MÔ HÌNH KẾT HỢP SMOTE-MLP TRONG VIỆC NÂNG CAO ĐỘ CHÍNH XÁC DỰ ĐOÁN XÁC SUẤT KHÔNG GIAN SẠT LỞ ĐẤT TỪ DỮ LIỆU MẤT CÂN BẰNG TẠI VÙNG NÚI TỈNH QUẢNG NAM

Nguyễn Bá Quang Vinh[1,2]

Tóm tắt: *Sự mất cân bằng trong dữ liệu sạt lở đất, khi số lượng điểm sạt lở thực tế rất hạn chế so với số điểm không sạt lở, gây ảnh hưởng lớn đến hiệu quả của các mô hình học máy trong dự báo xác suất không gian sạt lở đất. Nghiên cứu này đề xuất áp dụng kỹ thuật tăng cường mẫu thiểu số tổng hợp (SMOTE) để nội suy và tăng cường số lượng điểm sạt lở, kết hợp với mô hình mạng perceptron nhiều lớp (MLP) nhằm xây dựng bản đồ xác suất không gian sạt lở đất tại vùng núi tỉnh Quảng Nam. Tập dữ liệu ban đầu bao gồm 500 điểm sạt lở, được tăng cường bằng SMOTE với các mức lần lượt là 2,000; 5,000; 10,000; 20,000 và 50,000 điểm nội suy. Kết quả cho thấy AUC (Area Under the ROC Curve) trên tập xác thực tăng từ 0.846 (với 500 điểm gốc) lên 0.887 khi sử dụng 10,000 điểm nội suy, cho thấy hiệu quả rõ rệt của kỹ thuật SMOTE trong cải thiện độ chính xác mô hình. Tuy nhiên, khi tăng số lượng điểm nội suy lên 20,000 và 50,000, AUC có xu hướng giảm nhẹ xuống còn 0.868 và 0.866, cho thấy dấu hiệu suy giảm khả năng khái quát hóa. Do đó, lựa chọn số lượng điểm nội suy hợp lý đóng vai trò quan trọng nhằm cân bằng giữa cải thiện hiệu suất và tránh overfitting. Nghiên cứu này khẳng định tiềm năng của việc kết hợp SMOTE và MLP trong xây dựng bản đồ xác suất không gian sạt lở đất từ tập dữ liệu mất cân bằng.*

Từ khóa: Sạt lở đất, mạng perceptron nhiều lớp, mất cân bằng dữ liệu, tăng cường mẫu thiểu số tổng hợp.

1. GIỚI THIỆU

Đánh giá xác suất không gian sạt lở đất (Landslide spatial probability-LSP) là một phương pháp có giá trị trong việc xác định các khu vực có khả năng xảy ra sạt lở đất (Van Westen, 2000). Trong những thập kỷ gần đây, sự phát triển của công nghệ đã thúc đẩy sự ra đời của nhiều phương pháp khác nhau để dự đoán LSP, trong đó các kỹ thuật học máy (Machine learning-ML) đã được ứng dụng rộng rãi trong mô hình hóa xác suất nhằm dự báo nguy cơ sạt lở đất, với các thuật toán điển hình như Naïve Bayes (NB) (Nguyen et al. 2024; Nguyen and Kim, 2021; Nhu et al. 2020), cây quyết định (Decision trees-DT) (Hong et al. 2018; Pradhan, 2013), máy vector hỗ trợ (Support vector machines-SVM) (Nguyen and Kim, 2021; Pradhan, 2013), và mạng nơ-ron nhân tạo (Artificial neural networks-ANN) (Nguyen and Kim, 2021; Bragagnolo and Grzybowski, 2020). Nhiều nghiên cứu gần đây đã khẳng định tính hiệu quả của các phương pháp học máy trong dự đoán xác suất không gian sạt lở đất (Nguyen et al. 2024; Nguyen and Kim, 2021; Ali et al. 2021). Tuy nhiên, việc lựa chọn mô hình học máy phù hợp vẫn là một thách thức quan trọng nhằm đảm bảo độ chính xác cao trong dự báo. Bên cạnh đó, trong các nghiên cứu đự đoán LSP bằng học máy, một trong những thách thức lớn là sự mất cân bằng nghiêm trọng trong phân bố dữ liệu, khi số lượng điểm xảy ra sạt lở thực tế thu thập được thường rất ít so với số điểm không sạt lở. Tình trạng này ảnh hưởng trực tiếp đến hiệu suất của các mô hình phân loại, đặc biệt làm giảm độ nhạy trong việc nhận diện các vùng nguy cơ cao.

