TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Xây dựng bản đồ phân vùng nguy cơ sạt lở đất tại huyện Mường Chà, tỉnh Điện Biên sử dụng các kỹ thuật phân loại K-Nearest-Neighbor và Gradient Boosting

Vũ Cao Đạt[1*], Nguyễn Đức Đảm[1], Phạm Thái Bình[1]

[1] Trường Đại học Công nghệ GTVT, 54 Triều Khúc, Thanh Xuân, Hà Nội, Việt Nam; datvc@utt.edu.vn; damnd@utt.edu.vn; binhpt@utt.edu.vn

*Tác giả liên hệ: datvc@utt.edu.vn; Tel.: +84–384026586

Ban Biên tập nhận bài: 5/11/2022; Ngày phản biện xong: 23/12/2022; Ngày đăng bài: 25/12/2022

Tóm tắt: Bài báo tiến hành xây dựng bản đồ phân vùng nguy cơ sạt lở đất tại Huyện Mường Chà, tỉnh Điện Biên sử dụng các kỹ thuật phân loại K-Nearest-Neighbor (KNN) và Gradient Boosting (GB) - là những kỹ thuật học máy có khả năng phân tích và khai phá dữ liệu lịch sử để phân loại và dự báo. Dữ liệu không gian được xây dựng bao gồm 206 vị trí sạt lở đất xảy ra trong quá khứ và 10 tham số điều kiện gây ra sạt lở đất được thu thập. Để kiểm chứng và so sánh các mô hình, các chỉ tiêu đánh giá định lượng bao gồm đường cong ROC, độ chính xác (%) đươc sử dụng. Kết quả đánh giá và so sánh cho thấy cả hai mô hình KNN và GB có năng lực dự báo không gian sạt lở đất cao; trong đó, mô hình GB có năng lực dự báo cao hơn so với mô hình KNN. Bản đồ phân vùng nguy cơ sạt lở đất xây dựng từ mô hình GB có độ chính xác cao có thể được sử dụng vào mục đích lập quy hoạch sử dụng đất, phục vụ phòng và chống những tác hại gây ra bởi sạt lở đất.

Từ khóa: Sạt lở đất; K–Nearest–Neighbor; Gradient Boosting; Điện Biên; Việt Nam.

1. Giới thiệu

Khu vực miền núi Phía Bắc của Việt Nam là một trong những khu vực chịu ảnh hưởng nghiêm trọng bởi sạt lở đất hàng năm [1]. Khu vực này bao gồm 15 tỉnh trong đó có tỉnh Điện Biên là tỉnh chiếm 28/8% diện tích tự nhiên của Việt Nam và có địa hình chủ yếu là dãy núi cao có độ dốc lớn và nền địa chất yếu, Vì vậy, dưới tác động của biến đổi khí hậu và quá trình đô thị hóa diễn ra mạnh mẽ trong thời gian gần đây các hiện tượng thiên tại như sạt lở đất, lũ quét và lũ ống xảy ra ngày càng nhiều và mức độ nghiêm trọng ngày càng gia tăng. Vì vậy, cần phải có những công cụ, giải pháp cần thiết và kịp thời để giảm thiểu những thiệt hại gây ra bởi thiên tai sạt lở đất.

Xây dựng bản đồ phân vùng nguy cơ sạt lở đất để xác định các khu vực có xắc xuất xảy ra sạt lở đất cao là nhiệm vụ cần thiết và là công cụ hữu ích để nâng cao hiệu quả phòng và chống thiên tại sạt lở đất [2]. Nghiên cứu và dự báo không gian sạt lở đã được thực hiện tại rất nhiều khu vực trên thế giới trong đó có Việt Nam. Nói chung, có hai cách tiếp cận chính nghiên cứu về dự báo không gian sạt lở đất bao gồm "định lượng" và "định tính" [3]. Cách tiếp cận định tính dựa vào quan điểm của các chuyên gia để xác định các trọng số cho các tham số thành phần để xác định xắc xuất xảy ra sạt lở đất ở một khu vực nghiên cứu nhất định. Cách tiếp cận định lượng là cách tiếp sử dụng các hàm hoặc công thức toán học dựa trên xắc xuất thống kê để xác định các trọng số. Cách tiếp cận định lượng được xem xét là cách tiếp cận có tính khách quan hơn và cho kết quả có độ chính xác cao hơn so với cách tiếp cận định tính [4].

Trong một vài thập kỷ gần đây, học máy (trí tuệ nhân tạo) được biết đến như là một phương pháp tính toán định lượng tiên tiến giải quyết rất nhiều các bài toán dự báo trong đó có dự báo không gian sạt lở đất, với nhiều kết quả có độ chính xác và hiệu quả cao. [5] kết hợp thuật toán trọng số lớp và các mô hình trí tuệ nhân tạo (hồi quy logistic, rừng ngẫu nhiên, máy học tăng cường độ dốc ánh sáng) trong dự báo không gian sạt lở đất khu vực hồ Tam Hiệp, Trung Quốc. [6] phát triển các mô hình lại giữa học sâu và phương pháp tập đồng bộ không đồng nhất để dự báo không gian sạt lở đất khu vực hồ Tam Hiệp, Trung Quốc. [7] xây dựng công cụ kỹ thuật tính toán bán tự động mã nguồn mở và miễn phí trong lập bản đồ phân vùng nguy cơ sạt lở đất sử dụng một vài thuật toán trí tuệ nhân tạo như máy véc tơ hỗ trợ (SVM), rừng ngẫu nhiên (RF) và XGBoost. Ở Việt Nam, một số nghiên cứu xây dựng bản đồ phân vùng nguy cơ sạt lở đất sử dụng các mô hình trí tuệ nhân tạo đã được thực hiện ở một số khu vực [8–10]. Nhìn chung, các mô hình học máy được đánh giá là cách tiếp có độ chính xác cao và phù hợp trong xây dựng bản đồ phân vùng nguy cơ sạt lở đất.

