớp phủ mặt đất trong nghiên cứu sử dụng thuật toán Random Forest được tiến hành theo hai hướng: (1) sử dụng bốn kênh ảnh có độ phân giải 10m (kênh 2, kênh 3, kênh 4, kênh 8 - tương ứng kênh lam, lục, đỏ và cận hồng ngoại) của Sentinel-2, (2) sử dụng bốn kênh ảnh có độ phân giải 10m và bổ sung thêm các ảnh chỉ số phổ. Nhiều nghiên cứu đã sử dụng các chỉ số phổ để cải thiện kết quả phân loại: nghiên cứu chỉ sử dụng chỉ số NDVI [14, 19], trong khi nghiên cứu sử dụng kết hợp nhiều loại chỉ số phổ khác nhau [20].

Trong nghiên cứu này, các chỉ số phổ được sử dụng bao gồm:

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

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

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad (3)$$

Trong đó NDVI (*Normalized Difference Vegetation Index*): chỉ số thực vật khác biệt chuẩn hóa [21]. Giá trị NDVI nằm trong khoảng từ -1 đến 1, giúp định lượng sức khỏe và mật độ của thảm thực vật. NDVI cao đại diện cho khu vực có độ che phủ thảm thực vật cao, giá trị âm thể hiện khu vực mặt nước; NDWI (*Normalized Difference Water Index*): chỉ số nước khác biệt chuẩn hóa. Giá trị NDWI giúp phân biệt giữa vùng nước so với các đối tượng thực vật [22]; NDBI (*Normalized Difference Built-up Index*): chỉ số xây dựng khác biệt chuẩn hóa. Giá trị NDBI dùng để nhấn mạnh các khu vực xây dựng [23]; NIR là giá trị bức xạ của sóng hồng ngoại gần; RED là giá trị bức xạ của bước sóng đỏ; GREEN là giá trị bức xạ của bước sóng màu lục; SWIR là giá trị bức xạ của hồng ngoại sóng ngắn.

## 3. Kết quả và thảo luận

### *3.1. Thực hiện phân loại*

Python đã tích hợp sẵn một số thư viện trong Google Colab như Geopandas/panda - Matplotlib…Bên cạnh đó, Google Colab kết nối với Google Earth Engine để lấy dữ liệu ảnh Sentinel-2 cho khu vực nghiên cứu, với thời gian là tháng 8 năm 2022, độ phủ mây dưới 30%. Kết quả hiển thị trên geemap - gói Python để phân tích và trực quan hóa không gian địa lý. Từ tập dữ liệu ảnh thu thập được, tiến hành loại bỏ mây dựa trên kênh ảnh "probability" (COPERNICUS/S2_CLOUD_PROBABILITY). Sau đó sử dụng hàm "median" và "clip" để ghép các ảnh trong bộ sưu tập và cắt thành một ảnh theo ranh giới khu vực nghiên cứu.

Trong nghiên cứu này, gần 1800 điểm được lựa chọn trong quá trình phân loại, trong đó sử dụng 70% làm dữ liệu huấn luyện (training data) và 30% được sử dụng làm dữ liệu đánh giá (validation data). Tỷ lệ này cũng thường được sử dụng trong các bài toán phân loại bằng Random Forest [20]. Số lượng cây được lựa chọn là 300 trên cơ sở các thử nghiệm của nhóm nghiên cứu. Như vậy theo hướng (1), phân loại RF với việc lựa chọn 4 kênh ảnh dùng để huấn luyện được thể hiện như hình 5 (trên nền tảng Google Colab).

```
[ ]  # Select bands for training
     bands4 =['B2', 'B3', 'B4','B8']
     # Get training
     trainingsample4 = S2_StudyArea.select(bands4).sampleRegions(**{
         'collection':trainingfc,
         'properties': ['Class'],
         'scale': 10
     })
     validationsample4 = S2_StudyArea.select(bands4).sampleRegions(**{
         'collection':testingfc,
         'properties': ['Class'],
         'scale': 10
     })
     ####### RF Classifier Model Building
     # ee.Classifier.smileRandomForest(numberOfTrees, variablesPerSplit, minLeafPopulation, bagFraction, maxNodes, se
     #numberOfTrees: 300,variablesPerSplit:null, minLeafPopulation: 1, maxNodes: null})
     RFclassifier4 = ee.Classifier.smileRandomForest(300).train(**{
         'features': trainingsample4,
         'classProperty': 'Class',
         'inputProperties': bands4,
     })
     # Classify the image
     RFClassified4 = S2_StudyArea.select(bands4).classify(RFclassifier4)
```

