Tạp chí Khoa học Công nghệ Xây dựng, ĐHXDHN, 2026, 20 (5V): 188–204

# XÂY DỰNG BẢN ĐỒ NHẠY CẢM TRƯỢT LỞ ĐẤT SỬ DỤNG MÔ HÌNH HỌC MÁY VÀ DỮ LIỆU KHÔNG GIAN ĐA NGUỒN

Lưu Thị Diệu Chinh[a], Hà Thị Hằng[b,*], Bùi Duy Quỳnh[b], Vũ Đình Chiều[b], Lê Văn Quang[c], Lương Ngọc Dũng[b]

[a]*Khoa Công trình Thủy, Trường Đại học Xây dựng Hà Nội, 55 đường Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam*
[b]*Khoa Cầu đường, Trường Đại học Xây dựng Hà Nội, 55 đường Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam*
[c]*Ban Đào tạo Kỹ sư chất lượng cao, Trường Đại học Xây dựng Hà Nội, 55 đường Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam*

*Nhận ngày 13/4/2026, Sửa xong 09/6/2026, Chấp nhận đăng 17/6/2026*

**Tóm tắt**

Nghiên cứu này đề xuất phát triển hai thuật toán học máy, bao gồm Categorical Boosting (CatBoost) và Random Forest (RF) để xây dựng bản đồ nhạy cảm trượt lở đất cho tỉnh Thái Nguyên (cũ). Một bộ dữ liệu hiện trạng trượt lở đất trong lịch sử và mười lăm yếu tố đại diện cho các đặc điểm địa hình, địa chất, môi trường, khí hậu của khu vực nghiên cứu đã được sử dụng làm dữ liệu đầu vào. Hiệu suất dự báo của các mô hình học máy được so sánh và đánh giá dựa trên các chỉ số thống kê phổ biến. Kết quả phân tích đường cong ROC trên tập dữ liệu kiểm tra cho thấy mô hình CatBoost có hiệu suất dự báo vượt trội so với mô hình Random Forest với giá trị AUC là 0,950. Theo đó, mô hình CatBoost đã được lựa chọn để xây dựng bản đồ nhạy cảm trượt lở đất tốt nhất. Ngoài ra, bản đồ này cũng được kiểm chứng với các vị trí trượt lở đất mới nhất từ trang web của Hệ thống giám sát thiên tai Việt Nam. Kết quả đạt được này không chỉ thúc đẩy các giải pháp phòng chống trượt lở đất một cách chủ động, mà còn khẳng định tính hiệu quả và độ tin cậy của các mô hình học máy tiên tiến, hiện đại trong lĩnh vực quản lý và giảm thiểu rủi ro thiên tai.

*Từ khoá*: bản đồ nhạy cảm trượt lở đất; học máy; Categorical Boosting; Random Forests; tỉnh Thái Nguyên (cũ); Việt Nam.

LANDSLIDE SUSCEPTIBILITY MAPPING USING MACHINE LEARNING MODELS AND MULTI-SOURCE GEOSPATIAL DATA

**Abstract**

This study developed two machine learning algorithms, Categorical Boosting (CatBoost) and Random Forest (RF), to construct landslide susceptibility maps for Thai Nguyen province (old). A landslide inventory database and fifteen factors representing the topographic, geological, environmental, and climatic characteristics of the study area were used as input data. The predictive performance of ML models was compared using popular statistical metrics. Analysis of the ROC curve on the testing dataset indicated that the CatBoost model had superior predictive performance compared to the RF model, with an AUC value of 0.950. Accordingly, the CatBoost model was selected to generate the best landslide susceptibility map. Furthermore, this map was also validated against the latest landslide locations from the Vietnam Disaster Monitoring System website. This achievement not only promotes proactive landslide-prevention solutions but also confirms the effectiveness and reliability of advanced ML models in disaster risk management and mitigation.

*Keywords*: landslide susceptibility map; machine learning; Categorical Boosting; Random Forests; Thai Nguyen province (old); Vietnam.

https://doi.org/10.31814/stce.huce2026-20(5V)-14 © 2026 Trường Đại học Xây dựng Hà Nội (ĐHXDHN)

## 1. Giới thiệu

Trượt lở đất (TLĐ) là một trong những hiểm họa tự nhiên phổ biến nhất ở các khu vực vùng núi [1]. Do tính chất xuất hiện đột ngột và khó dự đoán nên TLĐ thường gây ra thiệt hại rất lớn về người, nhà cửa, cơ sở hạ tầng và gián đoạn sự phát triển kinh tế. Thống kê trong hơn thập kỷ qua cho thấy

*Tác giả đại diện. *Địa chỉ e-mail:* hanght@huce.edu.vn (Hằng, H. T.)

các vụ TLĐ xảy ra thường xuyên hơn, quy mô lớn hơn, cũng như gây thiệt hại nghiêm trọng hơn [2]. Xu hướng này còn có khả năng gia tăng trong những năm tới do tình trạng phá rừng, quá trình đô thị hóa nhanh chóng và các hình thái thời tiết cực đoan gia tăng do biến đổi khí hậu [3]. Các bản đồ nguy cơ TLĐ được xem là một công cụ trợ giúp hữu ích cho chính quyền và cộng đồng địa phương trong việc cung cấp thông tin không gian về các khu vực tiềm năng xảy ra trượt lở trong tương lai, cho phép con người chủ động phòng ngừa và thích ứng một cách chủ động hơn [4]. Theo đó, chúng có thể dự đoán các khu vực tiềm năng trong tương lai bằng cách xem xét các địa điểm đã từng xảy ra trượt lở trong lịch sử và các yếu tố kích hoạt có liên quan đến địa hình, địa chất, môi trường, khí hậu [5]. Tuy nhiên, kết quả của các bản đồ nguy cơ TLĐ có thể bị ảnh hưởng bởi các mô hình dự đoán được sử dụng, điều đó có nghĩa là việc lựa chọn cách tiếp cận và phương pháp nghiên cứu phù hợp là rất quan trọng để đảm bảo tính khoa học và ứng dụng thực tiễn của các bản đồ nguy cơ TLĐ này [1].

Ngày nay, sự phát triển mạnh mẽ của công nghệ trí tuệ nhân tạo, đặc biệt là các thuật toán học máy (Machine Learning), đã mở ra hướng tiếp cận mới đầy triển vọng trong lập bản đồ nguy cơ TLĐ. Với khả năng tự động học từ dữ liệu, nhận diện các mẫu ẩn và mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các biến đầu vào, công nghệ học máy ngày càng được ứng dụng rộng rãi trong các nghiên cứu về TLĐ [4]. Đặc biệt, các thuật toán học máy tổ hợp, như Categorical Boosting (CatBoost) và Random Forest (RF) là những thuật toán học máy mạnh mẽ, có khả năng làm việc với cả biến liên tục và biến phân loại, rất phù hợp với những kiểu dữ liệu đa dạng như trong đánh giá nguy cơ TLĐ [6, 7]. Hơn nữa, cả hai thuật toán này đều ít nhạy cảm với hiện tượng quá khớp (overfitting) so với các mô hình học máy đơn lẻ khác do chúng tổ hợp kết quả từ nhiều cây quyết định để đưa ra dự báo cuối cùng, thay vì chỉ dựa vào một mô hình duy nhất. Chính vì vậy, hai thuật toán CatBoost và RF được sử dụng khá nhiều trong các nghiên cứu thành lập bản đồ nguy cơ TLĐ trong những năm gần đây [6, 7].

Một số nghiên cứu chỉ thử nghiệm thuật toán RF cùng với một số thuật toán học máy khác, cụ thể, Kim và cs. đã so sánh và đánh giá khả năng dự báo của hai thuật toán học máy là RF và Boosted Tree (BT) [8]. Youssef và cs. đã so sánh và đánh giá hiệu suất dự báo của 4 mô hình học máy khác nhau, bao gồm RF, Boosted Regression Tree (BRT), Classification And Regression Tree (CART) và General Linear Model (GLM) [6]. Trong khi đó, Sun và cs. thì phân tích và so sánh hiệu suất dự báo của hai thuật toán là RF và Boosted Regression Tree (BRT) [9]. Trong một nghiên cứu khác, Sun và cộng sự chỉ sử dụng mô hình RF để lập bản đồ nguy cơ TLĐ dựa trên tối ưu hóa siêu tham số sử dụng thuật toán Bayes [10]. Trong khi đó, Sahin thì tập trung so sánh độ chính xác dự báo của 3 mô hình kết hợp dựa trên cây hồi quy, bao gồm gradient boosting machines (GBM), extreme gradient boosting (XGBoost), và RF [11]. Mới đây nhất, Zhang và cs. so sánh và phân tích hiệu suất dự báo của hai thuật toán học máy là RF và Extreme Gradient Boosting (XGBoost) trong việc lập bản đồ nguy cơ TLĐ [12]. Tương tự như vậy, một số nghiên cứu khác lại chỉ thử nghiệm mỗi thuật toán CatBoost cùng với một số thuật toán học máy khác, điển hình là nghiên cứu của Sahin đã so sánh hiệu suất dự báo của 4 thuật toán tăng cường độ dốc khác nhau, bao gồm Gradient Boosting Machine (GBM), Categorical Boosting (CatBoost), Extreme Gradient Boosting (XGBoost), và Light Gradient Boosting Machine (LightGBM) [7].

