XÂY DỰNG BẢN ĐỒ CẢNH BÁO SẠT LỞ ĐẤT THEO THỜI GIAN THỰC CHO TỈNH LÀO CAI SỬ DỤNG CÁC NGUỒN DỮ LIỆU MỞ VÀ CÔNG NGHỆ HỌC MÁY

TRẦN MẠNH CƯỜNG[1], TRẦN ANH PHƯƠNG[1], NGUYỄN ANH ĐỨC[1], TRẦN VĂN TÚ[1], TRẦN BẢO CHUNG[1]

[1] Viện Khoa học tài nguyên nước

Tóm tắt

Trong bối cảnh sạt lở đất ngày càng gia tăng về tần suất và mức độ tàn phá trên địa bàn tỉnh Lào Cai – nơi có địa hình dốc và bị chia cắt mạnh, việc xây dựng một hệ thống cảnh báo sớm sạt lở đất là nhu cầu cấp thiết nhằm giảm thiểu thiệt hại về người và tài sản. Tuy nhiên, công tác cảnh báo hiện nay gặp nhiều khó khăn do hạn chế về dữ liệu quan trắc truyền thống. Nghiên cứu khai thác các nguồn dữ liệu mở theo thời gian thực kết hợp với mô hình học máy Random Forest – một thuật toán ensemble learning hiệu quả trong phân loại - để xây dựng bản đồ cảnh báo nguy cơ sạt lở đất cho tỉnh Lào Cai. Bộ dữ liệu đầu vào bao gồm điểm sạt lở quan sát được cùng các yếu tố điều kiện như độ dốc, hướng dốc, lượng mưa, độ ẩm đất, thổ nhưỡng, lớp phủ bề mặt, khoảng cách tới đường và sông suối, cùng với lượng mưa dự báo theo thời gian thực. Kết quả mô hình cho thấy, độ chính xác tổng thể đạt 85% và hệ số Kappa 0,69, các kết quả theo phương pháp ROC và PRC đạt độ tin cậy cao, chứng minh tính khả thi và hiệu quả của cách tiếp cận này. Kết quả nghiên cứu khẳng định tiềm năng ứng dụng học máy dựa trên dữ liệu nguồn mở trong việc phát triển hệ thống cảnh báo sớm sạt lở đất cho các khu vực miền núi còn hạn chế dữ liệu quan trắc như Lào Cai.

Từ khóa: *Sạt lở đất; Random Forest, dữ liệu nguồn mở, Lào Cai, cảnh báo nguy cơ.*

Ngày nhận bài: *2/8/2025****; Ngày sửa chữa:*** *28/8/2025****; Ngày duyệt đăng:*** *20/9/2025.*

Application of the random forest machine learning model in developing landslide susceptibility maps in Lao Cai province

Abstract

In the context of increasing frequency and destructive intensity of landslides in Lào Cai Province - an area characterized by steep and highly dissected terrain - the development of an early warning system for landslides is an urgent requirement to mitigate losses of life and property. However, current warning efforts face considerable challenges due to the limited availability of traditional observation data. This study leverages open-access, near real-time datasets in combination with the Random Forest machine learning algorithm - an efficient ensemble learning method for classification - to construct a landslide hazard warning map for Lào Cai Province. The input dataset includes observed landslide locations along with conditioning factors such as slope, aspect, rainfall, soil moisture, soil type, land cover, distance to roads and rivers, as well as real-time rainfall forecasts. Model evaluation shows an overall accuracy of 85% and a Kappa coefficient of 0.69, with ROC and PRC analyses indicating high reliability, thereby demonstrating the feasibility and effectiveness of this approach. The findings highlight the potential of open-source data-driven machine learning in developing early warning systems for landslide-prone mountainous regions where observational data remain scarce, such as Lào Cai.

Keywords: *Landslide; Random Forest; Lào Cai; Hazard Warning.*

JEL Classifications: *O13, Q15, R00.*

1. MỞ ĐẦU

Sạt lở đất, một hiểm họa địa chất tự nhiên phổ biến, gây ra những tổn thất đáng kể về người và tài sản trên toàn cầu [1]. Với địa hình nhiều đồi núi và khí hậu nhiệt đới ẩm gió mùa, Việt Nam là một quốc gia thường xuyên phải đối mặt với hiện tượng sạt lở đất, tập trung ở các tỉnh miền núi phía Bắc và khu vực miền Trung - Tây Nguyên [2][3]. Lào Cai, với đặc điểm địa hình dốc, bị chia cắt mạnh và chế độ mưa cường độ lớn, là một ví dụ điển hình về khu vực thường xuyên phải đối mặt với nguy cơ này [4].

Trong bối cảnh biến đổi khí hậu và mưa cực đoan ngày càng gia tăng, nhu cầu xây dựng các hệ thống cảnh báo sạt lở đất theo thời gian thực trở nên cấp thiết nhằm hỗ trợ công tác quản lý rủi ro thiên tai và đảm bảo an toàn cho cộng đồng. Các bản đồ cảnh báo sạt lở theo thời gian thực cho phép cập nhật nguy cơ theo điều kiện khí tượng tức thời, từ đó nâng cao hiệu quả

ứng phó so với các bản đồ nguy cơ truyền thống vốn chỉ thể hiện trạng thái tĩnh. Tại Việt Nam, các nghiên cứu về bản đồ cảnh báo sạt lở đất đã được thực hiện từ đầu những năm 2000, tuy nhiên chủ yếu là các bản đồ tĩnh. Những năm gần đây, một số nghiên cứu trong nước đã tiến hành sử dụng mô hình học máy để xây dựng bản đồ sạt lở đất. Trên thế giới, công tác cảnh báo sạt lở đất theo thời gian thực đã được quan tâm từ khá sớm, đặc biệt tại các quốc gia có địa hình phức tạp và chịu ảnh hưởng mạnh của biến đổi khí hậu.

Bản đồ cảnh báo sạt lở đất có thể được xây dựng dựa trên các phương pháp khác nhau như: Phân tích trọng số - GIS, mô hình thống kê, mô hình vật lý, mô hình học máy. Trong khi các phương pháp phân tích trọng số - GIS, mô hình thống kê, mô hình vật lý chỉ xây dựng được bản đồ cảnh báo sạt lở dưới dạng tĩnh thì phương pháp sử dụng công nghệ học máy cho phép cung cấp thông tin sạt lở theo thời gian thực.

Một trong những thách thức lớn trong công tác nghiên cứu và cảnh báo sạt lở đất tại các khu vực như Lào Cai là sự thiếu hụt dữ liệu chi tiết và đồng bộ, đặc biệt là dữ liệu là theo thời gian thực. Hạn chế về dữ liệu khiến cho công tác cảnh báo sạt lở gặp nhiều khó khăn, chưa đảm bảo độ chính xác và tin cậy. Trong bối cảnh đó, việc khai thác và tích hợp các nguồn dữ liệu mở, như dữ liệu địa hình số, lớp phủ bề mặt, thổ nhưỡng, mưa vệ tinh, đã mở ra nhiều cơ hội mới cho nghiên cứu sạt lở đất [5][6][7]. Đặc biệt, với sự phát triển của các công nghệ học máy (machine learning), việc ứng dụng các nguồn dữ liệu này trong công tác cảnh báo sạt lở đất ngày càng trở nên phổ biến. Nhiều công trình gần đây cho thấy các nguồn dữ liệu mở có thể được kết hợp hiệu quả với những mô hình học máy để nâng cao độ chính xác trong đánh giá và dự báo nguy cơ sạt lở, ngay cả ở những khu vực thiếu dữ liệu tại chỗ [8].