Mục tiêu chính của nghiên cứu này là xây dựng bản đồ phân vùng nguy cơ sạt lở đất khu vực huyện Mường Chà, tỉnh Điện Biên sử dụng các kỹ thuật học máy điển hình như kỹ thuật phân loại K-Nearest-Neighbor (KNN) và Gradient Boosting (GB). Khu vực huyện Mường Chà, tỉnh Điện Biên được lựa chọn nghiên cứu là vùng có địa lý đồi núi hiểm trở và thường xuyên phải hứng chịu nhiều thiệt hại về người và của do sạt lở đất gây ra hàng năm. Kỹ thuật đường cong ROC và các chỉ số thống kê đánh giá định lượng được sử dụng để đánh giá và so sánh độ chính xác của các mô hình dự báo. Các công cụ như ArcGIS và Python được dùng để xây dựng cơ sở dữ liệu và mô hình hóa.

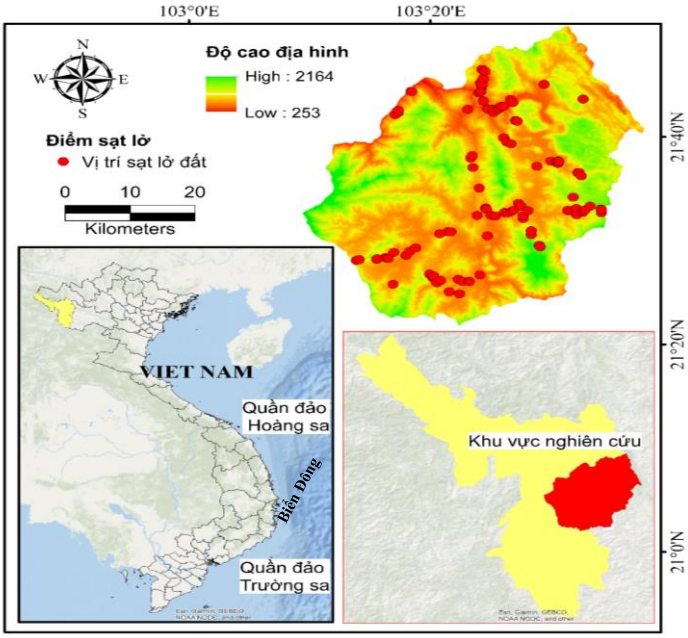
2. Dữ liệu và phương pháp nghiên cứu

2.1. Đặc điểm của khu vực nghiên cứu

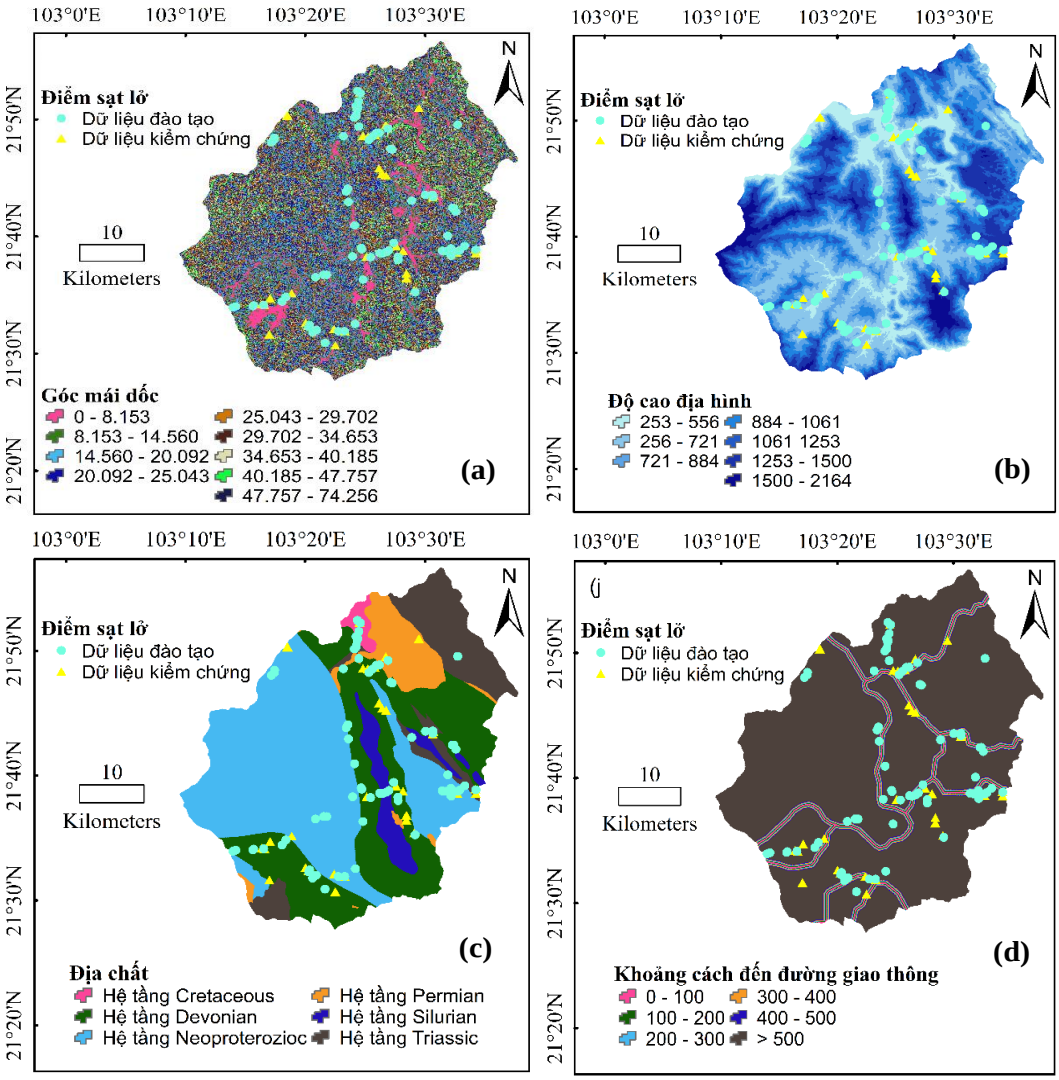
Huyện Mường Chà là một huyện miền núi thuộc vùng Tây Bắc, tỉnh Điện Biên có tọa độ địa lý kinh độ 103°49’ Đông, vĩ độ 21°40’ Bắc. Phía Tây Nam giáp với cộng hòa dân chủ nhân dân Lào, phía Tây giáp huyện Mường Nhé, phía Đông giáp huyện Tủa Chùa và Tuần Giáo, phía Nam giáp huyện Điện Biên, và phía Bắc giáp thị xã Mường Lay. Mường Chà đường biên giới Việt - Lào dài 56 km với tổng diện tích khoảng 1200 km^2, gồm 14 xã trong đó có 6 xã biên giới và 1 thị trấn. Các xã trong huyện đều là những khu vực vùng cao và sâu không thuận lợi trong giao thông, sự phân bố dân cư không tập trung. Địa hình huyện Mường Chà chủa yếu là địa hình đồi núi có độ cao trung bình so với mặt nước biển từ 350-1.500m, hướng của địa hình nghiêng dần theo hướng Tây Bắc - Đông Nam. Ngoài ra, địa hình bị chia cắt và mức độ chênh lệch địa hình lớn. Về thủy văn, huyện Mường Chà nằm trong phạm vi đầu nguồn của lưu vực sông Đà, nhiệt độ trung bình là 22°C đến 25°C, lượng mưa trung bình cả năm là 2.432 mm. Mùa mưa chủ yếu gia tăng từ tháng 4 đến cuối tháng 9 (http://dienbien.gov.vn/portal/pages/print.aspx?p=12962).