Việt Nam nằm trong khu vực nhiệt đới và chịu ảnh hưởng của nhiều kiểu khí hậu như cận nhiệt đới ẩm, xavan nhiệt đới và gió mùa nhiệt đới. Mỗi năm, Việt Nam hứng chịu hơn 10 cơn bão nhiệt đới cùng với nhiều đợt mưa lớn, đặc biệt ảnh hưởng nghiêm trọng đến các tỉnh miền Trung. Địa hình Việt Nam chủ yếu là đồi núi, chiếm khoảng 3/4 diện tích cả nước; trong đó, 85% có độ cao dưới 1,000 m và chỉ khoảng 1% vượt quá 2,000 m. Sạt lở đất là hiện tượng thiên tai tái diễn thường xuyên, gây thiệt hại lớn về người với trung bình hơn 100 ca tử vong mỗi năm. Theo đề án "Điều tra, đánh giá và phân vùng cảnh báo nguy cơ trượt lở đất đá các vùng núi Việt Nam" của Viện Khoa học địa chất và khoáng sản (Bộ TN&MT), từ năm 2012 đến 2020 đã ghi nhận 13,233 vị trí sạt lở trên địa bàn 21 tỉnh, trong đó Quảng Nam, Quảng Trị và Thừa Thiên Huế được xác định là các khu vực có nguy cơ cao với mật độ sạt lở dày đặc. Thực trạng đáng lo ngại này cho thấy sự cần thiết cấp bách của các nghiên cứu chuyên sâu về đánh giá nguy cơ sạt lở đất và xây dựng chiến lược giảm thiểu rủi ro hiệu quả, đặc biệt tại khu vực miền Trung Việt Nam. Nhiều nghiên cứu trong nước đã sử dụng các mô hình trọng số dẫn chứng (WoE), chỉ số thống kê (SI), phân tích thứ bậc (AHP) và các phương pháp phân vùng nguy cơ để xây dựng bản đồ sạt lở (Đỗ et al. 2022; Hoàng and Võ, 2021; Nguyen and Dang, 2024). Tuy nhiên, các cách tiếp cận này chủ

[1]*Khoa Kỹ thuật và Quản lý Xây dựng, Trường Đại học Quốc tế, TP. Hồ Chí Minh, Việt Nam*
[2]*Đại học Quốc Gia TP. Hồ Chí Minh, Việt Nam*

yếu dựa trên phân tích thống kê tuyến tính và gán trọng số chuyên gia, nên còn hạn chế trong việc xử lý mối quan hệ phi tuyến và dữ liệu mất cân bằng giữa điểm sạt lở và không sạt lở.

Từ những phân tích trên, nghiên cứu này đề xuất một phương pháp nhằm nâng cao độ chính xác và độ tin cậy trong đánh giá LSP tại khu vực đồi núi tỉnh Quảng Nam thông qua việc tích hợp kỹ thuật cân bằng dữ liệu và mô hình học máy. Cụ thể, một quy trình kết hợp giữa kỹ thuật tăng cường mẫu thiểu số tổng hợp (Synthetic Minority Over-sampling Technique-SMOTE) và mô hình mạng perceptron nhiều lớp (Multilayer Perceptron-MLP) được đề xuất nhằm giải quyết vấn đề mất cân bằng nghiêm trọng trong dữ liệu sạt lở và cải thiện hiệu suất dự báo. Bản đồ xác suất sạt lở thu được cho thấy độ chính xác phân loại được cải thiện đáng kể, khẳng định hiệu quả của việc kết hợp SMOTE và MLP trong dự báo xác suất không gian sạt lở đất.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Nghiên cứu này đề xuất một quy trình mô hình hóa tích hợp, kết hợp kỹ thuật SMOTE và mô hình MLP nhằm nâng cao độ chính xác trong dự đoán LSP. Trong đó, SMOTE được sử dụng để tạo thêm các điểm sạt lở nhân tạo từ số lượng điểm quan sát gốc hạn chế, qua đó giúp cân bằng lại tập dữ liệu huấn luyện. Mô hình MLP sau đó được huấn luyện trên tập dữ liệu đã được tăng cường, tận dụng khả năng học phi tuyến của mạng nơ-ron để nắm bắt các mối quan hệ phức tạp liên quan đến sự xuất hiện của sạt lở đất. Phương pháp đề xuất bao gồm bảy bước chính, được minh họa trong Hình 1 và mô tả chi tiết như sau:

(i) Xây dựng tập dữ liệu sạt lở, trong đó các điểm sạt lở thực tế được thu thập để tạo lập tập huấn luyện và tập xác thực.

(ii) Lựa chọn các yếu tố tác động (conditioning factors - CFs) đến hiện tượng sạt lở đất, dựa trên kiến thức chuyên ngành và khả năng thu thập dữ liệu.

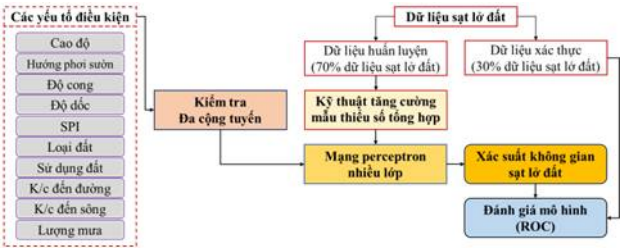
(iii) Phân tích đa cộng tuyến giữa các CFs nhằm phát hiện và loại bỏ những yếu tố dư thừa hoặc tương quan cao, đảm bảo tính độc lập trong mô hình hóa.