Các mô hình học máy như Random Forest (RF), SVM, XGBoost, Deep Learning… được sử dụng để kết hợp nhiều yếu tố địa hình, đất đá, thảm phủ, mưa vệ tinh, dữ liệu viễn thám, đã cho thấy độ chính xác cao trong xác định các vị trí có nguy cơ sạt lở đất. Trong số các phương pháp học máy, Random Forest (RF) đã chứng minh được tính ưu việt nhờ khả năng xử lý dữ liệu đa nguồn, kháng nhiễu tốt và cho kết quả ổn định trong các bài toán phân loại cũng như dự báo [9]. Mô hình RF có khả năng xác định mức độ quan trọng của từng yếu tố đầu vào, giúp làm rõ vai trò của các yếu tố như địa hình, thổ nhưỡng, lớp phủ thực vật hay lượng mưa trong việc kích hoạt sạt lở đất.

So với các mô hình thống kê truyền thống như hồi quy logistic, RF không đòi hỏi giả định phân bố dữ liệu và có khả năng mô hình hóa tốt các mối quan hệ phi tuyến, đa chiều giữa các yếu tố như địa hình, thổ nhưỡng, thảm phủ và lượng mưa. RF cho kết quả ổn định hơn trong không gian dữ liệu nhiều chiều và không yêu cầu quá trình chuẩn hóa phức tạp như với Support Vector Machine (SVM). Mặc dù mạng nơ-ron nhân tạo (ANN) có thể đạt độ chính xác cao nếu có tập dữ liệu lớn, nhưng thường tốn nhiều chi phí tính toán và thiếu tính minh bạch trong giải thích mô hình. RF có ưu thế về tốc độ huấn luyện và độ ổn định, đồng thời ít nhạy cảm với điều chỉnh siêu tham số so với các phương pháp boosting như XGBoost. Nhiều nghiên cứu quốc tế đã ứng dụng thành công RF để xây dựng bản đồ phân vùng nguy cơ và hệ thống cảnh báo sớm, cho thấy tiềm năng lớn của phương pháp này khi áp dụng tại các khu vực có dữ liệu hạn chế [10][11]. Tại Việt Nam, dù đã có những nghiên cứu sử dụng viễn thám và dữ liệu mở để phân vùng nguy cơ sạt lở [3], việc ứng dụng các mô hình học máy tiên tiến như RF trong cảnh báo theo thời gian gần thực vẫn còn hạn chế.

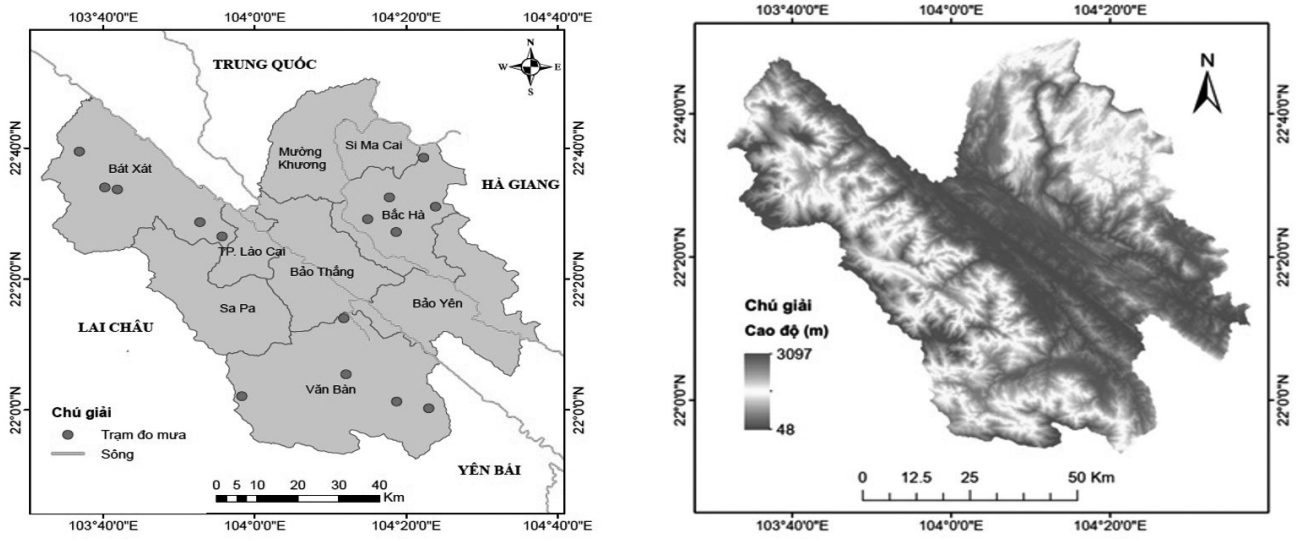
Hiện nay, nền tảng Google Earth Engine (GEE) đã được ứng dụng nhiều nơi trên thế giới để kết hợp mô hình học máy và các dữ viễn thám. Với thư viện dữ liệu mở toàn cầu (bao gồm ảnh vệ tinh, mô hình số độ cao, dữ liệu mưa vệ tinh, thảm phủ đất) và khả năng tính toán song song trên hạ tầng đám mây, GEE đã trở thành công cụ quan trọng trong nghiên cứu môi trường, tài nguyên và thiên tai. Việc triển khai RF trực tiếp trên GEE cho phép kết hợp dữ liệu đa nguồn, huấn luyện và dự báo trên phạm vi không gian rộng, đồng thời đảm bảo khả năng cập nhật liên tục và hiệu quả tính toán cao.

Bài báo tập trung vào việc kết hợp khai thác các nguồn dữ liệu mở toàn cầu với ứng dụng mô hình học máy RF để đánh giá và dự báo nguy cơ sạt lở đất tại tỉnh Lào Cai theo thời gian thực. Với khả năng phân tích mạnh mẽ của RF, nghiên cứu hướng đến việc xây dựng một công cụ hỗ trợ hữu ích cho công tác phòng chống và giảm thiểu rủi ro thiên tai, đồng thời đóng góp một khung phương pháp có thể mở rộng và áp dụng cho nhiều khu vực miền núi khác có điều kiện tương tự.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Giới thiệu khu vực nghiên cứu

Lào Cai là tỉnh miền núi phía Tây Bắc, phía Đông giáp tỉnh Hà Giang, phía Tây giáp tỉnh Lai Châu, phía Nam giáp tỉnh Yên Bái, phía Bắc giáp tỉnh Vân Nam (Trung Quốc) (Hình 1). Tọa độ địa lý của tỉnh là từ 21°40’56" đến 22°52’ vĩ độ Bắc và từ 103°30’24" đến 104°38’21" kinh độ Đông. Là tỉnh vùng cao biên giới, nằm chính giữa vùng Đông Bắc và vùng Tây Bắc của nước ta, Lào Cai cách Hà Nội 296 km theo đường sắt và 265 km theo đường bộ. Diện tích tự nhiên của toàn tỉnh là 6.383,88 km^2 (chiếm 2,44% diện tích cả nước, là tỉnh có diện tích lớn thứ 19/64 tỉnh, thành phố cả nước).



