

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



BÁO CÁO HỌC PHẦN TRÍ TUỆ NHÂN TẠO

**TỪ DỮ LIỆU VỀ TINH SENTINEL 2 TÍNH BẢN ĐỒ
PHÂN LOẠI CÁC LỚP PHỦ SỬ DỤNG PHƯƠNG PHÁP
HỌC MÁY**

Nhóm sinh viên: Trần Duy Tuấn Anh - 22028228
Hoàng Đức Dương - 22028259
Ngô Lê Hoàng - 22028042
Nguyễn Đức Khánh - 22028196

Giảng viên: PGS. TS. Nguyễn Thị Nhật Thanh

HÀ NỘI - 2025

Tóm tắt

Trong bối cảnh biến đổi môi trường toàn cầu và các hoạt động của con người đang tác động sâu rộng đến hệ sinh thái, việc giám sát và đánh giá chính xác sự thay đổi lớp phủ bề mặt Trái Đất là then chốt trong quản lý tài nguyên thiên nhiên bền vững và ứng phó biến đổi khí hậu. Các phương pháp truyền thống như điều tra thực địa hoặc phân tích ảnh hàng không đòi hỏi nguồn lực đáng kể và hạn chế khả năng giám sát liên tục cùng phạm vi không gian. Sự ra đời của vệ tinh Sentinel-2 đã tạo bước ngoặt quan trọng, cung cấp dữ liệu ảnh đa phổ độ phân giải cao, chu kỳ quay lại ngắn và hoàn toàn miễn phí.

Mục tiêu của đề tài là phát triển một mô hình phân loại lớp phủ bề mặt Trái Đất với độ chính xác cao, dựa trên dữ liệu Sentinel-2. Bắt đầu từ việc trích xuất dữ liệu thông qua bộ ảnh vệ tinh Sentinel-2 Level-2A (COPERNICUS/S2_HARMONIZED) chưa qua hiệu chỉnh khí quyển cho tỉnh Thanh Hóa. Sau đó là quá trình khai phá dữ liệu bao gồm phân tích đơn biến (trực quan hóa phân bố, xử lý lệch, nhận diện ngoại lai) và phân tích đa biến (tương quan giữa các đặc trưng, giảm chiều t-SNE để khám phá cấu trúc dữ liệu). Ngoài ra, bổ sung các chỉ số quang phổ như NDVI, NDWI, NDMI, NDBI, BSI, SAVI để làm nổi bật các đặc tính vật lý của đối tượng. Cuối cùng là chọn lựa đặc trưng thông qua việc loại bỏ các đặc trưng không phù hợp, đánh giá độ quan trọng bằng SHAP, và tìm số lượng đặc trưng tối ưu dùng RFE.

Nghiên cứu thử nghiệm đa dạng các mô hình học máy, từ truyền thống như Support Vector Machine (SVM) và K-Nearest Neighbors (KNN), các mô hình dựa trên cây (Tree-based Model) như Random Forest và XGBoost, đến các kỹ thuật học máy tổ hợp (Ensemble Learning) như Voting Classifier và Stacking Ensemble, cũng như mô hình Deep Learning (MLP Classifier, TabPFN). Tối ưu hóa siêu tham số được thực hiện bằng RandomizedSearchCV và Optuna để cải thiện độ chính xác, giảm hiện tượng quá khớp (overfit) và tăng tốc độ huấn luyện. Trong đó, việc sử dụng kỹ thuật bình chọn (voting) kết hợp tinh chỉnh siêu tham số đem lại kết quả tốt nhất với độ chính xác lên tới (0.882) đạt các chỉ số khác ở mức tốt như Precision (0.878), Recall (0.882) và điểm F1 (0.875). Cuối cùng, báo cáo phác thảo quy trình tạo bản đồ dự đoán lớp phủ cho tỉnh Thanh Hóa để trực quan hóa kết quả đạt được từ mô hình.

Tên	Task	% đóng góp
Ngô Lê Hoàng	Thu thập dữ liệu, Phân tích dữ liệu, Thực hiện các chạy thử nghiệm ban đầu, Xây dựng các kịch bản thực nghiệm, Thủ nghiệm 1: Quan sát hiệu suất mô hình phân loại dựa trên khoảng cách trước/sau chuẩn hoá dữ liệu, Thủ nghiệm 2: Hiệu suất mô hình với tham số cơ bản và tối ưu, Thủ nghiệm 5: Quan sát hiệu suất trước/sau khi tăng cường đặc trưng, Thủ nghiệm 6: Kỹ thuật lựa chọn đặc trưng, Viết báo cáo	25%
Hoàng Đức Dương	Thu thập dữ liệu, Thực hiện các chạy thử nghiệm ban đầu, Thủ nghiệm 2: Hiệu suất mô hình với tham số cơ bản và tối ưu, Thủ nghiệm 4: Sử dụng mô hình Voting Classifier, Thủ nghiệm 7: Thủ nghiệm các mô hình học sâu (MLP, TabPFN), Viết báo cáo, Tạo bản đồ dự đoán	25%
Nguyễn Đức Khánh	Thu thập dữ liệu, Phân tích dữ liệu, Thực hiện các chạy thử nghiệm ban đầu, Thủ nghiệm 3: Sử dụng kỹ thuật Stacking, Thủ nghiệm 5: Quan sát hiệu suất trước/sau khi tăng cường đặc trưng, Viết báo cáo, Tạo pipeline scale dữ liệu, SHAP	25%
Trần Duy Tuần Anh	Phân tích dữ liệu, Thực hiện các chạy thử nghiệm ban đầu, Xây dựng các kịch bản thực nghiệm, Thủ nghiệm 6: Kỹ thuật lựa chọn đặc trưng, Thủ nghiệm 7: Thủ nghiệm các mô hình học sâu (MLP, TabPFN), Viết báo cáo, SHAP	25%

Mục lục

Tóm tắt	1
Chương 1 Giới thiệu	7
1.1 Giới thiệu bài toán	7
1.2 Mục tiêu của đề tài	8
Chương 2 Dữ liệu	9
2.1 Thu thập dữ liệu	9
2.2 Khai phá dữ liệu	13
2.2.1 Phân tích đơn biến	14
2.2.2 Phân tích đa biến	16
2.2.3 Tăng cường đặc trưng	20
2.3 Chọn lựa đặc trưng	22
2.3.1 Loại bỏ đặc trưng không phù hợp	23
2.3.2 Dánh giá độ quan trọng đặc trưng bằng SHAP	23
2.3.3 Tìm số lượng đặc trưng tối ưu bằng RFE	24
Chương 3 Mô hình	26
3.1 Mô hình hóa	26
3.1.1 Support Vector Machine (SVM)	26
3.1.2 K-Nearest Neighbors (KNN)	27
3.1.3 Tree-based Model	27
3.1.4 Voting Classifier	29
3.1.5 Stacking Ensemble	29
3.1.6 Mô hình học sâu	30
3.2 Tối ưu hóa siêu tham số	32

Chương 4 Thực nghiệm và đánh giá	33
4.1 Phương pháp thực nghiệm	33
4.1.1 Chiến thuật thực nghiệm	33
4.1.2 Chỉ số đánh giá	35
4.2 Kết quả thực nghiệm	36
4.2.1 Thực nghiệm 1: Quan sát hiệu suất các mô hình phân loại dựa trên khoảng cách (distance-based model) trước và sau khi chuẩn hoá dữ liệu	36
4.2.2 Thực nghiệm 2: Quan sát hiệu suất các mô hình với tham số cơ bản và tham số tối ưu	36
4.2.3 Thực nghiệm 3: Sử dụng kĩ thuật Stacking	37
4.2.4 Thực nghiệm 4: Sử dụng mô hình Voting Classifier	38
4.2.5 Thực nghiệm 5: Quan sát hiệu suất các mô hình trước và sau khi tăng cường đặc trưng	38
4.2.6 Thực nghiệm 6: Thủ nghiệm kĩ thuật lựa chọn đặc trưng (Feature Selection)	39
4.2.7 Thực nghiệm 7: Thủ nghiệm các mô hình học sâu: MLP và TabPFN	40
4.3 Tạo bản đồ dự đoán	42
Kết luận	45
Tài liệu tham khảo	46

Danh sách bảng

2.1	Các dải phổ của Sentinel-2 và ứng dụng	9
2.2	Bảng các chỉ số thực vật và ứng dụng	20
4.1	So sánh hiệu suất các mô hình trước và sau khi chuẩn hoá dữ liệu	36
4.2	So sánh hiệu suất các mô hình trước và sau khi tối ưu tham số	37
4.3	So sánh hiệu suất của mô hình tốt nhất trước đó với kĩ thuật Stacking .	38
4.4	So sánh hiệu suất của mô hình tốt nhất trước đó với Voting Classifier .	39
4.5	So sánh hiệu suất các mô hình trước và sau khi tăng cường đặc trưng .	40
4.6	So sánh hiệu suất các mô hình trước và sau khi áp dụng kĩ thuật lựa chọn đặc trưng	41
4.7	Sử dụng các mô hình học sâu	42

Danh sách hình vẽ

2.1	Quy trình thu thập dữ liệu	13
2.2	Hiện tượng phân phối lệch phải ở phô 3 và phô 4	15
2.3	Boxplot của phô B4 ứng với các nhãn	16
2.4	Biểu đồ nhiệt tương quan	18
2.5	Kết quả giảm chiều t-SNE	19
2.6	Ví dụ kết quả sử dụng SHAP với mô hình LightGBM	25
4.1	Hiệu suất mô hình Random Forest Classifier qua từng giai đoạn RFE . .	41
4.2	Ảnh khu vực thị xã Nghi Sơn	43
4.3	Ảnh khu vực thị xã Nghi Sơn sau khi áp dụng mô hình phân loại lớp phủ đất	43
4.4	Ảnh phân loại bằng XGBoost sử dụng dataset phía tây tỉnh Thanh Hóa	44
4.5	Ảnh phân loại bằng Multi-Layer Perceptron sử dụng dataset tỉnh Thanh Hóa	44

Chương 1

Giới thiệu

1.1 Giới thiệu bài toán

Sự biến đổi môi trường toàn cầu đang diễn ra với tốc độ chưa từng có trong lịch sử, trong đó các hoạt động như đô thị hóa, khai thác rừng và mở rộng đất nông nghiệp đang tạo ra những tác động sâu rộng đến các hệ sinh thái trên quy mô cảnh quan. Các biến đổi này không chỉ làm thay đổi cấu trúc và chức năng của hệ sinh thái mà còn ảnh hưởng trực tiếp đến các dịch vụ hệ sinh thái thiết yếu cho sự sống còn và phát triển của con người. Trong bối cảnh đó, việc giám sát và đánh giá chính xác sự thay đổi lớp phủ bề mặt Trái Đất đóng vai trò then chốt trong việc xây dựng các chính sách quản lý tài nguyên thiên nhiên bền vững và chiến lược ứng phó với biến đổi khí hậu.

Phân loại lớp phủ bề mặt Trái Đất là một trong những công cụ quan trọng nhất trong việc theo dõi những biến đổi này. Truyền thống, quá trình phân loại lớp phủ đã dựa chủ yếu vào các phương pháp điều tra thực địa hoặc phân tích ảnh hàng không, những phương pháp đòi hỏi nguồn lực đáng kể về thời gian, nhân lực và tài chính. Hạn chế này không chỉ làm giảm khả năng giám sát liên tục mà còn hạn chế phạm vi không gian và độ chính xác của các nghiên cứu. Tuy nhiên, với sự phát triển vượt bậc trong công nghệ quan sát Trái Đất trong thập kỷ gần đây đã mang lại những đột phá quan trọng trong lĩnh vực này. Đặc biệt, sự ra đời của vệ tinh Sentinel-2 do Cơ quan Vũ trụ Châu Âu (ESA) phát triển trong khuôn khổ chương trình Copernicus

đã tạo ra một bước ngoặt quan trọng. Hệ thống này cung cấp dữ liệu ảnh đa phổ (multi-spectral) với độ phân giải không gian cao (từ 10m đến 60m), chu kỳ quay lại ngắn, và đặc biệt quan trọng là hoàn toàn miễn phí cho cộng đồng nghiên cứu toàn cầu.

1.2 Mục tiêu của đề tài

Nghiên cứu này tập trung vào việc phát triển một phương pháp phân loại lớp phủ bề mặt Trái Đất với độ chính xác cao, dựa trên nguồn dữ liệu ảnh vệ tinh đa phổ do hệ thống Sentinel-2 cung cấp. Mục tiêu chính là xây dựng một mô hình phân tích mạnh mẽ, có khả năng khai thác hiệu quả các đặc trưng phổ học của ảnh vệ tinh để nhận diện và phân loại chính xác các kiểu lớp phủ bề mặt khác nhau trên quy mô rộng lớn, trước hết hướng tới ứng dụng thực tiễn tại tỉnh Thanh Hóa. Cụ thể, nhóm nghiên cứu tập trung vào ba vấn đề cốt lõi trong quá trình xây dựng và tối ưu hóa mô hình như sau:

Thứ nhất, đề tài đặt mục tiêu xây dựng một mô hình phân loại đa lớp, có khả năng tự động nhận diện và phân định các điểm ảnh thành mười loại hình lớp phủ bề mặt phổ biến trên địa bàn tỉnh. Các lớp phủ này bao gồm: khu vực dân cư, vùng trồng lúa, đất nông nghiệp khác, đồng cỏ, đất trống, vùng cây bụi, khu vực rừng, đất ngập nước, mặt nước tự nhiên và vùng nuôi trồng thủy sản. Việc xác định chính xác các loại lớp phủ này đóng vai trò quan trọng trong việc hỗ trợ công tác quản lý, giám sát tài nguyên và quy hoạch sử dụng đất tại địa phương.

Thứ hai, đề tài chú trọng vào việc lựa chọn, tối ưu hóa và tinh chỉnh các thuật toán học máy hiện đại, nhằm nâng cao hiệu quả và độ tin cậy của mô hình phân loại. Quá trình này bao gồm thử nghiệm nhiều cấu hình mô hình, lựa chọn siêu tham số phù hợp, cũng như áp dụng các chiến lược huấn luyện tối ưu để thích ứng tốt với điều kiện dữ liệu thực tế của tỉnh Thanh Hóa.