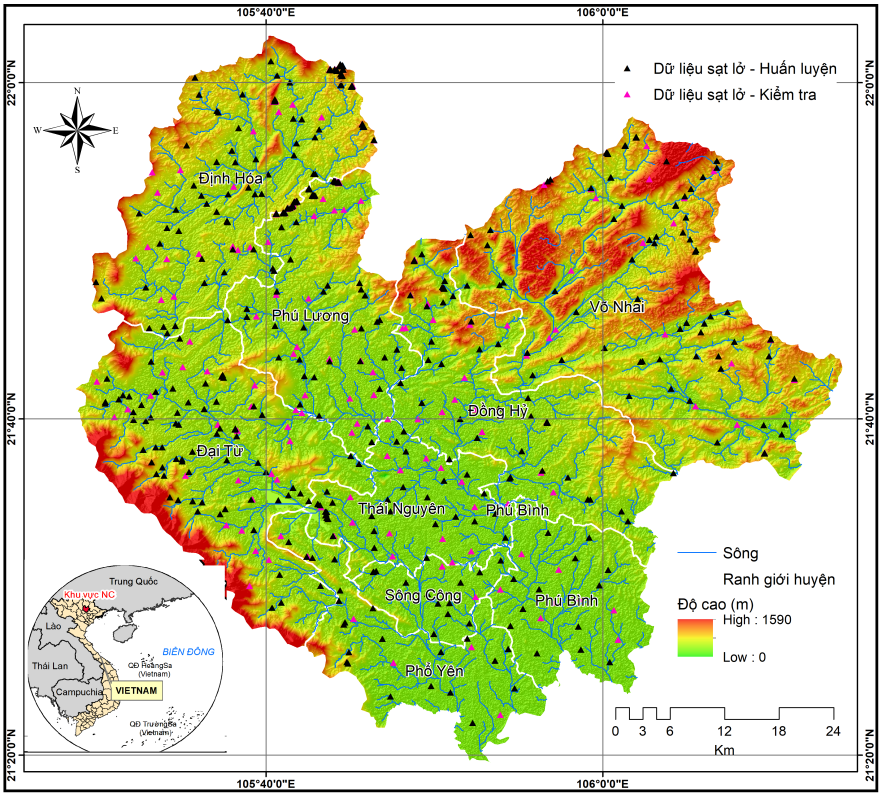
Ở Việt Nam, hai thuật toán học máy RF và CatBoost cũng được sử dụng tách biệt trong một số nghiên cứu về nguy cơ TLĐ. Đối với thuật toán RF, Biên và cs. đã phát triển và áp dụng các mô hình học máy lai mới, RF-FPA và MAB-FPA, kết hợp với hai kỹ thuật tập hợp là rừng ngẫu nhiên (RF) và multiboost (MAB), để xây dựng bản đồ nguy cơ TLĐ cho tỉnh Điện Biên [13]. Đối với thuật toán CatBoost, Hà và cs. đã so sánh hiệu suất dự đoán của các thuật toán tăng cường độ dốc tiên tiến để mô hình hóa khả năng trượt lở đất cho tỉnh Bắc Kạn (cũ), bao gồm Gradient Boosting (GB), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), Categorical

Boosting (CatBoost) và Natural Gradient Boosting (NGBoost) [14].

Nhìn chung, trên thế giới cũng như ở Việt Nam, mặc dù hai thuật toán RF và CatBoost được ứng dụng rất rộng rãi trong xây dựng bản đồ nguy cơ TLĐ, song ít có nghiên cứu trực tiếp so sánh, phân tích và đánh giá hiệu suất dự báo của hai thuật toán này một cách đồng thời. Trong khi đó, do hệ thống các yếu tố môi trường phức tạp của TLĐ tại các khu vực nghiên cứu khác nhau, dẫn tới độ chính xác và khả năng áp dụng thực tế của mỗi bản đồ nguy cơ TLĐ trong từng thuật toán áp dụng là vô cùng quan trọng [1, 9]. Việc thử nghiệm và so sánh hiệu suất dự báo, xem xét bao gồm cả ưu điểm và khả năng áp dụng của từng thuật toán, là rất quan trọng để đạt được một bản đồ nguy cơ TLĐ chính xác hơn. Chính vì vậy, nghiên cứu này đề xuất sử dụng hai thuật toán học máy phổ biến là CatBoost và RF trong đánh giá nguy cơ TLĐ cho địa bàn tỉnh Thái Nguyên (cũ). Kết quả của nghiên cứu này không chỉ nhằm phân tích và lựa chọn được bản đồ nguy cơ TLĐ chính xác hơn, mà còn làm phong phú hơn cách tiếp cận và tạo động lực thử nghiệm nhiều mô hình học máy truyền thống trong đánh giá nhạy cảm trượt lở.

## 2. Khu vực và quy trình nghiên cứu

### *2.1. Khu vực nghiên cứu*




Hình 1. Bản đồ các điểm TLĐ cập nhật mới nhất trên địa bàn tỉnh Thái Nguyên (cũ)

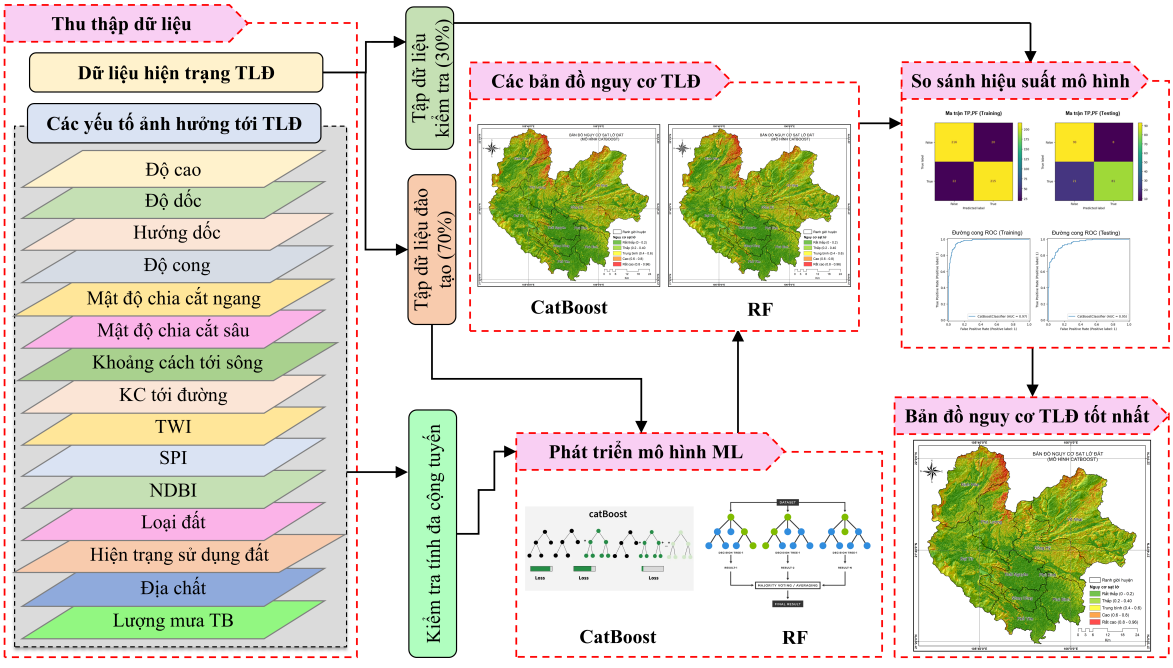
Thái Nguyên (cũ) là tỉnh tiếp giáp với Hà Nội và cách thủ đô Hà Nội 79 km về phía Bắc. Tỉnh nằm trong quy hoạch vùng Thủ đô Hà Nội và được xác định là một trong những trung tâm vùng của các tỉnh Trung du miền núi Bắc Bộ nước ta với diện tích tự nhiên 3.562,82 km$^2$ vào năm 2023. Tỉnh Thái Nguyên (cũ) có tọa độ địa lý từ 21°20′ đến 22°03′ vĩ độ Bắc và từ 105°52′ đến 106°14′ kinh độ Đông. Phía Bắc tiếp giáp với tỉnh Bắc Kạn, phía Tây giáp với các tỉnh Vĩnh Phúc, Tuyên Quang, phía Đông giáp với các tỉnh Lạng Sơn, Bắc Giang và phía Nam tiếp giáp với thủ đô Hà Nội. Tỉnh Thái

Nguyên (cũ) có 9 đơn vị hành chính, bao gồm thành phố Thái Nguyên, thành phố Sông Công, thị xã Phổ Yên và 6 huyện (Phú Bình, Đồng Hỷ, Võ Nhai, Định Hóa, Đại Từ, Phú Lương). Tổng số có 178 đơn vị hành chính cấp xã, bao gồm 32 phường, 9 thị trấn và 137 xã [15].