2.2. Cơ sở dữ liệu

Cơ sở dữ liệu sử dụng trong nghiên cứu này bao gồm hai dạng dữ liệu chính: hiện trạng sạt lở đất và các bản đồ thành phần các yếu tố điều kiện gây ra sạt lở đất. Trong đó, hiện trạng sạt lở đất được xây dựng từ dữ liệu thu thập từ Sở Tài nguyên và Môi trường tỉnh Điện Biên kết hợp với sử dụng ảnh Google Earth (Hình 1). Có tổng cộng 206 vụ sạt lở đất trong quá khứ đã được nhận diện và thu thập. Trong đó, 70% (144) ví trí được sử dụng để xây dựng dữ liệu đào tạo và 30% (62) vị trí còn lại được sử dụng để xây dựng dữ liệu kiểm chứng. Quá trình xảy ra sạt lở đất thường chịu tác động bởi các yếu tố nguyên nhân liên quan đến địa các yếu tố liên quan đến các hoạt động của con người, sử dụng đất, địa chất–thủy văn, và hình địa mạo [9].



Hình 1. Vị trí khu vực nghiên cứu và hiện trạng sạt lở đất.



Hình 2. Một số bản đồ tham số điều kiện gây ra sạt lở đất: (a) Góc mái dốc; (b) Độ cao địa hình; (c) Địa chất; (d) Khoảng cách đến đường giao thông.

Trong nghiên cứu này, căn cứ vào cơ chế xảy ra sạt lở đất trong quá khứ và giả thiết rằng các vụ sạt lở đất xảy ra trong tương lai sẽ xảy ra dưới sự tác động của cùng các yếu tố nguyên nhân gây ra các vụ sạt lở đất trong quá khứ, tổng cộng 10 tham số điều kiện (hình dáng bề mặt địa hình, độ cao địa hình, hướng mái dốc,góc mái dốc, chỉ số bao phủ thực vật (NDVI), độ ẩm địa hình, địa chất, khoảng cách tới đường giao thông, khoảng cách tới các đứt gãy, và khoảng cách đến sông suối) đã được xác định và lựa chọn để xây dựng cơ sở dữ liệu cho bài toán dự báo không gian sạt lở đất. Trong đó, các tham số địa hình–địa mạo như độ cao địa hình, khoảng cách tới sông suối, hướng mái dốc, độ ẩm địa hình, góc mái dốc, hình dáng bề mặt địa hình được trích xuất và xây dựng từ mô hình số độ cao (DEM) với độ phân giải 30m được tải từ cơ sở dữ liệu của Hội địa chất Hoa Kỳ (https://earthexplorer.usgs.gov), các tham số địa chất và khoảng cách tới đứt gãy được trích xuất và xây dựng từ bản đồ địa chất Việt Nam tỷ lệ 1:200.000 thu thập từ Tổng cục Địa chất và Khoáng sản Việt Nam. Chỉ số NDVI được trích xuất từ cơ sở dữ liệu của Hội địa chất Hoa Kỳ (https://earthexplorer.usgs.gov), khoảng cách tới đường giao thông được xây dựng từ hệ thống đường trích xuất từ bản đồ kỹ thuật số của thế giới (https://www.diva-gis.org/gdata). Bản đồ của các tham số điều kiện được xây dựng trên nền tảng ứng dụng ArcGIS (Hình 2) và được chồng lấn với bản đồ hiện trạng sạt lở đất để xây dựng cơ sở dữ liệu cho mô hình dự báo.