**Hình 5.** Code trên Google Colab phân loại theo hướng (1) sử dụng 4 kênh ảnh 10m của Sentinel-2.

Theo hướng (2), tính toán bổ sung các chỉ số đã nêu trong mục 2.3.3 với code thể hiện trong hình 6. Hình 7 thể hiện code phân loại theo hướng (2).

```
##### Image Classification with indices: NDVI, NDWI, BSI
# NDVI=(NIR-R)/(NIR+R)
NDVI=S2_StudyArea.normalizedDifference(['B8','B4']).rename('NDVI')
#NDWI = (G-NIR)/(G+NIR)
NDWI=S2_StudyArea.normalizedDifference(['B3','B8']).rename('NDWI')
#NDBI = (NIR-SWIR1)/(NIR+SWIR1)
NDBI=S2_StudyArea.normalizedDifference(['B11','B8']).rename('NDBI')
```

**Hình 6.** Code tính toán các chỉ số NDVI, NDWI, NDBI.

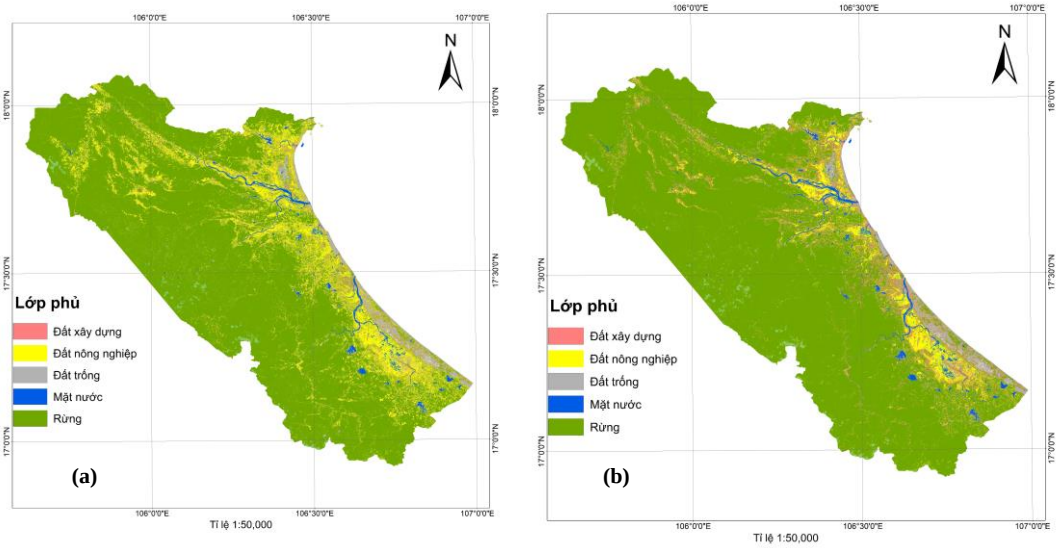
```
# Select 4 bands and 3 spectral indice for training
bands7 =['B2', 'B3', 'B4','B8', 'NDVI', 'NDWI', 'NDBI']
# Get training
trainingsample7 = S2_StudyArea3.select(bands7).sampleRegions(**{
    'collection':trainingfc,
    'properties': ['Class'],
    'scale': 10
})
validationsample7 = S2_StudyArea3.select(bands7).sampleRegions(**{
    'collection':testingfc,
    'properties': ['Class'],
    'scale': 10
})

####### RF Classifier Model Building
# ee.Classifier.smileRandomForest(numberOfTrees, variablesPerSplit, minLeafPopulation, bagFraction, maxNodes, seed)
#numberOfTrees: 300,variablesPerSplit:null, minLeafPopulation: 1, maxNodes: null})
RFclassifier7 = ee.Classifier.smileRandomForest(300).train(**{
    'features': trainingsample7,
    'classProperty': 'Class',
    'inputProperties': bands7,
})
# Classify the image
RFClassified7 = S2_StudyArea3.select(bands7).classify(RFclassifier7)
```

**Hình 7.** Code phân loại theo hướng (2) sử dụng 4 kênh ảnh 10m của Sentinel-2 và 3 chỉ số phổ.

### *3.2. Kết quả phân loại*

Hình 8 và hình 9 tương ứng là ảnh Sentinel-2 cắt theo khu vực tỉnh Quảng Bình trên giao diện Geemap trong Google Colab và bản đồ kết quả phân loại lớp phủ mặt đất theo hai hướng tiếp cận.