Địa hình tỉnh Thái Nguyên chủ yếu là đồi núi thấp, chiếm khoảng 2/3 diện tích, với xu hướng thấp dần từ Bắc xuống Nam và từ Tây sang Đông. Về mặt cấu trúc, địa hình tỉnh chịu ảnh hưởng rõ rệt của hai hệ thống cánh cung lớn, bao gồm cánh cung Sông Gâm (ở phía Tây) và cánh cung Ngân Sơn (ở phía Đông). Khí hậu chia làm 2 mùa rõ rệt, mùa mưa thường bắt đầu từ tháng 5 đến tháng 10 và mùa khô bắt đầu từ tháng 10 đến tháng 5 năm sau. Lượng mưa trung bình hàng năm khoảng 2.000 đến 2.500 mm, trong đó, cao nhất vào tháng 8 và thấp nhất vào tháng 1. Chính những điều kiện địa hình và khí hậu đặc thù này khiến cho tỉnh Thái Nguyên thường xuyên phải đối mặt với trượt lở đất vào mỗi mùa mưa bão [16].

### 2.2. *Quy trình nghiên cứu*

Quy trình được thực hiện trong nghiên cứu này bao gồm 6 bước chính như sau: (1) Chuẩn bị và thu thập dữ liệu, (2) Kiểm tra hiện tượng đa cộng tuyến của các yếu tố ảnh hưởng đến trượt lở đất, (3) Dữ liệu hiện trạng trượt lở đất được chia thành hai nhóm theo phương pháp lấy mẫu ngẫu nhiên: 70% tập dữ liệu đào tạo (tương ứng với 237 điểm trượt lở) được sử dụng để xây dựng mô hình dự báo nhạy cảm trượt lở và 30% tập dữ liệu kiểm tra (tương ứng với 101 điểm trượt lở) được sử dụng để kiểm chứng mô hình dự báo, (4) Phát triển các mô hình học máy tiên tiến, bao gồm CatBoost và RF, (5) Xây dựng các mô hình nhạy cảm trượt lở đất, và (6) So sánh hiệu suất dự báo của các mô hình để lựa chọn và thành lập bản đồ nhạy cảm trượt lở có độ chính xác và độ tin cậy cao nhất (Hình 2).



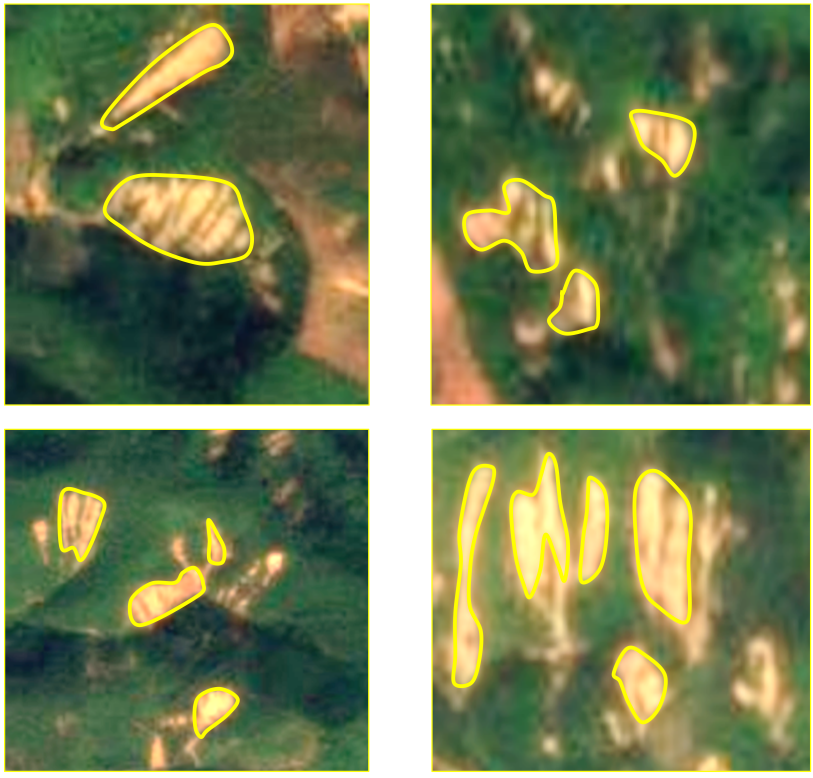

Hình 2. Quy trình nghiên cứu

### 2.3. *Dữ liệu sử dụng*

a. Dữ liệu hiện trạng trượt lở đất

Dữ liệu hiện trạng trượt lở đất đóng vai trò vô cùng quan trọng, là nguồn dữ liệu đầu vào không thể thiếu trong các nghiên cứu đánh giá nhạy cảm trượt lở đất [11]. Chúng không chỉ cung cấp các

thông tin về vị trí, thời điểm xảy ra trượt lở đất trong lịch sử, mà còn đóng vai trò quan trọng trong việc kiểm chứng độ chính xác của các mô hình dự báo [4]. Trong nghiên cứu này, dữ liệu hiện trạng trượt lở đất được thu thập từ nhiều nguồn khác nhau, trong đó, có 224 điểm được thừa hưởng từ kết quả khảo sát của đề án do Viện Khoa học địa chất và Khoáng sản chủ trì thực hiện vào năm 2017, tại trang web http://canhbaotruotlo.vn/phanvungcactinh.html, và 114 điểm được bổ sung từ giải đoán ảnh vệ tinh độ phân giải không gian cao PlanetScope và ảnh Google Earth kết hợp với mô hình số độ cao (DEM) trong phần mềm ArcGIS Pro (Bảng 1). Theo đó, các khối trượt sẽ được nhận diện và đối chiếu với tiêu chí độ dốc từ 10° trở lên, mỗi một khối trượt sẽ được chuyển đổi thành một điểm đại diện trọng tâm cho khối trượt đó (Hình 3). Trên cơ sở đó, bản đồ hiện trạng trượt lở đất khu vực tỉnh Thái Nguyên bao gồm 338 điểm trượt lở (Hình 1).



Hình 3. Kết quả nhận biết các khối trượt trên ảnh vệ tinh PlanetScope và Google Earth kết hợp với DEM trong việc xác định vị trí các điểm TLĐ

Để đảm bảo tính cân bằng cho dữ liệu đầu vào của mô hình, nghiên cứu tiến hành lựa chọn các điểm không trượt lở với tỷ lệ 1:1 so với điểm trượt lở. Theo đó, có 338 điểm không trượt lở đã được lựa chọn. Những điểm này được lấy mẫu ngẫu nhiên tại những khu vực được xem là ổn định trên địa bàn tỉnh Thái Nguyên thông qua việc kiểm tra kỹ lưỡng một chuỗi ảnh vệ tinh PlanetScope và Google Earth trong giai đoạn 2017 – 2024, đảm bảo rằng tại các vị trí này chưa từng xuất hiện dấu hiệu trượt lở trong quá khứ.

b. Các yếu tố ảnh hưởng tới trượt lở đất

Trong nghiên cứu này, 15 yếu tố ảnh hưởng tới trượt lở đất được lựa chọn dựa trên tính sẵn có và khả năng tiếp cận, bao gồm: độ cao, độ dốc, hướng sườn, độ cong bề mặt, mật độ chia cắt ngang, mật độ chia cắt sâu, khoảng cách đến sông suối, khoảng cách đến đường giao thông, Chỉ số độ ẩm địa

Bảng 1. Các nguồn dữ liệu được sử dụng trong nghiên cứu

| TT | Dữ liệu thu thập | Nguồn thu thập | Năm thu thập | Định dạng dữ liệu |
|---|---|---|---|---|
| 1 | Hiện trạng trượt lở đất | Viện Khoa học địa chất và Khoáng sản; Giải đoán trên ảnh vệ tinh và DEM | 2017–2025 | Point |
| 2 | Mô hình số độ cao DEM | https://gee-community-catalog.org/projects/fabdem/#source-data-structure | 12–2024 | Grid |
| 3 | Lượng mưa trung bình năm | 17 trạm quan trắc lượng mưa trên địa bàn tỉnh | 2011–2024 | Point |
| 4 | Ảnh vệ tinh Landsat-8 | https://earthexplorer.usgs.gov/ | 07–2022 | Grid |
| 5 | Bản đồ địa hình | Nhà xuất bản Bản đồ, Bộ Tài nguyên và Môi trường | 2021 | Vector |
| 6 | Hiện trạng sử dụng đất | Bộ Tài nguyên và Môi trường | 2021 | Vector |
| 7 | Loại đất | Bộ Tài nguyên và Môi trường | 2021 | Vector |
| 8 | Địa chất | Tổng cục địa chất và khoáng sản, Bộ Tài nguyên và Môi trường | 2021 | Vector |