2.3. Phương pháp nghiên cứu

2.3.1. Kỹ thuật phân loại K–Nearest neighbors (KNN)

KNN là một thuật toán trí tuệ nhân tạo phân loại dựa trên khoảng cách Euclide giữa các trường hợp [11]. KNN phân loại và dự đoán các nhãn lớp cho các trường hợp khác nhau bằng cách đo khoảng cách Euclide ngắn nhất của nó từ các trường hợp khác [12]; trong đó, khoảng cách Euclide được tính khi xem xét tất cả các tính năng hoặc thuộc tính dưới dạng thứ nguyên [13]. Trong bài báo này, mô hình KNN được sử dụng để dự báo không gian sạt lở đất dựa trên việc phân loại nhị phân 2 nhãn: nhãn "1" thể hiện các vị trí có xảy ra sạt lở đất và nhãn "0" thể hiện các vị trí không xảy ra sạt lở đất.

2.3.2. Kỹ thuật phân loại Gradient Boosting (GB)

GB là một trong những phương pháp trí tuệ nhân tạo điển hình được sử dụng để phát triển các mô hình phân loại và hồi quy nhằm tối ưu hóa quá trình học của mô hình để giải quyết các vấn đề phi tuyến tính [13]. GB được biết đến rộng rãi hơn với tên gọi cây quyết định hoặc cây hồi quy. GB được đào tạo và xây dựng bằng cách bằng cách thêm người học mới theo cách tuần tự dần dần từ đó nhóm các mô hình dự đoán yếu, ví dụ, cây quyết định, thông qua các các nút và lá của cây quyết định, và kết quả dự đoán cuối cùng được xác định dựa trên các nút quyết định [14]. Các cây quyết định riêng lẻ là những mô hình yếu, nhưng khi được xem như một tập hợp (GB), độ chính xác của chúng được cải thiện nhiều [15]. Vì vậy, các quần thể được xây dựng dần dần theo cách tăng dần sao cho mọi quần thể sẽ sửa lỗi trong quần thể trước đó, từ đó nâng cao độ chĩnh xác trong quá trình đào tạo mô hình.

2.3.3. Phương pháp đánh giá độ chính xác

Trong nghiên cứu này, các kỹ thuật như đường cong ROC và các chỉ số thống kê định lượng bao gồm chỉ số giá trị dự đoán âm (NPV), Kappa (K), giá trị dự đoán dương (PPV), độ chính xác (ACC), căn của sai số toàn phương trung bình gốc (RMSE), độ đặc hiệu (SPF), độ nhạy (SST), sai số tuyệt đối trung bình (MAE) được lựa chọn để đánh giá độ chính xác của các mô hình học máy. Lý thuyết và công thức tính các chỉ số này được trình bày cụ thể chi tiết trong nghiên cứu Đức, Thanh [1]. Nhìn chung, giá trị diện tích dưới đường cong ROC (AUC), K, PPV, NPV, ACC, SPF, SST càng cao thể hiện độ chính xác của mô hình là càng tốt. Ngược lại, các giá trị MAE, RMSE càng thấp thì độ chính xác của mô hình càng thấp.

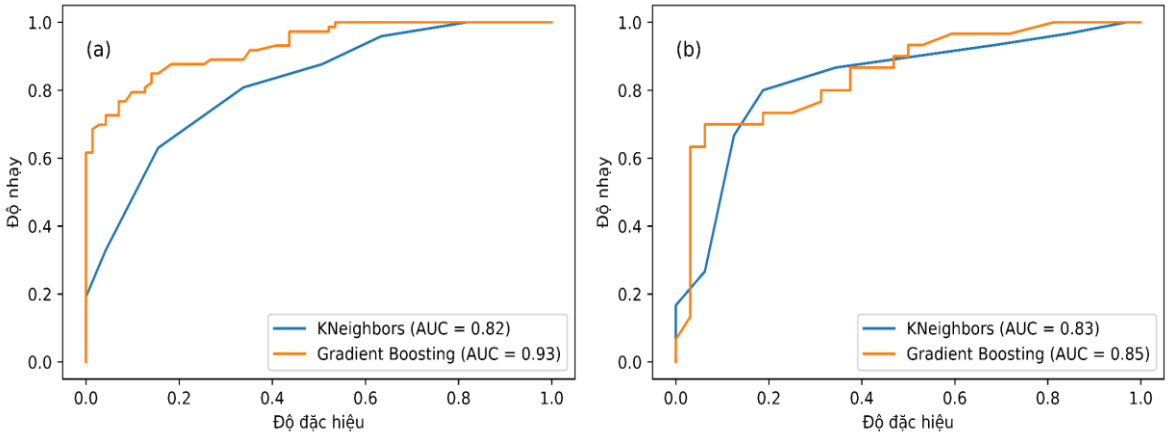
3. Kết quả và thảo luận

3.1. Đánh giá độ chính xác của các mô hình

Mô hình dự báo không gian sạt lở đất sử dụng kỹ thuật KNN và GB được xây dựng trên bộ dữ liệu đào tạo và được kiểm chứng trên bộ dữ liệu kiểm chứng và kết quả năng lực dự báo của các mô hình được thể hiện trên Hình 3, Hình 4 và Bảng 1. Kết quả dự báo sử dụng kỹ thuật đường cong ROC (Hình 3) thể hiện rằng giá trị AUC của cả hai mô hình KNN và GB đều cao cho cả bộ dữ liệu đào tạo và kiểm chứng. Cụ thể, Giá trị AUC của mô hình KNN và GB cho bộ dữ liệu đào tạo lần lượt là 0,82 và 0,93 trong khi đó với bộ dữ liệu kiểm chứng lần lượt là 0.83 và 0.85. Tuy nhiên, giá trị AUC của mô hình KNN cao hơn so với mô hình GB cho cả hai bộ dữ liệu đào tạo và dữ liệu kiểm chứng.