**Hình 8.** Ảnh Sentinel-2 khu vực tỉnh Quảng Bình vào tháng 8 năm 2022 (Tổ hợp màu thực) trên nền Geemap.




**Hình 9.** Kết quả phân loại lớp phủ mặt đất tỉnh Quảng Bình với Random Forest: (a) Trường hợp 1: sử dụng bốn kênh ảnh B2, B3, B4, B8; (b) Trường hợp 2: sử dụng bốn kênh ảnh B2, B3, B4, B8 và 3 ảnh chỉ số NDVI, NDWI. NDBI.

Hình 9 thể hiện kết quả phân loại của mô hình Random Forest của khu vực nghiên cứu trong hai trường hợp đã trình bày ở phần 2.3.3. Các lớp phủ được phân loại bao gồm: Rừng, Đất nông nghiệp (gồm đất trồng lúa, hoa màu, cây ăn quả…), Đất trống (gồm đất trống, cát, đá sỏi, đồi trọc …), Mặt nước (sông, hồ, kênh mương) và Đất xây dựng (gồm dân cư và cơ sở hạ tầng). Nhìn chung kết quả phân loại cho thấy diện tích rừng tỉnh Quảng Bình chiếm tỉ lệ lớn, tập trung chủ yếu ở vùng núi phía Tây tỉnh Quảng Bình. Vùng đất trống (cát) nằm ven biển phía Đông của tỉnh. Các lớp khác như đất nông nghiệp, đất xây dựng phân bố rải rác; vùng mặt nước như sông, hồ, kênh mương được phân loại chi tiết trên ảnh. Tuy nhiên với kết quả phân loại trong trường hợp (1), phần diện tích đất nông nghiệp lớn hơn so với trường hợp (2). Để kiểm tra kết quả phân loại nào tốt hơn, nhóm đã tiến hành đánh giá độ chính xác và đánh giá trực quan trên hai ảnh phân loại.

### *3.3. Đánh giá độ chính xác*

Độ chính xác của hai hướng tiếp cận trên được đánh giá bằng ma trận nhầm lẫn (*Confusion Matrix*), độ chính xác tổng thể (*OA-Overall Accuracy*), Kappa, độ chính xác của nhà sản xuất (PA) và độ chính xác của người dùng (*UA- user's accuracy*).

Ma trận nhầm lẫn là bảng tóm tắt hiệu suất của thuật toán phân loại, trong đó bao gồm giá trị phân loại và giá trị kiểm tra/tham chiếu. Mỗi hàng và mỗi cột sẽ tương ứng với từng lớp phủ được phân loại và lớp kiểm tra/tham chiếu. Đường chéo chính của ma trận liệt kê số lượng pixel được phân loại chính xác. Các giá trị còn lại cho biết mỗi lớp phủ được phân loại thế nào khi so với lớp kiểm tra, lớp nào được phân loại đúng nhiều nhất và lớp nào bị phân loại nhầm vào lớp khác.

Hệ số Kappa [24] là hệ số trong thống kê, dùng để đo đạc độ đồng thuận giữa các thành phần định tính (phân loại), là thước đo mức độ phù hợp tổng thể của ma trận.

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

Trong đó $P_o$ là giá trị đồng thuận quan sát được giữa các biến đánh giá; $P_e$ là xác suất của sự đồng thuận ngẫu nhiênPo và Pe được tính toán dựa trên dữ liệu trong bảng ma trận để tính toán xác suất ngẫu nhiên cho mỗi nhóm.

Một thước đo cơ bản là độ chính xác tổng thể, được tính bằng cách chia các pixel được phân loại chính xác (tổng các giá trị trong đường chéo chính) cho tổng số pixel được kiểm tra.

$$\text{OA} = \frac{N}{T} \tag{5}$$

Trong đó N là Số lượng các pixel được phân loại chính xác; T là Tổng số lượng các pixel được kiểm tra.