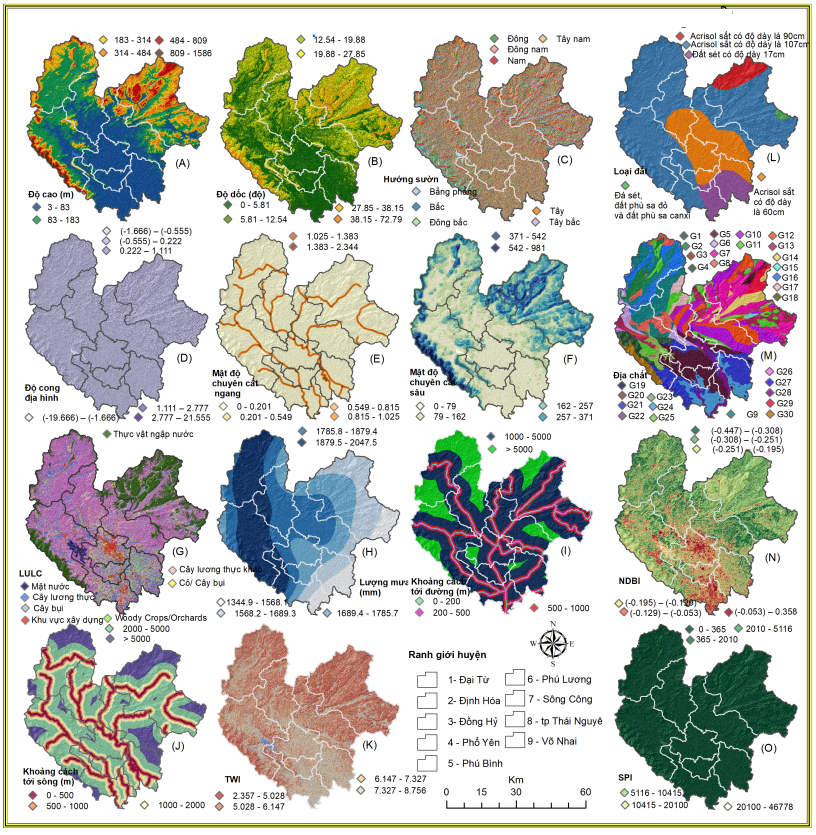
hình (TWI), Chỉ số năng lượng dòng chảy (SPI), Chỉ số khác biệt chuẩn hóa xây dựng (NDBI), loại đất, hiện trạng sử dụng đất, địa chất và lượng mưa trung bình. Trong đó, dữ liệu độ cao được khởi tạo từ dữ liệu FABDEM (Forest And Buildings removed Copernicus DEM), tải xuống vào tháng 12 năm 2024 từ trang web https://gee-community-catalog.org/projects/fabdem/#source-data-structure, với độ phân giải không gian là 30 m. Các yếu tố địa hình khác, như góc dốc, hướng sườn, độ cong bề mặt, tích tụ dòng chảy, mật độ chia cắt sâu, TWI và SPI, được tính toán từ dữ liệu FABDEM trong môi trường GIS, với độ phân giải không gian 30 m. Bản đồ địa chất khu vực nghiên cứu được thu thập từ Tổng cục Địa chất và Khoáng sản Việt Nam vào năm 2021. Lượng mưa trung bình từ năm 2011 đến năm 2024 được thu thập từ 17 trạm đo mưa của Trung tâm Dự báo Khí tượng Thủy văn Quốc gia Việt Nam. Dữ liệu NDBI được thu thập và tính toán dựa trên ảnh Landsat-8 vào tháng 7 năm 2022. Bản đồ mật độ chia cắt ngang, loại đất và hiện trạng sử dụng đất được cung cấp bởi Bộ Tài nguyên và Môi trường Việt Nam (MONRE) vào năm 2021. Thông tin chi tiết về 15 yếu tố ảnh hưởng đến trượt lở đất được thể hiện chi tiết trong Bảng 2.

Bảng 2. Các yếu tố ảnh hưởng tới trượt lở đất

| STT | Các yếu tố lựa chọn | Nguồn gốc | Độ phân giải không gian |
|---|---|---|---|
| 1 | Độ cao | Mô hình số độ cao DEM | 30 m |
| 2 | Độ dốc | Dẫn xuất từ DEM | 30 m |
| 3 | Chỉ số độ ẩm địa hình (TWI) | Dẫn xuất từ DEM | 30 m |
| 4 | Chỉ số năng lượng dòng chảy (SPI) | Dẫn xuất từ DEM | 30 m |
| 5 | Hướng sườn địa hình | Dẫn xuất từ DEM | 30 m |
| 6 | Độ cong bề mặt địa hình | Dẫn xuất từ DEM | 30 m |
| 7 | Mật độ chia cắt sâu | Dẫn xuất từ DEM | 30 m |
| 8 | Hiện trạng sử dụng đất | Bộ Tài nguyên và Môi trường | 1:50 000 |

| STT | Các yếu tố lựa chọn | Nguồn gốc | Độ phân giải không gian |
|---|---|---|---|
| 9 | Loại đất | Bộ Tài nguyên và Môi trường | 1:50 000 |
| 10 | Địa chất | Tổng cục địa chất và khoáng sản | 1:50 000 |
| 11 | Lượng mưa trung bình năm | Số liệu tại 17 trạm quan trắc lượng mưa trên địa bàn | – |
| 12 | Chỉ số khác biệt chuẩn hóa xây dựng (NDBI) | Ảnh vệ tinh Landsat-8 | 30 m |
| 13 | Mật độ chia cắt ngang | Bản đồ địa hình | 1:50 000 |
| 14 | Khoảng cách đến các sông | Bản đồ địa hình | 1:50 000 |
| 15 | Khoảng cách đến đường giao thông | Bản đồ địa hình | 1:50 000 |

Tất cả các yếu tố ảnh hưởng tới trượt lở đất này đều được chuyển đổi sang định dạng raster với độ phân giải không gian 30 m trong phần mềm ArcGIS Pro (Hình 4).




Hình 4. Bản đồ chuyên đề 15 lớp yếu tố ảnh hưởng tới trượt lở đất.

## 3. Phương pháp sử dụng

### *3.1. Kiểm tra tính đa cộng tuyến của các yếu tố ảnh hưởng*

Trong các nghiên cứu đánh giá nhạy cảm trượt lở đất, rất nhiều biến độc lập (chính là các yếu tố ảnh hưởng) được sử dụng [17]. Hiện tượng đa cộng tuyến giữa các biến độc lập xảy ra khi có sự tương quan mạnh mẽ giữa hai hoặc nhiều biến độc lập trong mô hình hồi quy đa biến, điều này dẫn tới việc

khó xác định chính xác tác động riêng biệt của từng biến độc lập lên mô hình dự báo (biến phụ thuộc) [18]. Kết quả của quá trình này có thể dẫn đến các ước lượng hệ số hồi quy không ổn định và không chính xác, làm sai lệch quá trình diễn giải và dự báo mô hình [4]. Do đó, việc nhận diện hiện tượng đa cộng tuyến giữa các yếu tố ảnh hưởng đến trượt lở đất là bước không thể thiếu nhằm đảm bảo tính tin cậy và độ chính xác của các mô hình dự báo. Trong phần lớn các nghiên cứu về trượt lở đất, hai chỉ số được sử dụng phổ biến nhất để kiểm tra hiện tượng này là Hệ số khuếch đại phương sai (Variance Inflation Factor – VIF) và Hệ số dung sai (Tolerance – TOL) [4]. Khi giá trị VIF vượt ngưỡng 10 hoặc giá trị TOL thấp hơn 0,1, đó là dấu hiệu cho thấy hiện tượng đa cộng tuyến đang tồn tại giữa các biến đầu vào liên quan đến trượt lở đất [4, 17].

### *3.2. Các thuật toán học máy sử dụng*

a. Thuật toán Categorical Boosting (CatBoost)

Thuật toán Categorical Boosting (CatBoost) là một thuật toán tăng cường gradient dựa trên các cây quyết định được xây dựng tuần tự để giảm thiểu lỗi và cải thiện dự đoán [7]. Quá trình này bắt đầu bằng cách xây dựng một cây quyết định và đánh giá mức độ lỗi trong dự đoán [14]. Sau khi cây đầu tiên được xây dựng, cây tiếp theo được tạo ra để sửa lỗi do cây trước đó gây ra. Quá trình này tiếp tục lặp đi lặp lại, với mỗi cây mới tập trung vào việc cải thiện dự đoán của mô hình bằng cách giảm thiểu lỗi trước đó. Quá trình này sẽ dừng lại cho đến khi đạt được số lần lặp được xác định trước [7]. Kết quả là một tập hợp các cây quyết định hoạt động cùng nhau để cung cấp các dự đoán chính xác. Về mặt toán học, nội dung của thuật toán này có thể được tóm lược như sau [14]:

- Cho một tập dữ liệu huấn luyện $D$ bao gồm $N$ mẫu quan sát: $D = \{(x_i, y_i)\}_{i=1}^{N}$, trong đó $x_i$ là vectơ đặc trưng chứa cả các biến liên tục và biến phân loại (địa chất, loại đất,...), và $y_i$ là lớp nhãn (trượt lở, không trượt lở).

- Thuật toán CatBoost chuyển đổi các biến phân loại thành giá trị số bằng phương pháp thống kê mục tiêu có sắp xếp (Ordered Target Statistics). Sau đó, thuật toán này xây dựng một chuỗi các cây quyết định đối xứng (symmetric trees) $h_t(x)$ một cách tuần tự. Tại mỗi một bước lặp $t$, mô hình mới tạo ra sẽ học từ phần dư (residuals) của mô hình trước đó để tối ưu hóa hàm mất mát.

- Bộ phân loại cuối cùng $F(x)$ được tổng hợp bằng cách cộng gộp kết quả của các cây quyết định thành phần $h_t(x)$ với tốc độ học $\eta$, được thể hiện qua phương trình (1):

$$F(x) = \sum_{t=1}^{T} \theta \cdot h_t(x) \tag{1}$$

trong đó, $T$ là tổng số vòng lặp (iterations), $\theta$ là số tốc độ học (learning rate), và $h_t(x)$ là kết quả dự báo của cây quyết định đối xứng tại bước lặp thứ $t$.