(a) *(a)*

Hình 1. (a) Bản đồ hành chính tỉnh Lào Cai và (b) bản đồ địa hình tỉnh Lào Cai

2.2. Dữ liệu nghiên cứu

Quá trình xây dựng một mô hình phân loại sạt lở bắt đầu với việc thu thập dữ liệu đào tạo. Dữ liệu đầu vào bao gồm các dữ liệu tĩnh và dữ liệu động (Bảng 1):

- *Dữ liệu tĩnh:* Là dữ liệu ít biến đổi theo thời gian. Các dữ liệu này bao gồm: Độ cao, độ dốc địa hình, phương diện, độ cong mặt phẳng (bề mặt đất), độ cong theo phương dọc (bề mặt đất), thảm phủ bề mặt, chỉ số NDVI, khoảng cách tới sông, khoảng cách tới đường.

- *Dữ liệu động:* Là các dữ liệu biến đổi liên tục theo thời gian thực bao gồm: Độ ẩm đất, lượng mưa 1 ngày, lượng mưa 3 ngày và lượng mưa 15 ngày và cường độ mưa trung bình 3 giờ tới.

Bảng 1. Các dữ liệu đầu vào cho mô hình học máy

Yếu tố	Mô tả	Thời gian thu thập	Nguồn
Độ cao	Độ cao bề mặt địa hình	2000	NASA
Độ dốc địa hình	Góc nghiêng của sườn đất	2000	NASA
Phương diện	Hướng của sườn đất	2000	NASA
Độ cong mặt phẳng	Độ cong vuông góc với sườn đất, chỉ ra bề mặt lõm hoặc lồi	2000	NASA
Độ cong theo phương dọc	Độ cong song song với sườn đất, chỉ ra bề mặt lõm hoặc lồi	2000	NASA
Thảm phủ bề mặt	Đối tượng trên bề mặt đất	2024	ESA
Chỉ số NDVI	Mức độ che phủ của thảm thực vật	2024	ESA
Khoảng cách tới sông	Khoảng cách tới sông		
Khoảng cách tới đường	Khoảng cách tới đường		
Độ ẩm đất	Lượng nước trong đất	2024	NASA
Lượng mưa 1 ngày tới thời điểm dự báo	Lượng mưa tích lũy trong 1 ngày trước thời điểm dự báo	9/9/2024	CHIRPS
Lượng mưa 3 ngày tới thời điểm dự báo	Lượng mưa tích lũy trong 3 ngày trước thời điểm dự báo	7-9/9/2024	CHIRPS
Lượng mưa 15 ngày tới thời điểm dự báo	Lượng mưa tích lũy trong 7 ngày trước thời điểm dự báo	26/8-9/9/2024	CHIRPS
Cường độ mưa dự báo 3 ngày tới	Cường độ mưa (mm/h) dự báo trong 3 ngày tới	10/9-12/9/2024	GFS
Vị trí sạt lở trong thực tế trong đợt mưa lũ 8-10/9/2024		8-10/9/2024	Thực địa

Dữ liệu độ cao mặt đất được thu thập từ dữ liệu mô hình số độ cao (DEM) của tổ chức NASA, đây là dữ liệu có độ tin cậy cao và được sử dụng rộng rãi trong nhiều nghiên cứu trên thế giới. Các dữ liệu độ dốc địa hình, phương diện, độ cong mặt phẳng (bề mặt đất), độ cong theo phương dọc (bề mặt đất) được tính từ dữ liệu độ cao. Dữ liệu thảm phủ bề mặt được thu thập từ World Cover 10m, và dữ liệu NDVI được tính toán từ ảnh vệ tinh Sentinel 2, cả hai nguồn dữ liệu này đều được cung cấp bởi tổ chức vũ trụ châu Âu (ESA). Khoảng cách tới sông và khoảng cách tới đường được xây dựng dựa vào khoảng cách từ ô lưới bất kỳ đến mạng lưới sông, đường được thu thập trên địa bàn tỉnh Lào Cai. Dữ liệu độ ẩm đất được thu thập từ nguồn SPL4SMGP.007 SMAP L4 Global của NASA. Dữ liệu mưa được thu thập từ nguồn CHIRPS Daily, được hiệu chỉnh với số liệu mưa thực đo. Lượng mưa 1 ngày, lũy tích 7 ngày và 15 ngày cho từng ô lưới được tính từ dữ liệu mưa CHIRPS này.

Độ dốc là yếu tố địa hình quan trọng nhất ảnh hưởng đến sự ổn định mái dốc. Khi độ dốc càng lớn, thành phần lực trượt song song mặt dốc tăng, trong khi lực chống trượt giảm, làm nguy cơ sạt lở gia tăng. Phương diện ảnh hưởng đến sự phân bố mưa, bức xạ mặt trời và thảm thực vật, các sườn dốc hướng nắng nhiều có thể làm đất khô, giảm độ ổn định của kết cấu đất đá theo thời gian. Độ cong mặt phẳng phản ánh sự phân kỳ hoặc hội tụ của dòng chảy bề mặt, khi dòng chảy hội tụ, dễ tích tụ nước, bão hòa và làm giảm ổn định mái dốc. Độ cong theo phương dọc thể hiện sự thay đổi độ dốc dọc theo chiều dòng chảy, khi bề mặt lõm, tăng tốc dòng chảy, tập trung năng lượng xói mòn, dễ gây mất ổn định và kích hoạt sạt lở. Thảm phủ bề mặt (loại đất sử dụng, rừng, nông nghiệp, đất trống…) quyết định khả năng bảo vệ mái dốc và NDVI là chỉ số phản ánh độ che phủ xanh trên bề mặt đất. Khu vực gần sông, suối chịu tác động xói lở chân dốc và bão hòa nước do mực nước thay đổi, làm giảm ổn định sườn dốc, do đó, nguy cơ sạt lở thường cao ở những vị trí gần sông. Các hoạt động làm đường thường cắt xẻ taluy, thay đổi địa hình và kết cấu mái dốc, gây mất ổn định, vùng gần đường thường có nguy cơ sạt lở cao hơn so với vùng xa đường. Độ ẩm cao làm tăng áp lực lỗ rỗng, giảm lực ma sát và lực dính kết của đất → mái dốc dễ bị trượt. Các nghiên cứu cho thấy sạt lở thường xảy ra khi độ ẩm đạt tới ngưỡng bão hòa. Mưa là tác nhân kích hoạt chính của sạt lở ở vùng nhiệt đới gió mùa. Cơ chế ảnh hưởng của mưa đến sự ổn định mái dốc chủ yếu thông qua hai quá trình: (i) Thấm nước vào đất làm tăng áp lực lỗ rỗng, từ đó giảm lực ma sát và lực dính kết của hạt đất đá; (ii) tích tụ nước trên bề mặt và trong tầng phong hóa gây bão hòa, gia tăng trọng lượng khối trượt và giảm khả năng chống cắt của vật liệu. Mưa cường độ lớn trong thời gian ngắn thường kích hoạt các vụ sạt lở quy mô nhỏ nhưng diễn ra nhanh, trong khi mưa kéo dài nhiều ngày có thể gây bão hòa diện rộng, dẫn tới những sự kiện sạt lở nghiêm trọng hơn.