Thứ ba, để nâng cao tính tổng quát và khả năng mở rộng của mô hình trên nhiều khu vực địa lý khác nhau, nhóm triển khai các kỹ thuật tiền xử lý dữ liệu, chọn và chuẩn hóa đặc trưng tốt nhất trong khả năng. Những kỹ thuật này giúp giảm thiểu ảnh hưởng của nhiễu trong dữ liệu và tăng tính ổn định của mô hình khi áp dụng cho các vùng có đặc trưng sinh thái, thổ nhưỡng và điều kiện địa hình đa dạng.

Chương 2

Dữ liệu

2.1 Thu thập dữ liệu

Sentinel-2 [6] là một hệ thống vệ tinh quan sát Trái Đất thuộc chương trình Copernicus của Liên minh Châu Âu, được phát triển bởi Cơ quan Vũ trụ Châu Âu (ESA). Hệ thống bao gồm hai vệ tinh Sentinel-2A và Sentinel-2B, hoạt động phối hợp để thu nhận dữ liệu hình ảnh quang học với độ phân giải không gian cao, bao phủ toàn bộ bề mặt đất liền của Trái Đất mỗi 5 ngày. Cảm biến chính trên vệ tinh là thiết bị đa phổ (Multispectral Instrument - MSI) với 13 dải phổ có ứng dụng như sau:

Bảng 2.1: Các dải phổ của Sentinel-2 và ứng dụng

Tên dải phổ	Bước sóng (nm)	Độ phân giải không gian (m)	Ứng dụng
B1 - Coastal aerosol	443	60	Giảm nhiễu khí quyển, hỗ trợ phân tích vùng ven biển, theo dõi chất lượng nước.
Tiếp tục ở trang sau			

Bảng 2.1 – Tiếp theo

Tên dải phổ	Bước sóng (nm)	Độ phân giải không gian (m)	Ứng dụng
B2 - Blue	490	10	Tạo ảnh RGB tự nhiên, phân biệt nước với các bề mặt khác, hỗ trợ phân tích chất lượng nước.
B3 - Green	560	10	Dánh giá tình trạng thực vật, sử dụng trong chỉ số NDWI để xác định vùng có nước.
B4 - Red	665	10	Hỗ trợ NDVI, nhạy với sự hấp thụ của diệp lục, giúp phân biệt thực vật với đất trống.
B5 - Red Edge 1	705	20	Phản ánh tình trạng thực vật, quan trọng trong phân tích sức khỏe cây trồng và phát hiện căng thẳng thực vật.
B6 - Red Edge 2	740	20	Hỗ trợ phân biệt thực vật và theo dõi quá trình quang hợp.
B7 - Red Edge 3	783	20	Tăng cường khả năng phát hiện thực vật, sử dụng trong phân tích nông nghiệp và lâm nghiệp.
B8 - NIR 1 (Near Infrared)	842	10	Dánh giá sức khỏe thực vật, dùng trong các chỉ số sinh thái như NDVI, SAVI.
Tiếp tục ở trang sau			

Bảng 2.1 – Tiếp theo

Tên dải phổ	Bước sóng (nm)	Độ phân giải không gian (m)	Ứng dụng
B8A - NIR 2	865	20	Bổ sung phân tích NIR, sử dụng trong NDMI để xác định độ ẩm thực vật.
B9 - Water vapor	945	60	Hấp thụ hơi nước trong khí quyển, giảm nhiễu khí quyển khi phân tích ảnh vệ tinh.
B10 - SWIR 1 (Short-wave Infrared)	1375	60	Nhạy với hơi nước trong khí quyển, giúp phát hiện và giảm ảnh hưởng của mây mỏng.
B11 - SWIR 2	1610	20	Xác định độ ẩm của đất và thực vật, sử dụng trong NDMI và BSI để phân biệt đất trống với thực vật.
B12 - SWIR 2	2190	20	Hỗ trợ phân biệt đất trống, đá và khu vực đô thị, dùng trong NDBI và BSI.

Trong Google Earth Engine Dataset, Sentinel-2 có 2 bộ dữ liệu là COPERNICUS/S2_SR_HARMONIZED và COPERNICUS/S2_HARMONIZED, trong đó:

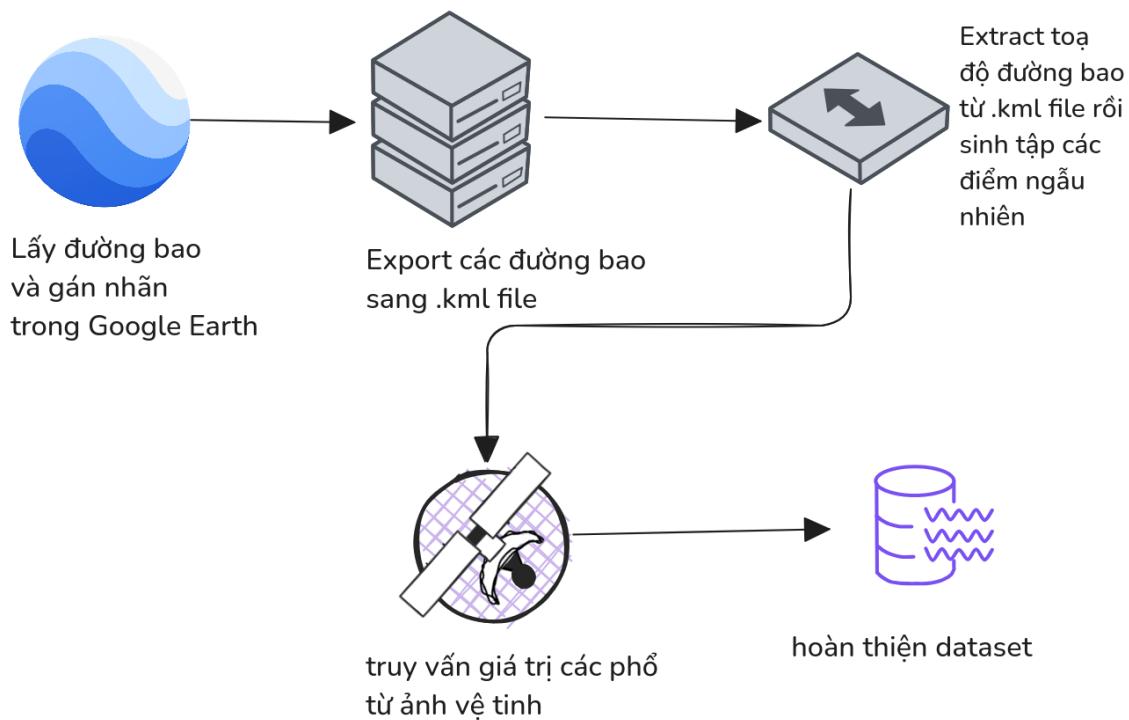
- **COPERNICUS/S2_SR_HARMONIZED:** Tập ảnh đã qua hiệu chỉnh khí quyển, **loại bỏ phổ 10 (B10)** do B10 được thiết kế để phát hiện mây mỏng ở tầng cao, không phù hợp với việc phân tích phản xạ bề mặt.
- **COPERNICUS/S2_HARMONIZED:** Tập ảnh chưa loại bỏ ảnh hưởng khí quyển, gồm mọi band (kể cả B10).

NNhóm lựa chọn sử dụng bộ dữ liệu COPERNICUS/S2_HARMONIZED đồng

thời chọn tỉnh **Thanh Hóa** làm khu vực nghiên cứu vì địa phương này có sự đa dạng cao về lớp phủ mặt đất. Sự phong phú này giúp nghiên cứu có dữ liệu bao quát nhiều loại địa hình, tạo điều kiện thuận lợi cho việc phân loại và đánh giá tính chính xác của mô hình.

Quy trình thu thập dữ liệu gồm 5 bước chính (Hình 2.1) được mô tả cụ thể như sau:

1. Trong giai đoạn đầu tiên, sử dụng Google Earth Pro vẽ các đường ranh giới (polygon boundary) để xác định các khu vực địa lý cụ thể. Mỗi vùng được đánh dấu gán một nhãn (label) tương ứng với loại lớp phủ bề mặt như rừng, sông hồ, đất trống, v.v.
2. Sau khi hoàn thành quá trình phân định và gán nhãn, các đa giác cùng với thông tin nhãn được xuất ra định dạng Keyhole Markup Language (.kml). Định dạng này được Google Earth hỗ trợ rộng rãi và có khả năng lưu trữ đồng thời cả thông tin không gian địa lý và thuộc tính mô tả.
3. Từ tệp .kml, trích xuất tọa độ chính xác của tất cả các đa giác. Mỗi đa giác bao gồm một tập hợp các điểm tọa độ trong hệ quy chiếu địa lý (latitude, longitude). Tiếp theo, tiến hành lấy mẫu ngẫu nhiên các điểm nằm trong phạm vi các đa giác này, tạo thành một tập dữ liệu điểm đại diện với nhãn tương ứng theo lớp phủ bề mặt.
4. Với tập điểm địa lý đã được xác định, truy vấn và trích xuất giá trị phổ từ tập dữ liệu ảnh vệ tinh COPERNICUS/S2_SR_HARMONIZED thông qua giao diện lập trình ứng dụng Python (Python SDK) của Google Earth Engine. Quá trình này cho phép lấy được các giá trị phổ đa kênh tại từng vị trí tọa độ đã được xác định.
5. Cuối cùng, kết quả trích xuất được tổ chức và xuất ra dưới dạng tệp giá trị phân cách bằng dấu phẩy (.csv). Tệp dữ liệu này bao gồm các thông tin về tọa độ địa lý, giá trị phổ đa kênh, và nhãn phân loại tương ứng cho mỗi điểm, tạo thành bộ dữ liệu hoàn chỉnh phục vụ cho quá trình huấn luyện và đánh giá mô hình phân loại.



Hình 2.1: Quy trình thu thập dữ liệu

2.2 Khai phá dữ liệu

Sau khi hoàn tất việc thu thập dữ liệu từ vệ tinh Sentinel-2, nhóm nghiên cứu tiến hành khai phá dữ liệu nhằm xây dựng tập dữ liệu dạng bảng có cấu trúc hoàn chỉnh và hữu ích hơn trong việc huấn luyện mô hình. Bộ dữ liệu thô ban đầu bao gồm các thông số tọa độ không gian (kinh độ, vĩ độ) kết hợp với 13 dải phổ từ B1 đến B12 và tương ứng với mười nhăm lớp phủ tối tỷ lệ khác nhau phản ánh đặc điểm địa lý của tỉnh Thanh Hóa. Với kích thước **11.400 mẫu cho tập huấn luyện và 2.850 mẫu cho tập kiểm tra**, nhóm đã triển khai đa dạng phương pháp phân tích nhằm tối ưu hóa hiệu suất mô hình và nâng cao hiểu biết về ý nghĩa vật lý của dữ liệu. Các kỹ thuật được áp dụng bao gồm phân tích tương quan đơn biến và đa biến để xác định mối liên hệ giữa các đặc trưng, tăng cường đặc trưng mới giàu thông tin và chọn lọc đặc trưng tối ưu cho mô hình. Với những phân tích này nhóm mong muốn không chỉ cải thiện độ chính xác của các mô hình chuyên dụng cho dữ liệu dạng bảng mà còn

cung cấp những hiểu biết về bản chất của dữ liệu viễn thám trong nghiên cứu.¹

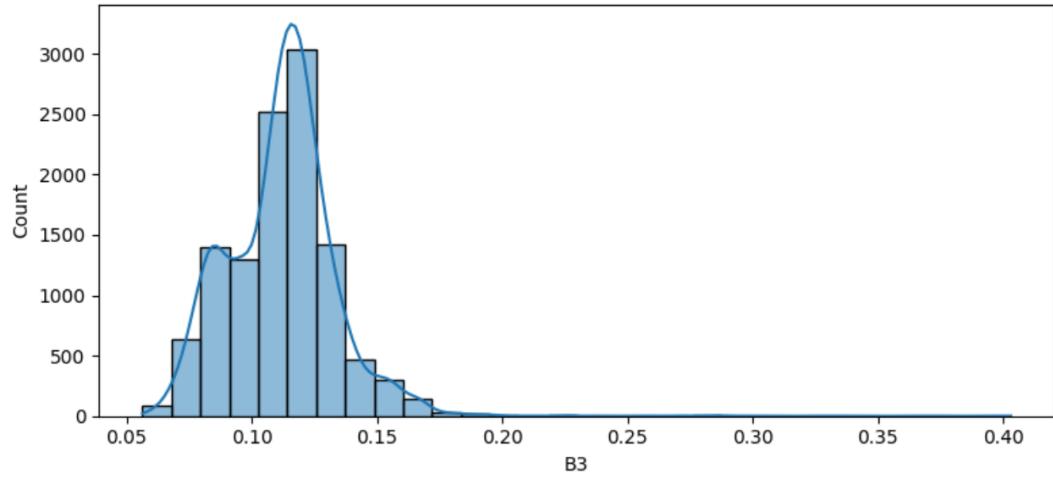
2.2.1 Phân tích đơn biến

Trước hết, nhóm trực quan hóa phân bố của dữ liệu nhằm hiểu rõ bản chất thống kê của từng đặc trưng phổ trước khi tiến hành xây dựng mô hình. Phương pháp này bao gồm việc **phân tích biểu đồ tần suất** (histogram) và các chỉ số thống kê mô tả như trung bình, độ lệch chuẩn, giá trị cực tiểu và cực đại cho mỗi dải phổ. Qua đó nhận thấy đa số các phổ đều có phân bố gần phân bố chuẩn, tuy nhiên có phổ thứ 10 xảy ra hiện tượng đa đỉnh giải thích cho việc này có thể đến từ việc dữ liệu phụ thuộc vào vị trí có mây hay không nếu giá trị này thấp đồng nghĩa với việc có nhiều mây. Bên cạnh đó các biểu đồ cho thấy tất cả các phổ đều xảy ra hiện tượng **phân phối lệch phải** (right skewed) - Phần lớn dữ liệu tập trung ở bên trái, nhưng có một số ít giá trị cực lớn kéo đuôi phân phối sang phải (Hình 2.2). Nhóm nghiên cứu đã áp dụng phép biến đổi logarithm (Log Transform) để khắc phục hiện tượng phân phối lệch phải, đồng thời sử dụng bộ chuẩn hóa Z-score (Standard Scaler) nhằm đưa các biến về phân phối gần với chuẩn tắc Gauss. Các bước tiền xử lý này được thực hiện trước khi huấn luyện các mô hình dựa trên khoảng cách (distance-based models) và các mô hình tuyến tính yêu cầu giả định về phân phối chuẩn của dữ liệu.