(iv) Áp dụng kỹ thuật SMOTE để tạo thêm các điểm sạt lở nhân tạo, qua đó làm cân bằng lại tập dữ liệu vốn bị lệch nghiêm trọng do số lượng điểm không sạt lở chiếm ưu thế.

(v) Huấn luyện mô hình MLP bằng cách sử dụng tập dữ liệu đã được tăng cường, với đầu vào bao gồm các CFs gốc. Mô hình học được mối liên hệ phi tuyến giữa các yếu tố địa môi trường và nguy cơ sạt lở.

(vi) Tạo bản đồ xác suất sạt lở đất bằng đầu ra của mô hình MLP, phản ánh xác suất xảy ra sạt lở tại từng vị trí không gian trên toàn khu vực nghiên cứu.

(vii) Đánh giá hiệu suất mô hình thông qua đường cong ROC (Receiver Operating Characteristic) và chỉ số AUC (Area Under the Curve), nhằm xác định mức độ chính xác phân loại của mô hình theo từng kịch bản dữ liệu đầu vào.



Hình 1. Sơ đồ quy trình xây dựng bản đồ xác suất không gian sạt lở đất bằng SMOTE-MLP

2.1. Kỹ thuật tăng cường mẫu thiểu số tổng hợp-SMOTE

SMOTE là một kỹ thuật phổ biến được sử dụng để xử lý các bài toán học máy có tập dữ liệu mất cân bằng (Chawla et al., 2002). Phương pháp này hoạt động bằng cách tổng hợp các điểm dữ liệu mới thuộc lớp thiểu số thông qua phép nội suy tuyến tính giữa một điểm gốc và các điểm lân cận gần nhất trong không gian đặc trưng. Trong nghiên cứu này, SMOTE được áp dụng để nội suy dữ liệu sạt lở từ 500 điểm gốc ban đầu lên các mức 2,000, 5,000, 10,000, 20,000 và 50,000 điểm, làm đầu vào huấn luyện cho mô hình MLP nhằm xây dựng bản đồ xác suất không gian sạt lở đất.

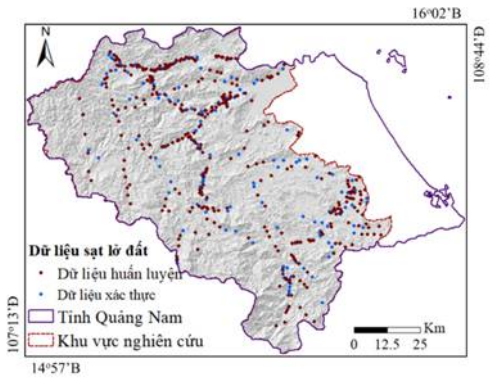
2.2. Mạng perceptron nhiều lớp-MLP

MLP là một trong những kiến trúc cơ bản của học sâu và được ứng dụng rộng rãi trong các bài toán hồi quy và phân loại, đặc biệt hiệu quả khi có sẵn tập dữ liệu huấn luyện lớn (Thirugnanam, 2023). Mô hình bao gồm một lớp đầu vào, một hoặc nhiều lớp ẩn, và một lớp đầu ra. Các lớp ẩn đóng vai trò là lõi tính toán, gồm nhiều neuron kết nối với nhau, trong đó số lượng neuron tại lớp đầu vào tương ứng với số yếu tố điều kiện sạt lở được chọn. Lớp đầu ra có hai neuron, đại diện cho hai lớp: sạt lở và không sạt lở. Cấu hình số lớp ẩn, số lượng neuron trong mỗi lớp, cũng như số vòng lặp huấn luyện được xác định dựa trên tập dữ liệu cụ thể. Quá trình tối ưu hóa các siêu tham số này được thực hiện theo phương pháp thử-sai nhằm đạt hiệu suất tốt nhất mà vẫn tránh hiện tượng quá khớp (Nguyen and Kim, 2021).