2.3. Phương pháp nghiên cứu

a. Tổng quan về mô hình Random Forest

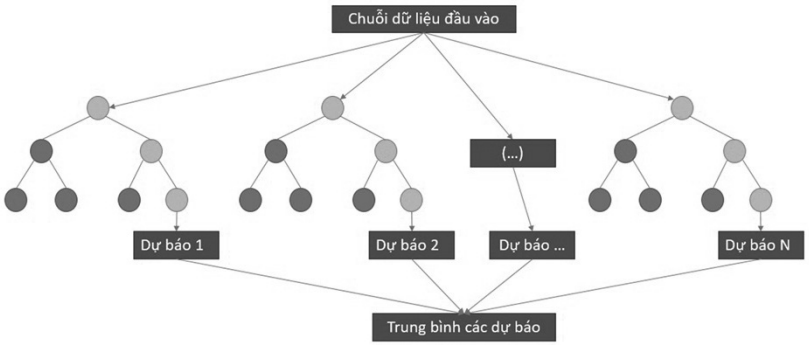
Một trong những yêu cầu cảnh báo sạt lở đòi hỏi thời gian dự báo nhanh để có thời gian chuẩn bị ứng phó với sạt lở dài hơn. So với phương pháp mô hình số trị, phương pháp sử dụng mô hình học máy cho kết quả dự báo tức thì và do đó tăng thời gian dự báo. Trong nghiên cứu này, thuật toán Random Forest được sử dụng. Đây là một thuật toán học máy thuộc nhóm mô hình cây quyết định (decision tree), hoạt động bằng cách tạo ra một tập hợp các cây quyết định (decision trees) độc lập, và kết quả dự báo được lấy là trung bình của các cây quyết định. Sơ đồ các bước triển khai thuật toán này được trình bày ở Hình 2. RF cho thấy hiệu quả hơn so với các phương pháp phân loại thường được sử dụng vì có khả năng tìm ra thuộc tính nào quan trọng hơn so với những thuộc tính khác. RF là một phương pháp thống kê mô hình hóa bằng máy (machine learning statistic) dùng để phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác bằng cách xây dựng nhiều cây quyết định (Decision tree). Cây quyết định bao gồm ba thành phần: Nút quyết định, nút lá và nút gốc. Thuật toán cây quyết định chia tập dữ liệu huấn luyện thành các nhánh, các nhánh này sẽ tách biệt thành các nhánh nhỏ hơn. Trình tự này tiếp tục cho đến khi đạt được một nút lá. Nút lá không thể được phân tách thêm. Các nút trong cây quyết định đại diện cho các thuộc tính được sử dụng để dự đoán kết quả. Các nút quyết định cung cấp một liên kết đến các lá.

RF thiết lập kết quả dựa trên các dự đoán của cây quyết định, dự đoán bằng cách lấy giá trị trung bình của kết quả từ các cây khác nhau. Tăng số lượng cây làm tăng độ chính xác kết quả của phương pháp RF. Quá trình huấn luyện của RF cho các cây ra quyết định diễn ra trong máy phân loại RF. Mọi cây ra quyết định đều bao gồm các nút quyết định, nút nhánh và nút gốc. Nút nhánh của mỗi cây là đầu ra cuối cùng do cây ra quyết định cụ thể tạo ra. Việc lựa chọn đầu ra cuối cùng tuân theo đa số. Phương pháp hoạt động của RF theo 4 bước sau:

(1) Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho;

(2) Thiết lập cây quyết định cho từng đối tượng và nhận kết quả dự đoán từ mỗi quyết định cây;

(3) Xác lập kết quả dự đoán;



Hình 2. Sơ đồ áp dụng thuật toán Random Forest cảnh báo sạt lở

(4) Chọn kết quả cuối cùng là kết quả trung bình của các cây quyết định (Hình 2).

Random Forest có nhiều ưu điểm:

- Khả năng xử lý dữ liệu không đồng nhất: Nó có thể xử lý tốt các dữ liệu có cấu trúc khác nhau, bao gồm dữ liệu phân loại và hồi quy.

- Khả năng giảm thiểu overfitting: Random Forest thông qua việc xây dựng nhiều cây quyết định và trung bình hóa kết quả giúp giảm bớt hiện tượng overfitting mà một cây quyết định đơn lẻ có thể gặp phải.

- Khả năng đánh giá các yếu tố quan trọng: Random Forest có thể giúp đánh giá mức độ quan trọng của các yếu tố đầu vào trong việc dự báo nguy cơ sạt lở.

b. Đầu vào cho mô hình Random Forest

Mô hình Random Forest cần một tập hợp các yếu tố đầu vào (features) để xây dựng các cây quyết định. Với bài toán cảnh báo sạt lở, đầu vào của mô hình RF là các yếu tố gây sạt lở. Các yếu tố này có thể bao gồm:

- Độ ẩm đất: Sử dụng thông tin độ ẩm đất ở các thời điểm khác nhau (theo ngày, tuần hoặc tháng).

- Lượng mưa: Lượng mưa trong khu vực nghiên cứu có thể được thu thập từ các dữ liệu quan trắc hoặc các mô hình khí hậu. Lượng mưa bao gồm cả lượng mưa lũy tích từ các ngày trước đó và cường độ mưa dự báo tại thời điểm dự báo sạt lở.

- Địa hình: Các dữ liệu địa hình bao gồm (Độ cao, độ dốc địa hình, phương diện, độ cong mặt phẳng (bề mặt đất), độ cong theo phương dọc (bề mặt đất). Các dữ liệu này được tính toán từ dữ liệu DEM.

- Loại đất và thành phần đất: Các thông số này có thể được lấy từ cơ sở dữ liệu đất đai (ví dụ: FAO, USDA).

- Lớp phủ thực vật: Thông tin về lớp phủ thực vật có thể giúp xác định mức độ bảo vệ của cây cối đối với sườn đất.

- Khoảng cách đến sông: Khoảng cách từ ô lưới đang xem xét đến sông.

- Khoảng cách đến đường giao thông: Khoảng cách từ ô lưới xem xét đến đường giao thông.

c. Xây dựng và huấn luyện mô hình Random Forest

Sau khi chuẩn bị dữ liệu đầu vào, mô hình Random Forest sẽ được huấn luyện với các dữ liệu đầu vào này. Quá trình huấn luyện bao gồm:

- Chia dữ liệu thành tập huấn luyện và tập kiểm tra (train-test split).

- Huấn luyện mô hình trên tập huấn luyện với các yếu tố đầu vào.

- Điều chỉnh các tham số của mô hình Random Forest (số lượng cây, độ sâu của cây, tỷ lệ phân chia…) để đạt được kết quả tốt nhất.

d. Đánh giá mô hình

Sau khi huấn luyện, mô hình cần được đánh giá bằng cách sử dụng tập kiểm tra để xem độ chính xác của dự báo. Các chỉ số đánh giá có thể bao gồm:

- Accuracy: Đo lường độ chính xác của mô hình trong việc phân loại khu vực có nguy cơ sạt lở.

- Precision, Recall và F1-score: Những chỉ số này giúp đánh giá khả năng của mô hình trong việc dự báo đúng các khu vực có nguy cơ sạt lở (chính xác) mà không bỏ sót.

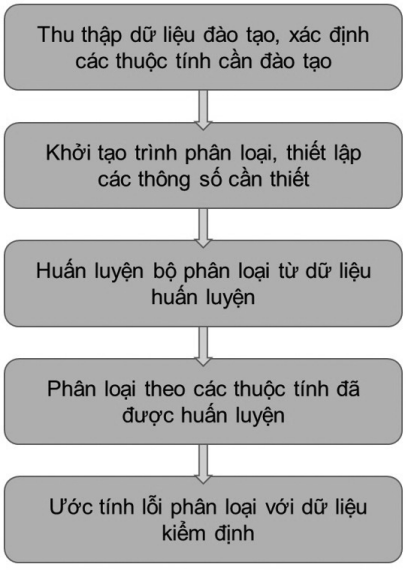
- ROC-AUC: Đo lường khả năng phân biệt giữa các khu vực có và không có nguy cơ sạt lở.