Thuật toán CatBoost đặc biệt phù hợp với các tập dữ liệu quy mô lớn với nhiều đặc trưng độc lập. Không giống như các thuật toán tăng cường gradient khác, thuật toán CatBoost được thiết kế đặc biệt để xử lý cả các biến liên tục và biến phân loại một cách liền mạch mà không cần mã hóa đặc trưng thủ công [7].

b. Thuật toán Random Forest (RF)

Thuật toán Random Forest (RF) được giới thiệu lần đầu tiên bởi Breiman vào năm 2001 [19]. Thuật toán này sử dụng nhiều cây quyết định để đưa ra dự đoán tốt hơn [11, 20]. Mỗi cây xem xét các phần ngẫu nhiên khác nhau của dữ liệu và kết quả của chúng được kết hợp bằng cách bỏ phiếu cho phân loại hoặc tính trung bình cho hồi quy, điều này làm cho nó trở thành một kỹ thuật học tập kết hợp, giúp cải thiện độ chính xác và giảm lỗi trong dự đoán [10]. Nội dung của thuật toán này có thể được tóm tắt như sau [19]:

- Dữ liệu đầu vào: Cho một tập dữ liệu huấn luyện $D$ bao gồm $N$ mẫu quan sát: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, trong đó $x$ là vectơ đặc trưng các yếu tố ảnh hưởng đến trượt lở và $y$ là lớp nhãn (trượt lở, không trượt lở).

- Tạo mẫu và xây dựng cây: Thuật toán RF sử dụng kĩ thuật lấy mẫu có hoàn lại để tạo ra $K$ tập mẫu bootstrap từ tập dữ liệu gốc $D$ ban đầu. Tuy nhiên, trong quá trình xây dựng mỗi cây quyết định $h_k(x)$, tại mỗi nút phân chia, thuật toán này chỉ chọn ngẫu nhiên một tập hợp con các đặc trưng ($m$ đặc trưng) từ tổng số $M$ đặc trưng đầu vào ($m < M$) thay vì xem xét toàn bộ.

- Tổng hợp kết quả: Bộ phân loại cuối cùng $C^*$ được tổng hợp bằng cách kết hợp kết quả dự báo của các cây quyết định thành phần $h_1, h_2, \ldots, h_K$ thông qua cơ chế bầu chọn đa số (majority voting), được thể hiện qua phương trình (2):

$$C^*(x) = \arg\max \sum_{k=1}^{K} I(h_k(x) = y) \tag{2}$$

trong đó, $I(\cdot)$ là hàm chỉ thị (indicator function), hàm bằng 1 nếu cây quyết định thứ $k$ dự báo lớp $y$ và 0 nếu ngược lại.

c. Quá trình lựa chọn siêu tham số

Trong nghiên cứu này, quy trình lựa chọn siêu tham số (hyperparameters) cho các mô hình CatBoost và Random Forest được tối ưu bằng phương pháp `RandomizedSearchCV` thông qua thử nghiệm ngẫu nhiên nhiều tổ hợp tham số kết hợp với kiểm định chéo (cross-validation). Đối với mô hình CatBoost, các siêu tham số được tối ưu bao gồm số vòng lặp (`iterations`), tốc độ học (`learning_rate`), độ sâu cây (`depth`), hệ số regularization L2 (`l2_leaf_reg`) và số ngưỡng phân chia của biến liên tục (`border_count`). Quá trình tối ưu hóa tham số được thực hiện bằng `RandomizedSearchCV` với số tổ hợp tham số thử nghiệm `n_iter` = 50, kết hợp kiểm định chéo 10-fold (`cv` = 10), trong đó bộ dữ liệu được chia thành 10 phần bằng nhau; mỗi lần lặp sử dụng 9 phần để huấn luyện và 1 phần để kiểm định mô hình, cùng với giá trị `random_state` = 42 (Bảng 3).

Đối với mô hình Random Forest, các siêu tham số được tối ưu bao gồm số lượng cây trong rừng (`n_estimators`), độ sâu tối đa của cây (`max_depth`), số lượng biến được xem xét tại mỗi lần phân tách (`max_features`), số mẫu tối thiểu để tách nút (`min_samples_split`), số mẫu tối thiểu tại nút lá (`min_samples_leaf`), phương pháp lấy mẫu bootstrap và tiêu chí đánh giá chất lượng phân tách nút (`criterion`). Quá trình tối ưu hóa tham số của Random Forest cũng được thực hiện bằng `RandomizedSearchCV` với số tổ hợp tham số thử nghiệm `n_iter` = 50, kết hợp kiểm định chéo 10-fold (`cv` = 10), trong đó dữ liệu được chia thành 10 phần để đánh giá tính ổn định và khả năng khái quát hóa của mô hình, cùng với giá trị `random_state` = 42 (Bảng 3).

Bảng 3. Các siêu tham số tối ưu trong thuật toán CatBoost và RF

| Các siêu tham số | Mô hình CatBoost | Mô hình RF |
|---|---|---|
| Tốc độ học (`learning_rate`) | 0,01 | – |
| Số vòng lặp (`iterations`) | 400 | – |
| Độ sâu cây (`depth`) | 3 | – |
| Hệ số `l2_leaf_reg` | 0,01 | – |
| Số ngưỡng phân chia của biến liên tục (`border_count`) | 32 | – |
| Số lượng cây trong rừng (`n_estimators`) | – | 100 |

| Các siêu tham số | Mô hình CatBoost | Mô hình RF |
|---|---|---|
| Độ sâu tối đa của cây (`max_depth`) | – | `None` |
| Số lượng biến được xem xét tại mỗi lần phân tách (`max_features`) | – | 0,5 |
| Số mẫu tối thiểu để tách nút (`min_samples_split`) | – | 6 |
| Số mẫu tối thiểu tại nút lá (`min_samples_leaf`) | – | 9 |
| Phương pháp lấy mẫu bootstrap | – | `True` |
| Chất lượng phân tách nút (`criterion`) | – | `gini` |
| `random_state` | 42 | 42 |

### *3.3. Kiểm chứng độ chính xác dự báo của các thuật toán học máy*

Cần phải đánh giá hiệu suất của các mô hình dự báo trượt lở đất nhằm đảm bảo độ tin cậy cũng như khả năng áp dụng của chúng trong thực tế [7]. Trong nghiên cứu này, các chỉ số thống kê phổ biến đã được sử dụng để đánh giá hiệu suất dự báo của mô hình học máy, bao gồm: Độ bao phủ (Recall) độ nhạy cảm (Sensitivity), Độ chính xác dự đoán (Precision), Độ chính xác tổng thể (Accuracy), Giá trị F - F-measure, chỉ số Jaccard, Sai số trung phương - Root Mean Squared Error (RMSE), đường cong ghi nhận tín hiệu ROC - Reciever Operating Characteristic (ROC) curve, và diện tích dưới đường cong ROC - Area Under the ROC Curve (AUC). Những chỉ số này được tính theo những công thức từ (3) đến (8) như sau:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F-measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{TP} + \text{FP}} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (X_{\text{Act.}} - X_{\text{Pred.}})^2}{N}} \tag{8}$$

trong đó TP (True Positive) là số pixel trượt lở đất được phân loại đúng là trượt lở, TN (True Negative) là số pixel không trượt lở đất được phân loại đúng là không trượt lở, FP (False Positive) là số pixel không trượt lở đất nhưng bị mô hình phân loại nhầm thành trượt lở, FN (False Negative) là số pixel trượt lở đất nhưng bị mô hình phân loại nhầm thành không trượt lở, $N$ biểu thị tổng số trường hợp trong tập dữ liệu đào tạo/kiểm tra, $X_{\text{Act.}}$ là giá trị thực của trường hợp thứ $i$ và $X_{\text{Pred.}}$ biểu thị giá trị dự đoán của trường hợp thứ $i$.

Đường cong đặc tính hoạt động của bộ phân loại (Receiver Operating Characteristic – ROC) là đồ thị biểu diễn mối quan hệ giữa tỷ lệ dương tính thật (True Positive Rate – TPR) và tỷ lệ dương tính giả (False Positive Rate – FPR) trên toàn bộ dải ngưỡng xác suất phân loại của mô hình. Do đó, ROC cho phép đánh giá tổng quát khả năng phân biệt giữa hai lớp xảy ra trượt lở đất và không xảy ra trượt lở đất [4]. Bên cạnh đó, diện tích dưới đường cong ROC (AUC) là một thước đo định lượng quan trọng, giúp xác định mức độ tin cậy của mô hình dự báo. Nếu giá trị AUC dao động trong khoảng

[0,5 – 0,6] phản ánh mô hình có hiệu suất dự báo rất kém; AUC dao động trong khoảng [0,6 – 0,7] tương ứng với mô hình dự báo kém; AUC dao động trong khoảng [0,7 – 0,8] cho thấy mô hình dự báo ở mức trung bình; AUC dao động từ [0,8 – 0,9] thể hiện mô hình có khả năng dự báo ở mức tốt; và AUC dao động trong khoảng [0,9 – 1,0] thì mô hình dự báo ở mức xuất sắc [16]. Trong nghiên cứu này, ngưỡng phân loại được sử dụng là 0,5 cho cả hai mô hình CatBoost và Random Forest. Theo đó, các mẫu có xác suất dự báo ≥ 0,5 được phân loại là "trượt lở đất", trong khi các mẫu có xác suất < 0,5 được phân loại là "không trượt lở đất".