Kết quả dự báo của hai mô hình sử dụng các chỉ số thống kê khác được thể hiện ở Bảng 2. Giá trị các chỉ số thông kê của mô hình KNN lần lượt là PPV = 66,20%, NPV = 80,82%, SST = 77,05%, SPF = 71,08%, ACC = 73,61% và K = 0,471 sử dụng bộ dữ liệu đào tạo và PPV = 81,25%, NPV = 80%, SST = 81,25%, SPF = 80%, ACC = 80,65% và K = 0,613 sử dụng bộ dữ liệu kiểm chứng. Giá trị các chỉ số thông kê của mô hình GB lần lượt là PPV = 85,92%, NPV = 82,19%, SST = 82,43%, SPF = 85,71%, ACC = 84,03% và K = 0,681 sử dụng bộ dữ liệu đào tạo và PPV = 90,63%, NPV = 70%, SST = 76,32%, SPF = 87,50%, ACC = 80,65% và K = 0,610 sử dụng bộ dữ liệu kiểm chứng. Hình 3 thể hiện sự phân bố giá trị lỗi bình phương trung bình gốc (RMSE) của mô hình KNN và GB sử dụng bộ dữ liệu đào tạo và bộ dữ liệu kiểm chứng.

Nhìn chung, kết quả cho thấy cả hai mô hình KNN và GB có năng lực dự báo tốt; trong đó độ chính xác của mô hình GB tốt hơn so với mô hình KNN trong việc dự báo không gian sạt lở đất. Kết quả này phù hợp với kết quả của các nghiên cứu đã công bố [5–6].

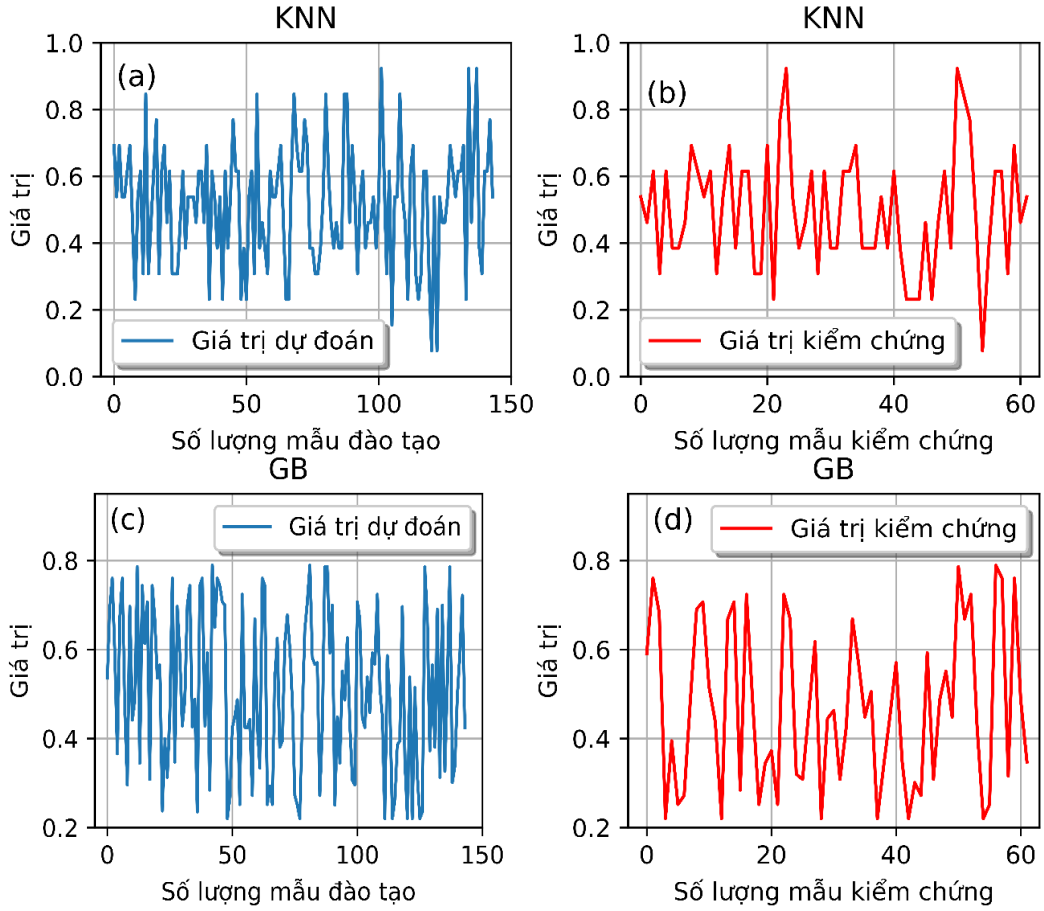


Hình 3. Giá trị AUC của các mô hình KNN và GB sử dụng: (a) Dữ liệu đào tạo; (b) Dữ liệu kiểm chứng.