3. DỮ LIỆU NGHIÊN CỨU

Nghiên cứu được thực hiện tại khu vực miền núi thuộc tỉnh Quảng Nam, miền Trung Việt Nam. Về mặt địa lý, Quảng Nam nằm trong khoảng từ vĩ độ 14°57′B đến 16°02′B và kinh độ 107°13′Đ đến 108°44′Đ. Địa hình của tỉnh chủ yếu là đồi núi, chiếm khoảng 72% tổng diện tích, với đặc điểm địa hình phức tạp gồm núi cao, trung bình và thấp. Độ cao phổ biến dao động từ 500 đến 1,000 mét. Khu vực này có khí hậu nhiệt đới gió mùa, phân hóa rõ rệt thành hai mùa: mùa khô và mùa mưa với lượng mưa lớn và kéo dài. Theo đề án "Điều tra, đánh giá và phân vùng cảnh báo nguy cơ trượt lở đất đá các vùng núi Việt Nam" do Viện Khoa học địa chất và khoáng sản (Bộ TN&MT) thống kê, trong giai đoạn 2012–2020, toàn tỉnh ghi nhận 1,286 vụ sạt lở đất. Các điểm sạt lở thường xuất hiện dọc

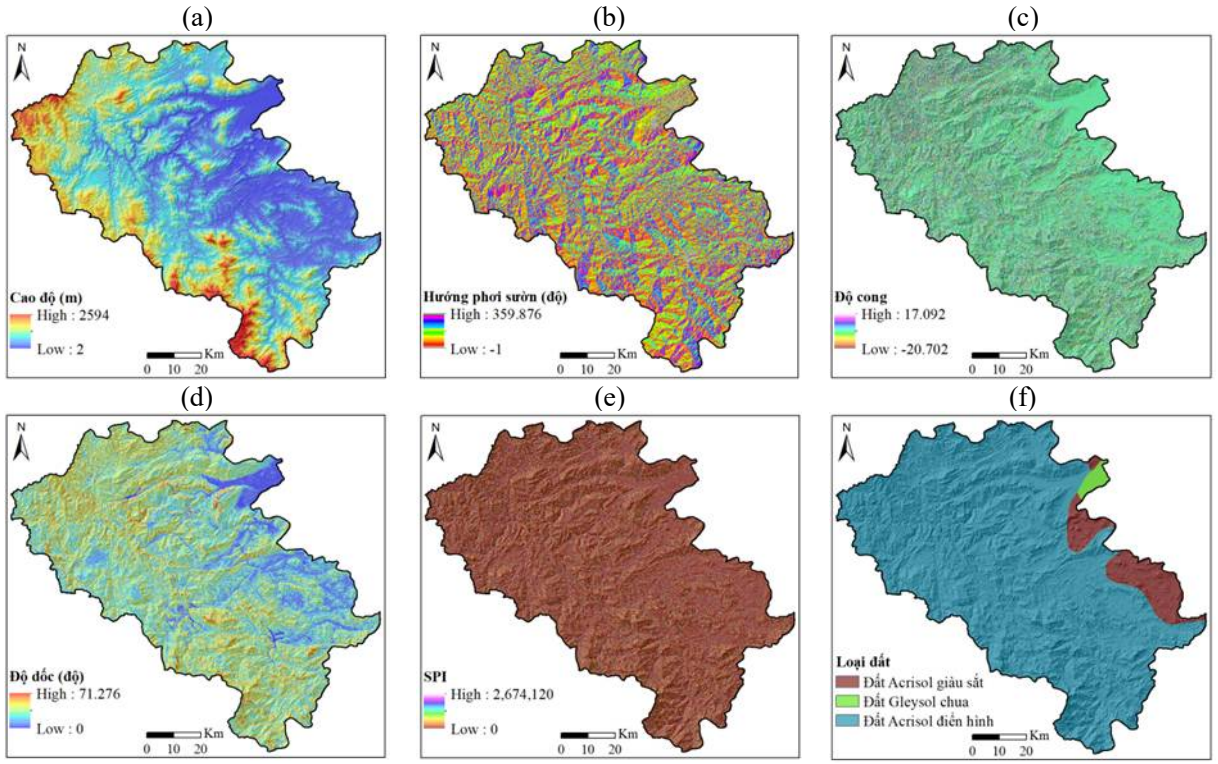
theo các tuyến giao thông chính như đường Hồ Chí Minh, đường Trường Sơn Đông và Quốc lộ 40B, gây gián đoạn giao thông và đe dọa nghiêm trọng đến cộng đồng dân cư lân cận. Đặc biệt, các huyện Bắc Trà My, Nam Trà My, Tiên Phước và Phước Sơn được xác định là các khu vực có nguy cơ sạt lở cao do địa hình dốc, địa chất yếu và lượng mưa lớn. Với tần suất sạt lở cao trong mùa mưa cùng điều kiện địa hình và khí hậu điển hình, khu vực miền núi thuộc tỉnh Quảng Nam được lựa chọn là khu vực nghiên cứu trong đề tài này.