Do các dữ liệu để triển khai mô hình học máy là các dữ liệu bản đồ (dạng ô lưới). Do đó, để thuận tiện cho việc xử lý dữ liệu, nghiên cứu triển khai huấn luyện mô hình học máy cảnh báo sạt lở đất trên nền tảng Google Earth Engine sử dụng ngôn ngữ lập trình Python để tận dụng khả năng phân tích không gian nhanh chóng và CSDL địa lý có sẵn trên Google (Hình 3).

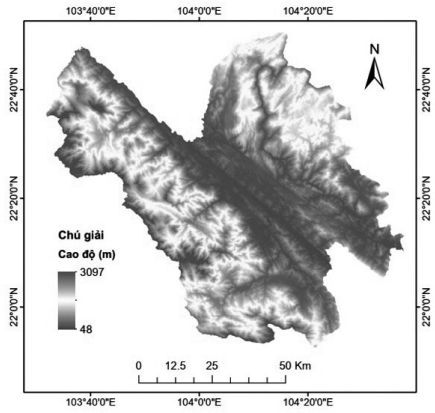
3. KẾT QUẢ VÀ THẢO LUẬN

1) Thiết lập dữ liệu đào tạo, xác định các thuộc tính cần đào tạo

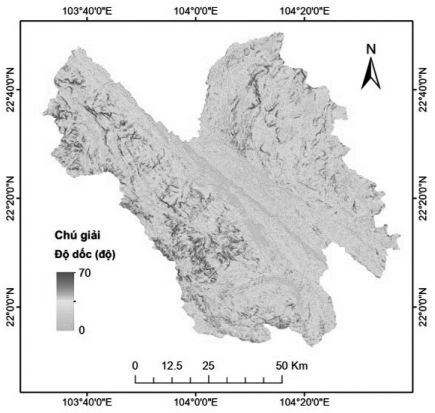
Hình 4 và Hình 5 tương ứng minh họa dữ liệu tĩnh vào dữ liệu động được sử dụng trong nghiên cứu này. Các dữ liệu này được truy xuất, tính toán và xử lý hoàn toàn trên GEE.



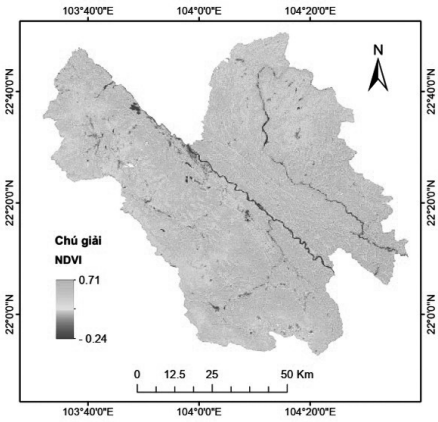
Hình 3. Sơ đồ xây dựng mô hình học máy cảnh báo sạt lở