### *3.4. Phân tích tầm quan trọng của các yếu tố đầu vào*

Phân tích tầm quan trọng của các yếu tố đầu vào giúp xác định các biến có tác động tới kết quả dự báo cũng như nâng cao tính chính xác, minh bạch của mô hình được sử dụng [21]. Hiện nay, kỹ thuật SHAP (SHapley Additive exPlanations) mang đến một cách tiếp cận mang tính toán học chặt chẽ, dựa trên lý thuyết giá trị Shapley trong lý thuyết trò chơi, giúp định lượng mức độ ảnh hưởng của từng yếu tố đầu vào tới kết quả dự báo của mô hình [21]. Theo đó, SHAP tính toán giá trị đóng góp của mỗi biến trong mô hình dựa trên mức trung bình có trọng số của tất cả các khả năng loại bỏ hoặc thêm biến đó vào mô hình [22]. Trên cơ sở đó, giá trị SHAP của một biến đầu vào $j$ được tính dựa trên giá trị trung bình của tất cả các kịch bản có thể xảy ra khi biến $j$ có mặt hoặc không có mặt trong mô hình theo công thức (9) sau:

$$\Phi_j = \sum_{S \subseteq P \setminus \{j\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} [f(S \cup \{j\}) - f(S)] \tag{9}$$

trong đó $\Phi_j$ là đóng góp của yếu tố cuối cùng; $S$ là tập con các đặc trưng ngoại trừ biến $j$; $P$ đại diện cho tập hợp tất cả các đặc trưng; $f(S)$ là dự đoán của mô hình sử dụng tập con $S$; $f(S \cup \{j\})$ biểu thị dự đoán khi đặc trưng $j$ được thêm vào tập con.

## 4. Kết quả và thảo luận

### *4.1. Kiểm tra tính đa cộng tuyến của các yếu tố ảnh hưởng*

Trong nghiên cứu này, hệ số phương sai (VIF) và dung sai (TOL) được sử dụng để kiểm tra tính đa cộng tuyến của 15 yếu tố ảnh hưởng tới trượt lở đất. Kết quả tính toán cho thấy tất cả các yếu tố này đều có VIF < 10 và dung sai TOL > 0,1, do đó, 15 yếu tố ảnh hưởng tới trượt lở đất đều được sử dụng để xây dựng mô hình dự báo trượt lở (Bảng 4).

Bảng 4. Kết quả kiểm tra đa cộng tuyến của 15 yếu tố ảnh hưởng tới TLĐ

| STT | Các yếu tố ảnh hưởng | VIF | TOL |
|---|---|---|---|
| 1 | Độ dốc | 2,30 | 0,435 |
| 2 | Chỉ số ẩm địa hình (TWI) | 2,22 | 0,450 |
| 3 | Độ cao | 2,01 | 0,498 |
| 4 | Mật độ chia cắt ngang | 1,70 | 0,588 |
| 5 | Khoảng cách đến sông | 1,68 | 0,595 |
| 6 | Hiện trạng sử dụng đất | 1,52 | 0,658 |
| 7 | Mật độ chia cắt sâu | 1,49 | 0,671 |
| 8 | Độ cong bề mặt | 1,47 | 0,680 |
| 9 | Chỉ số khác biệt chuẩn hóa xây dựng (NDBI) | 1,39 | 0,719 |
| 10 | Khoảng cách đến đường | 1,23 | 0,813 |

| STT | Các yếu tố ảnh hưởng | VIF | TOL |
|---|---|---|---|
| 11 | Loại đất | 1,16 | 0,862 |
| 12 | Địa chất | 1,14 | 0,877 |
| 13 | Lượng mưa | 1,11 | 0,901 |
| 14 | Chỉ số công suất dòng chảy (SPI) | 1,08 | 0,926 |
| 15 | Hướng sườn | 1,05 | 0,952 |

*4.2. So sánh hiệu suất của các mô hình dự báo trượt lở*

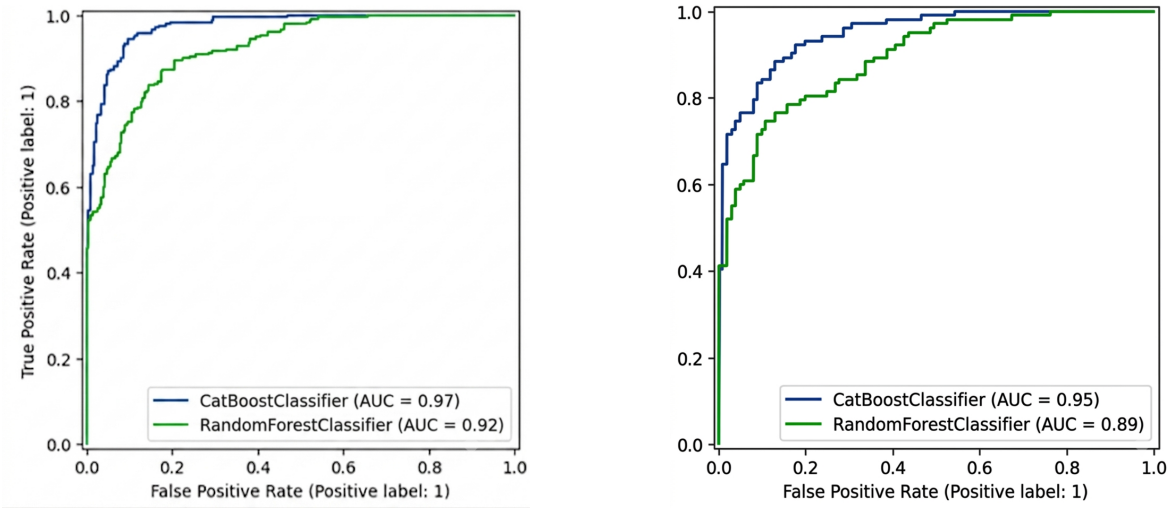
Nghiên cứu này sử dụng 02 mô hình học máy, bao gồm CatBoost và Random Forest, để xây dựng bản đồ nhạy cảm TLĐ cho khu vực tỉnh Thái Nguyên (cũ). Việc so sánh và đánh giá hiệu suất dự báo của các mô hình này được thực hiện thông qua các chỉ số thống kê, ma trận nhầm lẫn của từng mô hình, và giá trị AUC của đường cong ghi nhận tín hiệu ROC (Bảng 5).

Bảng 5. Các chỉ số thống kê dùng trong đánh giá hiệu suất mô hình dự báo

| Các chỉ số thống kê | Mô hình CatBoost | | Mô hình Random Forest | |
|---|---|---|---|---|
| | Dữ liệu đào tạo | Dữ liệu kiểm tra | Dữ liệu đào tạo | Dữ liệu kiểm tra |
| TN | 216 | 93 | 202 | 91 |
| FP | 20 | 8 | 34 | 10 |
| FN | 22 | 21 | 40 | 28 |
| TP | 215 | 81 | 197 | 74 |
| Accuracy (%) | 91,12 | 85,71 | 84,36 | 81,28 |
| Precision (%) | 91,49 | 91,01 | 85,28 | 88,10 |
| Recall (%) | 90,72 | 79,41 | 83,12 | 72,55 |
| F1-Score (%) | 91,10 | 84,82 | 84,19 | 79,57 |
| RMSE | 0,298 | 0,378 | 0,396 | 0,433 |

Đối với tập dữ liệu đào tạo, mô hình CatBoost đạt hiệu suất vượt trội hơn hẳn so với mô hình Random Forest trên tất cả các chỉ số đánh giá. Cụ thể, CatBoost có độ chính xác cao nhất (91,12%), độ Precision (91,49%), độ Recall (90,72%), chỉ số F1-Score (91,10%), và RMSE đạt 0,298, trong khi Random Forest chỉ đạt lần lượt 84,36%, 85,28%, 83,12%, 84,19%, và 0,396. Kết quả trên tập dữ liệu huấn luyện cũng cho thấy mô hình CatBoost có khả năng học và phân loại tốt hơn, với tỷ lệ phân loại sai thấp hơn đáng kể (FP = 20, FN = 22) so với Random Forest (FP = 34, FN = 40). Tương tự như vậy, đối với tập dữ liệu kiểm tra, mô hình CatBoost tiếp tục khẳng định ưu thế với độ chính xác đạt 85,71%, độ Precision (91,01%), độ Recall (79,41%), chỉ số F1-Score (84,19%), và RMSE đạt 0,378, trong khi mô hình Random Forest chỉ đạt lần lượt 81,28%, 88,10%, 72,55%, 79,57%, và 0,433.

Kết quả đánh giá đường cong ROC trên Hình 5 cho thấy sự khác biệt rõ rệt về khả năng phân loại giữa hai mô hình CatBoost và Random Forest trên cả tập dữ liệu đào tạo và kiểm tra. Đối với mô hình CatBoost, giá trị AUC đạt 0,97 trên tập dữ liệu đào tạo và 0,95 trên tập dữ liệu kiểm tra. Đối với mô hình Random Forest, giá trị AUC đạt 0,92 trên tập dữ liệu đào tạo và giảm xuống còn 0,89 trên tập dữ liệu kiểm tra. Đây đều là những giá trị AUC rất cao, phản ánh khả năng dự báo xuất sắc của cả hai mô hình học máy được sử dụng. Tuy nhiên, mô hình CatBoost vẫn có hiệu suất dự báo vượt trội hơn hẳn so với mô hình Random Forest trên cả hai tập dữ liệu đào tạo và kiểm tra, với AUC cao hơn lần lượt 0,05 điểm trên tập dữ liệu đào tạo (Hình 5(a)) và 0,06 điểm trên tập dữ liệu kiểm tra (Hình 5(b)), tương ứng với các giá trị là 0,97 và 0,95.