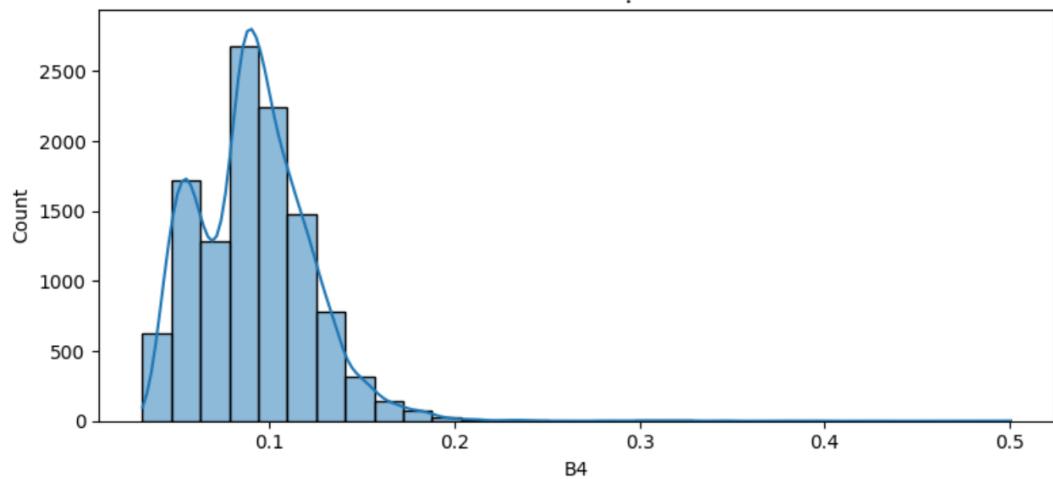
Bên cạnh đó, một điều đáng chú ý của dữ liệu vệ tinh này là sự xuất hiện của nhiều điểm ngoại lai (Outliers) đặc biệt tập trung ở nhãn đất thổ cư (Residential Land) trên tất cả các phổ (ví dụ như trong phổ B4 Hình 2.3), điều này có thể được giải thích bởi tính không đồng nhất trong cấu trúc không gian của khu vực dân cư, từ nông thôn đến thành thị, cùng với sự đa dạng trong vật liệu xây dựng có đặc tính phản xạ khác biệt như mái tôn, bê tông, gạch ngói hoặc các vật liệu truyền thống khác.Thêm vào đó, các hiệu ứng khí quyển phức tạp do hoạt động công nghiệp, giao thông và xây dựng tại các khu đô thị càng làm gia tăng biến động trong giá trị phổ thu nhận được. Việc xử lý các điểm ngoại lai này đặt ra thách thức đáng kể trong quá trình tiền xử lý dữ liệu, bởi loại bỏ hoàn toàn chúng có thể dẫn đến mất thông tin giá trị về tính đa dạng thực sự của lớp phủ đất thổ cư, làm giảm khả năng tổng quát hóa của mô hình trên các khu vực có đặc điểm tương tự trong tương lai, vì vậy nhóm

¹Xem chi tiết tại: Exploratory Data Analysis Notebook

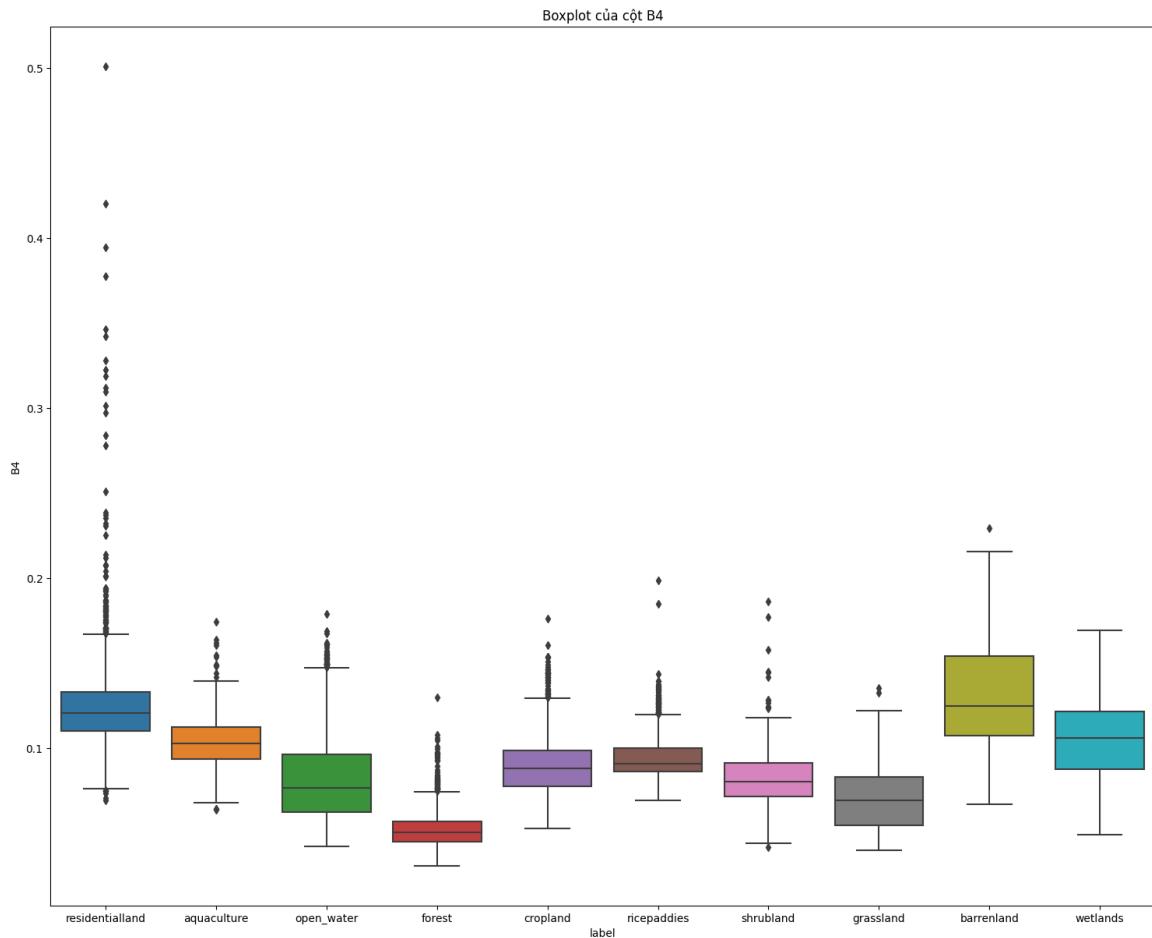
Phân bố của cột: B3



Phân bố của cột: B4



Hình 2.2: Hiện tượng phân phối lệch phải ở phẩ 3 và phẩ 4



Hình 2.3: Boxplot của phô B4 ứng với các nhãn

đã lựa chọn phương pháp Log Transform để giảm thiểu ảnh hưởng của các điểm này thay vì loại bỏ hoàn toàn. Ngoài ra, phân tích phân phối thống kê cũng cho thấy các trung vị (median) ứng với từng nhãn lớp phủ trong các phô 6, phô 11, và phô 12 có sự khác biệt rõ rệt, chứng tỏ tiềm năng của các dải phô này trong việc phân biệt hiệu quả các lớp phủ.

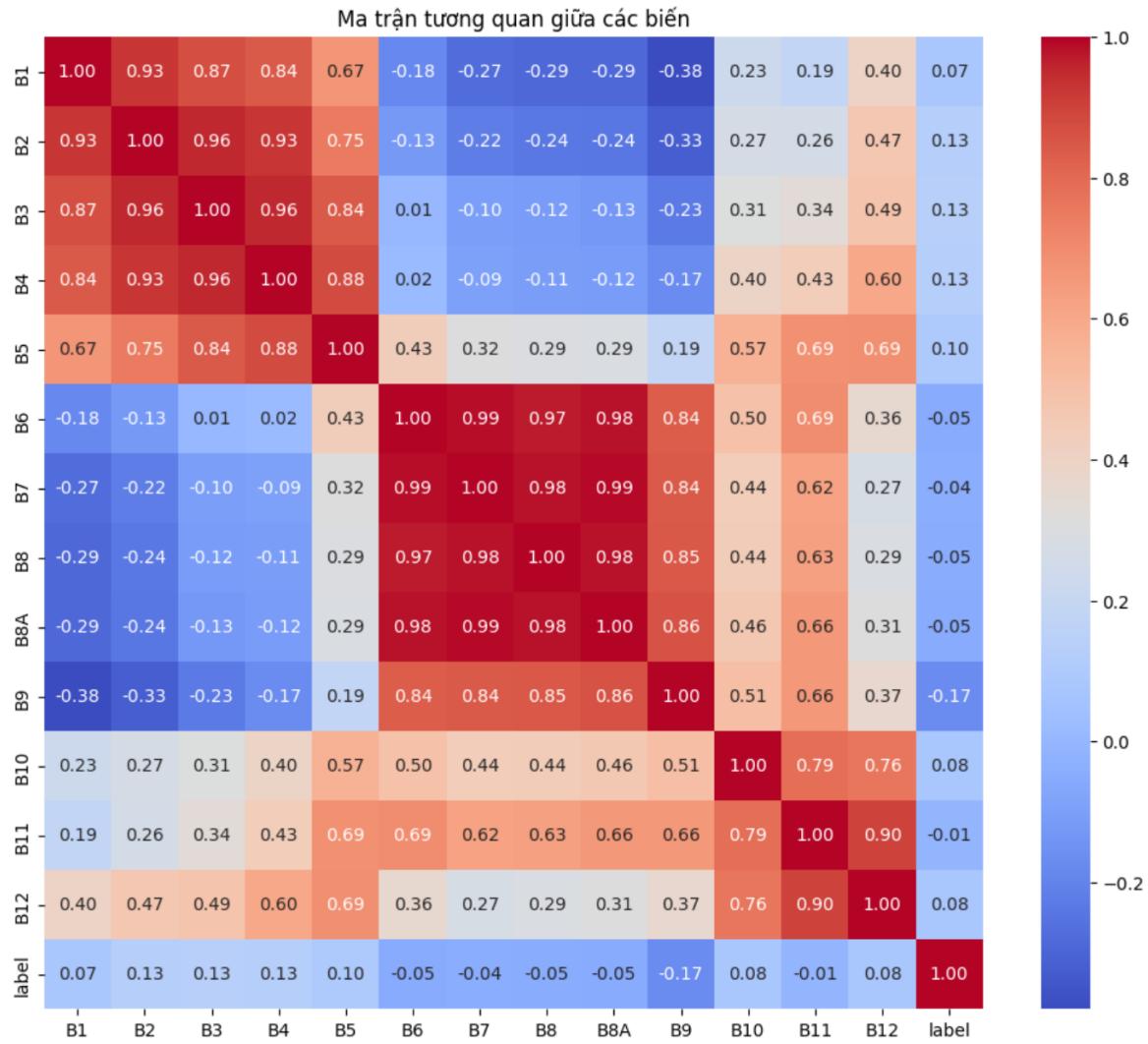
2.2.2 Phân tích đa biến

Bên cạnh việc tìm hiểu phân bố của từng phô, nhóm tiến hành phân tích đa biến nhằm xác định mối quan hệ giữa các đặc trưng quang phổ và khám phá cấu trúc dữ liệu tiềm ẩn. Qua đó, nhóm có thể hiểu rõ hơn về tính chất của dữ liệu và đưa ra các quyết định hợp lý trong quá trình tiền xử lý và lựa chọn mô hình. Các kỹ thuật phân

tích như biểu đồ nhiệt tương quan và giảm chiều t-SNE được áp dụng để đánh giá toàn diện bộ dữ liệu, từ đó làm cơ sở cho các bước phân tích và thiết kế mô hình tiếp theo. Phương pháp tiếp cận này góp phần làm cơ sở giúp tối ưu hóa việc lựa chọn đặc trưng và cung cấp cái nhìn về mối liên hệ giữa các kênh quang phổ trong dữ liệu ảnh vệ tinh Sentinel-2.

Trước tiên, nhóm sử dụng **biểu đồ nhiệt** (Heatmap) để tìm ra tương quan giữa các đặc trưng với nhau và với nhãn. Từ biểu đồ nhiệt (Hình 2.4) và kiến thức miền nhóm đưa ra những phân tích sau:

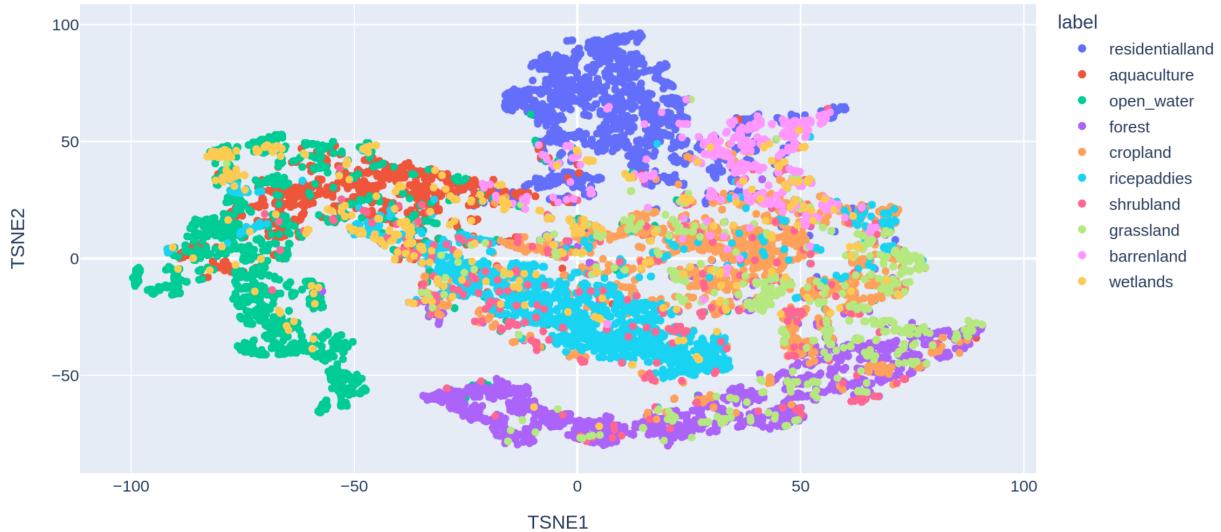
- Mặc dù các phổ B2, B3, và B4 có độ tương quan cao rất cao (lên tới 0.96) nhưng không thể loại bỏ vì vai trò vật lý của từng quang phổ là rất khác nhau trong việc phân biệt lớp phủ.
- Tương tự, các phổ B6 và B7 thể hiện mối tương quan cao về mặt dữ liệu (0.99). Tuy nhiên phổ B6 (vùng giữa Red Edge) cho thấy khả năng nhận diện tốt các biến đổi nhẹ trong thảm thực vật, trong khi phổ B7 (gần vùng cận hồng ngoại - Near Infrared) thể hiện độ nhạy cao đối với sự hiện diện và tình trạng của thảm thực vật phù hợp cho bài toán phải phân biệt lên tới 5 nhãn liên quan trực tiếp tới thực vật.
- Hai phổ B8 và B8A thể hiện mức độ tương quan cao (0.98) do cả hai đều thuộc vùng cận hồng ngoại và có khả năng cung cấp thông tin tương tự nhau trong phân tích thảm thực vật. Trong bài này, phổ B8A được sử dụng vì mặc dù nó cung cấp độ phân giải quang phổ cao hơn trong phạm vi hẹp, lợi thế này thực sự cần thiết trong các ứng dụng yêu cầu phân tích chi tiết về thảm thực vật.
- Dù phổ B11 và phổ B12 có mối tương quan cao (0.90) do cả hai đều thuộc vùng hồng ngoại trung, hơn nữa bài toán đặt ra cần phân biệt các nhãn liên quan tới đất trống và đất ngập nước. Phổ 11 giúp phân biệt đất ngập nước hoặc thực vật ẩm tốt hơn, trong khi đó phổ 12 lại hỗ trợ rất tốt trong việc phân loại đất trống nên nhóm quyết định giữ lại cả hai phổ này.
- Phần lớn các hệ số tương quan giữa các đặc trưng và nhãn cho thấy mức độ rất thấp, điều này chỉ ra rằng mối quan hệ tuyến tính giữa từng đặc trưng riêng lẻ



Hình 2.4: Biểu đồ nhiệt tương quan

và nhãn là rất yếu. Do đó, nên sử dụng thêm các đặc trưng phi tuyến và các mô hình có khả năng nắm bắt mối quan hệ phi tuyến và phức tạp, như các mô hình dựa trên cây quyết định (Random Forest, XGBoost).