Bên cạnh độ chính xác tổng thể, độ chính xác phân loại của từng lớp riêng lẻ có thể được tính theo cách tương tự: độ chính xác của người dùng (còn gọi là Độ chính xác sai sót/sử dụng) và độ chính xác của nhà sản xuất PA - Producer's accuracy (còn gọi là Độ chính xác thực hiện). Độ chính xác của nhà sản xuất được tính bằng cách chia số pixel chính xác trong một lớp chia cho tổng số pixel được lấy từ dữ liệu tham chiếu. Độ chính xác của người dùng UA-User's accuracy, về cơ bản cho biết mức độ thường xuyên mà lớp trên bản đồ sẽ thực sự hiện diện trên mặt đất, được tính là các pixel được phân loại chính xác trong một lớp chia cho tổng số pixel được phân loại trong lớp đó [2]. Độ chính xác của kết quả phân loại theo hai hướng tiếp cận trên được thể hiện trong bảng 1 đến bảng 4.

**Bảng 1.** Ma trận nhầm lẫn của phương pháp phân loại RF 4 kênh ảnh (đơn vị: pixel).

| Lớp phủ | Đất xây dựng | Đất nông nghiệp | Đất trống | Mặt nước | Rừng |
|---|---|---|---|---|---|
| Đất xây dựng | 167 | 0 | 3 | 0 | 1 |
| Đất nông nghiệp | 5 | 97 | 0 | 0 | 6 |
| Đất trống | 3 | 5 | 64 | 1 | 1 |
| Mặt nước | 0 | 2 | 0 | 77 | 0 |
| Rừng | 2 | 5 | 0 | 0 | 98 |

**Bảng 2.** Ma trận nhầm lẫn của phương pháp phân loại RF sử dụng 4 kênh ảnh và 3 chỉ số phổ (đơn vị: pixel).

| Lớp phủ | Đất xây dựng | Đất nông nghiệp | Đất trống | Mặt nước | Rừng |
|---|---|---|---|---|---|
| Đất xây dựng | 166 | 1 | 3 | 0 | 1 |
| Đất nông nghiệp | 5 | 98 | 0 | 1 | 4 |
| Đất trống | 0 | 5 | 67 | 1 | 1 |
| Mặt nước | 0 | 1 | 0 | 78 | 0 |
| Rừng | 1 | 1 | 0 | 0 | 103 |

**Bảng 3.** Độ chính xác tổng thể (OA) và hệ số kappa trong 2 trường hợp.

| Trường hợp phân loại RF | Độ chính xác tổng thể (OA) % | Kappa |
|---|---|---|
| (1) 4 kênh ảnh | 93,7 | 0,92 |
| (2) 4 kênh ảnh và 3 kênh ảnh chỉ số | 95,3 | 0,94 |

**Bảng 4.** Độ chính xác thực hiện PA và độ chính xác sử dụng UA trong 2 trường hợp.

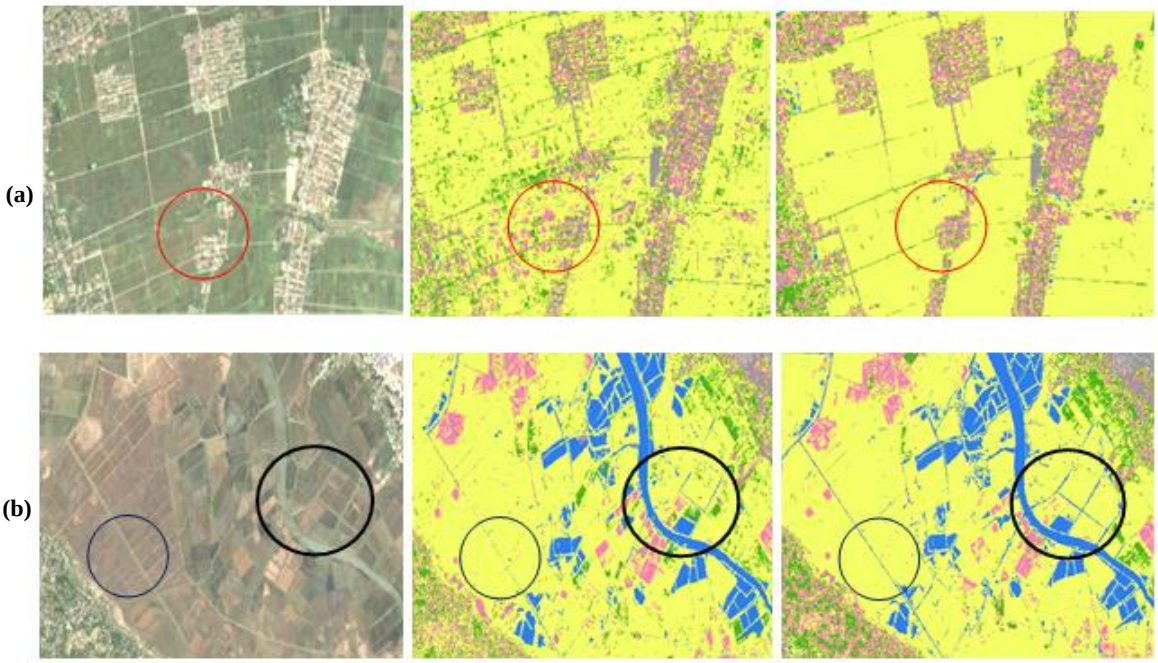
| Trường hợp phân loại RF / Lớp phủ | RF 4 kênh ảnh | | RF 4 kênh ảnh và 3 kênh chỉ số | |
|---|---|---|---|---|
| | UA (%) | PA (%) | UA (%) | PA (%) |
| Đất xây dựng | 94,4 | 97,7 | 96,5 | 97,1 |
| Đất nông nghiệp | 89,0 | 89,8 | 92,5 | 90,7 |
| Đất trống | 95,5 | 86,5 | 95,7 | 90,5 |
| Mặt nước | 98,7 | 97,5 | 97,5 | 98,7 |
| Rừng | 92,5 | 93,3 | 94,5 | 98,1 |