(a) Tập dữ liệu đào tạo (b) Tập dữ liệu kiểm tra

Hình 5. Kết quả phân tích đường cong ROC

### *4.3. Lựa chọn và xây dựng bản đồ nhạy cảm trượt lở đất tốt nhất*

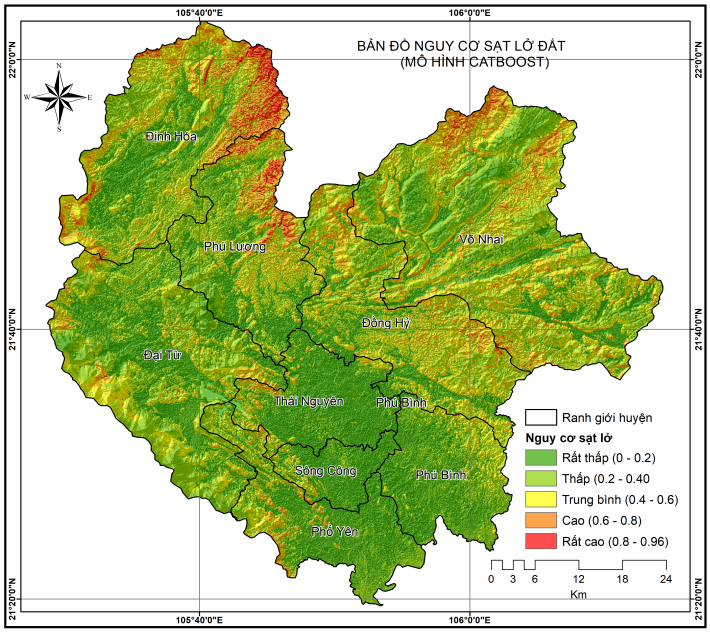
Mô hình CatBoost là mô hình có độ chính xác dự báo tốt nhất trong nghiên cứu này, do đó, nó được sử dụng để xây dựng bản đồ nhạy cảm trượt lở đất cho tỉnh Thái Nguyên (cũ). Theo đó, bản đồ nhạy cảm trượt lở đất tỉnh Thái Nguyên được phân loại thành năm cấp bằng kỹ thuật Quantile trong phần mềm ArcGIS Pro, bao gồm rất cao, cao, trung bình, thấp và rất thấp (Hình 6). Kết quả thống kê trên bản đồ nhạy cảm trượt lở đất theo mô hình CatBoost cho thấy 130666,3 ha (37,12%) nằm trong các khu vực nhạy cảm rất thấp; 111292,6 ha (31,62%) nằm trong các khu vực nhạy cảm thấp, 77442,93 ha (22,00%) nằm trong các khu vực nhạy cảm trung bình, 28884,06 ha (8,21%) nằm trong các khu vực nhạy cảm cao, và 3727,62 ha (1,06%) nằm trong các khu vực nhạy cảm rất cao (Hình 6).

Bản đồ nhạy cảm trượt lở đất theo mô hình CatBoost cho thấy các khu vực nằm trong vùng nhạy cảm trượt lở cao và rất cao tập trung chủ yếu ở phía Bắc, Tây Bắc và Đông Bắc của tỉnh Thái Nguyên, bao gồm các huyện Định Hoá, Võ Nhai, Đồng Hỷ, Phú Lương và Đại Từ, đây là những nơi có địa hình đồi núi hiểm trở, độ dốc cao. Ngược lại, các huyện nằm ở phía Nam của tỉnh Thái Nguyên, như Phổ Yên, Sông Công, Phú Bình và khu vực thành phố Thái Nguyên, chủ yếu nằm trong vùng nhạy cảm trượt lở rất thấp đến thấp, do địa hình bằng phẳng hơn. Nhận định này cũng phù hợp với những kết quả đã được công bố trước đây của nhóm Hà và cs. [16].

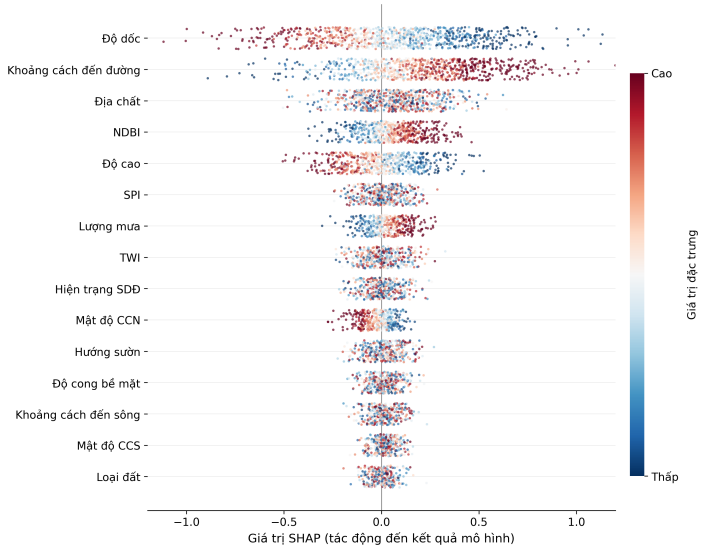
### *4.4. Phân tích tầm quan trọng của các yếu tố đầu vào*

Trong nghiên cứu này, kỹ thuật SHAP được sử dụng để định lượng mức độ ảnh hưởng của từng yếu tố đầu vào đối với bản đồ nhạy cảm trượt lở đất tốt nhất theo mô hình CatBoost. Thang màu biểu thị giá trị SHAP cho thấy những yếu tố nào tác động tích cực tới TLĐ (màu đỏ) và những yếu tố nào có tác động tiêu cực tới TLĐ (màu xanh). Trong nghiên cứu này, giá trị SHAP theo thứ tự giảm dần về mức độ ảnh hưởng của 15 yếu tố đầu vào tới TLĐ thể hiện như sau: Độ dốc > Khoảng cách đến đường > Địa chất > NDBI > Độ cao > SPI > Lượng mưa > TWI > Hiện trạng sử dụng đất > Mật độ chia cắt ngang > Hướng sườn > Độ cong bề mặt > Khoảng cách đến sông > Mật độ chia cắt sâu > Loại đất (Hình 7).

Trong số các biến số ảnh hưởng đến mô hình dự báo TLĐ trong khu vực nghiên cứu thì độ dốc, khoảng cách đến đường, địa chất, NDBI, và độ cao là những yếu tố có tác động lớn nhất, trong khi




Hình 6. Bản đồ nhạy cảm trượt lở đất tỉnh Thái Nguyên (cũ) theo mô hình CatBoost



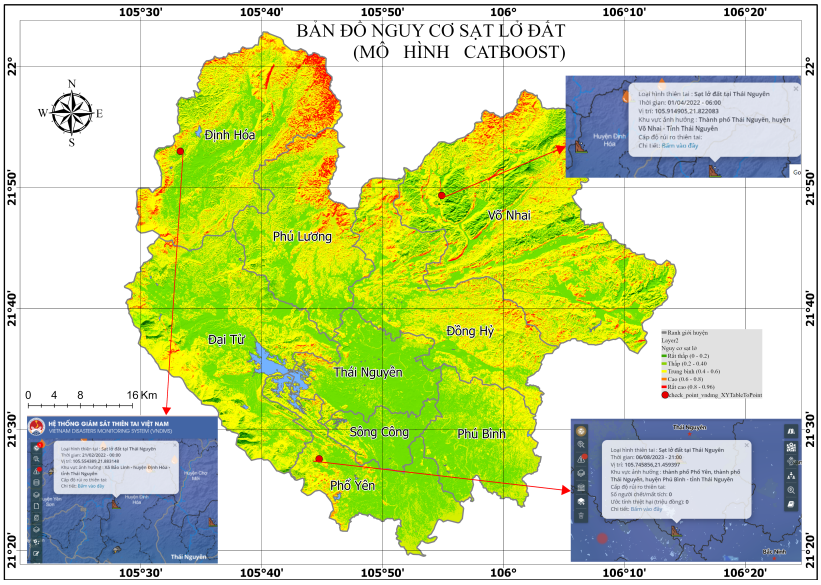

Hình 7. Giá trị SHAP của 15 yếu tố ảnh hưởng đến TLĐ trong mô hình CatBoost

đó, hướng sườn, độ cong bề mặt, khoảng cách đến sông, mật độ chia cắt ngang, và loại đất là những yếu tố có tác động ít nhất. Với địa hình chủ đạo là đồi núi thấp, chiếm khoảng 2/3 diện tích toàn khu vực, cùng với tiến trình đô thị hóa đang diễn ra mạnh mẽ thông qua những tuyến đường mới mở, mật độ xây dựng tăng cao và các hoạt động khai thác khoáng sản phục vụ phát triển kinh tế địa phương có thể được xem là những nguyên nhân chính hình thành nên TLĐ trên địa bàn tỉnh Thái Nguyên (cũ) trong mùa mưa.