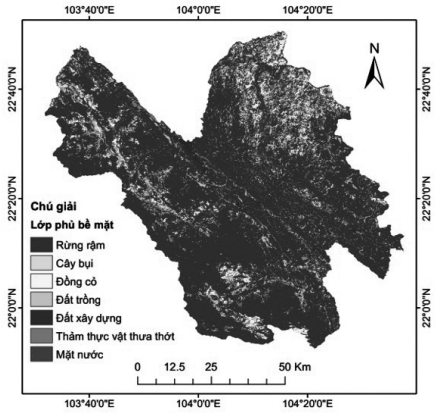
DEM



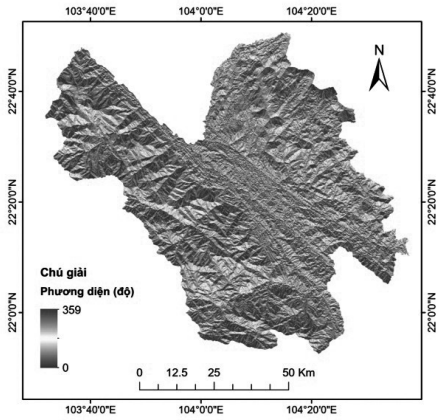
Độ dốc



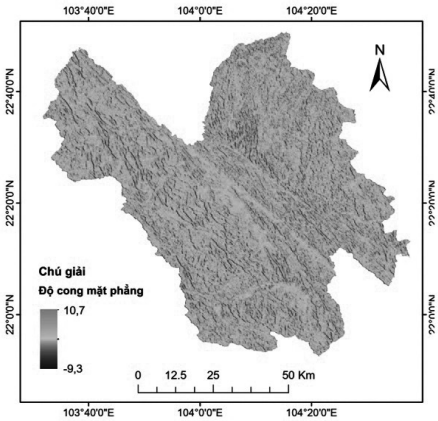
NDVI



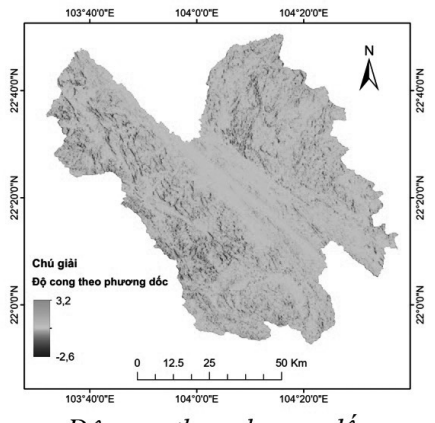
Thảm phủ



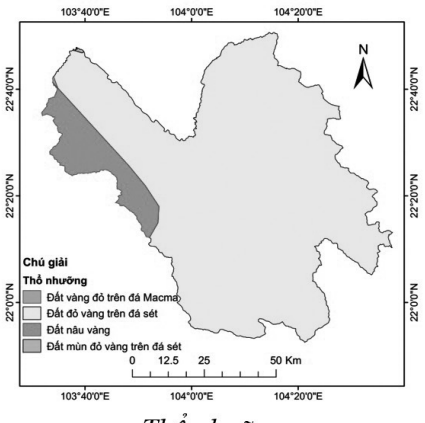
Phương diện



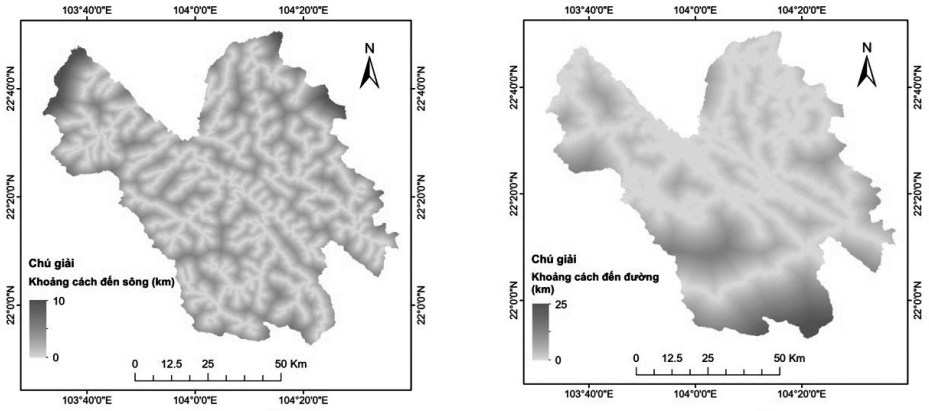
Độ cong mặt phẳng



Độ cong theo phương dốc

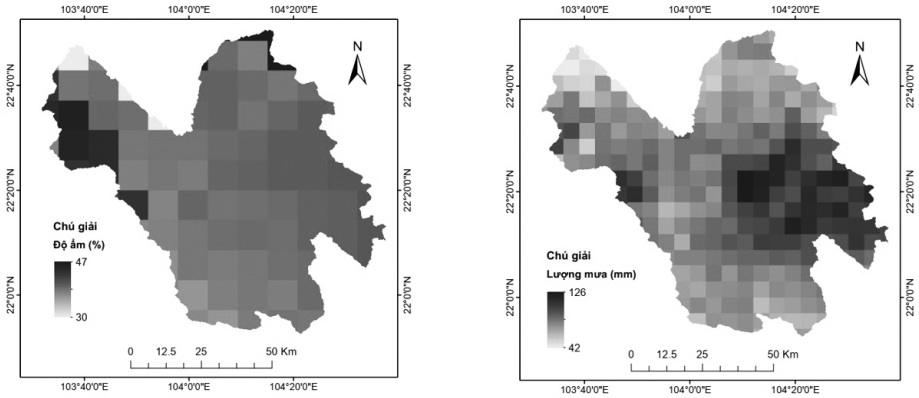


Thổ nhưỡng



Khoảng cách đến sông *Khoảng cách đến đường*

Hình 4. Một số dữ liệu đầu vào tĩnh của mô hình học máy cảnh báo sạt lở



Độ ẩm đất SMAP Dữ liệu mưa vệ tinh CHIRPS

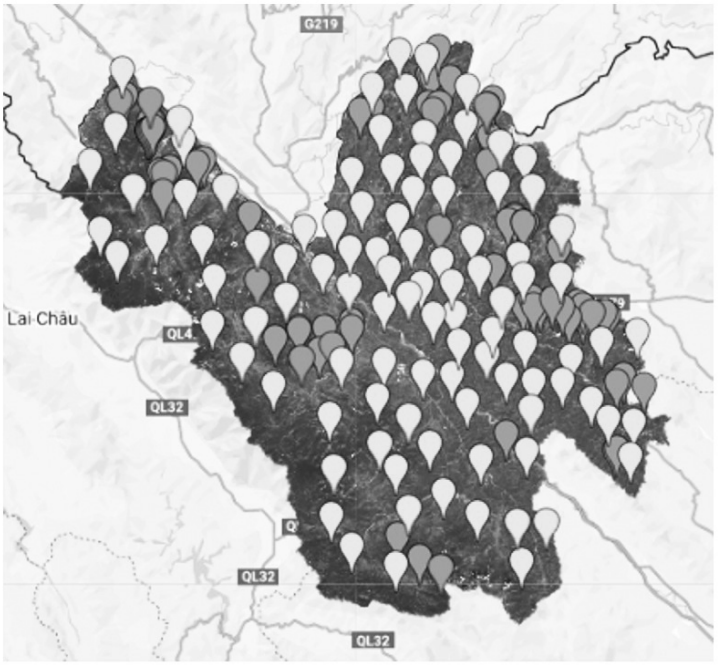
Hình 5. Minh họa dữ liệu đầu vào động của mô hình học máy (độ ẩm đất SMAP và dữ liệu mưa vệ tinh CHIRPS)

Nhiệm vụ của mô hình học máy là xác định xem với các dữ liệu đầu vào như trên, 1 vị trí (ô lưới) có khả năng bị sạt lở hay không. Để làm được điều này mô hình học máy cần phải được huấn luyện và kiểm tra sử dụng mẫu số liệu trong quá khứ. Ở nghiên cứu này các vị trí (ô lưới) sạt lở và không sạt lở đất (tổng cộng 180 vị trí) được trên địa bàn tỉnh Lào Cai (Hình 6). Tại mỗi vị trí, các yếu tố đầu ở Bảng 1 vào sẽ được trích xuất từ các bản đồ tương ứng (Hình 6).

2) Khởi tạo trình phân loại, thiết lập các thông số cần thiết

Ở bước này, thuật toán Random Forest được khởi tạo trong GEE để xây dựng mô hình phân loại:

- Khởi tạo bộ phân loại Random Forest: Trong GEE, có thể sử dụng phương thức ee.Classifier.smileRandomForest để khởi tạo bộ phân loại. Random Forest là một thuật toán học máy mạnh mẽ dựa trên việc tạo ra nhiều cây quyết định và sử dụng phương pháp bỏ phiếu để phân loại.



Hình 6. Bản đồ 180 vị trí mẫu phục vụ huấn luyện và kiểm tra mô hình học máy

- Thiết lập các thông số: Các thông số cần thiết cho thuật toán Random Forest bao gồm:

+ Số lượng cây quyết định (number of trees): Đây là số lượng cây quyết định trong Random Forest. Thông thường, một giá trị từ 100 đến 500 cây sẽ giúp đạt được hiệu quả tốt.

+ Số lượng thuộc tính đầu vào (number of features per tree): Số lượng thuộc tính được sử dụng tại mỗi nút phân tách trong mỗi cây quyết định.

+ Số lượng điểm mẫu (sample size): Lựa chọn số lượng mẫu từ mỗi lớp dữ liệu trong quá trình huấn luyện.

3) Huấn luyện bộ phân loại từ dữ liệu huấn luyện

Trong bước này, bộ phân loại Random Forest sẽ được huấn luyện từ dữ liệu đào tạo đã thu thập:

- Xác định lớp và nhãn dữ liệu: Các điểm sạt lở và không sạt lở sẽ được phân loại vào các lớp khác nhau. Mỗi điểm trong bộ dữ liệu đào tạo sẽ có nhãn tương ứng (ví dụ: 1 cho sạt lở và 0 cho không sạt lở).

- Áp dụng thuật toán học máy: Sử dụng phương thức classifier.train() để huấn luyện bộ phân loại với dữ liệu đào tạo. Thuật toán Random Forest sẽ học cách phân loại các điểm sạt lở dựa trên các thuộc tính (như độ cao, độ dốc, NDVI, ...).

- Tạo mô hình phân loại: Sau khi huấn luyện, mô hình Random Forest sẽ được tạo ra, có khả năng phân loại các điểm sạt lở mới từ các thuộc tính môi trường.

4) Phân loại theo các thuộc tính đã được huấn luyện

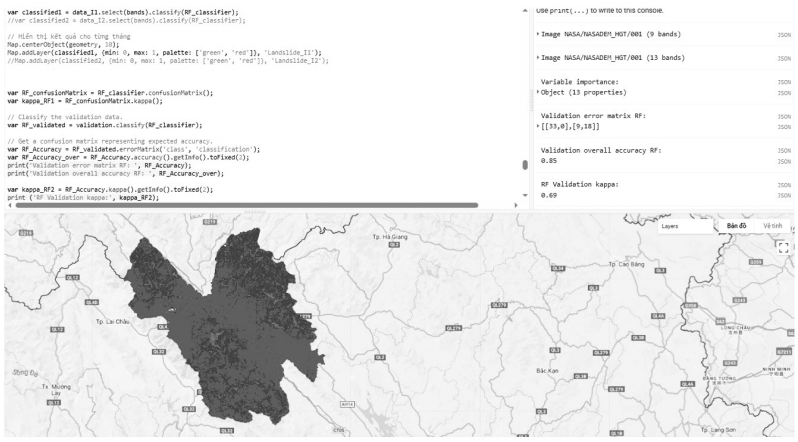
Khi mô hình đã được huấn luyện, mô hình này có thể được sử dụng để phân loại các khu vực sạt lở trên bản đồ:

- Áp dụng phân loại: Sử dụng phương thức classifier.classify để phân loại các vùng chưa biết (vùng cần phân tích sạt lở). Dữ liệu đầu vào sẽ là các thuộc tính môi trường như độ cao, độ dốc, độ che phủ thực vật…, và mô hình đã huấn luyện sẽ phân loại khu vực đó là sạt lở hay không sạt lở.