Bảng 2. Hiệu suất của mô hình.

STT	Tham số	Dữ liệu đào tạo		Dữ liệu kiểm chứng	
		KNN	GB	KNN	GB
1	TP	47	61	26	29
2	TN	59	60	24	21
3	FP	24	10	6	3
4	FN	14	13	6	9
5	PPV (%)	66,20	85,92	81,25	90,63
6	NPV (%)	80,82	82,19	80,00	70,00
7	SST (%)	77,05	82,43	81,25	76,32
8	SPF (%)	71,08	85,71	80,00	87,50

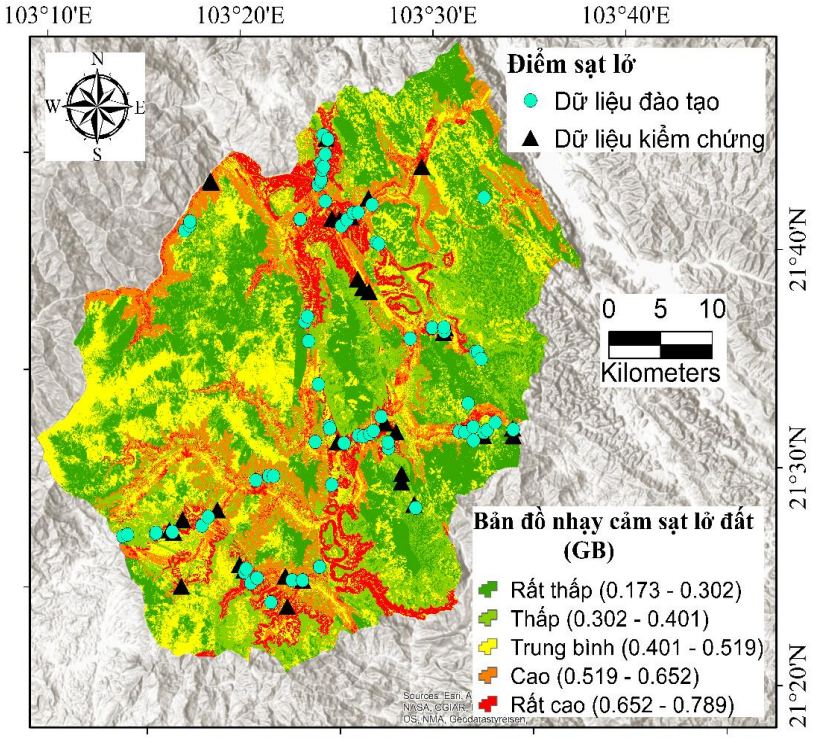
STT	Tham số	Dữ liệu đào tạo		Dữ liệu kiểm chứng	
		KNN	GB	KNN	GB
9	ACC (%)	73,61	84,03	80,65	80,65
10	K	0,471	0,681	0,613	0,610
11	RMSE	0,513	0,400	0,440	0,440
12	MAE	0,264	0,16	0,194	0,194



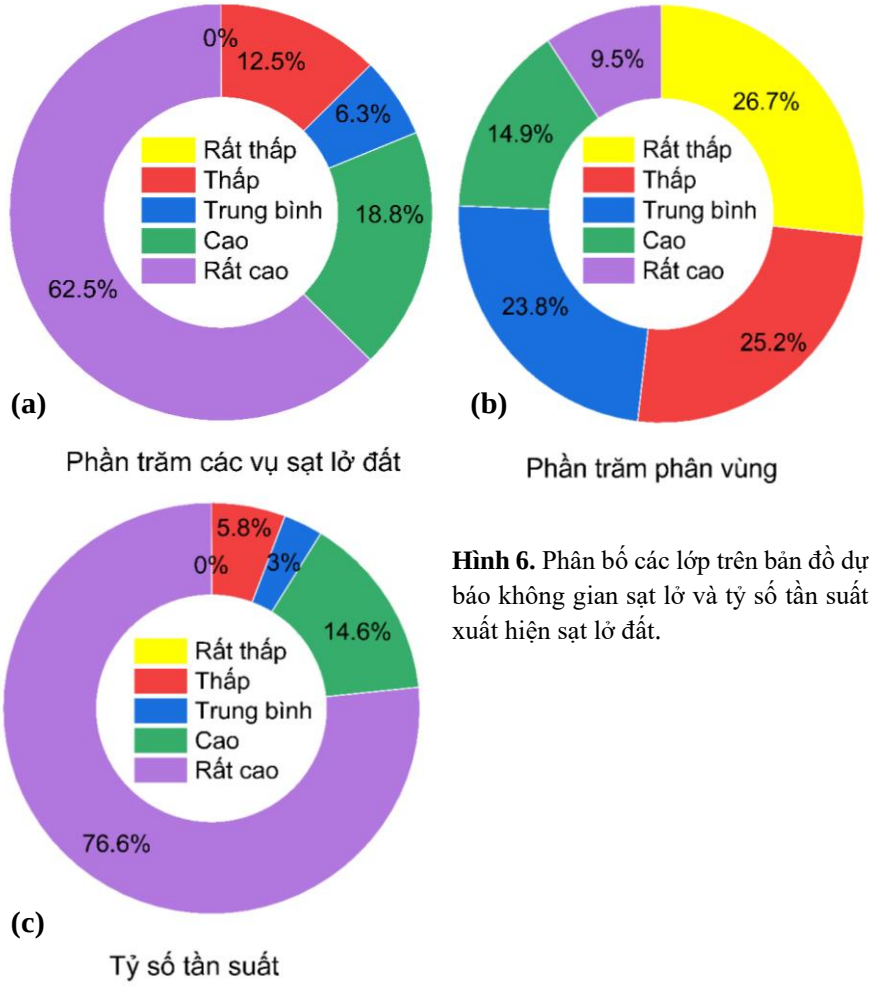
Hình 4. Giá trị lỗi bình phương trung bình gốc (RMSE) của các mô hình: (a) Đào tạo KNN; (b) Kiểm chứng KNN; (c) Đào tạo GB; (d) Kiểm chứng GB.