Ngoài ra, nhóm sử dụng thêm **t-SNE** (t-distributed Stochastic Neighbor Embedding) là một kỹ thuật giảm chiều phi tuyến để khám phá cấu trúc phân bố dữ liệu trong không gian thấp hơn. Phương pháp này đặc biệt phù hợp cho việc trực quan hóa dữ liệu đa chiều vì t-SNE có khả năng bảo toàn cấu trúc cục bộ của dữ liệu, tức là các điểm dữ liệu gần nhau trong không gian đa chiều sẽ có xu hướng vẫn gần nhau trong không gian biểu diễn thấp hơn. Nhóm lựa chọn t-SNE thay vì các phương pháp giảm



Hình 2.5: Kết quả giảm chiều t-SNE

chiều tuyến tính như PCA (Principal Component Analysis) vì bộ dữ liệu có mối quan hệ phi tuyến khá nhiều giữa các đặc trưng quang phổ và nhãn lớp phủ bề mặt. Kết quả biểu diễn t-SNE (Hình 2.5) cho thấy:

- Lớp nước mở (open water) và khu dân cư (residential land) thể hiện sự phân tách rõ rệt nhất, tạo thành một nhánh riêng biệt trên biểu đồ, cho thấy đặc tính quang phổ khác biệt của nước so với các lớp phủ bề mặt khác khiến cho nhãn này dễ nhận biết hơn.
- Các lớp thảm thực vật như rừng (forest), đồng cỏ (grassland) và đất cây bụi (shrubland) có sự phân bố chồng lấp đáng kể dọc theo các đường cong, phản ánh sự tương đồng trong các đặc tính quang phổ của chúng, điều này có thể khiến mô hình nhầm lẫn trong quá trình phân loại.
- Lớp ruộng lúa (rice paddies) cũng cho thấy một sự phân bố đặc biệt, vừa gần với các lớp thảm thực vật khác nhưng cũng bị đan xen với lớp đất ngập nước (wetlands), phản ánh đặc tính “vừa là thảm thực vật vừa chứa nước” của loại hình canh tác này.

Kết quả phân tích t-SNE này góp phần cung cấp thêm việc giữ lại tất cả các kênh quang phổ trong quá trình phân tích là cần thiết để nắm bắt đầy đủ sự biến đổi tinh

té giữa các lớp phủ bề mặt. Đồng thời cũng gợi ý sử dụng các mô hình phi tuyến phức tạp có khả năng học được các ranh giới quyết định phức tạp sẽ phù hợp hơn cho bài toán này so với các mô hình tuyến tính hoặc đơn giản.

2.2.3 Tăng cường đặc trưng

Dựa vào các nghiên cứu của các học giả về lĩnh vực này, nhóm bổ sung thêm các chỉ số quang phổ để nâng cao khả năng phân biệt giữa các lớp mặt đất. Các chỉ số này được tính toán từ tỷ lệ giữa các phổ khác nhau, cho phép làm nổi bật những đặc vật lý cụ thể của đối tượng. Cụ thể, nhóm đã tích hợp các chỉ số có trong Bảng 2.2

Bảng 2.2: Bảng các chỉ số thực vật và ứng dụng

Chỉ số	Công thức	Mục đích
NDVI (Normalized Difference Vegetation Index)[8]	$NDVI = \frac{B8 - B4}{B8 + B4}$	Xác định mật độ thảm thực vật xanh
NDWI (Normalized Difference Water Index) [4]	$NDWI = \frac{B3 - B8}{B3 + B8}$	Xác định vùng có nước
NDMI (Normalized Difference Moisture Index)[3]	$NDMI = \frac{B8A - B11}{B8A + B11}$	Đánh giá độ ẩm thảm thực vật

Tiếp tục ở trang sau

Bảng 2.2 – Tiếp theo

Chỉ số	Công thức	Mục đích
NDBI (Normalized Difference Built-up Index)[9]	$NDBI = \frac{B11 - B8}{B11 + B8}$	Xác định khu vực đô thị hóa
BSI (Bare Soil Index)[5]	$BSI = \frac{(B12 + B8) - (B4 + B3)}{(B12 + B8) + (B4 + B3)}$	Xác định bề mặt đất trống
SAVI (Soil-Adjusted Vegetation Index)[2]	$SAVI = \frac{B8 - B4}{B8 + B4 + L} \times (1 + L), \quad L = 0.5$	Xác định mật độ thảm thực vật, giảm ảnh hưởng của đất nền

Giá trị **NDVI** nằm trong khoảng từ -1 đến 1, với các giá trị cao hơn (thường lớn hơn 0.5) thể hiện thảm thực vật dày và khỏe mạnh, các giá trị thấp hơn (khoảng từ 0.2 đến 0.5) thể hiện thảm thực vật thừa hoặc cây bụi, và các giá trị gần 0 hoặc âm thường chỉ báo các khu vực không có thực vật như nước, đất trống, hoặc khu vực xây dựng.

Các vùng nước thường có giá trị **NDWI** dương, trong khi đất và thực vật thường có giá trị âm. Điều này tạo ra sự phân biệt rõ ràng giữa các thủy vực và các lớp phủ khác, đặc biệt hữu ích trong việc xác định vùng ngập nước, đất ngập nước, và ranh giới giữa đất liền và nước.

Giá trị **NDMI** nằm trong khoảng từ -1 đến 1. Các giá trị cao hơn (thường lớn hơn 0.3) cho thấy thảm thực vật có độ ẩm cao, như rừng xanh hoặc cây trồng phát triển tốt. Ngược lại, các giá trị thấp hơn hoặc âm thường biểu thị thảm thực vật bị khô hạn, cây trồng gặp cǎng thẳng do thiếu nước hoặc khu vực đất trống, không có thảm phủ thực vật.

NDBI cũng có giá trị trong khoảng từ -1 đến 1. Các giá trị dương (thường lớn hơn 0.2) thường biểu thị các khu vực xây dựng như nhà cửa, đường xá hoặc cơ sở hạ tầng. Trong khi đó, các giá trị gần 0 hoặc âm cho thấy những vùng ít chịu tác động của đô thị hóa, có thể là đất nông nghiệp, thảm thực vật hoặc mặt nước.

Chỉ số **BSI** nhằm mục đích nhận diện các khu vực có bề mặt đất trống, không bị che phủ bởi thực vật. Chỉ số này được xây dựng từ các kênh phổ nhạy cảm với đất và thực vật. Giá trị BSI cao (thường trong khoảng từ 0.2 đến 0.5) cho thấy sự hiện diện của đất trống, đất nông nghiệp mới cày xới, khu vực khai thác hoặc nơi bị phá rừng. Ngược lại, các giá trị thấp hoặc âm thường liên quan đến thảm thực vật dày đặc hoặc mặt nước.

Chỉ số **SAVI** nhằm mục đích đánh giá mật độ thảm thực vật xanh, đồng thời giảm thiểu ảnh hưởng của đất nền trong các khu vực có thảm thực vật thưa thớt. Chỉ số này được điều chỉnh từ NDVI bằng cách sử dụng hằng số L (thường là 0.5) để hiệu chỉnh tác động của đất. Giá trị SAVI cao (thường trong khoảng từ 0.3 đến 0.8) cho thấy sự hiện diện của thảm thực vật khỏe mạnh, chẳng hạn như rừng hoặc đồng cỏ dày đặc. Ngược lại, các giá trị thấp hoặc gần 0 thường liên quan đến đất trống, bề mặt đô thị hoặc khu vực có thảm thực vật rất thưa.

2.3 Chọn lựa đặc trưng

Chọn lựa đặc trưng là một bước quan trọng trong quy trình xây dựng mô hình phân loại lớp phủ bề mặt từ dữ liệu ảnh vệ tinh Sentinel-2. Quá trình này không chỉ giúp loại bỏ những đặc trưng không mang nhiều thông tin hoặc gây nhiễu, mà còn cải thiện hiệu suất tính toán và tăng khả năng giải thích của mô hình. Trong nghiên cứu này, nhóm tiếp cận việc chọn lựa đặc trưng theo hai giai đoạn chính. Đầu tiên là **loại bỏ các đặc trưng không phù hợp** dựa trên kiến thức chuyên môn và phân tích tương quan. Tiếp theo đó là đánh giá mức độ đóng góp của các đặc trưng còn lại thông qua các phương pháp định lượng như **SHAP** (SHapley Additive exPlanations) [7] và **loại bỏ tính năng đê quy** (Recursive Feature Elimination - RFE) [10].

2.3.1 Loại bỏ đặc trưng không phù hợp

Trong quá trình phân tích phân loại lớp phủ, không phải tất cả các phẩn của Sentinel-2 đều mang thông tin hữu ích cho mục tiêu nghiên cứu. Một số phẩn được thiết kế đặc biệt để phát hiện các hiện tượng khí quyển nhất định mà không liên quan trực tiếp đến đặc tính của lớp phủ bề mặt. Cụ thể, nhóm đã loại bỏ các phẩn sau:

- Phẩn 8 (B8): Như đã trình bày trong phần phân tích tương quan đa biến, phẩn 8 thể hiện mức độ tương quan rất cao với phẩn 8A (hệ số tương quan 0.98). Tuy nhiên, do phẩn 8A được đánh giá là phù hợp hơn cho bài toán, nên phẩn 8A được lựa chọn để sử dụng trong các bước phân tích và xây dựng mô hình tiếp theo.
- Phẩn 10 (B10): Được thiết kế đặc biệt để phát hiện mây ti, phẩn này có giá trị trong việc xác định và loại bỏ nhiễu do mây mỏng, nhưng không cung cấp thông tin phẩn cần thiết cho việc phân loại lớp phủ.
- Kinh độ và Vĩ độ (Longitude, Latitude): Mặc dù chứa thông tin về vị trí không gian, các đặc trưng này không trực tiếp hữu ích trong quá trình phân loại dựa trên đặc tính phản xạ của đối tượng.

2.3.2 Đánh giá độ quan trọng đặc trưng bằng SHAP

SHapley Additive exPlanations (SHAP) được sử dụng trong nghiên cứu này như một công cụ mạnh mẽ để đánh giá độ quan trọng của các đặc trưng một cách khách quan và định lượng. Phương pháp này dựa trên lý thuyết trò chơi của Shapley, đo lường mức đóng góp của từng đặc trưng đối với dự đoán của mô hình. Điểm mạnh của SHAP là khả năng giải thích được mối quan hệ phức tạp giữa các đặc trưng và kết quả dự đoán, cung cấp cả mức độ quan trọng tổng thể và tác động cục bộ của từng đặc trưng đối với từng dự đoán cụ thể. SHAP giúp làm rõ vai trò của từng kênh phẩn trong việc phân biệt các loại bề mặt khác nhau, đặc biệt là khi các đặc trưng có mối tương quan phức tạp và tương tác phi tuyến với nhau. Điều này không chỉ cải thiện hiệu suất mô hình thông qua việc chọn lọc đặc trưng tối ưu mà còn nâng cao khả năng giải thích của mô hình. Sau khi thử nghiệm SHAP với các mô hình khác nhau¹ và thu được 10 phẩn có mức độ ảnh hưởng nhất bao gồm: B1, B2,