### *3.4. Thảo luận*

Qua số liệu từ bảng 1 đến bảng 4, dễ dàng nhận thấy kết quả phân loại đều đạt mức cao (trên 90% với độ chính xác tổng thể OA và trên 0,9 với Kappa) khi sử dụng thuật toán Random Forest. Các lớp đối tượng về cơ bản có sự phân tách với các lớp còn lại. Trong đó, cả hai trường hợp đều cho thấy kết quả lớp mặt nước được đánh giá có mức độ phân loại đạt cao nhất, số pixel lẫn rất ít so với lớp đất nông nghiệp và không lẫn bất kỳ pixel nào với các lớp khác (Bảng 1 và Bảng 2), giá trị PA và UA gần như đạt mức tuyệt đối (Bảng 4). Điều này thể hiện rằng mẫu của lớp mặt nước thể hiện đúng đặc trưng riêng.

Các thứ tự mức chính xác giảm dần từ rừng, đất xây dựng đến đất nông nghiệp và đất trống.Về đất nông nghiệp, pixel lớp này bị nhận nhầm với pixel rừng do phản xạ phổ nên màu sắc đối tượng khá giống nhau; còn một phần bị nhầm với đất xây dựng vì đất này bao gồm cả đất dân cư, có thể xảy ra trường hợp nhận lớp thực vật trong khu dân cư thành đất nông nghiệp. Đối với vùng đất trống, giá trị PA ở hai trường hợp đều thấp hơn, pixel bị lẫn chủ yếu với đất nông nghiệp (Bảng 1, Bảng 2) do ảnh thu nhận trong thời kỳ mùa khô, một số ruộng khô trên ảnh có màu gần tương đồng với màu đất trống (cụ thể là các vùng đồi trọc), tuy nhiên mức độ lẫn không lớn, và các chỉ số đánh giá độ chính xác đều cao (trên 85%). Một số pixel đất trống lẫn với đất xây dựng trong trường hợp (1), nhưng trong trường hợp (2) thì không lẫn do có sử dụng thêm chỉ số NDBI. Tương tự với lớp Rừng, trong trường hợp (2) pixel rừng lẫn với đất nông nghiệp ít hơn trường hợp (1) nhờ vào việc kết hợp chỉ số NDVI.

Với hai hướng tiếp cận, kết quả tốt hơn khi phân loại với thuật toán Random Forest trong trường hợp có sử dụng thêm các chỉ số phổ ngoài các kênh ảnh có sẵn trên ảnh Sentinel-2. Cụ thể trong bảng 3, kết quả mô hình (1) RF 4 kênh ảnh có độ chính xác tổng thể OA là 93,7% và hệ số Kappa là 0,92, thấp hơn so với mô hình (2) (OA = 95,3%, Kappa = 0,94). Bảng 4 cho thấy các giá trị UA và PA trong trường hợp (2) cơ bản cao hơn trường hợp (1). Nghiên cứu tiến hành so sánh hai ảnh kết quả phân loại như trong hình 10.