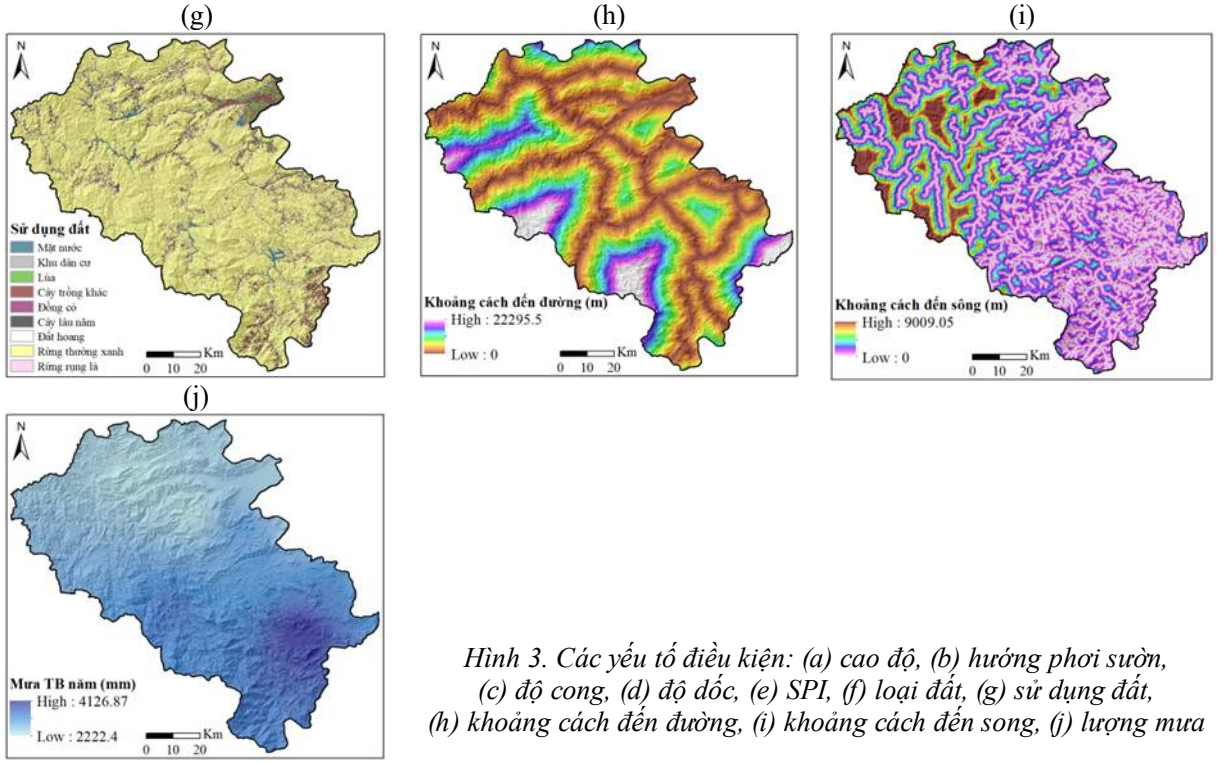


Hình 2. Dữ liệu kiểm kê sạt lở đất ở tỉnh Quảng Nam giai đoạn 1981-2015

Để xây dựng tập dữ liệu sạt lở cho khu vực miền núi tỉnh Quảng Nam, tổng cộng 500 vị trí sạt lở do mưa lớn đã được xác định (Hình 2). Tập dữ liệu này được thiết lập bằng cách kết hợp giữa ảnh viễn thám và khảo sát thực địa được xây dựng và thu thập thông qua luận văn Tiến sỹ của tác giả Nguyễn Thị Thu Hiền, (2017). Dữ liệu các điểm sạt lở thực tế phân bố dày đặc dọc theo những tuyến giao thông huyết mạch như đường Hồ Chí Minh, đường Trường Sơn Đông và Quốc lộ 40B. Đây là các tuyến đường đi qua khu vực địa hình dốc, mưa lớn kéo dài nên dễ xảy ra sạt lở, gây gián đoạn nghiêm trọng cho giao thông và đe dọa an toàn của các khu dân cư nằm ven đường. Tập dữ liệu sau đó, được chia ngẫu nhiên thành hai phần: 70% số điểm phục vụ huấn luyện mô hình, và 30% còn lại dùng để đánh giá độc lập hiệu suất mô hình.

Trong nghiên cứu này, 10 yếu tố điều kiện (Hình 3) ảnh hưởng đến sự xuất hiện sạt lở đất được lựa chọn dựa trên mức độ liên quan đã được chứng minh trong các nghiên cứu trước. Các yếu tố bao gồm: cao độ, hướng phơi sườn, độ cong, độ dốc, chỉ số năng lượng dòng chảy (Stream power index-SPI), loại đất, hiện trạng sử dụng đất, khoảng cách đến đường, khoảng cách đến sông và lượng mưa. Mô hình số độ cao (Digital elevation model-DEM) với độ phân giải 30m, thu thập từ nền tảng NextGIS.com, được sử dụng làm nguồn dữ liệu chính để trích xuất các yếu tố địa hình gồm hướng phơi sườn, độ cong, độ dốc, SPI và khoảng cách đến sông. Các yếu tố về loại đất, sử dụng đất, khoảng cách đến đường được thu thập từ Cơ quan thám hiểm hàng không vũ trụ Nhật Bản (https://www.eorc.jaxa.jp/). Trong khi, lượng mưa trung bình năm trong giai đoạn 1981-2015 được thu thập từ nghiên cứu của Nguyễn Thị Thu Hiền, (2017). Tất cả các lớp dữ liệu được xử lý và chuẩn hóa về cùng một độ phân giải không gian trước khi đưa vào mô hình phân tích.