3.2. Xây dựng bản đồ phân vùng nguy cơ sạt lở đất

Bản đồ phân vùng nguy cơ sạt lở đất được xây dựng sử dụng kết quả đào tạo của mô hình GB và được thể hiện trên Hình 5. Cụ thể, giá trị xác xuất xảy ra sạt lở đất cho các pixel trong khu vực nghiên cứu được xác định thông qua quá trình đào tạo mô hình GB. Các giá trị này sau đó được phân loại thành 5 lớp bao gồm: rất cao, cao, trung bình, thấp, và rất thấp sử dụng phương pháp phân loại điểm nghỉ tự nhiên được tích hợp trong ứng dụng ArcGIS [16]. Hình 6a và 6b thể hiện sự phân bố các vụ sạt lở đất trong quá khứ trên các lớp phân vùng của bản đồ phân vùng nguy cơ sạt lở đất. Để đánh giá độ chính xác của bản đồ dự báo, các vụ sạt lở đất trong dữ liệu kiểm chứng được chồng lấn lên các lớp của bản đồ phân vùng và xác định tỷ số tần suất xuất hiện, kết quả thể hiện trên Hình 6c. Kết quả đánh giá cho thấy hầu hết các vụ sạt lở đất trong quá khứ xảy ra ở lớp xắc xuất rất cao và cao với giá trị tỷ số tần suất là cao nhất: Rất cao (76,6%) và cao (14,6%). Điều này chứng tỏ, bản đồ dự báo không gian sạt lở đất xây dựng từ kết quả mô hình GB có độ chính xác cao và có thể sử dụng trong việc hỗ trợ giảm thiểu tác động gây ra bởi sạt lở đất.



Hình 5. Bản đồ dự báo không gian sạt lở đất sử dụng mô hình GB.



Hình 6. Phân bố các lớp trên bản đồ dự báo không gian sạt lở và tỷ số tần suất xuất hiện sạt lở đất.

4. Kết luận

Bản đồ phân vùng nguy cơ sạt lở đất là công cụ hữu ích phục vụ cho quá trình lập quy hoạch sử dụng đất hiệu quả giảm thiểu các tác động gây ra bởi thiên tai sạt lở đất. Bài báo tiến hành sử dụng các kỹ thuật tiên tiến trí tuệ nhân tạo: KNN và GB để xây dựng bản đồ dự báo không gian sạt lở đất khu vực huyện Mường Chà, tỉnh Điện Biên. Bản đồ hiện trạng sạt lở đất đã được xây dựng với tổng cộng 206 vụ sạt lở đất trong quá khứ. Có tổng cộng 10 tham số nguyên nhân sạt lở đất đã được lựa chọn để xây dựng cở sở dữ liệu sử dụng cho mô hình dự báo. Các kỹ thuật đánh giá định lượng như đường cong ROC đã được sử dụng để đánh giá và so sánh độ chính xác của các mô hình.

Kết quả của nghiên cứu chỉ ra rằng cả hai mô hình KNN và GB có có độ chính xác cao trong xây dựng bản đồ phân vùng nguy cơ sạt lở đất; Tuy nhiên, mô hình GB có độ chính xác cao hơn mô hình KNN. Vì vậy, mô hình GB có thể dùng như một công cụ tiềm năng trong xây dựng bản đồ phân vùng nguy cơ sạt lở đất. Bản đồ phân vùng nguy cơ sạt lở đất khu vực huyện Mường Chà được xây dựng có độ chính xác cao, có thể được dùng trong việc quy hoạch sử dụng đất và ra quyết định liên quan đến quản lý thiên tai sạt lở đất. Trong nghiên cứu này, các tham số liên quan đến địa hình-địa mạo, địa chất, … đã được sử dụng; Tuy nhiên, các tham số liên quan đến thủy văn như sự phân bố nước ngầm và tham số mưa chưa được xem xét. Các kỹ thuật KNN và GB được kiểm chứng có thể được áp dụng cho các khu vực khác khi xem xét đến tính đặc thù và đặc điểm riêng của từng khu vực.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: P.T.B., V.C.Đ., N.Đ.Đ.; Xử lý số liệu: N.Đ.Đ., V.C.Đ.; Chạy mô hình: N.Đ.Đ.; Viết bản thảo bài báo: P.T.B., V.C.Đ., N.Đ.Đ.; Chỉnh sửa bài báo: P.T.B., V.C.Đ..

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Trường Đại học Công nghệ Giao thông Vận tải trong đề tài "Nghiên cứu ứng dụng một số thuật toán học máy trong phân vùng nguy cơ sạt lở đất khu vực miền núi" mã số ĐTTĐ2022–16.