**Hình 10.** (a, b) So sánh kết quả phân loại lớp phủ mặt đất: Cột 1 là ảnh chụp một phần ảnh Sentinel-2, cột 2 là kết quả phân loại RF 4 kênh ảnh (RF trường hợp 1), cột 3 là kết quả phân loại RF 4 kênh ảnh và 3 chỉ số phổ NDVI, NDWI, NDBI (RF trường hợp 2). Các vòng tròn màu đỏ làm nổi bật các vùng có sự khác biệt giữa kết quả phân loại RF trường hợp (1) và (2) liên quan đến dân cư-đất nông nghiệp. Các vòng tròn màu đen thể hiện sự khác biệt nhận biết lớp mặt nước giữa RF trường hợp (1) và (2).

Hình 10a mô tả một phần khu vực với đất nông nghiệp và đất xây dựng. Có thể nhận thấy rằng, mô hình RF với trường hợp 1 không phân loại tốt: đất nông nghiệp lẫn với màu xanh của rừng, pixel dân cư rải rác, lẫn vùng dân cư và đất nông nghiệp; trong khi trường hợp có bổ sung các chỉ số phổ cho kết quả phân loại rõ ràng (khu vực được khoanh hình tròn đỏ). Hình 10b thể hiện đối tượng mặt nước được phân loại khá rõ trong cả hai trường hợp. Tuy nhiên, tại khu vực vòng tròn đen, kết quả phân loại trong trường hợp sử dụng 4 kênh ảnh cho thấy một số vùng mặt nước, cụ thể là các kênh mương, bị nhầm lẫn với lớp đất nông nghiệp. Mặt khác, mô hình RF (2) sử dụng 4 kênh ảnh Sentinel-2 và 3 chỉ số phổ đã phân loại chính xác các lớp phủ này do được bổ sung chỉ số về mặt nước NDWI. Việc đánh giá trực quan cho thấy, sự nhầm lẫn phổ giữa các lớp phủ mặt đất có mức độ cao hơn trong mô hình RF 4 kênh phổ so với kết quả của mô hình RF trong trường hợp còn lại.

So sánh kết quả đạt được với nghiên cứu [25], mặc dù sử dụng dữ liệu ảnh khác nhau, nhưng kết quả phân loại lớp phủ tỉnh Quảng Bình đạt được độ chính xác cao khi sử dụng thuật toán Random Forest. Nghiên cứu [25] cho độ chính xác tổng thể 88,8% và hệ số kappa 0.85 trong khi RF (1) là 93,7% và 0,92, RF (2) là 95,3% và 0,94. Các kết quả của [25] của cũng khá tương đồng với bài báo này ở độ chính xác PA và UA: phân loại RF cho kết quả tốt nhất ở đối tượng mặt nước, rừng, còn đất xây dựng và đất nông nghiệp có chỉ số PA và UA thấp hơn.

Nhìn chung, trên cơ sở so sánh độ chính xác và so sánh trực quan hai hướng tiếp cận của thuật toán Random Forest trong bài toán phân loại (Hình 10), khẳng định rằng việc sử

dụng các chỉ số quang phổ giúp làm tăng thông tin và cải thiện kết quả phân loại. Trong khi đó, công cụ Google Colab đã hoàn thành nhiệm vụ chạy các tập lệnh với thời gian là 308 giây. Con số này cũng thể hiện lợi thế về thời gian tính toán của Google Colab [18], đồng thời cho thấy tiềm năng của nền tảng này trong phân loại lớp phủ mặt đất.

## 4. Kết luận

Nghiên cứu đã phân loại 5 nhóm lớp phủ mặt đất tỉnh Quảng Bình vào tháng 8 năm 2022 sử dụng ảnh Sentinel-2 và thuật toán Random Forest, đồng thời cũng đánh giá hiệu quả của việc kết hợp các băng tần khác nhau của ảnh Sentinel-2 với các chỉ số phổ đặc trưng. Với các điểm đào tạo (*training data*) trong nghiên cứu, kết quả thử nghiệm cho thấy rằng mô hình RF (2) với 4 kênh phổ (các kênh lam, xanh lục, đỏ và cận hồng ngoại) kết hợp với các chỉ số phổ NDVI, NDWI, NDBI đã đạt được độ chính xác phân loại tổng thể là 95,3% (kappa = 0,94), tốt hơn so với mô hình RF 4 kênh phổ. Nhìn chung, với hướng đi mới này trong phân loại lớp phủ mặt đất, cần có thêm các nghiên cứu toàn diện hơn như sử dụng thêm các ảnh độ phân giải cao, các chỉ số phổ, hoặc kết hợp các phương pháp, mô hình khác nhau.