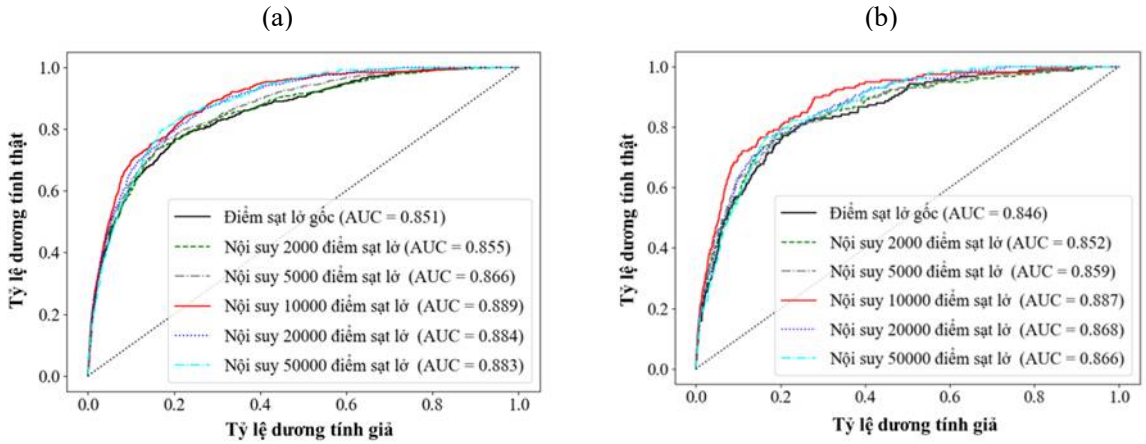
Hình 3. Các yếu tố điều kiện: (a) cao độ, (b) hướng phơi sườn, (c) độ cong, (d) độ dốc, (e) SPI, (f) loại đất, (g) sử dụng đất, (h) khoảng cách đến đường, (i) khoảng cách đến song, (j) lượng mưa

4. KẾT QUẢ VÀ THẢO LUẬN

Kết quả phân tích đa cộng tuyến đối với các yếu tố điều kiện gây sạt lở trong khu vực nghiên cứu được trình bày trong Bảng 1. Dựa trên các giá trị hệ số phóng đại phương sai (VIF) dao động từ 1.006 đến 1.571 và giá trị tolerance (TOL) trong khoảng 0.637 đến 0.994, không phát hiện thấy hiện tượng đa cộng tuyến đáng kể giữa các yếu tố.

Bảng 1. Kết quả phân tích đa cộng tuyến

Các yếu tố điều kiện	VIF	TOL
Cao độ	1.571	0.637
Hướng phơi sườn	1.007	0.993
Độ cong	1.006	0.994
Độ dốc	1.239	0.807
SPI	1.006	0.994
Loại đất	1.133	0.883
Sử dụng đất	1.156	0.865
Khoảng cách đến đường	1.266	0.790
Khoảng cách đến sông	1.449	0.690
Lượng mưa	1.236	0.809

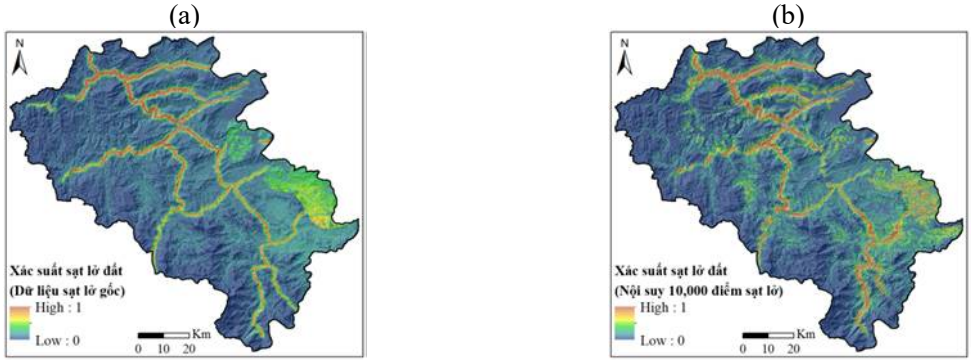


Hình 4. Đánh giá hiệu suất mô hình MLP qua chỉ số AUC theo từng kịch bản dữ liệu SMOTE (a) AUC trên tập huấn luyện; (b) AUC trên tập xác thực

Hình 4a thể hiện giá trị AUC thu được từ tập huấn luyện tương ứng với sáu kịch bản dữ liệu đầu vào: dữ liệu sạt lở gốc, và các tập dữ liệu được tăng cường bằng SMOTE với số lượng điểm sạt lở tăng dần: 2,000; 5,000;