### *4.5. Kiểm chứng bản đồ nhạy cảm trượt lở đất tốt nhất*

Nghiên cứu này đã sử dụng dữ liệu về các vụ trượt lở gần đây nhất từ trang web Hệ thống giám sát thiên tai Việt Nam (http://vndms.dmptc.gov.vn/) để kiểm chứng bản đồ nhạy cảm trượt lở đất tốt

nhất theo mô hình CatBoost. Qua đối soát, vị trí trượt lở đất dự đoán tại các khu vực nhạy cảm cao đã được xác nhận vào ngày 01 tháng 04 năm 2022 tại huyện Võ Nhai với tọa độ ở 21°49′19,50″ vĩ độ Bắc và 105°54′53,66″ kinh độ Đông; tiếp theo là ngày 21 tháng 02 năm 2022 tại huyện Định Hoá ở toạ độ 21°52′59,33″ vĩ độ Bắc và 105°33′15,80″ kinh độ Đông; ngày 06 tháng 8 năm 2023 tại huyện Phú Bình với tọa độ 21°27′33,83″ vĩ độ Bắc và 105°44′45,08″ kinh độ Đông (Hình 8).




Hình 8. Kiểm chứng bản đồ nhạy cảm trượt lở đất theo mô hình CatBoost

Kết quả đối sánh bước đầu cho thấy phần lớn các vị trí trượt lở đất mới cập nhật nằm trong các khu vực có mức nhạy cảm cao và rất cao theo mô hình CatBoost, qua đó cung cấp thêm minh chứng hỗ trợ về tính hợp lý không gian của bản đồ nhạy cảm trượt lở đất xây dựng được trong nghiên cứu này.

## 5. Kết luận

Nghiên cứu này đề xuất sử dụng hai mô hình học máy tiên tiến để xây dựng bản đồ nhạy cảm trượt lở đất cho tỉnh Thái Nguyên (cũ), bao gồm CatBoost và Random Forest (RF). Dữ liệu đầu vào bao gồm bộ dữ liệu hiện trạng trượt lở đất trong quá khứ và 15 yếu tố ảnh hưởng tới trượt lở đất, bao gồm độ cao, độ dốc, hướng sườn, độ cong bề mặt, mật độ chia cắt ngang, mật độ chia cắt sâu, khoảng cách đến sông suối, khoảng cách đến đường giao thông, Chỉ số độ ẩm địa hình (TWI), Chỉ số năng lượng dòng chảy (SPI), Chỉ số khác biệt chuẩn hóa xây dựng (NDBI), loại đất, lớp phủ bề mặt, địa chất và lượng mưa trung bình. Độ chính xác và độ tin cậy của các mô hình học máy được đánh giá bằng cách so sánh các chỉ số thống kê tiêu chuẩn khác nhau và giá trị diện tích dưới đường cong AUC. Kết quả đạt được cho thấy mô hình CatBoost có hiệu suất dự báo vượt trội hơn hẳn so với mô hình RF với giá trị AUC = 0,950 theo phân tích đường cong ROC trên tập dữ liệu kiểm tra. Trên cơ sở đó, mô hình CatBoost được lựa chọn để xây dựng bản đồ nhạy cảm trượt lở đất tốt nhất cho khu vực nghiên cứu. Bản đồ nhạy cảm trượt lở đất cung cấp các thông tin không gian về các khu vực có nguy cơ trong tương lai, làm cơ sở khoa học cho việc đề xuất các giải pháp quy hoạch vùng, sử dụng đất, cũng như định hướng phát triển kinh tế địa phương một cách lâu dài, bền vững. Ngoài ra, nghiên cứu này còn sử dụng các vị trí trượt lở đất mới nhất từ trang web của Hệ thống giám sát thiên tai Việt Nam để đối sánh và kiểm chứng độ tin cậy của bản đồ nhạy cảm trượt lở đất theo mô hình CatBoost. Những kết quả đạt được này không chỉ khích lệ việc thử nghiệm các mô hình học máy tiên tiến khác nhau trong

đánh giá nhạy cảm trượt lở đất nhằm đạt được các mô hình dự báo có độ chính xác cao hơn, mà còn thúc đẩy việc tạo dựng một thư viện học máy và cách tiếp cận, đánh giá đầy đủ hơn, phong phú hơn, phục vụ quá trình quản lý và giảm thiểu rủi ro đa thiên tai trong tương lai.

**Lời cảm ơn**

Nghiên cứu này được tài trợ bởi Bộ Giáo dục và Đào tạo trong đề tài mã số B2024-XDA-06.

**Tuyên bố về việc sử dụng trí tuệ nhân tạo tạo sinh và các công nghệ hỗ trợ bởi AI**

Các tác giả xác nhận không sử dụng bất kỳ công cụ trí tuệ nhân tạo tạo sinh hoặc công nghệ hỗ trợ bởi AI nào trong quá trình chuẩn bị và hoàn thiện bản thảo.

**Tài liệu tham khảo**

[1] Rong, G., Alu, S., Li, K., Su, Y., Zhang, J., Zhang, Y., Li, T. (2020). Rainfall-induced landslide susceptibility mapping based on Bayesian optimized random forest and gradient boosting decision tree models - A case study of Shuicheng County, China. *Water*, 12(11):3066.

[2] Hằng, H. T., Đông, K. T., Thao, L. V., Phương, N. T., Thảo, Đ. T. P. (2023). Nghiên cứu ứng dụng phương pháp tỷ số tần suất kết hợp GIS trong xây dựng bản đồ nguy cơ trượt lở đất huyện Pác Nặm - tỉnh Bắc Kạn. *Tạp chí Khoa học Công nghệ Xây dựng (TCKHCNXD) - ĐHXDHN*, 17(1V):75–90.

[3] Arabameri, A., Saha, S., Roy, J., Chen, W., Blaschke, T., Bui, D. T. (2020). Landslide susceptibility evaluation and management using different machine learning methods in the Gallicash River watershed, Iran. *Remote Sensing*, 12(3):475.

[4] Bui, Q. D., Ha, H., Khuc, D. T., Nguyen, D. Q., von Meding, J., Nguyen, L. P., Luu, C. (2023). Landslide susceptibility prediction mapping with advanced ensemble models: Son La province, Vietnam. *Natural Hazards*, 116(2):2283–2309.

[5] Youssef, A. M., Pourghasemi, H. R. (2021). Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geoscience Frontiers*, 12(2):101074.

[6] Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, 13(5):839–856.

[7] Sahin, E. K. (2022). Comparative analysis of gradient boosting algorithms for landslide susceptibility mapping. *Geocarto International*, 37(9):2441–2465.

[8] Kim, J. C., Lee, S., Jung, H. S., Lee, S. (2018). Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. *Geocarto International*, 33(9):1000–1015.

[9] Park, S., Kim, J. (2019). Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. *Applied Sciences*, 9(5):942.

[10] Sun, D., Wen, H., Wang, D., Xu, J. (2020). A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. *Geomorphology*, 362:107201.

[11] Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, 2 (7):1308.

[12] Zhang, W., He, Y., Wang, L., Liu, S., Meng, X. (2023). Landslide susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing. *Geological Journal*, 58(6): 2372–2387.

[13] Bien, T. X., Nguyen, Q. H., Pham, T. T., Nguyen, T. A., Bui, X. K., Prakash, I., Bui, D. T. (2023). Integration of rotation forest and multiboost ensemble methods with forest by penalizing attributes for spatial prediction of landslide susceptible areas. *Stochastic Environmental Research and Risk Assessment*, 37(12):4641–4660.

[14] Ha, H., Bui, Q. D., Tran, D. T., Nguyen, D. Q., Bui, H. X., Luu, C. (2025). Improving the forecast performance of landslide susceptibility mapping by using ensemble gradient boosting algorithms. *Environment, Development and Sustainability*, 27(8):18409–18443.

[15] Le, T. T. T., Kawagoe, S. (2018). Impact of the landslide for a relationship between rainfall condition and land cover in North Vietnam. *Journal of Geological Resource and Engineering*, 6:239–250.

[16] Ha, H., Bui, Q. D., Vu, D. C., Hoang, T., Pham, D. H., Nguyen, V. P., Do, M. H., Luu, C. (2024). Landslide spatial prediction utilizing fuzzy unordered rules induction ensemble models: A case study in Thai Nguyen, Vietnam. *Modeling Earth Systems and Environment*, 10(2):2201–2219.

[17] Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2):39–42.

[18] Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, 949 (1):012009.

[19] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[20] Chinh, L. T. D., Hang, H. T. (2023). An analysis of the relative variable importance to flood fatality using a machine learning approach. *Journal of Science and Technology in Civil Engineering (JSTCE) - HUCE*, 17(1):125–136.

[21] Luu, C., Ha, H., Tran, X. T., Vu, T. H., Bui, Q. D. (2024). Landslide susceptibility and building exposure assessment using machine learning models and geospatial analysis techniques. *Advances in Space Research*, 74(11):5489–5513.

[22] Lee, Y. G., Oh, J. Y., Kim, D., Kim, G. (2023). SHAP value-based feature importance analysis for short-term load forecasting. *Journal of Electrical Engineering & Technology*, 18(1):579–588.