Bên cạnh việc khẳng định tính khả thi của thuật toán Random Forest trong bài toán phân loại, nghiên cứu cũng đánh giá hiệu quả xử lý của nền tảng Google Colab với thời gian nhanh chóng. Đây là một công cụ thể hiện sự vượt trội trong lĩnh vực viễn thám, khi người dùng có thể bắt đầu mã hóa các mô hình khoa học sử dụng ngôn ngữ lập trình Python thông qua các trình duyệt.

Nghiên cứu đã thực hiện đầy đủ theo mục tiêu đề ra, tuy nhiên còn có một số hạn chế như sau: Google Colab miễn phí nhưng bị giới hạn về thời gian, bị ngắt kết nối khi không có tương tác của người dùng, phụ thuộc vào Internet và Google Drive. Vì vậy, việc sử dụng một phiên bản khác như Google Colab Pro sẽ mang lại hiệu quả hơn trong tương lai.

**Đóng góp của tác giả:** Xây dựng ý tưởng nghiên cứu: P.T.T.H.; Thu thập dữ liệu: L.T.N., N.M.H.; Xử lý số liệu: V.N.Q., Đ.T.N.P.; Thực hiện lập trình: P.T.T.H.; Viết bản thảo bài báo: P.T.T.H., V.N.Q.; Chỉnh sửa bài báo: P.T.T.H., N.M.H.

**Lời cảm ơn:** Bài báo hoàn thành nhờ vào kết quả của đề tài nghiên cứu khoa học cấp cơ sở mã số T23 - 41 được hỗ trợ kinh phí từ Trường Đại học Mỏ Địa chất, Hà Nội: "Nghiên cứu ứng dụng phương pháp Machine Learning với thuật toán Random Forest trong thành lập bản đồ lớp phủ bề mặt".

**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. Comber, A.; Fisher, P.; Wadsworth, R. What is Land Cover? *Environ. Plann. B:: Plann. Des.* **2005**, *32*, 199–209.
2. Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **1997**, *62(1)*, 77–89.
3. Aplin, P. Remote sensing: Land cover. *Prog. Phys. Geogr.* **2004**, *28*, 283–293.
4. Liu, J.; Zhuang, D.; Luo, D.; Xiao, X. Land-cover classification of China: Integrated analysis of AVHRR imagery and geophysical data. *Int. J. Remote Sens.* **2003**, *24*, 2485–2500.
5. Peng, X.; Wang, J.; Raed, M.; Gari, J. Land cover mapping from RADARSAT stereo images in a mountainous area of southern Argentina. *Can. J. Remote Sens.* **2003**, *29(1)*, 75–87.