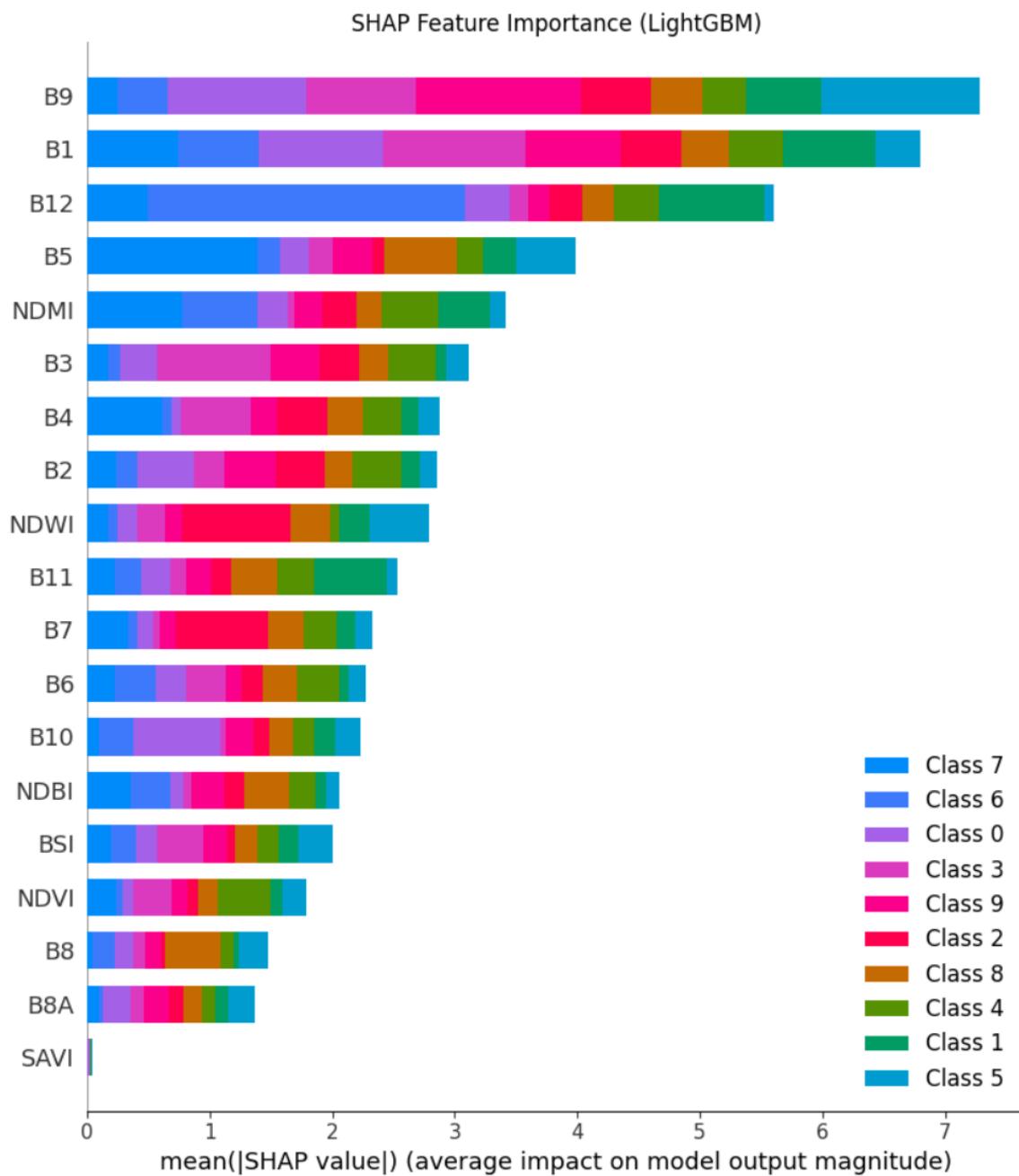
¹Xem chi tiết tại: SHAP Selection Notebook

B3, B4, B5, B9, B11, B12, NDWWI, NDMI (Hình 2.6)

2.3.3 Tìm số lượng đặc trưng tối ưu bằng RFE

Recursive Feature Elimination (RFE) là một phương pháp chọn lựa đặc trưng theo hướng tiếp cận loại bỏ đệ quy, trong đó mô hình được huấn luyện lặp đi lặp lại và các đặc trưng ít quan trọng nhất được loại bỏ trong mỗi lần lặp cho đến khi đạt được số lượng đặc trưng mong muốn. Trong dữ liệu viễn thám, các kênh phổ thường có mối liên hệ phức tạp và phụ thuộc lẫn nhau khi phản ánh đặc tính của các đối tượng bề mặt. RFE kết hợp với các mô hình cơ sở như Random Forest cho phép đánh giá tổng thể hiệu suất của các tổ hợp đặc trưng khác nhau, giúp xác định bộ đặc trưng tối ưu đại diện cho đặc tính phổ của các lớp phủ bề mặt mà không làm mất thông tin quan trọng hoặc đưa vào nhiều không cần thiết.

Kết quả được thể hiện trong Hình 4.1 cho thấy rằng hiệu suất của mô hình (đo bằng điểm xác thực chéo) tăng dần khi số lượng đặc trưng được chọn tăng lên, và đạt giá trị tối ưu khi sử dụng 10 đặc trưng. Ở trường này, mô hình đạt điểm trung bình là 0.858, cho thấy đây là tập con đặc trưng tốt nhất về mặt hiệu suất dự đoán mà không làm tăng độ phức tạp không cần thiết. Cụ thể 10 đặc trưng tốt nhất mà kỹ thuật này mang lại bao gồm: B1, B3, B4, B5, B8A, B9, B11, B12, NDWI, NDMI.



Hình 2.6: Ví dụ kết quả sử dụng SHAP với mô hình LightGBM

Chương 3

Mô hình

3.1 Mô hình hóa

Để giải quyết bài toán phân loại đa lớp một cách hiệu quả, nhóm thử nghiệm cách tiếp cận mô hình hóa đa tầng từ các kỹ thuật học máy truyền thống đến các mô hình học sâu. Các chiến thuật được chọn bao gồm sử dụng mô hình truyền thống SVM, Tree-based, và kết hợp với kỹ thuật voting. Ngoài ra nhóm sử dụng kỹ thuật K-Fold Cross Validation để chia dữ liệu huấn luyện thành nhiều phần

3.1.1 Support Vector Machine (SVM)

Trước tiên, nhóm lựa chọn sử dụng mô hình SVC, một mô hình phân loại thuộc họ SVM như một phương pháp cơ sở (baseline) cho bài toán. SVC là mô hình truyền thống thường cho kết quả tốt khi các lớp trong dữ liệu có ranh giới phân tách rõ ràng trong không gian đặc trưng. Mặc dù Linear Kernel có ưu điểm là nhanh và đơn giản, nhưng lại bị giới hạn trong các bài toán tuyến tính. Do đó, nhóm sử dụng RBF Kernel, một hàm kernel phi tuyến, giúp mô hình học được các mối quan hệ phức tạp hơn giữa các điểm dữ liệu. Hàm kernel RBF được định nghĩa như sau:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

Trong đó:

- x, x' là hai vector đầu vào,

- $\|x - x'\|^2$ là bình phương khoảng cách Euclidean giữa x và x' ,
- $\gamma > 0$ là tham số điều chỉnh độ ảnh hưởng của một điểm dữ liệu.

Ở mức cơ bản nhóm thiết lập các tham số:

1. `kernel='rbf'`: Sử dụng RBF Kernel.
2. `C=1`: Set tham số regularization (tham số này kiểm soát mức độ phạt cho các điểm bị phân loại sai)

3.1.2 K-Nearest Neighbors (KNN)

Dựa trên phân tích phân bố dữ liệu thông qua biểu đồ scatter plot và t-SNE, chúng tôi quan sát thấy xu hướng phân cụm rõ ràng trong tập dữ liệu. Điều này gợi ý khả năng ứng dụng hiệu quả của phương pháp K-Nearest Neighbors (KNN), một thuật toán thuộc nhóm Distance-based model tương tự như Support Vector Machine (SVM). Mô hình KNN được triển khai với các thông số cấu hình cơ bản như sau:

1. Số lượng láng giềng (`n_neighbors=5`): Thuật toán xác định 5 điểm dữ liệu gần nhất với điểm cần dự đoán và sử dụng thông tin phân loại của các điểm này để quyết định nhãn đầu ra.
2. Hệ số trọng số (`weights='uniform'`): Tất cả các láng giềng được xét đến đều có mức độ ảnh hưởng ngang nhau trong quá trình ra quyết định phân loại, không phụ thuộc vào khoảng cách tương đối so với điểm cần dự đoán.
3. Phương pháp đo khoảng cách (`metric='minkowski'`, `p=2`): Khoảng cách giữa các điểm dữ liệu được tính toán theo công thức khoảng cách Euclidean, là trường hợp đặc biệt của khoảng cách Minkowski khi $p=2$.

3.1.3 Tree-based Model

Khác với SVM, Tree-based Model dễ hiểu hơn và có thể phân loại các tập dữ liệu phi tuyến tính mà không cần phải chuyển đổi dữ liệu thành không gian đặc trưng tuyến tính. Tree-based Model cũng cung cấp thêm thông tin về độ quan trọng của các đặc trưng thông qua quá trình phân chia.

Random Forest: Nhằm cải thiện độ chính xác của mô hình và hạn chế hiện tượng overfitting, nhóm đã chọn thử nghiệm mô hình này. Random Forest kết hợp nhiều cây quyết định để tránh sai sót mà một cây quyết định đơn lẻ mắc phải, kết quả cuối cùng được đưa ra thông qua cơ chế bỏ phiếu (voting). Ngoài ra, Random Forest còn có khả năng đánh giá mức độ quan trọng của từng đặc trưng (feature importance), giúp nhóm có thể hiểu rõ hơn về các yếu tố ảnh hưởng đến kết quả phân loại. Random Forest sử dụng phương pháp **Bagging** tạo nhiều cây độc lập từ các tập con dữ liệu được lấy mẫu ngẫu nhiên và bình chọn (vote) kết quả. Hướng làm này giúp giảm phương sai (variance) giúp mô hình ổn định.

XGBoost: Bên cạnh mô hình Random Forest, nhóm cũng tiến hành thử nghiệm với XGBoost, một thuật toán tăng cường hiện đại được thiết kế nhằm tối ưu hóa cả hiệu suất tính toán và độ chính xác dự đoán. XGBoost áp dụng phương pháp **Boosting**, xây dựng các cây quyết định theo thứ tự tuần tự, trong đó mỗi cây mới được huấn luyện để sửa các sai lệch còn tồn tại của tổ hợp các cây trước. Điều này góp phần làm giảm cả phương sai và độ lệch trong mô hình. Ngoài ra, XGBoost tích hợp nhiều kỹ thuật cải tiến như chính quy hóa L1 và L2, cắt tỉa sớm (early stopping) và shrinkage, giúp nâng cao khả năng tổng quát hóa của mô hình. Các thông số được thiết lập cơ bản gồm:

- Số lượng ước lượng cơ sở (`n_estimators=100`): Nhóm thiết lập số lượng cây quyết định trong quá trình boosting là 100, một giá trị đủ lớn để mô hình có thể học được các mối quan hệ phức tạp trong dữ liệu. Tuy nhiên, giá trị này được giữ ở mức độ vừa phải để tránh hiện tượng quá khớp (overfitting), đảm bảo mô hình không học quá sâu vào đặc điểm nhiều của tập huấn luyện.
- Độ sâu tối đa của cây (`max_depth=4`): Mỗi cây quyết định trong tập hợp được giới hạn ở độ sâu tối đa là 4. Việc giới hạn này giúp mỗi cây đủ phức tạp để nắm bắt các mối quan hệ trong dữ liệu, nhưng không quá chi tiết để dẫn đến việc mất khả năng tổng quát hóa.
- Tốc độ học (`learning_rate=0.1`): Giá trị này đảm bảo mô hình hội tụ ổn định và không bị dao động mạnh trong quá trình huấn luyện.
- Phương pháp đánh giá (`eval_metric='logloss'`): Nhóm sử dụng logarithmic

loss làm thước đo đánh giá trong quá trình huấn luyện. Phương pháp này đo lường mức độ khác biệt giữa xác suất dự đoán của mô hình và nhãn thực tế của dữ liệu. Log loss đặc biệt phù hợp với bài toán phân loại, vì nó không chỉ đánh giá độ chính xác của dự đoán mà còn xem xét độ tin cậy của các dự đoán đó thông qua xác suất.

3.1.4 Voting Classifier

Voting Classifier giúp kết hợp các mô hình thành viên lại với nhau, tận dụng điểm mạnh của từng mô hình đồng thời bù đắp cho những điểm yếu của chúng. Nhóm sử dụng Voting Classifier với 3 mô hình thành viên là: SVM, Random Forest, XGBoost. Phương pháp voting được áp dụng là “soft voting”, trong đó mỗi mô hình thành viên đóng góp tỷ lệ xác suất dự đoán cho kết quả cuối cùng, và tất cả các mô hình thành phần đều có trọng số bằng nhau.

3.1.5 Stacking Ensemble

Để khai thác tối đa sức mạnh dự đoán từ nhiều loại mô hình khác nhau, nhóm tiếp tục áp dụng kỹ thuật xếp chồng (Stacking hay Stacked Generalization). Mục tiêu của Stacking là huấn luyện một mô hình cấp cao (meta-model) để học cách kết hợp các dự đoán từ một tập hợp các mô hình cơ sở (base models) một cách tối ưu, thay vì chỉ dựa vào các quy tắc đơn giản như bỏ phiếu đa số hay trung bình xác suất như trong Voting Classifier. Kỹ thuật có thể mang lại hiệu suất phân loại cao hơn do khả năng tận dụng và tổng hợp thông minh các điểm mạnh đặc thù của từng mô hình thành phần.

Quy trình hoạt động của Stacking thường bao gồm hai cấp độ chính:

1. **Cấp độ 0:** Tại tầng này, nhiều mô hình cơ sở đa dạng (ví dụ: SVM với RBF kernel, K-Nearest Neighbors, Random Forest, XGBoost) được huấn luyện độc lập trên tập dữ liệu huấn luyện gốc. Để tạo ra đầu vào cho tầng tiếp theo mà không bị “rò rỉ” thông tin (tức là mô hình meta không học trên chính dự đoán của base model trên dữ liệu mà nó đã thấy), kỹ thuật K-Fold Cross-Validation thường được sử dụng. Cụ thể, tập huấn luyện được chia thành K phần (folds). Với mỗi fold, các base model được huấn luyện trên K-1 folds còn lại và sau đó

đưa ra dự đoán trên fold thứ K (fold được giữ lại). Quá trình này được lặp lại K lần, đảm bảo mỗi mẫu trong tập huấn luyện gốc đều có một dự đoán được tạo ra bởi một mô hình chưa từng “nhìn thấy” nó trong quá trình huấn luyện của fold đó. Tập hợp các dự đoán này được gọi là dự đoán “out-of-fold” (OOF).

2. **Cấp độ 1:** Một meta-model (thường là một mô hình tương đối đơn giản như Logistic Regression để tránh overfitting, hoặc đôi khi là một mô hình mạnh hơn như XGBoost nếu cần thiết) được huấn luyện. Đầu vào đặc trưng cho meta-model này chính là tập hợp các dự đoán OOF được tạo ra bởi tất cả các base model ở cấp độ 0. Nhãn mục tiêu (target) cho việc huấn luyện meta-model chính là nhãn lớp phủ thực tế của tập dữ liệu huấn luyện gốc. Meta-model học cách phân tích các dự đoán từ tầng dưới và đưa ra quyết định phân loại cuối cùng.