10,000; 20,000 và 50,000. Kết quả cho thấy AUC có xu hướng tăng ổn định từ 0.851 (dữ liệu sạt lở gốc) lên đến 0.889 với 10,000 điểm nội suy, cho thấy việc sử dụng SMOTE đã cải thiện đáng kể khả năng học của mô hình MLP. Tuy nhiên, khi số lượng điểm nội suy tiếp tục tăng lên 20,000 và 50,000, AUC đạt giá trị lần lượt là 0.884 và 0.883, cho thấy dấu hiệu quá khớp nhẹ khi mô hình học quá nhiều từ dữ liệu tổng hợp. Hình 4b trình bày kết quả AUC trên tập xác thực độc lập, phản ánh trực tiếp năng lực tổng quát hóa của mô hình. Kết quả cho thấy AUC tăng từ 0.846 (dữ liệu gốc) lên đến 0.887 khi sử dụng 10,000 điểm SMOTE, chứng minh hiệu quả của việc tăng cường dữ liệu đối với hiệu suất phân loại thực tế. Tuy nhiên, tương tự tập huấn luyện, AUC trên tập xác thực giảm nhẹ xuống còn 0.868 (20,000 điểm) và 0.866 (50,000 điểm), cho thấy rằng việc sinh quá nhiều điểm nhân tạo có thể làm giảm tính khái quát của mô hình. Tổng thể, cho thấy 10,000 điểm nội suy là mức tăng cường tối ưu, giúp mô hình đạt được sự cân bằng giữa khả năng học và tổng quát hóa. Điều này nhấn mạnh tầm quan trọng của việc lựa chọn hợp lý số lượng điểm SMOTE để đảm bảo hiệu suất dự báo tốt nhất.

Hình 5a trình bày bản đồ LSP được xây dựng từ tập dữ liệu gốc với 500 điểm sạt lở quan sát thực tế. Mặc dù mô hình MLP đã nhận diện được một số khu vực nguy cơ cao, bản đồ vẫn thể hiện sự phân bố rời rạc trong việc khoanh vùng các vùng có khả năng xảy ra sạt lở. Điều này phản ánh hạn chế về khả năng học của mô hình do thiếu dữ liệu lớp thiểu số, dẫn đến kết quả dự báo còn chưa thật sự ổn định và độ chính xác phân loại thấp. Ngược lại, Hình 5b thể hiện bản đồ LSP được tạo ra từ mô hình huấn luyện trên tập dữ liệu đã được tăng cường bằng SMOTE lên 10,000 điểm sạt lở. Bản đồ cho thấy sự cải thiện rõ rệt về mặt liên tục không gian, với các vùng nguy cơ cao được xác định rõ ràng hơn, đặc biệt tập trung tại các khu vực địa hình dốc, gần sông và đường giao thông, những nơi vốn có điều kiện thuận lợi để xảy ra sạt lở. Sự cải thiện này cho thấy việc sử dụng kỹ thuật SMOTE giúp mô hình học sâu MLP khai thác tốt hơn các mối quan hệ phi tuyến giữa các yếu tố điều kiện và sự xuất hiện của sạt lở, từ đó nâng cao chất lượng bản đồ dự báo.



Hình 5. Bản đồ xác suất không gian sạt lở đất (a) dựa trên dữ liệu sạt lở gốc, (b) dựa trên 10,000 điểm sạt lở nội suy

5. KẾT LUẬN

Nghiên cứu này cho thấy việc kết hợp kỹ thuật SMOTE và mô hình học sâu MLP là một hướng tiếp cận hiệu quả nhằm cải thiện độ chính xác trong đánh giá xác suất không gian sạt lở đất, đặc biệt trong bối cảnh dữ liệu mất cân bằng nghiêm trọng. Cụ thể, kết quả AUC trên tập huấn luyện tăng dần từ 0.851 (mô hình gốc) và đạt cao nhất 0.889 khi nội suy 10,000 điểm sạt lở, tương ứng mức tăng 4.46% so với ban đầu. AUC trên tập xác nhận cũng tăng dần từ 0.846 và đạt cao nhất 0.887, cải thiện 4.85%. Điều này chứng tỏ việc tăng cường dữ liệu bằng SMOTE không chỉ khắc phục mất cân bằng mà còn nâng cao chất lượng phân loại và độ tin cậy của bản đồ xác suất sạt lở đất. Các kết quả cho thấy phương pháp đề xuất có tiềm năng ứng dụng thực tiễn cao trong cảnh báo sớm và quy hoạch giảm nhẹ rủi ro sạt lở tại các khu vực miền núi như Quảng Nam, Việt Nam.

Lời cảm ơn: Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số DS2025-28-01.