6. Makinde, E.O.; Oyelade, E.O. Land cover mapping using Sentinel-1 SAR and Landsat 8 imageries of Lagos State for 2017. *Environ. Sci. Pollut. Res.* **2020**, *27(1)*, 66–74.
7. Song, K.; Wang, Z.; Liu, Q.; Liu, D.; Ermoshin, V.V.; Ganzei, S.S.; Zhang, B.; Ren, C.; Zeng, L.; Du, J. Land use/land cover (LULC) classification with MODIS time series data and validation in the Amur River Basin. *Environ. Sci. Pollut. Res.* **2011**, *32(1)*, 9–15.
8. Sen, J.; Mehtab, S.; Sen, R.; Dutta, A.; Kherwa, P.; Ahmed, S.; Berry, P.; Khurana, S.; Singh, S.; Cadotte, D.; Anderson, D.; Ost, K.; Akinbo, R.; Daramola, O.; Lainjo, B. Machine Learning: Algorithms, Models, and Applications. IntechOpen Series Artificial Intelligence, 2022, 7, pp. 133.
9. Bansal, R.; Singh, J.; Kaur, R. Machine learning and its applications: A Review. *J. Appl. Sci. Comput.* **2019**, *6(6)*, 1392–1398.
10. Yuh, Y.G.; Tracz, W.; Matthews, H.D.; Turner, S.E. Application of machine learning approaches for land cover monitoring in northern Cameroon. *Ecol. Inf.* **2023**, *74*, 101955.
11. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104.
12. Tokar, O.; Olena, V.; Lubov, K.; Havryliuk, S.; Korol, M. Using the Random Forest Classification for Land Cover Interpretation of Landsat Images in the Prykarpattya Region of Ukraine. Proceeding of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT) 2018, pp. 241–244.
13. Kulkarni, A.D.; Lowe, B. Random Forest Algorithm for Land Cover Classification. *Int. J. Recent Innovation Trends Comput. Commun.* **2016**, *4(3)*, 58–63.
14. Tran, V.A.; Le, M.H.; Tran, H.H.; Le, T.N.; Tran, T.A.; Nguyen, C.C.; Ha, T.K. Land cover mapping in Camau province by machine learning algorithms using Sentinel-2 imagery. Proceeding of the 43th Asian Conference on Remote Sensing (ACRS 2022), 2022.
15. Pessoa, T.; Medeiros, R.; Nepomuceno, T.; Bian, G.B.; Albuquerque, V.H.C.; Filho, P.P. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *IEEE Access.* **2018**, *6*, 61677–61685.
16. Bisong, E. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Apress Publishers, 2019, pp. 709.
17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45(1)*, 5–32.
18. Hải, P.M.; Quang, N.N. Nghiên cứu thử nghiệm kết hợp môi trường làm việc Google Colaboratory và phương pháp học máy (Machine learning) trong phân loại ảnh viễn thám. *Tạp chí Khoa học Đo đạc và Bản đồ* **2020**, *43*, 13–17.
19. Jin, Y.; Liu, X.; Chen, Y.; Liang, X. Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: a case study of central Shandong. *Int. J. Remote Sens.* **2018**, *39(23)*, 8703–8723.
20. Tassi, A.; Gigante, D.; Modica, G.; Di Martino, L.; Vizzari, M. Pixel- vs. Object-Based Landsat 8 Data Classification in Google Earth Engine Using Random Forest: The Case Study of Maiella National Park. *Remote Sens.* **2021**, *13(12)*, 2299. Doi: 10.3390/rs13122299.
21. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8(2)*, 127–150.
22. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Remote Sens. Environ.* **1996**, *17(7)*, 1425–1432.

23. Zha, Y.; Gao, J.; Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Remote Sens.* **2003**, *24(3)*, 583–594.
24. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
25. Phong, D.H.; Huệ, N. Giám sát và kiểm kê phát thải khí nhà kính ($CO_2$ tương đương) trên cơ sở phân loại lớp phủ bằng ảnh Sentinel 1 tỉnh Quảng Bình. *Tạp chí Khí tượng Thủy văn* **2022**, *735*, 63–73.

# Research on the potential of applying Random Forest algorithm and Sentinel-2 to classify land cover in Quang Binh province on the Google Colab platform

**Pham Thi Thanh Hoa[1,2]*, Vu Ngoc Quang[3], Le Thanh Nghi[1,2], Doan Thi Nam Phuong[1,2], Nguyen Minh Hai[1]**

[1] Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology; phamthithanhhoa@humg.edu.vn; lethanhnghi@humg.edu.vn; doanthinamphuong@humg.edu.vn, nguyenminhhai@humg.edu.vn
[2] Geomatics in Earth Sciences Research Group, Hanoi University of Mining and Geology; phamthithanhhoa@humg.edu.vn; lethanhnghi@humg.edu.vn; doanthinamphuong@humg.edu.vn
[3] Engineering Faculty, University of Transport Technology; quangvn@utt.edu.vn

**Abstract:** In the era of technology, Machine learning has gradually replaced traditional methods in the field of remote sensing. One of the supervised machine learning algorithms which has high accuracy in land cover classification is Random Forest. Besides, instead of classifying images by commercial software, the cloud platform - Google Colab helps optimize image processing with a massive variety of libraries and is especially appropriate for machine learning methods. Thus, the article classified land cover using the Random Forest algorithm on the Google Colab platform, a case study in Quang Binh province in August 2022. Sentinel-2 image was chosen since it has a spatial resolution higher than other free images. At the same time, a comparison was performed in two cases: (1) using four image bands with 10m spatial resolution, (2) using four image bands with 10m spatial resolution and NDVI, NDWI, NDBI indices. The results with two approaches have an overall accuracy higher than 90% and Kappa above 0,9, showing the feasibility of Random Forest algorithm. Regarding accuracy, the second case has better results, confirming that the use of spectral indices helps increase information and improve classification results.

**Keywords:** Random Forest; Sentinel-2; Landcover; Google Colab.