Sau khi meta-model đã được huấn luyện, để dự đoán trên tập dữ liệu kiểm tra (test set) mới, quy trình như sau: đầu tiên, tất cả các base model (đã được huấn luyện lại một lần nữa trên toàn bộ tập huấn luyện gốc để có hiệu năng tốt nhất) sẽ đưa ra dự đoán của chúng trên tập test. Sau đó, các dự đoán này được dùng làm đầu vào cho meta-model đã được huấn luyện để tạo ra kết quả phân loại cuối cùng. Bằng cách này, Stacking cho phép hệ thống học được cách kết hợp phức tạp và hiệu quả các dự đoán từ nhiều nguồn, thường dẫn đến mô hình tổng thể mạnh mẽ và có khả năng tổng quát hóa tốt hơn.

3.1.6 Mô hình học sâu

MLP MLP là một phiên bản mở rộng của ANN, trong đó số lượng lớp ẩn được tăng lên để cải thiện khả năng học các đặc trưng phi tuyến tính. MLP là một mạng nơ-ron truyền thẳng (feedforward neural network), nghĩa là dữ liệu di chuyển từ lớp đầu vào qua các lớp ẩn và đến lớp đầu ra mà không có vòng lặp ngược. Điểm mạnh của MLP so với ANN cơ bản nằm ở khả năng xử lý các bài toán phi tuyến tính phức tạp hơn nhờ nhiều lớp ẩn, với mỗi lớp áp dụng một phép biến đổi phi tuyến thông qua hàm kích hoạt. Tuy nhiên, MLP vẫn có những hạn chế nhất định, chẳng hạn như yêu cầu số lượng lớn dữ liệu huấn luyện để tránh hiện tượng quá khớp (overfitting) và thời gian tính toán tăng đáng kể khi quy mô mạng lớn hơn. Để tối ưu hóa MLP, các tham số quan trọng cần điều chỉnh bao gồm:

- Số lượng lớp ẩn (number of hidden layers): Quyết định độ sâu của mạng, ảnh hưởng đến khả năng học các đặc trưng phức tạp.
- Số nơ-ron mỗi lớp (number of neurons per layer): Ảnh hưởng đến độ chi tiết của các đặc trưng được học.
- Tốc độ học (learning rate): Điều chỉnh tốc độ cập nhật trọng số trong quá trình huấn luyện, tương tự như trong Gradient Boosting.
- Hàm kích hoạt (activation function): Lựa chọn giữa ReLU, sigmoid hoặc tanh để phù hợp với đặc điểm dữ liệu.

Như vậy MLP là một trong những lựa chọn phù hợp cho các bài toán phân loại ảnh vệ tinh đơn giản, đặc biệt khi dữ liệu đã được tiền xử lý tốt và không yêu cầu trích xuất đặc trưng không gian phức tạp như trong các mô hình học sâu tiên tiến hơn (chẳng hạn như Convolutional Neural Networks - CNN).

TabPFN[1] là một mô hình tiên tiến được đào tạo trước cho dữ liệu dạng bảng, sử dụng cơ chế học trong ngữ cảnh (In-context learning - ICL) tương tự như mô hình ngôn ngữ lớn. Thay vì được huấn luyện cho từng tập dữ liệu cụ thể, TabPFN học cách trở thành một thuật toán tổng quát có thể áp dụng cho nhiều bộ dữ liệu khác nhau.

Quy trình phát triển TabPFN bao gồm hai giai đoạn chính. Đầu tiên tạo dữ liệu tổng hợp dựa trên mô hình nhân quả có cấu trúc (Structural Causal Models - SCMs) thông qua đồ thị tuần hoàn có hướng (Directed Acyclic Graph - DAG), áp dụng các biến đổi đa dạng và kỹ thuật hậu xử lý để tạo ra khoảng 100 triệu tập dữ liệu. Sau đó, huấn luyện một mô hình transformer (PFN) để dự đoán các giá trị mục tiêu bị che giấu trên các tập dữ liệu này. Mô hình sử dụng kiến trúc transformer được điều chỉnh với cơ chế chú ý hai chiều cho dữ liệu bảng, bắt biến với thứ tự mẫu và đặc trưng. Khi áp dụng cho dữ liệu thực tế, TabPFN chỉ cần một lần chuyển tiếp duy nhất, sử dụng mẫu huấn luyện có nhãn như ngữ cảnh để dự đoán. Đối với bài toán hồi quy, mô hình sử dụng phân phối đều ra hàng số từng phần để mô hình hóa sự không chắc chắn. Hiệu suất được tối ưu hóa thông qua kỹ thuật tập hợp (ensemble) với xáo trộn thứ tự đặc trưng và lưu trữ trạng thái huấn luyện để tăng tốc độ suy luận.

3.2 Tối ưu hóa siêu tham số

RandomizedSearchCV là một phương pháp tìm kiếm ngẫu nhiên các tổ hợp siêu tham số trong không gian tìm kiếm được định nghĩa trước. Thay vì kiểm tra toàn bộ các tổ hợp như trong Grid Search, phương pháp này chỉ chọn một số lượng tổ hợp ngẫu nhiên (xác định bởi tham số `n_iter`), từ đó đánh giá hiệu suất mô hình dựa trên tiêu chí đã chọn. Cách tiếp cận này giúp giảm đáng kể chi phí tính toán, đặc biệt khi không gian siêu tham số quá lớn. Quy trình tối ưu hóa siêu tham số bằng RandomizedSearchCV bao gồm các bước sau:

1. Xác định không gian siêu tham số và số lượng tổ hợp cần thử (`n_iter`)
2. Chọn mô hình cần tối ưu siêu tham số
3. Dùng RandomizedSearchCV để tìm kiếm ngẫu nhiên các siêu tham số tốt
4. Đánh giá kết quả từ RandomizedSearchCV để xác định mô hình tốt nhất

Optuna là một thư viện tối ưu hóa siêu tham số dựa trên phương pháp Bayesian, cung cấp khả năng tìm kiếm hiệu quả và hỗ trợ kỹ thuật *pruning* để dừng sớm các thử nghiệm không tiềm năng. Việc sử dụng Optuna chỉ yêu cầu định nghĩa một hàm mục tiêu (`objective`) và chỉ rõ các siêu tham số cần tối ưu. Trong nghiên cứu này, nhóm đã sử dụng Optuna để tối ưu hóa mô hình XGBoost với các siêu tham số nhằm cải thiện ba khía cạnh chính:

1. Cải thiện độ chính xác: `n_estimators`, `max_depth`, `learning_rate`.
2. Giảm hiện tượng quá khớp (overfitting): `subsample`, `colsample_bytree`, `gamma`, `min_child_weight`.
3. Tăng tốc độ huấn luyện và độ ổn định mô hình: `learning_rate`, `gamma`.

Chương 4

Thực nghiệm và đánh giá

4.1 Phương pháp thực nghiệm

4.1.1 Chiến thuật thực nghiệm

Để kiểm chứng được việc xử lý dữ liệu bằng các kiến thức miền và tương quan của dữ liệu, nhóm chia ra các trường hợp thực nghiệm khác nhau như sau:

- **Thực nghiệm 1: Quan sát hiệu suất các mô hình phân loại dựa trên khoảng cách (distance-based) trước và sau khi chuẩn hoá dữ liệu.**
Mô hình phân loại dựa trên khoảng cách được lựa chọn để thử nghiệm là SVC và KNN. Thử nghiệm gồm hai lần chạy, lần chạy thứ nhất sử dụng dữ liệu nguyên bản và lần chạy thứ hai sử dụng dữ liệu sau khi giảm lệch phai trên phân bố và được chuẩn hoá thông qua Log Transform và Standard Scaler.
- **Thực nghiệm 2: Quan sát hiệu suất các mô hình với tham số cơ bản và tham số tối ưu.**
Các mô hình SVC, KNN, Random Forest và XGBoost được lựa chọn để thử nghiệm. Trong đó SVC, KNN và Random Forest được tối ưu hóa tham số bằng cách sử dụng RandomizedSearchCV, còn XGBoost được sử dụng Optuna để tối ưu hóa tham số.
- **Thực nghiệm 3: Sử dụng kĩ thuật Stacking**
Nhóm lựa chọn một tập hợp các mô hình cơ sở đa dạng đã được thử nghiệm

trước đó, bao gồm SVC (với RBF kernel), K-Nearest Neighbors, Random Forest và XGBoost, làm các mô hình ở Cấp độ 0. Để huấn luyện meta-model ở Cấp độ 1 (nhóm chọn sử dụng XGBoost cho tầng này do khả năng xử lý mối quan hệ phức tạp giữa đầu ra các base model). Nhóm cũng áp dụng kỹ thuật K-Fold Cross-Validation (với K=5) trên tập huấn luyện để tạo ra các dự đoán “out-of-fold” (OOF). Các dự đoán OOF này được sử dụng làm đặc trưng đầu vào cho meta-model, giúp nó học được cách trọng số hóa và tổng hợp các dự đoán từ tầng dưới một cách hiệu quả mà không bị rò rỉ thông tin.

- **Thực nghiệm 4: Sử dụng mô hình Voting Classifier.**

Nhóm sử dụng Voting Classifier với ba mô hình thành viên, trong đó hai thành viên thuộc Tree-based model là Random Forest, XGBoost và một thành viên thuộc Distance-based model là SVC. Kiểu voting được cấu hình là “soft”, kết hợp dự đoán của cả ba mô hình bằng trung bình trọng số của xác suất dự đoán.

- **Thực nghiệm 5: Quan sát hiệu suất các mô hình trước và sau khi tăng cường đặc trưng.**

Thử nghiệm toàn bộ mô hình SVC, KNN, Random Forest, XGBoost, và Voting Classifier. Thử nghiệm gồm hai lần chạy, lần chạy thứ nhất sử dụng dữ liệu nguyên bản và lần chạy thứ hai sử dụng dữ liệu sau khi tăng cường đặc trưng.

- **Thực nghiệm 6: Thủ nghiệm kĩ thuật lựa chọn đặc trưng (Feature Selection).**

Nhóm sử dụng mô hình Random Forest và XGBoost để so sánh hiệu suất trước và sau khi giảm chiều dữ liệu. Các đặc trưng với độ quan trọng thấp sẽ được loại bỏ qua các lần lặp trong RFE. Nhóm sẽ lựa chọn ra một tập đặc trưng từ hai tập đặc trưng quan trọng thu được từ quá trình RFE với hai mô hình để áp dụng lên mô hình Voting Classifier.

- **Thực nghiệm 7: Thủ nghiệm thêm với các mô hình học sâu: Deep Learning (MLP) và TabPFN**

Nhóm sử dụng mô hình Multi-layer perceptron (MLP) và mô hình TabPFN từ báo cáo “Accurate predictions on small data with a tabular foundation model” [1] tuy nhiên vì giới hạn của mô hình TabPFN chỉ cho phép tối đa 10,000 điểm

dữ liệu nên nhóm giảm bớt dữ liệu đối với thử nghiệm mô hình này.

4.1.2 Chỉ số đánh giá

Đánh giá Vì tập dữ liệu khá cân đối, sau khi huấn luyện, mô hình được đánh giá trên tập kiểm thử với tiêu chí là độ chính xác (Accuracy). Qua đó, mô hình có độ chính xác lớn nhất sẽ được đánh giá qua ma trận nhầm lẫn (Confusion matrix).

Accuracy là tỷ lệ giữa số dự đoán đúng trên tổng số mẫu dữ liệu. Công thức tính:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Trong đó:

- **TP**: True Positive (Dự đoán đúng là dương tính).
- **TN**: True Negative (Dự đoán đúng là âm tính).
- **FP**: False Positive (Dự đoán sai là dương tính).
- **FN**: False Negative (Dự đoán sai là âm tính).

Precision là tỷ lệ giữa số dự đoán đúng là dương tính trên tổng số dự đoán dương tính. Công thức tính:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall là tỷ lệ giữa số dự đoán đúng là dương tính trên tổng số mẫu thực sự dương tính. Công thức tính:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score F1-Score là thước đo cân bằng giữa **Precision** và **Recall**, đặc biệt hữu ích khi tập dữ liệu bị mất cân đối (số lượng mẫu của các lớp chênh lệch lớn). F1-Score đạt giá trị cao nhất (bằng 1) khi cả Precision và Recall đều cao, và thấp nhất (bằng 0) khi một trong hai chỉ số bằng 0. Đây là chỉ số phù hợp khi cần cân bằng giữa việc tránh bỏ sót mẫu *False Negative* và tránh dự đoán sai mẫu *False Positive*. F1-Score được tính bằng trung bình điều hòa của Precision và Recall, với công thức:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.2 Kết quả thực nghiệm

4.2.1 Thực nghiệm 1: Quan sát hiệu suất các mô hình phân loại dựa trên khoảng cách (distance-based model) trước và sau khi chuẩn hoá dữ liệu

Từ Bảng 4.1, nhóm nhận thấy hiệu suất tăng trên cả hai mô hình khi được huấn luyện trên tập dữ liệu đã được chuẩn hoá. Kết quả này cũng củng cố luận điểm các mô hình phân loại dựa trên khoảng cách rất nhạy cảm với độ lệch chuẩn và đơn vị đo của từng đặc trưng đầu vào. Khi chưa chuẩn hoá, các đặc trưng có giá trị lớn hơn có thể chi phối khoảng cách giữa các điểm dữ liệu, dẫn đến ranh giới phân lớp không tối ưu. Việc chuẩn hoá giúp đưa các đặc trưng về cùng một thang đo, từ đó cải thiện độ chính xác và khả năng khái quát của mô hình.

Một quan sát khác nhóm nhận thấy là mức độ cải thiện phụ thuộc vào mô hình: trong khi mô hình KNN tăng Accuracy 2.4% thì mô hình SVC tăng Accuracy 4.7%. Điều này cho thấy SVC được hưởng lợi nhiều hơn từ chuẩn hoá so với KNN, lý do nhóm đưa ra là vì SVC tối ưu biện phân cách trong không gian đặc trưng, rất nhạy với tỷ lệ giữa các trục.