- Kết quả phân loại: Sau khi phân loại xong, bản đồ phân loại sẽ được tạo ra, trong đó các khu vực sạt lở sẽ được xác định rõ ràng.

5) Ước tính lỗi phân loại với dữ liệu kiểm định

Để đánh giá hiệu quả của mô hình phân loại, cần thực hiện bước kiểm định:



Hình 7. Giao diện nền tảng GEE thực hiện tính toán, xây dựng bản đồ cảnh báo sạt lở đất

- Tạo dữ liệu kiểm định: Dữ liệu kiểm định là một bộ dữ liệu độc lập không tham gia vào quá trình huấn luyện. Bộ dữ liệu này thường được chia từ bộ dữ liệu gốc gồm 70% dữ liệu huấn luyện và 30% dữ liệu kiểm định.

- Ước tính độ chính xác sử dụng các phương thức như confusionMatrix và các chỉ số như độ chính xác (accuracy) để đánh giá hiệu quả của mô hình.

- Phân tích kết quả: Dựa trên ma trận nhầm lẫn và các chỉ số đánh giá, có thể điều chỉnh các tham số của mô hình hoặc cải thiện dữ liệu đầu vào để tăng cường độ chính xác và khả năng phân loại.

Quá trình tính toán cho thấy mô hình học máy có khả năng phân loại tốt với độ chính xác tổng thể đạt 85% và hệ số Kappa là 0,69 (hệ số Kappa trong khoảng từ 0,61 đến 0,8 thì mô hình được đánh giá là tốt). Các kết quả đánh giá theo ROC và SRC-PRC đều đạt kết quả tốt. Tuy nhiên, cần lưu ý là độ chính xác trên được xây dựng dựa trên các yếu tố đầu vào động (độ ẩm đất và lượng mưa) là các yếu tố quan trắc từ viễn thám. Các sai số do mưa dự báo trong vận hành thực tế chưa được tính đến.

Hình 7 thể hiện bản đồ cảnh báo sạt lở cho tỉnh Lào Cai trong đợt mưa lũ ngày 8-10/9/2024. Kết quả tính toán sạt lở đất cho thấy các vị trí có khả năng sạt lở tập trung chủ yếu ở các huyện Bát Xát, Sa Pa, Văn Bàn, Mường Khương, Si Ma Cai, Bắc Hà. Đây là những khu vực có địa hình đồi núi với độ dốc lớn. Các huyện, thành phố khác gồm thành phố Lào Cai, Bảo Thắng, Bảo Yên có ít vị trí có khả năng sạt lở hơn (Bảng 2, Hình 8).

Trong khi các nghiên cứu xây dựng bản đồ cảnh báo, dự báo sạt lở đất đã được thực hiện trước đây chủ yếu xây dựng cảnh báo theo phạm vi khu vực hành chính (thường là cấp xã) thì nghiên cứu trong bài báo cung cấp phương pháp mới mang tính linh động hơn. Nghiên cứu được thực hiện cung cấp thông tin cảnh báo sạt lở với phạm vi nhỏ hơn, thuộc một khu vực trong xã, hoặc vùng liên xã. Tuy vậy, phạm vi cảnh báo có độ chi tiết chưa cao do độ

phân giải của dữ liệu đầu vào không cao. Trong tương lai, khi các nghiên cứu tiếp theo có thể tiếp cận với các nguồn dữ liệu thương mại – nguồn dữ liệu có độ phân giải cao thì kết quả phạm vi cảnh báo sẽ được chi tiết hơn.

4. KẾT LUẬN

Nghiên cứu đã chứng minh tính khả thi và hiệu quả của việc ứng dụng mô hình học máy Random Forest trong xây dựng bản đồ cảnh báo nguy cơ sạt lở đất tại tỉnh Lào Cai. Với nguồn dữ liệu mở từ viễn thám và các đặc trưng địa hình – khí tượng (như độ dốc, lượng mưa, độ ẩm đất, thảm phủ, thổ nhưỡng, khoảng cách tới sông và đường giao thông), mô hình đã đạt được độ chính xác tổng thể 85% và hệ số Kappa 0,69, các kết quả theo phương pháp ROC và PRC đạt độ tin cậy cao, phản ánh khả năng phân loại tốt.

Kết quả thử nghiệm cho thấy, các khu vực có nguy cơ sạt lở cao tập trung ở những huyện có địa hình dốc và bị chia cắt mạnh như Bát Xát, Sa Pa, Văn Bàn, Mường Khương, Si Ma Cai và Bắc Hà. Điều này phù hợp với thực tế quan sát, khẳng định tính tin cậy của mô hình trong công tác cảnh báo. Nghiên cứu góp phần mở ra hướng tiếp cận mới trong việc ứng dụng học máy kết hợp dữ liệu mở cho cảnh báo sớm sạt lở đất, đặc biệt hữu ích đối với các khu vực miền núi còn hạn chế về dữ liệu quan trắc truyền thống. Trong tương lai, việc tích hợp thêm các nguồn dữ liệu thời gian thực và hiệu chỉnh sai số mưa dự báo có thể giúp nâng cao hơn nữa độ chính xác và khả năng ứng dụng của hệ thống cảnh báo. Ngoài ra, khi các nghiên cứu tiếp theo có thể tiếp cận với các nguồn dữ liệu thương mại – nguồn dữ liệu có độ phân giải cao thì kết quả phạm vi cảnh báo sẽ được chi tiết hơn.

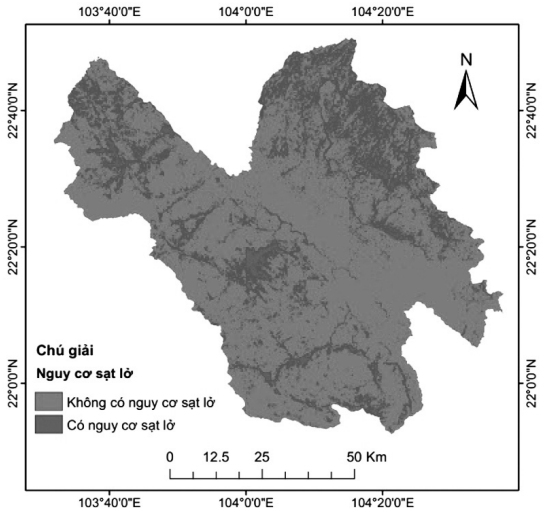
Đóng góp của tác giả: *Xây dựng ý tưởng và phương pháp nghiên cứu: Trần Anh Phương, Trần Văn Tú, Trần Mạnh Cường, Nguyễn Anh Đức, Trần Bảo Chung; Xử lý số liệu: Trần Văn Tú, Trần Mạnh Cường, Trần Bảo Chung; Tính toán: Trần Bảo Chung; Phân tích kết quả và viết bản thảo bài báo: Trần Anh Phương, Trần Văn Tú, Trần Mạnh Cường; Chỉnh sửa bài báo: Trần Văn Tú, Trần Mạnh Cường; Kiểm soát, định hướng nghiên cứu: Trần Anh Phương, Nguyễn Anh Đức.*

Lời cảm ơn: *Bài báo dựa trên kết quả Đề tài khoa học công nghệ cấp Bộ "Nghiên cứu công nghệ tích hợp số liệu độ ẩm đất viễn thám SMAP, mô hình mô phỏng dòng chảy và biến động sườn dốc phục vụ xây dựng hệ thống cảnh báo sạt lở đất và lũ bùn đá" do Viện Khoa học tài nguyên nước chủ trì thực hiện.*