TÀI LIỆU THAM KHẢO

Đỗ, C. T., Phạm, T. B., & Nguyễn, Đ. Đ. (2022). *Ứng dụng mô hình trọng số dẫn chứng (woe) trong xây dựng bản đồ nguy cơ sạt lở tại tỉnh Quảng Nam.* Tạp Chí Khoa Học Công Nghệ Xây Dựng, 16 (2V), 139–152.

Hoàng, N. T., & Võ, T. T. (2021). *Nghiên cứu xây dựng bản đồ phân vùng nguy cơ sạt lở đất cho khu vực miền núi tỉnh Quảng Nam.* Tạp Chí Khoa Học Và Công Nghệ Thủy Lợi, 68.

Nguyễn, T. T. H. (2017). *Đánh giá điều kiện hình thành và nguy cơ trượt lở đất trong bối cảnh biến đổi khí hậu ở tỉnh Quảng Nam.*

Ali, S. A., Parvin, F., Vojteková, J., Costache, R., Linh, N. T. T., Pham, Q. B., Vojtek, M., Gigović, L., Ahmad, A., & Ghorbani, M. A. (2021). *GIS-based landslide susceptibility modeling: A comparison between fuzzy multi-criteria and machine learning algorithms.* Geoscience Frontiers, 12(2), 857–876.

Bragagnolo, L., da Silva, R. V., & Grzybowski, J. M. V. (2020). *Landslide susceptibility mapping with r. landslide: A free open-source GIS-integrated tool based on Artificial Neural Networks.* Environmental Modelling & Software, 123, 104565.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 16, 321–357.

Hong, H., Liu, J., Bui, D. T., Pradhan, B., Acharya, T. D., Pham, B. T., Zhu, A.-X., Chen, W., & Ahmad, B. Bin. (2018). *Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China).* Catena, 163, 399–413.

Nguyen, B.-Q.-V., & Kim, Y.-T. (2021). *Landslide spatial probability prediction: a comparative assessment of naive Bayes, ensemble learning, and deep learning approaches.* Bulletin of Engineering Geology and the Environment, 80, 4291–4321.

Nguyen, T. T. H., & Dang, T. H. (2024). *Landslide susceptibility assessment in quang nam province using statistical index and analytical hierarchical process.* Hnue Journal Of Science, 69(1), 144–160.

Nguyen, B. Q. V, Ho, L. H. P., & Kim, Y. T. (2024). *An ensemble model of logistic regression, Naïve Bayes, and adaboost for assessing the landslide spatial probability-study case: Phuoc Son, Quang Nam, Vietnam and Umyeon, Seoul, Korea.* Civ. Eng. Archit, 12(3), 2010–2028.

Nhu, V.-H., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., Jaafari, A., Chen, W., Miraki, S., & Dou, J. (2020). *Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms.* International Journal of Environmental Research and Public Health, 17(8), 2749.

Pradhan, B. (2013). *A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS.* Computers & Geosciences, 51, 350–365.

Thirugnanam, H. (2023). Deep Learning in Landslide Studies: *A Review. Progress in Landslide Research and Technology,* Volume 1 Issue 2, 2022, 247–255.

Van Westen, C. J. (2000). *The modeling of landslide hazards using GIS.* Surveys in Geophysics, Van Westen, C. J. (2000). The modeling of landslid. https://doi.org/10.1023/A:1006794127521

Abstract:

A SMOTE-MLP INTEGRATED MODEL FOR ENHANCING THE ACCURACY OF LANDSLIDE SPATIAL PROBABILITY PREDICTION FROM IMBALANCED DATA IN THE MOUNTAINOUS REGION OF QUANG NAM PROVINCE

The imbalance in landslide inventory datasets, where the number of actual landslide points is significantly lower than non-landslide points, greatly affects the performance of machine learning models in landslide spatial probability (LSP) prediction. This study proposes the use of the Synthetic Minority Over-sampling Technique (SMOTE) to interpolate and augment the number of landslide points, in combination with a Multilayer Perceptron (MLP) model, to generate LSP maps for the mountainous region of Quang Nam Province, Vietnam. The original dataset, consisting of 500 landslide points, was augmented using SMOTE at five levels: 2,000; 5,000; 10,000; 20,000; and 50,000 synthetic samples. The results show that the AUC (Area Under the ROC Curve) on the validation set improved from 0.846 (with 500 original points) to 0.887 when 10,000 synthetic points were used, demonstrating the effectiveness of SMOTE in enhancing model accuracy. However, when the number of synthetic points increased to 20,000 and 50,000, the AUC slightly decreased to 0.868 and 0.866, respectively, indicating a decline in generalization performance. Therefore, selecting an appropriate number of synthetic samples is crucial to balancing performance improvement and avoiding overfitting. This study confirms the potential of integrating SMOTE and MLP for constructing reliable landslide susceptibility maps from imbalanced datasets.

Keywords: Landslide, multilayer perceptron, data imbalance, synthetic minority oversampling technique.

Ngày nhận bài: 20/7/2025

Ngày chấp nhận đăng: 28/8/2025