Bảng 4.1: So sánh hiệu suất các mô hình trước và sau khi chuẩn hoá dữ liệu

Dữ liệu	Mô hình	Accuracy	Precision	Recall	F1 Score
Nguyên bản	SVC	0.774	0.751	0.774	0.747
	KNN	0.824	0.819	0.824	0.813
Sau khi chuẩn hoá	SVC	0.821	0.832	0.821	0.808
	KNN	0.848	0.845	0.848	0.839

4.2.2 Thực nghiệm 2: Quan sát hiệu suất các mô hình với tham số cơ bản và tham số tối ưu

Theo kết quả thu được trong Bảng 4.2, hiệu suất tất cả các mô hình thực nghiệm đều tăng sau khi tối ưu hoá bộ tham số với XGBoost là mô hình đạt độ chính xác cao nhất (**0.874**).

Việc tối ưu hoá tham số giúp cho hiệu suất của SVC tăng vượt bậc so với trước, cụ thể tham số điều chỉnh hoá (regularization parameter) C ở bộ tham số tối ưu là 96.63, lớn hơn rất nhiều so với giá trị mặc định của C là 1. Điều này cho thấy mô hình SVC đạt hiệu suất tốt hơn khi giảm mức độ phạt khi vi phạm margin, chọn một giá trị C lớn giúp mô hình khai thác được cấu trúc dữ liệu rõ ràng hơn.

Mặc dù việc tối ưu hoá tham số giúp cải thiện hiệu suất cho hầu hết các mô hình, tuy nhiên các mô hình dựa trên cây quyết định như Random Forest và XGBoost chỉ tăng nhẹ. Nguyên nhân là do Random Forest và XGBoost vốn là những mô hình mạnh, thường cho hiệu suất khá tốt ngay cả với tham số mặc định nhờ khả năng tự điều chỉnh qua bagging (Random Forest) và boosting (XGBoost). Vì vậy, việc tối ưu tham số chỉ giúp mô hình điều chỉnh chi tiết hơn chứ không tạo ra cải tiến đáng kể. Theo luận giải của nhóm, cả Random Forest và XGBoost đều có xu hướng đạt điểm bão hòa sau một mức độ phức tạp nhất định, điều chỉnh số lượng cây hoặc độ sâu không mang lại nhiều cải thiện sau khi mô hình đã nắm bắt tốt mẫu huấn luyện ban đầu.

Bảng 4.2: So sánh hiệu suất các mô hình trước và sau khi tối ưu tham số

Bộ tham số	Mô hình	Accuracy	Precision	Recall	F1 Score
Cơ bản	SVC	0.774	0.751	0.774	0.747
	KNN	0.824	0.819	0.824	0.813
	Random Forest	0.869	0.868	0.869	0.861
	XGBoost	0.851	0.846	0.851	0.841
Tối ưu	SVC	0.821	0.832	0.821	0.808
	KNN	0.848	0.845	0.848	0.839
	Random Forest	0.872	0.867	0.872	0.865
	XGBoost	0.874	0.867	0.874	0.866

4.2.3 Thực nghiệm 3: Sử dụng kỹ thuật Stacking

Bảng 4.3 trình bày kết quả thu được từ việc áp dụng kỹ thuật Stacking Ensemble. Mô hình này đạt được độ chính xác tổng thể là 0.880, cùng với các chỉ số Precision,

Recall và F1-score tương ứng. Khi so sánh với hiệu suất của các mô hình đơn lẻ tốt nhất đã được tối ưu hóa trong Thực nghiệm 2 (ví dụ: XGBoost với accuracy 0.874), mô hình Stacking cho thấy một sự cải thiện nhỏ. Sự khác biệt này có thể được lý giải bởi khả năng của meta-model (ở đây là XGBoost) trong việc học cách kết hợp các dự đoán từ nhiều mô hình cơ sở một cách thông minh, thay vì chỉ dựa vào các quy tắc kết hợp đơn giản. Meta-model đã học được cách trọng số hóa và tổng hợp đầu ra từ các mô hình Cấp độ 0 dựa trên hiệu suất và đặc tính dự đoán của chúng trên dữ liệu “out-of-fold”, qua đó khai thác hiệu quả điểm mạnh của từng thuật toán thành phần.

Bảng 4.3: So sánh hiệu suất của mô hình tốt nhất trước đó với kỹ thuật Stacking

Mô hình	Accuracy	Precision	Recall	F1 Score
XGBoost (Tuning)	0.874	0.867	0.874	0.866
Stacking	0.875	0.872	0.875	0.869
Stacking (Tuning)	0.880	0.874	0.880	0.873

4.2.4 Thực nghiệm 4: Sử dụng mô hình Voting Classifier

Bảng 4.4 cho thấy Voting Classifier giúp cải thiện hiệu suất so với mô hình đơn tốt nhất trước đó là XGBoost với bộ tham số tối ưu. Khi sử dụng bộ tham số tối ưu thì Voting Classifier càng tăng hiệu suất với tất cả các chỉ số đánh giá đều cao đồng đều. Điều này cho thấy phối hợp nhiều mô hình có thể khai thác điểm mạnh riêng của từng mô hình con, giúp hệ thống tổng thể tổng hợp tốt hơn và khái quát hóa tốt hơn trên dữ liệu. Với hiệu suất tốt nhất và khả năng cân bằng ánh tượng, Voting Classifier (tuning) được lựa chọn là mô hình cuối cùng để triển khai khi thực hiện tạo bản đồ dự đoán.

4.2.5 Thực nghiệm 5: Quan sát hiệu suất các mô hình trước và sau khi tăng cường đặc trưng

Từ bảng 4.5, các mô hình khi tăng số lượng đặc trưng có hiệu suất giảm nhẹ, điều này gợi ý rằng các đặc trưng thêm vào mang thông tin trùng lặp với các đặc trưng hiện có, không đóng góp thêm giá trị phân biệt cho mô hình. Các đặc trưng mới được

Bảng 4.4: So sánh hiệu suất của mô hình tốt nhất trước đó với Voting Classifier

Mô hình	Accuracy	Precision	Recall	F1 Score
XGBoost (Tuning)	0.874	0.867	0.874	0.866
Stacking (Tuning)	0.88	0.874	0.88	0.873
Voting	0.875	0.872	0.875	0.867
Voting (Tuning)	0.882	0.878	0.882	0.875

thêm vào làm mô hình bị nhiễu và overfit nhẹ.

Voting Classifier có hiệu suất vượt trội hơn các mô hình đơn lẻ trong cả hai trường hợp, cho thấy sức mạnh của tổ hợp nhiều mô hình đa dạng, giảm thiểu sai lệch và phương sai đồng thời giúp cải thiện tổng thể hiệu suất phân loại. Kết quả này cũng cố sự đa dạng giữa các mô hình thành phần là một yếu tố then chốt giúp tổ hợp đạt hiệu suất cao vì việc tổng hợp quyết định sẽ làm mờ đi các sai số riêng lẻ.

Random Forest có thể ảnh hưởng tiêu cực bởi đặc trưng dư thừa và nhiều hơn XGBoost. Một phần nguyên nhân có thể đến từ cơ chế chọn đặc trưng ngẫu nhiên tại mỗi node của Random Forest: khi số đặc trưng tăng lên, xác suất chọn nhầm đặc trưng kém liên quan cũng tăng, dẫn đến việc phân tách không hiệu quả trong các cây. Ngược lại, XGBoost sử dụng cơ chế boosting tuần tự và đánh giá độ quan trọng đặc trưng kỹ hơn tại mỗi bước, giúp nó bỏ qua phần lớn các đặc trưng không hữu ích, từ đó giữ được hiệu suất ổn định hơn.

4.2.6 Thực nghiệm 6: Thử nghiệm kĩ thuật lựa chọn đặc trưng (Feature Selection)

Từ Bảng 4.6, nhóm nhận thấy với 10 đặc trưng chọn lựa thay vì 19, độ chính xác của Random Forest thậm chí tăng nhẹ từ 0.866 lên 0.870, điều này cho thấy Random Forest có khả năng ít bị ảnh hưởng bởi nhiễu hơn và có thể tận dụng tốt các đặc trưng quan trọng mà không cần toàn bộ tập đầu vào.

Ví dụ trong lúc thực hiện RFE trên mô hình Random Forest, chỉ với 10 đặc trưng được lựa chọn gồm có B1, B3, B4, B5, B8A, B9, B11, B12, NDWI, NDMI độ chính xác của mô hình là 0.870. Kết quả này cho thấy phần lớn thông tin phân biệt có thể

Bảng 4.5: So sánh hiệu suất các mô hình trước và sau khi tăng cường đặc trưng

Kích thước	Mô hình	Accuracy	Precision	Recall	F1 Score
13 đặc trưng	SVC	0.865	0.859	0.865	0.857
	KNN	0.859	0.852	0.859	0.851
	Random Forest	0.872	0.867	0.872	0.865
	XGBoost	0.874	0.867	0.874	0.866
	Voting	0.882	0.878	0.882	0.875
19 đặc trưng	SVC	0.863	0.859	0.863	0.856
	KNN	0.848	0.845	0.848	0.839
	Random Forest	0.866	0.861	0.866	0.859
	XGBoost	0.872	0.866	0.872	0.865
	Voting	0.880	0.875	0.880	0.872

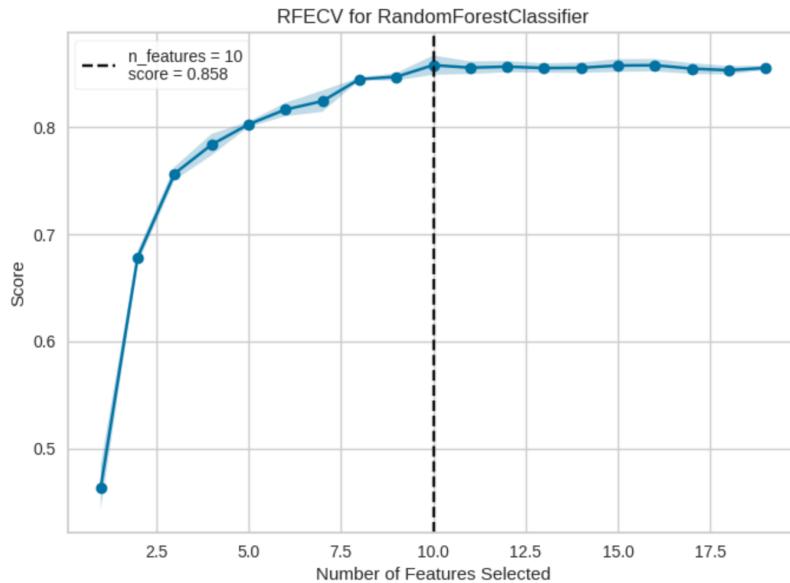
đã được nắm bắt chỉ từ một tập con nhỏ hơn. Ngoài ra, kết quả này gần giống hoàn toàn kết quả thu được từ đánh giá độ quan trọng đặc trưng bằng SHAP cho thấy sự thống nhất giữa hai phương pháp lựa chọn đặc trưng khác nhau và góp phần củng cố độ tin cậy của việc rút gọn đặc trưng trong bài toán này.

Hình 4.1 cho thấy giai đoạn 10 đặc trưng tới 19 đặc trưng hiệu suất duy trì ở quanh 0.84-0.86, việc này chứng tỏ thêm các đặc trưng ngoài ảnh hưởng rất ít tới hiệu suất của mô hình. Giai đoạn cải thiện hiệu suất thực sự chỉ sụt mạnh ở giai đoạn từ 7 đặc trưng về 2 đặc trưng, đây là những đặc trưng đóng góp đáng kể vào khả năng dự đoán của mô hình.

XGBoost bị ảnh hưởng đáng kể khi giảm số đặc trưng, cho thấy mô hình XGBoost phụ thuộc nhiều hơn vào thông tin từ toàn bộ đặc trưng, đặc biệt trong giai đoạn boosting (Việc loại bỏ một số đặc trưng có thể khiến cây ở các vòng boosting sau không còn đủ thông tin để sửa lỗi từ vòng trước).

4.2.7 Thực nghiệm 7: Thử nghiệm các mô hình học sâu: MLP và TabPFN

Từ Bảng 4.7, nhóm nhận thấy được các mô hình học sâu hoạt động khá tốt trên dữ liệu nguyên bản, độ chính xác của các mô hình này khá cao, nằm trong top đầu



Hình 4.1: Hiệu suất mô hình Random Forest Classifier qua từng giai đoạn RFE

Bảng 4.6: So sánh hiệu suất các mô hình trước và sau khi áp dụng kỹ thuật lựa chọn đặc trưng

Kích thước	Mô hình	Accuracy	Precision	Recall	F1 Score
19 đặc trưng	Random Forest	0.866	0.861	0.866	0.859
	XGBoost	0.872	0.866	0.8872	0.865
10 đặc trưng	Random Forest	0.870	0.862	0.870	0.862
	XGBoost	0.823	0.813	0.823	0.811

của các mô hình thử nghiệm. Cụ thể, TabPFN đạt độ chính xác 89%, cao nhất trong số các mô hình được thử nghiệm. So với các phương pháp học máy truyền thống, các mô hình học sâu cho thấy lợi thế rõ rệt khi làm việc với dữ liệu nguyên bản. Đặc biệt, TabPFN đã thể hiện hiệu suất vượt trội, chứng minh khả năng học các mối quan hệ phức tạp trong dữ liệu mà không cần đến các kỹ thuật tăng cường đặc trưng thủ công. Tuy nhiên, do một số lý do liên quan đến việc giới hạn đầu vào của mô hình thử nghiệm, nhóm sẽ không sử dụng TabPFN cho phần dự đoán lớp phủ, tạo bản đồ dự đoán.