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

Tài liệu tham khảo

1. Long, D.V.; Cong, N.C.; Cuong, N.T.; Binh, N.Q.; Phuoc, V.N.D. An Assessment of Terrain Quality and Selection Model in Developing Landslide Susceptibility Map–A Case Study in Mountainous Areas of Quang Ngai Province, Vietnam. In: Modern mechanics and applications, Springer, 2022, pp. 959–970.
2. Trinh, T.; Luu, B.T.; Le, T.H.T.; Nguyen, D.H.; Van, T.T.; Van, N.T.H.; Nguyen, K.Q.; Nguyen, L.T. A comparative analysis of weight–based machine learning methods for landslide susceptibility mapping in Ha Giang area. *Big Earth Data* **2022**, 1–30.
3. Zhang, W.; Liu, S.; Wang, L.; Samui, P.; Chwała, M.; He, Y. Landslide susceptibility research combining qualitative analysis and quantitative evaluation: A case study of Yunyang County in Chongqing, China. *Forests* **2022**, *13(7),*1055.
4. Yong, C.; Jinlong, D.; Fei, G.; Bin, T.; Tao, Z.; Hao, F.; Li, W.; Qinghua, Z. Review of landslide susceptibility assessment based on knowledge mapping. *Stochastic Environ. Res. Risk Assess* **2022**, 1–19.
5. Zhang, H.; Song, Y.; Xu, S.; He, Y.; Li, Z.; Yu, X.; Liang, Y.; Wu, W.; Wang, Y. Combining a class–weighted algorithm and machine learning models in landslide susceptibility mapping: A case study of Wanzhou section of the Three Gorges Reservoir, China. *Comput. Geosci*. **2022**, *158*, 104966.

6. Lv, L.; Chen, T.; Dou, J.; Plaza, A. A hybrid ensemble–based deep–learning framework for landslide susceptibility mapping. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102713.
7. Sahin, E.K. Implementation of free and open–source semi–automatic feature engineering tool in landslide susceptibility mapping using the machine–learning algorithms RF, SVM, and XGBoost. *Stochastic Environ. Res. Risk Assess* **2022**, 1–26.
8. Bien, T.X.; Truyen, P.T.; Phong, T.V.; Nguyen, D.D.; Amiri, M.; Costache, R.; Duc, D.M.; Le, H.V.; Nguyen, H.B.T.; Prakash, I. Landslide susceptibility mapping at sin Ho, Lai Chau province, Vietnam using ensemble models based on fuzzy unordered rules induction algorithm. *Geocarto Int.* **2022**, 1–22.
9. Đức, Đ.N.; Thanh, T.N.; Văn, P.T.; Thái, B.P. Phát triển mô hình học máy cây quyết định và cây quyết đinh xen kẽ thành lập bản đồ dự báo không gian sạt lở đất tại huyện Mường Nhé, tỉnh Điện Biên, Việt Nam. *Tạp chí điện tử Khoa học và Công nghệ Giao thông* **2022**, 36–56.
10. Bui, Q.D.; Ha, H.; Khuc, D.T.; Nguyen, D.Q.; von Meding, J.; Nguyen, L.P.; Luu, C. Landslide susceptibility prediction mapping with advanced ensemble models: Son La province, Vietnam. *Nat. Hazard* **2022**, 1–27.
11. Betgeri, S.N.; Vadyala, S.R.; Matthews, J.C.; Madadi, M.; Vladeanu, G. Wastewater pipe condition rating model using K–nearest neighbors. *Tunnelling Underground Space Technol.* **2023**, *132*, 104921.
12. Abu Alfeilat, H.A.; Hassanat, A.B.; Lasassmeh, O.; Tarawneh, A.S.; Alhasanat, M.B.; Eyal Salman, H.S.; Prasath, V.S. Effects of distance measure choice on k–nearest neighbor classifier performance: a review. *Big Data* **2019**, *7(4)*, 221–248.
13. Chakrabarty, N.; Kundu, T.; Dandapat, S.; Sarkar, A.; Kole, D.K. Flight arrival delay prediction using gradient boosting classifier. In: Emerging technologies in data mining and information security. *Springer* **2019**, 651–659.
14. Khan, M.S.I.; Islam, N.; Uddin, J.; Islam, S.; Nasir, M.K. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34(8)*, 4773–4781.
15. Lusa, L. Gradient boosting for high–dimensional prediction of rare events. *Comput. Stat. Data Anal.* **2017**, *113*, 19–37.
16. Roy, S.; Pandit, S.; Papia, M.; Rahman, M.M.; Ocampo, J.C.O.R.; Razi, M.A. Fraile–Jurado, P.; Ahmed, N.; Hoque, M.A.A.; Hasan, M.M. Coastal erosion risk assessment in the dynamic estuary: The Meghna estuary case of Bangladesh coast. *Int. J. Disaster Risk Reduct.* **2021**, *61*, 102364.

Landslide susceptibility mapping at Muong Cha district, Dien Bien Province, Vietnam province using machine learning classifiers K-Nearest-Neighbor and Gradient Boosting

Vu Cao Dat[1*], Nguyen Duc Dam[1], Pham Thai Binh[1]

[1] University of Transport and Technology, datvc@utt.edu.vn; binhpt@utt.edu.vn; damnd@utt.edu.vn

Abstract: In this research, the main objective is to build landslide susceptibility map at Muong Cha, Dien Bien province using classifiers such as K-Nearest-Neighbor (KNN) and Gradient Boosting (GB) - machine learning (artificial intelligence) techniques. Databese used in this study includes 206 past and present landslide locations and 10 landslide conditioning factors collected from various sources. To validate and compare the models, quantitative indicators including ROC curve and accuracy (%) were used. The results

showed that both KNN and GB performed well for landslide susceptibility modeling and mapping but the GB model outperforms the KNN model. Landslide susceptibility map constructed from the GB model with high performance can be used for effective land use planning and better landslide hazard management at the study area.

Keywords: Landslide; K–Nearest–Neighbor; Gradient Boosting; Dien Bien; Vietnam.