Lời cam đoan: *Tập thể tác giả cam đoan bài báo, chưa từng công bố trước đó, không sao chép, đạo văn; không có sự tranh chấp lợi ích trong nhóm tác giả*■

Bảng 2. Bảng kết quả đánh giá độ chính xác của kết quả huấn luyện mô hình Random Forest

	Không sạt lở	Sạt lở	Tổng
Không sạt lở	33	0	33
Sạt lở	9	18	27
Tổng	42	18	60
Độ chính xác tổng thể	85%		
Hệ số Kappa	0,69		
ROC	FPR = 0; TPR = 0,667		
PRC	Precision = 1.00, Recall = 0.667		



Hình 8. Bản đồ cảnh báo sạt lở tỉnh Lào Cai tính toán thử nghiệm cho đợt mưa lũ tháng 9/2024

TÀI LIỆU THAM KHẢO

1. Alcántara-Ayala, I. (2025). Landslides in a changing world. Landslides, 1-15.

2. Đạt, V. C., Đảm, N. Đ., & Bình, P. T. 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.

3. Tuấn, H. N., & Tuyết, V. 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.

4. Van Tran, A., Nguyen, B. A., Dinh, T., Nguyen, Y. H. T., & Le, N. T. (2020). Landslides detection in Bat Xat district, Lao Cai province, Vietnam using the Alos PalSAR time-series imagery by the SBAS method. Journal of Mining and Earth Sciences Vol, 61(4), 1-10.

5. Arnone, E., Noto, L. V., Lepore, C., & Bras, R. L. (2011). Physically-based and distributed approach to analyze rainfall-triggered landslides at watershed scale. Geomorphology, 133(3-4), 121-131.

6. Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a

new environmental record for monitoring extremes. Scientific data, 2(1), 1-21.
7. García, M., Riaño, D., Chuvieco, E., Salas, J., & Danson, F. M. (2011). Multispectral and LiDAR data fusion for fuel type mapping using Support Vector Machine and decision rules. Remote Sensing of Environment, 115(6), 1369-1379.
8. Segoni, S., Lagomarsino, D., Fanti, R., Moretti, S., & Casagli, N. (2015). Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system. Landslides, 12, 773-785.
9. Yaser Peiro 1,2 , Evelina Volpe 1, Luca Ciabatta 2 and Elisabetta Cattoni 1, (2024). High Resolution Precipitation and Soil Moisture Data Integration for Landslide Susceptibility Mapping.*
10. Pack, R. T., Tarboton, D. G., & Goodwin, C. N. (1998). SINMAP, a stability index approach to terrain stability hazard mapping. SINMAP user's manual, Terratech Consulting Ltd.
11. Baum, R.L., Savage, W.Z., & Godt, J.W. (2002). TRIGRS—A Fortran program for transient rainfall infiltration and grid-based regional slope-stability analysis. USGS Open-File Report 02–424.
12. Baum, R. L., Savage, W. Z., & Godt, J. W. (2008). TRIGRS: a Fortran program for transient rainfall infiltration and grid-based regional slope-stability analysis, version 2.0 (p. 75). Reston, VA, USA: US Geological Survey.
13. Dikshit, A., Satyam, N., & Pradhan, B. (2019). Estimation of rainfall-induced landslides using the TRIGRS model. Earth Systems and Environment, 3, 575-584.
14. Bordoni, M., Meisina, C., Valentino, R., Bittelli, M., & Chersich, S. (2015). Site-specific to local-scale shallow landslides triggering zones assessment using TRIGRS. Natural Hazards and Earth System Sciences, 15(5), 1025-1050.
15. Fang, L. (2025). Prediction of Rainfall-Induced Landslide-Mudslide Hazard Chain Using Coupled TRIGRS and RAMMS Models. International Journal of Information System Modeling and Design (IJISMD), 16(1), 1-23.
16. Đức, Đ. N., Thanh, T. N., Văn, P. T., & Thái, B. P. (2022). 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, 36-56.
17. Bảo, H. Đ., Anh, N. C. T., & Ngọc, Q. N. T. (2023). Phân vùng và dự báo nguy cơ sạt lở đất trên địa bàn huyện Krông Bông, tỉnh Đắk Lắk trong điều kiện biến đổi khí hậu. Tạp chí Khoa học Tây Nguyên, 17(59).
18. Ngọc, T. T. H., Long, Đ. V., & Công, N. C. (2022). Dự báo nguy cơ trượt lở đất cho huyện A Lưới, tỉnh Thừa Thiên Huế sử dụng mô hình Logistic regression. Tạp chí Khoa học và Công nghệ-Đại học Đà Nẵng, 5-10.
19. Gessler, P. E., Chadwick, O. A., Chamran, F., Althouse, L., & Holmes, K. (2000). Modeling soil–landscape and ecosystem properties using terrain attributes. Soil Science Society of America Journal, 64(6), 2046–2056.
20. Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (1997). The soil production function and landscape equilibrium. Nature, 388(6640), 358–361.
21. Tran, T. V., Alvioli, M., Lee, G., & An, H. U. (2018). Three-dimensional, time-dependent modeling of rainfall-induced landslides over a digital landscape: a case study. Landslides, 15, 1071-1084.
22. Schilirò, L., Cepeda, J., Devoli, G., & Piciullo, L. (2021). Regional analyses of rainfall-induced landslide initiation in Upper Gudbrandsdalen (South-Eastern Norway) using TRIGRS model. Geosciences, 11(1), 35.
23. Song, K., Han, L., Ruan, D., Li, H., & Ma, B. (2023). Stability prediction of rainfall-induced shallow landslides: A case study of mountainous area in China. Water, 15(16), 2938.
24. Iverson, R.M. (2000). Landslide triggering by rain infiltration. Water Resources Research, 36(7), 1897–1910.
25. Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. Water Resources Research, 45(10).
26. Rawls, W. J., & Brakensiek, D. L. (1982). Estimating soil water retention and hydraulic properties. Advances in Soil Science, 1, 213–234.
27. Zhao, Y., Peth, S., Krümmelbein, J., Horn, R., Wang, Z., Steffens, M., ... & Peng, X. (2006). Spatial variability of soil properties affected by grazing intensity in Inner Mongolia grassland. Ecological Modelling, 198(1-2), 178–189.
28. Wegman, K. (2008). Mechanics of materials and soils in mountainous terrain. Journal of Geotechnical Studies, 12(3), 145–158.
29. Cosby, B. J., Hornberger, G. M., Clapp, R. B., & Ginn, T. R. (1984). A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resources Research, 20(6), 682–690.
30. Clapp, R. B., & Hornberger, G. M. (1978). Empirical equations for some soil hydraulic properties. Water Resources Research, 14(4), 601–604.
31. Das, B. M. (2013). Principles of Geotechnical Engineering. 8th Edition, Cengage Learning.
32. Coduto, D.P., Yeung, M.C.R., & Kitch, W.A. (2010). Geotechnical Engineering: Principles and Practices, 2nd Edition, Pearson Education.
33. Vernimmen, R. R. E., Hooijer, A., Aldrian, E., & Van Dijk, A. I. J. M. (2012). Evaluation and bias correction of satellite rainfall data for drought monitoring in Indonesia. Hydrology and Earth System Sciences, 16(1), 133-146.
34. Lafon, T., Dadson, S., Buys, G., & Prudhomme, C. (2013). Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. International journal of climatology, 33(6), 1367-1381.