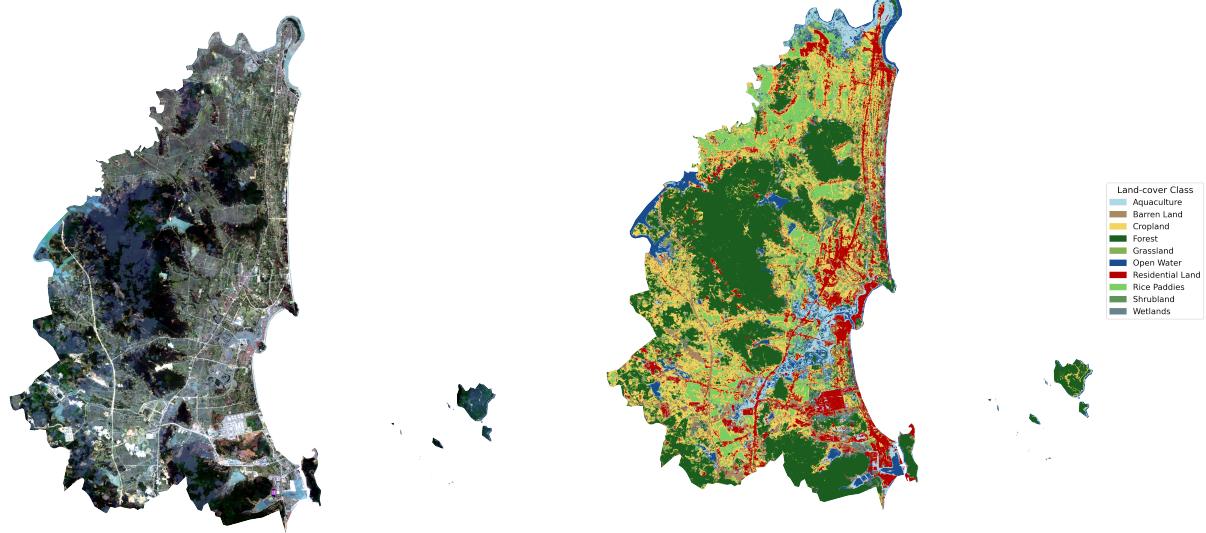
Bảng 4.7: Sử dụng các mô hình học sâu

Mô hình	Accuracy	Precision	Recall	F1 Score
MLP	0.87	0.87	0.87	0.86
TabPFN	0.89	0.89	0.89	0.88

4.3 Tạo bản đồ dự đoán

Nhóm sử dụng ảnh vệ tinh Sentinel-2 để tạo bản đồ phân loại lớp phủ cho thị xã Nghi Sơn (trước đây là huyện Tĩnh Gia), tỉnh Thanh Hóa với mục tiêu phân loại 10 lớp phủ đất, sử dụng mô hình được đánh giá là tốt nhất: Voting (Tuning) trong thực nghiệm 5 với quy trình tạo bản đồ sẽ được thực hiện theo các bước sau:

- 1. Thu thập và tiền xử lý ảnh:** Quá trình phân loại được khởi đầu bằng việc thu thập ảnh vệ tinh Sentinel-2 từ nền tảng Google Earth Engine (GEE), sử dụng bộ dữ liệu COPERNICUS/S2_HARMONIZED - một tập dữ liệu đã được chuẩn hóa nhằm đảm bảo tính nhất quán giữa các cảnh ảnh. Nhóm nghiên cứu sẽ lọc ảnh theo khoảng thời gian (năm 2023), loại bỏ mây và bóng mây, tính toán giá trị trung bình của các ảnh, trích xuất 13 band quang phổ (B1-B12, B8A) và xuất ảnh đã xử lý ra định dạng GeoTIFF.
- 2. Xử lý theo phương pháp phân đoạn:** Để tối ưu hóa bộ nhớ và tăng tốc quá trình xử lý, ảnh vệ tinh sẽ được chia thành các đoạn (chunk) nhỏ theo chiều dọc. Mỗi đoạn có kích thước cố định và được xử lý tuần tự, giúp giảm thiểu yêu cầu về bộ nhớ RAM. Các pixel không hợp lệ (có giá trị NaN) sẽ được xác định và loại trừ khỏi quá trình dự đoán.
- 3. Phân loại và tạo bản đồ kết quả:** Nhóm nghiên cứu sẽ áp dụng mô hình học máy đã được huấn luyện trước đó để phân loại lớp phủ đất cho từng pixel hợp lệ trong mỗi đoạn. Kết quả dự đoán được chuyển đổi về định dạng raster và ghi vào file GeoTIFF mới. Khác với ảnh đầu vào có 13 band, ảnh kết quả chỉ có một lớp duy nhất chứa thông tin về lớp phủ đất đã được phân loại, với kiểu dữ liệu là số nguyên không dấu 8-bit (uint8) đại diện cho các mã lớp phủ đất.



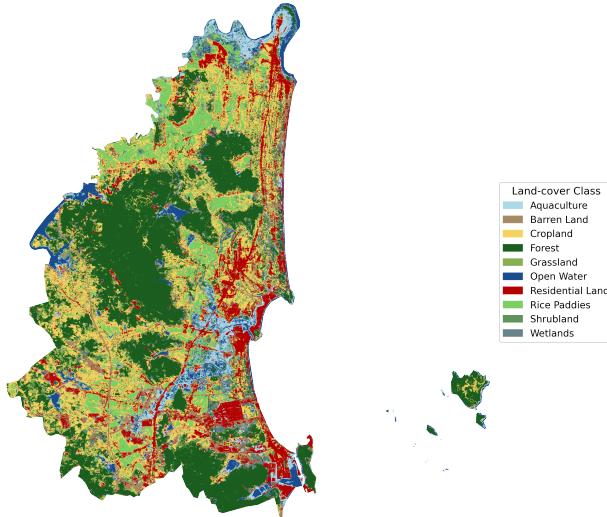
Hình 4.2: Ảnh khu vực thị xã Nghi Sơn

Hình 4.3: Ảnh khu vực thị xã Nghi Sơn sau khi áp dụng mô hình phân loại lớp phủ đất

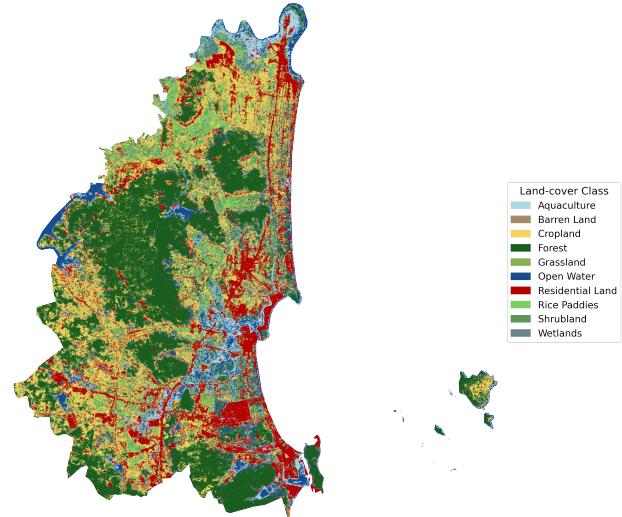
4. Hiển thị và phân tích kết quả: Sau khi hoàn thành quá trình phân loại, bản đồ lớp phủ đất sẽ được hiển thị với bảng màu tùy chỉnh cho 10 lớp phủ đất: aquaculture, barrenland, cropland, forest, grassland, open_water, residentialland, ricepaddies, shrubland và wetlands. Sau đó, nhóm xuất bản đồ thành file hình ảnh.

Sau khi thực hiện quá trình thu được hình ảnh của thị xã Nghi Sơn (trước đây là huyện Tĩnh Gia) trước (Hình 4.2) và sau (Hình 4.3) khi áp dụng mô hình phân loại lớp phủ (voting) với tỷ lệ các lớp phủ như sau: aquaculture (5.59%), barren land (5.46%), cropland (23.05%), forest (33.00%), grassland (1.66%), open water (5.23%), residential land (10.63%), rice paddies (8.96%), shrubland (2.47%), wetlands (3.93%).

Kết quả phân loại lớp phủ ở thị xã Nghi Sơn cho thấy rừng và đất nông nghiệp chiếm tỉ lệ lớn nhất (gồm cropland và rice paddies), phản ánh đúng đặc điểm tự nhiên và sản xuất của địa phương. Tỉ lệ đất ở và nuôi trồng thủy sản cũng khá cao, phù hợp với xu hướng đô thị hóa và phát triển kinh tế biển của khu vực. Các loại đất khác như đất trống, đất cây bụi, đất ngập nước chiếm tỉ lệ nhỏ, cho thấy mô hình đã nhận diện tương đối hợp lý các loại lớp phủ, phù hợp với thực tế sử dụng đất ở Nghi Sơn hiện nay.



Hình 4.4: Ảnh phân loại bằng XGBoost sử dụng dataset phía tây tỉnh Thanh Hóa



Hình 4.5: Ảnh phân loại bằng Multi-Layer Perceptron sử dụng dataset tỉnh Thanh Hóa

Ngoài ra, để tăng tính khách quan và kiểm chứng hiệu quả của mô hình, nhóm cũng đã tiến hành thử nghiệm các mô hình phân loại khác trên các bộ dữ liệu mà nhóm tự thu thập được từ nhiều nguồn khác nhau. Sau quá trình thống kê và so sánh các nhãn phân loại giữa các mô hình, nhóm nhận thấy có một số thay đổi nhỏ về tỉ lệ diện tích các lớp phủ đất. Cụ thể, có một số thay đổi có thể kể đến như diện tích rừng tăng giảm trong khoảng 1-3%, diện tích đất trống tăng khoảng 2-4%, đất cây bụi thay đổi tăng giảm gần 0,5%, ... so với kết quả của mô hình chính.

Những kết quả này cho thấy, mặc dù có sự khác biệt nhất định giữa các mô hình cũng như giữa các bộ dữ liệu đầu vào, nhưng nhìn chung, mô hình phân loại lớp phủ đất mà nhóm xây dựng vẫn đảm bảo được tính ổn định và độ tin cậy cao. Sự biến động tỉ lệ các lớp phủ đất qua các lần thử nghiệm đều nằm trong ngưỡng cho phép và không ảnh hưởng lớn đến tổng thể kết quả phân loại. Điều này chứng tỏ mô hình có khả năng tổng quát hóa tốt, phù hợp để ứng dụng cho các khu vực nghiên cứu khác hoặc các bộ dữ liệu tương tự. Như vậy, các kết quả phân loại thu được hoàn toàn có thể sử dụng làm cơ sở cho các phân tích tiếp theo.

Kết luận

Nghiên cứu này đã phát triển một mô hình phân loại lớp phủ bề mặt Trái Đất với độ chính xác cao dựa trên dữ liệu ảnh vệ tinh đa phổ Sentinel-2. Thông qua quá trình thử nghiệm toàn diện với nhiều mô hình khác nhau, từ các thuật toán dựa trên khoảng cách (SVC, KNN), mô hình dựa trên cây quyết định (Random Forest, XGBoost), đến các phương pháp học kết hợp (Ensemble Learning) và học sâu (MLP, TabPFN), chúng tôi đã xác định được rằng mô hình voting sau khi được tinh chỉnh tham số cho hiệu suất vượt trội nhất. Mô hình này đạt độ chính xác 88% khi phân loại mười lớp phủ bề mặt khác nhau (khu vực dân cư, vùng canh tác lúa, đất nông nghiệp khác, đồng cỏ, đất trống, vùng cây bụi, khu vực rừng, đất ngập nước, mặt nước tự nhiên và các vùng nuôi trồng thủy sản) trên địa bàn tỉnh Thanh Hóa, cải thiện đáng kể so với phương pháp SVC truyền thống.

Nghiên cứu cũng đã áp các kỹ thuật tiền xử lý dữ liệu tiên tiến, đặc biệt là việc ứng dụng SHAP để giải thích và chọn lọc đặc trưng dựa trên mức độ quan trọng. Mặc dù các phân tích cho thấy việc bổ sung đặc trưng phi tuyến chưa mang lại cải thiện đột phá về hiệu năng, nhưng việc loại bỏ các đặc trưng dư thừa đã góp phần tối ưu hóa quy trình huấn luyện, giảm thiểu thời gian xử lý.

Tuy đạt được những kết quả khả quan, nhóm nhận thấy mô hình vẫn còn những hạn chế cần được khắc phục trong các nghiên cứu tiếp theo. Để nâng cao hiệu suất và khả năng ứng dụng thực tiễn, nhóm đề xuất một số hướng phát triển quan trọng. Trước hết, việc tích hợp dữ liệu đa nguồn bao gồm ảnh radar Sentinel-1, mô hình số độ cao và dữ liệu khí hậu sẽ bổ sung thông tin phong phú cho quá trình phân loại. Tiếp theo, hướng đến phát triển phân tích chuỗi thời gian nhằm theo dõi biến động lớp phủ trong dài hạn, giúp nắm bắt được xu hướng thay đổi và các yếu tố tác động.

Cuối cùng, áp dụng các kiến trúc học sâu như CNN hoặc U-Net sẽ giúp khai thác tốt hơn thông tin không gian trong dữ liệu ảnh, nâng cao khả năng nhận diện các đặc trưng phức tạp.

Nghiên cứu này hi vọng có thể hỗ trợ các nhà hoạch định chính sách trong việc đưa ra quyết định dựa trên dữ liệu, đồng thời tạo tiền đề cho việc phát triển các hệ thống giám sát thay đổi lớp phủ bề mặt theo thời gian thực trong tương lai, góp phần vào công tác quản lý đất đai bền vững và ứng phó hiệu quả với các thách thức môi trường đang ngày càng gia tăng.

Tài liệu tham khảo

- [1] Noah Hollmann **and others**. “Accurate predictions on small data with a tabular foundation model”. *inNature*: 637.8045 (2025), **pages** 319–326. ISSN: 1476-4687. DOI: 10 . 1038 / s41586 - 024 - 08328 - 6. URL: <https://doi.org/10.1038/s41586-024-08328-6>.
- [2] Alfredo R Huete. “A soil-adjusted vegetation index (SAVI)”. *inRemote sensing of environment*: 25.3 (1988), **pages** 295–309.
- [3] Suming Jin **and** Steven A Sader. “Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances”. *inRemote sensing of Environment*: 94.3 (2005), **pages** 364–372.
- [4] Stuart K McFeeters. “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features”. *inInternational journal of remote sensing*: 17.7 (1996), **pages** 1425–1432.
- [5] Can Trong Nguyen **and others**. “A modified bare soil index to identify bare land features during agricultural fallow-period in southeast Asia using Landsat 8”. *inLand*: 10.3 (2021), **page** 231.
- [6] Darius Phiri **and others**. “Sentinel-2 data for land cover/use mapping: A review”. *inRemote sensing*: 12.14 (2020), **page** 2291.
- [7] Benedek Rozemberczki **and others**. *The Shapley Value in Machine Learning*. 2022. arXiv: 2202 . 05594 [cs.LG]. URL: <https://arxiv.org/abs/2202.05594>.

- [8] John RG Townshend **and** CO Justice. “Analysis of the dynamics of African vegetation using the normalized difference vegetation index”. **in***International journal of remote sensing*: 7.11 (1986), **pages** 1435–1445.
- [9] Yong Zha, Jay Gao **and** Shaoxiang Ni. “Use of normalized difference built-up index in automatically mapping urban areas from TM imagery”. **in***International journal of remote sensing*: 24.3 (2003), **pages** 583–594.
- [10] Xinzhi Zhou **and others**. “Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization”. **in***Geoscience Frontiers*: 12.5 (2021), **page** 101211.